Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

RICARDO DE FARIAS SANTOS

Detecção de Anomalias Estatísticas nos dados de produção Ambulatorial do SUS

Brasília

2020

RICARDO DE FARIAS SANTOS

Detecção de Anomalias Estatísticas nos dados de produção Ambulatorial do SUS

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle. Realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.
Orientador: Prof. Me. Saul Campos Berardo

Brasília

2020

REFERÊNCIA BIBLIOGRÁFICA

SANTOS, Ricardo de Farias. **Detecção de Anomalias Estatísticas nos dados de produção Ambulatorial do SUS**. 2020. Trabalho de Conclusão de Curso (Especialização em Análise de Dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF.

CESSÃO DE DIREITOS

NOME DO AUTOR: Ricardo de Farias Santos
TÍTULO: Detecção de Anomalias Estatísticas nos dados de produção Ambulatorial do SUS
GRAU/ANO: Especialista/2020

_____

Ricardo de Farias Santos
santosrf@tcu.gov.br

Ficha catalográfica

Santos, Ricardo de Farias
Detecção de Anomalias Estatísticas nos dados de produção Ambulatorial do SUS / Ricardo de Farias Santos; orientador, Saul Campos Berardo, 2020.

103 p.

Monografia (especialização) - Escola Superior do Tribunal de Contas da União, Curso de Especialização em Análise de Dados para o Controle, Brasília, 2020.

Inclui referências.

1.Análise de Dados. 2. Detecção de *Outliers*. 4. Sistema de Informação de Saúde. 5. Metodologia CRISP-DM. I. Escola Superior do Tribunal de Contas da União. Especialização em Análise de Dados para o Controle.

RICARDO DE FARIAS SANTOS

TÍTULO: Detecção de Anomalias Estatísticas nos dados de produção Ambulatorial do SUS

Trabalho de conclusão do curso de pós-graduação lato sensu em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Brasília, 27 de março de 2020.

Banca Examinadora:

_____

Prof. Saul Campos Berardo, Me.

Orientador

Instituto Serzedello Correa

_____

Prof. Edans Flavius de Oliveira Sandes, Dr.

Instituto Serzedello Correa

Dedico este trabalho aos meus avós, Oswaldo e Marina (*in memorian*), que sempre foram grandes incentivadores da minha carreira acadêmica.

AGRADECIMENTOS

Agradeço, antes de tudo, a Deus por me dar a oportunidade de mais uma vez poder enfrentar o desafio do aprendizado e de aumentar o meu conhecimento.

À minha família pela compreensão, carinho e sacrifício das horas e finais de semana em que poderíamos estar juntos e que foram investidas neste curso, em especial à minha esposa Raquel por ter dado todo o apoio necessário para que eu pudesse me dedicar às atividades do curso, à minha filha Marina e ao meu filho Rafael que sempre foram compreensivos por não terem a minha companhia em vários momentos em que não pude estar com eles.

Ao colega, professor e orientador Saul Berardo pelo apoio, condução, sugestões e pelo desafio ao propor este trabalho e ao professores e colegas Edans Sandes e João Batista Rodrigues pelos aconselhamentos técnicos nos momentos de aperto.

Ao meu antigo secretário, Marcelo Eira por ter me apoiado em fazer esta especialização, ao meu atual secretário Carlos Caixeta e chefe Alison Aparecido por todo apoio e pela compreensão.

Aos colegas e amigos de TCU, Walter Venson, Raul Mascarenhas e Eduardo Junio pelos esclarecimentos técnicos na confecção do TCC.

Aos colegas do curso de pós-graduação pela companhia na caminhada nesses dois anos, compartilhando as angústias de finais de semana e as alegrias do sucesso alcançado na busca do conhecimento, em especial à Sarah e ao Marcelo Chaves por terem me acompanhado em temas relacionados ao TCC, ao Helton por ter sido meu colega e amigo fiel de dupla, à Renata e ao Borela pelas trocas conhecimento e ajuda que me deram durante todo o curso.

Aos colegas do Instituto Serzedello Corrêa que utilizaram de todo o seu talento e comprometimento para nos entregar um curso de altíssima qualidade, bem como a todos os responsáveis pela criação deste curso.

A todos os professores do curso pelo seu empenho, dedicação e conhecimento transmitido.

"*Só se pode alcançar um grande êxito quando nos mantemos fiéis a nós mesmos*"

Friedrich Nietzsche (1844 – 1900)

"*Julgue seu sucesso pelas coisas que você teve que renunciar para conseguir. "*

Dalai Lama (1935 – atual)

"*O que sabemos é uma gota,*

o que ignoramos é um oceano."

Isaac Newton (1643 – 1727)

RESUMO

A crescente disponibilização de dados abertos e adoção de políticas de transparência pelos governos de vários países democráticos proporcionou um campo fértil para a utilização de ferramentas de análise de dados. No entanto, o volume de dados que se coloca para análise é gigantesco e há poucas ferramentas que possam ajudar a entendê-lo. Os órgãos que executam atividades de controle, como o Tribunal de Contas da União (TCU) podem fazer uso desses dados para potencializar as suas ações de controle. Diante da impossibilidade de lidar com o grande volume de dados, faz-se necessário utilizar ferramentas que permitam tirar proveito das informações disponibilizadas. Na área da saúde pública, o TCU, por meio da Secretaria de Controle Externo da Saúde (SecexSaúde), realiza auditorias nos dados disponibilizados pel Sistema Único de Saúde (SUS) como o Sistema de Informações Hospitalares (SIH), Sistema de Informações Ambulatoriais (SIA) e o cadastro nacional de Estabelecimentos de Saúde (CNES). Este trabalho tem a intenção de oferecer uma análise de outliers sobre os dados ambulatorias do SIA de forma a servir de ferramenta para o trabalho do auditor. Este trabalho utilizou a metodologia C*ross-Industry Standard Process for Data Mining* (CRISP-DM). Foi realizada a internalização dos dados gerais dos atendimentos ambulatoriais do SIA de 2016 a 2018, limpeza e inclusão de campos, além da aplicação de técnicas de detecção de *outliers* para a análise dos procedimentos. As anomalias foram calculadas em relação aos municípios e aos estabelecimentos para os dados de 2018. Por fim, os dados foram gravados em um banco de dados e apresentados em um painel para consulta. Como resultado, foram analisados 1.054 procedimentos e encontradas anomalias em6.620 casos relacionados aos municípios e 9.938 casos relacionados aos estabelecimentos.

Palavras-chave: Mineração de dados. Análise de dados. Ciência de dados. Detecção de Outliers. Detecção de Anomalias. Teste de Normalidade. CRISP-DM. Sistema Único de Saúde. Sistema de Informação Ambulatorial.

ABSTRACT

The increasing availability of open data and the adoption of transparency policies by the governments of several democratic countries has provided a fertile ground for the adoption of data analysis tools. However, the volume of data that is placed for analysis is gigantic, and a few tools can help to understand it. The auditing organizations such as the Federal Court of Accounts (TCU) can make use of this data to enhance their control actions. Given the impossibility of manually dealing with the large volume of data, it is necessary to use tools that allow taking advantage of the information made available. In the area of public health, TCU, through the Secretariat of External Health Control (SecexSaúde), performs audits on the data made possible by the Brazilian Integrated Health System (SUS) such as the Hospital Information System (SIH), Ambulatory Information System (SIA) and the National Health Establishments Register (CNES). This work intends to offer an analysis of outliers on the ambulatory procedures data of the SIA to serve as a tool for the auditor's work. This work used the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Internalization of the SIA data, cleaning, and inclusion of fields was carried out, in addition to the application of outlier detection techniques for the analysis of the procedures. We calculated anomalies for towns and healthcare facilities. Finally, we recorded the data from 2016 to 2018 in a database and presented on an online panel for utilization by auditors. As a result, we analyzed 1,054 procedures and presented 6,620 cases related to the towns and 9,938 cases related to the healthcare facilities on the online panel.

***Keywords*:** *Data mining. Data analysis. Data science. Outlier detection. Anomaly detection. CRISP-DM, Brazilian Integrated Health System, Ambulatory Information System.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIH – Autorização de Internação Hospitalar

ANS – Agência Nacional de Saúde

APAC – Autorização de Procedimento Ambulatorial

BPA – Boletim de Produção Ambulatorial

BPA-C – Boletim de Produção Ambulatorial – Consolidado

BPA-I – Boletim de Produção Ambulatorial – Individualizado

CADSUS – Sistema de Cadastramento de usuários do SUS

CBO – Classificação Brasileira de Ocupações

CIB – Comissão Intergestores Bipartite

CID-10 – Classificação Internacional de Doenças e Problemas Relacionados à Saúde

CIT – Comissão Intergestores Tripartite

CNES – Cadastro Nacional de Estabelecimentos de Saúde

CNS – Cartão Nacional de Saúde

Conasems – Conselho Nacional de Secretarias Municipais de Saúde

Conass – Conselho Nacional de Secretário de Saúde

Cosems – Conselhos de Secretarias Municipais de Saúde (Cosems)

CRISP-DM – Cross-Industry Standard Process for Data Mining

DATASUS – Departamento de Informática do Sistema Único de Saúde do Brasil

FAEC – Fundo de Ações Estratégicas e Compensação

FPO – Programação Físico-Orçamentaria

GAP – Guia de Autorização de Pagamento

IBGE – Instituto Brasileiro de Geografia e Estatística

IQR – *Interquartile Range* (Amplitude Interquartil)

LOF – *Local Outlier Factor*

MAD – Median of the absolute deviation

OCDE – Organização para a Cooperação e Desenvolvimento Econômico

ONG – Organização Não Governamental

OPM – Órteses, próteses e materiais especiais

PGASS – Programação Geral das Ações e Serviços de Saúde

PPI – Programação Pactuada e Integrada

RAAS – Registro das Ações Ambulatoriais de Saúde

RAAS-AD – Registro das Ações Ambulatoriais de Saúde – Atenção Domiciliar

RAAS-PSI – Registro das Ações de Saúde – Psicossocial

SES – Secretaria Estadual de Saúde

SIA – Sistema de Informações Ambulatoriais

SIAB – Sistema de Informação da Atenção Básica

SICAPS – Sistema de Informação e Controle Ambulatorial da Previdência Social

SIGTAP – Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do

SIH – Sistema de Informações Hospitalares

SIHD – Sistema de Informações Hospitalares Descentralizado

SMS – Secretaria Municipal de Saúde

SQL – Structured Query Language

SUS – Sistema Único de Saúde

TCU – Tribunal de Contas da União

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Nos últimos anos, o volume de dados disponíveis mundialmente cresceu enormemente com o uso de aplicativos em celulares por uma grande parcela da população. Este fenômeno veio acompanhado pela crescente disponibilização de dados abertos e adoção de políticas de transparência pelos governos de vários países democráticos.

No Brasil, o advento da Lei da Transparência e da Lei de Acesso à Informação promoveram o acesso do cidadão a dados da gestão e dos gastos do Governo. No entanto, o volume de dados que se coloca para análise é gigantesco e há poucas ferramentas que possam ajudar a entendê-lo. Este não é um problema só para o cidadão como também para os órgãos que executam atividades de controle, como o Tribunal de Contas da União (TCU).

Como é humanamente impossível lidar com a grande massa de dados, faz-se necessário utilizar ferramentas que apoiem o auditor na tarefa de análise, permitindo tirar proveito das informações disponibilizadas.

No âmbito da temática de saúde pública, o TCU, por meio da Secretaria de Controle Externo da Saúde (SecexSaúde), realiza auditorias no Sistema Único de Saúde (SUS). Uma das fontes de informação para o trabalho desta unidade são os Sistemas de Informações Hospitalares (SIH), Ambulatoriais (SIA) e o cadastro nacional de Estabelecimentos de Saúde (CNES). Atualmente não há nenhuma ferramenta que apoie o trabalho daquela unidade na exploração dessas bases de dados para a identificação de indícios de irregularidades na produção registrada (e, consequentemente, na remuneração de prestadores de serviços vinculados ao SUS); e identificação de problemas de qualidade de dados.

Soma-se a este contexto, o fato de que houve nos últimos anos a aposentadoria de aproximadamente 10% do pessoal do TCU sem a devida reposição do quadro por conta das restrições orçamentárias. Portanto, toda ferramenta que aumente a produtividade dos auditores por meio de automatização de tarefas e aumento da eficiência é bem-vinda.

O documento está organizado da seguinte forma. Seção 2 – Metodologia Adotada no Projeto. Seção 3 – Fase de Entendimento do Negócio. Seção 4 – Fase de Entendimento dos Dados. Seção 5 – Fase de Preparação dos Dados. Seção 6 – Fase de Modelagem. Seção 7 – Fase de Avaliação. Seção 8 – Validação em um Estudo de Caso. Seção 9 – Conclusão.

1.2 PROBLEMA E JUSTIFICATIVA

A promulgação da Constituição Federal de 1988 trouxe em seu texto a criação do SUS – Sistema Único de Saúde como o sistema público de saúde no Brasil. A ideia era oferecer à população brasileira o acesso universal ao cuidado em saúde a ser financiado pelo orçamento do Governo Federal, dos Estados, do Distrito Federal e dos Municípios.

Portanto, toda a população do Brasil, aproximadamente 211 milhões de habitantes (IBGE, 2020), está coberta pelo SUS. Assim, a atenção à saúde da população gera um grande volume de dados sobre os serviços de saúde públicos oferecidos o que demanda a automatização da análise para fins de auditoria. Para se ter uma ideia da ordem de grandeza do sistema, foram mais de 3,7 bilhões de atendimentos ambulatoriais em 2019 com o financiamento do SUS, cujos dados detalhados podem ser obtidos publicamente.

Esses dados podem ser utilizados para suportar as atividades de auditoria do Tribunal de Contas da União no âmbito das atividades relativas à área de saúde. Conforme a portaria PORTARIA-SEGECEX Nº 3, DE 14 DE JANEIRO DE 2019, a Secex-Saúde é a secretaria do Tribunal de Contas da União que é responsável pelas ações de controle relacionadas à área de saúde na Esfera Federal. Como parte integrante da sua estrutura tem-se o Núcleo de Tratamento de Dados e Informações, cuja competência, segundo o Art. 4º, inciso IV da PORTARIA-SECEXSAÚDE N. 3, DE 10 DE JUNHO DE 2019 (BRASIL, 2019), é tratar os dados dos bancos de dados do Ministério da Saúde e outros, bem como propor o encaminhamento dos resultados obtidos.

Dentre os bancos de dados de interesse desta Secretaria, estão os bancos de dados relativos ao SUS. Tais bases, são numerosas e provenientes de várias soluções, muitas vezes não integradas entre si o que acaba dificultando a análise dos dados provenientes delas. Dentre elas pode-se citar o SIA/SUS – Sistema de Informações Ambulatoriais do SUS, SIH/SUS – Sistema de Informações Hospitalares do SUS e o CNES – Cadastro Nacional de Estabelecimentos de Saúde. Neste trabalho pretende-se trabalhar com a base de atendimentos ambulatoriais, o SIA/SUS.

O SIA/SUS é parte de uma solução que foi implantada em todo o território nacional em 1995 e que permite aos gestores locais o processamento das informações de atendimento ambulatorial registrados nos aplicativos de captação do atendimento ambulatorial pelos prestadores públicos e privados contratados ou conveniados pelo SUS.

A intenção do presente trabalho é estudar, documentar, internalizar, bem como analisar os dados do SIA/SUS em conjunto com os dados do CNES (Cadastro Nacional de

Estabelecimentos de Saúde) e do IBGE com a finalidade de encontrar anomalias estatísticas que possam servir de insumo aos trabalhos de auditoria realizados pela SecexSaúde no TCU.

1.3 OBJETIVOS GERAIS

Auxiliar a SecexSaúde do TCU a avaliar e levantar indícios de objetos que possam ser interessantes ao controle realizado por aquela Secretaria no que tange à produção do SUS relacionada ao atendimento ambulatorial

1.4 OBJETIVOS ESPECÍFICOS

1. Internalização dos dados do SIA/SUS e demais bases de dados relacionadas (ex.: CNES e IBGE);

2. Documentar os dados internalizados para disseminar e facilitar o seu uso por parte dos auditores;

3. Realizar análise exploratória dos dados do SIA;

4. Realizar prova de conceito com técnicas de mineração de dados para avaliar a viabilidade da identificação de discrepâncias estatísticas nos registros de internações ambulatoriais do SUS;

5. Construir um painel por meio do qual seja possível ao auditor da área de saúde avaliar as anomalias identificadas.

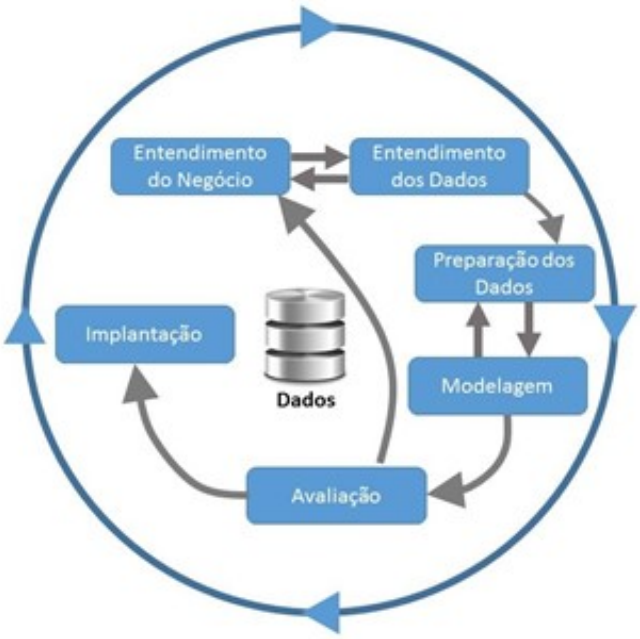
2 METODOLOGIA ADOTADA NO PROJETO

Para desenvolver este trabalho utilizou-se a metodologia de referência *Cross-Industry Standard Process for Data Mining* (CRISP-DM), comumente utilizada para projetos de mineração de dados (CHAPMAN, 2000) e que vem sendo utilizada por mais de uma década como um padrão para processos de análise de dados. Trata-se de um padrão aberto, que pode ser usado livremente, desenvolvido por um consórcio europeu formado por mais de 200 organizações e financiado pela União Europeia,

A sequência de fases não deve ser interpretada de maneira rigorosa e há sempre a possibilidade de livre movimentação entre elas, de acordo com a necessidade de cada projeto. O modelo é flexível e pode ser facilmente customizado (IBM, 2019b). Por exemplo, se a organização visa detectar casos de lavagem de dinheiro, é provável que seus analistas vão vasculhar enormes quantidades de dados sem determinar um objetivo de modelagem específico. Ao invés de modelagem, o trabalho irá focar na exploração e visualização de dados para revelar padrões suspeitos em dados. Em tal tipo de situação, as fases de modelagem, avaliação e implantação podem vir a ser menos relevantes que as fases de entendimento e preparação de dados. No entanto, ainda assim é importante considerar algumas questões levantadas durante estas fases em termos de planejamento de longo prazo e para objetivos de mineração futuros.

O ciclo de vida da metodologia é apresentado na Figura 1. O círculo exterior no diagrama simboliza a natureza cíclica da atividade de mineração de dados, cujo processo continua após a implantação da solução. As setas representadas no diagrama de processo indicam as dependências mais importantes e frequentes entre as fases. As lições aprendidas durante o processo podem disparar novas questões mais focadas no negócio, bem como beneficiar o desenvolvimento de novos processos de mineração.

Figura 1. O ciclo de vida do CRISP-DM



Fonte: IBM (2019a).

A fase de **entendimento do negócio** foca na compreensão dos objetivos e requisitos do projeto a partir da perspectiva de negócio de forma a definir um problema de mineração de dados e um plano preliminar do projeto.

A fase de **entendimento dos dados** começa pela coleta dos dados e prossegue com atividades com a finalidade de tornar-se familiar com os dados, identificar problemas com a qualidade dos dados, descobrir os primeiros insights nos dados ou ainda detectar subconjuntos interessantes para formar hipóteses para a informação que está sendo avaliada.

De acordo com o CRISP-DM, a fase de **preparação dos dados** cobre todas as atividades envolvidas com a construção do conjunto de dados final obtido a partir dos dados brutos originais, de forma a prepará-los para o processamento desejado. As tarefas de preparação de dados serão executadas várias vezes e não precisam seguir nenhuma ordem pré-determinada. Dentre elas, podemos listar: detecção de dados faltantes e incoerentes, levantamento de dados redundantes, levantamento de campos importantes para a análise e consequente exclusão de dados que não serão utilizados, agregação de dados, derivação de novos campos, integração com outras bases, etc. Uma boa preparação dos dados é crucial para que o resultado do projeto seja válido e confiável, além de economizar tempo e esforço.

Após a preparação dos dados, passa-se à fase de **modelagem**. Várias técnicas de modelagem de dados são selecionadas, aplicadas e validadas. Como cada técnica possui suas

exigências no formato dos dados a serem utilizados, seus parâmetros são calibrados para os valores adequados após vários testes. Portanto, revisitar a fase de preparação dos dados pode ser necessário algumas vezes. Exemplos de técnicas são: classificação, agrupamento ou regressão.

Na fase de **avaliação**, o projeto terá um ou mais modelos bem desenvolvidos. Assim, esta fase se dedica a avaliar a qualidade dos modelos desenvolvidos sob a perspectiva da análise de dados, bem como verificar se eles atendem adequadamente aos objetivos do negócio.

Na fase de **implantação**, o modelo é colocado à disposição da organização para que possa gerar os resultados planejados

3 FASE DE ENTENDIMENTO DO NEGÓCIO

Dado o enorme volume de dados da área de saúde disponível para análise, é impossível realizar o seu tratamento de forma manual. Assim, a SecexSaúde necessita de ferramentas que possam automatizar esta tarefa. Como exemplo de ferramenta, pode-se citar o projeto InfoSAS que foi uma tentativa do Ministério da Saúde em conjunto com a Universidade Federal de Minas Gerais – UFMG para detectar anomalias em dados de internações hospitalares e de atendimentos ambulatoriais (CARVALHO, 2016).

O InfoSAS é um sistema de detecção de anomalias estatísticas nos registros da produção do SUS. O seu objetivo é encontrar taxas de atendimentos por habitante muito superiores à média nacional, ou valores de internação bem acima dos praticados pela maioria dos estabelecimentos para um mesmo procedimento. Os resultados encontrados indicam que centenas de milhões de reais gastos pelo SUS são destinados a atendimentos considerados anômalos por critérios conservadores. Anomalias estatísticas podem ser provocadas por fraudes, mas também por mutirões de saúde, epidemias, ou má distribuição do atendimento. Em qualquer caso, anomalias graves devem ser investigadas ou explicadas.

Para se ter ideia da ordem de grandeza do problema, se considerarmos o número de procedimentos possíveis (mais de 5000) que podem ser realizados por aproximadamente 6000 prestadores e, considerando apenas períodos de 12 meses, em 3 anos de produção, seriam 36 janelas de tempo possíveis. Um cálculo simples mostra que se tem, literalmente, bilhões de possibilidades de análise das bases de dados examinadas. O InfoSAS utiliza diversos algoritmos que procuram capturar discrepâncias, produzindo escores que permitem ordenação e priorização dos achados.

Para a seleção, o InfoSAS também permite ao usuário concentrar-se em áreas definidas por filtros geográficos, por período de análise e por alvos, pois um profissional de controle e

avaliação tem, muitas vezes, sua atenção dirigida para setores específicos da saúde, como cardiologia ou ortopedia e, claro, maior interesse em sua região de atuação.

O InfoSAS analisa as séries temporais de valor médio mensal por procedimento e de produção mensal em cada alvo desejado. Tais séries são calculadas por estabelecimento e por município de residência dos pacientes. Vários algoritmos de detecção de anomalias são utilizados e cada um deles calcula um escore que são posteriormente combinados em forma de resultado a ser utilizado para permitir o planejamento de ações de controle.

Esta abordagem viabiliza a utilização das bases de dados disponíveis na área de saúde para levantar casos, dentre as centenas de milhões que foram registrados, que merecem ser investigados de maneira mais cuidadosa pela área responsável.

No entanto, o projeto do InfoSAS não possui código aberto e, por causa disso, não pode ser adaptado para uso pelo TCU. Assim, foi utilizado como inspiração para este trabalho de tal forma a criar uma prova de conceito de ferramenta que possa auxiliar a SecexSaúde na análise dos dados ambulatoriais.

Diante disto, propôs-se tentar reproduzir produto similar que pudesse realizar análise dos dados das bases de dados de internação hospitalar, de atendimento ambulatorial e do cadastro de estabelecimentos do SUS com a intenção de encontrar discrepâncias estatísticas que pudessem servir de insumo para o trabalho da secretaria. Cada base seria analisada sob aspectos diferentes em três trabalhos diferentes. O presente trabalho focou na análise de anomalias estatísticas na base do SIA/SUS.

Nesta fase, foram realizadas as seguintes atividades:

a) Levantamento da legislação disponível para entendimento do SUS e sua estrutura;
b) Entendimento do funcionamento das plataformas do SUS relacionadas com o SIA, SIH e CNES;
c) Reuniões formais com o responsável pelo núcleo de dados da SecexSaúde;
d) Reunião com os responsáveis pelas bases de dados no Datasus.

3.1 O SISTEMA ÚNICO DE SAÚDE - SUS

O Sistema Único de Saúde (SUS) é um dos maiores e mais complexos sistemas de saúde pública do mundo, abrangendo desde o simples atendimento para avaliação da pressão arterial, por meio da Atenção Primária, até o transplante de órgãos, garantindo acesso integral, universal e gratuito para toda a população do país (BRASIL, MS, 2020). Composto por uma rede de atenção que assiste aos 27 entes federados e aos 5.570 municípios brasileiros e

prestando serviços a uma população estimada em 211.434.000 milhões de pessoas (IBGE, 2020), o SUS é considerado uma das mais amplas e importantes experiências de atenção à saúde no mundo. Com a sua criação, ele proporcionou o acesso universal ao sistema público de saúde, sem discriminação. Assim, atenção integral à saúde, e não somente aos cuidados assistenciais, passou a ser um direito de todos os brasileiros, desde a gestação e por toda a vida, com foco na saúde com qualidade de vida, visando a prevenção e a promoção da saúde.

A gestão das ações e dos serviços de saúde deve ser solidária e participativa entre os três entes da Federação: a União, os Estados e os municípios. A rede que compõe o SUS é ampla e abrange tanto ações quanto os serviços de saúde. Engloba a atenção primária, média e alta complexidades, os serviços urgência e emergência, a atenção hospitalar, as ações e serviços das vigilâncias epidemiológica, sanitária e ambiental e assistência farmacêutica.

3.1.1 Estrutura do SUS

O Sistema Único de Saúde (SUS) é composto pelo Ministério da Saúde, Estados e Municípios, conforme determina a Constituição Federal. Cada ente tem suas corresponsabilidades. (Figura 2)

Ministério da Saúde: gestor nacional do SUS, formula, normatiza, fiscaliza, monitora e avalia políticas e ações, em articulação com o Conselho Nacional de Saúde. Atua no âmbito da Comissão Intergestores Tripartite (CIT) para pactuar o Plano Nacional de Saúde. Integram sua estrutura, dentre outros: Fiocruz, Funasa, Anvisa, ANS, Hemobrás, Inca e Into.

Secretaria Estadual de Saúde (SES): participa da formulação das políticas e ações de saúde, presta apoio aos municípios em articulação com o conselho estadual e participa da Comissão Intergestores Bipartite (CIB) para aprovar e implementar o plano estadual de saúde.

Secretaria Municipal de Saúde (SMS): planeja, organiza, controla, avalia e executa as ações e serviços de saúde em articulação com o conselho municipal e a esfera estadual para aprovar e implantar o plano municipal de saúde.

Conselhos de Saúde: o Conselho de Saúde, no âmbito de atuação (Nacional, Estadual ou Municipal), em caráter permanente e deliberativo, órgão colegiado composto por representantes do governo, prestadores de serviço, profissionais de saúde e usuários, atua na formulação de estratégias e no controle da execução da política de saúde na instância correspondente, inclusive nos aspectos econômicos e financeiros, cujas decisões serão homologadas pelo chefe do poder legalmente constituído em cada esfera do governo.

Comissão Intergestores Tripartite (CIT): foro de negociação e pactuação entre gestores federal, estadual e municipal, quanto aos aspectos operacionais do SUS.

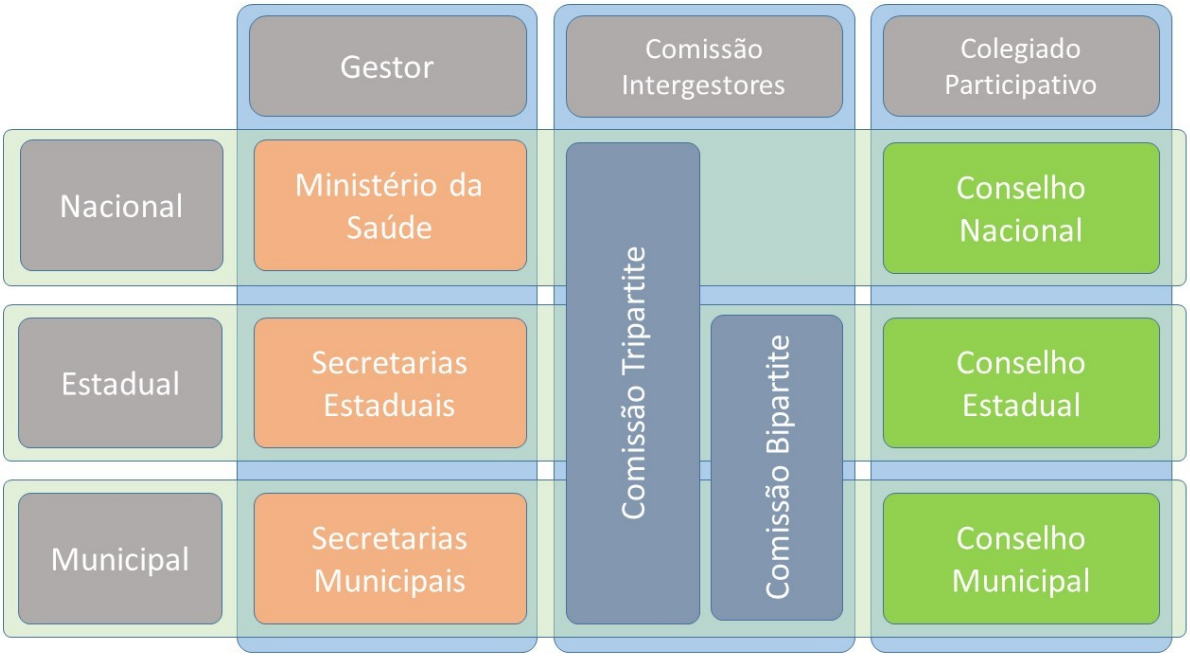
Comissão Intergestores Bipartite (CIB): foro de negociação e pactuação entre gestores estadual e municipais, quanto aos aspectos operacionais do SUS.

Conselho Nacional de Secretário da Saúde (Conass): entidade representativa dos entes estaduais e do Distrito Federal na CIT para tratar de matérias referentes à saúde.

Conselho Nacional de Secretarias Municipais de Saúde (Conasems): entidade representativa dos entes municipais na CIT para tratar de matérias referentes à saúde.

Conselhos de Secretarias Municipais de Saúde (Cosems): são reconhecidos como entidades que representam os entes municipais, no âmbito estadual, para tratar de matérias referentes à saúde, desde que vinculados institucionalmente ao Conasems, na forma que dispuserem seus estatutos

Figura 2. Estrutura do SUS



Fonte: Elaborado pelo Autor (2019), adaptado de SOUZA (2019).

3.1.2 Princípios do SUS

Universalização: a saúde é um direito de cidadania de todas as pessoas e cabe ao Estado assegurar este direito, sendo que o acesso às ações e serviços deve ser garantido a todas as pessoas, independentemente de sexo, raça, ocupação ou outras características sociais ou pessoais.

Equidade: o objetivo desse princípio é diminuir desigualdades. Apesar de todas as pessoas possuírem direito aos serviços, as pessoas não são iguais e, por isso, têm necessidades distintas. Em outras palavras, equidade significa tratar desigualmente os desiguais, investindo mais onde a carência é maior.

Integralidade: este princípio considera as pessoas como um todo, atendendo a todas as suas necessidades. Para isso, é importante a integração de ações, incluindo a promoção da saúde, a prevenção de doenças, o tratamento e a reabilitação. Juntamente, o princípio de integralidade pressupõe a articulação da saúde com outras políticas públicas, para assegurar uma atuação intersetorial entre as diferentes áreas que tenham repercussão na saúde e qualidade de vida dos indivíduos.

3.1.3 Responsabilidade dos Entes que compõem o SUS

3.1.3.1 União

A gestão federal da saúde é realizada por meio do Ministério da Saúde. O governo federal é o principal financiador da rede pública de saúde. Historicamente, o Ministério da Saúde aplica metade de todos os recursos gastos no país em saúde pública em todo o Brasil, e estados e municípios, em geral, contribuem com a outra metade dos recursos. O Ministério da Saúde formula políticas nacionais de saúde, mas não realiza as ações. Para a realização dos projetos, depende de seus parceiros (estados, municípios, ONGs, fundações, empresas, etc.). Também tem a função de planejar, elaborar normas, avaliar e utilizar instrumentos para o controle do SUS.

3.1.3.2 Estados e Distrito Federal

Os estados possuem secretarias específicas para a gestão de saúde. O gestor estadual deve aplicar recursos próprios, inclusive nos municípios, e os repassados pela União. Além de ser um dos parceiros para a aplicação de políticas nacionais de saúde, o estado formula suas próprias políticas de saúde. Ele coordena e planeja o SUS em nível estadual, respeitando a normatização federal. Os gestores estaduais são responsáveis pela organização do atendimento à saúde em seu território.

3.1.3.3 Municípios

São responsáveis pela execução das ações e serviços de saúde no âmbito do seu território. O gestor municipal deve aplicar recursos próprios e os repassados pela União e pelo estado. O município formula suas próprias políticas de saúde e também é um dos parceiros para a aplicação de políticas nacionais e estaduais de saúde. Ele coordena e planeja o SUS em nível

municipal, respeitando a normatização federal. Pode estabelecer parcerias com outros municípios para garantir o atendimento pleno de sua população, para procedimentos de complexidade que estejam acima daqueles que pode oferecer.

3.2 BASES DE DADOS DE SAÚDE NO BRASIL

O DATASUS é o órgão do governo que gerencia as bases de dados relacionadas à saúde pública. Dentre as informações disponibilizadas por ele temos dados sobre internações hospitalares, atendimentos ambulatoriais, procedimentos de alta complexidade, imunização, óbitos, nascimentos, vigilância em saúde, dentre outros. Em alguns casos, o histórico alcança mais de três décadas de atendimento.

A produção ambulatorial e hospitalar no país financiada pelo SUS é registrada de forma detalhada nos sistemas de dados médico administrativos do SUS: SIH – Sistema de Informação Hospitalar e SIA – Sistema de Informação Ambulatorial. Muito embora o registro seja feito visando principalmente a gestão da remuneração dos prestadores, os dados são utilizados como fonte para diversos tipos de pesquisas nos mais variados campos como gastos, oferta e demanda de serviços, cobertura, e também na construção de indicadores.

O SIH é o sistema que armazena dados de todas as internações realizadas no SUS cujo instrumento de coleta é a AIH – Autorização de Internação Hospitalar (BRASIL, MS, 2012a). O SIA, por sua vez, tem como finalidade registrar os atendimentos, procedimentos e tratamentos realizados em cada estabelecimento de saúde no âmbito ambulatorial e que tem como fonte de coleta o BPA – Boletim de Procedimento Ambulatorial, APAC – Autorização de Procedimentos Ambulatoriais e o RAAS – Registro de Ações Ambulatoriais em Saúde (BRASIL, MS, 2010).

Além desses, vale mencionar o CNES – Cadastro Nacional de Estabelecimentos de Saúde e o CADSUS – Cadastro dos Usuários do SUS. O primeiro gerencia todos os estabelecimentos de saúde no país com informações de cada estabelecimento como quantidade de leitos, equipamentos, profissionais disponíveis e especialidades atendidas. O segundo armazena todos os usuários do SUS, atribuindo um número a cada pessoa, conhecido como CNS – Cartão Nacional de Saúde. Este número é atribuído aos usuários, mesmo que usem planos privados de saúde. A ideia é facilitar a inclusão destes usuários no CADSUS. Por razões de sigilo sobre informações pessoais, este banco de dados não é aberto para download.

Há ainda um sistema que registra todos os procedimentos e materiais, que é utilizada na codificação dos serviços de saúde prestados no setor público, a SIGTAP – Sistema de Gerenciamento da Tabela de Procedimento, Medicamentos e OPM do SUS.

Além das soluções e bases já citadas, há uma miríade de outras soluções com bases de dados da saúde que não foram citadas neste trabalho, pois não serão utilizadas. No entanto, uma boa documentação sobre elas pode ser consultada em LIMA (2016).

4 FASE DE ENTENDIMENTO DOS DADOS

4.1 SIA/SUS – SISTEMA DE INFORMAÇÕES AMBULATORIAIS

O SIA/SUS foi criado em 1992 e implantado a partir de julho de 1994, nas Secretarias Estaduais com a finalidade de substituir os sistemas GAP – Guia de Autorização de pagamento e SICAPS – Sistema de Informações e Controle Ambulatorial da Previdência Social na gestão do financiamento dos atendimentos ambulatoriais. Em 1996 foi largamente implantado nas Secretarias Municipais de Saúde. Foi desenvolvido em FoxPro (plataforma 16 bits) e usa arquivos DBF.

Ele é o sistema responsável por processar as informações de atendimento ambulatorial no nível municipal e estadual (BRASIL, MS, 2012b). Os dados são registrados nos aplicativos de captação dos diferentes tipos de atendimento ambulatorial: APAC – Autorização de Procedimentos Ambulatoriais, BPA – Boletim de Produção Ambulatorial e RAAS – Registro das Ações Ambulatoriais de Saúde. Este atendimento é realizado por prestadores públicos e privados contratados ou conveniados pelos SUS.

O SIA/SUS recebe a transcrição de produção nos documentos correspondentes, faz consolidação, valida o pagamento contra parâmetros orçamentários estipulados pelo próprio gestor de saúde, antes de aprovar o pagamento – para isto utiliza-se do sistema FPO – Ficha de Programação Físico-Orçamentária.

As informações obtidas a partir do SIA são utilizadas como um instrumento de gestão que subsidia as ações de planejamento, programação, regulação, avaliação, controle e auditoria da assistência ambulatorial. Além disso, as informações possibilitam o acompanhamento e a análise da evolução dos gastos referentes à assistência ambulatorial, oferece subsídios para avaliação quantitativa e qualitativa das ações de saúde, além de poderem ser usadas para os processos da PGASS – Programação Geral das Ações e Serviços de Saúde.

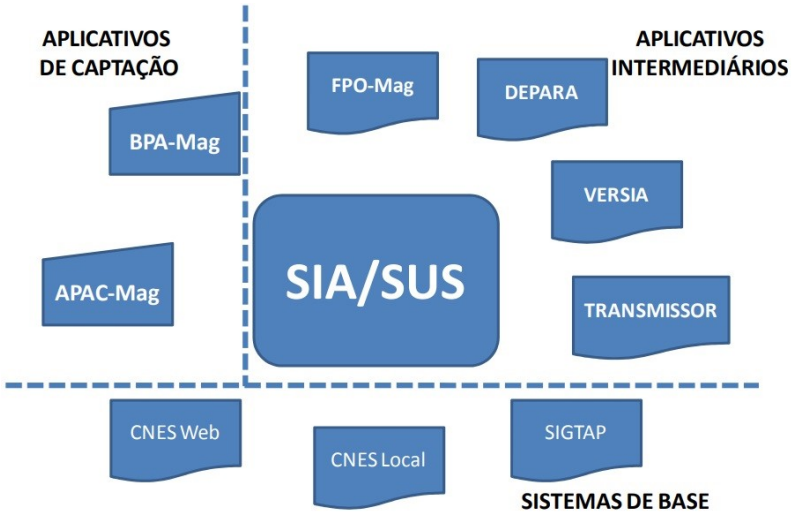
4.2 ESTRUTURA DO SIA/SUS

Para realizar a captação e processamento dos dados da produção ambulatoria do SUS, classificam-se os sistemas como:

- Sistema de processamento,
- Sistemas de base,
- Aplicativos de captação do atendimento, e
- Aplicativos intermediários.

A interação entre esses aplicativos pode ser visualizada na Figura 3

Figura 3. Estrutura do SIA/SUS



Fonte: BRASIL, MS, (2012b)

4.2.1 Aplicativos de Captação

BPA magnético (Boletim de Produção Ambulatorial): Aplicativo para entrada de dados referentes ao atendimento ambulatorial. Nele são registrados os procedimentos de atenção básica (AB) e média complexidade (MC). Dependendo do tipo de procedimento realizado, ele será preenchido no módulo de BPA consolidado (BPA-C) ou no módulo de BPA individualizado (BPA-I).

APAC magnética (Autorização de Procedimentos Ambulatoriais): Aplicativo de entrada de dados dos procedimentos ambulatoriais que exigem autorização prévia do gestor local para que o estabelecimento possa realizá-lo. Portanto, é o módulo onde são digitados a maioria dos procedimentos de alto custo (AC).

RAAS (Registro de Ações Ambulatoriais de Saúde): Aplicativo usado para a entrada de dados dos atendimentos relacionados com procedimentos de atenção psicossocial que são financiados por meio de incentivos da política das Redes de Atenção à Saúde.

4.2.2 Sistemas de Base

CNES (Cadastro Nacional dos Estabelecimentos de Saúde): É a ferramenta oficial do Ministério da Saúde para gerenciar o cadastro das informações de todos os estabelecimentos de saúde e de profissionais prestadores de serviço ao SUS, públicos ou privados. Neste sistema estão listados dados dos estabelecimentos de saúde concernentes à área física, recursos humanos, equipamentos, profissionais e serviços ambulatoriais e hospitalares.

SIGTAP (Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS): É a solução encarregada de gerenciar a tabela de procedimentos do SUS, com todos seus atributos aos estabelecimentos de saúde credenciados ao SUS.

4.2.3 Aplicativos Intermediários

FPO Magnético: aplicativo usado pelo gestor local para registrar a programação física orçamentária ambulatorial de cada estabelecimento de saúde conveniado ao SUS. A programação deve estar coerente com a Programação Pactuada e Integrada – PPI e baseada em contrato ou convênio com o SUS. Suas principais funcionalidades são:

- Registrar a programação físico orçamentária por grupo, subgrupo, forma de organização e/ou procedimento;
- Informar o limite financeiro de acordo com o tipo de financiamento (Atenção Básica – PAB, Média e Alta Complexidade – MAC, Fundo de Ações Estratégicas e Compensação – FAEC)
- Registrar os valores para os procedimentos de atenção básica;
- Importar e exportar dados.

DE PARA: efetua a comunicação do CNES com o SIA permitindo a inclusão de informações dos estabelecimentos de saúde nos bancos de dados do SIA para execução do processamento.

VERSIA: critica a produção processada pelo SIA e gera remessa da produção aprovada para o DATASUS.

BDSIA: pacote de tabelas com as atualizações mensais dos bancos de dados do SIA, principalmente os procedimentos do SIGTAP.

TRANSMISSOR: envia os arquivos de banco de dados dos sistemas (CIH, SIA/SUS, SIHD, SIAB) para alimentar o Banco de Dados Nacional desses Sistemas de Informação.

4.2.4 Sistemas de Processamento

SIA/SUS (Sistema de Informação Ambulatorial): é o sistema responsável pelo processamento da informação de produção ambulatorial da Atenção Básica e Especializada realizada pelos prestadores do SUS e foco do presente trabalho. Ele tem como principais funcionalidades:

- Importação do cadastro dos estabelecimentos de saúde por meio do "DEPARA";
- Importação da programação física orçamentária efetuada pelo FPO/Magnético;
- Importação da produção ambulatorial originária do BPA-Mag, APAC-Mag ou RAAS;
- Processamento da produção ambulatorial comparando os dados com os atributos da tabela de procedimentos; bem como com os dados de cadastro e da programação orçamentária;
- Consolidção da informação ambulatorial para posterior disseminação por meio do DATASUS;
- Geração de relatórios com valores brutos para o setor financeiro.

4.2.5 Instrumentos de entrada de dados

Para o presente trabalho, é importante entendermos como funciona a lógica por trás dos instrumentos de entrada de dados ambulatoriais. Como já mostrado, os dados sobre os procedimentos são cadastrados por meio dos instrumentos BPA-C, BPA-I, APAC e RAAS.

O BPA-C permite o cadastro consolidado de procedimentos. Assim, para estes casos, os registros não necessitam ser individualizados. Via de regra, digita-se o mês e ano de competência, o código do procedimento, o CBO[1] do profissional que realizou o procedimento e a quantidade de procedimentos realizados. Como o cadastro é consolidado, alguns campos perdem detalhamento da informação, o que é o caso de idade e sexo, que são gravados como 999 e 0, respectivamente. Por exemplo, o procedimento 0301100012 – Administração de

[1] CBO – Classificação Brasileira de Ocupações: norma de classificação numerativa e descritiva de atividades econômicas e profissionais determinada pela Comissão Nacional de Classificação para o uso de órgãos governamentais. Cada procedimento do SUS determina quais CBOs podem realizá-lo. Alguns procedimentos só podem ser realizados por médicos. Outros, por quaisquer outros profissionais de saúde.

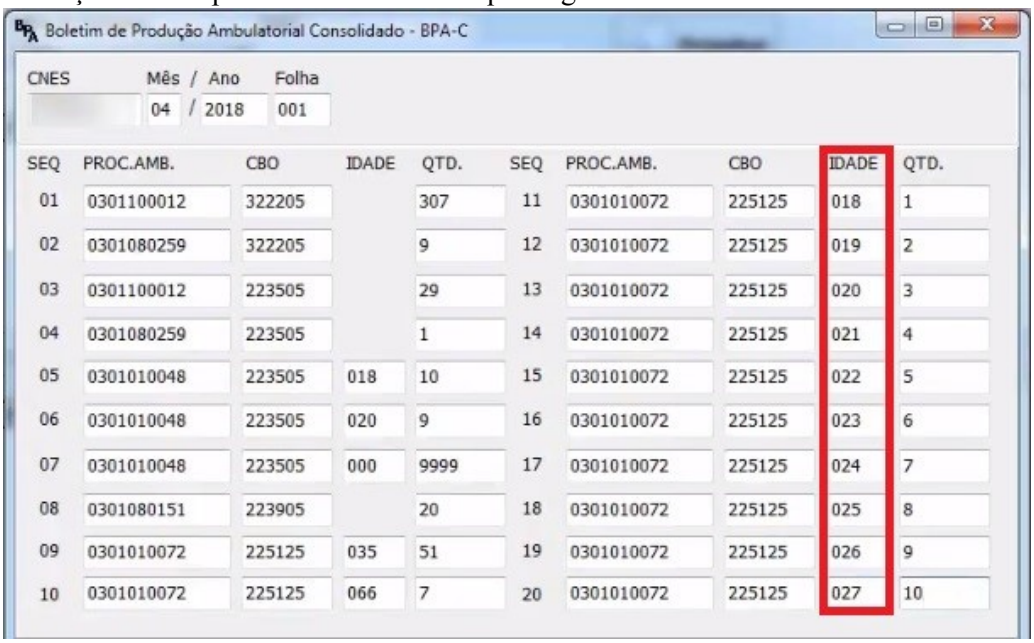
Medicamentos na Atenção Especializada (Por Paciente) pode ser preenchido conforme indicado na figura a seguir.

Figura 4. Preenchimento dos dados do BPA-C para o procedimento 0301100012 mostrando o campo idade desabilitado e CBOs diferentes

Fonte: BRASIL, MS, (2012a)

Em alguns casos, dependendo do procedimento, é necessário preencher a idade do paciente, mesmo se tratando de um registro consolidado. Nestes casos, o registro de quantitativo será agrupado pela idade dos usuários atendidos, como pode ser visto a seguir.

Figura 5. Lançamento de procedimento BPA-C que exige a idade.



Fonte: BRASIL, MS, (2012a)

Os procedimentos que exigem a idade são determinados por meio de portarias do Ministério da Saúde. Alguns exemplos podem ser vistos na Tabela 1, a seguir.

Tabela 1. Exemplos de procedimentos BPA-C que exigem o lançamento da idade

CÓDIGO	DESCRIÇÃO
03.01.01.001-3	Consulta ao usuário curado de tuberculose tratamento supervisionado
03.01.01.002-1	Consulta com identificação de casos novos de tuberculose
03.01.01.003-0	Consulta de profissionais de nível superior na atenção básica

Fonte: Elaborada pelo Autor (2020)

O BPA-I, por sua vez, já conta com dados individualizados como o CNS do paciente, o CNS do profissional de saúde, a data exata de atendimento, o CID10, município de residência, dentre outros. O número da autorização é facultativo para este tipo de instrumento. Tanto o BPA-C quanto o BPA-I não exigem autorização prévia.

Para o lançamento de procedimento APAC é necessário ter autorização prévia do gestor de saúde. Alguns destes permitem o lançamento de procedimentos secundários que não exigem autorização. Por se tratarem de procedimentos mais complexos, há o envio de informação complementar sobre os procedimentos. Trata-se, portanto, de uma ótima fonte de dados para pesquisa. Por exemplo, em no caso de radioterapia há dados sobre as localizações das lesões irradiadas, nomes dos locais do tumor primário ou metastático com o respectivo código CID. No caso de nefrologia tem-se informações sobre hemoglobina, glicose, albumina e índice de massa corpórea.

O RAAS, por sua vez, é o instrumento utilizado para entrada de dados referentes as ações ambulatoriais de saúde de atenção domiciliar e psicossocial, para fins de monitoramento e repasse de recursos. Anteriormente, estes procedimentos eram considerados como parte da APAC. No entanto, a partir de 2012, foram migrados para RAAS por ocasião da publicação da Portaria nº 276, de 30 de março de 2012 com o objetivo de incluir as necessidades relacionadas ao monitoramento das ações e serviços de saúde conformados em Redes de Atenção à Saúde. Os dados informados no RAAS são bem detalhados e contam com os dados pessoais do paciente como nome, data de nascimento, telefone celular, etnia, além do CID principal, CNS do profissional de saúde e CNES do estabelecimento, dentre outros.

Para este trabalho foram considerados os dados dos BPA-Cs, BPA-Is e APACs.

4.3 OBTENÇÃO DOS ARQUIVOS COM DADOS AMBULATORIAIS

O DATASUS disponibiliza toda a Produção Ambulatorial do SUS em uma única base de dados chamada Produção Ambulatorial, que contém todas as informações referentes aos atendimentos ambulatoriais registrados através do Boletim de Produção Ambulatorial, APAC e RAAS. Os arquivos são disponibilizados no formato .dbc no servidor FTP do SUS no endereço ftp://ftp.datasus.gov.br. Este servidor aceita conexões anônimas, não sendo necessário o cadastramento de usuário específico para o acesso aos dados.

Cada arquivo de produção ambulatorial (chamaremos daqui para frente de "arquivo PA") traz a produção ambulatorial de uma UF em um determinado mês. Desta forma, cada arquivo é nomeado segundo o padrão PAUFMMAA.dbc, onde:

- PA: é um texto fixo e significa Procedimento Ambulatorial
- UF: sigla da UF originária dos dados
- MM: mês de competência dos dados
- AA: ano de competência dos dados

É importante mencionar que há exceções a esta regra. No caso do estado de São Paulo (SP), os nomes dos arquivos acrescentam uma letra minúscula antes da extensão (PASP0120a.dbc, PASP0120b.dbc, etc.). Este detalhe não foi encontrado em nenhuma parte da documentação do DATASUS e deve ser considerado pela rotina de download dos arquivos para que ela funcione corretamente. Entende-se que esta solução foi adotada por causa dos tamanhos dos arquivos daquela UF serem muito grandes.

Para desenvolver o módulo de download dos arquivos foi feita uma pesquisa das opções disponíveis. A primeira possibilidade encontrada foi baixar os arquivos com um cliente de FTP (e.g.: FileZilla). Na pesquisa também foi encontrada uma biblioteca READ.DBC em R (PETRUZALEK, 2016) e outra em Python, denominada PySUS (PYSUS, 2001). Como a linguagem Python (PYTHON, 2019) foi dominante durante o curso, decidiu-se por utilizar esta última forma.

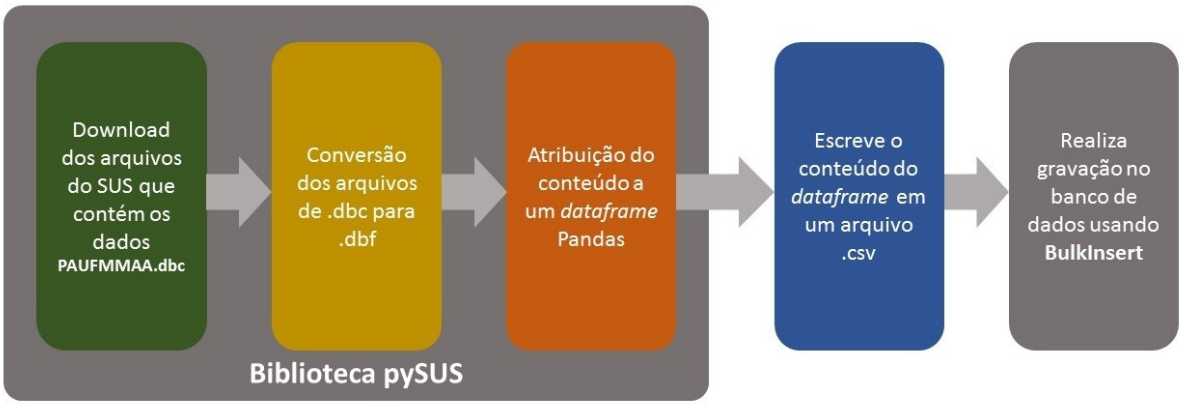
4.4 ESQUEMA DE FUNCIONAMENTO DO DOWNLOAD

Com a finalidade de realizar o download dos dados a serem utilizados pelo trabalho, criou-se uma rotina de download. A própria biblioteca PySUS baixa os arquivos conforme a configuração desejada, converte o arquivo do formato .dbc para .dbf e o atribui a um *dataframe* da biblioteca Pandas (THE PANDAS PROJECT, 2019), que pode ser armazenado em um banco de dados relacional ou usado em um programa Python para os cálculos desejados.

Importante registrar que a primeira estratégia utilizada foi usar o comando `to_sql()` do Python para gravar no banco de dados. No entanto, ela se mostrou muito lenta. Cada arquivo levava em média uma hora para ser baixado, convertido e gravado na tabela. Considerando um período de um ano para os dados, tem-se um arquivo por UF e por mês. Assim tem-se 27 arquivos vezes 12 meses, o que nos dá em torno de 324 arquivos ou a média de 324 horas (aproximadamente 14 dias) para armazenar os dados de um ano, na visão mais otimista. Isto, é claro, se não houvesse nenhum tipo de erro. Cenário este bastante improvável.

Desta forma, tentou-se utilizar o comando *BulkInsert* o qual se mostrou, em média, mais de duzentas vezes mais rápido que a forma anterior, diminuindo consideravelmente o período de carga dos arquivos. Para este trabalho, foram carregados todos os arquivos de procedimentos ambulatoriais para todas as UF para os anos de 2016, 2017 e 2018. Só para 2018 que foi o ano foco da análise realizada no estudo de caso, foram 355.830.512 de linhas incluindo BPAs, APACs e RAAS. Mais de 3,7 bilhões de procedimentos ambulatoriais totalizados. A Figura 6 mostra a ideia utilizada para a obtenção dos dados.

Figura 6. Esquema de download dos dados do SUS usando o PySUS



Fonte: Elaborada pelo Autor (2020)

4.5 ESTRUTURA DOS DADOS AMBULATORIAIS

Segundo BRASIL, MS(2019), dentro do arquivo PAUFAAMM.DBF constam dados dos procedimentos ambulatoriais obtidos através dos seguintes instrumentos de registro do SIA/SUS:

- APAC: Autorização de Procedimentos Ambulatoriais
- BPA-C: Boletim de Produção Ambulatorial Consolidado
- BPA-I: Boletim de Produção Ambulatorial Individualizado
- RAAS-AD: Registro das Ações Ambulatoriais de Saúde - Atenção Domiciliar
- RAAS-PSI: Registro das Ações Ambulatoriais de Saúde - Atenção Psicossocial

O preenchimento dos campos de cada registro do arquivo PA varia conforme o tipo de instrumento que o originou, o que é identificado pelo conteúdo do campo PA_DOCORIG. Assim, quando o campo estiver preenchido com:

- "P", o registro representa os dados do procedimento principal lançado por meio de uma APAC e "S" para os dados do procedimento secundário;
- "C", o registro representa os dados dos procedimentos obtidos por intermédio do BPA-C;
- "A", o registro representa os dados dos procedimentos obtidos por intermédio do RAAS-AD;
- "B", o registro representa os dados dos procedimentos obtidos por intermédio do RAAS – Psicossocial; e,
- "I", o registro representa os dados dos procedimentos obtidos por intermédio do BPA - Individualizado.

Uma vez obtidos os dados, passou-se para a fase de compreensão deles. Os dados ambulatoriais obtidos por meio dos arquivos do SUS possuem 60 colunas, as quais são detalhadas na Tabela 2. Descrição dos campos relacionados aos dados ambulatoriais e os instrumentos que os utilizam (*Legenda: C: BPA-C, I: BPA-I, P: APAC, A: RAAS – Atenção Domiciliar, B: RAAS – Psicossocial*).

Tabela 2. Descrição dos campos relacionados aos dados ambulatoriais e os instrumentos que os utilizam (*Legenda: C: BPA-C, I: BPA-I, P: APAC, A: RAAS – Atenção Domiciliar, B: RAAS – Psicossocial*).

SEQ	COLUNA	DESCRIÇÃO	Instrumento
1	PA_CODUNI	Código do Estabelecimento no CNES (Cadastro Nacional de Estabelecimentos de Saúde)	Todos
2	PA_GESTAO	Código da Unidade da Federação (IBGE) + Código do Município (IBGE) do Gestor, ou UF0000 se o estabelecimento estiver sob Gestão Estadual	Todos
3	PA_CONDIC	Sigla do Tipo de Gestão no qual o Estado ou Município está habilitado	Todos
4	PA_UFMUN	Unidade da Federação + Código do Município onde está localizado o estabelecimento	Todos
5	PA_REGCT	Código da Regra Contratual	Todos
6	PA_INCOUT	Incremento - Outros	P, C, S, I
7	PA_INCURG	Incremento Urgência	Nenhum
8	PA_TPUPS	Tipo de Estabelecimento	Todos
9	PA_TIPPRE	Tipo de Prestador	P, C, S, I
10	PA_MN_IND	Estabelecimento Mantido / Individual	Todos
11	PA_CNPJCPF	CNPJ do Estabelecimento executante	Todos
12	PA_CNPJMNT	CNPJ da Mantenedora do estabelecimento ou zeros, caso não a tenha	Todos
13	PA_CNPJ_CC	CNPJ do Órgão que recebeu pela produção por cessão de crédito ou zeros, caso não o tenha	Todos
14	PA_MVM	Data de Processamento / Movimento (AAAAMM)	Todos
15	PA_CMP	Data da Realização do Procedimento / Competência (AAAAMM)	Todos
16	PA_PROC_ID	Código do Procedimento Ambulatorial: Lei de formação: $N_1N_2N_3N_4N_5N_6XXXX$, onde: N_1N_2-GRUPO, $N_1N_2N_3N_4$-SUBGRUPO, $N_1N_2N_3N_4N_5N_6$-FORMA, XXXX-ID DO PROCEDIMENTO	Todos
17	PA_TPFIN	Tipo de Financiamento da produção	Todos

18	PA_SUBFIN	Subtipo de Financiamento da produção	P, C, S, I
19	PA_NIVCPL	Complexidade do Procedimento	Todos
20	PA_DOCORIG	Instrumento de Registro. *1 BPA-C (C), 2 BPA-I (I), 3 APAC (P,S), 4 RAAS – Atenção Domiciliar (A), 7 RAAS – Psicossocial (B)*	Todos
21	PA_AUTORIZ	Número da APAC ou número de autorização do BPA-I, conforme o caso. **No BPA-I, não é obrigatório, portanto, não é criticado**. Regra de formação: UFAATsssssssd, onde: UF – Unid. da Federação, AA – ano, T – tipo, sssssss – sequencial, d – dígito	P, C, S, I (não obrigatório no caso do BPA-I)
22	PA_CNSMED	Número do CNS (Cartão Nacional de Saúde) do profissional de saúde executante	A, I, B
23	PA_CBOCOD	Código da Ocupação do profissional na Classificação Brasileira de Ocupações8 (Ministério do Trabalho)	Todos
24	PA_MOTSAI	Motivo de saída **ou zeros, caso não tenha**	I, A, P, S, B
25	PA_OBITO	Indicador de Óbito	I, A, P, S, B
26	PA_ENCERR	Indicador de Encerramento	P, B, I, S
27	PA_PERMAN	Indicador de Permanência	A, P, S, B
28	PA_ALTA	Indicador de Alta	A, P, S, B
29	PA_TRANSF	Indicador de Transferência	P, S, B
30	PA_CIDPRI	CID Principal	I, A, P, S, B
31	PA_CIDSEC	CID Secundário	P, S, I
32	PA_CIDCAS	CID Causas Associadas	I, A, P, S, B
33	PA_CATEND	Caráter de Atendimento	I, A, P, S, B
34	PA_IDADE	Idade do paciente em anos	Todos com exceção dos BPA-C que só têm para algumas exceções
36	IDADEMIN	Idade mínima do paciente para realização do procedimento	?

37	IDADEMAX	Idade máxima do paciente para realização do procedimento	?
35	PA_FLIDADE	Compatibilidade com a faixa de idade do procedimento (SIGTAP – Sistema de Gerenciamento da Tabela de Procedimentos do SUS):	Todos
38	PA_SEXO	Sexo do paciente. M – Masculino, F – Feminino, 0 – se dados consolidados	Todos. No entanto, valores M e F só são preenchidos para I, A, P, S, B
39	PA_RACACOR	Raça/Cor do paciente	Todos. No entanto, para o tipo C, o valor é sempre '00'
40	PA_MUNPCN	Código da Unidade da Federação + Código do Município de residência do paciente ou do estabelecimento, caso não se tenha a identificação do paciente, o que ocorre no (BPA-C)	I, A, P, S, B
41	PA_QTDPRO	Quantidade Produzida (APRESENTADA)	Todos
42	PA_QTDAPR	Quantidade Aprovada do procedimento	Todos
43	PA_VALPRO	Valor Produzido (APRESENTADO)	Todos
44	PA_VALAPR	Valor Aprovado do procedimento	Todos
45	PA_UFDIF	Indica se a UF de residência do paciente é diferente da UF de localização do estabelecimento:	I, A, P, S, B
46	PA_MNDIF	Indica se o município de residência do paciente é diferente do município de localização do estabelecimento:	I, A, P, S, B
47	PA_DIF_VAL	Diferença do Valor Unitário do procedimento praticado na Tabela Unificada com Valor Unitário praticado pelo Gestor da Produção, multiplicado pela Quantidade Aprovada	I, C
48	NU_VPA_TOT	Valor Unitário do Procedimento da Tabela VPA	I, C, P, S
49	NU_PA_TOT	Valor Unitário do Procedimento da Tabela SIGTAP	I, C, P, S

50	PA_INDICA	Indicativo de situação da produção produzida:	Todos
51	PA_CODOCO	Código de Ocorrência	Todos
52	PA_FLQT	Indicador de erro de Quantidade Produzida	Todos
53	PA_FLER	Indicador de erro de corpo da APAC	Todos
54	PA_ETNIA	Etnia do paciente	I, A, P, S, B
55	PA_VL_CF	Valor do Complemento Federal	Todos
56	PA_VL_CL	Valor do Complemento Local	Todos
57	PA_VL_INC	Valor do Incremento	Todos
58	PA_SRC_C	Código do Serviço Especializado / Classificação CBO (de acordo com o CNES)	Todos

Fonte: Adaptado de BRASIL, MS (2019)

Como a tabela é usada para preencher dados de todos os instrumentos de coleta de informação, alguns deles só são preenchidos quando temos um BPA-I ou uma APAC. Isto torna a utilização dos dados um pouco confusa, pois, dependendo do tipo de instrumento usado, temos ao mesmo tempo, dados consolidados e individualizados. Assim, é importante tomar cuidado com esse detalhe ao realizar cálculos. Por exemplo, o campo sexo (PA_SEXO) é detalhado nos casos individualizados (valores F ou M). Mas, no caso da informação consolidada, o campo assume valor 0 (zero), que significa 'não requerido'.

Desta forma, fica impossível saber a distribuição exata de todos os atendimentos ambulatoriais em relação ao sexo. Há ainda casos em que determinados procedimentos podem ser lançados de forma consolidada ou individualizada, o que dificulta ainda mais o tratamento dos dados.

A quantidade de registros (linhas) varia conforme o tipo de instrumento. A saber:

- Um instrumento BPA-C gera diversos registros no arquivo PA: um para cada par [*código de procedimento, CBO*], correspondendo à respectiva linha no instrumento original em papel;
- Um instrumento BPA-I gera diversos registros no arquivo PA: um para cada atendimento. Há casos de registros de um ou mais atendimentos no BPA-I para o mesmo paciente. Neste caso, pode-se, inclusive, repetir o número da autorização dada pelo gestor de saúde. No BPA-I, o campo (PA_AUTORIZ) é não-obrigatório e, portanto, não é criticado. Assim, devem ser consideradas as limitações quanto à qualidade do seu preenchimento, caso este campo seja utilizado para algum tipo de controle;
- Um instrumento APAC gera diversos registros no arquivo PA: um para cada código de procedimento realizado na APAC, seja ele procedimento principal (P) ou secundário (S).

Importante notar que alguns dos campos têm os valores descritos em arquivos complementares com extensão .cnv e .dbf que também são fornecidos pelo DATASUS. A relação do campo com seu respectivo arquivo complementar é encontrada no arquivo "Produção Ambulatorial - 200801_.DEF". Estes arquivos foram importados para tabelas no banco de dados, de acordo com a necessidade.

Caso o procedimento seja do tipo APAC, há arquivos com informações complementares e específicas do procedimento. Para cada tipo detalhado há uma regra de formação do nome do arquivo a ser utilizado, segundo a Tabela 3.

Tabela 3. Regra de formação dos nomes de arquivos do SIA/SUS

APAC	
	Medicamentos: **AM**UFAAMM.DBF
	Nefrologia: **AN**UFAAMM.DBF
	Quimioterapia: **AQ**UFAAMM.DBF
	Radioterapia: **AR**UFAAMM.DBF
	Cirurgia Bariátrica: **AB**UFAAMM.DBF
	Confecção de Fístula Arteriovenosa: **ACF**UFAAMM.DBF
	Tratamento Dialítico: **ATD**UFAAMM.DBF
BPA-I	
	BIUFAAMM.DBF
RAAS	
	Atenção Domiciliar: **SAD***.DBF
	Psicossocial: **PS***.DBF

Fonte: Elaborada pelo Autor (2020)

5 FASE DE PREPARAÇÃO DOS DADOS

Após realizar o entendimento dos dados, passou-se à fase de preparação dos dados. Nesta etapa, realizou-se uma análise exploratória com a finalidade de checar a qualidade dos dados obtidos do DATASUS. Apesar de já se ter uma ideia antecipada de quais colunas seriam utilizadas, decidiu-se por realizar a análise para todas as colunas, com a intenção de achar algum dado interessante que não havia sido considerado. Para tal, utilizaram-se os dados do estado do Acre dos anos de 2017 e 2018.

5.1 PREPARAÇÃO DOS DADOS

No início realizou-se um comando *describe*() para verificar as colunas que possuem dados numéricos. O resultado pode ser visto na Figura 7.

Figura 7. Resultado do comando *describe*()

```
pd.set_option('display.max_columns', 999)
df.describe()
```

	PA_QTDPRO	PA_QTDAPR	PA_VALPRO	PA_VALAPR	PA_DIF_VAL	NU_VPA_TOT	NU_PA_TOT	PA_VL_CF	PA_VL_CL	PA_VL_INC
count	1.450474e+06	1.450474e+06	1.450474e+06	1.450474e+06	1450474.0	1450474.0	1.450474e+06	1450474.0	1450474.0	1.450474e+06
mean	1.679378e+01	1.613943e+01	8.868424e+01	8.648284e+01	0.0	0.0	2.616337e+01	0.0	0.0	1.310148e-01
std	2.947667e+02	2.881358e+02	1.067019e+03	1.041382e+03	0.0	0.0	1.142883e+02	0.0	0.0	6.230907e+00
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0	0.0	0.000000e+00	0.0	0.0	0.000000e+00
25%	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.0	0.0	0.000000e+00	0.0	0.0	0.000000e+00
50%	1.000000e+00	1.000000e+00	6.970000e+00	6.970000e+00	0.0	0.0	6.300000e+00	0.0	0.0	0.000000e+00
75%	4.000000e+00	4.000000e+00	3.000000e+01	3.000000e+01	0.0	0.0	1.500000e+01	0.0	0.0	0.000000e+00
max	1.303010e+05	1.303010e+05	2.692250e+05	2.692250e+05	0.0	0.0	1.164400e+04	0.0	0.0	4.488400e+02

Fonte: Elaborada pelo Autor (2020)

Em seguida, os dados foram testados em relação aos nulos ou brancos. Ao testar os nulos, não foi encontrado nenhum caso no conjunto de dados, o que pareceu ser estranho à primeira vista. Foi então feito um teste por valores em branco. No caso de valores numéricos, os brancos foram substituídos pelo valor 0 (zero).

Na sequência, os valores numéricos que constavam como *string* foram convertidos para os valores apropriados. Isso foi feito com as colunas PA_IDADE, IDADEMIN, IDADEMAX, PA_FLIDADE, PA_QTDPRO, PA_QTDAPR, PA_VALPRO e PA_VALAPR.

Como próximo passo, verificou-se se a regra da obrigatoriedade do preenchimento do número de autorização (PA_AUTORIZ) para APACs foi respeitada. Assim, agrupou-se pelo campo PA_DOCORIG e verificou-se que somente os tipos de instrumentos que não exigem o preenchimento dos números é que estão com as autorizações zeradas (Figura 8).

Figura 8. Valores de autorizações zerados para os diferentes instrumentos de captação de dados.

```
In [16]:
# Avalia quantas autorizações estão representadas por 0000000000000
df.groupby(['PA_AUTORIZ','PA_DOCORIG']).size().head(15)

# RAAS - Psicossocial (B) tem 40.000, BPA-C tem 609.173 e BPA-I tem 635.643 com o valor 0000000000000

Out[16]: PA_AUTORIZ     PA_DOCORIG
         0000000000000  B             40000
                        C            609173
                        I            635643
         0000000000001  I             21159
         0000000000002  I               153
         0000000000003  I               146
         0000000000004  I               116
         0000000000005  I               115
         0000000000006  I               115
         0000000000007  I               116
         0000000000008  I               114
         0000000000009  I               115
         000000000000Q  I                 3
         000000000000Z  I                 1
         0000000000010  I               114
         dtype: int64
```

Fonte: Elaborada pelo Autor (2020)

Um aspecto curioso encontrado foi que o dígito verificador variou de 0 a 9 e, por vezes, utilizou as letras Q e Z também (Figura 9). Não foi possível saber se isso é algo normal ou se se trata de um erro de digitação.

Figura 9. Resultado da avaliação do campo PA_AUTORIZ por valores zerados.

```
In [22]:
# Considerando que o último dígito é verificador, vamos verificar as demais autorizações
# no formato 000000000000X

# Primeiro vamos buscar os casos em que PA_AUTORIZ termina com uma letra
df[df['PA_AUTORIZ'].str.contains('^\d{12}\D$')].PA_AUTORIZ
```

```
Out[22]: 48887    000000000000Q
11178    000000000000Z
999      000000000000Q
60864    000000000000Q
Name: PA_AUTORIZ, dtype: object
```

```
In [23]:
# Verificamos dois casos em que o valor termina com 'Q'
# Agora vamos ver os demais casos que temos autorizaçoes zeradas que
# terminam com números que não zero
df[df['PA_AUTORIZ'].str.contains('^0{12}[1-9]$')].groupby(['PA_AUTORIZ']).size()
```

```
Out[23]: PA_AUTORIZ
0000000000001    21159
0000000000002      153
0000000000003      146
0000000000004      116
0000000000005      115
0000000000006      115
0000000000007      116
0000000000008      114
0000000000009      115
dtype: int64
```

Fonte: Elaborada pelo Autor (2020)

5.2 REMOÇÃO DE DADOS NULOS E EM BRANCO

Considerando o resultado do comando *describe* (Figura 7), as colunas PA_DIF_VAL, NU_VPA_TOT, PA_VL_CF e PA_VL_CL não apresentam nenhum valor. Além disso, a coluna PA_VL_INC apresenta o min, 25%, 50% e 75% iguais a zero. Entendeu-se que ela pode ser descartada em conjunto com as anteriores.

Continuando a avaliação dos dados, verificou-se ainda mais as seguintes colunas com valores zerados, vazios ou nulos em grande quantidade: PA_INCOUT, PA_INCURG, PA_TIPPRE, PA_FLER, PA_ETNIA, PA_SRV_C e PA_INE. Desta forma, essas colunas também foram excluídas da análise em conjunto com as anteriores.

Aparentemente, algumas colunas foram inseridas, mas ainda não são plenamente utilizadas ou não são corretamente preenchidas.

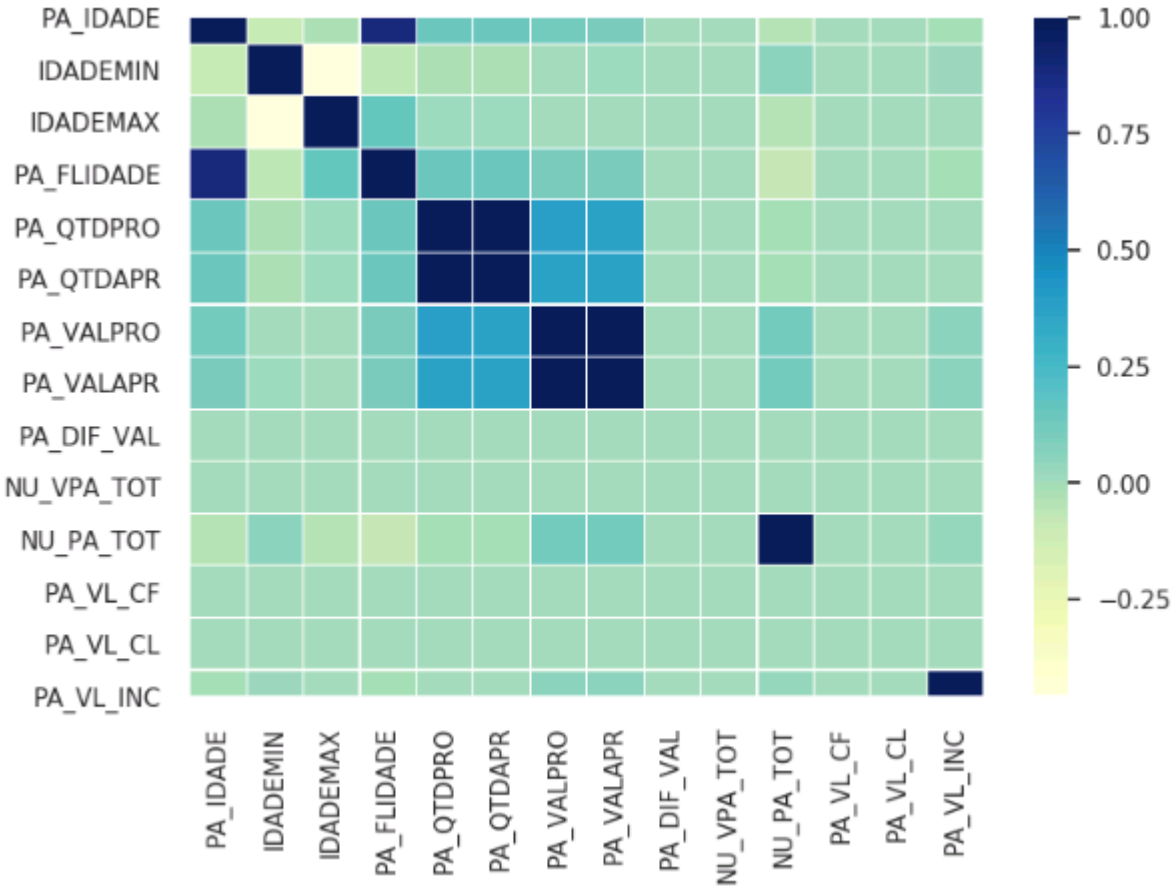
Verificou-se ainda se o campo PA_UFDIF que indica que o atendimento foi realizado em UF diferente da UF de residência do usuário do SUS. Este campo somente é preenchido nos registros individualizados. Logo, não pode ser considerado em estudos onde os dados consolidados sejam utilizados. Também existe o campo PA_MNDIF que indica se o município

de residência do paciente é diferente do município de localização do estabelecimento de atendimento, o qual segue a mesma ideia de preenchimento.

Além desses campos, a análise ainda mostrou que o campo PA_OBITO possui pouquíssimas linhas preenchidas. Normalmente o valor utilizado é zero. Desta forma, também foi deixado de fora.

Como passo final, plotou-se uma matriz de correlação das colunas (Figura 10) que não mostrou nenhum achado revelador em termos de correlação. Basicamente há alguma correlação entre os campos de idade e os campos de quantidade e valores. No caso destes últimos, são os campos que seriam utilizados inicialmente na análise proposta.

Figura 10. Matriz de correlação das variáveis do SIA/SUS



Fonte: Elaborada pelo Autor (2020)

Sendo assim, optou-se por realizar a análise em um *dataframe* reduzido com apenas 23 colunas seguintes:

'PA_CODUNI', 'PA_GESTAO', 'PA_CONDIC', 'PA_UFMUN', 'PA_CNPJ_CC', 'PA_MVM', 'PA_CMP', 'PA_PROC_ID', 'PA_DOCORIG', 'PA_AUTORIZ', 'PA_CNSMED',

'PA_CBOCOD', 'PA_CIDPRI', 'PA_IDADE', 'PA_QTDPRO', 'PA_QTDAPR', 'PA_VALPRO', 'PA_VALAPR', 'PA_UFDIF', ‘PA_MNDIF', 'PA_DIF_VAL', 'NU_VPA_TOT', 'NU_PA_TOT'

5.3 CAMPOS IMPORTANTES PARA A ANÁLISE

Dentre os campos pré-selecionados, o campo PA_PROC_ID merece atenção especial. É onde encontramos o número dos procedimentos realizados. Segundo a documentação do SUS, a identificação do procedimento segue o formato da Figura 11.

Figura 11. Formato do campo PA_PROC_ID

02.04.02.010-7	RADIOGRAFIA DE COLUNA TORACO-LOMBAR
02.XX.XX.XXX-X	Procedimentos com finalidade diagnóstica
02.04.XX.XXX-X	Diagnóstico por radiologia
02.04.02.XXX-X	Exames radiológicos da coluna vertebral

Fonte: Elaborada pelo Autor (2020)

Entender esta estrutura é importante porque a utilizaremos por todo o trabalho, uma vez que a análise poderá ser realizada por grupo, subgrupo, forma ou pelo procedimento específico. Por exemplo, pode-se em algum momento optar por avaliar todos os procedimentos do grupo 02 – Procedimentos com finalidade diagnóstica. Neste caso, as taxas serão calculadas para todos os procedimentos pertencentes ao grupo, ou seja, todos os procedimentos cujo PA_PROC_ID tenham o formato 02.xx.xx.xxx-x. Caso a escolha seja avaliar os procedimentos da forma 02.04.02 – Exames radiológicos da coluna vertebral, serão considerados todos os procedimentos cujo PA_PROC_ID tenham o formato 02.04.02.xxx-x, conforme pode-se verificar na Tabela 4.

Tabela 4. Procedimentos da forma 02.04.02 - Exames radiológicos da coluna vertebral

Código do procedimento	Descrição
02.04.02.001-8	Mielografia
02.04.02.002-6	Planigrafia de coluna vertebral
02.04.02.003-4	Radiografia de coluna cervical (ap + lateral + to + oblíquas)
02.04.02.004-2	Radiografia de coluna cervical (ap + lateral + to / flexão)
02.04.02.005-0	Radiografia de coluna cervical funcional / dinâmica
02.04.02.006-9	Radiografia de coluna lombo-sacra
02.04.02.007-7	Radiografia de coluna lombo-sacra (c/ obliquas)
02.04.02.008-5	Radiografia de coluna lombo-sacra funcional / dinâmica
02.04.02.009-3	Radiografia de coluna torácica (ap + lateral)
02.04.02.010-7	Radiografia de coluna tóraco-lombar
02.04.02.011-5	Radiografia de coluna tóraco-lombar dinâmica
02.04.02.012-3	Radiografia de região sacro-coccigea
02.04.02.013-1	Radiografia panorâmica de coluna total- telespondilografia (p/ escoliose)

Fonte: Elaborada pelo Autor (2020)

A avaliação dos procedimentos de um grupo, subgrupo ou forma poderá ainda ser feita de forma condensada ou individualizada. Assim, dependendo da escolha, os *outliers* poderão ser levantados considerando a taxa de atendimentos de todos os procedimentos em conjunto ou de cada um pertencente ao grupo, subgrupo ou forma escolhido.

Para entender melhor com um exemplo, caso a escolha fosse analisar os procedimentos da forma 02.04.02, uma análise condensada seria realizada com o cálculo da taxa com base na soma da quantidade de todos os atendimentos dos procedimentos listados na Tabela 4. Já no caso da análise individual da forma em questão, a análise de *outlier* seria realizada com base na taxa calculada para cada procedimento da Tabela 4.

Além do campo PA_PROC_ID, é importante dar uma atenção maior para o significado dos campos PA_CODUNI, PA_UFMUN, PA_CMP e PA_DOCORIG.

PA_CODUNI é o campo que contém a informação do estabelecimento que realizou o procedimento que está sendo analisado e que está cadastrado no CNES (Cadastro Nacional de Estabelecimentos de Saúde). Ela será utilizada para a análise de *outliers* por estabelecimento.

O código PA_UFMUN contém o código do município segundo a tabela do IBGE com a exceção do dígito verificador. Ele será utilizado para a análise de *outliers* por município.

O campo PA_CMP, por sua vez, traz a data de realização do procedimento. Ela será convertida em tipo *Date* para utilização nas análises.

Já o campo PA_DOCORIG armazena o tipo do instrumento utilizado para captar a informação sobre o procedimento ambulatorial e já foi detalhado na Seção 4.5.

Considerando ainda que a intenção é realizar uma análise de valores para encontrar distorções estatísticas em qualquer procedimento ambulatorial e, como os arquivos misturam dados consolidados e individualizados, entendeu-se que as variáveis que mais se mostraram interessantes para a nossa análise foram aquelas relacionadas com quantidade ('PA_QTDPRO' e 'PA_QTDAPR') e valores ('PA_VALPRO' e 'PA_VALAPR'). A variável PA_QTDPRO trata da quantidade produzida declarada pelo estabelecimento. PA_QTDAPR, por sua vez, é a quantidade aprovada com base no parâmetro anterior. O mesmo raciocínio vale por analogia para as duas variáveis de valor. A variável PA_QTDAPR será utilizada para calcular a taxa de atendimento a ser utilizada pelo método de detecção de *outliers*. Decidiu-se que esta variável é mais apropriada que PA_QTDPRO por ela ter a informação já aprovada pelo SUS.

5.4 INSERÇÃO DE CAMPOS COMPLEMENTARES

Como vários campos dos dados ambulatoriais vêm na forma de códigos, foi necessário incluir as respectivas descrições, provenientes de outras tabelas. Além disso, foi necessário incluir os dados populacionais dos municípios, UFs e do país com a finalidade de realizar os cálculos das taxas de atendimento. A tabela mostra esses campos, seus complementos e a origem dos dados.

Tabela 5. Campos adicionados à tabela de procedimentos ambulatoriais.

Campo	Descrição	Origem da Descrição
PA_UFMUN	Nome do município	Tabela POP_MUNICIPIO do IBGE
PA_CODUNI	Nome do estabelecimento	Tabela de Estabelecimentos de saúde do CNES
PA_PROC_ID	Nome do procedimento	Tabela TB_SIGTAP.DBF
FORMA	Descrição da forma	Tabela TB_FORMA.DBF
SUBGRUPO	Descrição do subgrupo	Tabela TB_SUBGR.DBF
GRUPO	Descrição do grupo	Tabela TB_GRUPO.DBF
POPULACAO_MUN	População do município	Tabela POP_MUNICIPIO do IBGE

POPULACAO_UF	População da UF	Tabela POP_MUNICIPIO do IBGE
POPULACAO_NAC	População do país	Tabela POP_MUNICIPIO do IBGE

Fonte: Elaborada pelo Autor (2020)

6 FASE DE MODELAGEM

6.1 ANÁLISE EXPLORATÓRIA

Após avaliar e selecionar as colunas a serem utilizadas, realizou-se uma rápida análise exploratória. Selecionaram-se informações associadas às consultas médicas realizadas no Estado do Acre. O comando SQL usado foi o mesmo do trabalho de análise dos indicadores de saúde da OCDE. No comando foram utilizados números de procedimentos ou formas que identificam as consultas médicas, limitadas pelos CBOs que identificam os profissionais que exercem a medicina.

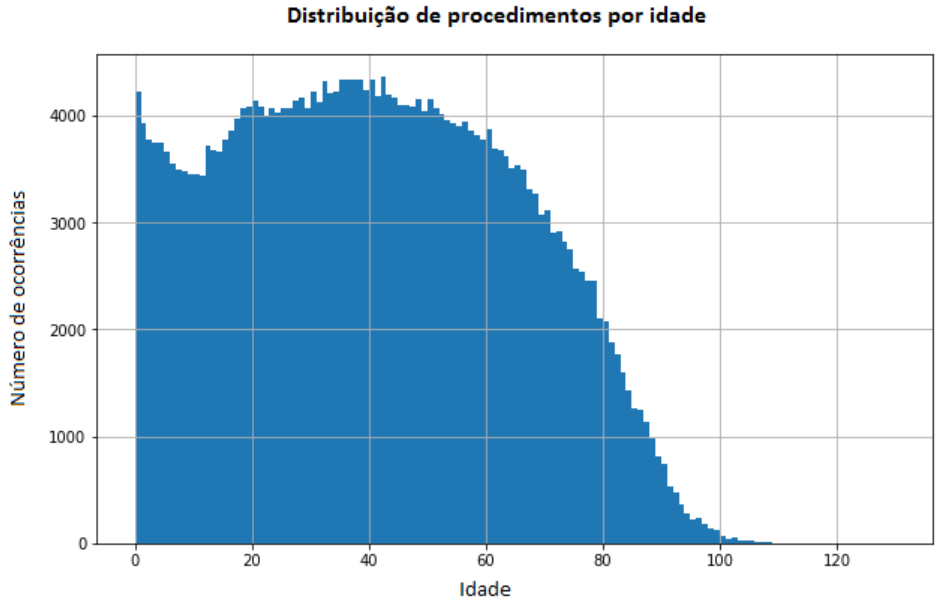
Figura 12. Comando SQL para selecionar as consultas médicas do estado do Acre.

```
# SQL para selecionar todas as consultas médicas
query = """select *
from [U_SANTOSRF].[dbo].[SIA_PA]
where (pa_proc_id in('0301040010', '0301040028', '0301040044')
or pa_proc_id like '030101%'
or pa_proc_id like '030102%'
or pa_proc_id like '030106%'
or pa_proc_id like '030107%'
or pa_proc_id like '030109%'
or pa_proc_id like '030111%'
or pa_proc_id like '030112%'
or pa_proc_id like '030113%')
and (pa_cbocod like '2231%'
or pa_cbocod like '2251%'
or pa_cbocod like '2252%'
or pa_cbocod like '2253%')
"""

connSQL = pyodbc.connect('DSN=LABCONTAS;DATABASE=U_SANTOSRF')
df = pd.read_sql_query(query, connSQL)
connSQL.close()
df.head()
```

Fonte: Elaborado pelo Autor (2020)

Com a finalidade verificar os dados, realizaram-se algumas observações sobre eles. A distribuição dos atendimentos por idade pode ser vista na Figura 13.

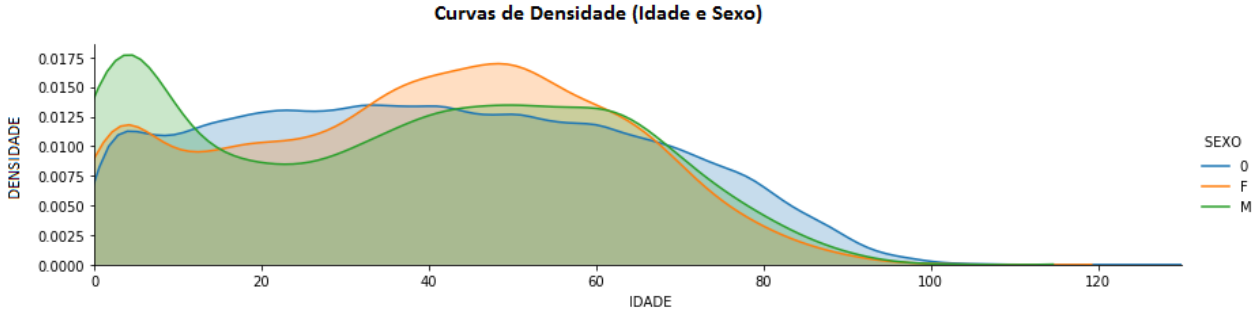
Figura 13. Distribuição por idade das consultas médicas no estado do Acre



Fonte: Elaborada pelo Autor (2020)

Também foi realizada a análise da distribuição dos dados por sexo e idade, a qual pode ser verificada na Figura 14. Importante notar que o terceiro valor para o campo PA_SEXO denota os casos de informação consolidada (BPA-C). Desta forma, a informação sobre o sexo dos pacientes atendidos é perdida.

Figura 14. Distribuição das consultas médicas por sexo e idade.

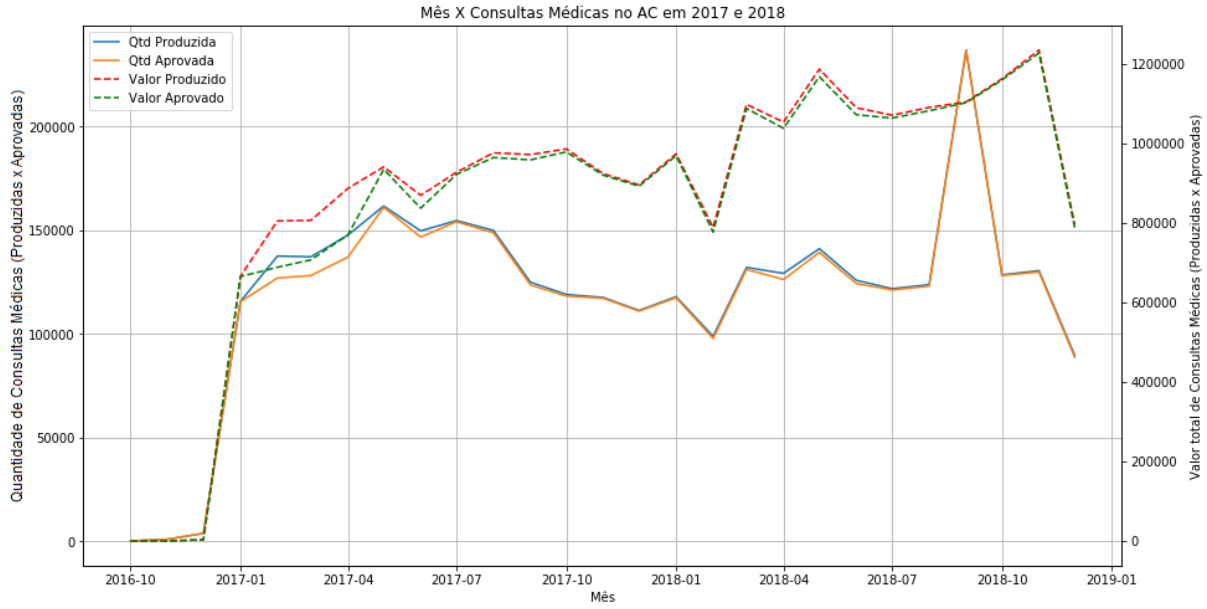


Fonte: Elaborada pelo Autor (2020)

Plotando um gráfico confrontando a quantidade produzida e aprovada com os valores produzidos e aprovados (Figura 15), verifica-se a evolução no decorrer dos meses de 2017 e 2018 para o estado do Acre.[2]

[2] O gráfico mostra alguns valores para 2016, pois a produção de um mês pode ser informada até três meses após a data de realização. Assim, dados de dezembro de 2016, por exemplo, pode ser lançado até março de

Figura 15. Confronto da quantidade produzida e aprovada com os valores produzidos e aprovados.



Fonte: Elaborada pelo Autor (2020)

Verifica-se imediatamente pelo gráfico na Figura 15 que há um pico no número de atendimentos no mês de setembro de 2018 que difere do comportamento esperado. Analisando os valores da UNIDADE DE SAUDE DA FAMILIA FRANCISCO PEREIRA DANTAS em Cruzeiro do Sul, Acre, verifica-se que existe um registro que realmente destoa dos demais (Linha 20 da Figura 16).

2017. Assim, apesar de os arquivos serem de 2017 e 2018, há dados residuais de 2016 que acabaram sendo plotados no gráfico da Figura 15.

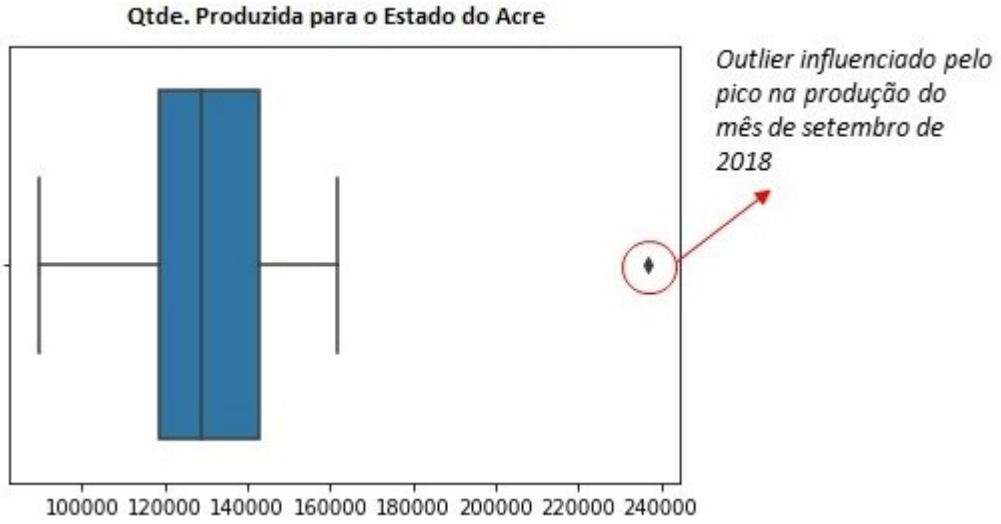
Figura 16. Listagem da quantidade de consultas médicas produzidas pela unidade de atendimento 6270638.

	Unidade de atendimento	PA_CMP	PA_QTDPRO
0	6270638	2017-01-01	207
1	6270638	2017-02-01	184
2	6270638	2017-03-01	199
3	6270638	2017-04-01	171
4	6270638	2017-05-01	256
5	6270638	2017-06-01	227
6	6270638	2017-07-01	212
7	6270638	2017-08-01	101
8	6270638	2017-09-01	121
9	6270638	2017-10-01	181
10	6270638	2017-11-01	166
11	6270638	2017-12-01	169
12	6270638	2018-01-01	181
13	6270638	2018-02-01	204
14	6270638	2018-03-01	116
15	6270638	2018-04-01	203
16	6270638	2018-05-01	207
17	6270638	2018-06-01	165
18	6270638	2018-07-01	125
19	6270638	2018-08-01	181
20	6270638	2018-09-01	101186
21	6270638	2018-10-01	209
22	6270638	2018-11-01	227
23	6270638	2018-12-01	150

Fonte: Elaborada pelo Autor (2020)

Ainda usando um gráfico do tipo BoxPlot, agora somando a produção de todas as unidades, podemos avaliar também o ponto que está fora da normalidade, para o Estado do Acre, que representa o valor fora da normalidade encontrado na linha 20, mostrado na Figura 16 e como ele impacta toda a produção da UF em questão.

Figura 17. Boxplot mostrando a unidade *outlier* dentre a produção das unidades do estado do Acre.



Fonte: Elaborada pelo Autor (2020)

Este seria um exemplo de uma distorção nos dados interessante de ser analisada. No entanto, dada a grande quantidade de dados, não é possível fazer isso sempre manualmente e visualmente.

Tomando como exemplo este caso, a proposta do trabalho é desenvolver uma rotina automática de verificação de *outliers* que, dado um procedimento, indique os municípios e estabelecimentos que apresentam valores fora da normalidade para as quantidades realizadas ou aprovadas.

Realizou-se então, ainda nesta análise exploratória, um teste com algoritmos de detecção de *outliers*: Z-score, Z-score Modificado, IQR – *Interquartile Range, Isolation Forest* e LOF – *Local Outlier Factor*, os quais serão explicados na seção 6.5. Nos primeiros testes, verificou-se que os métodos estavam retornando muitos resultados o que dificulta a análise.

Com vistas a reduzir este número decidiu-se criar um método que aplicasse todos os algoritmos sobre os valores e selecionasse apenas os *outliers* em comum nos 5 métodos.

Além disso, com a intenção de realizar uma comparação menos enviesada entre os municípios ou estabelecimentos, ao invés de usar o número de atendimentos diretamente no cálculo, utilizou-se a taxa de atendimentos aprovada (PA_QTDAPR) por um fator de 100 habitantes no município onde foi realizado o atendimento.

$$\text{Taxa}_{\text{Qtde}_{\text{Aprov}}} = \frac{\text{Qtde}_{\text{Aprovada}}}{\text{População}_{\text{Município}_{\text{Atend}}}} * 100$$

A título de exemplo do funcionamento do algoritmo de detecção, realizou-se o cálculo para todos os meses e aplicou-se o algoritmo conjunto de detecção de *outliers* sobre a taxa média mensal de consultas médicasdo Estado do Acre. Os resultados do mês de setembro de 2018 podem ser visualizados na Tabela 6 em conjunto com a Figura 18.

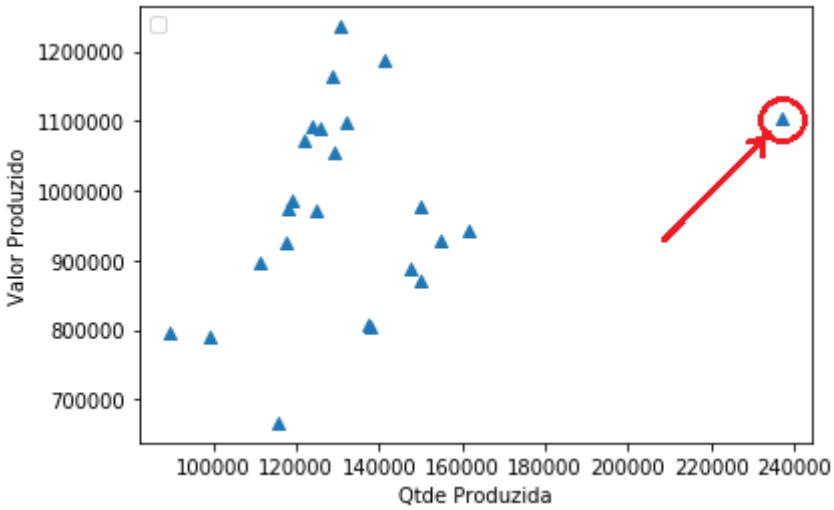
Tabela 6. Resultados dos algoritmos de detecção de *outliers*.

Z-Score	Z-Score Modificado	IQR	Isolation Forest	LOF
23: 30.52148	**23:** 30.52148	**23:** True	16: 12.743685 **23:** 30.521480 26: 11.544272	7: 20.829969 **23:** 30.52148 26: 11.54427

Fonte: Elaborada pelo Autor (2020)

O *outlier* em comum nos 5 métodos foi exatamente o mês de setembro de 2018 para a taxa de atendimento do Estado do Acre. (Registro de número 23), responsável pelo valor fora da normalidade visto anteriormente.[3]

Figura 18. Gráfico com os pontos relativos às quantidades produzidas, com o mês de setembro de 2018 detectado como *outlier*, marcado em vermelho



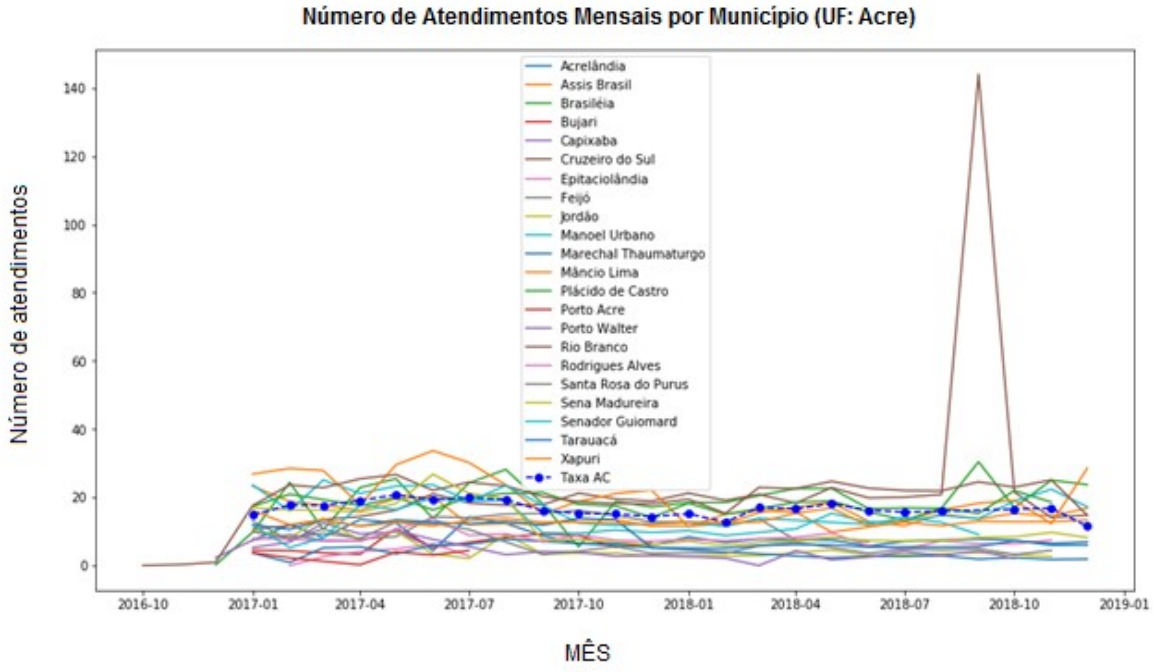
Fonte: Elaborada pelo Autor (2020)

No gráfico da Figura 19, a taxa de atendimento do município *outlier* (Cruzeiro do Sul – AC) pode ser comparada com a taxa dos demais municípios, bem como com a taxa de

[3] Neste caso, o registro 23 mostra a taxa média mensal para a UF do AC no mês de setembro de 2018. Este mês coincide com o pico de atendimento da linha 20 (Figura 16) da UNIDADE DE SAUDE DA FAMILIA FRANCISCO PEREIRA DANTAS em Cruzeiro do Sul, Acre

atendimento do estado em questão: Acre. O pico do mês de setembro de 2018, bem como o comportamento médio esperado ficam bem claros na figura.

Figura 19. Comparação das taxas de atendimento dos municípios do estado do Acre.



Fonte: Elaborada pelo Autor (2020)

A intenção é, portanto, expandir essa avaliação para qualquer procedimento ambulatorial considerando o desempenho em todos os municípios do país, identificando os municípios e estabelecimentos com *outliers* para um conjunto de procedimentos. Desta forma, a análise exploratória realizada serviu como um pequeno ensaio do trabalho de análise a ser realizado.

6.2 MODELAGEM

Após a análise exploratória, passou-se à fase de modelagem. Neste passo, aprimorou-se a rotina de análise definida anteriormente para o levantamento de *outliers* e, ao invés de rodar os dados apenas para os municípios de uma UF, rodou-se o procedimento para todos os municípios do país. O ano escolhido foi o de 2018, pois era, à altura, o ano mais recente com todos os meses fechados em relação aos dados.

A intenção é tentar reproduzir a ideia proposta no trabalho do InfoSAS onde um procedimento tem a taxa de atendimento calculada para todos os municípios do país e, esta taxa é utilizada para o cálculo das distorções.

O primeiro problema enfrentado foi o grande volume de dados. Como visto anteriormente, tem-se mais de 350 milhões de registros somente para o ano de 2018, o que torna difícil utilizá-los na totalidade para os cálculos do presente trabalho sem uma estratégia de administração de dados. Com isto, a complexidade do trabalho tenderia a aumentar. Decidiu-se, portanto, trabalhar apenas com um dos grupos para testar os conceitos. A quantidade de registros para cada grupo pode ser vista na Tabela 7.

Tabela 7. Dados sobre os grupos de procedimentos ambulatoriais.

Grupo	Descrição	Número de procedimentos	Quantidade de registros
01	Ações de promoção e prevenção em saúde	108	5.056.766
02	Procedimentos com finalidade diagnóstica	1.054	135.617.451
03	Procedimentos clínicos	825	171.186.865
04	Procedimentos cirúrgicos	1.840	8.650.864
05	Transplantes de órgãos, tecidos e células	139	1.869.352
06	Medicamentos	617	26.222.920
07	Órteses, próteses e materiais especiais	555	2.221.350
08	Ações complementares da atenção à saúde	44	5.004.944
	Total:	**5.182**	**355.830.512**

Fonte: Elaborada pelo Autor (2020)

Para o presente trabalho, escolheu-se o grupo **02 – Procedimentos com finalidade diagnóstica**, o qual é o segundo maior grupo e possui exames com uma grande abrangência em vários municípios e importância, tais como mamografias e vários tipos de radiografias. Isso ajudará a realizar uma prova de conceito com um volume menor de dados. Mesmo assim, a ideia é que se possa desenvolver um produto que permita ser utilizado com os demais grupos de procedimentos ambulatoriais realizando apenas uma configuração de qual grupo pode ser utilizado.

6.3 MÉTODOS DE DETECÇÃO DE *OUTLIERS*

Segundo (HAN, 2012), um *outlier* é um objeto de dados que desvia significantemente dos demais objetos, como se ele tivesse sido gerado por um mecanismo diferente.

Outliers são diferentes de ruídos. Um ruído é um erro aleatório ou a variância em uma variável medida. De maneira geral, ruído não é interessante na análise de dados, inclusive na detecção de *outliers*. Por exemplo, em um algoritmo de detecção de fraudes em cartões de crédito, o comportamento de compra do consumidor pode ser modelado como uma variável randômica. Um cliente pode gerar algumas 'transações ruidosas' tais como o pagamento de um almoço caro em um determinado dia. Este tipo de transação não deve ser tratado como *outliers*, caso contrário, a companhia acabará tendo um alto custo para verificar todas as transações. Na análise e mineração de dados o ruído deve ser removido antes da detecção de *outliers*.

6.4 TIPOS DE *OUTLIERS*

Os *outliers* podem ser classificados em três categorias: globais, contextuais (ou condicionais) e coletivos. (HAN, 2012)

6.4.1 *Outliers* Globais

Em um dado *dataset*, um objeto de dados é um *outlier* global se ele se desvia significantemente do restante do conjunto de dados. Este tipo de *outlier* é o mais simples de todos e também é chamado de anomalia pontual.

Para detectar os *outliers* globais, um detalhe crítico é encontrar um desvio de medida apropriado de acordo com a aplicação em questão. Várias formas de medição são propostas e, baseadas nelas, a detecção de *outliers* são divididas em categorias diferentes.

A detecção de *outliers* globais é importante em muitas aplicações. Por exemplo, a detecção de intrusos em redes de computadores, por exemplo. Se o comportamento de comunicação em um computador é muito diferente dos padrões normais (e.g.: um grande número de pacotes é transmitido em um curto espaço de tempo), este comportamento pode ser considerado como um *outlier* global e o computador dissonante é considerado suspeito de um ataque de *hacking*. Como outro exemplo, tem-se a auditoria em transações comerciais onde as transações que não seguem os normativos são consideradas *outliers* globais e são selecionadas para um exame mais detalhado.

6.4.2 *Outliers* Contextuais

Diferentemente dos *outliers* globais, neste caso, a decisão em se classificar um valor como *outlier* depende do contexto. Variáveis como data, localização ou outros fatores ajudam a determinar se se trata de *outlier* ou não. Definir se uma temperatura é fora do normal, por exemplo, depende do local e da época. Por exemplo, se faz 28°C no inverno em Toronto, trata-se de um *outlier*. Caso seja em um dia de verão na mesma cidade, trata-se de uma temperatura normal.

Portanto, em um dado conjunto de dados, um objeto de dados é um *outlier* contextual se ele se desvia significantemente com respeito a um contexto específico do objeto. Este tipo de *outlier* também é conhecido como condicional, pois eles são condicionais em relação a um determinado contexto. Desta forma, o contexto deve ser especificado como parte do problema no caso de detecção de *outliers* condicionais. Geralmente, os atributos dos objetos de dados em questão são divididos em dois grupos:

■ Atributos contextuais: Definem o contexto do objeto de dados. No exemplo anterior, os atributos podem ser data e localização.

■ Atributos comportamentais: estes definem as características do objeto e são utilizados para avaliar se o objeto é um *outlier* no contexto. No exemplo anterior, os atributos comportamentais podem ser temperatura, umidade e pressão.

Diferentemente da detecção de *outlier* global, na detecção de *outlier* contextual, definir se um objeto de dados é um *outlier* depende não somente dos atributos comportamentais, mas também dos contextuais. Uma configuração de atributos comportamentais pode não ser considerada um *outlier* em um contexto, mas pode sê-lo em outro (e.g. 28ºC ser ou não *outlier* em Toronto, dependendo da estação).

6.4.3 *Outliers* Coletivos

Dado um conjunto de dados, um subconjunto de objetos de dados forma um *outlier* coletivo se os objetos como um todo se desviam significantemente do conjunto de dados como um todo. Por exemplo, no caso de um departamento de entregas de uma loja, se o envio de uma encomenda estiver atrasado, não é motivo suficiente para ser considerado um *outlier*. No entanto, se 100 encomendas sofrerem atraso no mesmo dia, essas 100 encomendas irão formar um *outlier* coletivo e devem ser analisadas como tal para permitir o entendimento do problema.

6.5 MÉTODOS DE DETECÇÃO DE *OUTLIERS*

Neste trabalho, decidiu-se utilizar 5 métodos de detecção de *outliers*:

- Z-Score
- Z-Score modificado
- *Isolation Forest*
- *Local Outlier Factor* (LOF)
- *Interquartile Range* (IQR) *outlier detection*

6.5.1 Z-Score

O método z-Score dá a ideia do quão distante da media um ponto de dados está (Z-SCORE: DEFINITION, FORMULA AND CALCULATION, 2019). Mais tecnicamente, ele é uma medida de quantos desvios padrões acima ou abaixo da média populacional um determinado valor está. Por exemplo:

Um z-score de 1 apresenta 1 desvio padrão acima da média.

Um z-score de 2 representa 2 desvios padrão acima da média.

Um z-score de -1.8 representa 1.8 desvios padrão abaixo da média.

O z-Score pode ser colocado em uma curva de distribuição normal, indo de -3 até +3 desvios padrões. Este método é uma maneira de comparar os resultados aos de uma população "normal". Resultados analisados isoladamente, dificilmente têm algum significado. Eles devem ser comparados com as demais observações dentro do seu universo. Por exemplo, uma pessoa que pesa 120 Kg é uma informação sem grande significado. Mas, se compararmos este dado com o peso médio de uma população, podemos concluir que seu z-Score, por exemplo, é 2,0. Isso significa que está dois desvios padrão acima da média. Para ser mais exato, podemos utilizar a tabela z para encontrarmos a porcentagem da população que está acima e abaixo deste valor. A tabela mostra que um z-score de 2.0, nos dá o valor 0,9772 (ou 97.72%). Isto indica que 97.72% da população está abaixo do peso de 120Kg e que 100% – 97.72% = 2.28% dos valores estão posicionados acima deste valor. Ou seja, apenas 2.28% da população está com peso acima de 120kg. Após esta análise, passamos de um dado sem nenhum contexto para a situação em que o valor medido está acima do normal esperado naquela população.

6.5.2 Z-Score Modificado

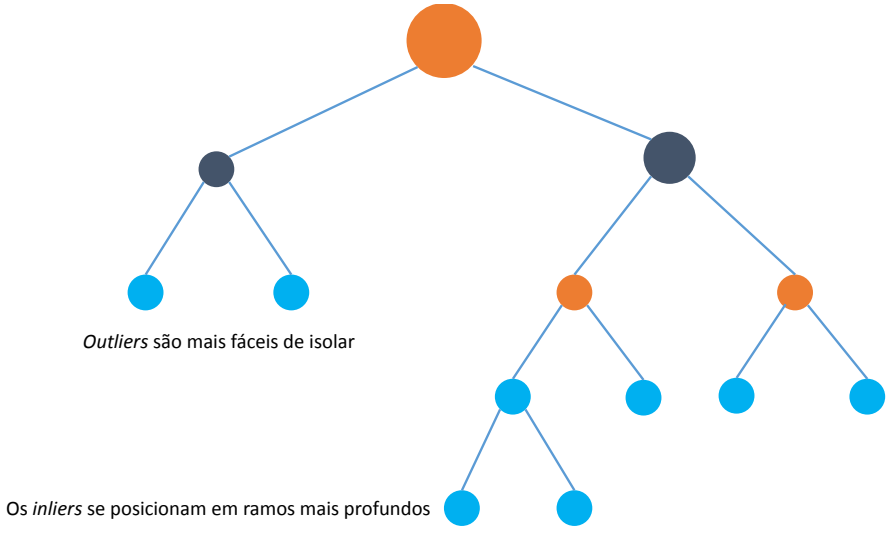
O Z-Score é calculado com base na relação de um ponto com a média e o desvio padrão de um grupo de pontos. Por isso, este cálculo pode ser influenciado por *outliers*. Assim, ele pode não se apresentar robusto o suficiente para a detecção de *outliers* em alguns casos. Como forma de ajustar isso, utiliza-se o Z-score Modificado (IGLEWICZ; HOAGLIN, 1993). Ele é calculado com base na mediana e o desvio absoluto da mediana (MAD — *Median of the absolute deviation*) ao invés da média e do desvio padrão usados no Z-score comum. Desta forma, com este método, é possível encontrar *outliers* que não seriam encontrados com o z-Score normal.

6.5.3 Isolation Forest

Isolation forest é um algoritmo não-supervisionado para detecção de anomalias que se baseia no princípio do isolamento das anomalias (LIU; TING; ZHOU, 2018).

Os autores tiraram vantagem de duas propriedades de pontos de dados anômalos em uma amostra. Primeiramente, *outliers* consistem de poucas instâncias. Em Segundo lugar, eles têm valores que são bem diferentes das instâncias normais. Desta forma, segundo a ideia deles, as anomalias são fáceis de serem isoladas em relação aos pontos que são considerados normais. O algoritmo então constrói um conjunto de "Árvores de Isolamento" para o conjunto de dados, sendo que as anomalias são os pontos com o menor caminho (Figura 20).

Figura 20. Esquema de funcionamento do algoritmo de *Isolation Tree*.



Fonte: Adaptado de VERBUS (2019)

Ele é bastante útil e é fundamentalmente diferente de todos os métodos existentes, pois introduz o uso do isolamento como um meio mais efetivo e eficiente de detecção que os

métodos que comumente utilizam a distância básica e as medidas de densidade. Além disso, este método é um algoritmo com uma complexidade linear baixa e baixos requisitos de memória. Como resultado, ele constrói um modelo com um bom desempenho com um pequeno número de árvores que utilizam pequenos subconjuntos de tamanho fixo, independentemente do tamanho do conjunto de dados.

O algoritmo de *Isolation Forest* isola as observações selecionando aleatoriamente uma característica e então selecionando aleatoriamente um valor de corte entre os valores mínimos e máximos da característica selecionada anteriormente. A ideia por trás disso funciona da seguinte maneira: isolar as observações anômalas é mais fácil, pois somente algumas poucas condições são necessárias para separá-las das observações normais. Por outro lado, isolar as observações normais requerem mais condições. Portanto, um valor anômalo pode ser calculado como um número de condições necessários para separar uma dada observação.

A maneira como o algoritmo constrói a separação é primeiro criando árvores isoladas, ou árvores de decisão aleatórias. Então, o escore é calculado como comprimento do caminho para isolar a observação.

6.5.4 *Local Outlier Factor* (LOF)

Este algoritmo foi proposto em (BREUNIG et al., 2000) para encontrar pontos de dados anômalos medindo o desvio local de um dado ponto de dados em relação aos seus vizinhos.

O LOF é baseado no conceito de densidade local, a qual é dada localmente pelo algoritmo kNN (k *Nearest Neighbors*), cuja distância é utilizada para estimar a densidade. Comparando a densidade local de um objeto às densidades locais dos seus vizinhos, pode-se identificar regiões de densidades similares e outros pontos que possuem uma densidade substancialmente menor que os seus vizinhos. Estes são considerados *outliers*.

A densidade local é estimada pela distância típica na qual um ponto pode ser alcançado a partir de seus vizinhos. A definição de 'distância de alcance' utilizada no algoritmo LOF é uma medida adicional para produzir resultados mais estáveis dentro dos clusters.

Enquanto a intuição geométrica do LOF só é aplicável a espaços vetoriais de baixa dimensão, o algoritmo pode ser aplicado em qualquer contexto onde uma função de similaridade pode ser definida. Ele desempenhou muito bem experimentalmente sempre se

mostrando melhor que seus competidores em problemas como detecção de invasão em redes ou sobre dados processados de benchmarks de classificação.

A família de métodos LOF pode ser facilmente generalizada e então aplicada a vários outros problemas, tais como detecção de *outliers* em dados geográficos, por exemplo.

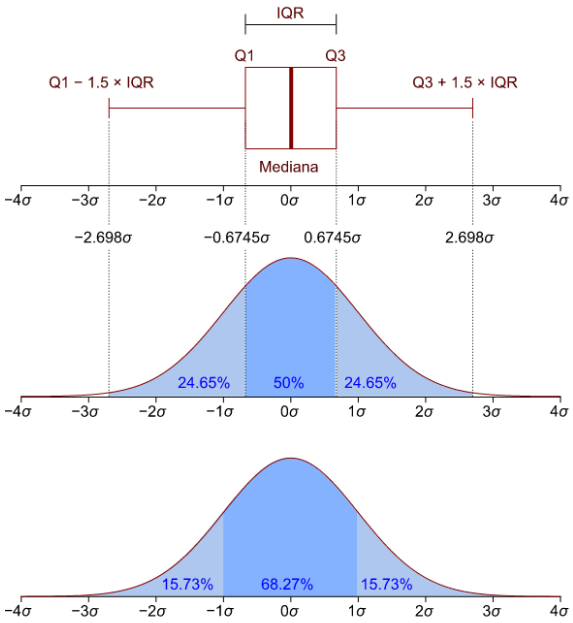
6.5.5 *Interquartile Range (IQR)*

O IQR é uma forma de descrever um conjunto de dados que pode ser aplicada na descoberta de padrões e *outliers* (MOLIN, 2019). A ideia se baseia na ideia da dispersão estatística em dividir os dados em cinco números que consistem em:

- Valor mínimo existente no conjunto de dados.
- O primeiro quartil Q1, é o valor abaixo do qual estão 25% dos valores no conjunto de dados
- A mediana dos dados (Q2) que indica o ponto central dos dados
- O terceiro quartil Q3, é o valor acima do qual estão os 25% dos valores restantes.
- O valor máximo dos dados.

A representação do IQR é geralmente feita pelo gráfico do tipo Boxplot (Figura 21). Neste é possível observar os valores mínimos, máximos e os quartis.

Figura 21. Boxplot representando os quartis de uma distribuição normal.



Fonte: Adaptada de < https://en.wikipedia.org/wiki/Interquartile_range#/media/File:Boxplot_vs_PDF.svg> Acesso em mar. 2020.

Estes cinco números são capazes de tornar a análise de um conjunto de dados muito mais intuitiva. Por exemplo, o intervalo, que é o mínimo subtraído do máximo, é um indicador do quão espalhados os dados estão no conjunto de dados. No entanto, por serem bastante influenciados pelos *outliers*, o mínimo e o máximo podem não ser uma boa estratégia para entender o espalhamento dos dados. Assim, a ideia do IQR é usar os valores Q1 e Q3 como medidas de intervalo, uma vez que eles são menos sensíveis aos *outliers*.

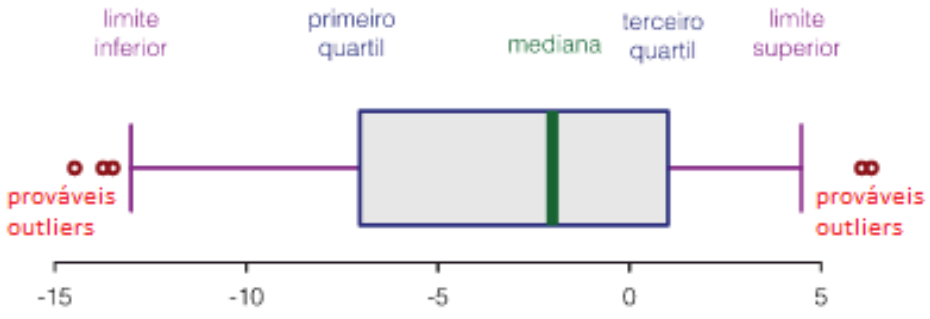
$$IQR = Q3 - Q1$$

A ideia do IQR é mostrar como os dados estão dispersos em relação à mediana (e não a média). Por isso, esta técnica é menos suscetível aos *outliers*. Apesar disso, o IQR pode ser utilizado para detectar *outliers* usando os seguintes passos:

- Calcular o intervalo interquartil (IQR = Q3 – Q1) para os dados
- Multiplicar o intervalo por 1,5 (constante utilizada para diferenciar *outliers*)
- Adicionar o valor ao Q3. Qualquer número maior que este valor resultante é um suspeito de *outlier*
- Subtrair o valor resultante do passo 2 de Q1. Qualquer número menor que isso é também um suspeito de *outlier*

A representar esta ideia com o gráfico de boxplot, os *outliers* são representados por pontos que são posicionados além dos limites inferiores e superiores, como na Figura 22.

Figura 22. Representação dos *outliers* em um boxplot



Fonte: Adaptada de < https://pt.wikipedia.org/wiki/Diagrama_de_caixa> Acesso em mar. 2020.

É importante ter a noção de que o IQR é somente um método que traça uma regra geral e não pode ser tido como verdade absoluta. Desta forma, cada *outlier* resultante deve ser

analisado segundo o seu próprio contexto para determinar se ele pode ser considerado uma anomalia.

7 FASE DE AVALIAÇÃO

A próxima fase é de avaliação. Aqui serão mostrados os resultados das análises exploratórias realizadas sobre os dados, bem como os resultados da detecção de anomalias usando os métodos já explicados.

Como estudo de caso, utilizaram-se os 1.054 procedimentos do grupo 02 – Procedimentos com finalidade diagnóstica para a avaliação. Os casos em que a quantidade total de procedimentos realizada em um ano for menor que 1.000 ou o comprimento do *dataframe* de análise de um procedimento tiver comprimento menor que 1.000 registros, a rotina desconsidera o procedimento para a análise. O fator utilizado para as taxas calculadas será de 100.000, ou seja, os valores obtidos para as taxas de atendimentos serão relativos a 100.000 habitantes.

Cada procedimento é analisado segundo os seguintes passos:

a. Plota a distribuição de atendimentos para o procedimento por região, considerando idade diferente de 999. Esta restrição da idade se faz necessária para evitar os lançamentos consolidados, pois seria impossível discernir os valores de idade desses lançamentos.
b. Plota a distribuição de atendimentos para o procedimento por sexo, considerando idade diferente de 999.
c. Plota distribuição de atendimentos, considerando idade diferente de 999
d. Plota a quantidade de atendimentos para o procedimento, por sexo e por mês
e. Calcula as taxas mensais nos níveis Municipal, Estadual, Regional e Nacional
f. Plota as taxas mensais nos níveis Municipal, Estadual, Regional e Nacional
g. Cria um *dataframe* com as taxas calculadas para que ele seja utilizado na rotina de detecção de *outliers*.
h. Calcula os *outliers* para os dados das taxas anuais de atendimentos por município.
i. Calcula as taxas mensais para cada estabelecimento nos níveis Municipal, Estadual, Regional e Nacional

j. Plota as taxas mensais para cada estabelecimento nos níveis Municipal, Estadual, Regional e Nacional
k. Cria um *dataframe* com as taxas calculadas para cada estabelecimento para que ele seja utilizado na rotina de detecção de *outliers*.
l. Calcula os *outliers* para os dados das taxas anuais de atendimentos por estabelecimento.
m. Acumula resultados no *dataframe* para posterior utilização
n. Grava dados dos *outliers* em tabelas no banco de dados, uma para municípios e outra para estabelecimentos.
o. Após a análise de todos os procedimentos, a rotina imprime sumário dos achados para o usuário

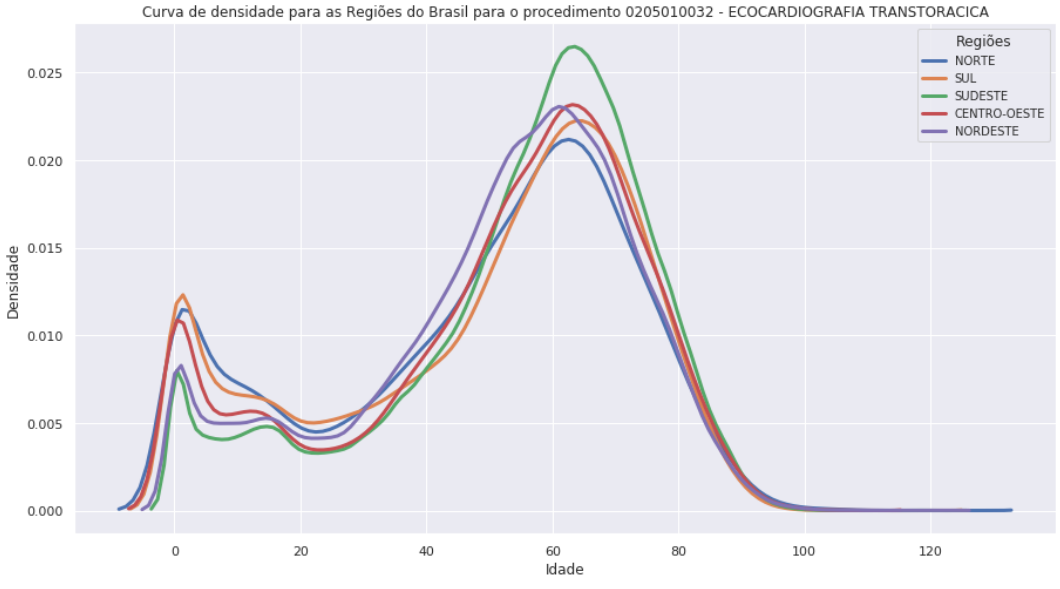
7.1 GRÁFICOS DE DISTRIBUIÇÃO DE ATENDIMENTOS

Com a finalidade de realizar uma análise exploratória, foram plotadas algumas distribuições dos procedimentos ambulatoriais do grupo 02.

7.1.1 Idade x Região

A primeira foi por idade e região. Na Figura 23, pode-se verificar que no caso do procedimento Ecocardiografia Transtorácica, a distribuição é semelhante nas diferentes regiões e se concentra na faixa de 0 a 3 e de 50 a 70 anos.

Figura 23. Curva de densidade para o procedimento Ecocardiografia Transtorácica por região

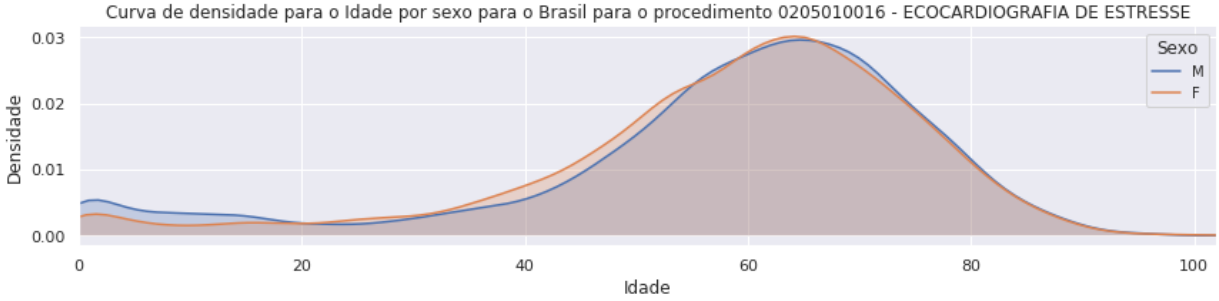


Fonte: Elaborada pelo Autor (2020)

7.1.2 Sexo x Idade

A próxima distribuição plotada foi por sexo e idade, considerando a idade diferente de 999. No gráfico da Figura 24, pode-se notar que, no caso da Ecocardiografia de Estresse, os procedimentos também se concentram na faixa entre os 50 e 70 anos e a distribuição é semelhante para ambos os sexos.

Figura 24. Curva de densidade para Idade x Sexo do procedimento Ecocardiografia de Estresse.

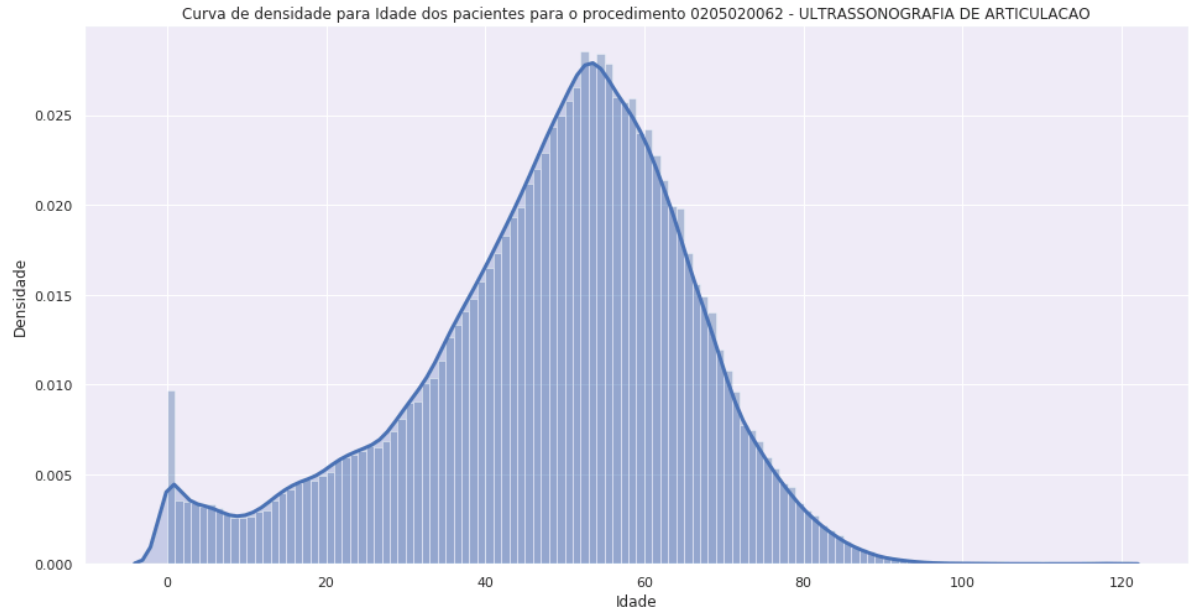


Fonte: Elaborada pelo Autor (2020)

7.1.3 Histograma de idade

Em seguida tem-se o histograma de idade com a respectiva curva, neste caso, para o procedimento Ultrassonografia de Articulação, como se pode notar na Figura 25. O histograma mostra que algumas idades têm picos, o que pode indicar problemas na qualidade dos dados. Muitas vezes, o responsável pela informação pode arredondar valores ou simplesmente usar um valor baseado em uma observação.

Figura 25. Histograma de distribuição de idade.

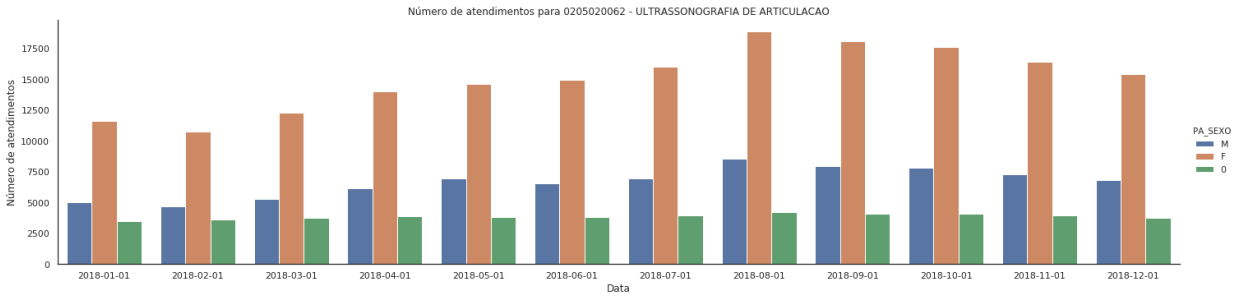


Fonte: Elaborada pelo Autor (2020)

7.1.4 Número de atendimentos x Mês x Sexo

A próxima análise realizada foi sobre o número de atendimentos. Neste caso, foi escolhido um procedimento que pode exemplificar que os dados podem ser preenchidos de forma consolidada ou individualizada para o mesmo tipo. No caso, novamente utilizou-se o procedimento de Ultrassonografia de Articulação. Pela Figura 26 pode-se visualizar que a maioria dos procedimentos é direcionada para o sexo feminino. No entanto, há uma parte dos procedimentos que é definida como sexo '0' (em verde), referente aos lançamentos consolidados.

Figura 26. Número de atendimentos x Mês x Sexo para o procedimento Ultrassonografia de Articulação.

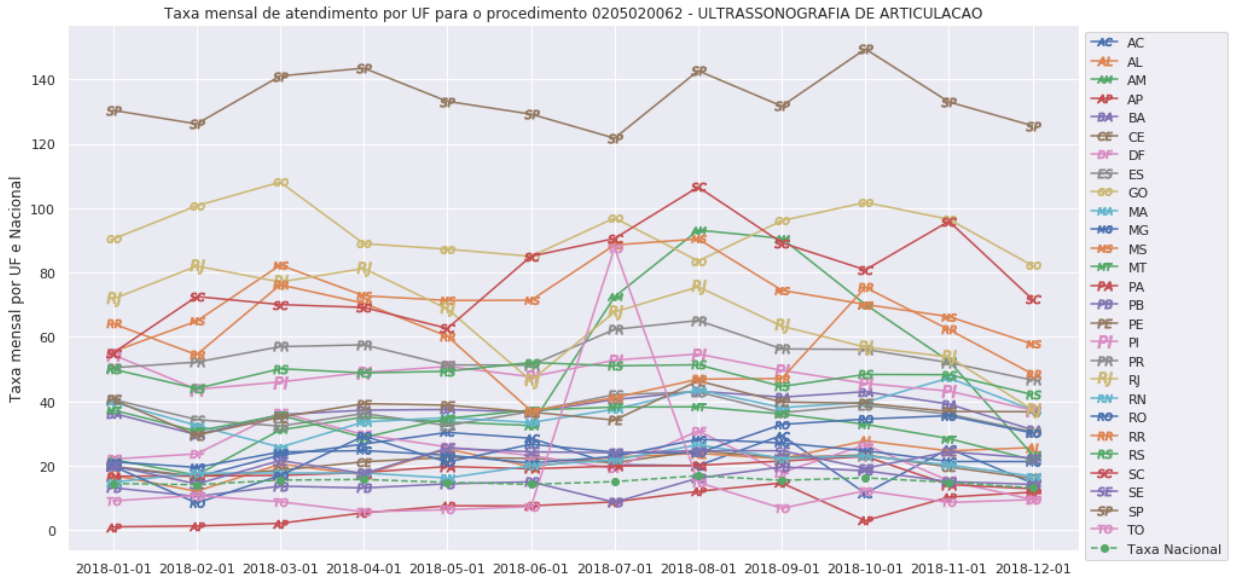


Fonte: Elaborada pelo Autor (2020)

7.1.5 Taxa mensal de atendimento por UF (Gráfico de linhas e boxplot)

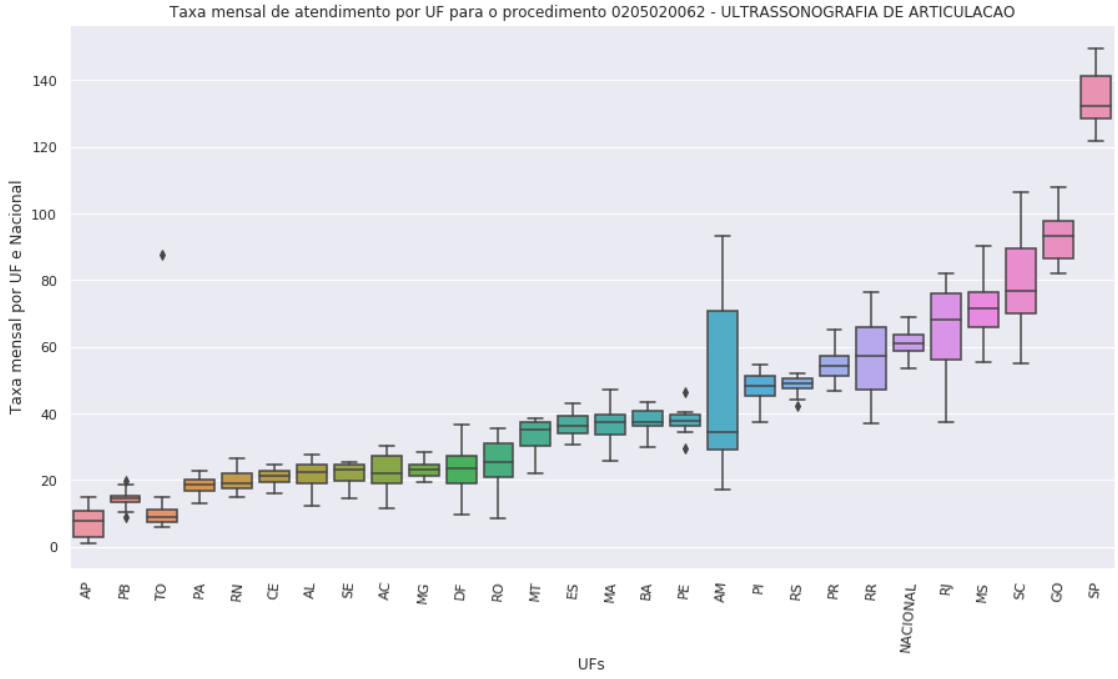
Como forma de analisar visualmente o resultado do cálculo das taxas mensais de atendimento plotaram-se dois gráficos um de linhas e outro com boxplots. Primeiramente, a análise foi feita levando em consideração os Estados e a taxa nacional. Nas figuras Figura 27 e Figura 28 pode-se notar a taxa nacional em linha tracejada verde e os estados que se apresentam bem acima desta taxa, como São Paulo, Santa Catarina e Goiás. Importante lembrar que as taxas são calculadas com base na população de cada UF.

Figura 27. Gráfico de linha para a taxa mensal de atendimento por UF para Ultrassonografia de Articulação



Fonte: Elaborada pelo Autor (2020)

Figura 28. Gráfico boxplot para a taxa mensal de atendimento por UF para Ultrassonografia de Articulação

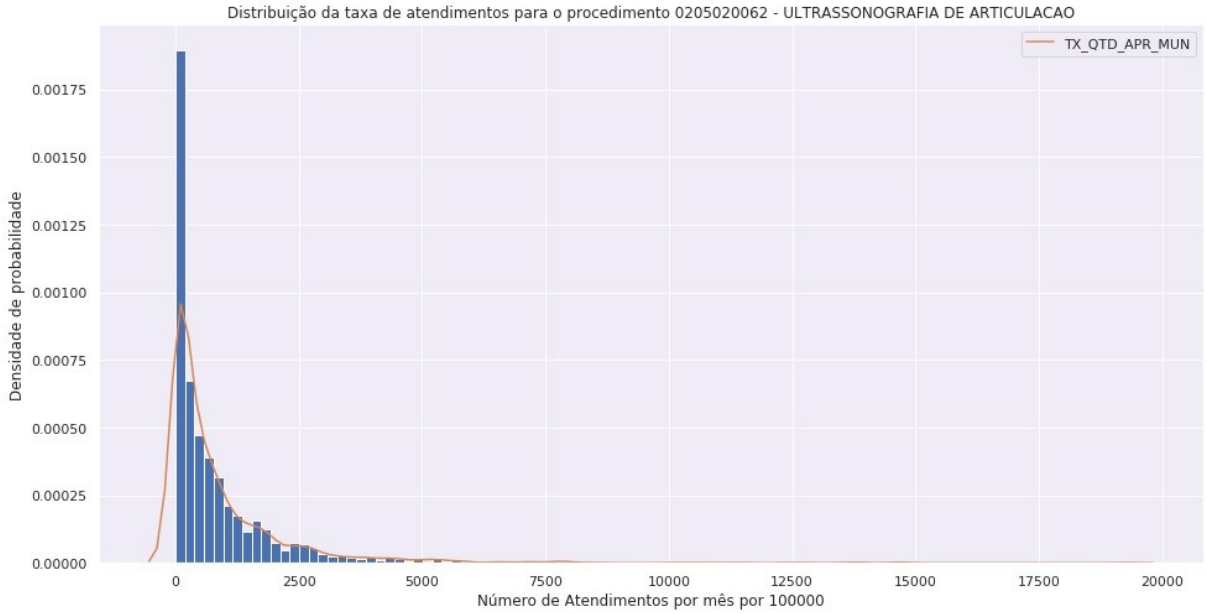


Fonte: Elaborada pelo Autor (2020)

7.1.6 Gráfico com a Distribuição da taxa de atendimentos

Por último, dentro da análise exploratória optou-se por plotar o gráfico com a distribuição da taxa média de atendimento anual a exemplo do que foi feito no InfoSAS. Verificou-se que no caso da Ultrassonografia de Articulação que se trata de uma distribuição log-normal (Figura 29). Este é o caso da maioria dos procedimentos analisados no grupo 02.

Figura 29. Distribuição da taxa média anual de atendimentos para Ultrassonografia de Articulação para o ano de 2018.



Fonte: Elaborada pelo Autor (2020)

7.2 DETECÇÃO DE ANOMALIAS PARA MUNICÍPIOS

Após a plotagem dos gráficos para a análise, passa-se para a detecção de anomalias que funciona conforme descrito em 6.3. Os resultados iniciais são exibidos para conferência, conforme a Figura 30.

Figura 30. Resumo dos resultados da análise de *outliers*.

```
--- Cálculo dos outliers finalizado ---  0205020062 - ULTRASSONOGRAFIA DE ARTICULACAO --

* Z-score  : 26
* Z-score 2: 34
* IQR      : 133
* Isolation: 151
* LOF      : 150

-- Outliers comuns encontrados --

- Procedimento:  0205020062 - ULTRASSONOGRAFIA DE ARTICULACAO --
* Outliers : 26
* Índices  : [101, 128, 134, 148, 736, 771, 810, 821, 859, 879, 904, 932, 938, 952, 996, 1023, 1045, 1095, 1106, 1149, 1233, 1270, 1276, 1284, 1349, 1458]
* Mediana da quantidade de procedimentos aprovados, desconsiderando os outliers ( 0205020062 - ULTRASSONOGRAFIA DE ARTICULACAO ): 375.6442736088058
* Média da quantidade de procedimentos aprovados, desconsiderando os outliers ( 0205020062 - ULTRASSONOGRAFIA DE ARTICULACAO ): 767.3605796879561
```

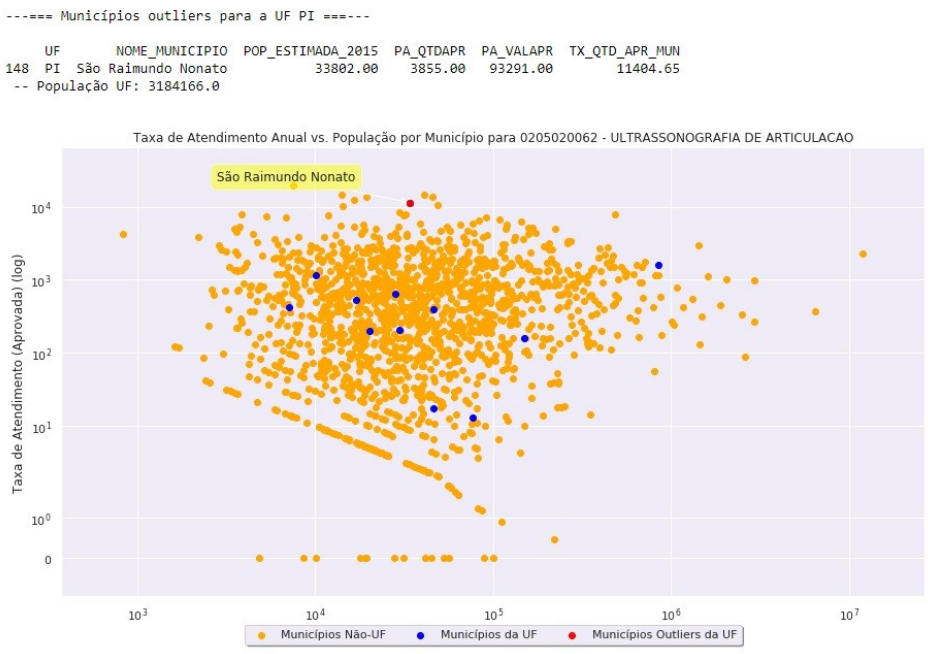
Fonte: Elaborada pelo Autor (2020)

O texto traz informações de quantos *outliers* cada método encontrou. Verifica-se que os métodos IQR, *Isolation Forest* e LOF trazem um número elevado de *outliers*, na casa das centenas. Fazendo a triagem dos *outliers* comuns, foram levantados 26 *outliers* em comum, cujos índices são listados para conferência, juntamente com o cálculo da mediana e da média.

Em seguida, a rotina lista detalhes dos *outliers* encontrados organizando-os por UF. Portanto, para cada UF são listados todos os municípios que apresentaram uma taxa de atendimento acima da normalidade. Na Figura 31, tem-se os detalhes para o estado do Piauí para o procedimento Ultrassonografia de Articulação. No caso, apenas um município foi encontrado nesta UF. O resumo dos dados é exibido antes da plotagem do gráfico. Tem-se como colunas o índice da linha selecionada, UF, Nome do Município, População para o município no ano de 2015, quantidade e valor aprovados pelo SUS e a taxa calculada para a quantidade aprovada.

Além disso, são plotados três gráficos que auxiliam na análise dos *outliers*. O primeiro é um gráfico de dispersão com eixos logarítmicos (Figura 31) que mostra os municípios da UF que realizaram o procedimento em análise (em azul) e os *outliers* encontrados (em vermelho) com uma legenda identificando o município em questão. Os pontos amarelos refletem todos os demais municípios do país, não pertencentes à UF em questão, que realizaram o procedimento. Como os *outliers* são destacados por UF, *outliers* de outras UF também são apresentados em amarelo e serão devidamente destacados na plotagem do gráfico da UF correspondente.

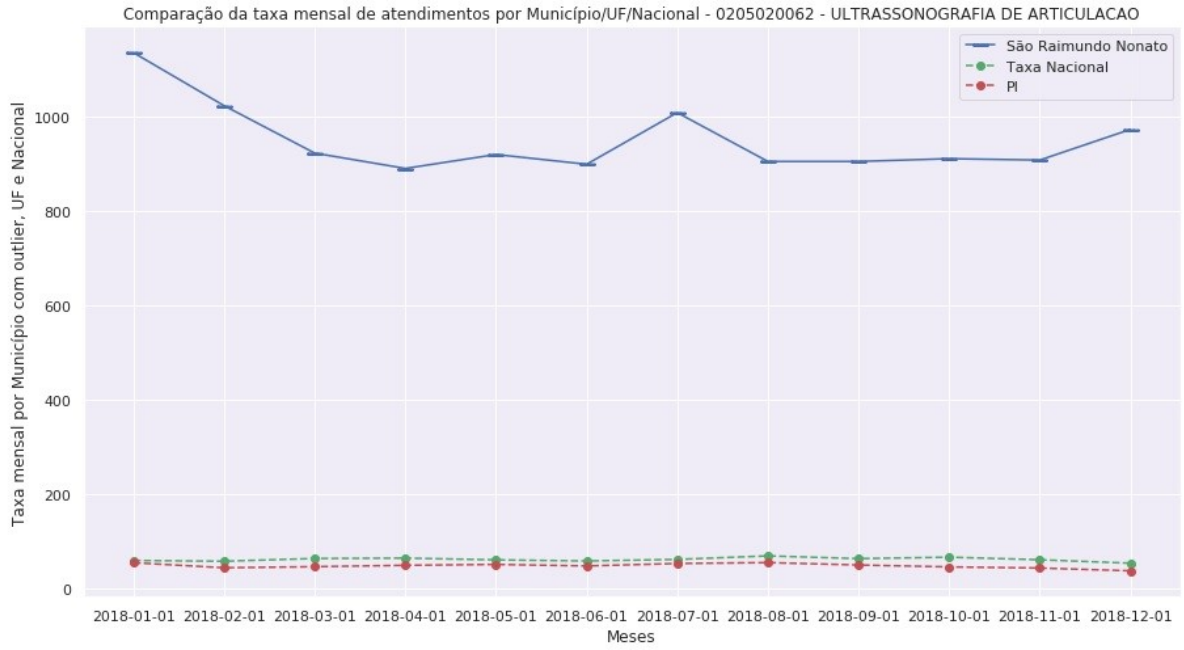
Figura 31. Dados para os municípios com *outliers* no Piauí e o gráfico de dispersão para a UF.



Fonte: Elaborada pelo Autor (2020)

Em seguida, tem-se os gráficos de linha (Figura 32) e boxplot (Figura 33) com as taxas para os municípios *outliers* da UF, a taxa da UF e a taxa Nacional, para efeitos de comparação. Os demais municípios da UF e do país não são listados para evitar que os gráficos fiquem poluídos.

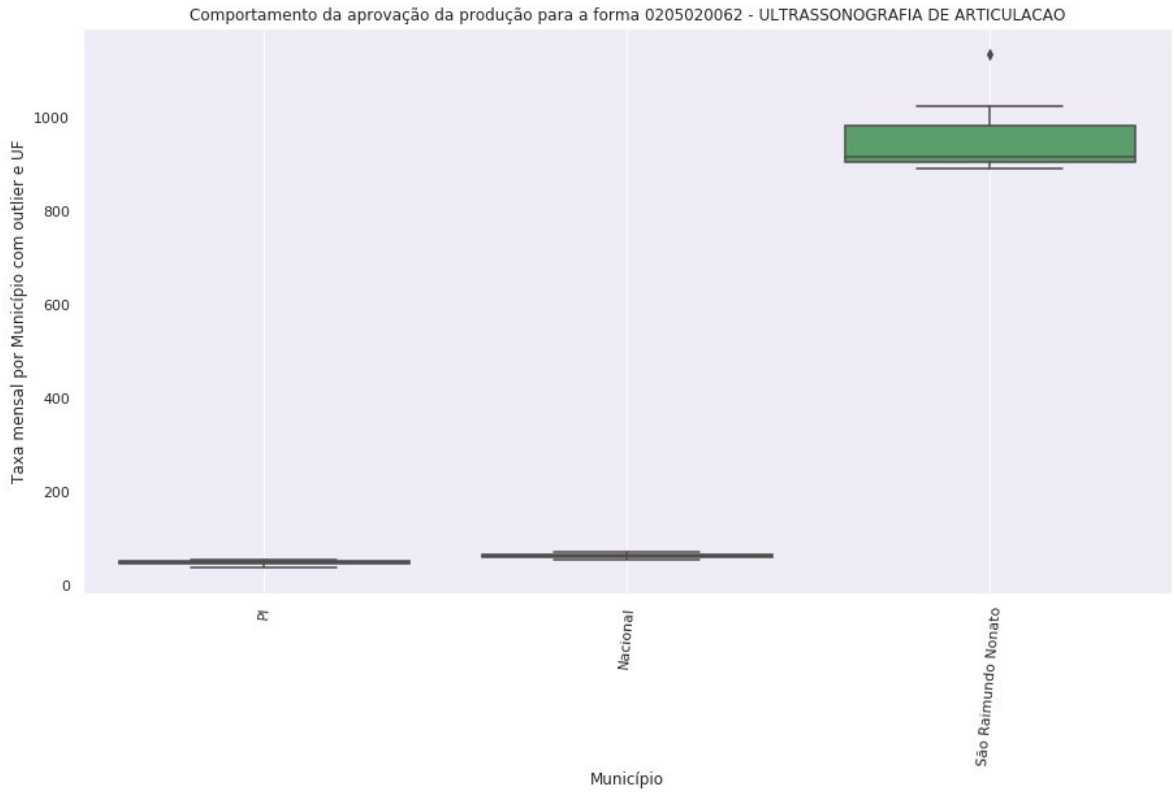
Figura 32. Gráfico de linha comparando *outliers* da UF com os valores de taxa da UF e nacional.



Fonte: Elaborada pelo Autor (2020)

Percebe-se que o gráfico boxplot apresenta o *outlier* de uma forma muito mais clara que o gráfico de linhas. Na Figura 33 é possível verificar que os valores para o município São Raimundo Nonato estão bem acima dos valores da UF e Nacional e também é possível visualizar um valor acima da normalidade comparado aos valores do próprio município representado por um ponto. Este valor é o que pode ser visualizado no mês de janeiro no gráfico de linhas.

Figura 33. Gráfico boxplot comparando *outliers* da UF com os valores de taxa da UF e nacional



Fonte: Elaborada pelo Autor (2020)

Este procedimento é repetido para cada UF até listar todos os municípios *outliers*. Ao final, um resumo com as informações de cada UF é exibido de acordo com o que se pode ver na Figura 34. Nele é possível visualizar os seguintes dados:

- Dados da UF
  - Taxa média de atendimentos para a UF por 100.000 habitantes: calculada dividindo o número de atendimentos para a UF (PA_QTDAPR) no período definido pela população da UF, multiplicado pelo fator de 100.000 habitantes.
  - População da UF
- Dados do município
  - Taxa média registrada por 100.000 habitantes: calculada dividindo o número de atendimentos para o município (PA_QTDAPR) no período definido pela população da UF, multiplicado pelo fator de 100.000 habitantes.

- Taxa média projetada por 100.000 habitantes: trata-se da taxa média esperada para o município com base na proporção sobre a taxa calculada para a UF.
- Taxa média projetada sem *outliers*: trata-se da taxa média esperada para o município com base na proporção sobre a taxa calculada para a UF desconsiderando os *outliers*.
- Proporção: indica quantas vezes a taxa encontrada para o município foi maior que a taxa média projetada. Ela dá uma ideia da dimensão do valor do *outlier*.
- População do município: dado obtido do IBGE.
- Valor Aprovado: Soma dos recursos aprovada (PA_VALAPR) para o procedimento em análise na janela de tempo analisada para o município.

Figura 34. Resumo da rotina de detecção de *outliers* para o estado do Maranhão para o procedimento Ultrassonografia de Articulação

```
##########################################################################################################################
Resumo dos MUNICÍPIOS levantados na detecção de outliers - 0205020062 - ULTRASSONOGRAFIA DE ARTICULACAO
##########################################################################################################################
-------------
UF: MA
-------------
- Taxa média de atendimentos para a UF: 1355.5500376536604 para 100000 hab.
- População da UF: 6794301.0 hab.
        * Axixá
        -  Taxa média registrada             : 7696.20253164557 atendimentos para 100000 hab.
        -  Taxa média projetada              : 2.3642267168021958 atendimentos para 100000 hab.
        -  Taxa média projetada sem outliers : 1.473714669458204 atendimentos para 100000 hab.
        -  Proporção                         : 3255.272633944047 vezes maior que a taxa projetada

        -  População do município            : 11850.0 hab.
        -  Valor Aprovado                    : R$ 22070.399999999998

        * Santo Antônio dos Lopes
        -  Taxa média registrada             : 10124.184382235318 atendimentos para 100000 hab.
        -  Taxa média projetada              : 2.843655982665122 atendimentos para 100000 hab.
        -  Taxa média projetada sem outliers : 1.7725616188850446 atendimentos para 100000 hab.
        -  Proporção                         : 3560.27045604397 vezes maior que a taxa projetada

        -  População do município            : 14253.0 hab.
        -  Valor Aprovado                    : R$ 34920.6

        * São Luís Gonzaga do Maranhão
        -  Taxa média registrada             : 5648.41188799914 atendimentos para 100000 hab.
        -  Taxa média projetada              : 3.7123347273872116 atendimentos para 100000 hab.
        -  Taxa média projetada sem outliers : 2.314042941317198 atendimentos para 100000 hab.
        -  Proporção                         : 1521.5254826912023 vezes maior que a taxa projetada

        -  População do município            : 18607.0 hab.
        -  Valor Aprovado                    : R$ 25434.199999999997
```

Fonte: Elaborada pelo Autor (2020)

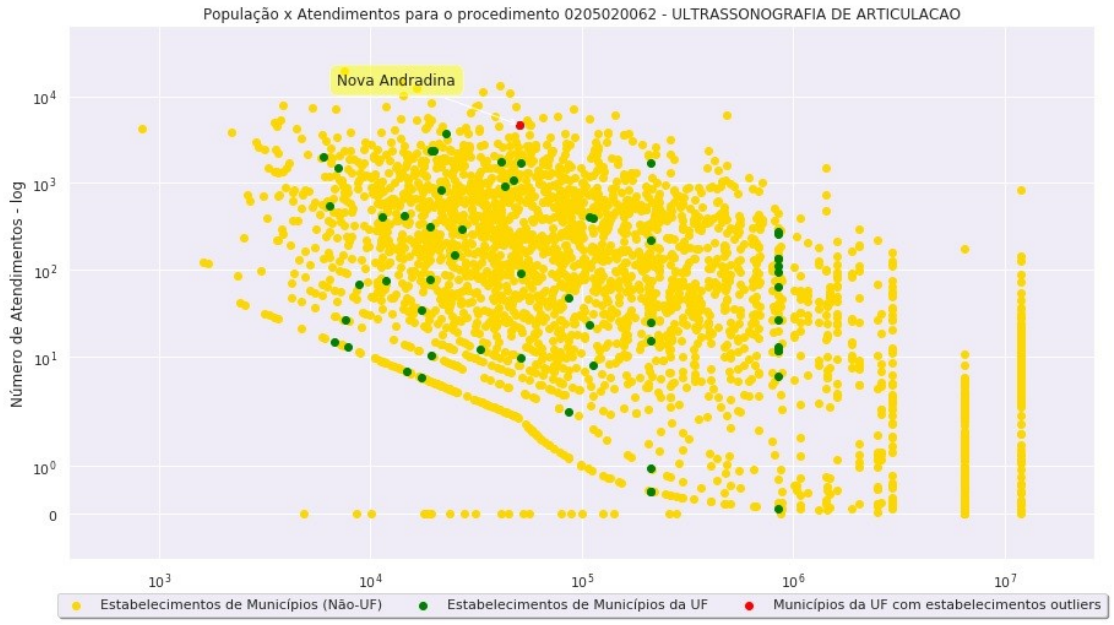
7.3 DETECÇÃO DE ANOMALIAS PARA ESTABELECIMENTOS

Em seguida, os mesmos passos realizados na detecção de anomalias para os municípios foram aplicados para os estabelecimentos. Isto se mostrou útil, pois, durante os estudos preliminares, verificou-se que alguns municípios não eram listados como *outliers*, mas possuíam estabelecimentos com taxa de atendimento muito alta. Por exemplo, um município pode apresentar um número total de atendimentos para um procedimento dentro do esperado. No entanto, neste município, um determinado estabelecimento pode ser um *outlier* dentre aqueles aptos a realizarem o procedimento. Assim, com este passo, foi possível encontrar municípios que não foram listados no passo anterior.

Os cálculos realizados são basicamente os mesmos do passo anterior agrupando os valores por estabelecimentos ao invés de municípios. Como exemplo para ilustrar o caso foi escolhido um estabelecimento de Nova Andradina, MS. Este município não foi detectado na rotina de municípios. As Figuras 35 a 38 mostram os gráficos para o estabelecimento Centro de Referência A Saúde Da Mulher.

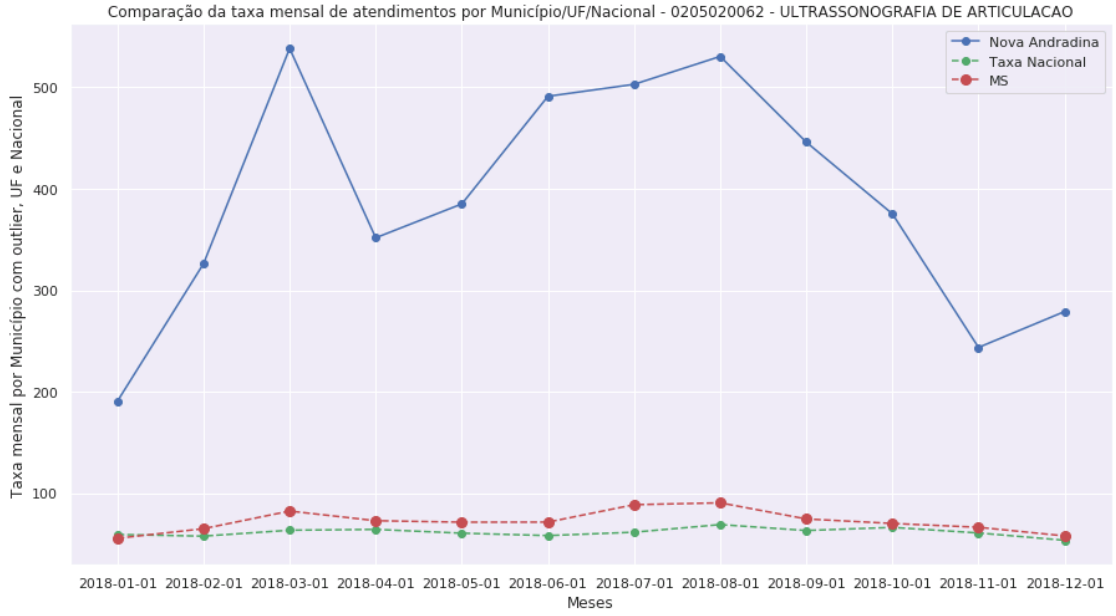
Figura 35. Dados para os municípios com estabelecimentos *outliers* no Mato Grosso do Sul e o gráfico de dispersão para a mesma UF.

```
---=== Estabelecimentos outliers para a UF MS ===---

     UF  NOME_MUNICIPIO                      DS_NOME_FANTASIA  POP_ESTIMADA_2015  PA_QTDAPR  PA_VALAPR  TX_QTD_APR_EST
987  MS  Nova Andradina  CENTRO DE REFERENCIA A SAUDE DA MULHER           50893.00    2372.00   57402.40         4660.76
 -- População UF: 2587269.0
```



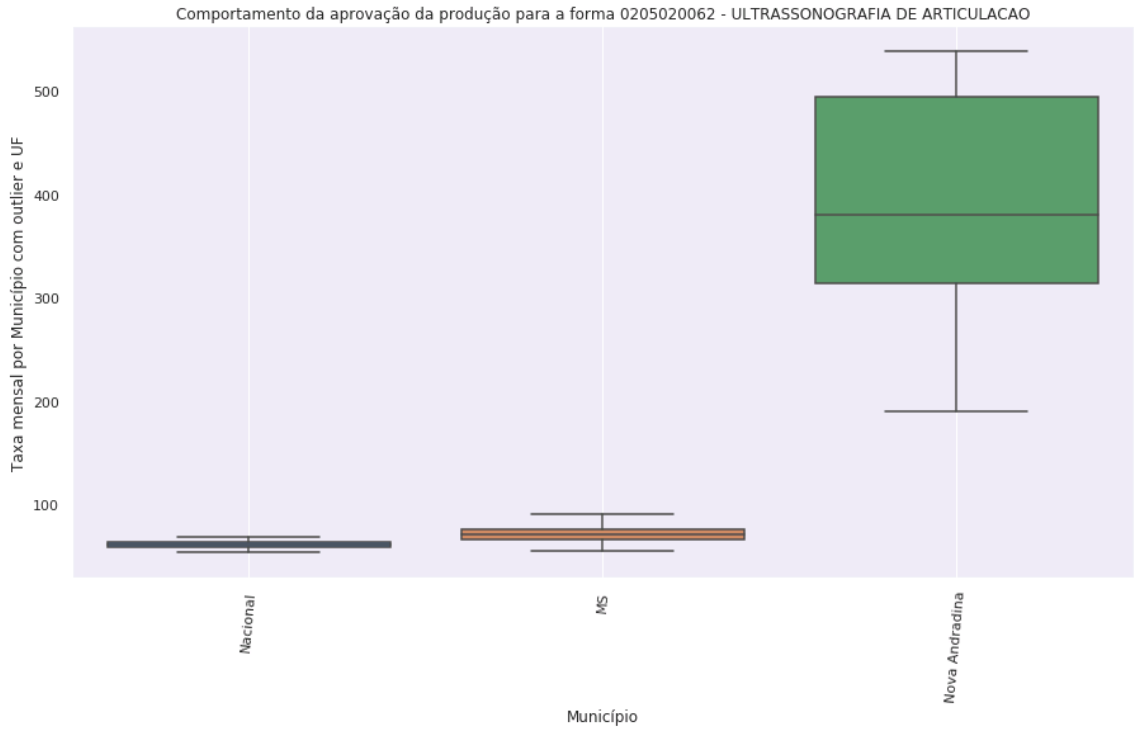
Fonte: Elaborada pelo Autor (2020)

Figura 36. Gráfico de linha comparando estabelecimentos *outliers* da UF com os valores de taxa da UF e nacional



Fonte: Elaborada pelo Autor (2020)

Figura 37. Gráfico boxplot comparando estabelecimentos *outliers* da UF com os valores de taxa da UF e nacional



Fonte: Elaborada pelo Autor (2020)

Figura 38. Resumo da rotina de detecção de estabelecimentos *outliers* para o estado do Mato Grosso do Sul para o procedimento Ultrassonografia de Articulação

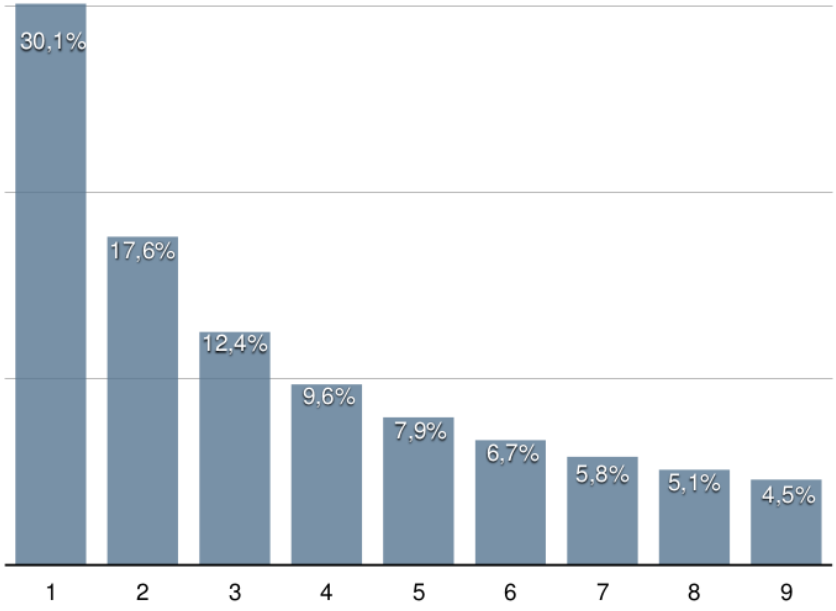
```
-------------
UF: MS
-------------
- Taxa média de atendimentos para a UF: 526.3508113926098 para 100000 hab.
- População da UF: 2587269.0 hab.
        * Município: Nova Andradina
        - Estabelecimento: CENTRO DE REFERENCIA A SAUDE DA MULHER
        -  Taxa média registrada            : 4660.7588469927105 atendimentos para 100000 hab.
        -  Taxa média projetada             : 10.353609092909972 atendimentos para 100000 hab.
        -  Taxa média projetada sem outliers : 8.874953216028375 atendimentos para 100000 hab.
        -  Proporção                        : 450.15789230292097 vezes maior que a taxa projetada

        -  População do município           : 50893.0 hab.
        -  Valor Aprovado                   : R$ 57402.39999999999
```

Fonte: Elaborada pelo Autor (2020)

7.4 APLICAÇÃO DA LEI DE BENFORD

Como passo seguinte, aplicou-se a Lei de Benford na análise dos dados (FEWSTER, 2019). Esta Lei, também conhecida como a Lei dos Primeiros Dígitos ou o Fenômeno de Dígitos Significativos, afirma que os primeiros algarismos dos números encontrados em séries de registros das mais variadas fontes não apresentam uma distribuição uniforme. Na realidade, o que ocorre é que os números são organizados de tal forma que o dígito "1" é o mais frequente, seguido de "2", "3", e assim sucessivamente até "9", que apresenta a menor frequência como o primeiro dígito, como pode ser visto na Figura 39.

Figura 39. Distribuição da frequência do primeiro dígito segundo a Lei de Benford



Fonte: BENFORD_PY, (2019)

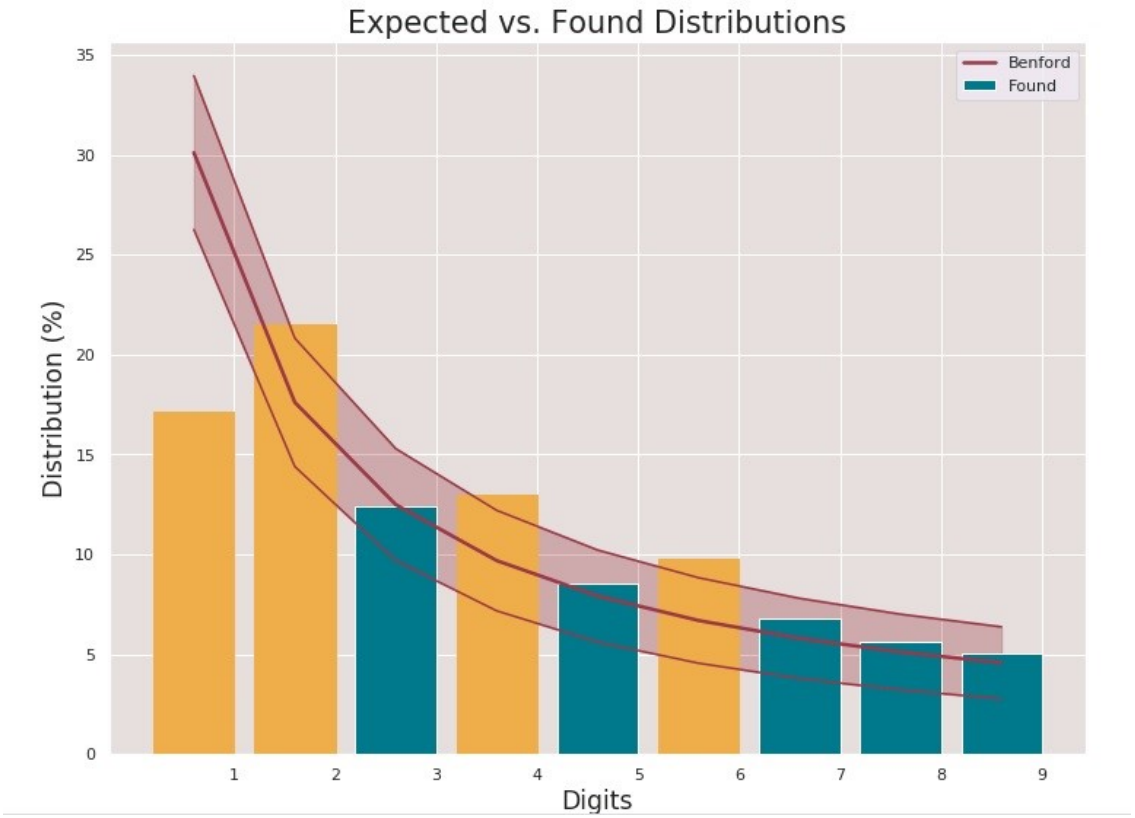
A Lei de Benford pode ser aplicada para detecção de fraude em eleições, relatórios financeiros e uma infinidade de aplicações. Contudo, a Lei de Benford não se aplica em todos os casos. Por exemplo, a Lei não se aplica à altura de todos os seres humanos, visto que a altura média do ser humano varia entre 1 a 2 metros. A Lei não se aplica também aos números de celulares, visto que os números começam geralmente com 9. Neste trabalho, ela só pode ser aplicada aos lançamentos consolidados, uma vez que as quantidades dos lançamentos para procedimentos individualizados geralmente e de uma ou duas unidades. Por esta razão, neste trabalho, a Lei de Benford só foi aplicada aos casos de *outliers* com dados consolidados.

Vale ressaltar que nem todos os casos em que a Lei de Benford não é seguida significa que se está diante de uma fraude. Em muitos casos, a distribuição favorece determinado dígito por uma imposição de uma regra do negócio. Por exemplo, se um determinado campo requer justificativa para valores a partir de 1000, vai ser muito comum encontrar uma distribuição maior para o número 9, pois, para evitar a justificativa, muitas pessoas irão optar por valores similares a 999.

No presente trabalho, utilizou-se a biblioteca Benford_py (BENFORD_PY, 2019) para analisar as quantidades lançadas para os procedimentos nos casos em que foram encontrados *outliers* e em casos em que tem-se números consolidados que permitam a análise. Como exemplo, pode-se ver na Figura 40 o resultado da aplicação da Lei para os resultados do município de Tupã-SP para o procedimento Ultrassonografia de Articulação. Verifica-se que, para os 571 registros testados, há desvio das distribuições dos dígitos 1, 2, 4 e 6 dentro do intervalo de confiança de 95%. Isso configura como mais um indício de que os números referentes a este município devem ser avaliados com maior proximidade.

Figura 40. Resultado da análise da Lei de Benford para o Município de Tupã-SP para o procedimento Ultrassonografia de Articulação

```
### Município Tupã - SP ###
Initialized sequence with 571 registries.

Test performed on 571 registries.
Discarded 0 records < 1 after preparation.

The entries with the significant positive deviations are:

             Expected  Found  Z_score
First_1_Dig
6                0.07   0.10     2.89
4                0.10   0.13     2.57
2                0.18   0.22     2.41
```



Fonte: Elaborada pelo Autor (2020)

7.5 RESUMO FINAL DOS ACHADOS

Após a realização da análise para todos os procedimentos listados, o sistema emite um resumo da avaliação baseado nos dados acumulados em dois *dataframes* de *outliers*: um para municípios e outro para estabelecimentos.

A ideia de acumular os valores surgiu como uma necessidade de possibilitar o entendimento da ordem de grandeza de valores e quantidades envolvidas para cada município ou estabelecimento. Olhando somente a informação de um procedimento isolado nos

deparamos muitas vezes com valores pequenos que nem chegam a justificar uma análise mais detalhada. No entanto, se conseguirmos ter uma visão totalizada (além da individualizada apenas), é possível ter uma ideia dos totais envolvidos para cada município ou estabelecimento.

As informações listadas são as seguintes:

- Municípios com o maior número de *outliers*
- Municípios com *outliers* com os maiores valores aprovados
- Procedimentos com maior número de *outliers*
- Procedimentos com *outliers* com os maiores valores aprovados
- Estabelecimentos com o maior número de *outliers*
- Estabelecimentos com *outliers* com os maiores valores aprovados

7.6 RESULTADOS

7.6.1 Municípios com o maior número de *outliers*

Segundo a Tabela 8, o município com o maior número de *outliers* foi Pariquera-Açu em SP. Somente para o grupo 02, foram 116 procedimentos testados com *outliers*.

Tabela 8. Os 10 municípios com o maior número de *outliers* para procedimentos do grupo 02

	NOME_MUNICIPIO	**UF**	**Qtde de *outliers***
1	Pariquera-Açu	SP	116
2	Botucatu	SP	89
3	Barretos	SP	87
4	Lindóia	SP	83
5	Campina Grande do Sul	PR	82
6	Cruzmaltina	PR	53
7	Barueri	SP	53
8	Dracena	SP	53
9	Porangatu	GO	51
10	Tenente Portela	RS	48

Fonte: Elaborada pelo Autor (2020)

7.6.2 Municípios com *outliers* com os maiores valores aprovados

Segundo a Tabela 9, o município com *outliers* que teve o maior valor autorizado pelo SUS para os procedimentos do grupo 02 foi Barretos-SP. Este município é o terceiro da lista na quantidade de *outliers* (Tabela 8). Sua população é de apenas 122.098 habitantes. No entanto, Barretos é conhecido por ser um centro de tratamento de câncer que atende pessoas de todo o Brasil. Assim, o movimento acima do esperado pode ser justificado pela localização do Hospital de Câncer de Barretos. Já Pariquera-Açu foi o 21º lugar na aprovação de recursos (R$ 4.319.686,50), valor este bastante alto considerando que os 20 primeiros municípios têm populações muito maiores.

Tabela 9. Os 10 municípios com *outliers* com o maior valor autorizado para procedimentos do grupo 02

	NOME_MUNICIPIO	UF	Valor
1	Barretos	SP	R$ 30.190.580,03
2	Salvador	BA	R$ 23.486.369,28
3	Recife	PE	R$ 19.024.389,92
4	Florianópolis	SC	R$ 16.057.169,05
5	Vitória	ES	R$ 15.161.993,46
6	Barueri	SP	R$ 14.630.752,55
7	Goiânia	GO	R$ 13.536.662,52
8	Belford Roxo	RJ	R$ 12.506.968,50
9	Botucatu	SP	R$ 11.838.742,65
10	Cuiabá	MT	R$ 8.364.286,78

Fonte: Elaborada pelo Autor (2020)

7.6.3 Procedimentos com maior número de *outliers*

Na Tabela 10 é possível visualizar os 20 procedimentos do grupo 02 com o maior número de *outliers*. O primeiro lugar é o exame de Determinação de Fator Reumatoide, com 48 municípios dos quais os 20 com o fator de relação mais alto podem ser conferidos na Tabela 11.

Tabela 10. Tabela com os 20 procedimentos do grupo 02 com o maior número de *outliers*.

	Código	Nome do Procedimento	Qtde
1	0202030075	Determinação de fator reumatoide	48
2	0202010600	Dosagem de potássio	43
3	0202010210	Dosagem de cálcio	41
4	0202020096	Determinação de tempo de sangramento -duke	39
5	0202010422	Dosagem de fosfatase alcalina	39
6	0202030300	Pesquisa de anticorpos anti-hiv-1 + hiv-2 (elisa)	39
7	0202010678	Dosagem de triglicerídeos	38
8	0204040124	Radiografia de punho (ap + lateral + obliqua)	37
9	0202040127	Pesquisa de ovos e cistos de parasitas	36
10	0202010473	Dosagem de glicose	36
11	0202010562	Dosagem de magnésio	36
12	0202010694	Dosagem de ureia	35
13	0204020069	Radiografia de coluna lombo-sacra	35
14	0205020038	Ultrassonografia de abdômen superior	35
15	0205020046	Ultrassonografia de abdômen total	34
16	0202010651	Dosagem de transaminase glutamico-piruvica (tgp)	34
17	0202010643	Dosagem de transaminase glutamico-oxalacetica (tgo)	33
18	0204040019	Radiografia de antebraço	33
19	0204040051	Radiografia de braço	33
20	0204010080	Radiografia de crânio (pa + lateral)	33

Fonte: Elaborada pelo Autor (2020)

Tabela 11. Os 20 municípios com as maiores relações para o procedimento Determinação de Fator Reumatoide.

	Município	UF	Taxa de atendimento por 100.000 hab.	Relação (quantas vezes maior que o esperado)
1	Sales	SP	9.030,32	66.032,75
2	Pejuçara	RS	19.630,08	65.738,45
3	Lindóia	SP	10.060,12	58.987,97
4	Cruzmaltina	PR	14.871,30	41.537,32
5	Santa Maria do Salto	MG	12.386,42	36.989,26
6	Santa Inês	PR	6.855,52	34.141,45
7	Taboleiro Grande	RN	27.243,84	33.252,38
8	São João das Missões	MG	23.869,74	30.384,25
9	Mato Verde	MG	24.195,42	30.218,43
10	Montezuma	MG	12.461,13	24.957,87

Fonte: Elaborada pelo Autor (2020)

7.6.4 Procedimentos com *outliers* com os maiores valores aprovados

Na Tabela 12 é possível conferir os procedimentos do grupo 02 que apresentaram *outliers* com os maiores valores aprovados. O procedimento Sorologia de doador de sangue apresentou o valor somado de R$ 13.838.775,00 para o ano de 2018. Os dois *outliers* identificados foram respectivamente em Vitória-ES e Florianópolis-SC. Importante notar que os valores apresentados se referem somente aos municípios com *outliers*. Para este mesmo procedimento, foi aprovada a soma total de R$ 302.749.875,00 com a quantidade total de 4.036.665 procedimentos

Tabela 12. Os 10 procedimentos do grupo 02 com os maiores valores aprovados.

	Código	Nome do Procedimento	Valor	Qtd. Aprovada	Qtd. de *Outliers*
1	0212010050	Sorologia de doador de sangue	R$ 13.838.775,00	184.517	2
2	0202030300	Pesquisa de anticorpos anti-HIV-1 + HIV-2 (Elisa)	R$ 11.420.830,00	1.142.087	39
3	0202030679	Pesquisa de anticorpos contra o vírus da hepatite c (anti-HCV)	R$ 11.191.066,60	603.301	23
4	0202030636	Pesquisa de anticorpos contra antígeno de superfície do vírus da hepatite b (anti-HBS)	R$ 10.269.150,15	553.593	26
5	0202030318	Pesquisa de anticorpos anti-HTLV-1 + HTLV-2	R$ 9.729.437,90	524.498	15

6	0211020052	Monitorização ambulatorial de pressão arterial (M.A.P.A)	R$ 7.715.946,17	766.231	6
7	0206030037	Tomografia computadorizada de pelve / bacia / abdômen inferior	R$ 6.858.719,25	49.475	11
8	0206030010	Tomografia computadorizada de abdômen superior	R$ 6.565.932,69	47.363	11
9	0206010079	Tomografia computadorizada do crânio	R$ 6.368.580,96	65.359	13
10	0202030784	Pesquisa de anticorpos IGG e IGM contra antígeno central do vírus da hepatite B (anti-HBC-total)	R$ 6.355.990,55	342.693	22

Fonte: Elaborada pelo Autor (2020)

7.6.5 Estabelecimentos com o maior número de *outliers*

Nesta Tabela 13, é possível visualizar os 20 estabelecimentos onde foi encontrado o maior número de *outliers* para os procedimentos do grupo 02. O Hospital das Clínicas de Botucatu apareceu como *outlier* em 113 procedimentos, totalizando o valor de R$ 14.014.546,20 somente para as anomalias. Interessante notar que dois estabelecimentos do município de Pariquera-Açu (linhas 5 e 12) foram listados.

Tabela 13. Os 20 estabelecimentos com o maior número de *outliers* levantados para os procedimentos do grupo 02.

	Estabelecimento	**Município**	**UF**	**Qtd. de *Outliers***	**Valor**	**Qtd. Aprovada**
1	HOSPITAL DAS CLINICAS DE BOTUCATU	Botucatu	SP	113	R$ 14.014.546,20	2.096.752
2	CONISCA	Lindóia	SP	105	R$ 1.905.418,99	321.826
3	HOSPITAL ANGELINA CARON	Campina Grande do Sul	PR	82	R$ 5.302.651,08	318.690
4	FUNDACAO PIO XII BARRETOS	Barretos	SP	80	R$ 28.646.353,03	1.167.332
5	AME PARIQUERA ACU	Pariquera-Açu	SP	67	R$ 2.641.862,53	128.740
6	HOSPITAL MUNICIPAL DE PORANGATU	Porangatu	GO	66	R$ 772.516,79	98.337

7	HOSPITAL SANTO ANTONIO TENENTE PORTELA	Tenente Portela	RS	65	R$ 2.496.359,69	275.058
8	LABORATORIO DE ANALISES CLINICAS JOAO ADROALDO	Cruzmaltina	PR	57	R$ 513.726,46	71.442
9	HOSPITAL NOSSA SENHORA DA POMPEIA	São Félix	BA	56	R$ 1.248.507,17	195.018
10	AME AMBULATORIO MED DE ESPECIALIDADES DRACENA	Dracena	SP	52	R$ 2.214.189,95	191.087
11	HOSPITAL MUNICIPAL SENHORA SANTANA	Brasília de Minas	MG	52	R$ 592.395,85	83.385
12	HOSPITAL DR LEOPOLDO BEVILACQUA	Pariquera-Açu	SP	51	R$ 946.756,95	137.877
13	HOSPITAL MUNICIPAL GETULIO VARGAS	Aragarças	GO	48	R$ 816.225,33	113.842
14	HOSPITAL MUNICIPAL SAO MARCOS CAMPINACU	Campinaçu	GO	47	R$ 128.338,54	33.549
15	LABORATORIO CENTRAL	Barueri	SP	45	R$ 11.241.214,76	2.874.587
16	HOSPITAL MUNICIPAL DE IPORA	Iporá	GO	45	R$ 821.684,44	169.770
17	HOSPITAL DE CARIDADE SAO ROQUE	Faxinal do Soturno	RS	42	R$ 2.842.259,17	165.140
18	AME AMB MEDICO DE ESP ELIANA N Z M GIANTOMASSI CASA BRANCA	Casa Branca	SP	42	R$ 880.825,62	37.579
19	HOSPITAL SAO GERALDO	São João da Ponte	MG	40	R$ 103.398,35	13.199
20	HOSPITAL REGIONAL DE MORROS	Morros	MA	39	R$ 415.533,97	88.129

Fonte: Elaborada pelo Autor (2020)

7.6.6 Estabelecimentos com *outliers* com os maiores valores aprovados

Por último, na Tabela 14 tem-se os estabelecimentos com *outliers* que apresentaram os 20 maiores valores aprovados para os procedimentos do grupo 02. A Fundação Pio XII em Barretos foi a que teve o maior valor listado sendo este o dobro do segundo colocado. Ela apresentou 80 procedimentos com *outlier*.

Tabela 14. Os 20 estabelecimentos com os maiores valores autorizados para os procedimentos do grupo 02.

	Estabelecimento	**Município**	**UF**	**Valor**	**Qtd. Aprovada**	**Qtd. de *Outliers***
1	FUNDACAO PIO XII BARRETOS	Barretos	SP	R$ 28.646.353,03	1.167.332	80
2	HOSPITAL DAS CLINICAS DE BOTUCATU	Botucatu	SP	R$ 14.014.546,20	2.096.752	113
3	HEMOSC	Florianópolis	SC	R$ 13.671.365,22	416.141	7
4	APAE GOIANIA	Goiânia	GO	R$ 11.936.836,21	870.582	10
5	LABORATORIO CENTRAL	Barueri	SP	R$ 11.241.214,76	2.874.587	45
6	CHECK UP LABORATORIO DE ANALISES CLINICAS	Uberlândia	MG	R$ 10.415.287,58	1.642.772	20
7	HEMOCENTRO DE RIBEIRAO PRETO	Ribeirão Preto	SP	R$ 9.856.605,00	246.803	3
8	HOSPITAL GERAL PIRAJUSSARA TABOAO DA SERRA	Taboão da Serra	SP	R$ 8.261.542,10	1.580.464	35
9	HOSPITAL AMARAL CARVALHO JAU	Jaú	SP	R$ 6.855.485,33	100.429	26
10	VITALAB MEDICINA DIAGNOSTICA	Feira de Santana	BA	R$ 6.693.254,63	481.054	8
11	HOSPITAL DA RESTAURACAO	Recife	PE	R$ 6.081.717,71	644.147	5
12	HEMOES	Vitória	ES	R$ 5.454.125,39	146.601	5
13	HOSPITAL ANGELINA CARON	Campina Grande do Sul	PR	R$ 5.302.651,08	318.690	82

14	LABORATORIO DO CENTRO DE SAUDE	Itaqui	RS	R$ 5.295.669,27	1.935.009	2
15	LABORATORIO DE ANALISES CLINICAS MUNICIPAL INDAIATUBA	Indaiatuba	SP	R$ 5.055.157,55	508.376	32
16	LABORATORIO CENTRAL	Belo Horizonte	MG	R$ 4.763.705,63	288.443	3
17	CENTRO DE HEMATOLOGIA E HEMOTERAPIA DO PIAUI HEMOPI	Teresina	PI	R$ 4.423.785,71	120.384	3
18	LABORATORIO CENTRAL MUNICIPAL	Vitória	ES	R$ 3.995.582,31	363.984	13
19	HOSPITAL DO ROCIO	Campo Largo	PR	R$ 3.638.818,91	156.422	17
20	HOSPITAL MUNICIPAL DE PAULINIA	Paulínia	SP	R$ 3.600.144,50	993.200	38

Fonte: Elaborada pelo Autor (2020)

7.7 PAINEL DE CONSULTA DOS DADOS

Após a construção da rotina de análise e obtenção dos resultados passou-se à criação de um painel para a exibição e avaliação dos dados obtidos. Criou-se então um painel simples, inicialmente como uma prova de conceito que permitisse que o usuário tivesse acesso aos dados dos *outliers*, bem como aos dados detalhados dos procedimentos. Desta forma, ao consultar os *outliers*, o usuário consegue explorar mais informações de um município ou de um estabelecimento de interesse.

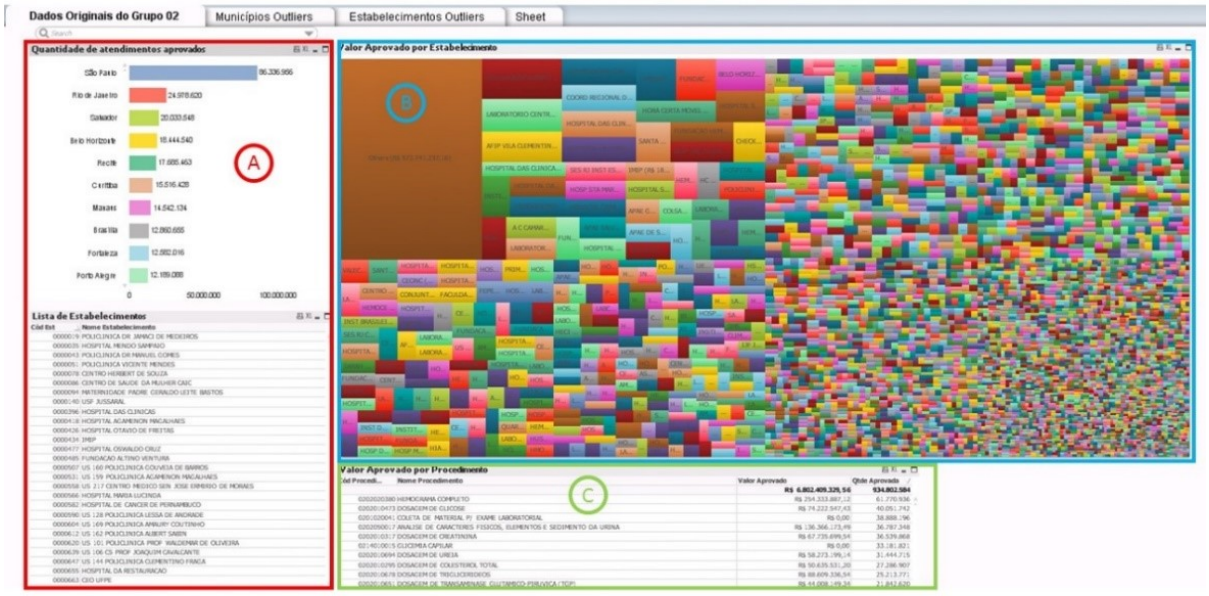
Depois de avaliar o SAS Visual Analytics e o QlikView, decidiu-se fazer a construção do painel utilizando esta última ferramenta, pois o SAS não foi capaz de suportar o volume de dados a ser carregado. O QlikView, por sua vez, conseguiu lidar com a massa de dados referente ao Grupo 02, que foi o recorte feito para este trabalho.

Para organizar os dados, criou-se uma aba com os dados originais e outras duas com os dados dos *outliers*: uma para municípios e outra para os estabelecimentos. Assim, o usuário pode analisar os dados da forma mais adequada. Por exemplo, ele pode consultar os municípios

outliers para um determinado procedimento e, a partir disto, tentar visualizar os dados originais para entender melhor a dinâmica daquele município, como por exemplo, o número de estabelecimentos existentes, outros procedimentos realizados, valores aprovados e quantidades de atendimentos realizados. Ou ainda, a partir de um município nos dados originais, o usuário pode verificar os dados dos *outliers* para aquele município.

A aba de dados originais se baseia na totalidade dos dados. Ela pode ser vista na Figura 41 com as áreas principais indicadas por cores e letras. Na área A (vermelha), tem-se a listagem da quantidade de atendimentos por município, ordenada decrescentemente. Logo abaixo está a Lista de Estabelecimentos. A área B (azul) possui um mapa com o valor aprovado para cada estabelecimento. Como não há nenhuma seleção inicial, são exibidos todos os estabelecimentos do país. Quanto maior o valor, maior a área exibida. Isto já dá uma amostra visual dos estabelecimentos com maior valor aprovado. A área C (verde) mostra os procedimentos juntamente com os valores e quantidades aprovados.

Figura 41. Aba de Dados Originais para o Grupo 02

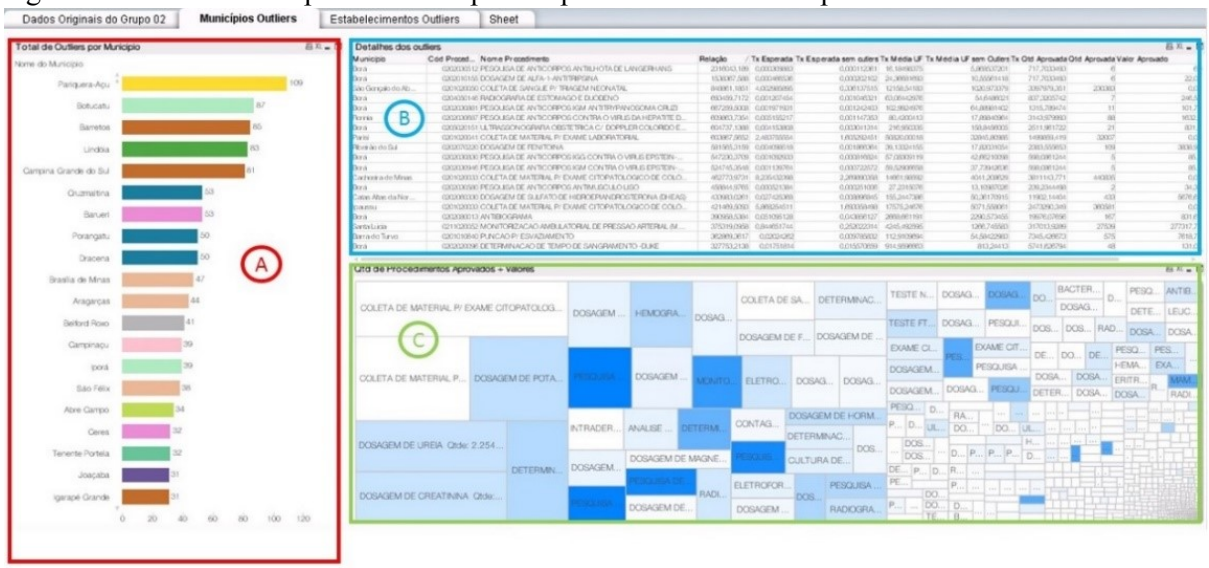


Fonte: Elaborada pelo Autor (2020)

A aba de Municípios *Outliers* pode ser vista na Figura 42. Ela, diferentemente da aba anterior, só carrega os dados dos *outliers*. Na área A (vermelha), tem-se os municípios com os respectivos números de *outliers* encontrados, em ordem decrescente. Na área B (azul), tem-se

os detalhes dos *outliers*. São as colunas: Município, código do procedimento, relação, taxa esperada, taxa esperada (sem *outliers*), taxa média da UF, taxa média da UF (sem *outliers*), taxa de atendimento, quantidade aprovada e valor aprovado. Na área C (azul) tem-se um mapa da distribuição dos procedimentos com *outliers*. Interessante notar que a área do retângulo está relacionada com a quantidade de procedimentos aprovados e a cor indica o valor. Quanto mais intensa a cor, maior o valor. Isto é útil pois nem sempre os procedimentos com uma grande quantidade são necessariamente os que apresentam maiores valores. Assim, o usuário pode rapidamente visualizar também aqueles que apresentam valores aprovados que podem ser interessantes para a análise.

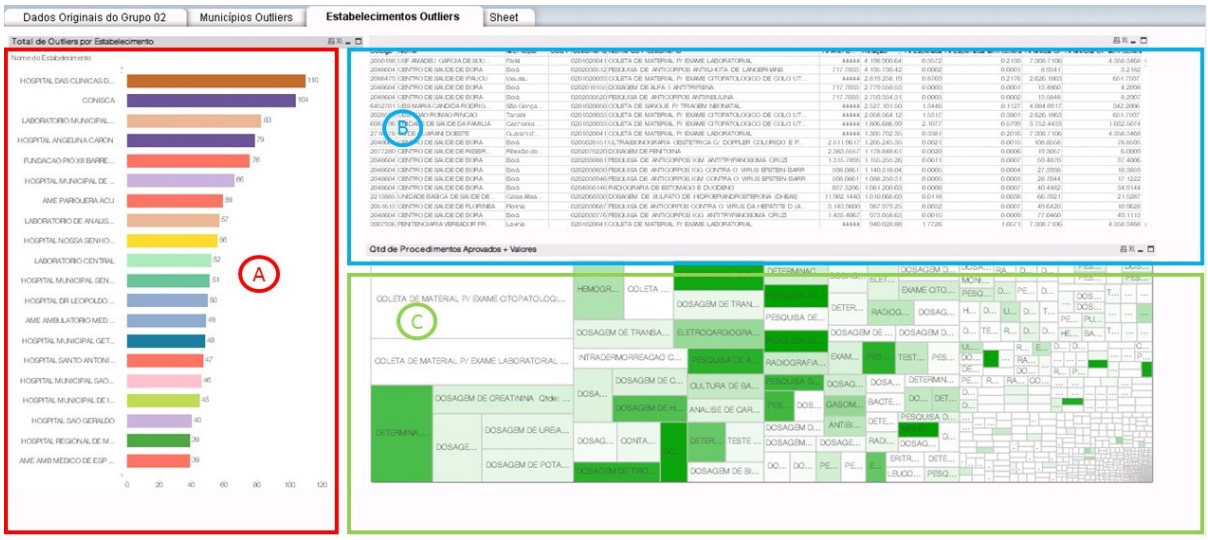
Figura 42. Aba de Municípios *Outliers* para os procedimentos do Grupo 02 do Painel de Resultados



Fonte: Elaborada pelo Autor (2020)

A ideia da última aba (Figura 43) é muito semelhante com a diferença de que os dados focam nos estabelecimentos ao invés dos municípios. Na área A (vermelha), tem-se os estabelecimentos com os respectivos números de *outliers* encontrados, em ordem decrescente. Na área B (azul), tem-se os detalhes dos *outliers*. São as colunas: Código e nome do estabelecimento, município, código do procedimento, relação, taxa esperada, taxa esperada (sem *outliers*), taxa média da UF, taxa média da UF (sem *outliers*), taxa de atendimento, quantidade aprovada e valor aprovado. Na área C (azul) tem-se um mapa da distribuição dos procedimentos com *outliers* nos estabelecimentos listados em A. Aqui também, a exemplo da aba anterior, utilizou-se o recurso da intensidade da cor para sinalizar o valor aprovado para o procedimento. A área do retângulo está relacionada com a quantidade de procedimentos aprovados para os estabelecimentos selecionados.

Figura 43. Aba de Estabelecimentos *Outliers* para os procedimentos do Grupo 02 do Painel de Resultados.



Fonte: Elaborada pelo Autor (2020)

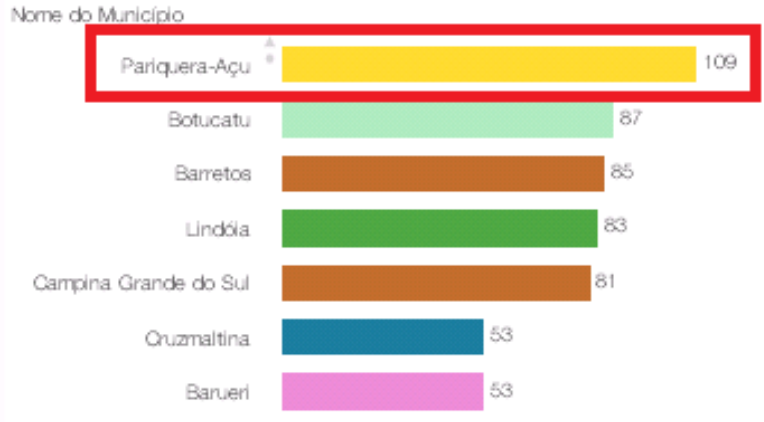
8 VALIDAÇÃO EM UM ESTUDO DE CASO

8.1 ESTUDO DE CASO

Com a finalidade de avaliar o modelo criado, decidiu-se selecionar um dos casos de *outlier* encontrados na avaliação dos procedimentos do grupo 02. Para tal, contamos com o auxílio do painel criado.

Analisando os municípios com *outliers*, decidiu-se por avaliar o caso de Pariquera-Açu em São Paulo. Ao avaliarmos o município no painel (Figura 44), percebe-se que ele apresentou *outliers* para 109 procedimentos. Avaliando a aba de dados originais, tem-se valores superlativos para o município. Foram 549.342 atendimentos e R$ 5.335.872,91 em valores aprovados, considerando todos os procedimentos e não só os que geraram *outliers*.

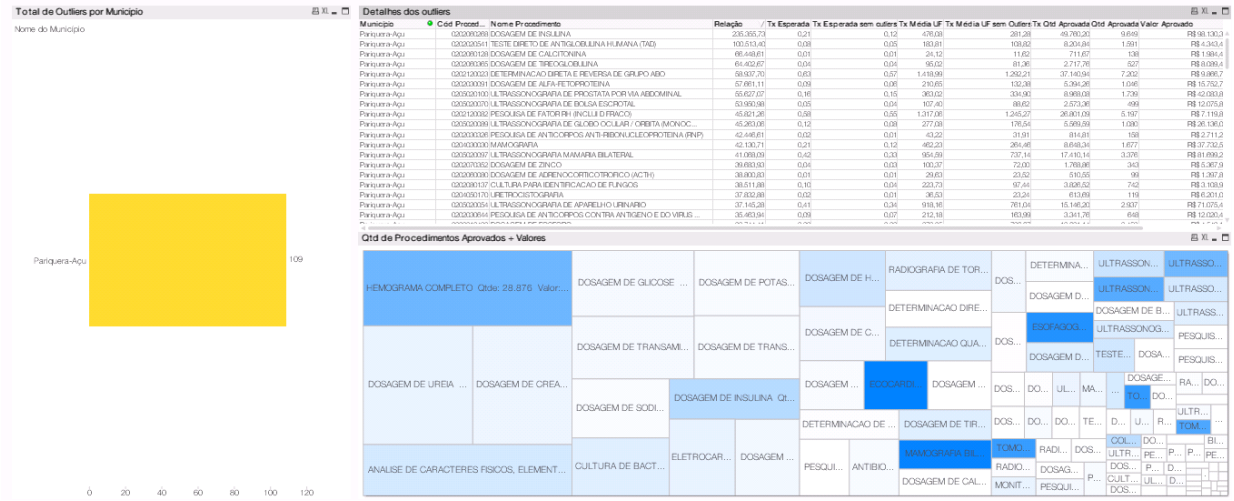
Figura 44. Recorte do painel mostrando os municípios com mais *outliers*.



Fonte: Elaborada pelo Autor (2020)

Para um município com a população de 19.391 habitantes (IBGE 2015) parece um resultado fora do normal. Ao selecionar o município no painel tem-se uma filtragem dos dados relacionados a ele, como se pode ver na Figura 45. A relação de procedimentos já vem ordenada por aqueles que tiveram a maior relação entre a taxa esperada e a taxa calculada.

Figura 45. Dados de *outliers* para o município de Pariquera-Açu



Fonte: Elaborada pelo Autor (2020)

Como exemplo, na Tabela 15 verifica-se que a taxa da Dosagem de Insulina é 235.355 vezes maior do que a taxa esperada com base no comportamento da UF. O valor aprovado para este procedimento no ano de 2018 foi de R$ 98.130,33. Não se trata de um valor alto, considerando o orçamento do SUS, mas ao ser somado aos demais 108 procedimentos *outliers*, tem-se o valor de R$ 3.698.292,26.

Tabela 15. Procedimentos da Pariquera-Açu com as maiores relações entre a taxa esperada e a calculada

Código	Procedimento	Relação	Tx Esperada	Tx Esperada sem *outliers*	Tx Média UF	Tx Média UF sem *Outliers*	Tx Qtd Aprovada	Qtd Aprovada	Valor Aprovado
0202060268	DOSAGEM DE INSULINA	235.355,73	0,21	0,12	476,08	281,28	49.760,20	9649,0	R$ 98.130,33
0202020541	TESTE DIRETO DE ANTIGLOBULINA HUMANA (TAD)	100.513,40	0,08	0,05	183,81	108,82	8.204,84	1591,0	R$ 4.343,43
0202060128	DOSAGEM DE CALCITONINA	66.448,61	0,01	0,01	24,12	11,62	711,67	138,0	R$ 1.984,44
0202060365	DOSAGEM DE TIREOGLOBULINA	64.402,67	0,04	0,04	95,02	81,36	2.717,76	527,0	R$ 8.089,45
0202120023	DETERMINACAO DIRETA E REVERSA DE GRUPO ABO	58.937,70	0,63	0,57	1.418,99	1.292,21	37.140,94	7202,0	R$ 9.866,74
0202030091	DOSAGEM DE ALFA-FETOPROTEINA	57.661,11	0,09	0,06	210,65	132,38	5.394,26	1046,0	R$ 15.752,76
0205020100	ULTRASSONOGRAFIA DE PROSTATA POR VIA ABDOMINAL	55.627,07	0,16	0,15	363,02	334,90	8.968,08	1739,0	R$ 42.083,80
0205020070	ULTRASSONOGRAFIA DE BOLSA ESCROTAL	53.950,98	0,05	0,04	107,40	88,62	2.573,36	499,0	R$ 12.075,80
0202120082	PESQUISA DE FATOR RH (INCLUI D FRACO)	45.821,26	0,58	0,55	1.317,06	1.245,27	26.801,09	5197,0	R$ 7.119,89
0205020089	ULTRASSONOGRAFIA DE GLOBO OCULAR / ORBITA (MONOCULAR)	45.263,06	0,12	0,08	277,08	176,54	5.569,59	1080,0	R$ 26.136,00

Fonte: Elaborada pelo Autor (2020)

Considerando os valores aprovados para os *outliers*, os 10 maiores valores são listados na Tabela 16. Verifica-se que foram realizadas 5.767 Ecocardiografias Transtorácicas, 4.934 Mamografias bilaterais para rastreamento e 3.709 Esofagogastroduodenoscopias. Somente esses 3 procedimentos, tem-se quase 14.500 procedimentos em um ano. O Hemograma Completo, por sua vez, conta com 28.876 procedimentos. É como se cada habitante tivesse realizado o exame 1,5 vez.

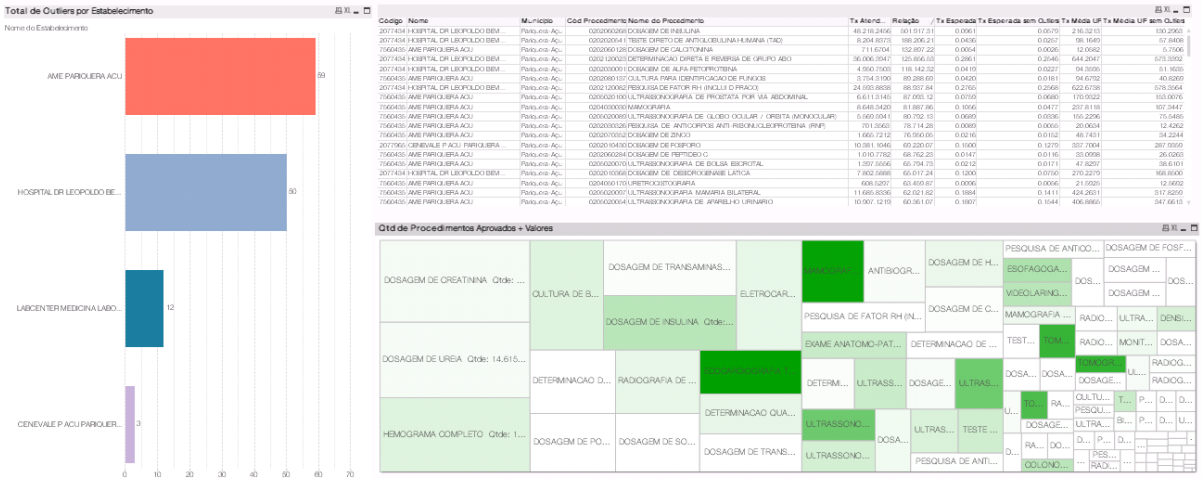
Tabela 16. Procedimentos de Pariquera-Açu com os 10 maiores valores aprovados.

Código	Procedimento	Relação	Tx Esperada	Tx Esperada sem *outliers*	Tx Média UF	Tx Média UF sem *Outliers*	Tx Qtd Aprovada	Qtd Aprovada	Valor Aprovado
0205010032	ECOCARDIOGRAFIA TRANSTORACICA	30.188,54	0,99	0,67	2.218,34	1.514,87	29.740,60	5767,0	R$ 230.333,98
0204030188	MAMOGRAFIA BILATERAL PARA RASTREAMENTO	12.482,26	2,04	1,67	4.590,14	3.751,32	25.444,79	4934,0	R$ 222.030,00
0209010037	ESOFAGOGASTRODUODENOSCOPIA	30.084,43	0,64	0,59	1.431,64	1.322,41	19.127,43	3709,0	R$ 178.625,44
0206030010	TOMOGRAFIA COMPUTADORIZADA DE ABDOMEN SUPERIOR	18.729,61	0,32	0,22	718,58	489,22	5.977,00	1159,0	R$ 160.672,17
0206010079	TOMOGRAFIA COMPUTADORIZADA DO CRANIO	15.547,40	0,52	0,46	1.177,86	1.043,08	8.132,64	1577,0	R$ 153.662,88
0206030037	TOMOGRAFIA COMPUTADORIZADA DE PELVE / BACIA / ABDOMEN INFERIOR	16.808,72	0,30	0,19	670,82	433,07	5.007,48	971,0	R$ 134.609,73
0205010040	ULTRASSONOGRAFIA DOPPLER COLORIDO DE VASOS	23.355,35	0,72	0,45	1.616,41	1.006,41	16.765,51	3251,0	R$ 128.739,60
0202020380	HEMOGRAMA COMPLETO	10.299,78	14,46	13,49	32.555,83	30.387,13	148.914,44	28876,0	R$ 118.680,36
0205020046	ULTRASSONOGRAFIA DE ABDOMEN TOTAL	20.420,82	0,78	0,71	1.767,37	1.594,96	16.028,05	3108,0	R$ 117.948,60
0202060268	DOSAGEM DE INSULINA	235.355,73	0,21	0,12	476,08	281,28	49.760,20	9649,0	R$ 98.130,33

Fonte: Elaborada pelo Autor (2020)

Verificando os estabelecimentos na Figura 46, percebe-se que os *outliers* estão distribuídos em 4 unidades. As duas mais significantes são o AME de Pariquera-Açu e o Hospital Dr. Leopoldo Bevilacqua.

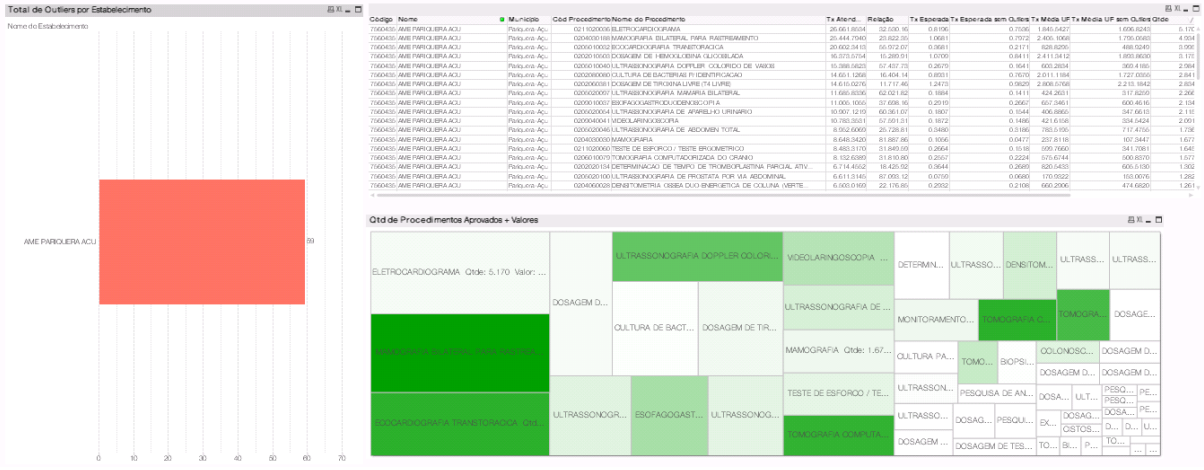
Figura 46. Estabelecimentos com *outliers* em Pariquera-Açu.



Fonte: Elaborada pelo Autor (2020)

O AME Pariquera-Açu (Figura 47) apresenta como procedimentos mais frequentes o Eletrocardiograma (5.170 atendimentos), Mamografia bilateral para rastreamento (4.934) e Ecocardiografia Transtorácica (3.995).

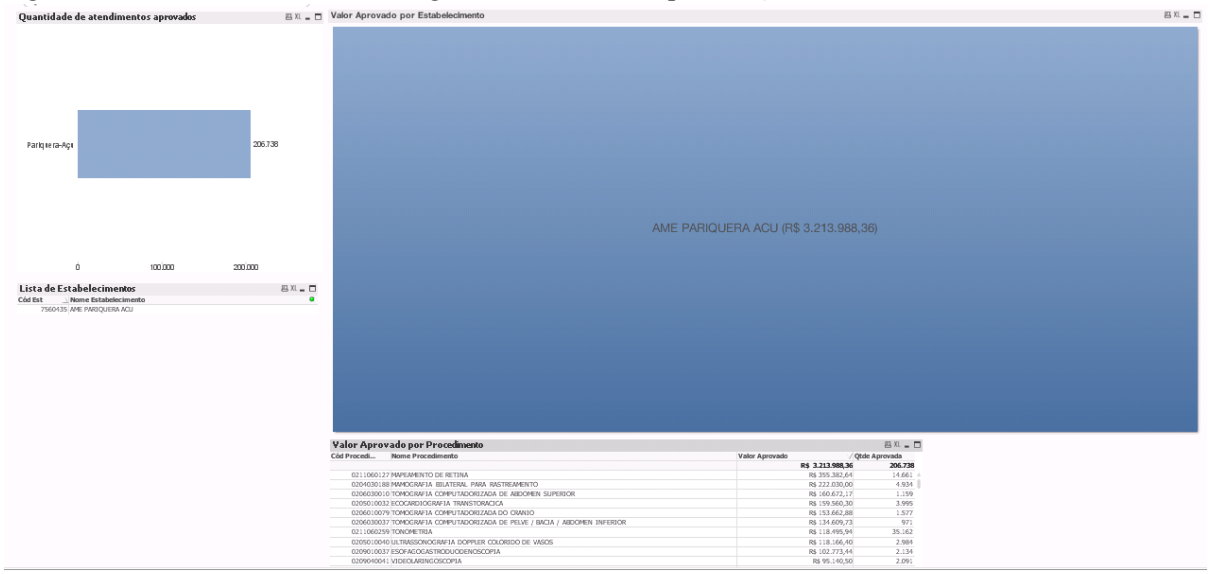
Figura 47. Detalhamento do estabelecimento AME Pariquera-Açu



Fonte: Elaborada pelo Autor (2020)

Ao selecionar a aba dos dados originais (Figura 48) para este estabelecimento, verifica-se que foram realizados um total de 206.738 procedimentos, gerando um valor aprovado de R$ 3.213.988,36. Já para o Para o Hospital Dr. Leopoldo Bevilacqua foram 196.456 procedimentos no total e R$ 1.494.764,95

Figura 48. Detalhamento dos dados gerais da AME Pariquera-Açu



Fonte: Elaborada pelo Autor (2020)

Ao realizar uma pesquisa rápida na Internet sobre essas duas unidades, foram levantadas informações de que são unidades de referência para a região do Vale do Ribeira. Esta região abrange 15 municípios: Barra do Turvo, Cajati, Cananéia, Eldorado, Iguape, Ilha Comprida, Iporanga, Itariri, Jacupiranga, Juquiá, Miracatu, Pariquera-Açu, Pedro de Toledo, Registro e Sete Barras. Segundo dados do IBGE, as populações desses municípios somadas dá 296.159 habitantes, mas o potencial dela pode chegar até a 500.000 habitantes, pois ainda pode atingir os municípios incluídos parcialmente pela região de abrangência do vale e também municípios do estado do Paraná. Esta pode ser uma explicação para o alto número de atendimentos.

Segundo ISG (2020), o Ambulatório Médico de Especialidades (AME) de Pariquera-Açu (Figura 49) é um ambulatório público, com atendimento gratuito e 100% SUS e está sob gestão do Instituto Sócrates Guanaes desde janeiro de 2020. Ele presta atendimento diagnóstico de alta resolubilidade em diagnóstico e orientação terapêutica para diferentes especialidades médicas à população da região do Vale do Ribeira.

Figura 49. AME de Pariquera-Açu.

Os atendimentos de Primeira Consulta Médica realizados no AME Pariquera-Açu são, na sua totalidade, eletivos, agendados pela Secretaria de Estado da Saúde de São Paulo (SES/SP). Os pacientes são encaminhados, principalmente pela Rede de Atenção Básica dos diversos municípios da região do Vale do Ribeira.

O AME de Pariquera-Açu dispõe de 12 consultórios médicos e 9 consultórios não médicos, 1 sala de procedimentos, 14 salas para realização de exames, além de centro de diagnóstico de imagem, laboratório com dois boxes e centro cirúrgico composto por duas salas cirúrgicas, 1 sala de procedimento e recuperação pós-anestésica. Tem capacidade mensal para cerca de 5 mil consultas médicas, 340 consultas não médicas, 340 cirurgias ambulatoriais de médio e pequeno portes e 275 exames externos de apoio diagnóstico e terapêutico.

Consultando o CNES, verificou-se ainda que o AME oferece os equipamentos para diagnóstico de imagem constantes na Tabela 17.

Tabela 17. Listagem de equipamentos de imagem do AME Pariquera-Açu

EQUIPAMENTOS DE DIAGNOSTICO POR IMAGEM			
Equipamento	**Existente**	**Em Uso**	**SUS**
MAMOGRAFO COMPUTADORIZADO	1	1	SIM
RAIO X ATE 100 MA	1	1	SIM
TOMÓGRAFO COMPUTADORIZADO	1	1	SIM
ULTRASSOM CONVENCIONAL	1	1	SIM
ULTRASSOM ECOGRAFO	2	2	SIM

Considerando que há apenas um mamógrafo disponível e que houve 4.934 procedimentos deste tipo, isso daria uma média de 13 procedimentos por dia. Este número parece razoável, ainda mais quando é possível colocar nesta conta alguma ação promocional de saúde que possa ter ocorrido.

Figura 50. Hospital Dr. Leopoldo Bevilacqua (antigo Hospital Regional do Vale do Ribeira)

Fonte: Hospital (2019)

Segundo HOSPITAL (2019), o Hospital Dr. Leopoldo Bevilacqua (Figura 50) é importante referência em saúde no Vale do Ribeira e é uma unidade da Secretaria de Estado da Saúde de São Paulo, administrada pelo CONSAÚDE, sendo voltada ao atendimento dos usuários do Sistema Único de Saúde (SUS) e abrangendo os 15 municípios pertencentes ao DRS XII de Registro (Departamento Regional de Saúde).

Só no ano passado, o Hospital Regional de Pariquera-Açu realizou um total de 2.011 partos, perfazendo uma média mensal de 168. No movimento geral de 2018, foram registradas 9.596 internações no total de 2018. O Pronto-Socorro do HRLB/CONSAÚDE registrou um total de 55.490 atendimentos, com movimento médio mensal de 3.624 pacientes. Os atendimentos ambulatoriais somaram 47.385 durante o ano de 2018. Foram realizados no total, 208.655 exames de laboratório. O hospital registrou ainda um total de 4.709 cirurgias nas especialidades de cirurgia geral, ginecologia, mastologia, neurologia, obstetrícia, oncologia, ortopedia, urologia e vascular.

Concluindo a análise, os números superlativos encontrados para o município de Pariquera-Açu podem ser explicados em parte pelo atendimento de referência para a região do Vale do Ribeira, mas precisam ser avaliados de maneira mais detalhada.

No entanto, a utilidade da ferramenta foi demonstrada. As informações da análise de *outliers* disponíveis por meio do painel permitiram uma atuação rápida já partindo de um município que apresenta anomalias no atendimento ambulatorial. Isso, por si só já permite a simplificação do trabalho com o enfoque em casos que realmente mereçam a atenção do auditor que realiza a tarefa

9 CONCLUSÃO

A utilização de ferramentas de análise de dados para auxiliar a atividade de Controle Externo a tirar proveito da grande quantidade de dados disponível é de suma importância para a nossa sociedade. Cada vez mais, devido à crescente informatização na área pública, tem-se acesso a uma massa de dados que muitas vezes tem o seu potencial inexplorado por falta de pessoal ou ferramentas disponíveis capazes de analisá-la.

Assim, diante disto, o desafio proposto neste trabalho foi criar uma ferramenta capaz de auxiliar a atuação da SecexSaúde na análise das bases de dados de procedimentos ambulatoriais disponibilizadas pelo SUS, permitindo que o auditor possa partir para a avaliação de indícios mais claro de irregularidades. A proposta foi então entender a base disponibilizada pelo SUS, importá-la para uma base local, tratá-la de forma a eliminar inconsistências e problemas nos dados e, por fim, aplicar uma rotina de detecção de anomalias que pudesse indicar procedimentos cujos dados de atendimentos possam apresentar problemas como forma de auxiliar a Secretaria nos seus trabalhos.

Apesar de este ser o produto principal, pode-se afirmar que as meras atividades de entendimento dos dados e de internalização, per se, são de grande valor para a instituição, pois os dados do Sistema de Informações Ambulatoriais ainda não eram dominados por completo pela instituição e este trabalho permitiu explorá-los e definir a sua utilidade no processo de trabalho do TCU.

O trabalho realizado propôs uma forma de detecção de anomalias nas quantidades aprovadas para os procedimentos ambulatoriais baseada na combinação de cinco algoritmos diferentes de detecção de anomalias: Z-score, Z-score Modificado, IQR – *Interquartile Range*, *Isolation Fores*t e LOF – *Local Outlier Factor*. Ao invés de compararmos diretamente o campo com a quantidade de atendimentos aprovados (PA_QTDAPR), utilizou-se uma taxa calculada

com base na quantidade de procedimentos e na população de cada município. Além disso, a Lei de Benford foi aplicada para aqueles procedimentos cujos dados são cadastrados de forma consolidada.

Por conta da ordem de grandeza da base do SIA, a aplicação do método foi feita sobre um recorte de dados baseado no grupo 02 de procedimentos. Os *outliers* foram detectados com base na quantidade de procedimentos realizados nos municípios e nos estabelecimentos. Cada um dos 1.054 procedimentos do Grupo 02 foi analisado e os resultados foram apresentados em forma de gráficos e de dados para análise por meio de um painel. Foram encontrados 6.620 casos de procedimentos *outliers* para os municípios e 9.938 para os estabelecimentos.

No entanto, é relevante notar que o trabalho foi realizado na forma de prova de conceito do que pode vir a se tornar um produto real, finalizado e implantado.

Desta forma, pode-se afirmar que cada um dos objetivos específicos listados em 1.4 foi devidamente alcançado, a saber:

1. Internalização dos dados do SIA/SUS e demais bases de dados relacionadas (ex.: CNES e IBGE) naquilo em que foi necessário para o trabalho;

2. Documentar os dados internalizados para disseminar e facilitar o seu uso por parte dos auditores;

3. Realizar análise exploratória dos dados do SIA;

4. Realizar prova de conceito com técnicas de mineração de dados para avaliar a viabilidade da identificação de discrepâncias estatísticas nos registros de internações ambulatoriais do SUS;

5. Construir um painel por meio do qual seja possível ao auditor da área de saúde avaliar as anomalias identificadas.

9.1 TRABALHOS FUTUROS

Este trabalho de forma alguma teve a intenção de esgotar o estudo da aplicação dos dados ambulatoriais do SIA/SUS. Portanto, seguem algumas sugestões para trabalhos futuros que podem ser interessantes como continuidade.

- Aplicar o método de detecção de anomalias para os demais Grupos de procedimento ambulatoriais e também para a variável valor aprovado (PA_VALAPR).

- Utilizar dados da clusterização da rede ambulatorial ou dos municípios para que a análise seja realizada entre municípios e estabelecimentos semelhantes entre si. Os dados do SUS que poderiam auxiliar nesta tarefa (IDSUS) não são atualizados desde 2011.
- Aplicar o método de detecção com base no Município de origem do paciente de forma a avaliar se os municípios e estabelecimentos foram classificados como *outliers* por causa da demanda somada dos municípios do seu entorno. Isso ajudaria a detectar casos de hospitais de referência como o Hospital de Câncer de Barreto que recebe pessoas de todo o país.
- Tornar a prova de conceito entregue em um produto final desenvolvido e implantado para uso pela SecexSaúde.
- Criar um método de detecção de anomalias mais robusto e tolerante a distorções nos dados.
- Utilizar os dados específicos dos procedimentos de alta complexidade para novas aplicações.
- Tentar obter os dados identificados por paciente para verificar se os procedimentos realizados por eles fazem sentido dentro do contexto do tratamento que está sendo aplicado ao paciente.

REFERÊNCIAS

BENFORD_PY, **Benford_py**. Disponível em: < https://github.com/milcent/benford_py>. Acesso em 12 dez 2019.

BRASIL, Ministério da Saúde. **Sistema Único de Saúde (SUS): estrutura, princípios e como funciona**. Disponível em: http:// https://www.saude.gov.br/sistema-unico-de-saude. Acesso em 11 jan 2020.

BRASIL, Tribunal de Contas da União. **Portaria-SECEXSAÚDE nº 3, de 10 de junho de 2019**. Disponível em <http://www.in.gov.br/web/dou/-/portaria-n-20-de-3-de-julho-de-2019-187435306>. Acesso em: 12 jan 2020.

BRASIL. Ministério da Saúde/Secretaria de Atenção à Saúde/ Departamento de Regulação, Avaliação e Controle/Coordenação Geral de Sistemas de Informação. **MANUAL TÉCNICO OPERACIONAL SIA/SUS - SISTEMA DE INFORMAÇÕES AMBULATORIAIS - Aplicativos de captação da produção ambulatorial APAC Magnético – BPA Magnético - VERSIA – DE-PARA – FPO Magnético,** 2010. Disponível em: < http://www1.saude.rs.gov.br/dados/1273242960988Manual_Operacional_SIA2010.pdf >. Acesso em: 12 jul 2019

BRASIL. Ministério da Saúde/Secretaria de Atenção à Saúde/ Departamento de Regulação, Avaliação e Controle/Coordenação Geral de Sistemas de Informação. **MANUAL TÉCNICO OPERACIONAL DO SISTEMA DE INFORMAÇÃO HOSPITALAR – ORIENTAÇÕES TÉCNICAS,** 2012a. Disponível em: < http://bvsms.saude.gov.br/bvs/publicacoes/manual_tecnico_sistema_informacao_hospitalar_sus.pdf>. Acesso em: 12 jul 2019

BRASIL. Ministério da Saúde/Secretaria de Atenção à Saúde/ Departamento de Regulação, Avaliação e Controle/Coordenação Geral de Sistemas de Informação. **SIA – Sistema de Informação Ambulatorial do SUS: Manual de Operação do Sistema,** 2012b. Disponível em: < http://www1.saude.rs.gov.br/dados/1273242960988Manual_Operacional_SIA2010.pdf >. Acesso em: 12 jul 2019

BRASIL. Ministério da Saúde/Secretaria Executiva/DATASUS/CGDIS. **DISSEMINAÇÃO DE INFORMAÇÕES EM SAÚDE SISTEMA DE INFORMAÇÕES AMBULATORIAIS DO SUS - SIASUS**. Disponível em: < ftp://ftp.datasus.gov.br/dissemin/publicos/SIASUS/200801_/Doc/Informe_Tecnico_SIASUS_2019_07.pdf>. Acesso em: 10 jul 2019

BREUNIG, M. M et al.. **LOF: Identifying Density-based Local *Outliers***. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD. (pp. 93–104), 2000.

CARVALHO, Osvaldo et al. **InfoSAS: um sistema de mineração de dados para controle da produção do SUS**. Revista do TCU, nº137, Setembro/Dezembro, 2016, p. 52-59. Disponível em: <https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/1378>. Acesso em: 14 mar 2019.

CHAPMAN, Pete et al. **CRISP-DM 1.0: Step-by-step data mining guide**. Tecnical report. The CRISP-DM consortium, 2000. Disponível em: <https://pdfs.semanticscholar.org/5406/1a4aa0cb241a726f54d0569efae1c13aab3a.pdf>. Acesso em: 29 jan 2020.

FEWSTER, R.M. **A Simple Explanation of Benford's Law**, The American Statistician. Disponível em: < https://www.stat.auckland.ac.nz/~fewster/RFewster_Benford.pdf>. Acesso em: 20 out 2019.

HAN, Jiawei et al. **Data Mining: Concepts and Techniques**. Elsevier, 2012. Disponível em: < https://learning.oreilly.com/library/view/data-mining-concepts/9780123814791/ >. Acesso em: 20 fev 2020.

HOSPITAL Regional completa 69 anos fazendo a diferença no Vale do Ribeira. **Registro Diário**, 2019. Disponível em: < http://www.registrodiario.com/noticia/4991/hospital-regional-completa-69-anos-fazendo-a-diferenca-no-vale-do-ribeira.html>. Acesso em: 20 de fev de 2020.

IBGE. Instituto Brasileiro de Geografia e Estatística. Disponível em <https://www.ibge.gov.br/apps/populacao/projecao/box_popclock.php>, 2020. Acesso em 02 abr 2020.

IBM. **IBM SPSS Modeler CRISP-DM Guide**. Disponível em: < https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm>, 2019b. Acesso em 29 jan 2020

IBM. **IBM SPSS Modeler CRISP-DM Guide**. Disponível em: <https://www.ibm.com/support/knowledgecenter/SS3RA7_18.2.1/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html>, 2019a. Acesso em 29 jan 2020.

IGLEWICZ, Boris; HOAGLIN, David C. **How to detect and handle *outliers***. ASQC Quality Press, 1993.

ISG. AME Pariquera-Açu, c2020. **Quem Somos**. Disponível em: < http://www.isgsaude.org/novo/amepariquera-acu/quem-somos-ame.php>. Acesso em: 04 mar 2020.

LIMA, Cecília Pessanha. **Comparando a saúde no Brasil com os países da OCDE: explorando dados de saúde pública,** 2016. Disponível em: < http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/16503/CeciliaMestrado.pdf?sequence=1&isAllowed=y> Acesso em: 25/06/2019.

LIU, Fei Tony; TING, Kai Ming e ZHOU, Zhi-Hua. **Isolation Forest**. Proceedings of the 8th IEEE International Conference on Data Mining (pp. 413-422). IEEE. Pisa, Italy, 2008.

MOLIN, Stefanie. **Hands-On Data Analysis with Pandas.** . Packt Publishing, 2019. Disponível em: < https://learning.oreilly.com/library/view/hands-on-data-analysis/9781789615326/>. Acesso em: 20 fev 2020.

PEDREGOSA, Fabian et all. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, 2011. volume 12, p. 2825–2830.

PETRUZALEK, Daniela. **READ.DBC - UM PACOTE PARA IMPORTAÇÃO DE DADOS DO DATASUS NA LINGUAGEM R**. XV Congresso Brasileiro de Informática em Saúde (CBIS). Goiânia, GO, Brasil, 2016.

PROJECT JUPYTER. **Project Jupyter®**. Disponível em: <https://jupyter.org>. Acesso em: 30 ago 2019.

PYSUS, PySUS. Disponível em: < https://pypi.org/project/PySUS/>. Acesso em 20 jun 2019.

PYTHON, Software Foundation. **Python**. Disponível em: <https://www.python.org/>. Acesso em: 30 ago 2019.

QUEIROZ, Christina. **Pesquisa FAPESP: Engrenagem Complexa. Alimentados por arrecadação tributária, regimes de financiamento à educação como o Fundeb, que expira em 2020, constituem desafio ao governo federal**. Disponível em: https://revistapesquisa.fapesp.br/2019/03/12/engrenagem-complexa>. Acesso em 02 fev 2020.

SEABORN. **SEABORN statistical data visualization**. Disponível em <https://seaborn.pydata.org/index.html>. Acesso em: 30 ago 2019.

SEN, Soumya. **Intercluster and Intracluster Distance**. Disponível em: <https://www.geeksforgeeks.org/ml-intercluster-and-intracluster-distance>. Acesso em: 20 dez 2019.

SOUZA, Renilson Rehem de. **O sistema público de saúde brasileiro**. Seminário Internacional: Tendências e desafios dos sistemas de saúde das Américas, p. 36. Disponível em: < http://portalarquivos.saude.gov.br/images/pdf/2016/janeiro/15/PGASS-Programa----o-Geral-das-A----es-e-Servi--os-de-Sa--de.pdf>. Acesso em: 19 jan 2019.

SRIVASTAVA, Shobhit. **Feature Scaling in Scikit-learn**. Disponível em: <https://medium.com/analytics-vidhya/feature-scaling-in-scikit-learn-b11209d949e7> Acesso em: 20 dez 2019.

THE PANDAS PROJECT. **Pandas - Python Data Analysis Library**. Disponível em: <https://pandas.pydata.org>. Acesso em: 30 ago 2019.

VERBUS, James. **Detecting and Preventing Abuse on LinkedIn Using Isolation Forests**, 13 ago 2019. Disponível em: < https://engineering.linkedin.com/blog/2019/isolation-forest > Acesso em: 23 jan 2020.

Z-SCORE: DEFINITION, FORMULA AND CALCULATION, Statistics How To. Statistics for the rest of us. Disponível em: <https://www.statisticshowto.com/probability-and-statistics/z-score/>. Acesso em: 19 out 2019.