PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA
GRADUAÇÃO EM CIÊNCIA DE COMPUTAÇÃO

UTILIZAÇÃO DE SOFTWARE LIVRE PARA ANÁLISE EXPLORATÓRIA DE DADOS EM SAÚDE - ESTUDO DE CASOS SOBRE COVID-19 NO MUNICÍPIO DE APARECIDA DE GOIÂNIA

IONÁ SANTANA

GOIÂNIA
2022

IONÁ SANTANA

UTILIZAÇÃO DE SOFTWARE LIVRE PARA ANÁLISE EXPLORATÓRIA DE DADOS EM SAÚDE - ESTUDO DE CASOS SOBRE COVID-19 NO MUNICÍPIO DE APARECIDA DE GOIÂNIA.

Trabalho de Conclusão de Curso apresentado à Escola Politécnica, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Talles Marcelo Gonçalves de Andrade Barbosa

GOIÂNIA

2022

IONÁ SANTANA

UTILIZAÇÃO DE SOFTWARE LIVRE PARA ANÁLISE EXPLORATÓRIA DE DADOS EM SAÚDE - ESTUDO DE CASOS SOBRE COVID-19 NO MUNICÍPIO DE APARECIDA DE GOIÂNIA.

Trabalho de Conclusão de Curso julgado adequado para obtenção do título de Bacharel em Ciência da Computação, e aprovado em sua forma final pela Escola Politécnica, da Pontifícia Universidade Católica de Goiás.

_____

Profa. Ma. Ludmilla Reis Pinheiro dos Santos
Coordenadora de Trabalho de Conclusão de Curso

Banca examinadora:

_____

Orientador: Prof. Dr. Talles Marcelo Gonçalves de Andrade Barbosa

_____

Prof. José Olímpio Ferreira

_____

Prof. Drª Kátia Kelvis Cassiano

_____

Prof. Olegario Correa da Silva Neto

GOIÂNIA
2022

DEDICATÓRIA

Dedico meu trabalho a minha mãe, por me apoiar e me deixar livre para seguir os meus sonhos.

AGRADECIMENTOS

Agradeço a todos os professores que contribuíram para a minha formação. Mas quero pontuar três professores: meu orientador Talles que acreditou em mim e me entendeu como sou, a coordenadora Carmen que me deu aula extra só para conseguir entender a matéria e por último o professor Alexandre, que me mostrou que todo problema tem começo, meio e fim. E por fim, gostaria de agradecer todas as pessoas que passaram na minha vida e me mostraram como pensar diferente é libertador.

EPÍGRAFE

"Já que sou, o jeito é ser"

Clarice Lispector

RESUMO

Este trabalho propõe uma metodologia para análise exploratória de dados para saúde, para ajudar pessoas de diversas áreas a analisar dados de saúde. Com intuito de democratizar a programação, usando *software* livre e a linguagem de programação Python, para decisões assertivas usando os dados disponíveis, para possibilitar melhorar diagnósticos, descobrir padrões em doenças e especializar o atendimento para cada pessoa. A análise está disponibilizada neste link: https://github.com/IonaSantana/UTILIZACAO-DE-SOFTWARE-LIVRE-PARA-ANALISE-EXPLORATORIA-DE-DADOS-EM-SAUDE---COVID-19

Palavras chaves: Análise Exploratória de Dados em Saúde, Python, Software Livre.

ABSTRACT

This work proposes a methodology for data exploratory analysis of health, to help people in a lot of areas analyze health data. To democratize the programming, it was used free software and Python language programming, for the right choices using available data, to enable better diagnostics, to find patterns in sickness, and to specialized attention to each person. The analysis is in the repository on this link: https://github.com/IonaSantana/UTILIZACAO-DE-SOFTWARE-LIVRE-PARA-ANALISE-EXPLORATORIA-DE-DADOS-EM-SAUDE---COVID-19

Keywords: Data Exploratory Analysis to Health, Python, Free Software.

LISTA DE ILUSTRAÇÕES

Figura 1 - Procedimento Metodológico Base de Dados COVID-19 da Itália 17
Figura 2 - Procedimento Metodológico Base de Dados incidentes de trânsito ocorridos em vias monitoradas pela CET-Rio 19
Figura 3 - Procedimento Metodológico Base de Dados Metabric Breast Cancer..... 20
Figura 4 - Proposta metodologia para Análise Exploratória de Dados..... 23
Figura 5 - Parte dos Dados 24
Figura 6 - Quantidade de Nulos por Variável 26
Figura 7 - Continuação quantidade de Nulos por Variável 27
Figura 8 - Correlação de Pearson 28
Figura 9 - Correlação de Phik (φk) 29

Gráfico 1 - Número de notificações do ano de 2020 dos meses de julho, agosto e setembro 32
Gráfico 2 - Proporção de profissionais da saúde na base 33
Gráfico 3 - Proporção de pessoas com comorbidades 34
Gráfico 4 - Quantidade por comorbidade 34
Gráfico 5 - Dias na UTI para pessoas com e sem comorbidades/risco 35
Gráfico 6 - Dias na UTI para pessoas com e sem comorbidades/risco, pelo menos 1 dia 36
Gráfico 7 - Quantidade por comorbidade com doenças cardíacas crônicas 38
Gráfico 8 - Quantidade por comorbidade com diabetes 39
Gráfico 9 - Quantidade por comorbidade com doenças respiratórias crônicas 40
Gráfico 10 - Proporção feminino e masculino 41
Gráfico 11 - Histograma geral coluna Outra Faixa etária 42
Gráfico 12 - Histograma feminino com a coluna idadeM 43
Gráfico 13 - Histograma feminino com a coluna Outra Faixa etária 43
Gráfico 14 - Boxplot idade mulheres 43
Gráfico 15 - Q-Q idade mulheres 45
Gráfico 16 - Histograma idade homens com a coluna idadeM 45
Gráfico 17 - Histograma idade homens com a coluna Outra Faixa etária 46

Gráfico 18 - Boxplot idade homens 46
Gráfico 19 - Q-Q homens 48
Gráfico 20 - Cura/Recuperado em relação ao sexo feminino e masculino 48
Gráfico 21 - Óbitos por COVID-19 relação ao sexo feminino e masculino..... 49
Gráfico 22 - Internado - Enfermaria relação ao sexo feminino e masculino..... 49
Gráfico 23 - Isolamento Domiciliar relação ao sexo feminino e masculino 50
Gráfico 24 - Internado - UTI relação ao sexo feminino e masculino 50
Gráfico 25 - Sintomas 51
Gráfico 26 - Histograma quantidade de sintomas..... 52
Gráfico 27 - Boxplot quantidade de sintomas 52
Gráfico 28 - Quantidade de sintomas em relação a Isolamento domiciliar 53
Gráfico 29 - Quantidade de sintomas em relação a Óbitos por Covid-19 54
Gráfico 30 - Quantidade de sintomas em relação a Cura/Recuperado..... 54
Gráfico 31 - Quantidade de sintomas em relação a internação enfermaria e UTI..... 55
Gráfico 32 - Vacinas tomadas 56

LISTA DE TABELAS

Tabela 1 - Trabalhos Relacionados 21
Tabela 2 - Estratificação Base Casos Covid-19 de Aparecida de Goiânia 30
Tabela 3 - Métricas pessoas com comorbidades/risco 36
Tabela 4 - Métricas pessoas sem comorbidades/risco 37
Tabela 5 - Evolução da contaminação de COVID-19 37
Tabela 6 - Métricas idade mulheres 44
Tabela 7 - Métricas idade homens 47
Tabela 8 - Métricas quantidade de sintomas 53

LISTA DE ABREVIATURAS E SIGLAS

IBM = *International Business Machines*

AED = Análise Exploratória de Dados

UTI = Unidade de Terapia Intensivo

SUMÁRIO

1. INTRODUÇÃO 14
2 REFERENCIAL TEÓRICO 17
3 MATERIAIS E MÉTODOS 23
3.1 Coleta de dados 24
3.2 Preparação da base 24
3.2.1 Retirada de dados sensíveis 25
3.2.2 Avaliação dos nulos 25
3.3.4 Correlação de variáveis 27
3.3.4 Estratificação da base 29
3.3 Análise e refinamento dos dados 31
3.4 Tecnologias utilizadas 31
4 RESULTADOS 32
4.1 Diária 32
4.2 Profissional da saúde 33
4.3 Comorbidade/risco 33
4.3.1 Geral 34
4.3.2 Dias na UTI 35
4.3.3 - Comorbidades com maior frequência 37
4.3.3.1 - Doenças cardíacas crônicas 38
4.3.3.2 - Diabetes 39
4.3.3.3 - Doenças respiratórias crônicas 40
4.4 Sexo 41
4.4.1 Geral 41
4.4.2 Idade 41
4.4.2.1 - Idade Feminino 42
4.4.2.1 - Idade Masculino 45
4.4.3 Evolução do paciente 48
4.5 Sintomas 51
4.5.1 Geral 51
4.5.2 Evolução do paciente 53
4.6 Vacinados 55
5 CONCLUSÕES 57
REFERÊNCIAS 58
APÊNDICES 60

Apêndice A – Códigos 60
Apêndice B – Manual de uso do Google Colab 65

1. INTRODUÇÃO

A partir do estudo levantado pela *International Business Machines* (IBM) (2020), os dados na área da saúde passaram de 500 petabytes em 2012 para 25.000 petabytes em 2020 (HIMSS Media, 2017). Esses dados além de serem coletados, gerenciados, armazenados, podem ser usados para tomadas de decisões, como: a precisão para decisões de tratamento de câncer, pois até 44% dos tratamentos iniciais são modificados no segundo curso de tratamento e a capacidade de acompanhar evidências, pois menos de 50% da medicina é baseada em evidências, o que levaria um epidemiologista gastar 167 horas em leituras semanais, para poder acompanhar novas percepções profissionais (HIMSS Media, 2017).

Esses dados são separados em estruturados e não estruturados. Dados estruturados são informações que podem ser dispostas em linhas e colunas, como, planilhas do Excel. E o termo dados não estruturados é usado quando não há organização (ou estrutura) integrada aos dados, por exemplo, coleção de arquivos de áudio ou postagens de mídia social (IBM, 2022).

Apesar da diferença, é possível analisar os dois, usando a Análise Exploratória de Dados (AED), sendo um campo usado para representar visualmente o conhecimento do conjunto de dados fornecidos. A técnica é utilizada para gerar inferências a partir de um determinado conjunto de dados (DSOUZA; VELAN, 2020, p.1). Podendo ajudar na identificação de erros óbvios, entender padrões presentes nos dados, detectar desvios ou eventos anômalos, encontrar relações entre as variáveis e testar hipóteses, usando medidas estatísticas (IBM CLOUD EDUCATION, 2020), (SURESH; AHMED, 2020).

As variáveis de um conjunto de dados, podem ser separadas, em:

Quantitativas, sendo as variáveis numéricas ou em qualitativas, sendo as variáveis não numéricas, também chamadas de categóricas (MAYER, 2016).

As variáveis numéricas são separadas em: discretas e contínuas. As discretas assumem valores inteiros (MAYER, 2016). Por exemplo, números de casos em um determinado dia ou quantidade de sintomas de um paciente. E, as contínuas,

assumem valores no intervalo dos números reais, como, peso ou altura de um paciente.

As variáveis categóricas são divididas em: nominais e ordinais. As nominais, não possuem ordem específica, por exemplos, nomes e sexo. E, as ordinais, podem ser ordenadas, como, idade (criança, adolescente, adulto) e grau de instrução (básico, médio, graduação) (MAYER, 2016).

Tanto para as numéricas, quanto para as categóricas é possível aplicar técnicas para visualização de dados. Sendo uma boa estratégia, pois "a compreensão humana é 60.000 vezes mais sensível aos dados visuais do que os dados em texto" (DSOUZA; VELAN, 2020, p.1).

Para a área de saúde, a AED permite ajudar a levantar inferências, como, no trabalho do Saini, et al (2020)**.** Neste trabalho, foi utilizado uma AED juntamente com uma técnica de agrupamento para agrupar os países com base no número de casos confirmados e de óbitos por Covid-19, obtendo 3 grupos: países que são moderadamente afetados, países severamente afetados e países com grande número de casos confirmados, mas com menor número de óbitos.

Outra possível aplicação é avaliar, constantemente, os pacientes por meio da telemedicina, onde os atendimentos são ajudados a distância, por exemplo, atendimentos feitos pela internet (BRITO; LEITÃO*,* 2020), sendo possível determinar precocemente alguma complicação ou até mesmo mortes, levando em consideração o tempo de assistência ser maior. (HAU et al., 2020).

Existem várias ferramentas para AED, Kapko (2019), mostra algumas gratuitas, como, DataMelt, Orange e KNIME. São ferramentas, com interface para AED que possuem a facilidade de trazer algumas funcionalidades e gráficos prontos. Porém algumas funcionalidades são limitadas impossibilitando de aprofundar algumas análises.

Por isso, para este trabalho foi utilizada a linguagem Python, pois a comunidade que o utiliza é composta por 8 milhões de usuários (BIELAK, 2021), tendo vários conteúdos, tornando mais fácil democratizar a programação e AED para pessoas de diferentes áreas.

A ferramenta escolhida é de *software* livre, pois possibilita manter a liberdade de quem usa, compartilhar, estudar e modificar o código fonte (FREE SOFTWARE FOUNDATION, 2019). Para o trabalho foi utilizado o JupyterLab juntamente com a interface do Jupyter Notebook, sendo um aplicativo baseado na Web que permite a criação de documentos permitindo código, texto, equações e visualizações, aceitando as linguagens de programação Python, R e Julia (JUPYTER, 2015 e 2019).

Este trabalho apresenta o procedimento de uma análise sobre o período de 18/02/2020 até 28/10/2021 sobre a base de casos de Covid-19 de Aparecida de Goiânia. Mostrando técnicas para identificar padrões, verificar hipóteses e visualizar tendências nos dados. Este trabalho tem a autorização do uso dos dados adquiridos na dissertação de mestrado "Estudo descritivo sobre o serviço de telemedicina no acompanhamento de pacientes diagnosticados com Covid-19 no município de Aparecida de Goiânia" (REZENDE e BARBOSA, 2022).

Os artigos em AED, são focados na análise dos dados e por muitas vezes os processos são descritos, mas os códigos não são compartilhados. Por isso, é importante ressaltar a necessidade de manuais e documentos técnicos que facilitem a utilização de ferramentas de *software* livre para AED, bem como, a definição de processos e procedimentos para uso adequado.

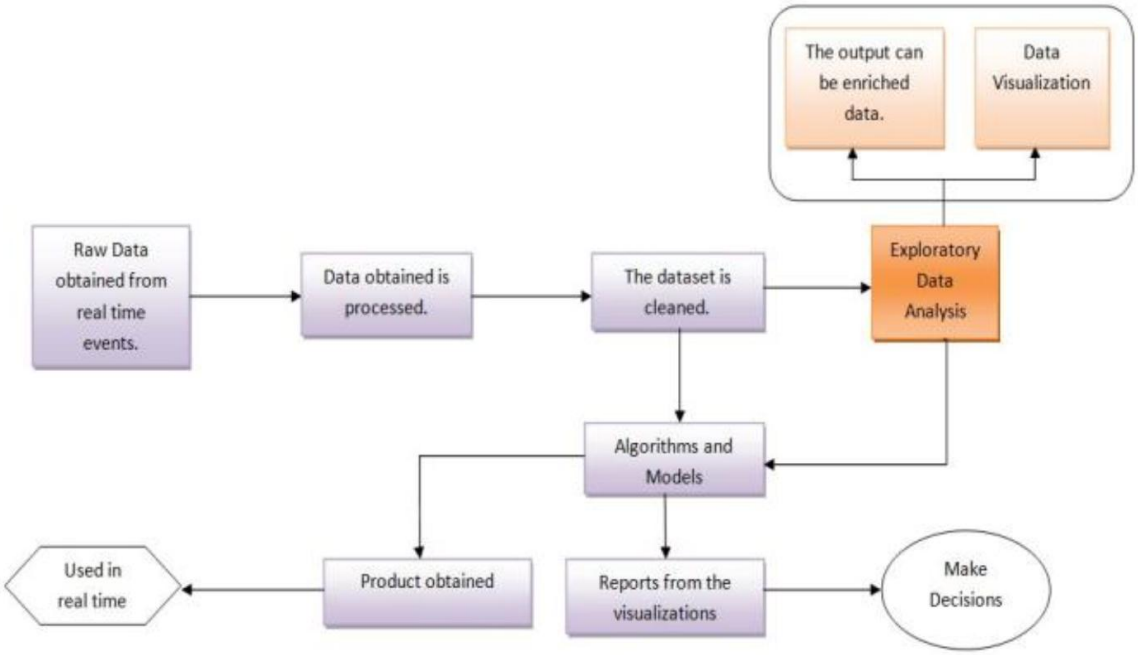
A organização restante deste trabalho está organizado do capítulo 2 ao 5, sendo: o capítulo 2, o referencial teórico com trabalho relacionados mostrando procedimentos metodológicos de várias análises exploratórias; capítulo 3, materiais e métodos, apresenta a metodologia proposta para AED; capítulo 4, a discussão dos resultados obtidos sobre o estudo de caso desse trabalho; capítulo 5, conclusão e trabalhos futuros;

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados os trabalhos relacionados, mostrando o procedimento metodológico feito por cada trabalho.

Dsouza e Velan (2020), realizaram a análise exploratória sobre a base de dados de COVID-19 da Itália em 2020. E seguiram o procedimento metodológico abaixo na Figura 1:

Figura 1 - Procedimento Metodológico Base de Dados COVID-19 da Itália



Fonte: Dsouza e Velan, 2020

Primeiro, os dados foram coletados, depois processados e então passaram pela limpeza. Após essa etapa, é feita a AED, para visualizar os dados e obter padrões, dessa forma, aplicar algoritmos, reportar a análise e obter um resultado, para fim de tomar decisões. Para AED, Dsouza e Velan (2020), usaram uma matriz de correlação, para entender a correlação entre as variáveis e gerar novas hipóteses.

Usou a linguagem Python e algumas de suas bibliotecas. Para mapear a base, a biblioteca Pandas, trazendo os dados em formatos de tabulares, e para os gráficos, as bibliotecas Matplotlib e Seaborn (DSOUZA e VELAN, 2020).

O trabalho Saini, et al. (2020), explora o conjunto de dados aberto de 2019-nCoV da Universidade Johns Hopkins. Essa base, tem atualizações diárias do total de casos e mortes do mundo todo. A AED baseou-se em números de casos confirmados, recuperados e óbitos e também realizou uma análise comparativa da taxa de mortalidade e recuperação para quase 222 nações do mundo.

Por fim, os países foram agrupados usando o algoritmo de agrupamento *K-means*, eficaz para conjuntos de dados grandes e possui como objetivo agrupar dados semelhantes e não estruturados (Taulli, 2019). Usou-se a premissa dos números de casos confirmados e de óbitos, com intuitos de avaliar o aumentos dos riscos em uma determinada área, usando análise visual, para comparar a contagem dos casos e para ajudar nas estratégias de controlar a disseminação global (SAINI, et al; 2020).

Como resultado, foram gerados 3 grupos: 1 - altos casos, mas baixo número de mortes; 2- grande número de casos confirmados e de óbitos; 3 - grande número de casos confirmados e de óbitos, mas inferior ao do grupo 2 (SAINI, et al; 2020).

Outro artigo teve a análise exploratória sobre a vacinação em diferentes países. Chen, et al. (2021), analisaram tendências globais, quantidade de dados e tipos de vacinação, e uma comparação mais aprofundada entre a China e a Índia.

O procedimento metodológico foi separado em: 1 - Análise emergencial; 2 - Divisão em duas fases: fase de exploração e fase de verificação. Na fase de exploração o foco foi descobrir padrões e na fase de validação, focaram em verificar se o novo modelo da etapa de exploração de dados estava correto (CHEN, ET AL; 2021).

Adquiriu a base de dados pelo Kaggle, uma plataforma popular para competições de ciência de dados. Com 11175 registros, em 14 de abril de 2021. Tanto para leitura do arquivo e pré-processamento dos dados usou-se a linguagem de programação Python (CHEN, ET AL; 2021)..

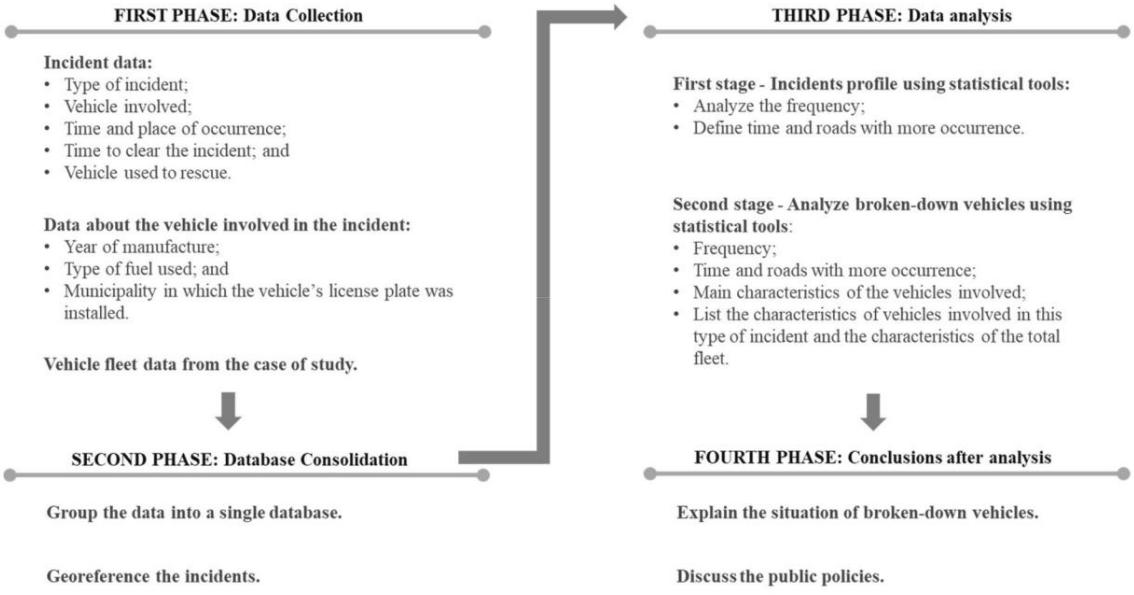
Para limpeza dos dados, Chen, et al. (2021), fez com intuito de identificar erros, padronizar operações e promover a qualidade dos dados, excluindo dados

duplicados, processando valores inválidos e ausentes. Após, a limpeza dos dados foi feita a estratificação dos dados, para responder as inferências levantadas.

Em outra AED os autores Baltar, Ribeiro e Santos (2022), exploraram o conjunto de dados históricos sobre incidentes de trânsito ocorridos em vias monitoradas pela CET-Rio, ocorridos entre 2015 a 2017. Com informações, como, tipo de incidente, veículo envolvido, hora e local do incidente e veículo usado para ajudar o motorista.

O procedimento metodológico utilizado na análise, foi dividido em quatro fases e fluxograma pode ser visto na Figura 2: primeira fase - Coleta dos dados; segunda fase - Base de dados consolidada; terceira fase - Análise dos dados e quarta fase - Conclusões sobre a análise (BALTAR; RIBEIRO; SANTOS, 2022).

Figura 2 - Procedimento Metodológico Base de Dados incidentes de trânsito ocorridos em vias monitoradas pela CET-Rio



Fonte: Baltar, Ribeiro e Santos, 2022

Na primeira fase, adquiriu-se a base de dados, com informações separadas, em: incidentes e data em que ocorreu o acidente com o veículo. Na segunda fase, foi a consolidação do banco de dados, agrupando as informação em um único lugar e guardando a referência local de cada um, também incluíram limpeza dos dados, como, retirada de nulos, retirada de registros com placas inconsistentes e placas de outros estados (BALTAR; RIBEIRO; SANTOS, 2022).

Na terceira fase, estratificou-se os dados em duas partes, perfil de incidentes e veículos quebrados e a análise sobre a estratificação foi feita usando ferramentas estatísticas. A última, explica sobre o resultados encontrados para veículos quebrados e propõe políticas (BALTAR; RIBEIRO; SANTOS, 2022).

O último trabalho relacionado, é dos autores Sweetlin, Saudia (2021), a AED usou o Conjunto de dados *Metabric Breast Cancer* para a visualização de Índice de Prognóstico de Nottingham (INP), a *Status* Sobrevivência Geral e *Status* Livre de Recaída para determinar a sobrevivência e recorrência da doença no câncer de mama e simplificou a visualização de vários aspectos necessários para determinar o período de 5 e 10 anos de sobrevida de pacientes com câncer de mama.

Sweetlin, Saudia (2021), o procedimento metodológico usado na AED foi separado em quatro etapas e pode ser vista na Figura 3:

Figura 3 - Procedimento Metodológico Base de Dados *Metabric Breast Cancer*

Fonte: Sweetlin, Saudia, 2021

Primeira é a aquisição dos dados e foi considerado apenas dados clínicos para análise, excluindo os dados genômicos. A segunda etapa é a análise exploratória, e em seguida é separada em duas etapas: visualização e refinamento dos dados, sendo um processo incremental entre as duas etapas.

Segue abaixo a tabela 1 com os procedimentos metodológicos de cada artigo.

Tabela 1 - Trabalhos Relacionados

Trabalhos	**Descrição**	**Metodologia**	**Tecnologias Usadas**
Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases	AED sobre a base de dados de COVID-19 da Itália em 2020	1 - Coleta dos dados; 2 - Correlação dos dados; 3 - Estratificação dos dados 4 - Análise do dados por meio de gráficos	Linguagem de programação Python. Bibliotecas Python: Pandas, Matplotlib e Seaborn
Visual Exploratory Data Analysis of COVID-19 Pandemic	AED sobre o conjunto de dados aberto de 2019-nCoV da Universidade Johns Hopkins.	1 - Coleta dos dados; 2 - Estratificação do dados; 3 - Análise do dados por meio de gráficos; 4 - Aplicação do algoritmo K-means.	Não menciona
Exploratory Data Analysis on the Usage of COVID-19 Vaccine	AED sobre a base de dados adquirida pelo Kaggle. Sobre a vacinação em diferentes países, como, tendências globais, quantidade de dados e tipos de vacinação.	1 - Coleta dos dados; 2 - Limpeza dos dados; 3 - Estratificação dos dados; 4 - Análise do dados por meio de gráficos	Linguagem de programação Python para leitura da base e pré-processamento. Não cita quais ferramentas foram utilizadas para os gráficos.
Exploratory analysis of incidents in an urban area	AED sobre o conjunto de dados históricos sobre incidentes de trânsito	1 - Coleta dos dados; 2 - Base	Não menciona

focused on broken-down vehicles: The case of Rio de Janeiro	ocorridos em vias monitoradas pela CET-Rio, ocorridos entre 2015 a 2017.	de dados consolidada; 3 - Análise dos dados e 4 - Conclusões depois da análise.	
Exploratory Data Analysis on Breast cancer dataset about Survivability and Recurrence	AED sobre a base de dados (Câncer de Mama, METABRIC, Nature 2012 e Comunicação Nat 2016)	1 - Coleta de dados; 2 - Análise de Dados Exploratória; 3 - Refinamento dos dados em conjunto com a visualização dos dados	Linguagem Python. Ferramenta de visualização do Python.

Fonte: autoria própria

3 MATERIAIS E MÉTODOS

Os trabalhos relacionados podem ser separados em quatro etapas:

A primeira etapa, a coleta de dados, sendo o carregamento da base para a análise, pode ser feita através de arquivos, consultas a banco de dados, leitura de arquivos separados por vírgula, entre outros.

A segunda etapa, chamada de preparação dos dados. Nos trabalhos relacionados às etapas antes da AED são, correlação dos dados, estratificação dos dados, base de dados consolidada, limpeza dos dados, como retirada de registros duplicados e/ou nulos. E, por isso, foram agrupadas em uma etapa. Porque, todas são feitas para preparar a base para análise, podendo ser chamada também de pré-processamento dos dados.

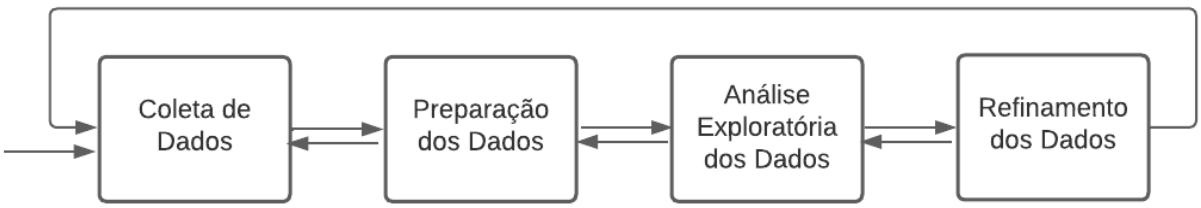
A terceira etapa, é a AED. Etapa onde são aplicadas as técnicas estatísticas e de visualização de dados, para encontrar padrões.

Também considerou-se uma quarta etapa, feita em conjunto com a AED, chamada de refinamento dos dados. Após a aplicação das técnicas de AED são feitas as conclusões e se necessário é feito o refinamento dos dados, senão a AED é concluída.

Todas as etapas podem ser voltadas, inclusive a coleta de dados, caso, por exemplo, algum dado foi identificado necessário e a base não o possui.

O diagrama na Figura 4, propõe uma metodologia de AED, de forma incremental, pois é possível encontrar padrões não vistos antes, sendo necessário outra coleta de dados e/ou os procedimentos seguintes.

Figura 4 - Proposta metodologia para Análise Exploratória de Dados



Fonte: autoria própria

Para ilustrar a metodologia, aplicou-se o procedimento na base de dados sobre casos de Covid-19 no Município de Goiânia. Os códigos podem ser visto no Apêndice A e o manual de uso do Google Colab no Apêndice B.

3.1 Coleta de dados

A primeira etapa é a coleta dos dados, para esse estudo de caso é apresentado as notificações de casos de Covid-19 (autorização do uso da base está no Anexo A), no período do dia 18 de fevereiro de 2020 a 28 de outubro de 2021. O seu tamanho é de 91.834 registros e possui 70 colunas, com informações, como, sintomas, idade e comorbidade. Parte dos dados, é mostrado na Figura 5:

Figura 5 - Parte dos Dados

Sexo	Unidade Notificadora	Data de Nascimento	IdadeM	Outra Faixa etária	Raça/Cor	Bairro	CCI	...	Data do Diagnóstico
Masculino	Hosp. Anis Rassi	1951-03-06	70	70 a 79 anos	Branca	Cidade Vera Cruz - Jardins Mônaco	46209	...	2020-03-18
Masculino	Hosp. Anis Rassi	1980-06-06	41	40 a 49 anos	Ignorado	Setor dos Afonsos	11958	...	2020-04-05
Feminino	Unimed GO	1977-12-30	43	40 a 49 anos	Ignorado	Vila Santo Antônio 2ºAcréscimo (Conj. Progresso)	10127	...	2020-04-09
Masculino	Hospital Amparo	1981-08-08	40	40 a 49 anos	Branca	Jardim Maria Inês	64806	...	2020-04-09
Masculino	HUAPA	1993-02-05	28	20 a 29 anos	Parda	Parque Veiga Jardim	117619	...	2020-04-09

Fonte: autoria própria

3.2 Preparação da base

A preparação da base consiste em avaliar os dados, como, valores em cada coluna, se estão consistentes com a realidade e, se não estiverem adaptá-los. Para isso, foi considerado as etapas: retirada de dados sensíveis, avaliação dos nulos, correlação entre as variáveis, técnicas de seleção de variáveis e estratificação da base.

3.2.1 Retirada de dados sensíveis

> "dado pessoal sensível: dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural;" (LGPD, 2018, cap 1, Art. 5º, item II)

Informações que identificam a pessoa, como, cpf, telefone, foram retiradas.

3.2.2 Avaliação dos nulos

Quantidade de nulos encontrados por cada variável, o valor está abaixo na Figura 6 e 7, essa avaliação é uma das maneiras de verificar quais variáveis possuem todas as informações e gerando maior consistência para análise.

Figura 6 - Quantidade de Nulos por Variável

	Coluna	Quantidade de nulo
0	Número do Registro	0
1	Notificações ESUS - SIVEP	12144
2	Sexo	0
3	Unidade Notificadora	6
4	Data de Nascimento	0
5	IdadeM	0
6	Outra Faixa etária	0
7	Raça/Cor	0
8	Bairro	0
9	CCI	73303
10	Data da Notificação:	0
11	É profissional de saúde ?	0
12	Profissão	0
13	2021-02-19 00:00:00	90120
14	Local 02	91721
15	Teve contato próximo com caso confirmado ou su...	12
16	Que contato?	85760
17	Data de início de sintomas	0
18	Febre	0
19	Tosse	0
20	Dispneia	0
21	Dor de Garganta	0
22	Fraqueza	16
23	Mialgia	16
24	Dor no Peito	16
25	Cefaleia	0
26	Assintomático	18
27	Coriza	0
28	Perda de Olfato	0
29	Perda de Paladar	0
30	Outros	84084
31	Tem comorbidades?	0
32	Doença Resp. Crônica	0
33	Doenças Renais Crônica em estágio Avançado	0
34	Gestante	0

Fonte: autoria própria

Figura 7 - Continuação quantidade de Nulos por Variável

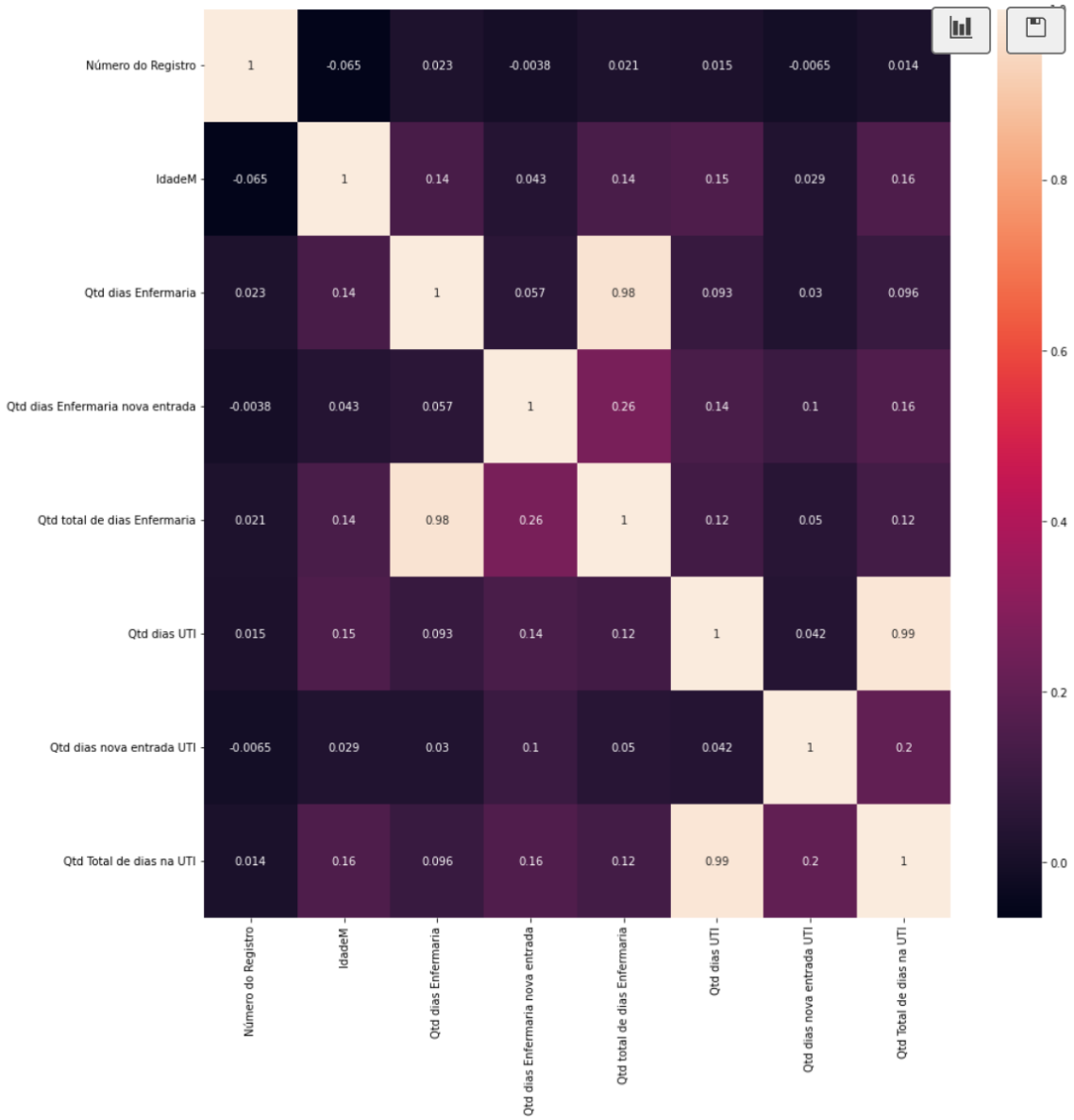
36	Doenças Cardíacas Crônicas	0
37	Imunossupressão	0
38	Diabetes	0
39	Outros?	89992
40	HOSPITALIZADO	10
41	Se Sim, Onde?	86343
42	Data de internação	86356
43	Enfermaria	56
44	Data entradaEnf	87718
45	Data SaídaEnf	87906
46	Qtd dias Enfermaria	0
47	Data nova entrada Enf	91643
48	Data nova saida Enf	91646
49	Qtd dias Enfermaria nova entrada	0
50	Qtd total de dias Enfermaria	0
51	UTI	13
52	Data entradaUti	89359
53	Data SaídaUti	89436
54	Qtd dias UTI	0
55	Data nova entrada UTI	91771
56	Data nova saida UTI	91770
57	Qtd dias nova entrada UTI	0
58	Qtd Total de dias na UTI	0
59	Evolução	0
60	Data do Diagnóstico	0
61	Data de Lançamento	0
62	Data da cura ou óbito	41421
63	Ja vacinou? Se sim, Qual Vacina?	90746
64	Data 1° Dose	90850
65	Data 2° Dose	91462
66	Fez sequenciamento? Se Sim, qual variante?	91833
67	Tipo de Teste / Método	4
68	Laboratório	5
69	Observação	54335

Fonte: autoria própria

3.3.4 Correlação de variáveis

“A correlação é uma análise bivariada que mede a força de associação entre duas variáveis e a direção da relação” (CENTRO DE ESTATÍSTICA APLICADA, 2022). Para variáveis quantitativas foi usada a correlação de Pearson, o coeficiente de correlação de Pearson (r) é uma medida de correlação linear entre duas variáveis. Seu valor está entre -1 e +1, -1 indicando correlação linear negativa total, 0 indicando nenhuma correlação linear e 1 indicando correlação linear positiva total (CENTRO DE ESTATÍSTICA APLICADA, 2022). A matriz resultante pode ser vista na Figura 8.

Figura 8 - Correlação de Pearson



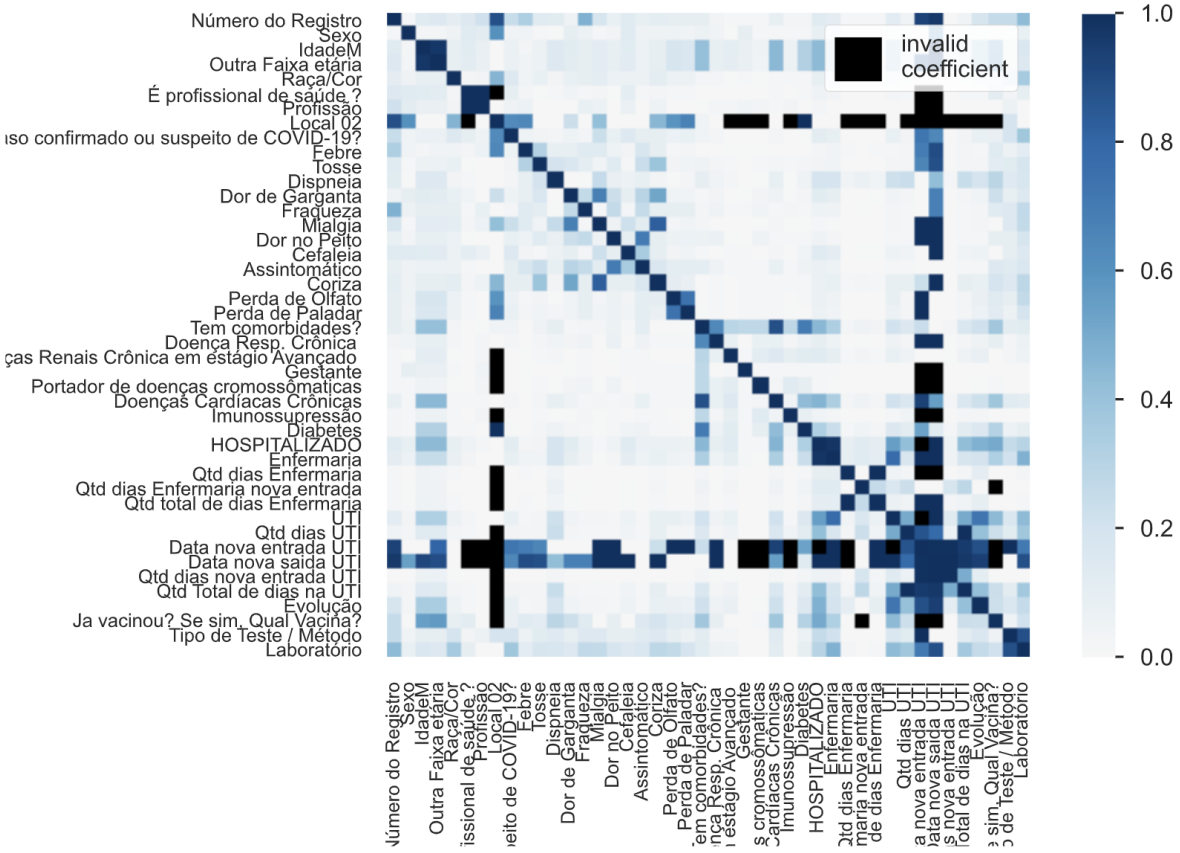
Fonte: autoria própria

A correlação alta nesta base está entre variáveis que possuem a mesma origem, por exemplo, 'Qtd dias UTI' e 'Qtd Total de dias na UTI', com o coeficiente de 0.99, os dois referentes aos dias que a pessoa ficou internada na UTI. Por isso, ao analisar a UTI, fica a critério qual variável usar na análise.

Para variáveis categóricas, ordinais e intervalares foi usada a correlação de de Phik (φk), sendo que o coeficiente de correlação captura dependência não linear e reverte para o coeficiente de correlação de Pearson no caso de uma distribuição

de entrada normal bivariada (CENTRO DE ESTATÍSTICA APLICADA, 2022). E quanto mais perto do valor 1, maior a correlação. A matriz resultante pode ser vista na Figura 9.

Figura 9 - Correlação de Phik (φk)



Fonte: autoria própria

As variáveis ‘Data nova entrada UTI’ e ‘Data nova saída UTI’, são correlacionadas com muitas outras, porém a nulidade delas chega a quase 100%. Alguns sintomas são correlacionados com outros sintomas, como, mialgia, coriza e dor de garganta.

3.3.4 Estratificação da base

Divisão da base em categorias menores considerando variáveis específicas para analisar a base de forma geral, estão dispostas na tabela 2. Foram criadas 7 categorias, comorbidades, diária, profissional da saúde, sexo, sintomas, vacinados,

evolução do paciente e internação, algumas foram usadas para apoio as inferências levantadas:

Tabela 2 - Estratificação Base Casos Covid-19 de Aparecida de Goiânia

Comorbidades/risco	'Tem comorbidades?', 'Doença Resp. Crônica ', 'Doenças Renais Crônica em estágio Avançado ', 'Gestante', 'Portador de doenças cromossômaticas', 'Doenças Cardíacas Crônicas', 'Imunossupressão', 'Diabetes', 'Outros?'
Diária	'Data da Notificação:'
Profissional da saúde	'É profissional de saúde ?'
Sexo	'Sexo', 'Outra Faixa etária', 'IdadeM'
Sintomas	'Febre', 'Tosse','Dispneia', 'Dor de Garganta', 'Fraqueza', 'Mialgia', 'Dor no Peito', 'Cefaleia', 'Assintomático', 'Coriza', 'Perda de Olfato', 'Perda de Paladar', 'Outros'
Vacinados	'Ja vacinou? Se sim, Qual Vacina?'
Evolução do Paciente	'Evolução'

Internação	'Qtd dias UTI'

Fonte: autoria própria

3.3 Análise e refinamento dos dados

Para variáveis numéricas: histograma e gráfico de barra para calcular frequência, gráfico de linha para distribuição entre dias.

Também foram utilizados a média aritmética simples, desvio padrão, quartis 25%, 50% e 75% e os valores mínimos e máximos.

Para variáveis categóricas: com valores "sim" e "não", usou-se o gráfico de pizza e para variáveis com mais valores o gráfico de barra.

Para verificar se a amostra segue uma distribuição normal, aplicou-se o gráfico Q-Q . Para tendência de centralidade, simetria e valores atípicos foram demonstradas através do gráfico *boxplot*.

O refinamento dos dados foi gerado, a partir da análise. Por exemplo, as três comorbidades com maiores porcentagens foram analisadas, em relação à evolução do paciente. E só foi possível detectar após a primeira análise.

3.4 Tecnologias utilizadas

Esta seção possui a descrição das tecnologias utilizadas para o desenvolvimento da análise exploratória.

Utilizou-se a linguagem Python, por sua comunidade ser grande e possuir várias bibliotecas. Para coleta dos dados, leitura da base, foi usada a biblioteca Pandas, que suporta dados tabulares, tendo uma interface familiar a planilha do Excel.

Para a preparação dos dados foi usado a biblioteca seaborn para correlação de pearson e para correlação de Phik (φk) e o mapa de calor da biblioteca pandas_profiling.

Para gráficos em geral utilizou-se MatplotLib, pois a mesma tem poder de criar gráficos estáticos, animados e interativos e para aplicação de técnicas estatísticas, a biblioteca seaborn. Para o teste de normalidade, aplicou-se o gráfico Q-Q da biblioteca statsmodels.

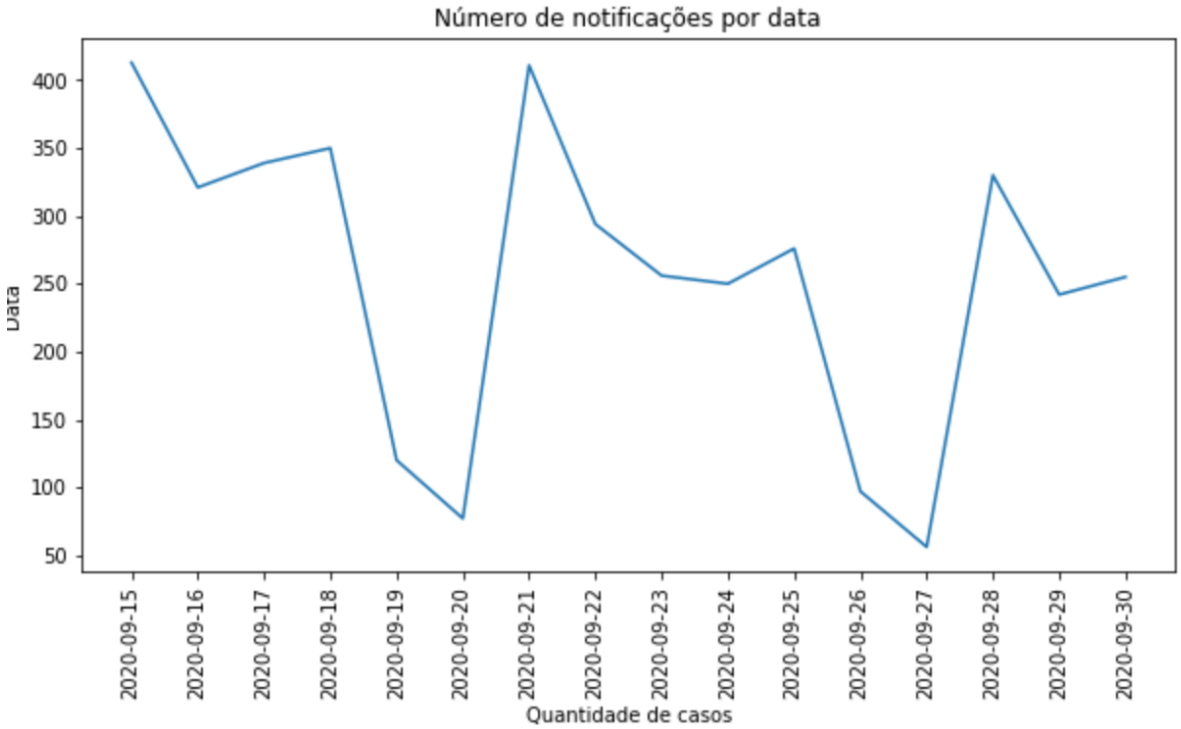
4 RESULTADOS

Os resultados da AED, usando o item 3.3.3 Estratificação de dados, são descritos abaixo:

4.1 Diária

Para encontrar picos de contágio, o gráfico diário mostra o número de notificações, por dia, mês e ano e é representado pelo Gráfico 1.

Gráfico 1 - Número de notificações do ano de 2020 dos 15 últimos dias do mês de setembro



Fonte: autoria própria

Há três picos com maior número de notificações no início de cada mês.

4.2 Profissional da saúde

Com apenas 3,3% da base, a análise parou após o resultado.

Gráfico 2 - Proporção de profissionais da saúde na base

Fonte: autoria própria

A quantidade de profissionais da saúde na base é de 3030 pessoas, representando apenas 3,3% do total.

4.3 Comorbidade/risco

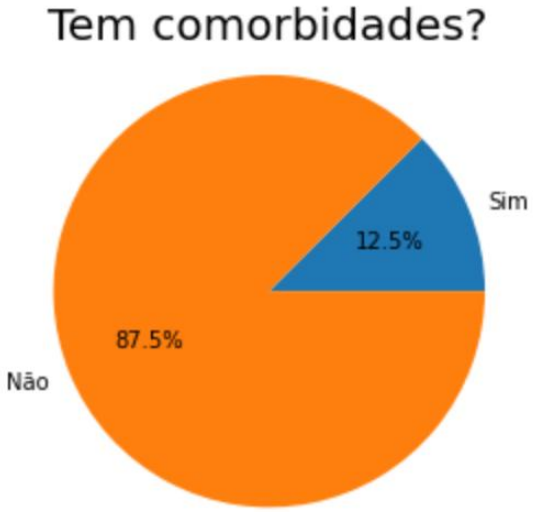
A AED sobre comorbidade/risco, constatou que 12,5% da base possuem comorbidades e que doenças cardíacas crônicas, diabetes e doenças respiratórias crônicas são as mais frequentes. Em relação aos número de dias na UTI, pessoas com e sem comorbidades, possuem métricas semelhantes.

Para as três comorbidades mais frequentes o número total de óbitos foi 921, 473, 168, respectivamente doenças cardíacas crônicas, diabetes e doenças respiratórias crônicas.

4.3.1 Geral

Para constatar a quantidade de pessoas comórbidas, filtrou-se a coluna ‘Tem comorbidades?’, resultando em 11449 pessoas, o que representa 12,5% da base e pode ser vista no Gráfico 3.

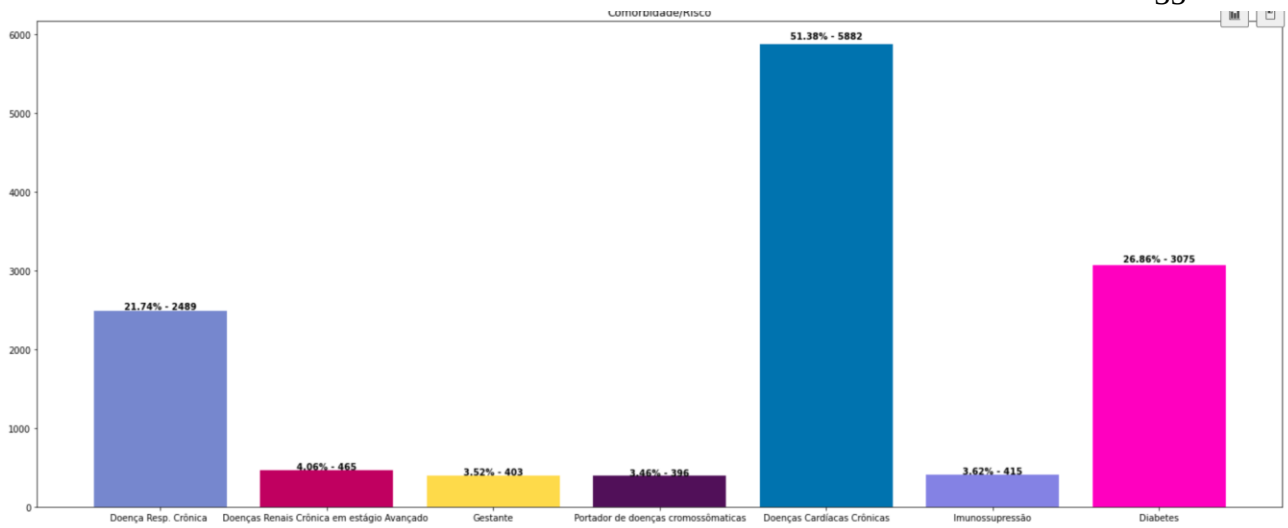
Gráfico 3 - Proporção de pessoas com comorbidades



Fonte: autoria própria

A partir dessa filtragem, considerando a estratificação comorbidades, foram encontrados as porcentagens no Gráfico 4:

Gráfico 4 - Quantidade por comorbidade



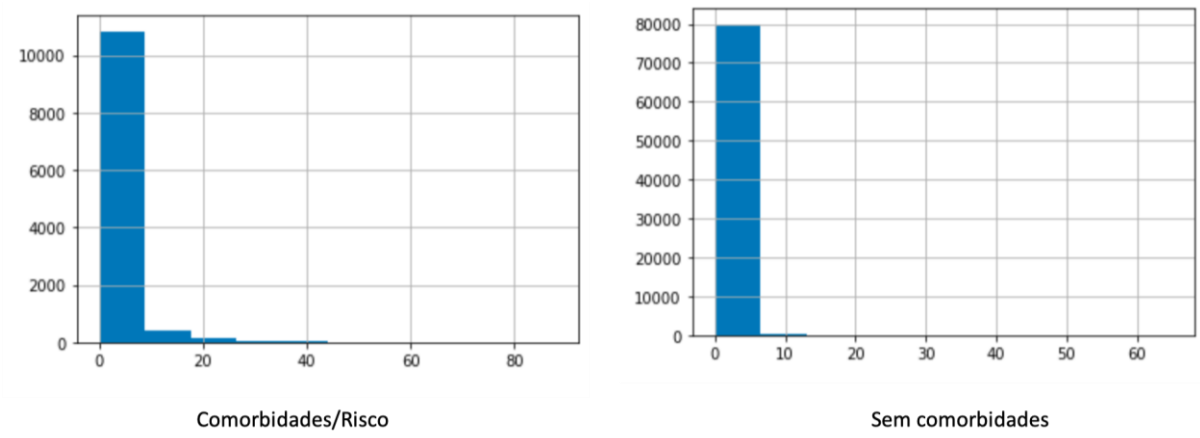
Fonte: autoria própria

Com base no gráfico temos que: das pessoas com comorbidades 465 pessoas, 4,1% possuem doenças renais crônicas em estágio avançado e 403 pessoas, ou seja, 3,5% são gestantes. 396 pessoas, 3,5% possuem doenças cromossomáticas e 5882 pessoas, ou seja, 48,6 % possuem doenças cardíacas crônicas. Doenças respiratória crônica, 21,7%, 2489 pessoas possuem a doença, 4,1%, sendo 465 pessoas com doenças renais crônicas em estágio avançado. E 26,9%, 3075 pessoas são diabéticas.

4.3.2 Dias na UTI

Em relação aos dias na UTI, é mostrado nos Gráficos 5 e 6.

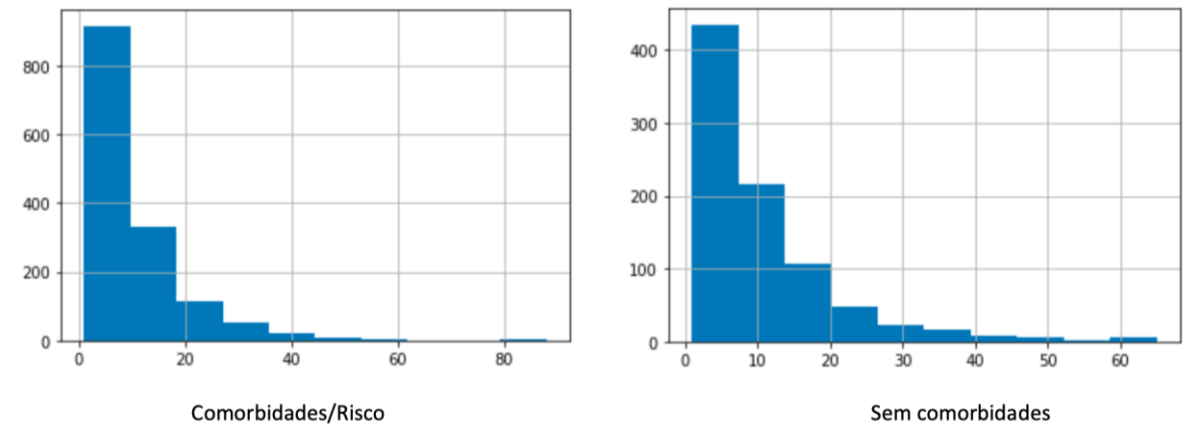
Gráfico 5 - Dias na UTI para pessoas com e sem comorbidades/risco



Fonte: autoria própria

Os histogramas estavam com frequência grande na região zero. Por isso, foi feito o refinamento abaixo considerando pelo menos 1 dia na UTI.

Gráfico 6 - Dias na UTI para pessoas com e sem comorbidades/risco, pelo menos 1 dia



Fonte: autoria própria

A partir do histograma verificamos que há mais pessoas com comorbidades que ficaram pelo menos 1 dia na UTI, os detalhes foram analisados em relação às métricas das tabelas 3 e 4.

Tabela 3 - Métricas pessoas com comorbidades/risco

Quantidade	1445.000000
Média	9.963322
Desvio Padrão	9.331649
Mínimo	1.000000
25%	4.000000
50%	7.000000
75%	13.000000
Máximo	88.000000

Fonte: autoria própria

Tabela 4 - Métricas pessoas sem comorbidades/risco

Quantidade	864.000000
Média	10.449074
Desvio Padrão	9.898193
Mínimo	1.000000
25%	4.000000
50%	7.000000
75%	13.000000
Máximo	65.000000

Fonte: autoria própria

É possível concluir que pessoas com comorbidades/risco em números e dias máximo é maior que pessoas sem comorbidades/risco, mas a média das pessoas sem comorbidades foi maior, sendo 10,45 contra 9,96 dias. O restante se manteve igual.

4.3.3 - Comorbidades com maior frequência

A partir da conclusão das comorbidades, a análise passou por um refinamento com as três comorbidades com maiores porcentagens, as doenças cardíacas crônicas - 48,6%, diabetes - 26,9% e doenças respiratórias crônicas - 21,7%. Com foco na evolução do paciente.

Os valores da coluna 'Evolução' estão na tabela 5, são informações referente a quantidade de pessoas curadas, falecidas, em isolamento domiciliar e internadas na enfermaria ou UTI:

Tabela 5 - Evolução da contaminação de COVID-19

Cura/Recuperado	89747
Óbito por COVID-19	1715

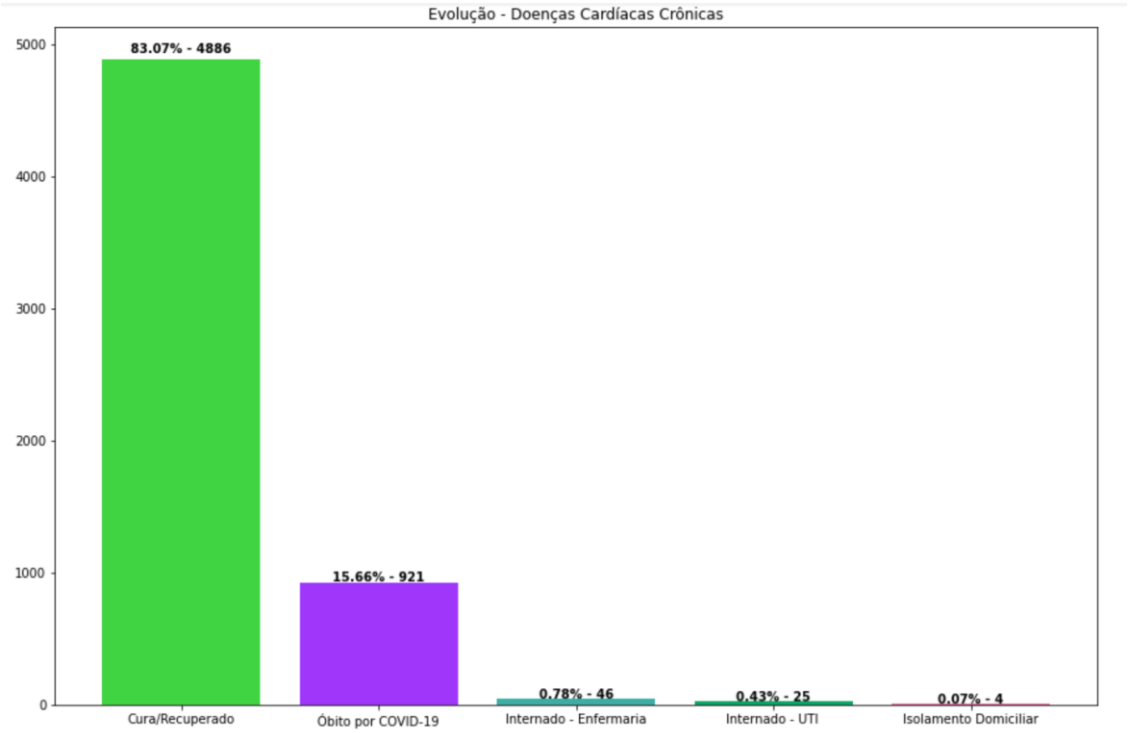
Internado - Enfermaria	170
Isolamento Domiciliar	120
Internado - UTI	82

Fonte: autoria própria

4.3.3.1 - Doenças cardíacas crônicas

No Gráfico 7, mostra quantidade de pessoas com doenças cardíacas crônicas em relação à evolução da pessoa.

Gráfico 7 - Quantidade por comorbidade com doenças cardíacas crônicas



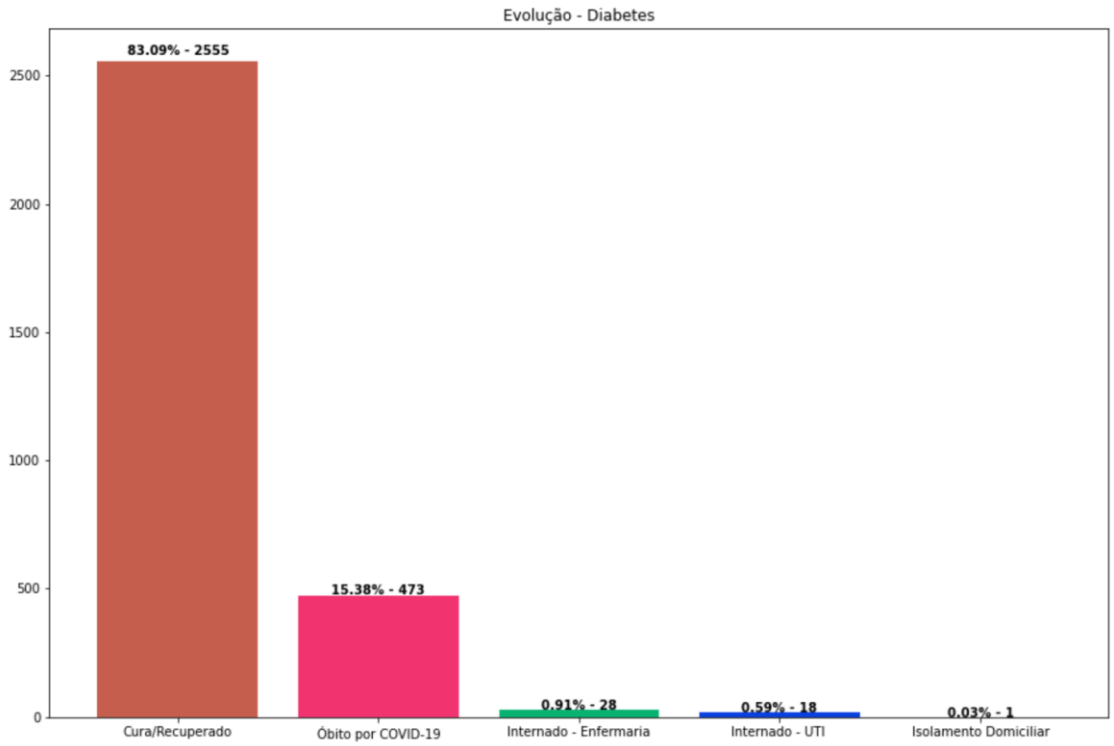
Fonte: autoria própria

Mais de 80% se recuperaram, mas 15,66% das pessoas com doença cardíaca crônica foram a óbito, no total de 921 pessoas.

4.3.3.2 - Diabetes

No Gráfico 8, mostra a quantidade de pessoas diabéticas em relação à evolução da pessoa.

Gráfico 8 - Quantidade por comorbidade com diabetes



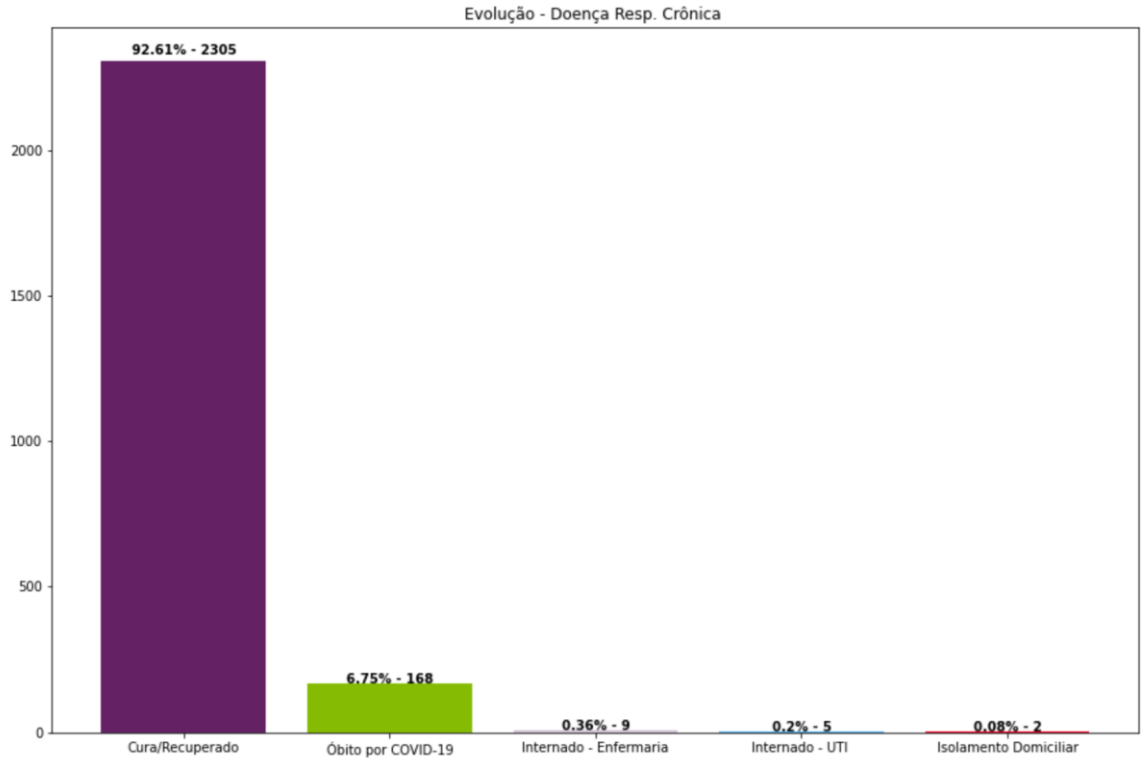
Fonte: autoria própria

Mais de 80% se recuperaram, mas 15,38% das pessoas com diabetes foram a óbito, no total de 473 pessoas.

4.3.3.3 - Doenças respiratórias crônicas

No Gráfico 9, mostra quantidade de pessoas com doenças respiratórias crônicas em relação à evolução da pessoa.

Gráfico 9 - Quantidade por comorbidade com doenças respiratórias crônicas



Fonte: autoria própria

Mais de 90% se recuperaram, a proporção de óbitos foi menor com 6,75% das pessoas com doença respiratória crônica, cerca de 168 pessoas.

A quantidade de óbitos é alta, considerando que 1715 é o total de óbitos da base. Porém, a análise não considerou a mesma pessoa ter duas ou mais dessas três comorbidades.

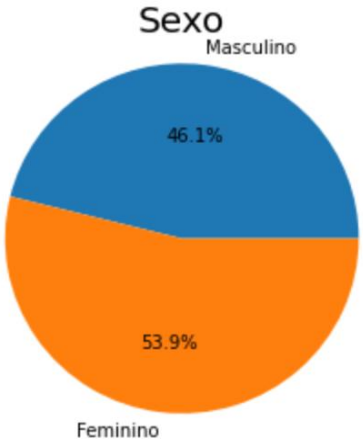
4.4 Sexo

A estratifição do sexo, masculino ou femino, foi analisado em relação a proporção entre os dois e depois a análise foi feita com base em idade e a evolução do paciente.

4.4.1 Geral

Proporção de mulheres e homens sobre a base é mostrada no Gráfico 10:

Gráfico 10 - Proporção feminino e masculino



Fonte: autoria própria

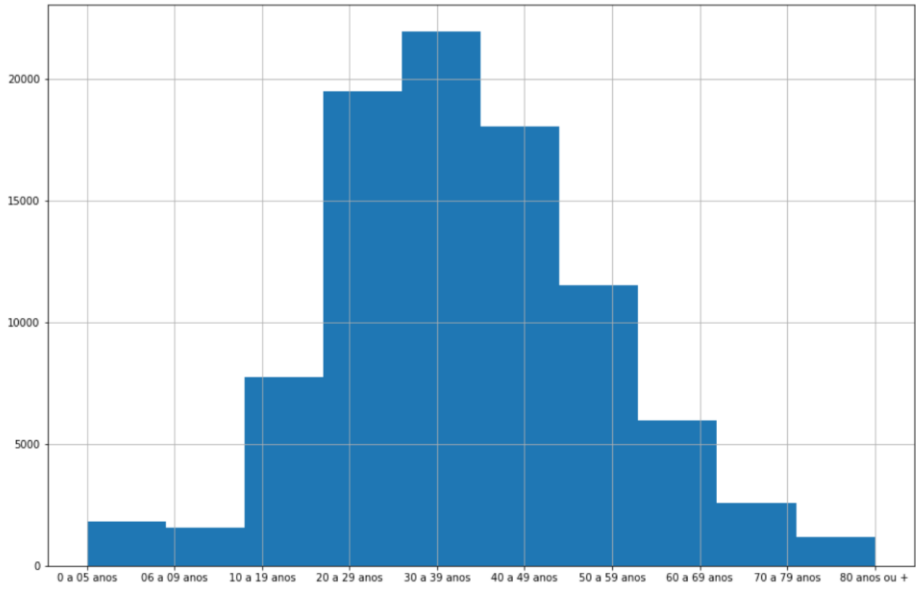
Há mais mulheres com COVID-19 do que homens, com a proporção de 53,9% do público feminino e 46,1% do público masculino.

4.4.2 Idade

Esta seção analisa a idade em geral e em relação ao sexo feminino e masculino, usando histograma, boxplot e gráfico Q-Q. O Gráfico 11 é um histograma

sobre a idade geral dos contaminados, mostra a frequência em relação a cada idade na base.

Gráfico 11 - Histograma geral coluna Outra Faixa etária



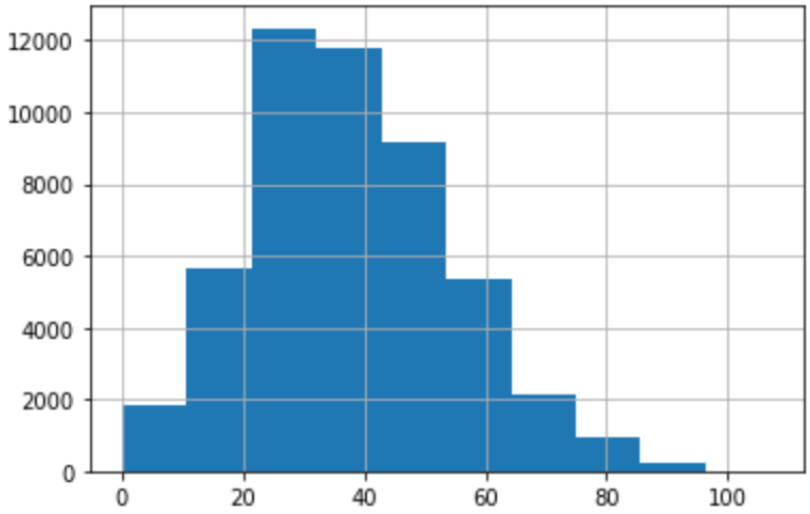
Fonte: autoria própria

As pessoas contaminadas com COVID-19 estão concentradas entre 10 a 60 anos, com maior concentração entre pessoas com 30 a 39 anos.

4.4.2.1 - Idade Feminino

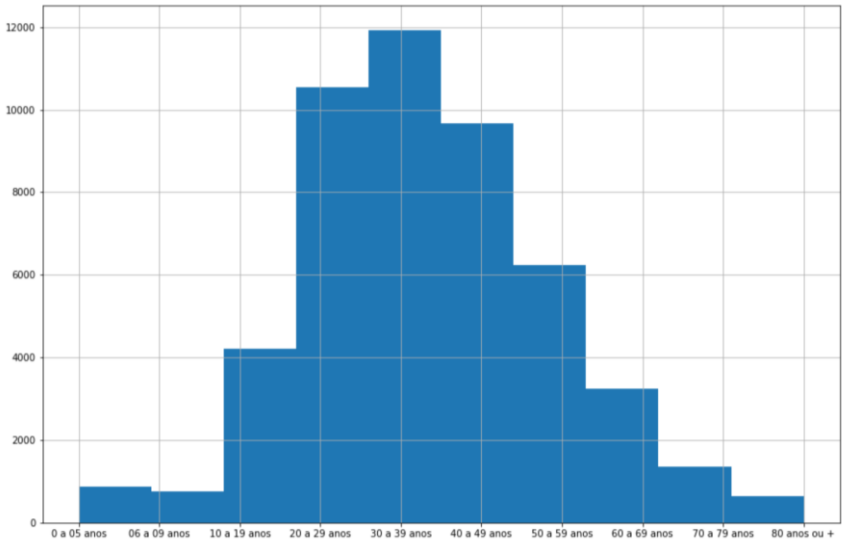
Abaixo os histogramas com relação às colunas idadeM e Outra Faixa etária, nos Gráficos 12 e 13 em relação ao sexo feminino, respectivamente.

Gráfico 12 - Histograma feminino com a coluna idadeM



Fonte: autoria própria

Gráfico 13 - Histograma feminino com a coluna Outra Faixa etária

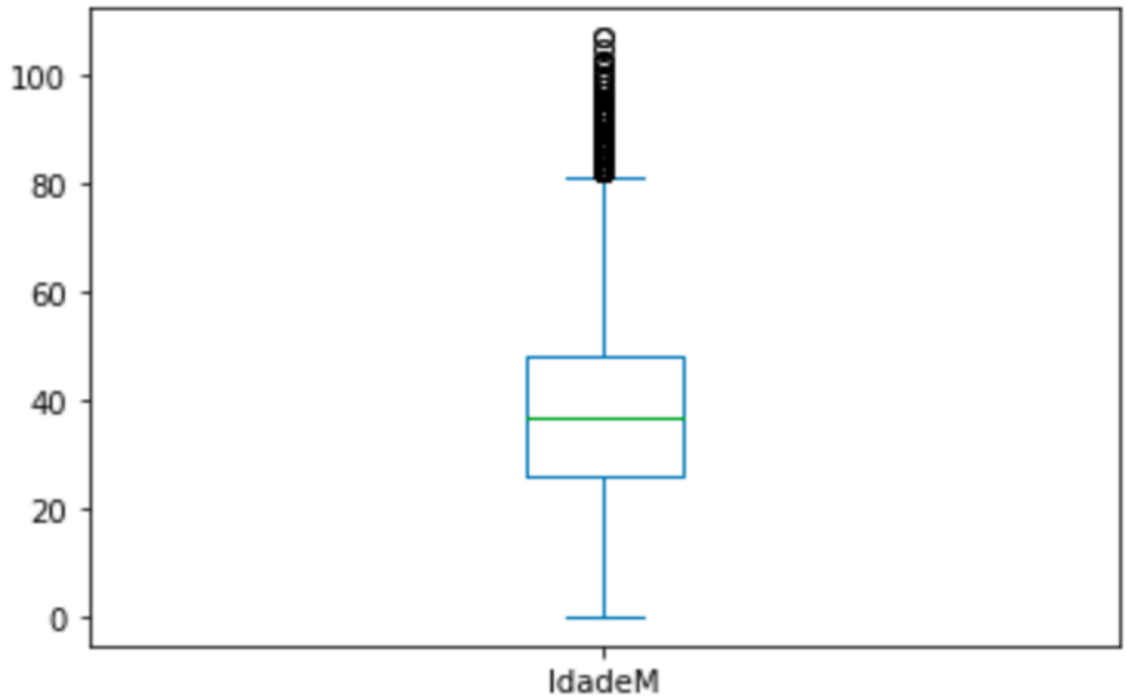


Fonte: autoria própria

As pessoas contaminadas com COVID-19 estão concentradas entre 10 a 60 anos. Em relação ao público feminino a tendência segue a mesma tendência da idade para o público geral.

Tendência de centralidade, simetria e valores atípicos, foi usado o Gráfico 14 e os valores das métricas a tabela 6.

Gráfico 14 - Boxplot idade mulheres



Fonte: autoria própria

Tabela 6 - Métricas idade mulheres

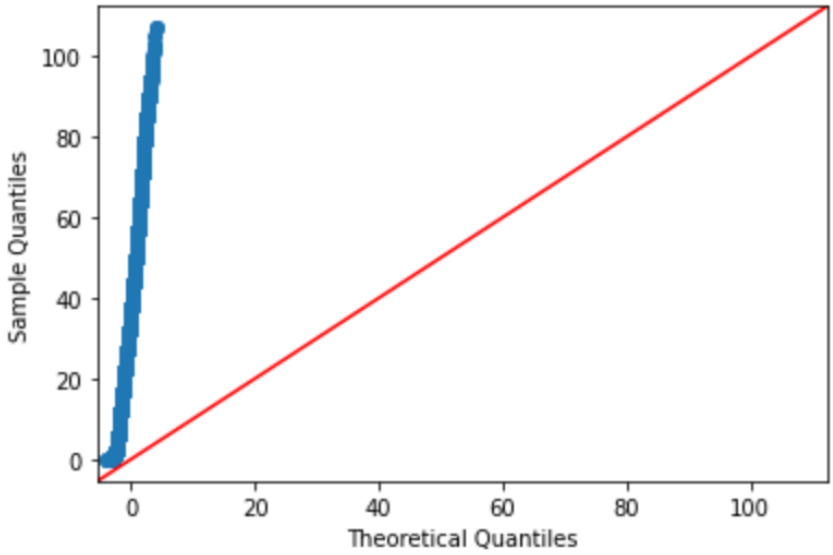
Média	37.896234
Desvio Padrão	16.606117
Mínimo	0.000000
25%	26.000000
50%	37.000000
75%	48.000000
Máximo	107.000000

Fonte: autoria própria

Em relação ao histograma, boxplot e as métricas da tabela 6, a média é de 38 anos, primeiro quartil 26 anos e terceiro quartil 48 anos. Mínimo de zero anos e máximo de 107 anos, com desvio padrão de 16,61. No boxplot, mostra mulheres com idades maiores de 80 anos, são valores atípicos na base.

Após, a análise da distribuição aplicou-se gráfico Q-Q, para verificar normalidade, sendo o Gráfico 15.

Gráfico 15 - Q-Q idade mulheres



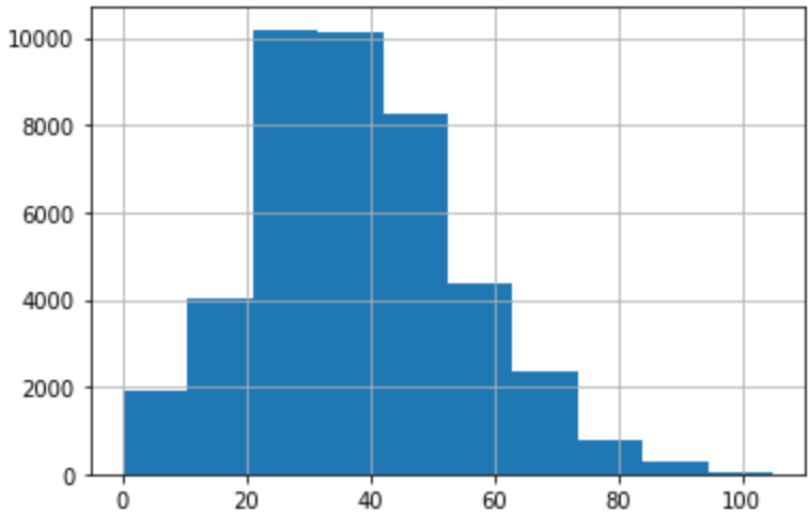
Fonte: autoria própria

Para a distribuição ser normal, os dados estariam seguindo a linha. Ou seja, não seguem uma distribuição normal.

4.4.2.1 - Idade Masculino

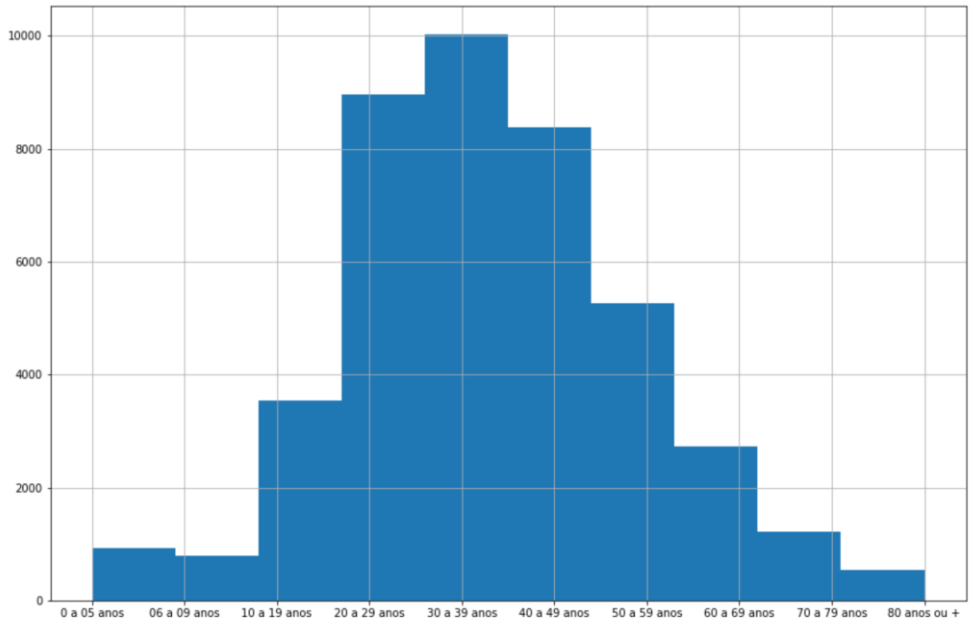
Abaixo os histogramas com relação às colunas idadeM e Outra Faixa etária, no Gráfico 16 e 17 em relação ao sexo masculino, respectivamente.

Gráfico 16 - Histograma idade homens com a coluna idadeM



Fonte: autoria própria

Gráfico 17 - Histograma idade homens com a coluna Outra Faixa etária

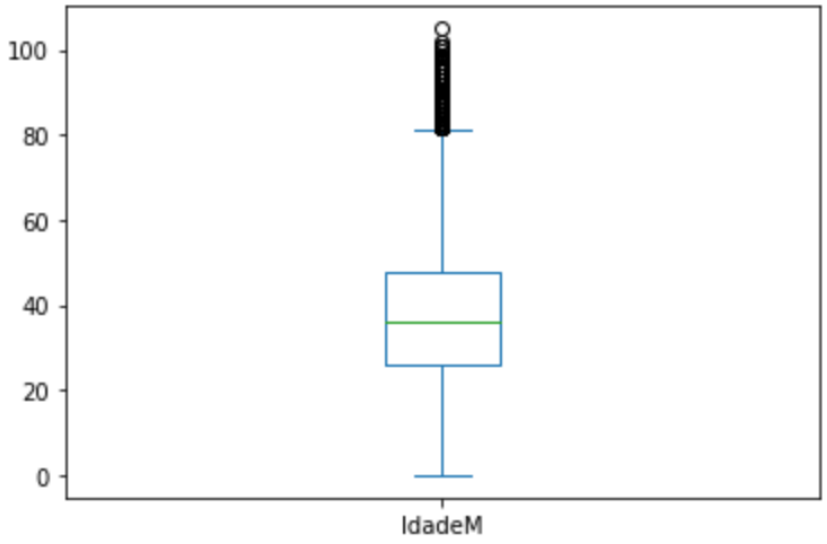


Fonte: autoria própria

Em relação ao público masculino também segue a tendência da idade para o público geral.

Para idade dos homens, a tendência de centralidade, simetria e valores atípicos, foi usado o Gráfico 18 e os valores das métricas a tabela 7.

Gráfico 18 - Boxplot idade homens



Fonte: autoria própria

Tabela 7 - Métricas idade homens

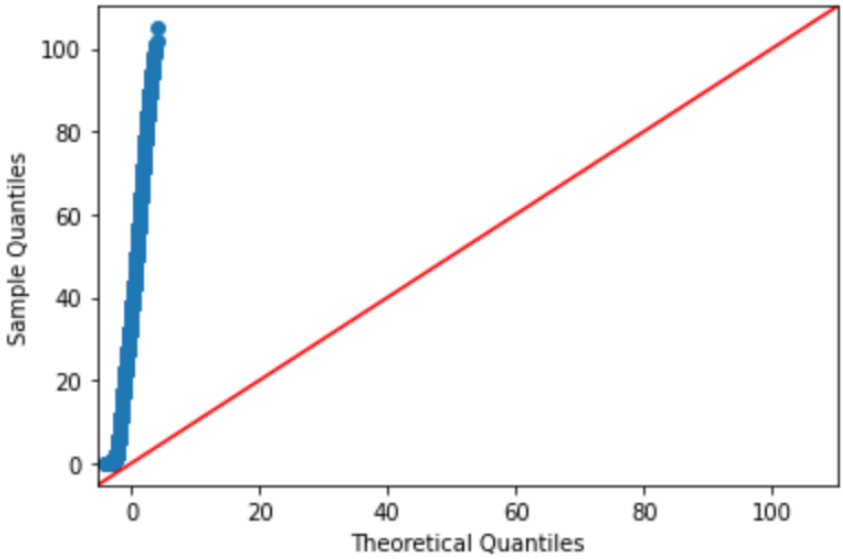
Média	37.641041
Desvio Padrão	16.786633
Mínimo	0.000000
25%	26.000000
50%	36.000000
75%	48.000000
Máximo	105.000000

Fonte: autoria própria

Em relação ao histograma, boxplot e as métricas da tabela 7, a média é um pouco abaixo de 40 anos, primeiro quartil 26 anos e terceiro quartil 48 anos. Mínimo de zero anos e máximo de 105 anos, com desvio padrão de 16,78.

Após, a análise da distribuição aplicou-se gráfico Q-Q, para verificar normalidade, sendo o Gráfico 19.

Gráfico 19 - Q-Q homens



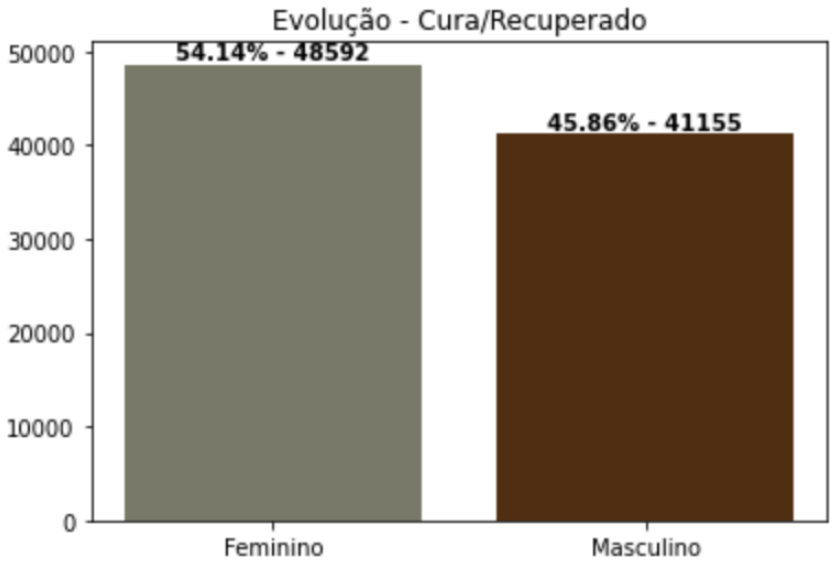
Fonte: autoria própria

Como os dados não estão sobre a linha vermelha, essa amostra não segue uma distribuição normal.

4.4.3 Evolução do paciente

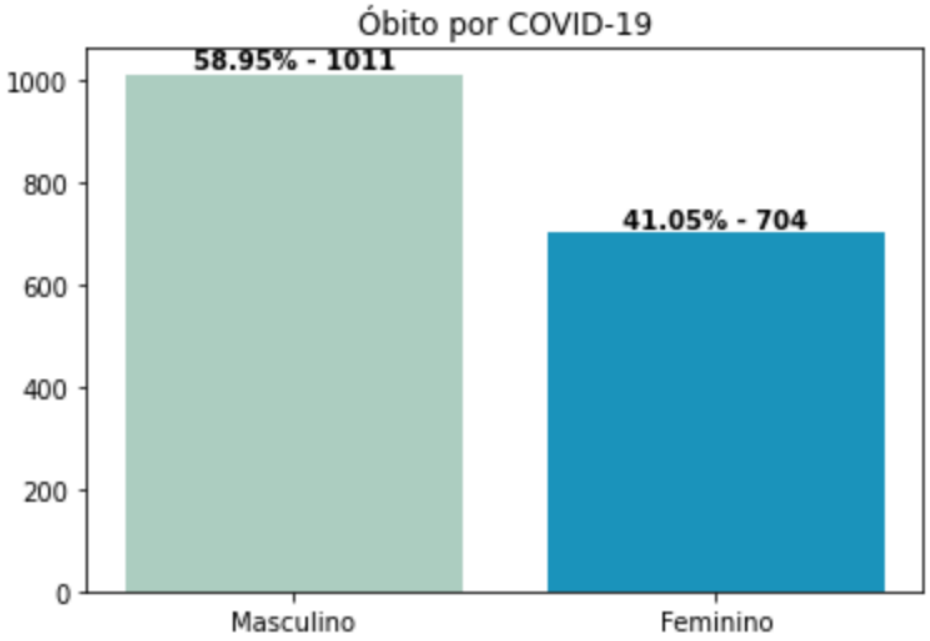
A proporção entre o sexo feminino e o sexo masculinho em relação da evolução da doenças estão nos gráficos 20 ao 24.

Gráfico 20 - Cura/Recuperado em relação ao sexo feminino e masculino



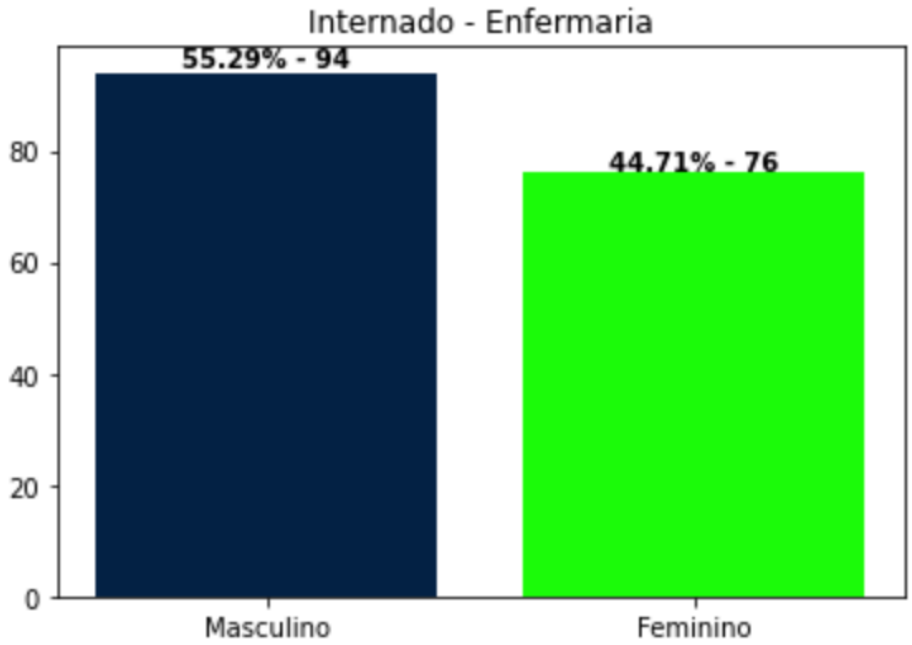
Fonte: autoria própria

Gráfico 21 - Óbitos por COVID-19 relação ao sexo feminino e masculino



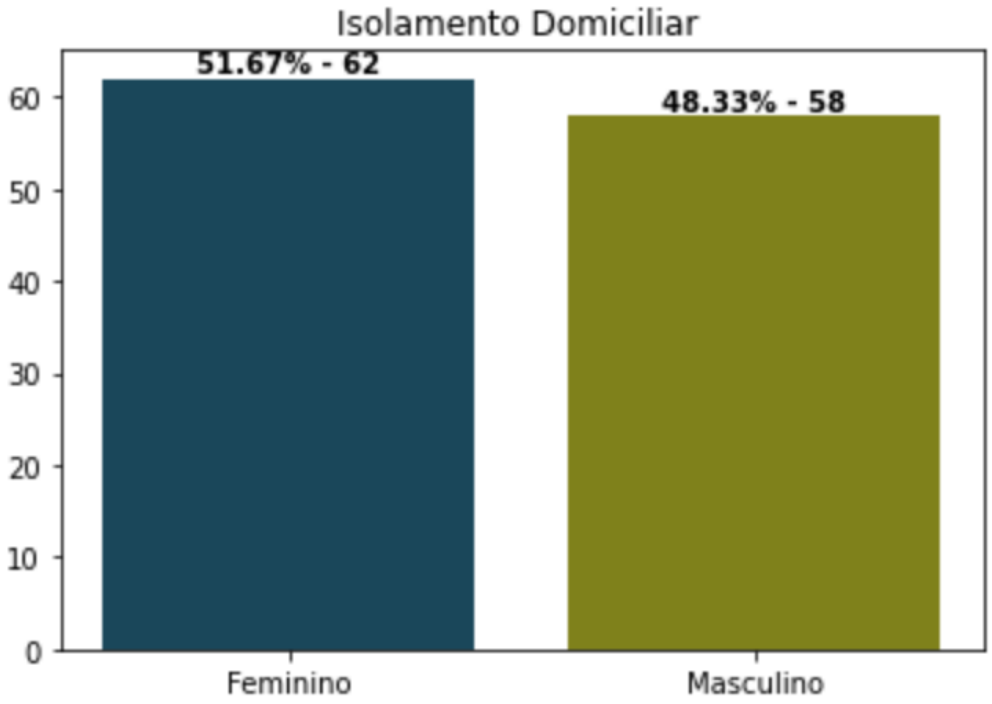
Fonte: autoria própria

Gráfico 22 - Internado - Enfermaria relação ao sexo feminino e masculino



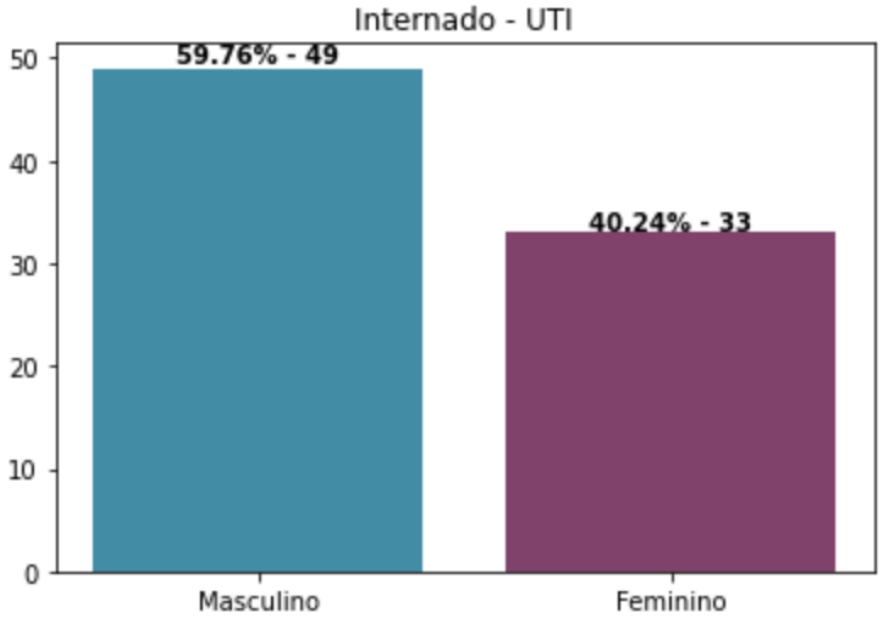
Fonte: autoria própria

Gráfico 23 - Isolamento Domiciliar relação ao sexo feminino e masculino



Fonte: autoria própria

Gráfico 24 - Internado - UTI relação ao sexo feminino e masculino



Fonte: autoria própria

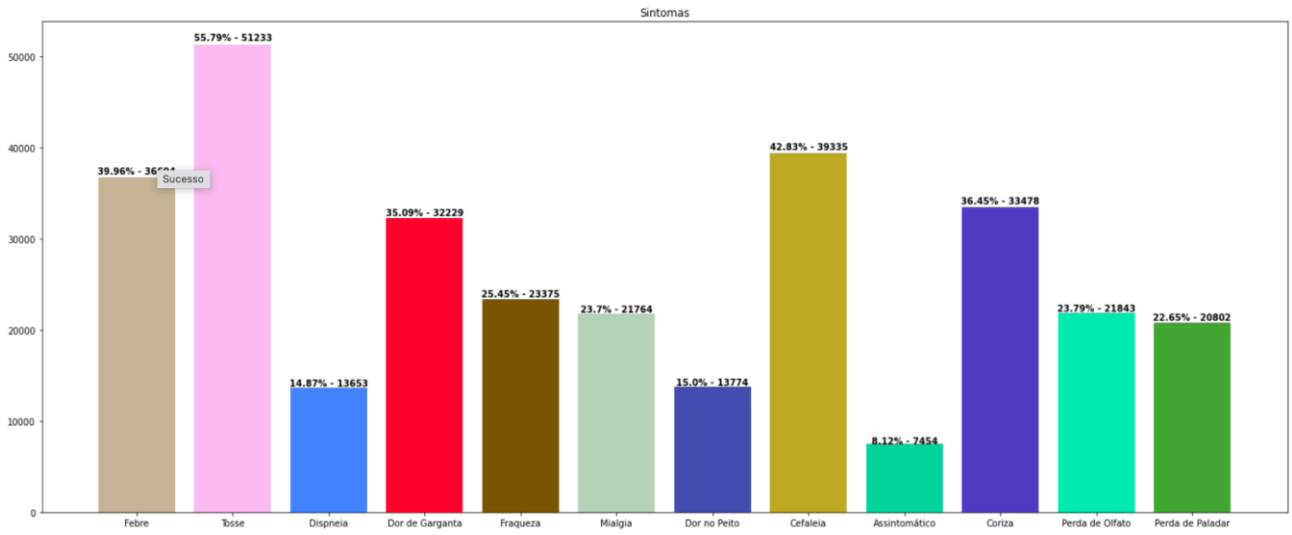
Em relação ao óbito por covid, internação enfermaria, UTI, o sexo masculino obteve a maior proporção, sendo respectivamente: 58,9%, 59,8% e 55,3% . Nas outras duas categorias: recuperados e no isolamento domiciliar, com proporção de 54,1% e 51,7%.

4.5 Sintomas

4.5.1 Geral

Na análise dos sintomas, usou-se a quantidade de sintomas que a pessoa tem, desconsiderando ser sintomas leves ou graves. Abaixo os gráficos dos sintomas no Gráfico 25 e o histograma em relação à quantidade de sintomas no Gráfico 26.

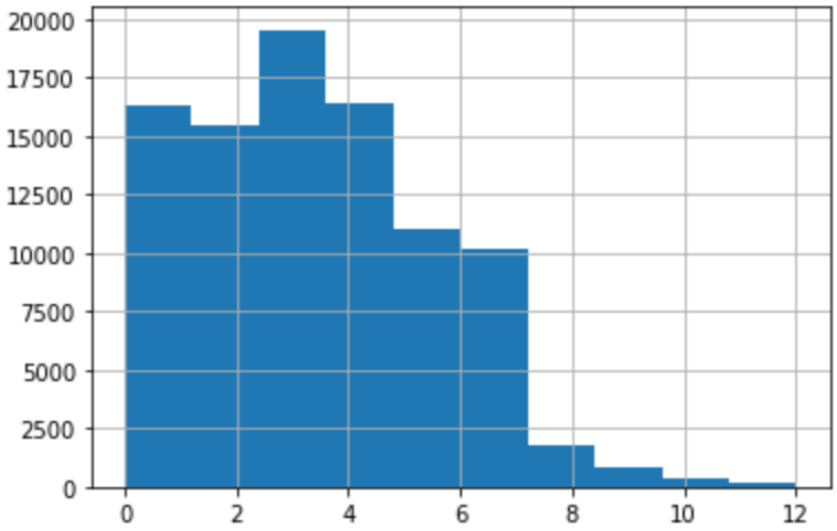
Gráfico 25 - Sintomas



Fonte: autoria própria

Sendo tosse, cefaléia e febre os sintomas mais comuns, com 55,79%, 42,83% e 39,96%, respectivamente.

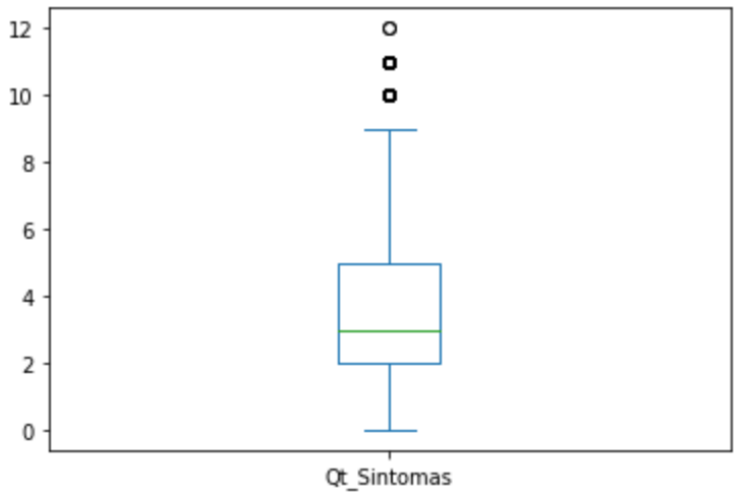
Gráfico 26 - Histograma quantidade de sintomas



Fonte: autoria própria

Para a quantidade de sintomas, a tendência de centralidade, simetria e valores atípicos, foi usado o Gráfico 27 e os valores das métricas a tabela 7.

Gráfico 27 - Boxplot quantidade de sintomas



Fonte: autoria própria

Tabela 8 - Métricas quantidade de sintomas

Média	3.437006
Desvio Padrão	1.956119
Mínimo	0.000000
25%	2.000000
50%	3.000000
75%	5.000000
Máximo	12.000000

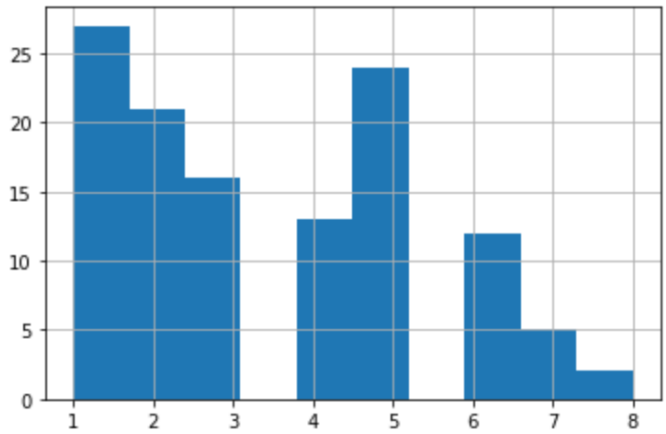
Fonte: autoria própria

Em relação ao histograma, boxplot e as métricas da tabela 8, a média é de 3 sintomas, mínimo de 0 sintomas e máximo de 12 sintomas. As outras métricas foram: aproximadamente 2 de desvio padrão, o primeiro quartil com 2 sintomas, mediana de 3 sintomas e o terceiro quartil com 5 sintomas, e os valores atípicos são sintomas de 9 para cima .

4.5.2 Evolução do paciente

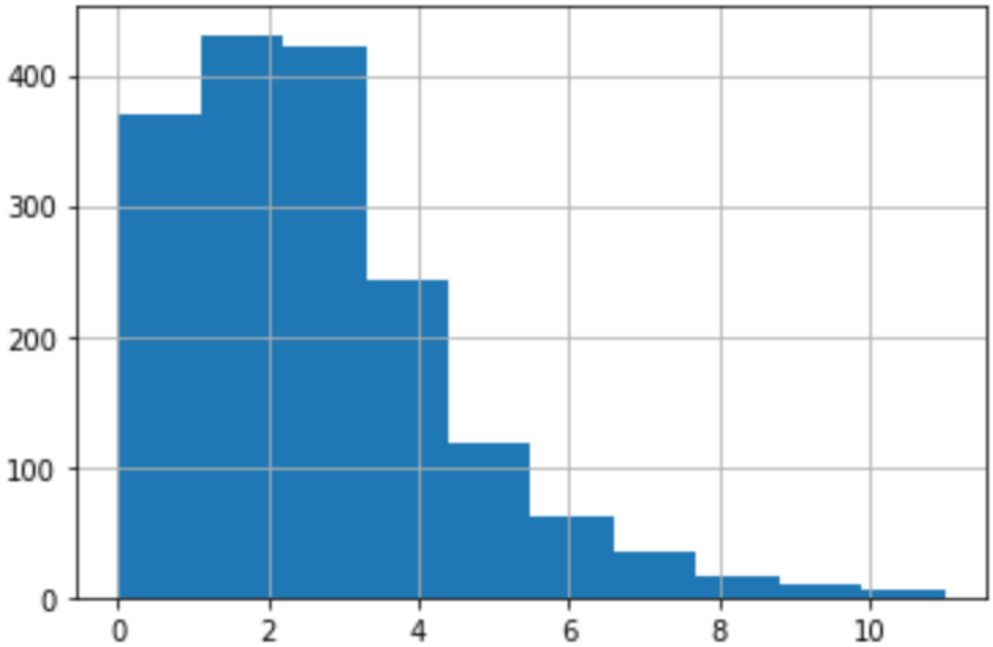
Analisa se a quantidade de sintomas influencia na evolução do paciente, como, recuperação e óbitos. Os gráficos 29 a 32, mostram essa evolução.

Gráfico 28 - Quantidade de sintomas em relação a Isolamento domiciliar



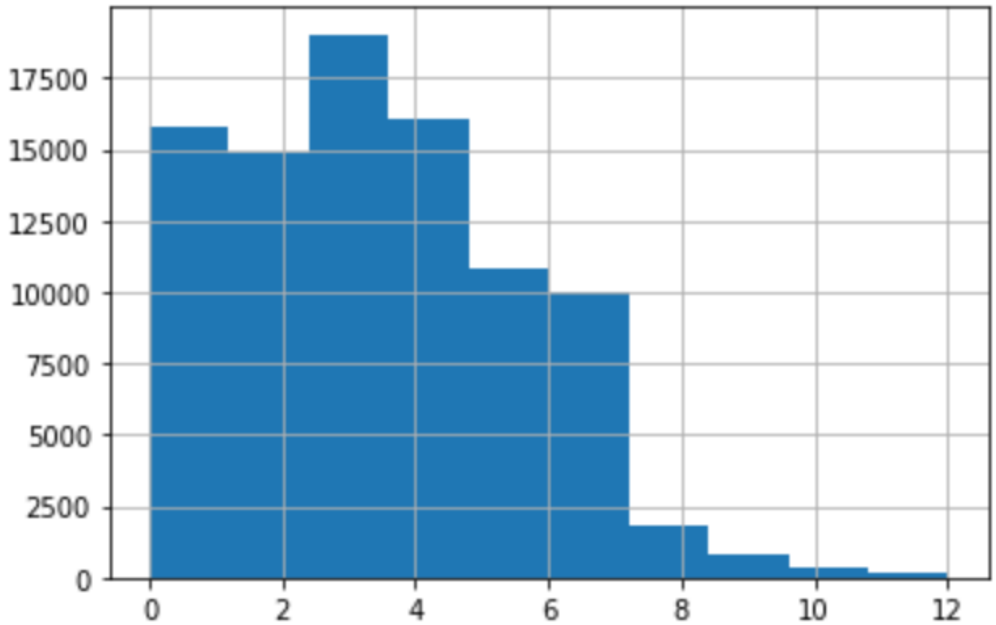
Fonte: autoria própria

Gráfico 30 - Quantidade de sintomas em relação a Óbitos por Covid-19



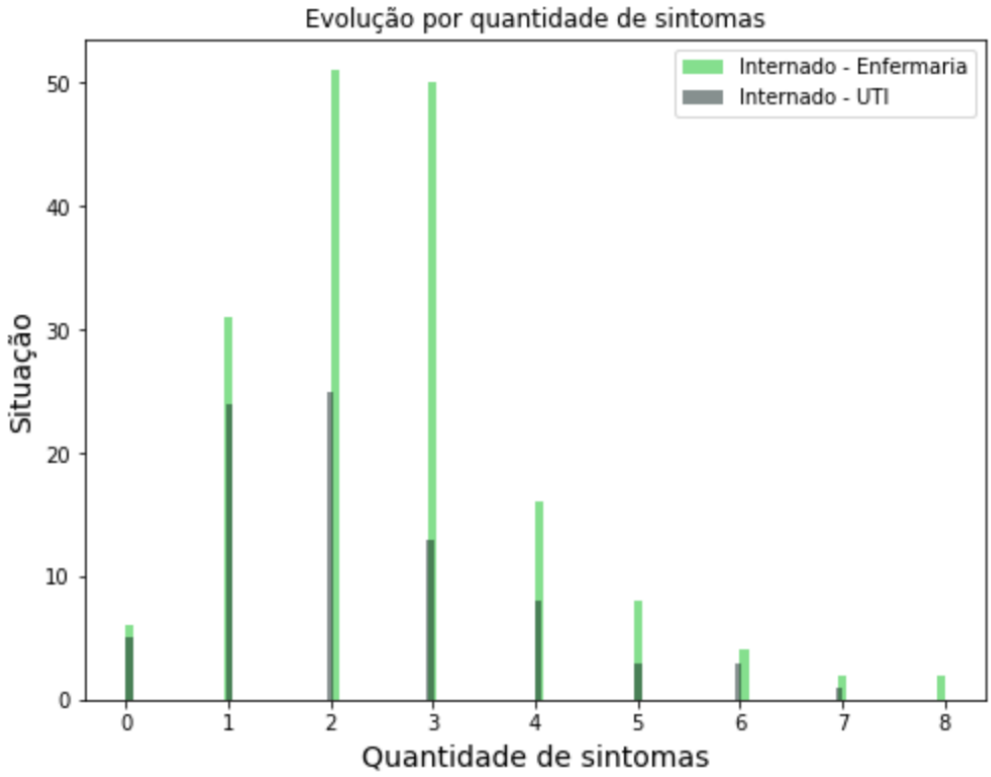
Fonte: autoria própria

Gráfico 29 - Quantidade de sintomas em relação a Cura/Recuperado



Fonte: autoria própria

Gráfico 30 - Quantidade de sintomas em relação a internação enfermaria e UTI



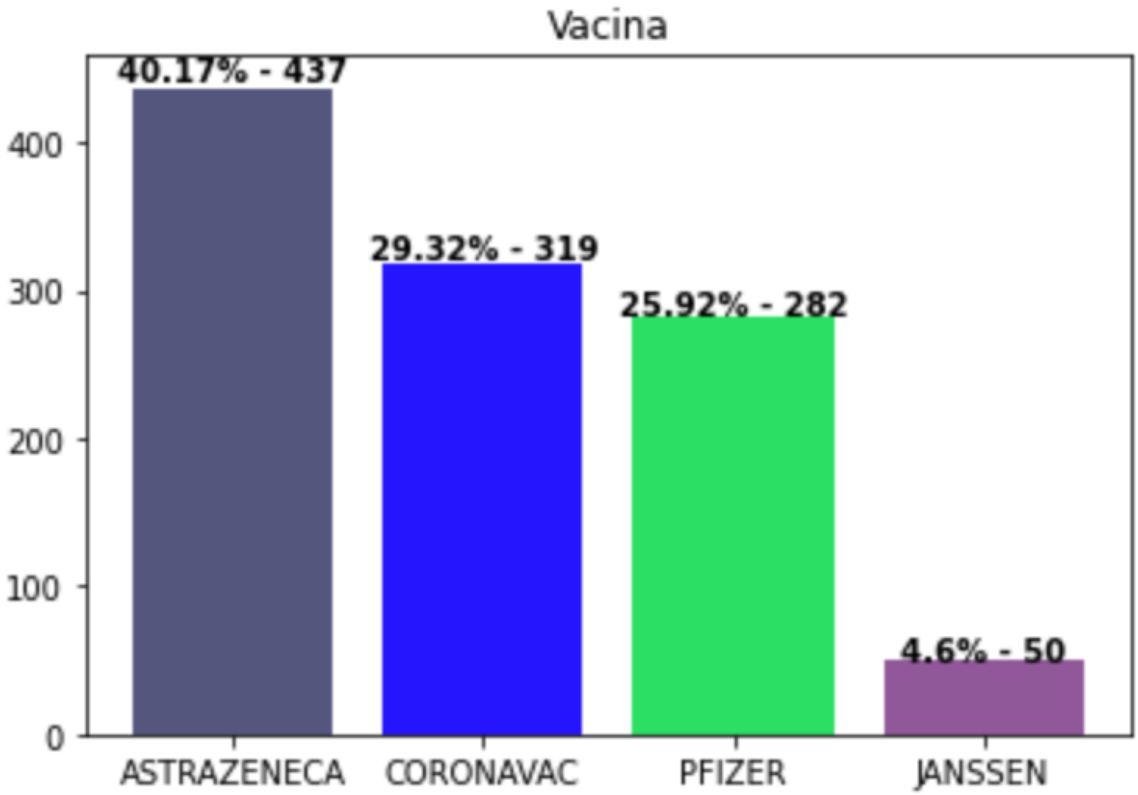
Fonte: autoria própria

Os dados são inconclusivos, era esperado que quanto mais sintomas, mais provável a internação ou o óbito, o que não foi obtido. Para uma próxima análise/ refinamento, considerar a gravidade dos sintomas.

4.6 Vacinados

Para o município de Aparecida de Goiânia, até o dia 28/10/2021, foram aplicadas quatro tipos de vacinas, sendo a mais frequente a Astrazeneca e pode ser visto no Gráfico 33.

Gráfico 31 - Vacinas tomadas



Fonte: autoria própria

Apenas 1088 pessoas da base tomaram pelo menos uma dose, representando 1,18% do total de registros, sendo a Astrazeneca a mais frequente com 437 pessoas, depois a Coronavac, com 319, Pfizer com 282 e por último a Janssen, com 50.

5 CONCLUSÕES

Esse trabalho propõe uma metodologia para análise exploratória de dados, para auxiliar profissionais de diferentes áreas a conseguirem produzir propostas de melhorias e soluções através dos dados. Os trabalhos em geral sobre AED, focam no estudo de caso e não na metodologia usada, enquanto este trabalho mostra como proceder uma AED e aplica sobre a base de dados de casos de COVID-19 no município de Aparecida de Goiânia. A AED é uma boa estratégia, pois a compreensão humana é mais sensível aos dados visuais do que os dados em texto. Esse trabalho tem como originalidade a proposta de uma metodologia geral, para profissionais da área da saúde conseguirem avaliar situações de risco e tomarem decisões rápidas, usando ferramentas de *software* livre, pois permite a facilidade de adaptação para diferentes tipos de dados.

Para a proposta, foram utilizadas o *software* livre jupyter notebook, a linguagem python e suas bibliotecas, por facilidade de instalação e adaptação em relação a diferentes bases de dados. Possibilitando democratizar a AED e a linguagem programação, permitindo que diferentes profissionais tomem decisões mais assertivas.

Em trabalhos futuros, serão comparadas diferentes plataformas e linguagens de programação, como, Matlab, R e Python, possibilitando a escolha do profissional. Aprofundamento sobre cada etapa da metodologia criada, explicando técnicas usadas por analistas e cientistas de dados para cada tipo de variável nas diferentes linguagens e plataformas. Além disso, pretende-se avaliar a metodologia, aplicando testes sobre um grupo de controle e um grupo experimental, com intuito de avaliar e comparar a curva de aprendizado dos dois grupos.

REFERÊNCIAS

Análise exploratória de dados. Disponível em: <http://leg.ufpr.br/~fernandomayer/aulas/ce001e-2016-2/02_Analise_Exploratoria_de_Dados.html>.

BIELAK, M. Python vs Scala - Know the Top 14 Differences, dez 2021. Disponível em:<https://www.netguru.com/blog/python-versus-scala>. Acesso em: 7 jun.2022.

BRITO, B.O., LEITÃO, L.P.C. Telemedicina no Brasil: Uma estratégia possível para o cuidado em saúde em tempo de pandemia? **RevistaSaúde em Redes**, v.6, p. 7-19,2020. Disponível em: http://revista.redeunida.org.br/ojs/index.php/rede-unida/article/viewFile/3202/550. Acesso em: 25 maio. 2022.

CENTRO DE ESTATÍSTICA APLICADA. **"Correlação (Pearson, Kendall, Spearman)**. Disponível em: <https://estatistica.pt/correlacao-pearson-kendall-spearman/>. Acesso em: 21 jun. 2022.

CHEN, L. et al. Exploratory Data Analysis on the Usage of COVID-19 Vaccine. Disponível em: <https://ieeexplore.ieee.org/document/9644489>. Acesso em: 18 abr. 2022.

DE BARROS BALTAR, M. L.; RIBEIRO, G. M.; SANTOS, R. F. Exploratory analysis of incidents in an urban area focused on broken-down vehicles: The case of Rio de Janeiro. Case Studies on Transport Policy, v. 10, n. 1, p. 723–731, mar. 2022. Disponível em: <https://bundles.yourlearning.ibm.com/ibm/ibm-ai-skills-academy/#EKEMRGDJVQYM22KW/JYNVWPRREQZR2EMW>. Acesso em: 25 maio. 2022.

DSOUZA, J.; VELAN S., S. Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases. Disponível em: <https://ieeexplore.ieee.org/document/9225621>.Acesso em: 30 maio. 2022.

FREE SOFTWARE FOUNDATION. **What is free software and why is it so important for society? — Free Software Foundation — working together for free software**. Disponível em: <https://www.fsf.org/about/what-is-free-software>.

HAU, Yong Sauk; KIM, Jeoung Kun; HUR, Jian; CHANG, Min Cheol. How about actively using telemedicine during the COVID-19 pandemic? **Journal of medical systems,** v. 44, n. 6, p. 1-2, 2020. Disponível em: https://link.springer.com/article/10.1007/s10916-020-01580-z Acesso em: 24 maio. 2022.

HIMSS Media. Modernizing healthcare technology for today's needs and tomorrow's possibilities. [s.l: s.n.]. Disponível em: <https://www.ibm.com/downloads/cas/Y4KZVK5O>.

IBM Cloud Education. O que é a Análise exploratória de dados? Disponível em: <https://www.ibm.com/br-pt/cloud/learn/exploratory-data-analysis>. Acesso em: 7 jun. 2022. Ano 2020

IBM; Set, 2016.<https://courses.yl.skillsnetwork.site/courses/course-v1:CognitiveClass+DS0103EN+v3/course/> Acesso em: 6 jun. 2022.

IBM. **Data Science for All IBMers**. Disponível em: <https://bundles.yourlearning.ibm.com/ibm/ibm-ai-skills-academy/#EKEMRGDJVQYM22KW/ZKVGDMPKGEDX1R6X>. Acesso em: 20 maio. 2022.

Install and Use —Jupyter Documentation 4.1.1 alpha documentation, 2015. Disponível em: <https://docs.jupyter.org/en/latest/install.html#jupyter-notebook-interface>.Acesso em: 25 maio. 2022.

JUPYTER. Project Jupyter, 2019.Disponível em: <https://jupyter.org/>. Acesso em: 25 maio. 2022.

KAPKO, M. **ferramentas gratuitas de análise de dados**. Disponível em: <https://itforum.com.br/noticias/7-ferramentas-gratuitas-de-analise-de-dados-que-voce-deve-conhecer/>.Acesso em: 7 jun. 2022.

L13709compilado. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm>. Acesso em: 7 jun. 2022.

Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: <https://www.gov.br/cidadania/pt-br/acesso-a-informacao/lgpd>. cap 1, Art.5º

Rezende, F. E; Barbosa, T. M. G. de A. **ESTUDO DESCRITIVO SOBRE O SERVIÇO DE TELEMEDICINA NO ACOMPANHAMENTO DE PACIENTES DIAGNOSTICADOS COM COVID-19 NO MUNICÍPIO DE APARECIDA DE GOIÂNIA. -** abril, 2022**.**

SAINI, S. K. et al. Visual Exploratory Data Analysis of COVID-19 Pandemic. Disponível em: <https://ieeexplore.ieee.org/document/9358331>. Acesso em: 30 ago. 2021.

SWEETLIN, E. J.; SAUDIA, S. Exploratory Data Analysis on Breast cancer dataset about Survivability and Recurrence. 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 13 maio 2021.

TAULLI, T. Artificial intelligence basics : a non-technical introduction. Berkeley, California: Apress, 2019. cap 3 - Machine Learning. Disponível em: <https://learning.oreilly.com/library/view/artificial-intelligence-basics/9781484250280/html/480660_1_En_3_Chapter.xhtml>. Acesso em: 25 maio. 2022.

What is free software and why is it so important for society? — Free Software Foundation — working together for free software. Disponível em: <https://www.fsf.org/about/what-is-free-software>.

APÊNDICES

Apêndice A – Códigos

Gráfico Diário – Seleciona ano, meses e dias para serem mostrados

Função para selecionar o período

```
def escolhe_ano_mes(ano, meses, df):
    '''
        Visualiza quantidade de casos por mês em um ano escolhido

        Retorna o Dataframe ordenado pelas datas
    '''
    filtro = pd.DataFrame()
    diaria = []
    for m in range(0,len(meses)):
        if(len(meses[m]) == 1):
            meses[m] = '0'+ meses[m]
    meses.sort()
    ano = sorted(ano)
    for a in ano:
        for m in meses:
            if(len(m) == 1):
                m = '0' + m
            filtro = df[df['Data da Notificação:'].str.startswith(f'{a}-{m}')]
            for dia in range(15,32): #seleciona os dias que quer do mês
                filtro1 = filtro[(filtro['Data da Notificação:'].str.startswith(f'{a}-{m}-{dia}')) | (filtro['Data da
Notificação:'].str.startswith(f'{a}-{m}-0{dia}'))]
                if(len(filtro1) > 0):
                    diaria.append([f'{a}-{m}-{dia}', len(filtro1)])
    df_ano_mes = pd.DataFrame(diaria, columns=['Periodo', 'Contagem'])
    # df_ano_mes.sort_values(by='Periodo') #ordena de acordo com o dia do mês e mês

    return df_ano_mes
```

Gera o gráfico diário

```
#tamanho do gráfico
plt.figure(figsize=(10,5))

#nome do eixo x e y, respectivamente
plt.xlabel('Quantidade de casos')
plt.ylabel('Data')

#rotacionar as datas
plt.xticks(rotation = 90)

#seleciona meses e o ano
lista_meses = ['9'] #coloque preferenciamente em sequência
df_visualiza = escolhe_ano_mes(['2020'], lista_meses , df_covid)
```

Gera o gráfico diário - continuação

```
#Valores únicos da coluna e conta quantas ocorrências tem
valores_x = (df_visualiza)
plt.plot(df_visualiza['Periodo'], df_visualiza['Contagem'])

# Título do gráfico
plt.title('Número de notificações por data')

#Mostra o gráfico
plt.show()
```

Gráfico de barras com porcentagem em cima de cada barra

Eixo x e y, gerado a partir da quantidade de cada valor por categoria de uma coluna do DataFrame.

Exemplo de chamada da função:

Filtra onde a coluna ‘Diabetes’ do DataFrame df_comorbidade for igual a ‘Sim’

```
grafico_barra_com_porcentagem(df_comorbidade[df_comorbidade['Diabetes'] == 'Sim'],'Evolução', 'Evolução - Diabetes', figsize = (15,10))
```

```
def grafico_barra_com_porcentagem(df, column_, titulo, figsize = (15,10)):
    '''
        Gera gráfico de barras, a partir do Dataframe e coluna escolhida

    '''
    # calcula a porcentagem de cada valor
    porcentagem = []
    eixo_y = list(df[column_].value_counts())
    eixo_x = df[column_].value_counts().index.tolist()
    porcentagem = list(df[column_].value_counts() / ((df[column_].dropna()).shape[0]) * 100)

    color = gerar_lista_cores(len(eixo_x))
    plt.figure(figsize=figsize)
    plt.title(titulo)
    grafico = plt.bar(eixo_x, eixo_y, color = color) #barras, eixo x e eixo y

    #para colocar a porcentagem em cima das barras
    i = 0
    for p in grafico:
        width = p.get_width()
        height = p.get_height()
        x, y = p.get_xy()
        plt.text(x+width/2,
                 y+height*1.01,
                 str(round(porcentagem[i],2))+f'% - {eixo_y[i]}',
                 ha='center',
                 weight='bold')
        i+=1

    plt.show()
    return None
```

Gráfico de barras com porcentagem em cima de cada barra

Eixo x e y, passados como parâmetros da função – eixo_x e eixo_y

Exemplo de chamada:

Criação de um DataFrame com a quantidade de cada sintoma e depois aplicado a função grafico_barra_com_porcentagem_2.

```
lista = []
for col in sintomas:
    aux = df_covid[df_covid[col] == 'Sim']
    lista.append([col, len(aux)])
df = pd.DataFrame(lista, columns=['Coluna', 'Quantidade'])
tamanho_do_dataframe = df_covid.shape[0]

grafico_barra_com_porcentagem_2(df['Coluna'], df['Quantidade'],'Sintomas',tamanho_do_dataframe ,figsize =
(25,10))
```

```
def grafico_barra_com_porcentagem_2(eixo_x,eixo_y, titulo, tamanho, figsize): #passando os valores de x e y

    # calcula a porcentagem de cada valor
    porcentagem = []
    porcentagem = list(eixo_y / tamanho * 100)

    color = gerar_lista_cores(len(eixo_x))
    plt.figure(figsize=figsize)
    plt.title(titulo)
    grafico = plt.bar(eixo_x, eixo_y, color = color) #barras, eixo x e eixo y

    #para colocar a porcentagem em cima das barras
    i = 0
    for p in grafico:
        width = p.get_width()
        height = p.get_height()
        x, y = p.get_xy()
        plt.text(x+width/2,
                 y+height*1.01,
                 str(round(porcentagem[i],2))+f'% - {eixo_y[i]}',
                 ha='center',
                 weight='bold')
        i+=1

    plt.show()
    return None
```

Gráfico Boxplot

A partir da coluna ‘Qt_Sintomas’, foi gerado o gráfico do tipo ‘box’

```
df_covid['Qt_Sintomas'].plot(kind = 'box')
```

Gráfico Histograma

```
df_saudavel['Qtd dias UTI'].hist()
```

Gráfico de Pizza

Exemplo de uso:

```
grafico_pizza(df_covid, 'Tem comorbidades?')
```

```
def grafico_pizza(df, column):
    columns = df[column].unique()
    lista_prop = []
    lista_nome = []
    for i in columns:
        lista_nome.append(i)
        lista_prop.append(len(df[df[column] == i])/ df.shape[0])
    fig, ax = plt.subplots()
    ax.pie(lista_prop, labels=lista_nome, autopct='%1.1f%%')
    ax.axis('equal')

    #título do gráfico
    plt.title(column, size = 20)

    #mostrar o gráfico
    plt.show()
    return None
```

Visualizar valores único de uma coluna

```
def visualizar_valores_unicos_coluna(df, colunas):
    for col in colunas:
        print(df[col].unique())
```

Uso da biblioteca pandas_profiling, só abrir o arquivo gerado ‘output.html’ em um navegador

```
prof = ProfileReport(df_covid)
prof.to_file(output_file='output.html')
```

Gráfico Correlação de Pearson

```
fig, x = plt.subplots(figsize=(15,15))
corrMatrix = df_covid.corr()
sns.heatmap(corrMatrix, annot = True)
plt.show()
```

Gráfico Q-Q

```
#gráfico q-q da idade masculina
sm.qqplot(df_masculino['IdadeM'], line='45')
pylab.show()
```

Histograma - quantidade de sintomas por internado na UTI e internado na enfermaria

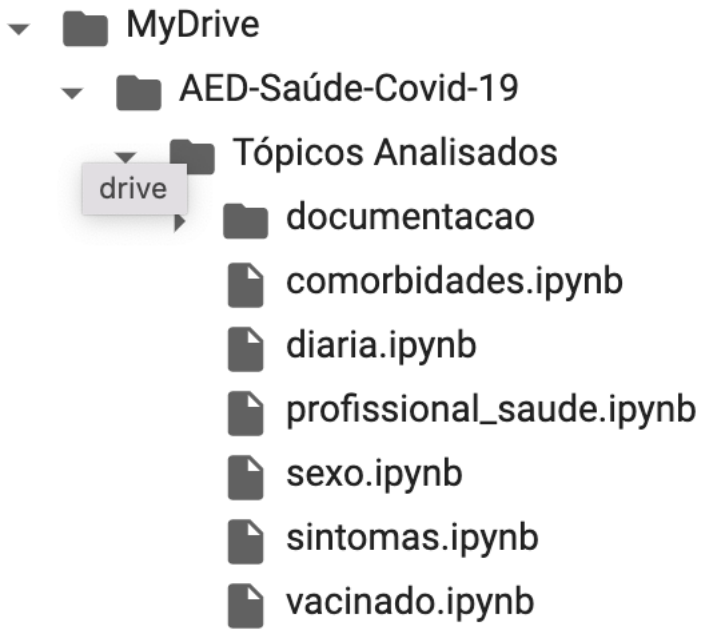
```
plt.figure(figsize=(8,6))
for col in evolucao:
    print(col)
    aux = df_covid[df_covid['Evolução'] == col]
    color = gerar_lista_cores(1)
    plt.hist(aux['Qt_Sintomas'], bins=100, alpha=0.5, label=col, color=color)

plt.xlabel("Quantidade de sintomas", size=14)
plt.ylabel("Situação", size=14)
plt.title("Evolução por quantidade de sintomas")
plt.legend(loc='upper right')
```

Gerar cores aleatórias

```
import random
def gerar_lista_cores(qt):
    color = []
    for i in range(0,qt):
        aux = list(np.random.choice(range(256), size=3))
        for i in range(0,3):
            aux[i] = aux[i]/255
        color.append(aux)
    return color
```

Apêndice B – Manual de uso do Google Colab

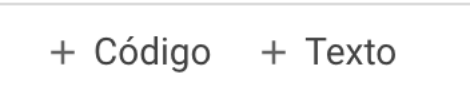
Ordem dos arquivos no Google Drive



Para copiar a pasta o link é:

https://drive.google.com/drive/folders/1vTL6XM_cmXq_rQ2DvmlVlAxwFD2PaMpb?usp=sharing

No notebook, podem ser criadas dois tipos de células, de código ou de texto.



Parar executar células de código, só clicar no botão mais a esquerda da célula

Para executar os arquivos, é necessário se conectar ao seu drive, é feito com o comando abaixo.

Obs: não esqueça de copiar a pasta AED-Saúde-Covid-19 para o seu drive

O comando abaixo para se conectar ao drive

```
[1] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Para as funções de processamento e análise, é necessário executar o arquivo 2_funcoes.ipynb

Executar todas funções necessárias para a análise

```
%run '/content/drive/MyDrive/AED-Saúde-Covid-19/Tópicos Analisados/documentacao/2_funcoes.ipynb'
```

Para a leitura da base de dados, é usado read_csv da biblioteca pandas

```
df_covid = pd.read_csv('/content/drive/MyDrive/AED-Saúde-Covid-19/Tópicos Analisados/documentacao/dataset/Planilha COVID19 - Aparecida 2021_02112021.csv')
```