UNIVERSIDADE FEDERAL DO PARANÁ

TASSO ELERO CERVI

ANÁLISE DESCRITIVA E EXPLORATÓRIA DOS DADOS DE COVID-19 EM CURITIBA ENTRE MARÇO DE 2020 E ABRIL DE 2024

CURITIBA

2024

TASSO ELERO CERVI

ANÁLISE DESCRITIVA E EXPLORATÓRIA DOS DADOS DE COVID-19 EM CURITIBA ENTRE MARÇO DE 2020 E ABRIL DE 2024

Trabalho de conclusão de curso apresentado como requisito parcial à obtenção do título de Graduado, Curso de Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná.

Orientador: Prof. Dr. José Marcelo Almeida Prado Cestari

CURITIBA
2024

RESUMO

O estudo analisou dados de casos confirmados de COVID-19 em Curitiba entre 11 de março de 2020 e 30 de abril de 2024, com o objetivo de realizar uma análise descritiva e exploratória dos dados buscando por padrões relevantes. Utilizou-se a base de dados do portal 156 da Prefeitura de Curitiba, composta por 638.651 registros. As etapas metodológicas incluíram o pré-tratamento dos dados, verificação de dados faltantes, padronização de colunas, criação de novas colunas para análise, e aplicação de técnicas de análise descritiva e visualização de dados. A matriz de correlação foi calculada utilizando o método de Spearman, após a verificação da distribuição dos dados pelo teste de Shapiro-Wilk. Os resultados mostraram variações significativas na distribuição de casos e óbitos por bairros, uma maior incidência de casos em pessoas de 21 a 50 anos, e uma maior letalidade em indivíduos com mais de 60 anos. A análise temporal indicou que, apesar de ondas de aumento no número de casos após 2021, a mortalidade e os internamentos não acompanharam esse crescimento, sugerindo a efetividade das vacinas na redução da gravidade e letalidade dos casos de COVID-19.

Palavras-chave: COVID-19. Análise de Dados. Saúde. Epidemiologia.

ABSTRACT

The study analyzed data on confirmed cases of COVID-19 in Curitiba between March 11, 2020 and April 30, 2024, with the aim of carrying out a descriptive and exploratory analysis of the data, looking for relevant patterns. The database of the 156 portal of the Curitiba City Hall was used, consisting of 638,651 records. The methodological steps included pre-processing the data, checking for missing data, standardizing columns, creating new columns for analysis, and applying descriptive analysis and data visualization techniques. The correlation matrix was calculated using the Spearman method, after checking the distribution of the data using the Shapiro-Wilk test. The results showed significant variations in the distribution of cases and deaths by neighborhood, a higher incidence of cases in people aged 21 to 50, and a higher lethality in individuals over 60. The temporal analysis indicated that, despite waves of increase in the number of cases after 2021, mortality and hospitalizations did not accompany this growth, suggesting the effectiveness of vaccines in reducing the severity and lethality of COVID-19 cases.

Key-words: COVID-19. Data Analysis. Health. Epidemiology.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

A pandemia de COVID-19, causada pelo coronavírus, trouxe desafios sem precedentes para a saúde pública em todo o mundo. Desde o surgimento dos primeiros casos na China, em dezembro de 2019, a doença se espalhou rapidamente, resultando em milhões de casos confirmados e óbitos globalmente.

> Em circunstâncias sem precedentes como a pandemia da COVID-19, além do surgimento de novas fontes e plataformas de dados, o que se tem verificado é excesso de óbitos, sobrecarga dos sistemas de saúde e atrasos no repasse das informações sobre óbitos. (PAES; FERREIRA; MOURA, 2023, p.2).

Governos e instituições de saúde enfrentaram a tarefa de monitorar a propagação do vírus, implementar medidas de controle e tratamento, e desenvolver vacinas em tempo recorde. No Brasil, a situação foi ainda mais desafiadora, especialmente nas grandes cidades, onde a alta densidade populacional e as disparidades socioeconômicas exacerbaram os problemas.

> A pandemia, causada pela Covid-19, manifestou-se como forma de ameaça ao Estado Democrático de Direito, atingindo a toda a população, porém com efeitos mais drásticos à população de baixa renda que não tem a possibilidade de permanecer em isolamento devido à situação econômica em que se encontram. (RONCATO; DE ANDRADE, 2021, p. 148).

Este trabalho se propõe a realizar uma análise descritiva e exploratória dos dados de COVID-19 fornecidos pelo portal 156 da Prefeitura de Curitiba. A análise se concentrará em explorar as características dos casos confirmados, identificar padrões temporais e espaciais, e investigar possíveis correlações entre variáveis relevantes, utilizando ferramentas de programação em *Python*.

1.1 OBJETIVOS

Analisar os dados de COVID-19 em Curitiba, fornecidos pelo portal 156, utilizando técnicas de programação em *Python* para descrever as características dos casos confirmados e identificar correlações que possam servir de base para futuras pesquisas científicas.

1.1.1 Objetivos Específicos

- Calcular e apresentar estatísticas descritivas das variáveis quantitativas e visualizar a distribuição dos dados através de gráficos.
- Investigar a correlação entre as variáveis da base.
- Interpretar os resultados obtidos e discutir as implicações dos achados para a saúde pública e futuras pesquisas científicas.

1.2 JUSTIFICATIVA

A pandemia trouxe desafios sem precedentes para a saúde pública global. Como salientam Vicentainer, Mattedi e Mello (2020, p. 206), “A necessidade controlar a crise causada pelo Covid-19 (Sars-Cov-2) fez emergir uma demanda, sem precedentes, por informações que possibilitam o monitoramento da propagação do vírus socialmente”. Em Curitiba, como em muitas outras cidades, a gestão e análise dos dados epidemiológicos se tornaram fundamentais para entender a progressão da doença, identificar áreas de maior risco e desenvolver estratégias eficazes de controle e prevenção.

> Os dados epidemiológicos são a principal referência quanto a eficiência das medidas controladoras implementadas, influenciando diretamente não apenas os gestores públicos para a tomada de decisão, mas também a sociedade para a prevenção. (VICENTAINER; MATTEDI; MELLO, 2020, p. 206).

Este estudo tem como objetivo realizar uma análise descritiva e exploratória dos casos confirmados de COVID-19 em Curitiba, entre 11 de março de 2020 e 30 de abril de 2024. A análise busca facilitar a visualização desses dados e encontrar padrões relevantes que possam servir de evidência para pesquisas futuras, seja com os padrões epidemiológicos encontrados ou com a metodologia de tratamento e visualização dos dados aplicada na área da saúde. Técnicas de pré-tratamento e higienização dos dados foram aplicadas, seguidas pela criação de visualizações em formatos de gráficos de barras, utilizando cores distintas (azul para casos confirmados e vermelho para óbitos confirmados) para facilitar a leitura e a identificação de padrões espaciais e temporais relevantes.

Durante as disciplinas cursadas na Gestão da Informação, conceitos e ferramentas para tratamento e análise de dados foram apresentados. Entre essas ferramentas, o Python destacou-se, especialmente na disciplina optativa de Tópicos Especiais em Linguagem de Programação na Gestão da Informação. A aplicação dessas técnicas reflete a interseção entre a gestão da informação e a saúde pública, mostrando a relevância de habilidades analíticas e de programação para a identificação de padrões e realização de análises que buscam auxiliar na tomada de decisão.

A análise de dados epidemiológicos de qualidade é uma ferramenta importante para auxiliar a Medicina Baseada em Evidências (MBE). El Dib (2007, p. 1) afirma que “a MBE utiliza provas científicas existentes e disponíveis no momento, com boa validade interna e externa, para a aplicação de seus resultados na prática clínica”.

> O que significa Medicina Baseada em Evidências? MBE se traduz pela prática da medicina em um contexto em que a experiência clínica é integrada com a capacidade de analisar criticamente e aplicar de forma racional a informação científica de forma a melhorar a qualidade da assistência médica. (LOPES, 2000, p. 285).

Considerando os pontos citados, uma análise detalhada dos dados fornecidos pelo portal 156 da Prefeitura de Curitiba pode agregar à literatura existente, permitindo explorar diversas dimensões dos casos confirmados de COVID-19, como distribuição por bairro, idade, gênero, e temporalidade. Este estudo pode encontrar padrões relevantes, como a relação entre variáveis demográficas e a gravidade da doença, além de destacar diferenças significativas na incidência e mortalidade em diferentes regiões da cidade.

Ao utilizar ferramentas de programação em *Python*, este trabalho não só realiza uma análise dos dados, mas também oferece uma metodologia replicável para futuras pesquisas. Os resultados obtidos podem servir de base para políticas públicas mais informadas e direcionadas, ajudando a melhorar a resposta a pandemias futuras e a alocar recursos de maneira mais eficiente.

Além disso, esta pesquisa também contribui para a literatura existente ao fornecer uma análise local detalhada, que pode ser comparada com estudos de outras regiões e contextos, ampliando a compreensão dos impactos e dinâmicas da COVID-19.

2 LITERATURA PERTINENTE

A COVID-19, doença causada pelo coronavírus (SARS-CoV-2), foi identificada pela primeira vez em *Wuhan* na China, em dezembro de 2019. A rápida disseminação do vírus levou à declaração de pandemia pela Organização Mundial da Saúde (OMS) em março de 2020. Desde então, a pandemia resultou em milhões de diagnósticos e mortes ao redor do mundo, afetando drasticamente a saúde pública, as economias e as sociedades globais. Siqueira et al. (2022, p. 1) observaram que "Territórios e continentes ao redor do mundo foram afetados pela pandemia causada pela COVID-19, com destaque para as Américas, onde ocorreram aproximadamente 39% dos casos e 47% das mortes".

O impacto da COVID-19 foi sentido em diversos níveis, desde a sobrecarga dos sistemas de saúde até os indicadores de qualidade de vida dos pacientes. Silva Filho et al. (2023, p. 2) destacam que "A mortalidade provocada pela pandemia da COVID-19 tem produzido impactos nos indicadores de Anos Potenciais de Vida Perdidos (APVP) e Expectativa de Vida (EV) em nível mundial". No Brasil, o cenário foi especialmente desafiador devido à falta de medidas governamentais rigorosas e ao negacionismo, gerando incertezas e dualidades na população sobre a obediência às medidas restritivas (SIQUEIRA et al., 2022). Essas decisões políticas divergiram do restante dos países.

> No entanto, diferentemente da maioria dos governos subnacionais que logo no início da pandemia fecharam escolas e comércios, cancelaram eventos, entre outras medidas não farmacológicas para reduzir a propagação de casos de Covid-19, Bolsonaro e seus apoiadores – incluindo parlamentares, familiares, ministros, empresários etc. – optaram por não seguir as diretrizes da Organização Mundial da Saúde (OMS) e as evidências científicas nesse sentido e desdenharam da pandemia, com a intenção de evitar consequências danosas à economia do País, que já cambaleava. (BRANDÃO; MENDONÇA; SOUSA, 2023, p. 59).

2.1 IMPORTÂNCIA DA ANÁLISE DE DADOS NA SAÚDE PÚBLICA

A análise de dados é um conjunto de técnicas e ferramentas multidisciplinares que podem ser aplicadas na saúde pública, permitindo a identificação de padrões, tendências e fatores de risco que são essenciais para a formulação de políticas e intervenções.

> Os dados em saúde são extremamente importantes para a tomada de decisões e para a melhoria da qualidade dos cuidados de saúde. Eles podem incluir informações sobre pacientes, tratamentos, resultados clínicos, custos de saúde e outros aspectos relacionados à saúde. (GASPAR et. al, 2023, p. 8).

A análise de dados em saúde pública envolve várias etapas, incluindo a coleta, limpeza, análise e interpretação dos dados. Métodos estatísticos e de aprendizado de máquina são frequentemente utilizados para extrair informações significativas dos dados. Gaspar et al. (2023, p. 8) destacam que “pode ajudar a identificar padrões, prever resultados e é fundamental para a melhoria da qualidade e eficiência dos serviços de saúde. No entanto, é importante garantir que esses dados sejam de alta qualidade, confiáveis e seguros”. Paes, Ferreira e Moura (2023, p. 2) enfatizam a importância de dados de alta qualidade sobre causas de óbitos, afirmando que "É sabido que dados de alta qualidade sobre causas de óbitos são uma fonte-chave de evidências úteis para a implementação de certas políticas e para tomadas de decisão que tenham o objetivo de melhorar a saúde da população".

> Considerando as diferentes políticas de testagem entre os estados, países e municípios, percebe-se o desafio de se aproximar da real situação epidemiológica e compará-las, já que a maioria dos locais realiza os testes para o Coronavírus apenas em pacientes sintomáticos, principalmente nos que apresentam sintomas mais graves. Desta forma, estima-se que ocorra subnotificação dos casos totais na ordem de sete a oito vezes. (FREDRICH, 2020, p. 64).

Um desafio encontrado na coleta de dados em saúde é a subnotificação, que pode mascarar a realidade dos dados e prejudicar a análise.

> Com a utilização de nosso método, concluímos que a taxa estimada de notificações de casos confirmados de COVID-19 no Brasil foi de cerca de 9,2% (IC95%8,8% - 9,5%). Assim, o número real de casos no Brasil foi cerca de 11 vezes mais alto do que o número oficial de casos notificados. (PRADO et. al, 2020, p. 225).

2.2 MÉTODOS DE ANÁLISE

A estatística descritiva é formada pelos procedimentos e técnicas da estatística que são utilizados para recolher, organizar, sintetizar e descrever os dados (SANTOS, 2007). O uso da estatística descritiva nas variáveis quantitativas é um passo importante na análise exploratória dos dados, possibilitando a visualização de padrões e descoberta de informações dentro de conjuntos de dados, que servem como base para análises estatísticas posteriores.

> A análise exploratória de dados nos fornece um extenso repertório de métodos para um estudo detalhado dos dados, antes de adaptá-los. Nessa abordagem, a finalidade é obter dos dados a maior quantidade possível de informação, que indique modelos plausíveis a serem utilizados numa fase posterior, a análise confirmatória de dados ou inferência estatística. (MEDRI, 2011).

Outra abordagem estatística aplicada neste estudo é o cálculo de correlação entre as variáveis, visando identificar a existência de correlações relevantes na base de dados. Figueiredo Filho (2009, p. 118) explica que “Em termos estatísticos, duas variáveis se associam quando elas guardam semelhanças na distribuição dos seus escores. Mais precisamente, elas podem se associar a partir da distribuição das frequências ou pelo compartilhamento de variância”. Para escolher o método adequado para o cálculo de correlação, foi necessário verificar se os dados seguem uma distribuição normal.

> A curva de distribuição mais utilizada é a curva normal ou curva de Gauss, cujas principais características são as seguintes: a área entre a curva e o eixo horizontal é igual a 100%; a curva é unimodal e simétrica em redor do ponto médio e possui a forma de sino. (SOUSA, 2019).

Para identificação da distribuição apresentada nos dados da base foi utilizado o teste de Shapiro-Wilk. Leotti, Birck e Riboldi (2005, p. 2) afirmam que “o teste de Shapiro-Wilk baseia-se nos valores amostrais ordenados elevados ao quadrado e tem sido o teste de normalidade preferido por mostrar ser mais poderoso que diversos testes alternativos”.

Na análise de dados epidemiológicos, métodos estatísticos robustos são necessários para lidar com outliers e distribuições não normais. Mukaka (2012) recomenda o uso do método de Spearman para calcular correlações quando uma ou mais variáveis não seguem uma distribuição normal, afirmando que este método é mais robusto a outliers do que o método de Pearson.

> O coeficiente de Spearman é calculado como o coeficiente de correlação entre as classificações das variáveis em ordem. Ele varia de -1 a 1, onde -1 indica uma correlação perfeitamente negativa, 1 indica uma correlação perfeitamente positiva e 0 indica ausência de correlação. (GASPAR et. al, 2023, p. 101).

2.3 IMPACTO SOCIOECONÔMICO E SEGREGAÇÃO ESPACIAL

Estudos indicam uma maior letalidade da COVID-19 em grupos de risco, como idosos e pessoas com comorbidades (DE ARAUJO FILHO, 2023). Além disso, a segregação socioespacial tem mostrado impactos significativos na disseminação da

COVID-19. Reinaldo Filho, Celestino e Tomazeli (2023) demonstram que regiões mais segregadas são mais prejudicadas pela pandemia.

3 METODOLOGIA

Para este estudo, utilizou-se a base de dados “Casos de COVID-19 em Curitiba”, disponível no portal de dados abertos da Prefeitura de Curitiba. Esta base de dados é composta por 638.651 registros, cada um representando um caso confirmado de COVID-19 em Curitiba entre 11 de março de 2020 e 30 de abril de 2024. A estrutura da tabela é descrita a seguir.

‘DATA INCLUSAO/NOTIFICACAO’ - Data em que o caso de COVID-19 foi incluído na base de dados e realizada a notificação. Os valores estão no formato DD/MM/YYYY (dia/mês/ano).

‘CLASSIFICACAO FINAL’ - Indica se o caso de COVID-19 foi confirmado. Os valores possíveis são: Confirmado.

‘IDADE (anos)’ - Idade do paciente no momento da notificação do caso de COVID-19. Valores numéricos representando a idade em anos completos.

‘SEXO’ – Gênero do paciente. Os valores possíveis são: F (Feminino) e M (Masculino).

‘BAIRRO’ - Bairro de residência do paciente no momento da notificação do caso de COVID-19. Os valores correspondem aos nomes dos bairros de Curitiba.

‘DISTRITO RESIDENCIA’ - Distrito sanitário de residência do paciente no momento da notificação do caso de COVID-19. Os valores possíveis são: DSBN (Distrito Sanitário Bairro Novo), DSBQ (Distrito Sanitário Boqueirão), DSBV (Distrito Sanitário Boa Vista), DSCIC (Distrito Sanitário Cidade Industrial de Curitiba), DSCJ (Distrito Sanitário Cajuru), DSMZ (Distrito Sanitário Matriz), DSPN (Distrito Sanitário Pinheirinho), DSPR (Distrito Sanitário Portão), DSSF (Distrito Sanitário Santa Felicidade) e DSTQ (Distrito Sanitário Tatuquara).

‘INTERNADO (SIM/NAO)’ - Indica se o paciente foi internado. Os valores possíveis são: SIM e NAO.

‘DATA COLETA EXAME’ - Data de coleta do exame do paciente. Os valores estão no formato DD/MM/YYYY (dia/mês/ano).

‘DATA OBITO’ - Data de óbito do paciente. Os valores estão no formato DD/MM/YYYY (dia/mês/ano).

‘ENCERRAMENTO’ – Conclusão do caso de COVID-19. Os valores possíveis são: ATIVO, RECUPERADO e OBITO CONF.

3.1 TRATAMENTO DE DADOS

O pré-tratamento e a higienização dos dados são etapas cruciais para assegurar a qualidade e a confiabilidade das análises realizadas. As seguintes correções e verificações foram realizadas para preparar a base de dados para análise.

3.1.1 Correções manuais

Foi utilizado o *LibreOffice Calc* para correções manuais, como remoção de caracteres acentuados corrompidos nos títulos das colunas e correção de nomes de bairros e valores categóricos.

3.1.2 Verificação de dados faltantes

Identificou-se que as colunas ‘BAIRRO’ e ‘DISTRITO RESIDENCIA’ possuíam 12.322 e 12.324 valores ausentes, respectivamente. Como esses valores representam uma pequena fração do total de registros decidiu-se não remover essas linhas para manter a integridade e representatividade da análise da base de dados.

3.1.3 Conversão de tipos e limpeza de dados

As colunas de datas (‘DATA INCLUSAO/NOTIFICACAO’, ‘DATA COLETA EXAME’ e ‘DATA OBITO’) foram convertidas para o tipo *datetime*, permitindo uma manipulação adequada dos dados e facilitando a realização de análises temporais.

A coluna ‘SEXO’ foi padronizada removendo espaços em branco e convertendo todos os valores para maiúsculas ('M' para masculino e 'F' para feminino). Esta padronização foi realizada para garantir a consistência dos dados e evitar discrepâncias na análise.

Foi criada uma nova coluna ‘FAIXA ETARIA’ para facilitar a análise por idade, separando os pacientes em intervalos de 10 anos, iniciando com a faixa de 0-10 anos e finalizando com a faixa de 90-100 anos. Esses intervalos foram definidos para permitir a análise da distribuição etária dos casos e identificar grupos etários com maior incidência e gravidade de COVID-19.

3.2 MÉTODOS ESTATÍSTICOS E TÉCNICAS DE VISUALIZAÇÃO DE DADOS

Para a análise descritiva dos dados de COVID-19 em Curitiba, foram aplicados os seguintes métodos estatísticos e técnicas de visualização de dados.

3.2.1 Análise estatística descritiva

Utilizando a biblioteca *Pandas* calculou-se estatísticas descritivas como média, mediana, moda, desvio padrão e percentis para variáveis numéricas.

3.2.2 Técnicas de visualização de dados

Com apoio da biblioteca *Matplotlib* foram criados gráficos informativos e esteticamente agradáveis. Os gráficos de barras foram utilizados para visualizar a distribuição de casos por bairros, distritos, gênero e faixas etárias. Enquanto os gráficos de barras empilhadas tiveram uso para comparar casos e óbitos confirmados em diferentes categorias, como faixa etária e status de internamento.

3.2.3 Cálculo de correlações

Foi utilizada a biblioteca *Pandas* para calcular a matriz de correlação visando entender a relação entre variáveis como idade, gênero, status de internamento e encerramento dos casos. Para realização da análise de correlação entre as variáveis, foi selecionada uma amostra aleatória de 5000 casos da base de dados e realizado o teste de Shapiro-Wilk para verificar a distribuição dos dados. A escolha de uma amostra foi baseada na recomendação de Gaspar et al. (2023, p. 64), que afirmam que o teste de Shapiro-Wilk “é um dos testes mais precisos para verificar a normalidade, mas é menos eficiente para grandes conjuntos de dados”.

4 RESULTADOS

A base de dados analisada conta com 638.651 registros, sendo estes separados em 629.687 pacientes recuperados, 8.907 pacientes com óbito confirmado e 57 pacientes com o caso ativo. Todos os códigos utilizados para realização desse estudo estão presentes no apêndice 1.

4.1 DISTRIBUIÇÃO DE CASOS POR BAIRRO

A análise da distribuição de casos por bairro mostrou uma variação significativa entre diferentes regiões de Curitiba. Os bairros com maior número de casos confirmados foram Cidade Industrial de Curitiba, Sítio Cercado e Cajuru. Esta distribuição pode refletir fatores como densidade populacional e condições socioeconômicas específicas de cada bairro.

FIGURA 1: DISTRIBUIÇÃO DE CASOS DE COVID-19 POR BAIRRO EM CURITIBA

FONTE: O autor (2024)

4.2 DISTRIBUIÇÃO DE ÓBITOS POR BAIRRO

A FIGURA 2 apresenta a contagem de óbitos confirmados por bairro. Bairros como Cidade Industrial de Curitiba, Sítio Cercado e Cajuru também apresentaram maior número de óbitos. A correlação entre a alta incidência de casos e o número de óbitos foi analisada, utilizando o método de Spearman o cálculo resultou em um coeficiente de correlação de 0.9731, indicando uma correlação muito forte. Este resultado sugere que bairros com mais casos tendem a ter proporcionalmente mais óbitos, o que reforça a importância de medidas de controle em áreas com alta incidência.

FIGURA 2: DISTRIBUIÇÃO DE CASOS COM ÓBITO CONFIRMADO POR BAIRRO EM CURITIBA

FONTE: O autor (2024).

TABELA 1: NÚMERO DE CASOS E ÓBITOS POR BAIRRO

Bairro	Nº de Casos	Nº de Óbitos	% de Óbitos
CIDADE INDUSTRIAL DE CURITIBA	62684	852	1,36%
SITIO CERCADO	39708	566	1,43%

CAJURU	32493	555	1,71%
UBERABA	25620	341	1,33%
BOQUEIRAO	25565	416	1,63%
TATUQUARA	20199	257	1,27%
XAXIM	20039	279	1,39%
PINHEIRINHO	18944	284	1,50%
ALTO BOQUEIRAO	17505	242	1,38%
BAIRRO ALTO	17303	255	1,47%
NOVO MUNDO	17172	277	1,61%
CAPAO RASO	14787	209	1,41%
SANTA CANDIDA	14764	201	1,36%
CENTRO	13993	218	1,56%
PORTAO	13964	170	1,22%
CAMPO DE SANTANA	13747	147	1,07%
AGUA VERDE	13681	208	1,52%
SANTA FELICIDADE	12831	165	1,29%
CAMPO COMPRIDO	10798	120	1,11%
PILARZINHO	10038	143	1,42%
BIGORRILHO	9840	118	1,20%
FAZENDINHA	9507	129	1,36%
BOA VISTA	9308	151	1,62%
SAO BRAZ	8640	140	1,62%
BACACHERI	8079	134	1,66%
UMBARA	7069	90	1,27%
BARREIRINHA	6569	122	1,86%
CAPAO DA IMBUIA	6388	103	1,61%
REBOUCAS	6284	83	1,32%
HAUER	6124	93	1,52%
ATUBA	5755	74	1,29%
GUAIRA	5711	70	1,23%
GANCHINHO	5248	77	1,47%

SANTA QUITERIA	5121	54	1,05%
BATEL	4847	97	2,00%
VILA IZABEL	4683	62	1,32%
TINGUI	4633	74	1,60%
MOSSUNGUE	4592	39	0,85%
ALTO DA RUA XV	4566	46	1,01%
CRISTO REI	4425	50	1,13%
MERCES	4355	50	1,15%
ABRANCHES	4353	58	1,33%
JARDIM DAS AMERICAS	4014	74	1,84%
BUTIATUVINHA	4010	58	1,45%
CACHOEIRA	3882	35	0,90%
VISTA ALEGRE	3820	45	1,18%
CABRAL	3621	46	1,27%
LINDOIA	3621	47	1,30%
GUABIROTUBA	3459	60	1,73%
JUVEVE	3310	67	2,02%
AHU	3270	44	1,35%
FANNY	3068	44	1,43%
PAROLIN	2947	50	1,70%
ORLEANS	2483	27	1,09%
SANTO INACIO	2470	28	1,13%
AUGUSTA	2333	32	1,37%
ALTO DA GLORIA	2299	39	1,70%
SAO FRANCISCO	2283	42	1,84%
CAMPINA DO SIQUEIRA	2129	33	1,55%
SEMINARIO	2075	32	1,54%
PRADO VELHO	1876	20	1,07%
JARDIM BOTANICO	1836	46	2,51%
BOM RETIRO	1825	22	1,21%
SAO LOURENCO	1796	28	1,56%
JARDIM SOCIAL	1795	20	1,11%
TARUMA	1794	30	1,67%

CENTRO CIVICO	1690	33	1,95%
CAXIMBA	1458	11	0,75%
SAO JOAO	1213	14	1,15%
HUGO LANGE	1019	20	1,96%
TABOAO	956	10	1,05%
CASCATINHA	902	12	1,33%
SAO MIGUEL	813	10	1,23%
LAMENHA PEQUENA	239	2	0,84%
RIVIERA	91	1	1,10%

FONTE: O autor (2024).

4.3 DISTRIBUIÇÃO DE CASOS E ÓBITOS POR DISTRITO SANITÁRIO DE RESIDÊNCIA

Esta análise apresenta o número de casos e de óbitos confirmados por distrito sanitário de residência, apresentando uma separação urbana diferente da separação por bairros, essas duas variáveis em conjunto podem auxiliar estudos de segregação socioespacial, por exemplo.

FIGURA 3: DISTRIBUIÇÃO DE CASOS E DE ÓBITOS POR DISTRITO SANITÁRIO DE RESIDÊNCIA

FONTE: O autor (2024).

4.4 DISTRIBUIÇÃO DE CASOS E ÓBITOS POR GÊNERO

A análise por gênero revela que as mulheres foram mais afetadas em termos de número de casos confirmados, totalizando 366.039 casos, enquanto os homens representam 272.612 casos. No entanto, a distribuição de óbitos por gênero mostrou que os homens apresentaram uma taxa de letalidade maior, com 4.994 mortes, correspondendo a 1,83% dos casos confirmados, em comparação com 3.913 óbitos entre as mulheres, representando 1,06% dos casos confirmados. Esses dados sugerem a hipótese de que existem diferenças comportamentais ou biológicas que podem influenciar a gravidade da COVID-19 em indivíduos do gênero masculino, sendo uma área de interesse para pesquisas futuras.

FIGURA 4: DISTRIBUIÇÃO DE CASOS E DE ÓBITOS POR GÊNERO

FONTE: O autor (2024).

4.5 DISTRIBUIÇÃO DE CASOS E ÓBITOS POR FAIXA ETÁRIA

A distribuição de casos por faixa etária indicou que pessoas entre 21 e 50 anos foram as mais afetadas. No entanto, a maior incidência de óbitos ocorreu entre indivíduos com mais de 60 anos. Esses dados de Curitiba corroboram com as descobertas de estudos que destacam a vulnerabilidade dos grupos etários mais velhos para a COVID-19.

FIGURA 5: DISTRIBUIÇÃO DE CASOS E DE ÓBITOS POR FAIXA ETÁRIA

FONTE: O autor (2024).

4.6 EVOLUÇÃO TEMPORAL DOS CASOS E ÓBITOS

A análise temporal dos casos e óbitos mostra a evolução da pandemia em Curitiba ao longo do período estudado. Observa-se um aumento significativo no número de casos e óbitos durante os picos da pandemia, refletindo as ondas epidêmicas locais. Após 2022, continuaram ocorrendo ondas de aumento no número de casos, mas a mortalidade não acompanhou esse aumento com a mesma intensidade, sugerindo a efetividade das vacinas. A redução na mortalidade, apesar do aumento no número de casos, indica que as campanhas de vacinação foram eficazes em diminuir a gravidade dos casos e a taxa de letalidade. Esses dados corroboram com a literatura existente, apontando uma possível evidência da efetividade das vacinas em reduzir a gravidade e a letalidade dos casos de COVID-19.

FIGURA 6: EVOLUÇÃO DOS CASOS POR MÊS AO LONGO DO PERÍODO

FONTE: O autor (2024).

FIGURA 7: EVOLUÇÃO DOS ÓBITOS POR MÊS AO LONGO DO PERÍODO

FONTE: O autor (2024).

4.7 EVOLUÇÃO DE INTERNAMENTOS AO LONGO DO PERÍODO

A análise temporal do número de internamentos por mês em Curitiba também reflete os efeitos positivos da vacinação na redução da gravidade dos casos e da letalidade. Assim como observado no gráfico de óbitos por mês, houve uma queda

acentuada no número de internamentos nos últimos meses de 2021, que se manteve em níveis baixos em comparação com os anos de 2020 e 2021. Esses gráficos demonstram que, embora continuemos a experimentar ondas de aumento no número de casos após o final de 2021, o impacto em termos de óbitos e internamentos não seguiu o mesmo padrão de aumento observado em 2020 e 2021. Notavelmente, o mês de janeiro de 2022 registrou o maior número de casos, mas os óbitos e internamentos não acompanharam esse aumento com a mesma intensidade que os meses anteriores, destacando a efetividade das vacinas em prevenir casos graves e internações.

FIGURA 8: EVOLUÇÃO DOS INTERNAMENTOS POR MÊS AO LONGO DO PERÍODO

FONTE: O autor (2024).

4.8 INTERNAÇÕES E ÓBITOS CONFIRMADOS

A comparação entre pacientes internados e não internados mostrou que a taxa de óbitos é consideravelmente maior entre os pacientes que necessitaram de internação, evidenciando a gravidade dos casos que requerem cuidados hospitalares. Esses dados reforçam a necessidade de instalações hospitalares adequadas e equipadas para atender a demanda.

FIGURA 9: COMPARAÇÃO DE ÓBITOS CONFIRMADOS ENTRE PACIENTES INTERNADOS E NÃO INTERNADOS.

FONTE: O autor (2024).

4.9 ANÁLISE DE CORRELAÇÃO ENTRE VARIÁVEIS

O teste de Shapiro-Wilk indicou que as variáveis BAIRRO, IDADE (anos) e DISTRITO RESIDÊNCIA não seguem distribuições normais. Com base nesses resultados, foi escolhido o método de Spearman para calcular a correlação entre as variáveis.

FIGURA 10: ANÁLISE DE CORRELAÇÃO ENTRE VARIÁVEIS

	IDADE (anos)	SEXO	BAIRRO	DISTRITO RESIDENCIA	INTERNADO (SIM/NAO)	ENCERRAMENTO
IDADE (anos)	1.000000	-0.045227	-0.027345	0.012391	0.199074	0.148483
SEXO	-0.045227	1.000000	-0.000971	0.004021	0.055891	0.032177
BAIRRO	-0.027345	-0.000971	1.000000	0.109327	0.005156	0.000331
DISTRITO RESIDENCIA	0.012391	0.004021	0.109327	1.000000	-0.000027	-0.000778
INTERNADO (SIM/NAO)	0.199074	0.055891	0.005156	-0.000027	1.000000	0.499622
ENCERRAMENTO	0.148483	0.032177	0.000331	-0.000778	0.499622	1.000000

FONTE: O autor (2024).

4.9.1 Correlação entre idade e variáveis

“IDADE (anos)” e “INTERNADO (SIM/NAO)” (0.199074) - Há uma baixa correlação positiva entre a idade dos pacientes e a necessidade de internação.

“IDADE (anos)” e “ENCERRAMENTO” (0.148483) - Existe também uma pequena correlação positiva entre a idade dos pacientes e o encerramento dos casos, sugerindo que a idade está relacionada com o resultado final dos casos, como a recuperação ou o óbito.

4.9.2 Correlação entre internamento e encerramento

“INTERNADO (SIM/NAO)” e “ENCERRAMENTO” (0.499622) - Esta é a correlação mais forte observada, indicando que a necessidade de internação tem uma relação moderada com o encerramento dos casos.

4.9.3 Correlação entre gênero e variáveis

“SEXO” e “INTERNADO (SIM/NAO)” (0.055891) e “SEXO” e “ENCERRAMENTO” (0.032177) - As correlações observadas entre o gênero dos pacientes e a necessidade de internação ou o encerramento dos casos são quase nulas, sugerindo que o gênero não está relacionado com a necessidade de internação ou com o óbito.

4.9.4 Observações gerais

A análise de correlação identifica relações entre variáveis que podem ser exploradas em estudos futuros. As correlações observadas, especialmente entre idade, internamento e encerramento, sugerem a existência de uma relação entre a idade do paciente e a gravidade do caso. Embora a correlação não implique causalidade, essas análises podem ser cruciais para direcionar políticas públicas e alocação de recursos de saúde, especialmente em estratégias de prevenção e tratamento para grupos mais vulneráveis.

5. CONSIDERAÇÕES FINAIS

Este estudo contribui para a literatura existente ao fornecer uma análise detalhada e local dos dados de COVID-19 em Curitiba, que pode ser comparada com estudos de outras regiões e contextos. As descobertas destacam a importância das análises de dados para a formulação de políticas públicas informadas, especialmente em estratégias de prevenção e tratamento para grupos mais vulneráveis.

Além disso, o uso de ferramentas de programação em *Python* demonstrou ser eficaz na realização de análises robustas e replicáveis, oferecendo uma metodologia que pode ser aplicada em futuras pesquisas epidemiológicas. Os resultados obtidos podem servir de base para políticas públicas mais informadas e direcionadas, ajudando a melhorar a resposta a pandemias futuras e a alocar recursos de maneira mais eficiente.

Os dados tratados e transformados em visualizações gráficas, facilitando a interpretação e utilização destes também foi alcançada, oferecendo evidências locais dos padrões de casos e óbitos por COVID-19 em Curitiba, agregando a literatura existente e auxiliando a medicina baseada em evidências.

Finalmente, este trabalho reforça a importância da coleta e análise contínua de dados epidemiológicos para responder de forma eficaz a crises de saúde pública e para melhorar a alocação de recursos e a formulação de políticas.

5.1 PRINCIPAIS ACHADOS

Os principais achados deste estudo podem ser separados entre os padrões espacias e temporais dos casos e óbitos, o impacto de variáveis demográficas e a correlação entre variáveis.

5.1.1 Distribuição espacial e temporal dos casos e óbitos

A análise revelou uma variação significativa na distribuição dos casos confirmados e óbitos por bairros de Curitiba, com bairros como Cidade Industrial de Curitiba, Sítio Cercado e Cajuru apresentando os maiores números de casos e óbitos.

A evolução temporal dos casos e óbitos mostrou picos distintos durante a pandemia, com uma diminuição no número de internamentos e na taxa de letalidade a partir de 2022, sugerindo a efetividade das campanhas de vacinação.

5.1.2 Impacto de variáveis demográficas

A análise por faixa etária indicou que indivíduos entre 21 e 50 anos foram os mais afetados em termos de casos confirmados, enquanto a maior incidência de óbitos ocorreu entre aqueles com mais de 60 anos, corroborando a maior vulnerabilidade dos grupos etários mais velhos.

Diferenças de gênero também foram observadas, com mulheres apresentando mais casos confirmados, mas homens exibindo uma taxa de letalidade mais elevada, sugerindo possíveis diferenças biológicas ou comportamentais na gravidade da doença. Apesar disso, a análise indicou uma baixa correlação entre o gênero e a necessidade de internação ou encerramento dos casos sugerindo que este não é um fator determinante significativo para essas variáveis

5.1.3 Correlação entre variáveis

A análise mostrou uma pequena correlação entre a idade dos pacientes e a necessidade de internação, bem como entre a idade e o encerramento dos casos. A correlação mais forte observada foi entre a necessidade de internação e o encerramento dos casos, indicando que a internação possui relação com a gravidade dos encerramentos.

5.2 DIREÇÕES PARA PESQUISAS FUTURAS

A análise realizada neste estudo abre caminho para diversas direções de pesquisa futura como, por exemplo, análises mais profundas das variáveis sociais e econômicas, investigando como fatores como renda, acesso a serviços de saúde e condições de moradia afetam a incidência e a gravidade dos casos de COVID-19. O presente estudo também pode auxiliar estudos longitudinais que acompanhem a evolução dos casos e a eficácia das vacinas ao longo do tempo.

REFERÊNCIAS

BRANDÃO, Celmário Castro; MENDONÇA, Ana Valéria Machado; SOUSA, Maria Fátima de. O Ministério da Saúde e a gestão do enfrentamento à pandemia de Covid-19 no Brasil. **Saúde em Debate,** v. 47, p. 58-75, 2023. Disponível em: <https://www.scielosp.org/pdf/sdeb/2023.v47n137/58-75/pt>. Acesso em: 10 jun. 2024.

DE ARAÚJO FILHO, Francisco José et al. Fatores que influenciam na adesão de idosos a vacina contra covid-19: revisão de escopo. **Nursing (São Paulo),** v. 26, n. 304, p. 9926-9931, 2023. Disponível em: <https://www.revistanursing.com.br/index.php/revistanursing/article/view/3130>. Acesso em: 16 jun. 2024.

EL DIB, Regina Paolucci. Como praticar a medicina baseada em evidências. **Jornal Vascular Brasileiro**, v. 6, p. 1-4, 2007. Disponível em <https://www.scielo.br/j/jvb/a/Dhy8TqBdZJgGcL7SwCmBK6r/?format=pdf&lang=pt>. Acesso em: 06 jun 2024.

FIGUEIREDO FILHO, Dalson Britto; SILVA JÚNIOR, José Alexandre. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). **Revista Política Hoje**, v. 18, n. 1, p. 115-146, 2009. Disponível em: <https://dirin.s3.amazonaws.com/drive_materias/1666287394.pdf>. Acesso em: 08 jun. 2024.

FREDRICH, Vanessa Cristine Ribeiro et al. Perfil de óbitos por Covid-19 no Estado do Paraná no início da pandemia: estudo transversal. **Revista de saúde pública do paraná**, v. 3, n. Supl., 2020. Disponível em: <http://revista.escoladesaude.pr.gov.br/index.php/rspp/article/view/409>. Acesso em: 25 abr. 2024.

GASPAR, Juliano de Souza et al. Introdução à análise de dados em saúde com Python. In: **Introdução à análise de dados em saúde com Python**. 2023. p. 130-130. Disponível em: <https://pesquisa.bvsalud.org/portal/resource/pt/biblio-1437637>. Acesso em: 20 abr. 2024.

LEOTTI, Vanessa Bielefeldt; BIRCK, Alan Rodrigues; RIBOLDI, João. Comparação dos Testes de Aderência à Normalidade Kolmogorov-smirnov, Anderson-Darling, Cramer–Von Mises e Shapiro-Wilk por Simulação. **Anais do 11º Simpósio de Estatística Aplicada à Experimentação Agronômica**, 2005. Disponível em: <https://www.inf.ufsc.br/~vera.carmo/Testes_de_Hipoteses/Testes_aderencia.pdf>. Acesso em: 06 jun. 2024.

LOPES, Anibal A. Medicina Baseada em Evidências: a arte de aplicar o conhecimento científico na prática clínica. **Revista da Associação Médica Brasileira**, v. 46, p. 285-288, 2000. Disponível em: <https://www.scielo.br/j/ramb/a/BBkKVMDFTg9BnkzdPqXKkGH/?format=pdf&lang=pt>. Acesso em: 06 jun. 2024.

MACEDO, Yuri Miguel; ORNELLAS, Joaquim Lemos; DO BOMFIM, Helder Freitas. COVID-19 nas favelas e periferias brasileiras. **Boletim de Conjuntura (BOCA)**, v. 2, n. 4, p. 50-54, 2020. Disponível em: <https://revista.ioles.com.br/boca/index.php/revista/article/view/149>. Acesso em: 16 jun. 2024.

MEDRI, Waldir. Análise exploratória de dados. **Londrina: Universidade Estadual de Londrina**, v. 8, n. 8, p. 151-170, 2011. Disponível em: <https://www.uel.br/pos/estatisticaquantitativa/textos_didaticos/especializacao_estatistica.pdf>. Acesso em: 06 jun. 2024.

MUKAKA, Mavuto M. A guide to appropriate use of correlation coefficient in medical research. **Malawi medical journal**, v. 24, n. 3, p. 69-71, 2012. Disponível em: <https://www.ajol.info/index.php/mmj/article/view/81576>. Acesso em: 19 jun. 2024.

PAES, Neir Antunes; FERREIRA, Assel Muratovna Shigayeva; MOURA, Lucas de Almeida. Proposta metodológica para avaliação de registros de óbitos por COVID-19. **Cadernos de Saúde Pública**, v. 39, p. e00096722, 2023. Disponível em: <https://www.scielo.br/j/csp/a/V6KqmMYGb6zvdQWRNKqpcFP>. Acesso em: 30 abr. 2024.

PRADO, Marcelo Freitas do et al. Análise da subnotificação de COVID-19 no Brasil. **Revista Brasileira de Terapia Intensiva**, v. 32, p. 224-228, 2020. Disponível em: <https://www.scielo.br/j/rbti/a/XHwNB9R4xhLTqpLxqXJ6dMx>. Acesso em: 25 abr. 2024.

REINALDO FILHO, A. B.; DA SILVA CELESTINO, Sáhira Michele; TOMAZELI, Maristella Rossi. A relação entre segregação socioespacial e os dados de positivos e óbitos por covid-19 na cidade de Passos (MG). **Oikos: Família e Sociedade em Debate**, v. 34, n. 3, 2023. Disponível em: <https://periodicos.ufv.br/oikos/article/view/15472>. Acesso em: 30 abr. 2024.

RONCATO, Carina Lamas; DE ANDRADE, Thaïs Savedra. Relação entre taxa de mortalidade da covid-19 e a desigualdade tributária no brasil. **Caderno PAIC**, v. 22, n. 1, p. 131-152, 2021. Disponível em: <https://cadernopaic.fae.emnuvens.com.br/cadernopaic/article/view/454>. Acesso em: 30 abr. 2024.

SANTOS, Carla. Estatística descritiva. **Manual de auto-aprendizagem**, v. 2, 2007. Disponível em: <https://static.fnac-static.com/multimedia/PT/pdf/9789726189688.pdf>. Acesso em: 06 jun. 2024.

SILVA FILHO, Aloisio Machado da et al. Anos Potenciais de Vida Perdidos devido à COVID-19, segundo a raça/cor e gênero, no Brasil, entre 2020 e 2021. **Ciência & Saúde Coletiva**, v. 29, p. e04702023, 2024. Disponível em: <https://www.scielo.br/j/csc/a/y3qxFnQK6JB5wsLvThqKTJm/>. Acesso em: 30 abr. 2024.

SIQUEIRA, Camila Alves dos Santos et al. COVID-19 no Brasil: tendências, desafios e perspectivas após 18 meses de pandemia. **Revista Panamericana de Salud Pública**, v. 46, p. e74, 2023. Disponível em: <https://www.scielosp.org/article/rpsp/2022.v46/e74/>. Acesso em: 20 abr. 2024.

SOUSA, Áurea. O papel da distribuição normal na Estatística. **Correio dos Açores**, p. 14-14, 2019. Disponível em: <https://repositorio.uac.pt/bitstream/10400.3/5363/1/Sousa_10%2520out%25202019.pdf>. Acesso em: 07 jun. 2024

VICENTAINER, Denis; MATTEDI, Marcos; MELLO, Bruno. Aplicação das Bibliotecas Python para tratamento de dados em tempo real: A análise dos dados de isolamento social em Santa Catarina. **Metodologias e Aprendizado**, v. 3, p. 206-217, 2020. Disponível em: <https://publicacoes.ifc.edu.br/index.php/metapre/article/view/1392>. Acesso em: 18 abr. 2024.

APÊNDICE 1 – CÓDIGOS DE PROGRAMAÇÃO PYTHON UTILIZADOS

Inicialmente, foi realizada a importação das bilbiotecas que serão usadas no estudo, usando a biblioteca *Pandas* foi importada a base de dados que está em formato csv e foi removida a coluna ‘CLASSIFICACAO FINAL’ que não terá uso. Também verificou-se quantos dados nulos haviam em cada coluna.

Foram realizadas duas verificações na base de dados, a primeira contando quantas vezes a coluna BAIRRO e a coluna DISTRITO RESIDENCIA aparecem nulas na mesma linha e a segunda contando quantas vezes a coluna ENCERRAMENTO aponta óbito confirmado mas a coluna DATA OBITO está em branco.

Durante o pré-tratamento dos dados as colunas de data foram convertidas para o tipo *datetime*, a coluna ‘SEXO’ teve seus valores padronizados, removendo espaços em branco e capitalizando os valores. Uma coluna chamada ‘FAIXA ETARIA’ separando os pacientes por intervalos de idade categorizados foi adicionada à base de dados.

Utilizando as funções da biblioteca Matplotlib foram criados dois gráficos de barras, o primeiro representando a contagem de casos por bairro (seção 4.1) e o segundo trazendo a contagem de óbitos confirmados por bairro (seção 4.2).

Para o cálculo da correlação entre o número de casos e o número de óbitos foi criado um *DataFrame* a partir de um dicionário onde a coluna ‘Casos’ é preenchida com as contagens de casos por bairro e a coluna ‘Obitos’ é preenchida com as contagens de óbitos confirmados por bairro. Os valores faltantes foram preenchidos com o valor ‘0’ garantindo que todos os bairros tenham um valor válido em ambas as colunas. Foi calculada a correlação de Spearman (seção 4.2) entre o número de casos e o número de óbitos por bairro, considerando que os dados não seguem uma distribuição normal.

Reajustando as configurações de exibição da biblioteca Pandas foram listados os valores de casos por bairro e óbitos por bairro que foram adicionados na Tabela 1 (Seção 4.2).

Os gráficos da FIGURA 3 (seção 4.3) foram feitos com o apoio da biblioteca Matplotlib, criando dois subplots para visualização das contagens de casos e óbitos por distrito sanitário de residência.

Da mesma forma, os dados de casos e óbitos por gênero, apresentados na FIGURA 4 (seção 4.4), foram gerados com dois subplots para visualização clara da distribuição dos casos e óbitos por gênero.

Fornece um gráfico de barras combinado para a visualização dos casos e óbitos por faixa etária (seção 4.5). Além disso, por se tratar de uma variável numérica foi calculado e exibido as estatísticas descritivas gerais da coluna.

Para análise dos óbitos entre pacientes internados e não internados (seção 4.8) foi filtrado o DataFrame entre os pacientes internados e não internados e contados os óbitos confirmados e não confirmados de cada grupo. Em cima disso, criou-se um gráfico com dois subplots para visualização dos dados.

Criou-se dois gráficos (seção 4.6) para visualização de casos e óbitos agrupados por mês ao longo do período estudado permitindo uma análise visual rápida das tendências temporais.

Para visualização do número de internamentos agrupado por mês ao longo do período (seção 4.7) foi filtrado o DataFrame apenas pelos pacientes internados, esses dados foram agrupados por mês e exibidos em um gráfico de barras.

Os dados de status de encerramento dos casos, apontados no inicio da seção 4, foram obtidos através da contagem dos valores únicos na coluna ‘ENCERRAMENTO’ e posterior exibição em formato de gráfico de barras horizontal.

Para realização do cálculo de correlação usando o método de Spearman (seção 4.9) foi criada uma cópia do DataFrame que teve suas variáveis categóricas convertidas para valores numéricos usando codificação ordinal e mapeamento binário. Colunas que não teriam uso na análise foram removidas e o cálculo e exibição da matriz de correlação foram executados.

Para realização dos testes de Shapiro-Wilk foram utilizadas amostras aleatórias de 5000 registros de cada coluna através do código abaixo.