COVID-19 NOS ANOS DE 2020 E 2021 – Uma análise estatística e visual

André Francis Soares Nunes Maria[1]
andre.maria@fatec.sp.gov.br
Antunes Sabino de Souza[2]
antunes.souza01@fatec.sp.gov.br
Leonardo dos Santos Freitas[3]
leonardo.freitas4@fatec.sp.gov.br
Lucas Lima[4]
lucas.lima142@fatec.sp.gov.br
Rafael Moraes de Camargo Clem[5]
rafael.clem@fatec.sp.gov.br

RESUMO: Este estudo analisou um dataset sobre casos de COVID-19 no Brasil entre 2020 e 2021, explorando relações entre variáveis como idade, gênero, sintomas, tempo de internação e necessidade de cuidados intensivos. Utilizando dados do Global.health, coletados até julho de 2021 por Anya Lindström Batalha da Universidade de Oxford. Investigou-se a correlação entre o início dos sintomas e o tempo de recuperação, além de explorar fatores de risco como obesidade, doenças relacionadas ao coração, diabetes, asma, e sua gravidade associados a covid-19. Os resultados destacaram que doenças cardiovasculares, diabetes e hipertensão estão fortemente correlacionadas com óbitos por COVID-19. Além disso verificou também se os pacientes que buscaram atendimento médico mais cedo tiveram menor tempo de internação do que os pacientes que demoraram mais. Enquanto a hipótese sobre o tempo até a hospitalização não se confirmou. Estados como São Paulo, Minas Gerais e Rio Grande do Sul lideraram em números absolutos de casos e recuperações, influenciados pelo tamanho populacional. A análise sublinhou a importância de direcionar medidas de saúde pública para mitigar os impactos da pandemia, especialmente entre populações vulneráveis e comorbidades específicas.

Palavras chaves: Covid-19; fatores de risco; dataset; internação;

[1] André Francis Soares Nunes Maria, discente na Fatec Cotia. Ciência de Dados, Cotia, Centro Paula Souza. Orientado (a) por Johne Marcus Jarske.

[2] Antunes Sabino de Souza, discente na Fatec Cotia. Ciência de Dados, Cotia, Centro Paula Souza. Orientado (a) por Johne Marcus Jarske.

[3] Leonado dos Santos Freitas, discente na Fatec Cotia. Ciência de Dados, Cotia, Centro Paula Souza. Orientado (a) por Johne Marcus Jarske.

[4] Lucas Lima, discente na Fatec Cotia. Ciência de Dados, Cotia, Centro Paula Souza. Orientado (a) por Johne Marcus Jarske

[5] Rafael Moraes de Camargo Clem, discente na Fatec Cotia. Ciência de Dados, Cotia, Centro Paula Souza. Orientado (a) por Johne Marcus Jarske.

1. Introdução

Este estudo explora um conjunto de dados, e aplica conceitos estatísticos para investigar os casos de covid-19 enfrentados pala população brasileira entre os anos de 2020 e 2021. Além disso, tenta encontrar relações entre diferentes regiões do Brasil, levando em consideração variáveis como, idade, gênero, sintomas, tempo de internação, necessidade de cuidados intensivos, entre outros.

O presente artigo se baseia em um conjunto de dados coletados do site Global.helth, de 26 de julho de 2021 sob a autoria da equipe Global.health, representado pela cientista de dados da Universidade de Oxford, Anya Lindström Batalha. Vale destacar que o conjunto de dados não se destina a rastrear todos os casos de SARS-CoV-2 relatados no Brasil, mas apenas aqueles com desfechos mais graves que requerem hospitalização, segundo os autores do dataset, representam 5,5% em comparação ao total de casos informados pela Organização Mundial da Saúde em 26 de julho de 2021.

O objetivo é entender o comportamento da doença no Brasil, qual a relação entre as variáveis do dataset. Como afirma o Ministério da Saúde do Brasil, a covid-19 é uma infecção respiratória aguda causada pelo coronavírus SARS-CoV-2, potencialmente grave, de elevada transmissibilidade e de distribuição global (MINSTÉRIO DA SAÚDE, BRASIL, 2020). O vírus foi descoberto em amostras de lavado broncoalveolar obtidas de pacientes com pneumonia de causa desconhecida na cidade de Wuhan, província de Hubei, China, em dezembro de 2019 (MINSTÉRIO DA SAÚDE, BRASIL, 2022) a rápida disseminação do vírus expôs vulnerabilidades em sistemas de saúde, economias e estruturas sociais em todo o mundo. Medidas de saúde, como lockdowns, distanciamento social e uso generalizado de máscaras foram implementadas em uma tentativa de conter a propagação do vírus e mitigar seu impacto.

A pesquisa se concentra em estudar a variação do tempo de internação hospitalar e a necessidade de cuidados intensivos, e explorar qual a relação entre o início dos sintomas e a resposta final ao tratamento.

Esta pesquisa busca entender também como os fatores de risco, como obesidade, colesterol, tabagismo, entre outros, poder estar relacionados com a ocorrência dos casos mais graves da covid-19.

Com base nessas premissas, formulamos as seguintes hipóteses e utilizamos de métodos estatísticos e visuais para fazer a análise delas:

Hipótese 1: Existe uma correlação entre o intervalo do início dos sintomas e o tempo de recuperação, sugerindo que casos com tratamentos iniciados precocemente possam ter uma recuperação mais rápida ou mais eficiente.

Hipótese 2: Há uma correlação entre a presença de fatores de risco preexistentes, e o desfecho do tratamento, indicando que certos fatores de risco podem ter um impacto maior na recuperação do paciente.

2. Referencial Teórico

A pandemia do novo Coronavírus (SARS-CoV-2), declarada pela Organização Mundial da Saúde em 11 de março de 2020 (OMS, 2020), impôs diversos desafios e preocupações em âmbito global. Este referencial teórico tem como objetivo abordar aspectos relevantes da etiologia, quadro clínico, complicações e prognóstico da COVID-19, com ênfase no contexto brasileiro.

O SARS-CoV-2 pertence à família Coronaviridae, caracterizada por sua aparência semelhante a uma coroa sob microscopia (MINISTÉRIO DA SAÚDE, BRASIL, 2020). A transmissão do vírus ocorre principalmente por meio de gotículas respiratórias expelidas por indivíduos infectados ao tossir, espirrar, falar ou respirar (VIEIRA ET AL., 2020). O contato direto com pessoas contaminadas ou com superfícies contaminadas também pode levar à infecção (MINISTÉRIO

DA SAÚDE, BRASIL, 2020). Estudos sugerem a possibilidade de transmissão aérea em ambientes fechados e mal ventilados (SANTANA ET AL., 2021).

A propagação da COVID-19 é influenciada por diversos fatores, incluindo a densidade populacional, mobilidade humana, padrões comportamentais e práticas de saúde pública (MINISTÉRIO DA SAÚDE, BRASIL, 2020). A falta de distanciamento físico, uso inconsistente de máscaras, aglomerações e ventilação inadequada aumentam o risco de transmissão (MINISTÉRIO DA SAÚDE, BRASIL, 2020).

> "Durante a pandemia, muitas dessas desigualdades puderam ser evidenciadas, mostrando que grupos historicamente excluídos e em desvantagem social encontravam-se mais vulneráveis, não só aos riscos associados ao novo vírus, mas também ao desemprego, evasão escolar, pobreza e violência" (COBO, CRUZ, DICK, 2021, p 4022).

A capacidade do vírus de se mutar, gerando variantes mais transmissíveis ou resistentes a vacinas, também influencia a propagação (FARIA ET AL., 2021). Além disso, a desinformação e as fake news dificultaram os esforços de controle da doença (ROSA, DELDUQUE, ALVES, 2023).

A prevenção da COVID-19 é fundamental para reduzir a transmissão do vírus e proteger a saúde pública. As principais medidas incluem:

- Distanciamento físico: Manter uma distância segura de outras pessoas, especialmente em ambientes fechados.
- Uso de máscaras: Utilizar máscaras faciais em locais públicos e situações que impedem o distanciamento físico.
- Higiene das mãos: Lavar as mãos frequentemente com água e sabão por pelo menos 20 segundos ou utilizar álcool em gel.
- Evitar aglomerações: Reduzir a participação em eventos com grande número de pessoas, principalmente em locais fechados e mal ventilados.

- Ventilação adequada: Manter ambientes fechados ventilados, abrindo janelas e portas.
- Vacinação: Vacinar-se contra a COVID-19 de acordo com o cronograma do Ministério da Saúde.

A prevenção da COVID-19 exige uma abordagem abrangente que combine medidas de distanciamento físico, uso de máscaras, higiene das mãos, controle de aglomerações, ventilação adequada e vacinação. A implementação dessas medidas de forma conjunta é recomendada para proteger a saúde pública e conter a disseminação do vírus no Brasil.

Além das ações preventivas a probabilidade de recuperação do paciente é maior o quanto antes iniciado o tratamento. “Menor letalidade e outros desfechos negativos podem estar associados a maior percepção dos sintomas da doença e rápida procura por serviços de saúde, tratando-se do sexo feminino. Indivíduos do sexo masculino tendem a buscar os serviços de saúde apenas nas fases mais graves da doença, quando geralmente são menores os recursos terapêuticos.” (MASCARELLO, VIEIRA, SOUZA, MARCARINI, BARAUNA, MACIEL, 2021, p.7)

Mesmo com os cuidados preventivos contra o coronavírus, existem fatores de risco que podem contribuir para o desenvolvimento da forma grave da Covid-19. Segundo o Instituto Federal de Santa Catarina, fatores de risco, “são uma série de condições e comorbidades, ou seja, doenças prévias que fazem com que a pessoa tenha maior probabilidade de desenvolver a forma grave da doença, de necessitar de hospitalização ou ter mais chances de morrer em comparação a quem não tem nenhum fator.” (INSTITUTO FEDERAL DE SANTA CATARINA, BRASIL, 2020), Ainda segundo a mesma pesquisa, os principais fatores de risco são: Cardiopatia, diabetes, doença renal, doenças neurológicas, pneumonia, imunodepressão, obesidade, asma, doença hepática e doença hematológica, (INSTITUTO FEDERAL DE SANTA CATARINA, BRASIL, 2020).

3. Metodologia da Pesquisa

O instrumental de pesquisa se constitui em um conjuntos de dados disponibilizados pelo grupo Global.health, em formato csv (1) “BR.csv”, disponível em “https://data.covid-19.global.health/”, contendo uma tabela com 3.853.973 (três milhões, oitocentos e cinquenta e três mil, novecentos e setenta e três) linhas e 76 (setenta e seis) colunas, e tamanho total do arquivo de 1.61 GB (um vírgula sessenta e um) gigabytes, acessados em 08 de abril de 2024. O processo de tratamento desses dados foi através da linguagem Python com a biblioteca Pandas, e para a criação dos gráficos foram usadas a linguagem Python e as bibliotecas pandas, numpy, matplotlib.pyplot, sklearn, scipy, statsmodels.stats.api. Os dados foram examinados, verificando-se a consistência em relação às variáveis utilizadas, quanto aos limites do dataset. Foram também verificados se existem valores anormais ou em desacordo nas variáveis do dataset. Após o processo de tratamento dos dados, o conjunto foi reduzido a um total de 878.589 (oitocentos e setenta e oito mil, quinhentos e oitenta e nove) linhas, e 35 (trinta e cinco) colunas, e um total de 211 MB (duzentos e onze) megabytes.

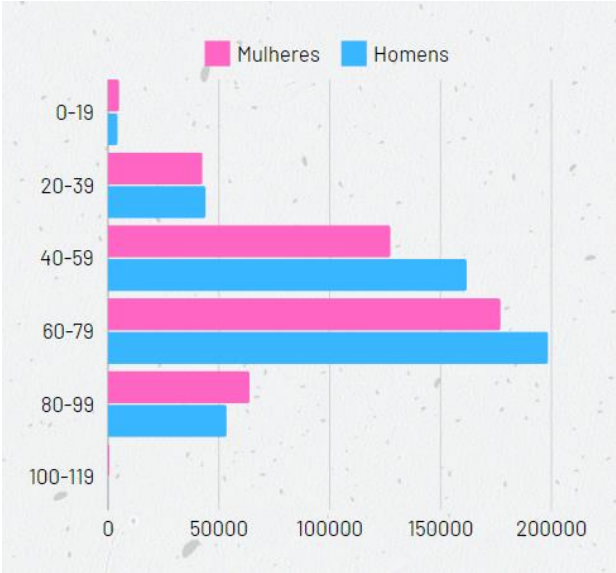
4. Análises

A seguir, vamos apresentar a composição dos gráficos derivados da análise dos conjuntos de dados que incluem 878.589 (duzentos e sete mil, trezentos e cinquenta e dois) registros de casos de COVID-19. Estas análises incluem uma exploração descritiva dos dados contidos no conjunto de dados em estudo.

4.1 Casos por idade

Utilizando o método estatístico descritivo, escolhemos o gráfico de pirâmide para representar os totais de casos separados por idade e sexo com um intervalo de 20 anos.

Gráfico 1: Casos por idade e sexo



Fonte: (Própria)

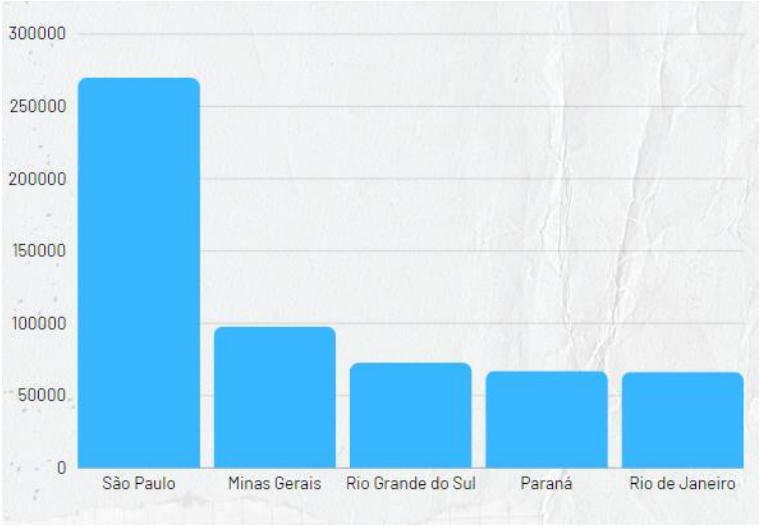
Pode-se observar que a distribuição de casos varia ao longo das diferentes faixas etárias. Essa análise também indica a necessidade de cuidados e prevenção em diferentes segmentos da população.

O gráfico oferece uma representação visual das tendências e variabilidades nos casos. Observa-se que a maior parte dos casos de covid-19 se concentra entre 60 e 79 anos.

4.2 Casos por estado

Utilizando o método estatístico descritivo, escolhemos o gráfico de barras na vertical para representar os cinco estados com maiores números de casos.

Gráfico 2: Total de casos por estado



Fonte: (Própria)

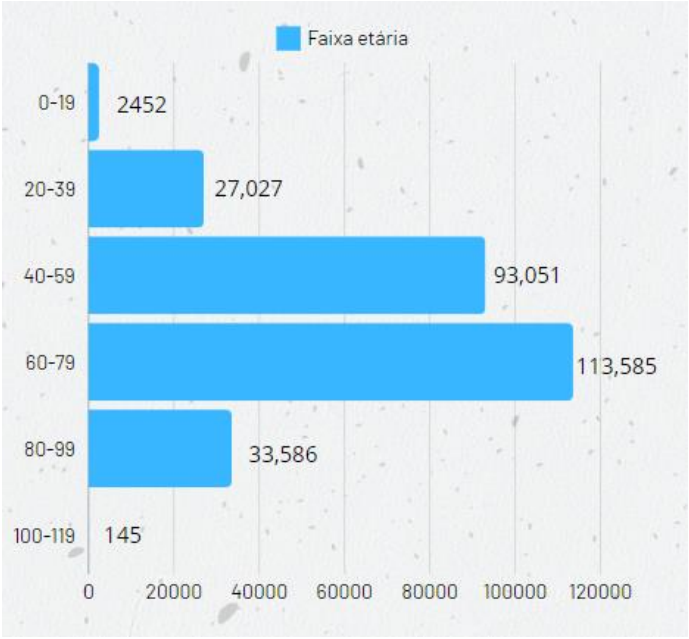
O gráfico fornece uma visualização comparativa da distribuição dos casos de COVID-19 entre os estados brasileiros.

Pode-se observar que os estados com maiores números de casos, destacando São Paulo, Minas Gerais e Rio Grande do Sul.

4.3 Total de casos no estado de São Paulo

Ainda utilizando o método estatístico descritivo, usamos o gráfico de barras na horizontal para expressar a distribuição de casos no estado de São Paulo, separados por faixa etária.

Gráfico 3: Total de casos em SP



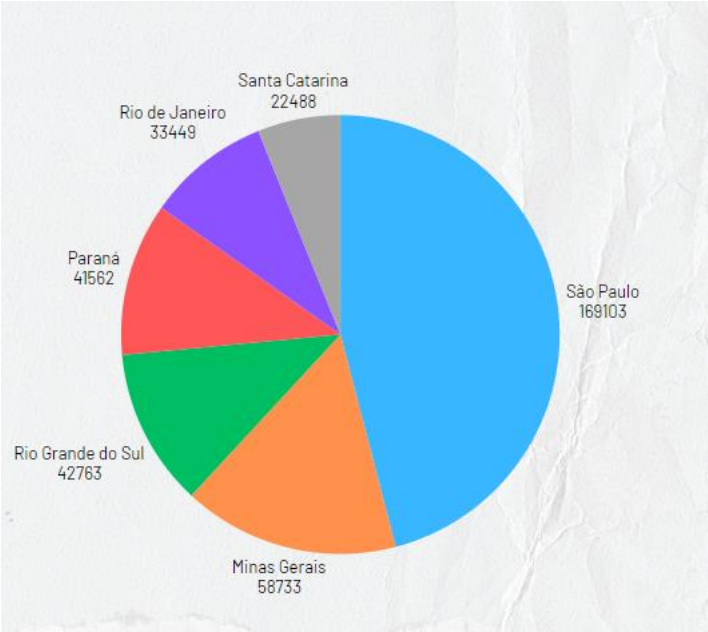
Fonte: (Própria)

Observou que assim como ocorre no total do Brasil, as faixas etárias com o maior número de ocorrências são aquelas entre 60 e 79 anos, e entre 40 e 59 anos.

4.4 Recuperados por estado

Utilizando o método estatístico descritivo, escolhemos o gráfico de setores para demonstrar o número de casos em que a vítima se recuperou dentro dos estados com maior número de ocorrências.

Gráfico 4: Recuperados por estado



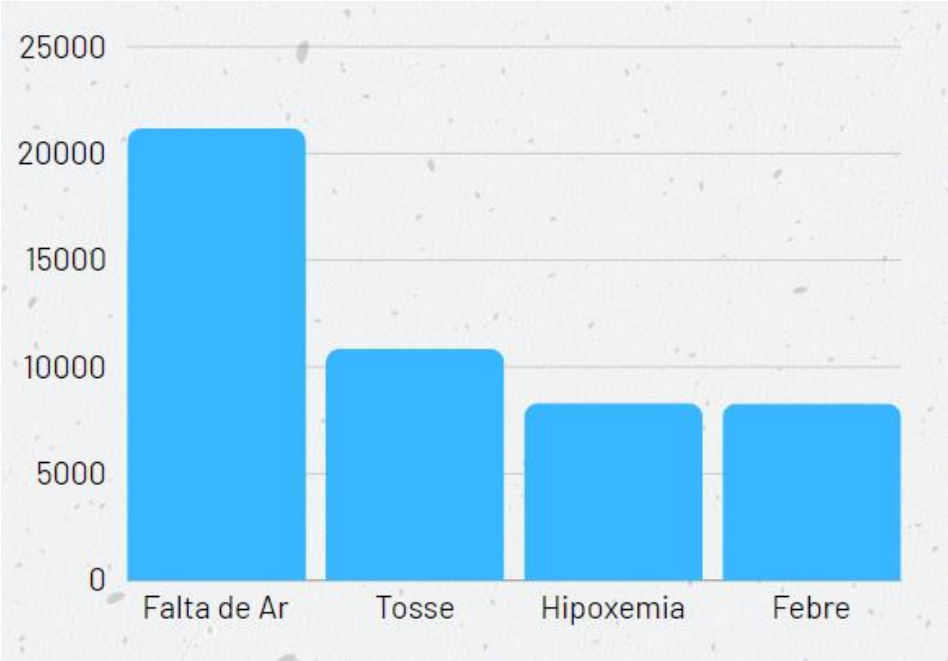
Fonte: (Própria)

Assim como São Paulo obteve o maior número de casos, ele também deteve do maior número de recuperados entre todos do Brasil.

4.5 Principais sintomas

Utilizando o método estatístico descritivo, foi elaborado um gráfico de barras horizontal para representar a distribuição dos casos de sintomas por faixa etária no estado de São Paulo. Este gráfico é uma ferramenta visual que permite comparar facilmente o número de casos em diferentes grupos etários.

Gráfico 5: Contagem de sintomas

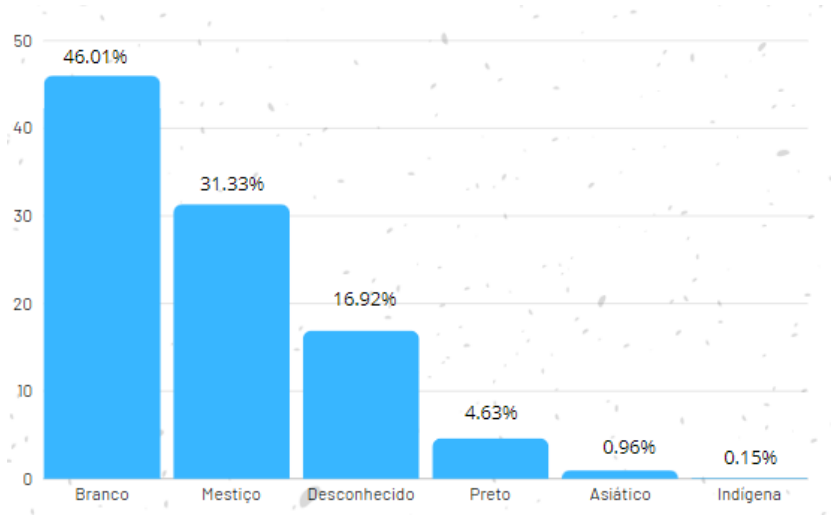


Fonte: (Própria)

Essa representação visual permite uma comparação direta entre a incidência de diferentes sintomas, destacando aqueles que são mais frequentemente associados à doença. Em que os casos mais comuns são falta de ar, tosse, hipoxemia e febre.

4.6 Distribuição dos casos por etnia

Gráfico 6: Casos por etnia



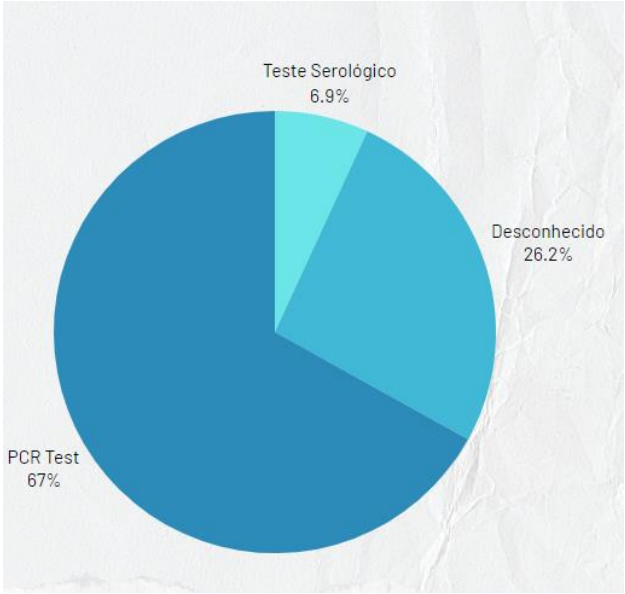
Fonte: (Própria)

O gráfico acima representa a etnia contemplada dentro de nossa base de dados. A grande parte dos pacientes pertencem a etnia branca, representando quase 50% dos dados contemplados. Temos nele que as populações com menores percentuais são a preta, asiática e indígena que representa apenas 0.15%.

4.7 Método de confirmação

Utilizamos um gráfico de setores para representar os métodos de confirmação, esse tipo de gráfico permite visualizar de forma clara e comparativa a distribuição dos tipos de testes.

Gráfico 7: Tipos de teste



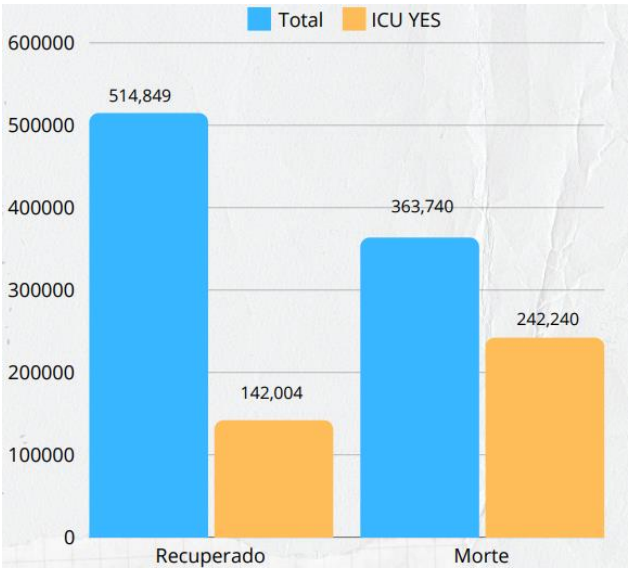
Fonte: (Própria)

O gráfico mostra o método de confirmação utilizado para diagnosticar a doença. O método mais usado foi o de PCR (Polymerase Chain Reaction), representando um total de 67% dos métodos utilizados. O PCR é uma tecnologia que consiste na amplificação de uma região específica do DNA, gerando a possibilidade de produzir uma enorme quantidade de material a ser analisado. Dentro da base, 26,2% foram catalogados como desconhecidos e, 6,9% foram diagnosticados através de teste serológico.

4.8 Ocorrência de necessidade de cuidados intensivos

Utilizamos os gráficos de barras duplas para representar a necessidade de cuidados intensivos, tanto dos pacientes recuperados, quanto dos pacientes falecidos.

Gráfico 8: Necessidade de cuidados intensivos



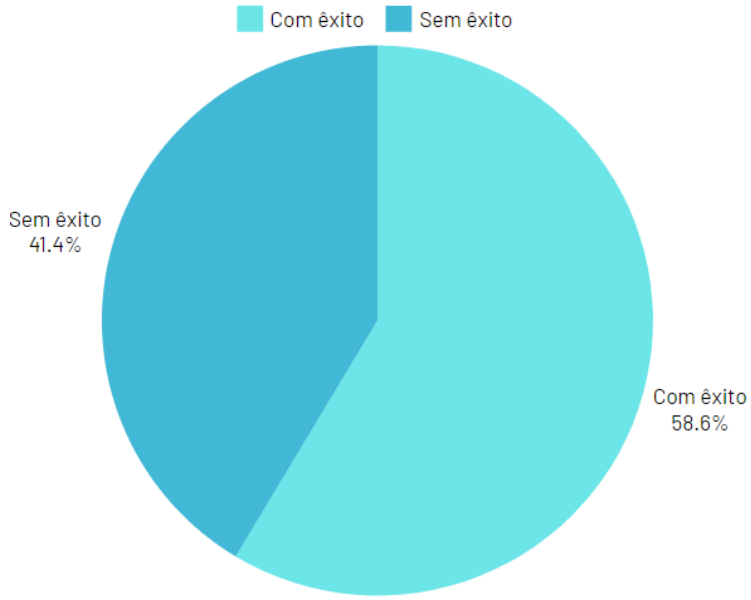
Fonte: (Própria)

No gráfico podemos notar que mesmo os pacientes recuperados tenham sido a maioria, os pacientes que foram a óbitos, tiveram um maior índice de necessidade de cuidados intensivos.

4.9 Recuperação

Novamente utilizado um gráfico de setores, identificamos a proporção da população que mais.

Gráfico 9: Taxa de recuperação



Fonte: (Própria)

Acima, podemos ver os casos que obtiveram êxito em sua recuperação, ou seja, tiveram alta do hospital, e os casos que não obtiveram, resultando em fatalidade. Analisando os casos registrados, foram confirmadas 363.739 mortes, representando 41,4% da base estudada. O número de altas representa 58,6%, ou seja, 514.848 recuperados da COVID-19.

4.10 Correlação entre início dos sintomas e a hospitalização por tempo de hospitalização.

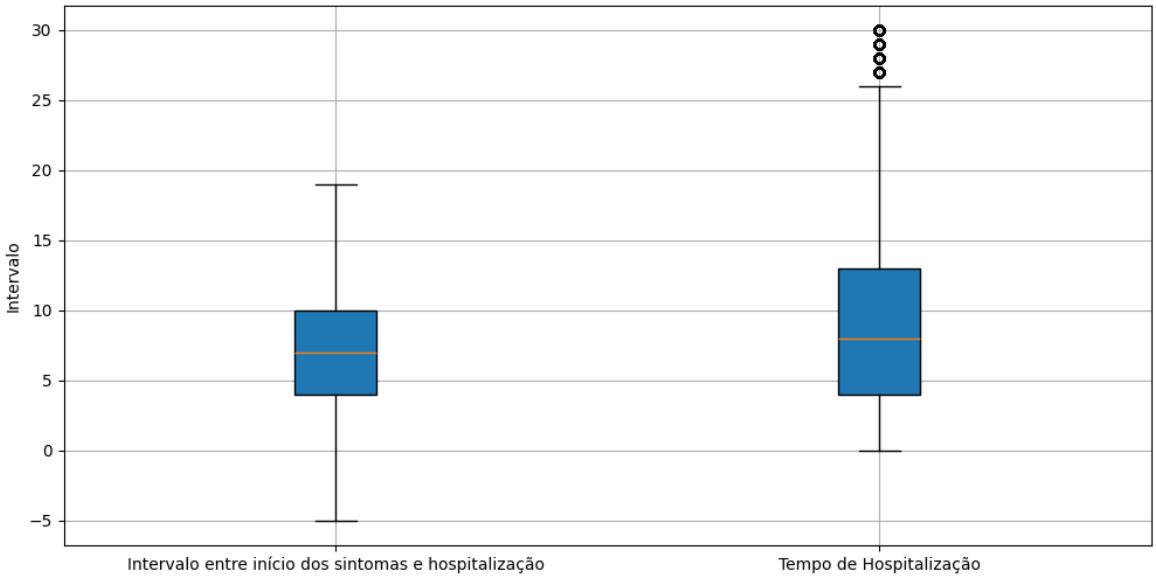
Para avaliar essa hipótese, serão analisados dados clínicos detalhados, incluindo a data de início dos sintomas e a duração da hospitalização de cada paciente.

O estudo envolverá a coleta e análise de informações de pacientes internados, buscando identificar padrões ou relações estatísticas entre a rapidez com que o paciente procurou um sistema de saúde, e o tempo que paciente permaneceu hospitalizado.

4.9.1 Intervalos entre início dos sintomas e a hospitalização, e, tempo de hospitalização

Para entender melhor como variavam os tempos de internação, recorremos a um gráfico de caixa. Esse tipo de representação nos possibilita comparar de maneira proporcional e em uma escala comum.

Gráfico 10: Comparação dos intervalos



Fonte: (Própria)

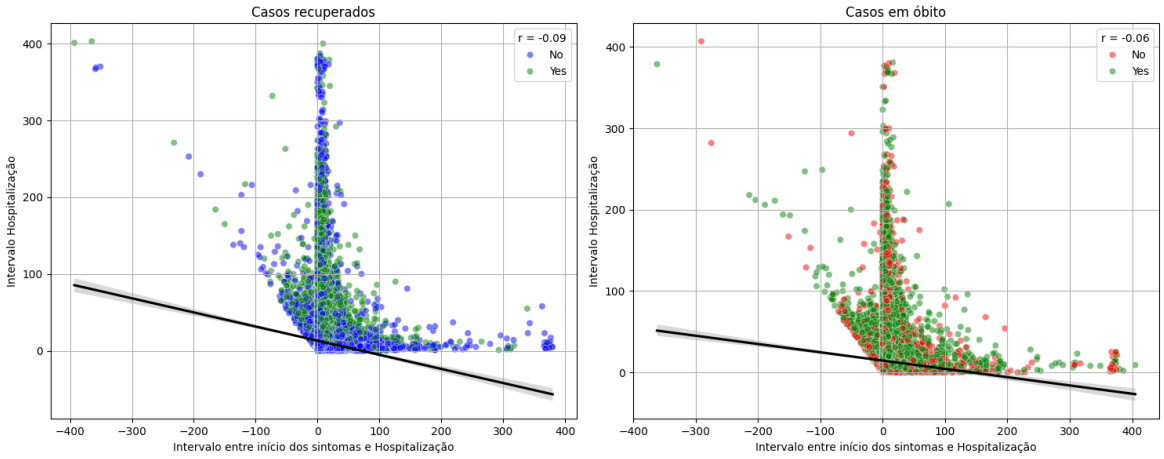
Nota-se que na população total da pesquisa o intervalo entre o início dos sintomas e a procura por um centro de saúde é menor do que o tempo de

hospitalização. Deve-se levar em consideração que existem mais outliers (ponto fora da curva), mas para melhorar a apresentação, o gráfico foi criado de forma com que as caixas possam ser mais bem visualizadas.

4.9.2 Correlação entre intervalos entre início dos sintomas e a hospitalização, e, tempo de hospitalização

Para verificar esta correlação foi um utilizado os métodos estatísticos descritivos e indutivos para, aplicando os valores em um gráfico de correlação, e traçado a reta r, do algoritmo de correlação de Pearson.

Gráfico 11: Correlação entre os intervalos



Fonte: (Própria)

O gráfico mostra a correlação entre as variáveis, indicando que a distribuição dos casos recuperados é parecida com a distribuição dos casos que foram a óbito, porém a ocorrência de necessidade de cuidados intensivos foi maior entre os pacientes falecidos. A reta de correlação de Pearson também foi

bom semelhante entre os dos gráficos, com um coeficiente de correlação de -0,09 nos casos recuperados e de -0,06 nos casos em óbito.

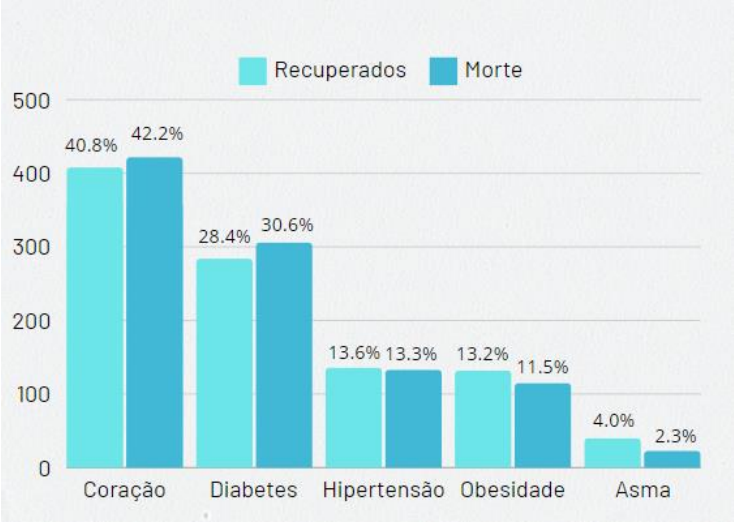
4.11 Identificar de acordo com o grupo de risco isolado, a taxa de mortalidade foi maior.

Para avaliar a segunda hipótese que é a de que a taxa de mortalidade é maior entre indivíduos que possuem fatores de risco isolados. Isso implica que, ao analisar separadamente cada condição pré-existentes, algum desses fatores implicaria mais na taxa de mortalidade dos pacientes.

O estudo deverá separar cada fator de risco e examinar separadamente se ele tem maior influência na taxa de mortalidade dos pacientes.

4.10.1 Fatores de risco mais comuns

Gráfico 12: Correlação entre os intervalos



Fonte: (Própria)

Os resultados das análises confirmam a hipótese, indicando que alguns fatores de risco pré-existentes têm maior impacto na mortalidade. O estudo revelou que os fatores de risco mais influentes são, em ordem de mortalidade crescente: pessoas com doenças cardíacas, diabetes, hipertensão, obesidade e asma. Os indivíduos com doenças cardíacas ou diabetes apresentaram a maior taxa de mortalidade entre os fatores de risco analisados. Esta condição foi identificada como a mais crítica, com uma taxa de mortalidade superior a de recuperados.

5.1 Considerações finais

Os testes de hipóteses realizados com algoritmos de correlação de Pearson forneceram insights sobre os fatores que influenciam as mortes por COVID-19 e outros aspectos relacionados à pandemia. Os resultados indicam que doenças cardiovasculares, diabetes e hipertensão são condições pré-existentes fortemente correlacionadas com as mortes por COVID-19, confirmando a hipótese inicial de que essas comorbidades impactam significativamente os casos fatais. Por outro lado, a hipótese de que o tempo entre o início dos sintomas e a hospitalização estaria correlacionado com a probabilidade de recuperação ou óbito foi refutada.

Em relação à distribuição dos casos, os estados de São Paulo, Minas Gerais e Rio Grande do Sul apresentaram os maiores números absolutos de casos e recuperações, o que pode ser atribuído ao tamanho de suas populações. Quanto aos métodos de detecção, o PCR foi o mais utilizado, seguido por métodos desconhecidos e, por último, pelo soro fisiológico.

Observou-se um equilíbrio no número de casos entre homens e mulheres em diversas faixas etárias, com a maior concentração de casos na faixa etária

de 60 a 79 anos. Esses achados destacam a importância de focar nas condições pré-existentes e nas populações mais vulneráveis para mitigar os impactos da COVID-19.

Referências

Brazil Covid-19 Line List [Dataset], 26 jul 2021. Global.health. Disponivel em [https://data.covid-19.global.health/]. Acesso em: 08 abr 2023.

BRASIL. Ministério da Cidadania. **Coronavírus: o que você precisa saber.** 2020.

BRASIL. Ministério da Saúde. **COVID-19: Perguntas e Respostas.** 2020.

BRASIL. Ministério da Saúde. **Plano Nacional de Operacionalização da Vacinação contra a COVID-19.** 2021.

COBO, Barbara; CRUZ, Claudia; DICK, Paulo C. **Desigualdades de gênero e raciais no acesso e uso dos serviços de atenção primária à saúde no Brasil**. Scielo Brasil, p. 4022, 2021

BRASIL. Instituto Federal de Santa Catarina. **Você tem um ou mais fatores de risco para a Covid-19?** 2020

ROSA, Tiago; DELDUQUE, Maria Célia; ALVES, Sandra Mara Campos. **A pandemia de covid-19 e as fakes news: uma revisão da literatura.** Scielo Saúde Pública, p. 1, 2023

FARIA, N. R. et al. **Genomic diversity and emergence of lineage B.1.1.7 of SARS-CoV-2 in Brazil.** Nature, v. 596, n. 7873, p. 1107-1111, 2021

MASCARELLO, Keila Cristina; VIEIRA, Anne Caroline Barbosa Cerqueira; SOUZA, Ana Sara Semeão de; MARCARINI, Wena Dantas; BARAUNA, Valério Garrone; MACIEL, Ethel Leonor Noia. **Hospitalização e morte por COVID-19 e sua relação com determinantes sociais da saúde e morbidades no Espírito Santo: um estudo transversal**. Scielo SP, p. 7, 2021

Organização Mundial da Saúde (OMS). **Declaração de Pandemia de COVID-19.** 11 de março de 2020

SANTANA, M. C. et al. **Transmissão aérea do SARS-CoV-2: uma revisão sistemática da literatura.** Revista Brasileira de Epidemiologia, v. 24, n. 1, p. e10249, 2021.

VIEIRA, R. L. et al. **Transmissão do SARS-CoV-2 e medidas de controle: uma revisão sistemática da literatura.** Revista Brasileira de Medicina, v. 100, n. 2, p. 189-202, 2020.

SISTEMA DE INFORMAÇÃO DA VIGILÂNCIA EPIDEMIOLÓGICA DA GRIPE. Dados atualizados em 07 de setembro de 2020