Análise de dados aplicada aos dados de vacinação e ocupação de leitos no estado de São Paulo

Daniel Paulo de Assis Júnior[1], Guilherme Silva Eduardo[2], Sandra Cristina Costa Prado[3]

[1] Discente do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas / daniel.assis2@fatec.sp.gov.br
[2] Discente do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas / guilherme.eduardo@fatec.sp.gov.br
[3] Docente do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas / sandra.prado01@fatec.sp.gov.br

RESUMO

O presente artigo tem como principal objetivo trazer informações referentes aos dados de vacinação e ocupação de leitos devido a complicações da COVID-19 de forma acessível, por meio de visualizações gráficas, visando correlacionar a vacinação ao número de internações em estado grave nos municípios do estado de São Paulo. Para efetuar a análise os dados foram extraídos do OpenDataSUS e armazenados em um banco de dados PostgreSQL, já a análise propriamente dita foi feita com a linguagem R utilizando bibliotecas de manipulação e análise de dados. Constatou-se que as cidades mais populosas foram as que mais vacinaram, porém, a ordem em que aparecem são distintas no que se refere a número de vacinados e número de habitantes. A ocupação de leitos de UTI por complicações ocasionadas pelo vírus não segue o mesmo padrão, mostrando que, em determinado grau, o número de vacinados está atrelado ao número de internados em estado grave.

Palavras-chave: Análise de dados; vacinação; ocupação de leitos.

1 INTRODUÇÃO

Tecnologia e saúde são campos que há muito tempo caminham juntos. Segundo um estudo realizado pela International Data Corporation (IDC), do final do ano de 2020 até o término do presente ano (2022), o investimento no setor da saúde na América Latina deve atingir US$ 1.931 bilhões de dólares, cerca de 10 bilhões de reais (na data desta publicação).

O aumento do investimento se deve, principalmente, à pandemia da COVID-19, que vem afetando o mundo todo, teve seu início com o surto no final de 2019 e foi declarada como pandemia pela Organização Mundial da Saúde (OMS) em março de 2020. Tal fato fomentou o crescimento de diversas áreas, fazendo com que

buscassem modernização com o intuito de se adaptar às demandas trazidas pela pandemia.

Um campo que se tornou de suma importância e teve um aumento significativo foi a ciência de dados. O termo Ciência de Dados diz respeito a uma área interdisciplinar, exigindo conhecimentos matemáticos, estatísticos, além de programação e conhecimento analítico, que vem se tornando cada dia mais importante, sendo usada em mercados como o financeiro, o esportivo e, devido ao atual momento, o da saúde e que tem como um dos principais objetivos auxiliar nas tomadas de decisão, tornando-as mais assertivas.

No âmbito da saúde, é usada desde suas análises de dados até o diagnóstico do vírus através de sistemas de inteligência artificial (IA). Seu uso data desde o início da pandemia com o levantamento de informações e acompanhamento dos casos por plataformas como Our World in Data e Worldmeter, além de competições do Kaggle e por órgãos de saúde ao redor do mundo, buscando informar e prever os possíveis rumos que o vírus tomaria, sendo assim tornar a luta mais efetiva.

O presente artigo tem como objetivo contextualizar e empregar o uso de ciência de dados, com enfoque na análise de dados aplicada à campanha de vacinação e seus impactos no número de casos e internações em decorrência das infecções por COVID-19.

2 REFERENCIAL TEÓRICO

A área de dados, assim como as demais áreas no âmbito da tecnologia possui um vocabulário à parte, parecendo até mesmo outro idioma e, de fato, é em algum grau, pois a maioria dos termos são derivados do inglês. Dentre eles destacam-se os que são de suma importância para esse estudo, sendo eles ETL e Data Pipeline.

A sigla ETL possui como significado, na sua língua original, extract, transform and load, o que em uma tradução livre seria extração, transformação e carregamento. O termo carga também é comumente utilizado como sinônimo do último item e será adotado como o termo em português no decorrer do artigo.

Tal processo é o ato de extrair o dado de sua origem, fazer as transformações necessárias para que as informações obtidas supram a necessidade e após isso carregar os dados transformados para que sejam empregados na aplicação.

Esse processo é um dos componentes de algo maior, chamado de Data Pipeline, a tradução de pipeline para o português varia do contexto no qual a palavra está inserida, no contexto atual e na área de computação de forma geral a tradução mais utilizada é segmentação de instruções, porém o mais comum é a utilização do termo em inglês, o que será feito no decorrer do artigo. Este termo diz respeito a uma série de etapas de processamento, consistindo em três elementos principais: uma fonte, o processamento e o destino (BLOG DATA SCIENCE ACADEMY, 2020).

3 METODOLOGIA

Para a realização do projeto se mostrou necessário a divisão das ferramentas em três grupos: o primeiro deles contém as ferramentas para operações com os dados, o segundo, o armazenamento e o terceiro, as tecnologias de análise.

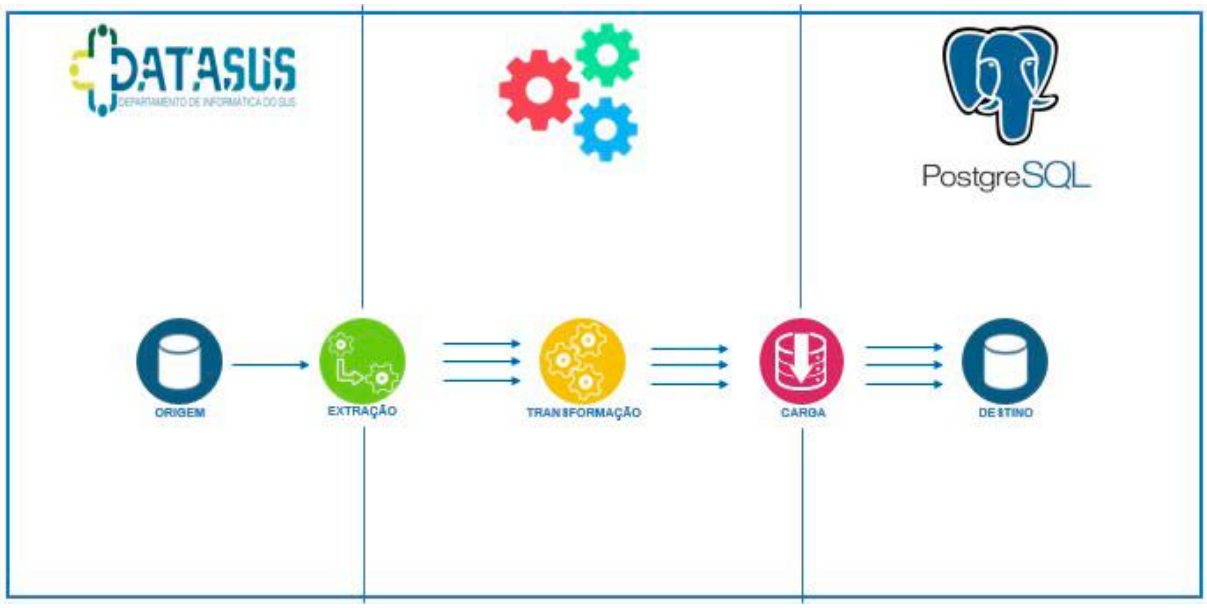
O primeiro grupo assume o papel de pilar de toda a estrutura do projeto, pois os dados são críticos para a realização do estudo, sendo assim devem ser sólidos e consistentes e principalmente confiáveis. Os dados utilizados referem-se à campanha de vacinação contra a COVID-19 e são encontrados no sistema de dados abertos da saúde, o OpenDataSUS, cobrindo todo o território nacional. Mas no presente artigo serão utilizados somente os dados referentes ao estado de São Paulo que, devido ao grande volume, encontram-se divididos em três partes e no formato de arquivo CSV. A extração de forma manual se mostra improdutiva por consumir tempo em demasia. Por tal fato optou-se pela criação de um pipeline de dados para promover o processo de extração, transformação e carga dos dados (ETL). Este processo será melhor explicado no decorrer deste artigo.

Para o armazenamento dos dados, o segundo do grupo, optou-se pelo armazenamento no banco de dados PostgreSQL, instalado em um servidor Windows Server 2016. Tal ferramenta foi escolhida devido a sua performance robusta, além de sua grande capacidade de armazenamento de forma gratuita, se mostrando como a melhor opção para o armazenamento e operação dos dados.

No processo de análise de dados, optou-se pela utilização da linguagem R, em seu ambiente RStudio, devido ao seu caráter analítico e, por conta disso, seu poderio frente aos desafios da análise de dados.

Após a definição das ferramentas, o processo prático foi iniciado com a criação do pipeline de dados, consistindo no processo de extrair os dados particionados do OpenDataSUS, aplicar o tratamento necessário, que de imediato seria a união dos três arquivos em um e, por fim, efetuar a carga no banco de dados, conforme representado no diagrama da arquitetura do pipeline mostrado na figura 1.

Figura 1 – Pipeline de dados 1.



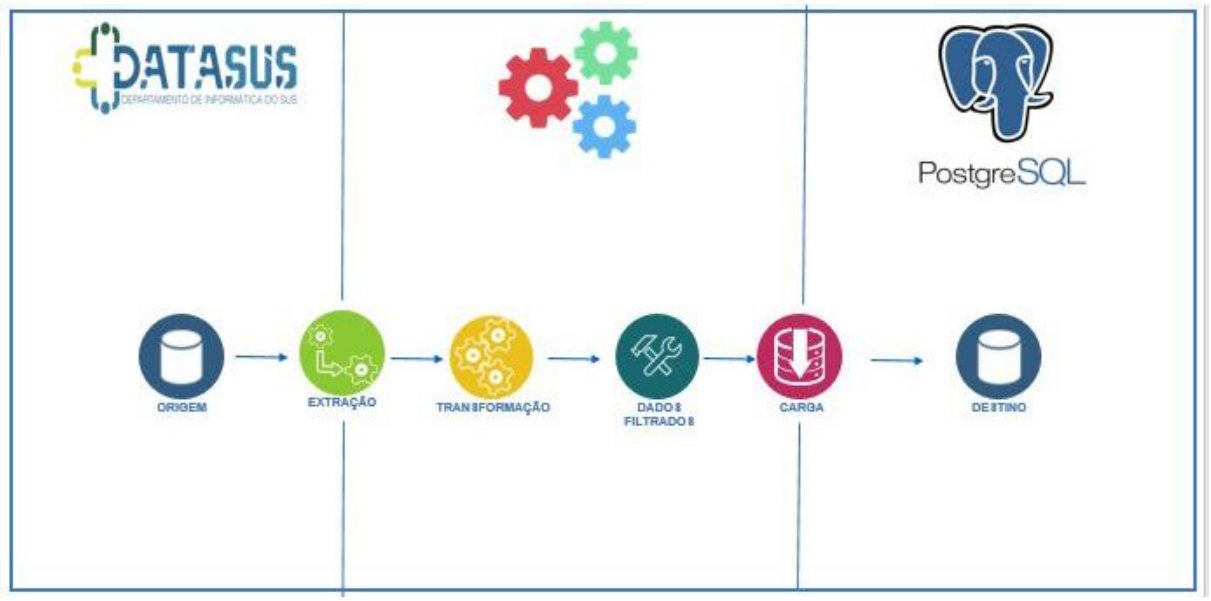
Fonte: Elaborado pelos autores**.**

Os dados originam-se do OpenDataSUS e sua extração ocorre através de um script escrito na linguagem de programação Python, com uso da biblioteca Selenium, utilizada para criação de robôs que emulam o comportamento humano em páginas web. Neste caso, o script realiza o download dos dados e, após esse processo, os dados são carregados pela função de transformação. Uma vez carregados, são renomeados para um padrão contendo a palavra de identificação "Dados" e um número de identificação indo de 1 a 3. Uma vez que a verificação foi realizada, a função responsável pela carga dos dados é acionada e se conecta com o servidor que hospeda o banco de dados. A carga dos dados é feita de modo individual em lotes de 250.000 linhas, uma vez que uma carga termina a seguinte é iniciada, isso se deve ao tamanho dos arquivos, pois cada um deles possui cerca de 15 GBs, logo a carga feita de forma única seria mais custosa. Uma vez que uma carga é feita seus dados

são alocados na tabela referente à campanha de vacinação, intitulada "vacinacao". Tal processo é feito até que a última carga seja efetuada.

Para uma melhor compreensão do impacto das vacinas no combate ao corona vírus, se faz necessário o levantamento de dados sobre o número de casos, internações e óbitos. Tais informações também são disponibilizadas pelo OpenDataSUS. Para a obtenção dos dados são referentes à ocupação hospitalar por complicações ocasionadas pelo coronavírus em todo território nacional no ano de 2022, também foi utilizado um pipeline de dados, porém com a etapa de transformação um pouco diferente, conforme representado na figura 2.

Figura 2 – Pipeline dados 2.



Fonte: Elaborado pelos autores.

O processo é muito similar ao pipeline responsável pelos dados da vacinação, principalmente na extração, diferindo-se no processo de transformação, nessa etapa ele passa por uma etapa de limpeza, pois, diferente dos dados da vacinação, os dados referem-se a todo o território nacional. Por conta disso, foi necessário um processo de limpeza para filtrar somente os dados do estado de São Paulo. Uma vez que o arquivo esteja "limpo", a função de carga é acionada e uma outra função realiza a conexão e a inserção dos dados na tabela destinada aos dados de ocupação dos leitos no estado de São Paulo, intitulada "ocupacao".

A estrutura do banco resultante da ingestão dos dados foram duas tabelas, uma contendo informações sobre a vacinação (vacinacao) e outra contendo informações sobre a ocupação de leitos e internações (ocupacao).

Após o processo de obtenção, iniciou-se a parte de análise dados, para a qual foi utilizada a linguagem R conforme explicado no parágrafo de abertura desta sessão, estabelecendo-se uma conexão entre o servidor contendo o banco de dados e o RStudio, devido ao grande volume de dados referentes a vacinação, que conta com mais de cem milhões de linhas, optou-se pelo uso de uma amostra de vinte milhões de linhas, pois a quantidade se mostrou válida para garantir a fidelidade ao conjunto total sem diminuir a qualidade dos resultados e sem apresentar problemas de desempenho durante o processo de análise.

A análise sustentou-se em três perguntas que visaram correlacionar a campanha de vacinação com os números da ocupação hospitalar em decorrência da corona vírus:

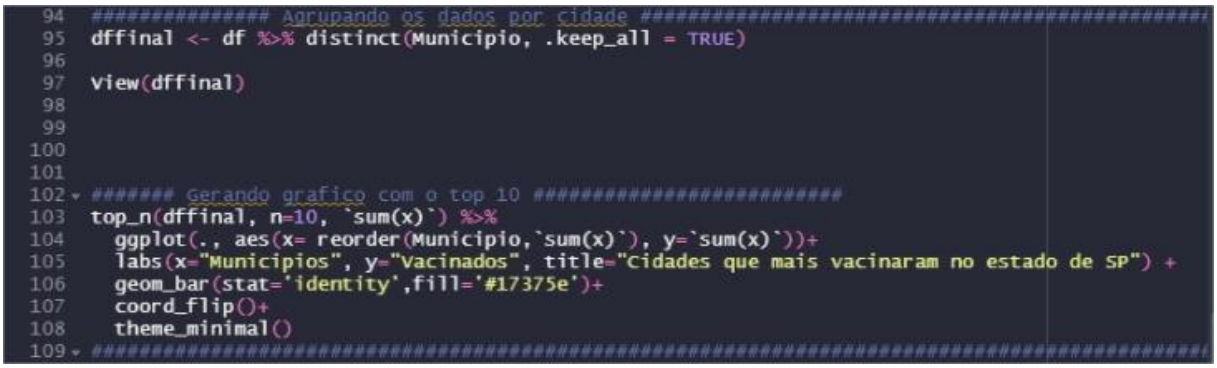
1. Quais cidades do estado mais vacinaram a população?
2. Quais cidades mais sofreram com internações devido ao vírus?
3. Qual o impacto das vacinações nas internações?

4 RESULTADOS E DISCUSSÃO

A primeira parte do processo de análise visou encontrar quais foram as cidades que mais vacinaram no estado de São Paulo, para tal foi necessária a manipulação dos dados de forma com que fossem agrupados por cidades e, então, computadas quantas vezes apareciam no dataset, pois cada linha indica uma vacina aplicada. Este processo foi executado pelo trecho de código mostrado na figura 3.

A análise dos dados de ocupação de UTI foi feita de forma similar: agrupamento dos dados por cidade, soma das quantidades referentes de pacientes e levantamento de quais cidades tiveram mais pacientes internados na UTI por complicações de corona vírus. Este processo foi executado pelo trecho de código mostrado na figura 4.

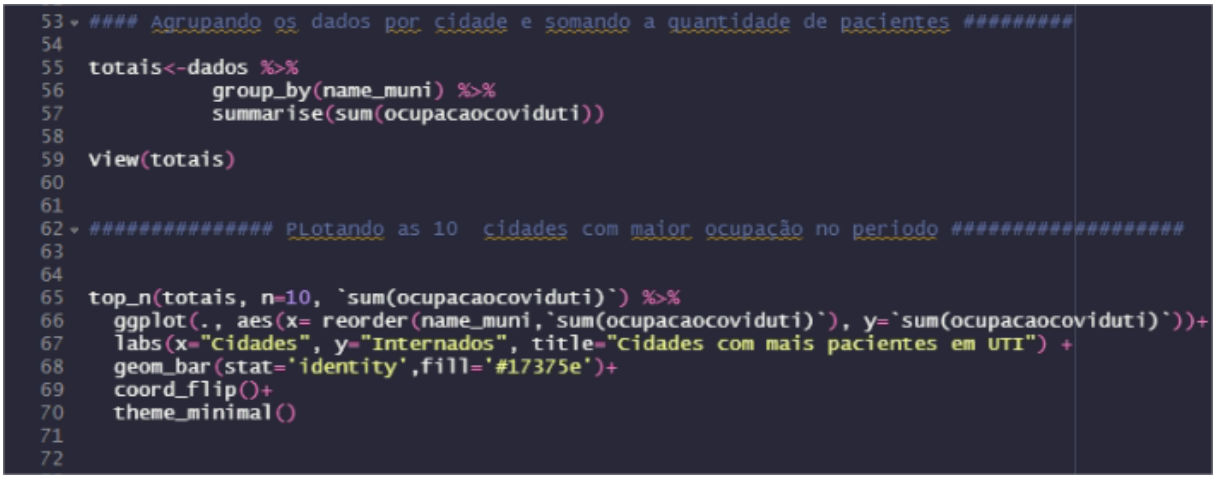
Figura 3 – Funções para contagem e criação do gráfico.



```
############### Agrupando os dados por cidade ##################################################
dffinal <- df %>% distinct(Municipio, .keep_all = TRUE)

View(dffinal)




######## Gerando grafico com o top 10 ############################
top_n(dffinal, n=10, `sum(x)`) %>%
  ggplot(., aes(x= reorder(Municipio,`sum(x)`), y=`sum(x)`))+
  labs(x="Municipios", y="Vacinados", title="Cidades que mais vacinaram no estado de SP") +
  geom_bar(stat='identity',fill='#17375e')+
  coord_flip()+
  theme_minimal()
################################################################################################
```

Fonte: Elaborado pelos autores.

Figura 4 – Agrupando e exibindo dados das cidades com maior ocupação.



```
#### Agrupando os dados por cidade e somando a quantidade de pacientes #########

totais<-dados %>%
          group_by(name_muni) %>%
          summarise(sum(ocupacaocoviduti))

View(totais)


############### PLotando as 10  cidades com maior ocupação no periodo ###################


top_n(totais, n=10, `sum(ocupacaocoviduti)`) %>%
  ggplot(., aes(x= reorder(name_muni,`sum(ocupacaocoviduti)`), y=`sum(ocupacaocoviduti)`))+
  labs(x="Cidades", y="Internados", title="Cidades com mais pacientes em UTI") +
  geom_bar(stat='identity',fill='#17375e')+
  coord_flip()+
  theme_minimal()
```

Fonte: Elaborado pelos autores.

As figuras 5 e 6 resumem os resultados destas análises em tabelas.

Adicionalmente, a tabela da figura 5 apresenta dados publicados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) indicando que nove das dez cidades apresentadas na figura 5 estão entre as dez mais populosas do estado, o que justifica o fato de serem as que mais vacinaram. Porém, dois fatos interessantes podem ser notados:

1. A ordem em que as cidades aparecem nos dados do IBGE e na figura 5 são idênticas até a quarta posição, mas a partir da quinta, a ordem passa a ser diferente.
2. Ausência da cidade de Mauá que, de acordo com informações do IBGE, é a décima cidade com mais habitantes no estado, porém não está entre as dez que

mais vacinaram. São José do Rio Preto, a décima primeira no ranking do IBGE, é a décima primeira que mais vacinou contra COVID-19.

As duas primeiras cidades, São Paulo e Campinas, são a primeira e terceira cidade mais populosas do estado, respectivamente, conforme mencionado anteriormente outras três das cidades mais populosas constam na lista, sendo elas Ribeirão Preto, Sorocaba e Osasco.

As figuras 7 e 8 resumem os dados em gráficos, para melhor compreensão dos dados, uma vez que o intervalo é de dez mil.

Figura 5 – As dez mais populosas versus dez que mais vacinaram.

Municipios	Posição no ranking de habitantes IBGE	Posição no ranking de vacinação
São Paulo	1	1
Guarulhos	2	2
Campinas	3	3
São Bernardo do Campo	4	4
São José dos Campos	5	9
Santo André	6	5
Ribeirão Preto	7	8
Osasco	8	6
Sorocaba	9	7
Mauá	10	NÃO APARECE
São José do Rio Preto	11	10

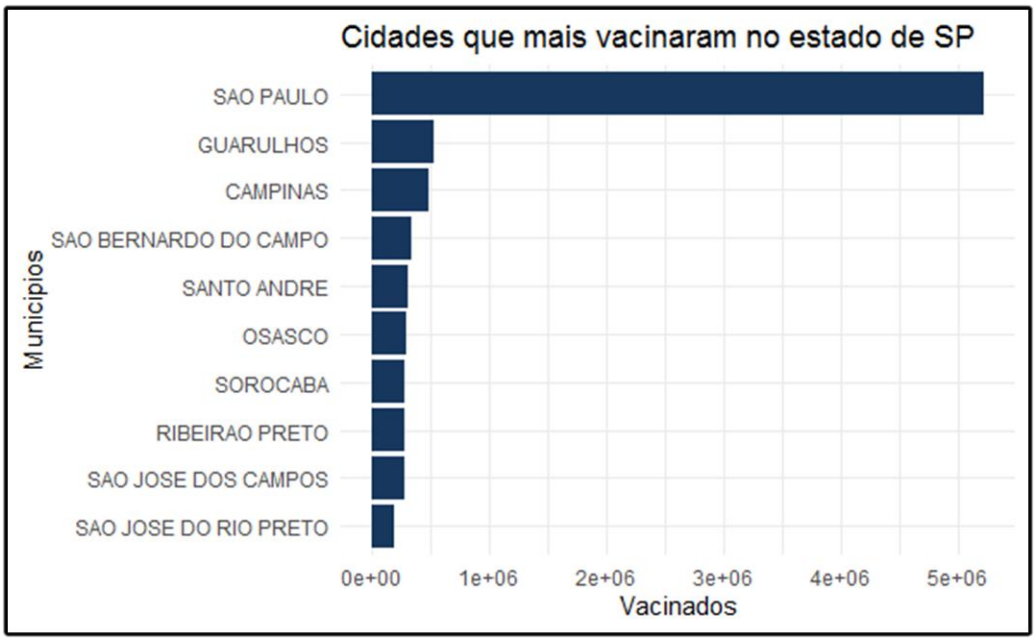
Fonte: Elaborado pelos autores**.**

Figura 6 – As dez cidades que mais vacinaram.

	name_muni	sum(ocupacaocoviduti)
1	São Paulo	34595
2	Campinas	5073
3	Marília	3191
4	Ribeirão Preto	2641
5	Sorocaba	2019
6	Santos	1878
7	Botucatu	1425
8	Presidente Prudente	1032
9	Osasco	983
10	Piracicaba	870

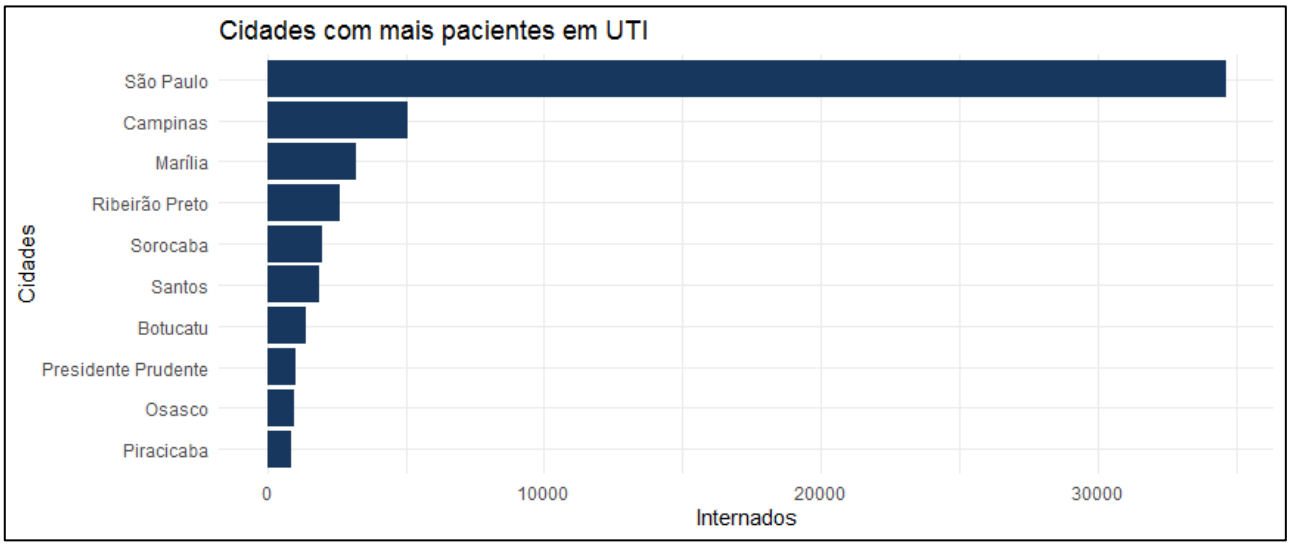
Fonte: Elaborado pelos autores.

Figura 7 – Cidades que mais vacinaram no estado SP.



Fonte: Elaborado pelos autores.

Figura 8 - Cidades que mais ocupação de UTI no estado SP.



Fonte: Elaborado pelos autores.

Com estas informações, também, foram gerados mapas geográficos com a biblioteca de visualização leaflet utilizando os códigos mostrados nas figuras 9 e 10 Os mapas, mostrados nas figuras 11 e 12, trazem, em tons mais escuros, as cidades mais populosas e com mais vacinados, respectivamente. Nota-se que algumas

cidades estão bem mais claras pois a amostra dos dados não comtempla todo o conjunto de dados por motivos citados anteriormente.

Figura 9 – Criação e configuração do mapa.

```
##########################################################################
map <-leaflet(dffinal$geometry)%>%
  addTiles()

map %>% addPolygons(
  weight = 1,
  opacity = 0.5,
  color = "red",
  dashArray = "1",
  fillOpacity = 0
)

####### Definindo cores ###################################
bins <- c(0,10,20,50,100,200,500,1000,Inf)
pal <- colorBin("YlOrRd", domain = dffinal$`sum(x)`, bins = bins)

############ Plotando Mapa ##############################

map %>% addPolygons(
  fillColor = pal(dffinal$`sum(x)`),
  weight = 2,
  opacity = 1,
  color = "white",
  dashArray = "1",
  fillOpacity = 0.7,
  highlight = highlightOptions(
    weight = 5,
    color = "#666",
    dashArray = "",
    fillOpacity = 0.7,
    bringToFront = TRUE))

################ Criando Label ############################

label <- sprintf(
  "<strong>%s</strong></br>%g Vacinados </br>",
  dffinal$Municipio, dffinal$`sum(x)`
)%>%lapply(htmltools::HTML)


################ Plotando Mapa Final ######################
map %>% addPolygons(
  fillColor = pal(dffinal$`sum(x)`),
  weight = 2,
  opacity = 1,
  color = "white",
  dashArray = "1",
  fillOpacity = 0.7,
  highlight = highlightOptions(
    weight = 5,
    color = "#666",
    dashArray = "",
    fillOpacity = 0.7,
    bringToFront = TRUE),
  label = label,
  labelOptions = labelOptions(
    style = list("font-weight" = "normal", padding = "3px 8px"),
    textsize = "15px",
    direction = 'auto'
  ))
```

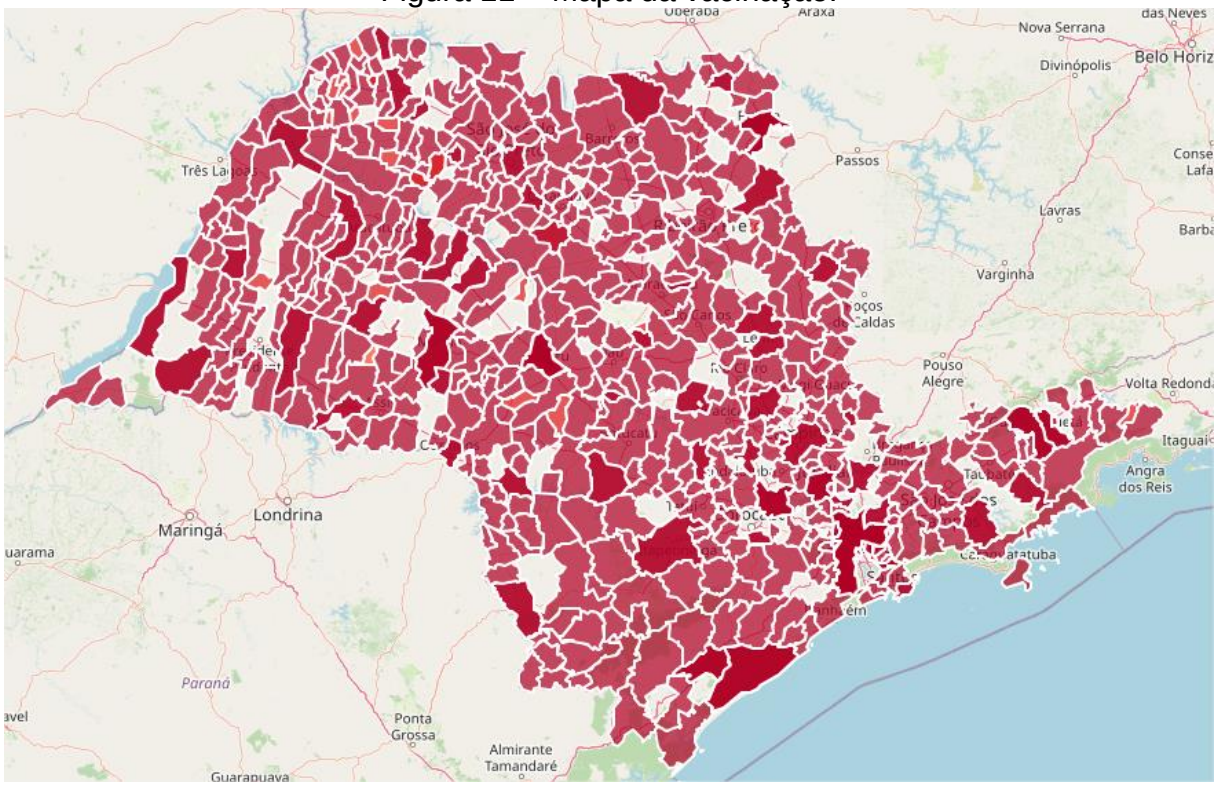
Fonte: Elaborado pelos autores.

Figura 10 – Filtrando os dados do mapa.

```
#### Agrupando os dados por cidade e somando a quantidade de pacientes #########

totais<-dados %>%
          group_by(name_muni) %>%
          summarise(sum(ocupacaocoviduti))

View(totais)


################ PLotando as 10  cidades com maior ocupação no periodo ####################


top_n(totais, n=10, `sum(ocupacaocoviduti)`) %>%
  ggplot(., aes(x= reorder(name_muni,`sum(ocupacaocoviduti)`), y=`sum(ocupacaocoviduti)`))+
  labs(x="Cidades", y="Internados", title="Cidades com mais pacientes em UTI") +
  geom_bar(stat='identity',fill='#17375e')+
  coord_flip()+
  theme_minimal()


```

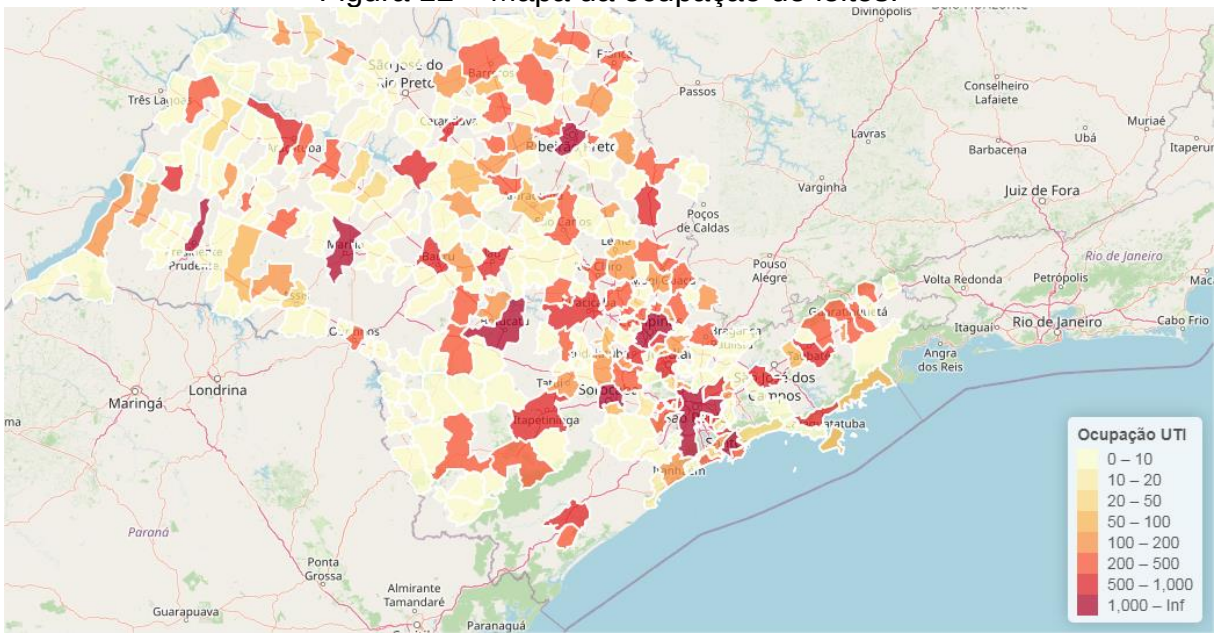
Fonte: Elaborado pelos autores.

Figura 11 – Mapa da vacinação.



Fonte: Elaborado pelos autores.

Figura 12 – Mapa da ocupação de leitos.



Fonte: Elaborado pelos autores.

5 CONSIDERAÇÕES FINAIS

A análise desenvolvida no presente artigo se sustentou em três perguntas, sendo elas: quais cidades do estado mais vacinaram a população, quais cidades mais sofreram com internações devido à COVID-19 e, finalmente, qual o impacto das vacinações nas internações. Além de responder as questões a análise se fundamentou em transformar os dados em informações que podem ser facilmente absorvidas por todos.

As duas primeiras questões foram respondidas e se correlacionam criando um panorama para responder a terceira. Porém, a amostra dos dados utilizados não permitiu um aprofundamento maior na terceira questão, mas os resultados abriram a possibilidade de uma análise mais profunda, pois constatou-se que entre as dez cidades que mais sofreram com internações no ano de 2022, de acordo com informações levantadas neste trabalho, somente quatro se encontram na lista das que mais vacinaram no estado, sendo que todas elas constam na lista de mais populosas do estado.

Portanto, conclui-se que o objetivo foi atingido, trazendo respostas de uma forma acessível a todos e criando um panorama promissor para uma análise mais

profunda na relação da vacinação com a queda da ocupação de leitos de UTI em decorrência de complicações da COVID-19.

REFERÊNCIAS

Pesquisa mostra que investimento em tecnologia na saúde atinge 10 bilhões de reais. Portal Hospitais Brasil,2021. Disponível em: https://portalhospitaisbrasil.com.br/pesquisa-mostra-investimento-em-tecnologia-no-setor-da-saude-atinge-10-bilhoes-de-reais/

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Censo Brasileiro de 2010. São Paulo: IBGE, 2022. Disponível em: https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama. Acesso em: 29 mai. 2022.

Campanha Nacional de Vacinação contra Covid-19. OpenDataSUS, 2022.Disponivel em: https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao

Registro de ocupação hospitalar COVID-19. OpenDataSUS, 2022.Disponivel em: https://opendatasus.saude.gov.br/dataset/registro-de-ocupacao-hospitalar-covid-19