Atividade 02 - Análise Exploratória

Outubro 2020

1 Tema

Proposição de prioridade da distribuição de vacinais de acordo com a vulnerabilidade de cidades por meio da análise de fatores como contaminação, IDH, índices econômicos e demográficos.

2 Equipe

Nome da equipe: DICE-Science
Integrantes:

- Otávio Thomas Bertucini, 2023970, otaviobertuccini, otaviobertucini@gmail.com, bacharelado em Sistemas de Informação, UTFPR
- Luan Carlos Klein, 2022613, luancklein, luanklein@alunos.utfpr.edu.br, bacharelado em Engenharia de Computação, UTFPR
- Gabriel Rauta Buiar, 2022532, gabrielbuiar, gabrielbuiar@gmail.com, bacharelado em Engenharia de Computação, UTFPR

3 Obtenção e processamento de dados

Os dados relativos ao número de casos e total de mortes do COVID 19 foram obtidos na plataforma *Brasil.io*, que por sua vez é um compilado dos dados das secretarias de saúde estaduais. Já os dados socioeconômicos foram obtidos no site Instituto de Pesquisa Econômica Aplicada (IPEA) que também é um compilado dos dados do censo de 2010.

Para o pré-processamento e limpeza dos dados, foram utilizadas diversas técnicas. Inicialmente foram retiradas algumas colunas da base de dados proveniente do *Brasil.io*, permanecendo apenas as colunas: estado, cidade, tipo do lugar, ultima quantidade confirmada de casos, quantidade de mortes, taxa de mortalidade e a estimativa do tamanho da população. Após isso, foi feita a verificação se não haviam dados faltantes, e constatou-se que haviam 19 linhas que haviam essa carência. Entretanto, isso se deve pela definição da linha em

si, na qual o nome da "cidade" é "Importados/Indefinidos". Para tal resolução, esses dados foram retirados.

Para a utilização dos dados referentes as cidades, tais como renda e IDH, foi utilizado os dados provenientes do Censo, realizado pela ultima vez pelo IBGE no ano de 2010. Nele também limitamos as colunas que iremos usar, que são: Nome do estado, município, expectativa de vida, porcentagem de extremamente pobres, renda per capita e IDH dos municípios. Além disso, foi necessário transformar os dados que estão em formato de *string* para *float* (nas colunas numéricas). Nessa base de dados não haviam dados faltantes. Para realizar a integração entre essas duas bases, foi necessário utilizar uma terceira base, proveniente do site de *basedosdados.github.io* (que posteriormente foi colocado no repositório de um dos integrantes para o acesso, visto que os dados presentes ali são fixos), pois os códigos do IBGE das cidades utilizados nas duas bases eram distintos, e essa terceira base fornece a ligação entre os dois códigos. Após isso, foi realizado o *merge* entre os três *datasets*, e dessa maneira os dados ficaram prontos para a análise. A quarta base de dados explorada foi a das ocupações hospitalares, na qual apresenta os dados do dia atual dos dados referentes as unidades hospitalares, como quantidade de internações devido a Síndrome Respiratória Aguda Grave (SRAG), vas disponíveis para isso, além de outros dados relevantes. Os dados foram coletados através do *elasctisearch*. Para o tratamento dos mesmos, foram removidas as colunas em branco (pois foi constatado que eles não faziam referencia a uma unidade hospitalar especifica), o que faz com que esteja fora do foco atual da análise. Os demais campos já estavam todos corretos, e não foi necessário nenhuma nova modificação.

A seguir, a fonte utilizada para cada a obtenção de cada *dataset*:

- Casos: https://brasil.io/api/v1/dataset/covid19/caso_full/data
- Censo: https://www.ipea.gov.br/ipeageo/arquivos/bases/IDH_2010.xls
- Internações: https://elastic-leitos.saude.gov.br/leito_ocupacao/_search

4 Cobertura e distribuição dos dados

Os histogramas da renda per capita e do IDH municipal serão analisados com o histograma de casos por população para verificar se há alguma correlação entre os fatores. Considerando esses três fatores espera-se poder criar um índice que indica a prioridade que a cidade terá para receber a vacina.

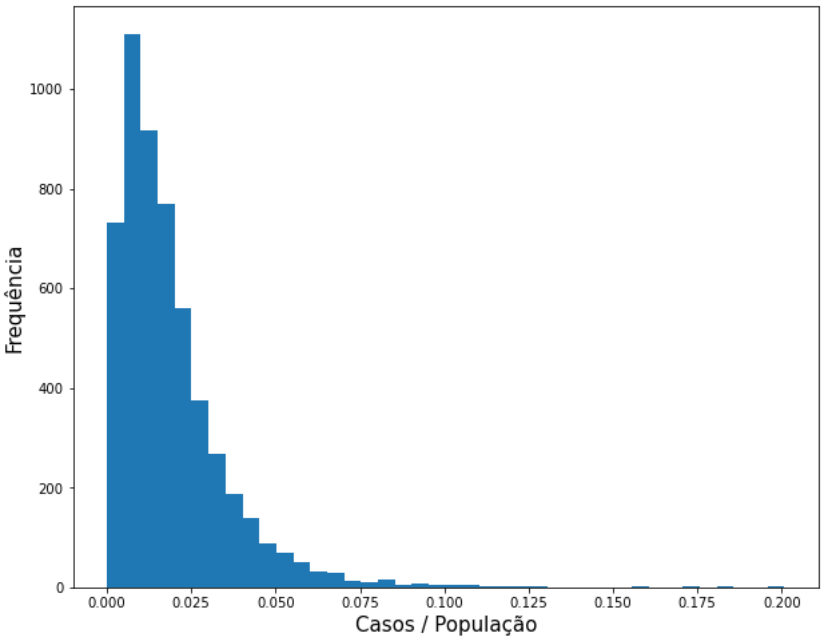

Figure 1: Histograma do número de casos pela população das cidades brasileiras

O primeiro índice demográfico que foi considerado é a renda *per capita*, com a análise desse histograma com o número de casos por população será possível explorar o quanto a renda *per capita* de cada cidade afeta a contaminação das pessoas, e se esse índice de fato está correlacionado com o número de casos.

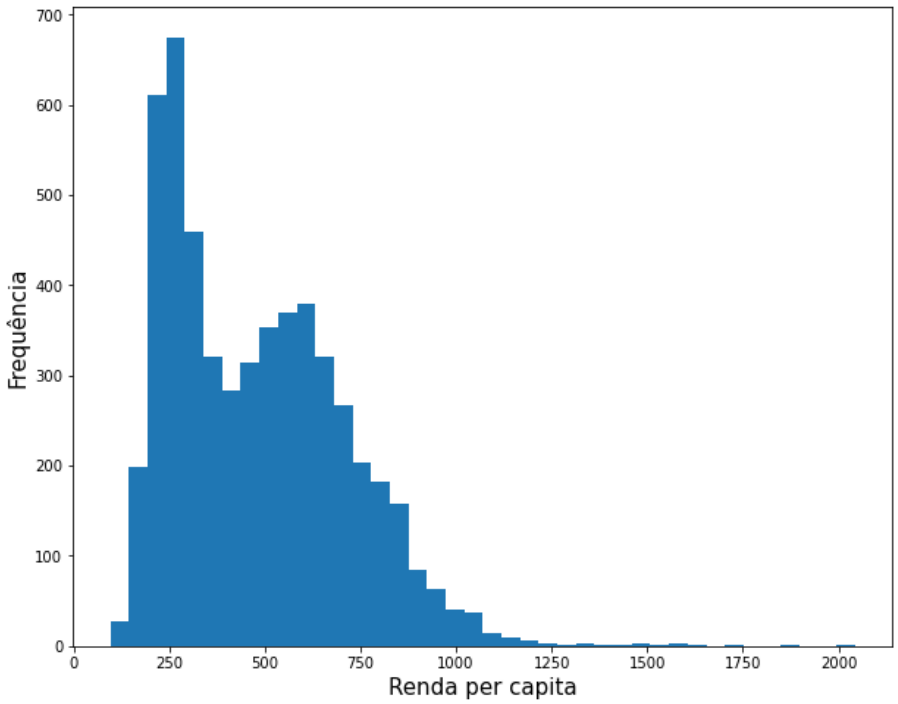

Figure 2: Histograma da Renda per capita das cidades brasileiras

O segundo índice demográfico é o IDH municipal mostrado no *boxplot* da Figura 3, assim como a renda *per capita*, com a análise desse índice será possível analisar o quanto o IDH dos municípios está relacionado ao número de casos e se ele é um fator relvante para considerar na criação do índice de prioridade para a distribuição da vacina.

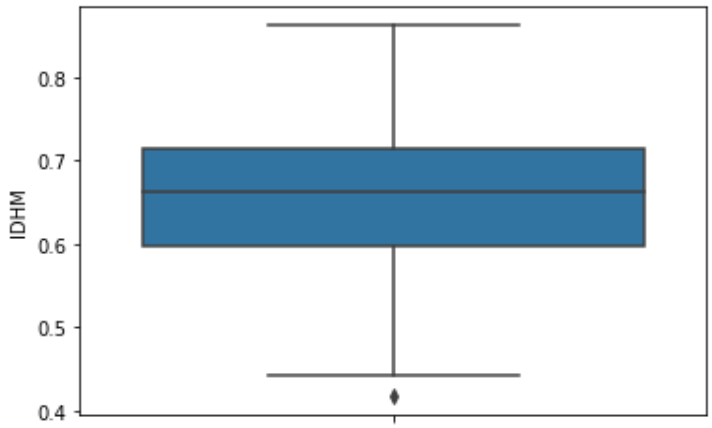

Figure 3: Boxplot dos IDHM das cidades brasileiras

Também foi observado a quantidade de internações devido a SRAG dentro de cada unidade hospitalar. Esse dado pode ser fundamental na análise de vulnerabilidade visto que ele indica como está sendo um dos aspectos que influencia na "gravidade" do covid-19 naquela região, além de saber qual a capacidade dos hospitais e atual situação dos mesmos. Na Figura 4, mostra-se o histograma das internações na UTI nas unidades hospitalares brasileiras.

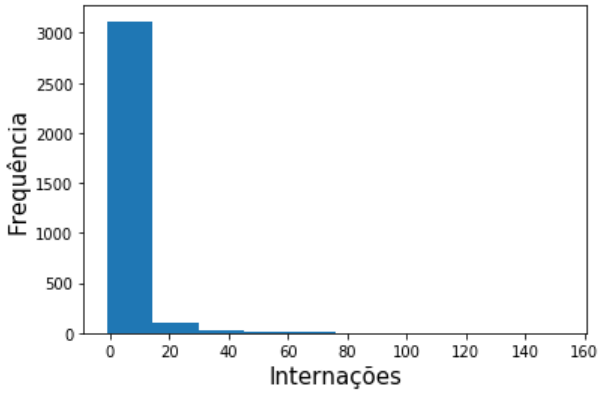

Figure 4: Histograma das internações devido a SRAG

5 Perguntas de pesquisa e explorações iniciais

A pesquisa tem propósito inicial responder as perguntas:

- Quais cidades apresentam maior vulnerabilidade em termos dos impactos na saúde referente ao coronavírus?

- Quais fatores podem influenciar nessa vulnerabilidade: IDH, densidade demográfica, PIB e renda *per capita*?

- É possível sugerir um índice que visa a facilitação na distribuição efetiva da vacina?

Diante das perguntas acimas, diversos gráficos simples foram geradas, criando algumas correlações entre os dados. Diante dos resultados, como apontam os gráficos das Figuras 5, 6 e 7 (*scatters plot* do IDH, renda *per capita* e PIB pela quantidade de casos confirmados, respectivamente), foi possível perceber, de maneira breve, que alguns aspectos tem uma alta correlação enquanto outros tem baixa. Porém, de maneira geral, acredita-se que será possível responder as perguntas descritas acima (com as possíveis limitações que os dados do CENSO desatualizado ocasionam).

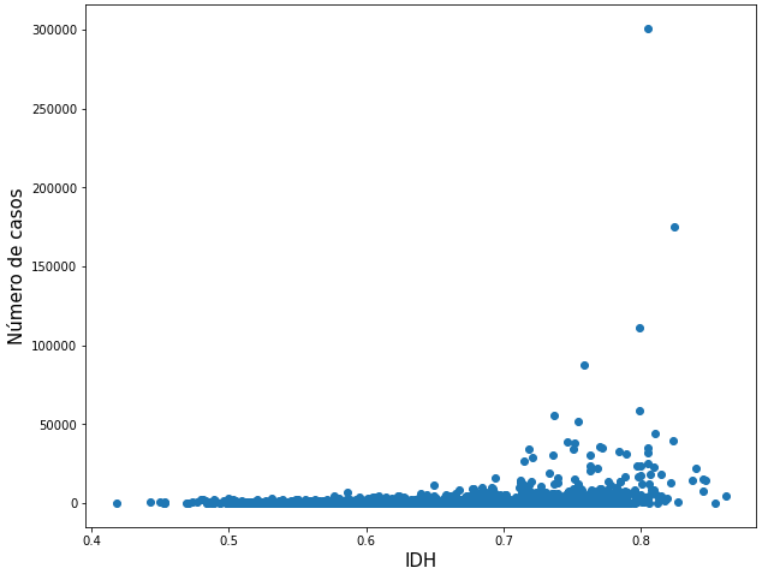

Figure 5: Scatter plot do IDH por casos de COVID 19 das cidades brasileiras

Um ponto de especial atenção foi observado na taxa de internações nos leitos de uti devido ao SRGA. No dia 21 de Outubro de 2020, mais de 100 unidades hospitalares apresentaram uma taxa superior a 1 (mais de uma internação por vaga disponível) indicando assim, uma possível superlotação do hospital. Um boxplot dessas cidades é apresentado na Figura 8. Vale destacar ainda que a maior taxa chegou a 7 (7 internações e 1 vaga disponível).

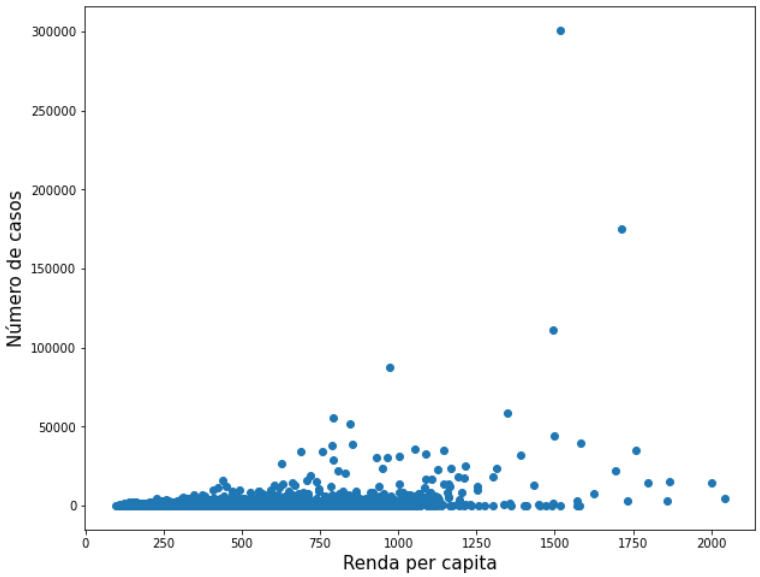

Figure 6: Scatter plot da renda per capita por casos de COVID 19 das cidades brasileiras

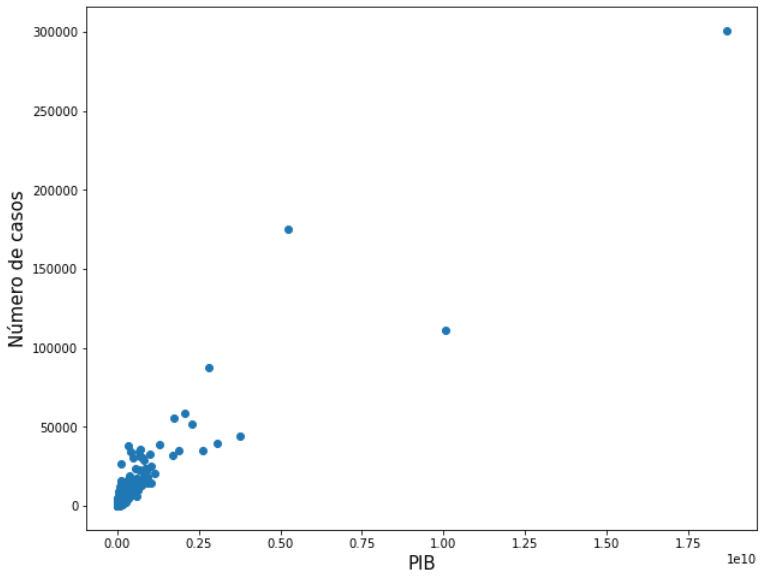

Figure 7: Scatter plot do PIB por casos de COVID 19 das cidades brasileiras

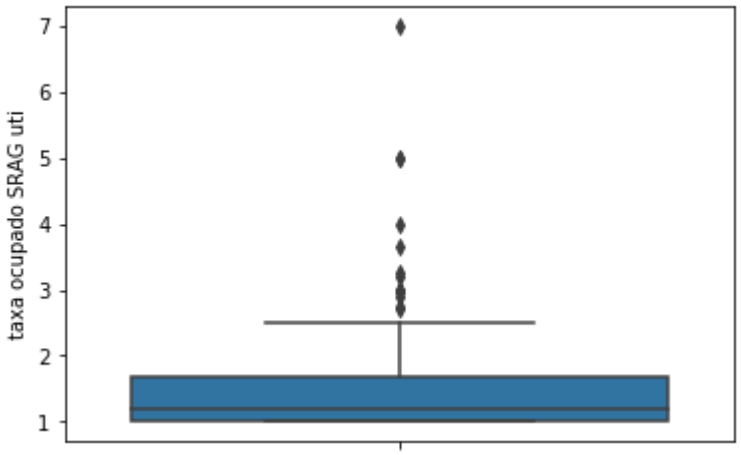

Figure 8: Taxa de internações devido a SRAG no tipo UTI em unidades com mais de 100% de ocupação

6 Discussão e próximos passos

Com a análise dos dados realizada até o momento, constatou-se que a maioria dos dados apresentou uma boa completude (nenhum dado fundamental estava faltando). Um dos pontos que merece atenção a falta de atualização dos dados do Censo do IBGE, que foi realizado pela ultima vez no ano de 2010. Estava previsto para uma nova realização do Censo no neste ano de 2020. Entretanto, acredita-se que os indicadores como Renda per capita e IDH mesmo estando desatualizados ainda podem trazer resultados satisfatórios.

Além disso, a análise exploratório feita no primeiro momento apresentou diversos resultados que podem auxiliar nas respostas as perguntas, como por exemplo, uma alta correlação entre o PIB das cidades e o número de casos na mesma, enquanto que outros fatores aparentemente foram quase irrelevantes em termos de correlação, como por exemplo renda per capita e IDH. Os indicadores da ocupação hospitalar também se demonstraram um elemento promissor na definição de um vulnerabilidade das cidades.

O grupo almeja nos próximos passos, encontrar de maneira mais satisfatória as correlações entre os dados, além de incluir novos indicadores, como índices de pobrezas nas cidades, e correlacionar também os índices de mortes nas cidades. Outros dados provavelmente deverão ser buscados pela equipe, visando outros indicadores que podem contribuir para a definição de um indicide de vulnerabilidade.