Analise Exploratória dos Dados

Andressa Marçal
a262878@dac.unicamp.br

Décio Gonçalves
d226072@dac.unicamp.br

Diego Alyson
d230640@dac.unicamp.br

I. Introdução

Este documento contém os dados e gráficos coletados das analises feitas para o Trabalho 02 da disciplina de Ciência de Dados e suas Aplicações em Cenários de Pandemia, disciplina ofertada pela Unicamp em parceria com professores da UTFPR no semestre de 2020.2.

II. Obtenção e processamento de dados

Nesta seção apresentamos as bases de dados utilizadas na etapa de exploração. No total exploramos dados provenientes de três bases de dados com o objetivo de obter dados de diferentes fontes que possam ser utilizados para validar hipóteses que surjam ao longo do trabalho. Para alcançar o objetivo exploratório de obter informações de possíveis correlações entre o clima e a propagação da Covid-19 utilizamos as bases INMET, Brasil.io e Google Mobility. Estas bases, respectivamente, contêm dados metereológicos, do avanço da covid-19 no e de mobilidade, todos referentes ao Brasil.

A. *Dataset INMET (Instituto Nacional de Meteorologia)*

Com o intuito de obter dados meteorológicos, como variação de temperatura e umidade, para responder os questionamentos iniciais do trabalho, a variação de casos e o impacto do clima, temperatura e umidade nesta variação, buscou-se diferentes datasets climáticos das regiões brasileiras. Nesta primeira busca, verificou-se que em sua grande maioria os dados relacionados à clima, nacional e internacional são disponibilizados de forma privada e paga. No INMET (Instituto Nacional de Meteorologia) obteve-se dados gratuitos de históricos em território nacional, por campos de estações meteorológicas do instituto, de acordo com a Imagem 1.

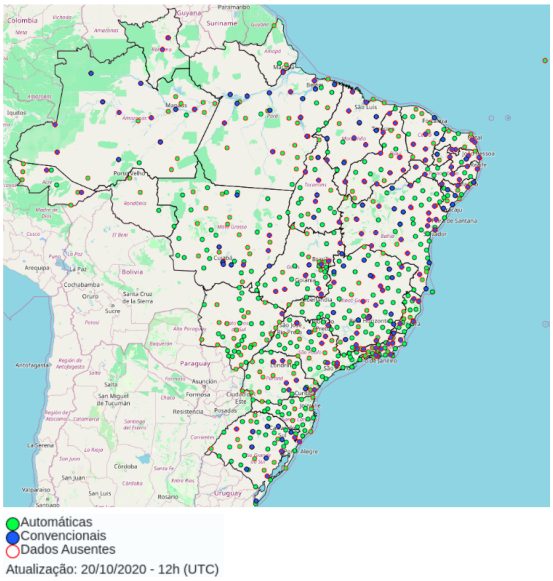

Fig. 1. Estações Meteorológicas Brasileiras.

Neste trabalho propomos realizar um estudo dos impactos meteorológicos de âmbito nacional e, como observações, algumas cidades selecionadas, considerando-se os diferentes climas e taxas de umidade do país. A base de dados do INMET disponibiliza estes dados por cidade e por histórico temporal, por hora, possibilitando sua utilização para esta pesquisa. Os são disponibilizados através da plataforma portal.inmet.gov.br, no formato zip. As informações são subdivididas em vários arquivos .CSV, separadas por cada estação disponível pelo INMET, como observado na imagem anterior. Devido esta grande segmentação, do número de cidades e diferentes horários, o primeiro tratamento necessário é a integração destas diferentes fontes em uma única base de informação.

Para realizar a integração, levou-se em consideração três diferentes níveis de observação. Em um primeiro nível, as informações de clima nacional, unificando os dados regionais e obtendo-se uma média nacional de mudanças meteorológicas. No segundo nível proposto, observou-se os dados regionalmente, por estados. Foram listados os 26 estados nacionais e o Distrito Federal. Por último, para uma melhor observação de localidade, foram selecionadas algumas diferentes cidades, com climas diferenciados.

Um segundo tratamento nos dados recebidos é a verificação dos seus tipos e observação de dados faltantes. Em uma primeira observação, verificou-se que as informações recebidas estavam em um formato diferente do necessário para a futura correlação. observa-se na imagem 2, um exemplo dos tipos de dados recebidos para uma única cidade.

Somente após este primeiro estudo e a realização das correções realizadas foi possível iniciar a observação de cobertura e distribuição. Segue na Imagem 3 uma observação geral dos dados tratados para a mesma cidade observada anteriormente.

Como estudo inicial, os gráficos temporais climáticos foram observados. Na imagem 4, a temperatura de bulbo seco, realiza através de termômetros comuns, é demonstrada em relação ao tempo para a média brasileira. Considerou-se inicialmente os dados de temperatura a partir do início da proliferação do vírus no Brasil até dia 30 de setembro.

Pode-se observar, através do gráfico apresentado grandes variações de temperatura durante o ano, decorrente das

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6576 entries, 0 to 6575
Data columns (total 20 columns):
 #   Column                                                 Non-Null Count  Dtype
---  ------                                                 --------------  -----
 0   Data                                                   6576 non-null   object
 1   Hora UTC                                               6576 non-null   object
 2   PRECIPITAÇÃO TOTAL, HORÁRIO (mm)                       6575 non-null   object
 3   PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)  6575 non-null   object
 4   PRESSÃO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)        6575 non-null   object
 5   PRESSÃO ATMOSFERICA MIN. NA HORA ANT. (AUT) (mB)       6575 non-null   object
 6   RADIACAO GLOBAL (Kj/m²)                                3459 non-null   object
 7   TEMPERATURA DO AR - BULBO SECO, HORARIA (°C)           6575 non-null   object
 8   TEMPERATURA DO PONTO DE ORVALHO (°C)                   6569 non-null   object
 9   TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)             6575 non-null   object
 10  TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)             6575 non-null   object
 11  TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (°C)       6575 non-null   object
 12  TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (°C)       6575 non-null   object
 13  UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)               6575 non-null   float64
 14  UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)               6575 non-null   float64
 15  UMIDADE RELATIVA DO AR, HORARIA (%)                    6569 non-null   float64
 16  VENTO, DIREÇÃO HORARIA (gr) (° (gr))                   6574 non-null   float64
 17  VENTO, RAJADA MAXIMA (m/s)                             6575 non-null   object
 18  VENTO, VELOCIDADE HORARIA (m/s)                        6575 non-null   object
 19  Unnamed: 19                                            0 non-null      float64
dtypes: float64(5), object(15)
memory usage: 1.0+ MB
```

Fig. 2. Tipos de Dados Recebidos, Pré-Processamento.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 274 entries, 0 to 273
Data columns (total 18 columns):
 #   Column                                                 Non-Null Count  Dtype
---  ------                                                 --------------  -----
 0   Data                                                   274 non-null    period[D]
 1   PRECIPITAÇÃO TOTAL, HORÁRIO (mm)                       274 non-null    float64
 2   PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)  274 non-null    float64
 3   PRESSÃO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)        274 non-null    float64
 4   PRESSÃO ATMOSFERICA MIN. NA HORA ANT. (AUT) (mB)       274 non-null    float64
 5   RADIACAO GLOBAL (Kj/m²)                                274 non-null    float64
 6   TEMPERATURA DO AR - BULBO SECO, HORARIA (°C)           274 non-null    float64
 7   TEMPERATURA DO PONTO DE ORVALHO (°C)                   274 non-null    float64
 8   TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)             274 non-null    float64
 9   TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)             274 non-null    float64
 10  TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (°C)       274 non-null    float64
 11  TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (°C)       274 non-null    float64
 12  UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)               274 non-null    float64
 13  UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)               274 non-null    float64
 14  UMIDADE RELATIVA DO AR, HORARIA (%)                    274 non-null    float64
 15  VENTO, DIREÇÃO HORARIA (gr) (° (gr))                   274 non-null    float64
 16  VENTO, RAJADA MAXIMA (m/s)                             274 non-null    float64
 17  VENTO, VELOCIDADE HORARIA (m/s)                        274 non-null    float64
dtypes: float64(17), period[D](1)
memory usage: 38.7 KB
```

Fig. 3. Tipos de Dados Recebidos, Pós-Processamento

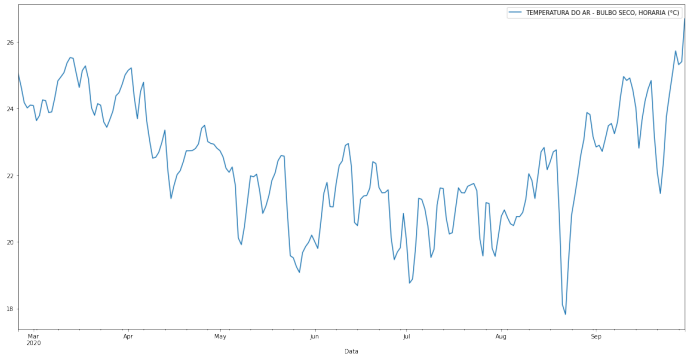

Fig. 4. Média de Temperatura Brasileira.

estações correntes e fatores externos. Em relação às estações, neste período, o país passou por duas estações diferentes e está iniciando uma terceira. Entre Março e Junho o outono, com variações de temperaturas entre 25ºC e um início de decréscimo na mesma. Em junho, início do inverno, as temperaturas já se encontram amenas, até o término da estação em setembro, início da primavera.

Como exemplo de fatores externos, as queimadas no pantanal no mês de setembro causaram um grande aumento na temperatura do país, culminando no fim da análise, com máxima de 26.69ºC para média nacional.

Como mencionado anteriormente, também foram realizadas análises em relação a umidade, para verificar a possível correlação com o contágio. A primeira análise, assim como na temperatura foi a média da umidade nacional em relação ao tempo na mesma janela anteriormente comentada, como observado na Imagem 5.

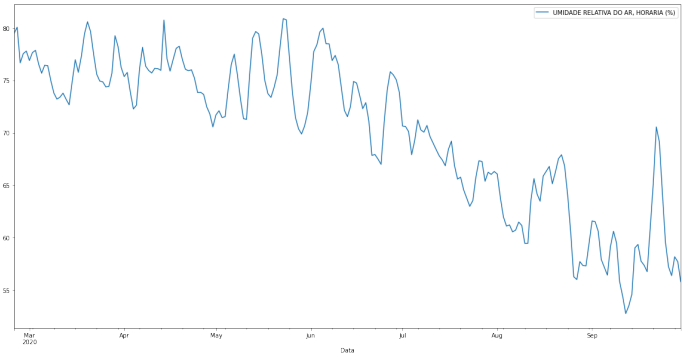

Fig. 5. Média da Umidade Relativa do Ar Brasileiro.

Os mesmo fatores para mudança na temperatura podem ser observados para variação da umidade, relação de clima por época do ano e os fatores externos que impactam na curva observada.

Também foram estudados a variação dos valores dentro do tempo estipulado. Para realizar esta observação, utilizou-se gráficos box plots, desta forma, pode-se observar a variação dos valores e possíveis outliers nos dados de entrada. Foram traçados dois gráficos distintos para a observação nacional, visto que os valores de temperatura e umidade estão em patamares diferentes. Segue a imagem dos valores no gráfico da Imagem 6.

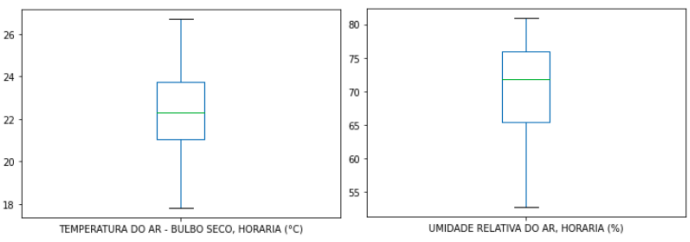

Fig. 6. Gráficos em Box Plot da Temperatura e Umidade Média - Brasil.

Estes resultados obtidos anteriormente para os valores nacionais, da média da variação da temperatura e umidade no

tempo, assim como os gráficos de box plot destas variações podem ser também gerados e observados para cada um dos 26 estados e Distrito Federal. Foram escolhidos devido cinco diferentes estados considerando suas grandes diferenças climáticas, tanto em temperatura quando umidade relativa do ar. As imagens 7 e 8

Fig. 7. Temperatura dos Cinco Estados Selecionados.

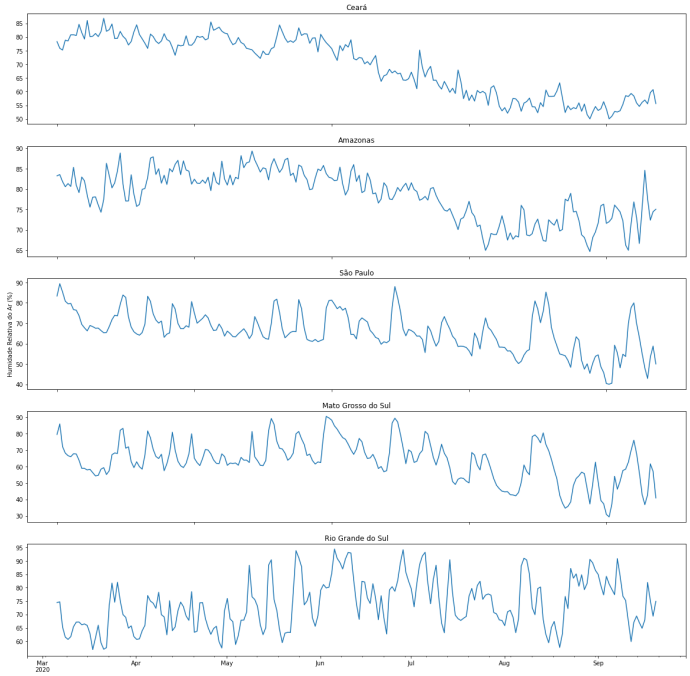

Fig. 8. Umidade Média dos Cinco Estados Selecionados.

O trabalho também suporta e observa estes valores a nível cidade, porém devido a grande quantidade de cidades possíveis, foram selecionadas algumas como valores amostrais por região do país, com variações de clima entre si.

Como demonstração dos resultados obtidos para estado e cidade, os gráficos da imagem 9 representam os valores de uma das cidades selecionada, neste caso são observações de Fortaleza-CE para temperatura média.

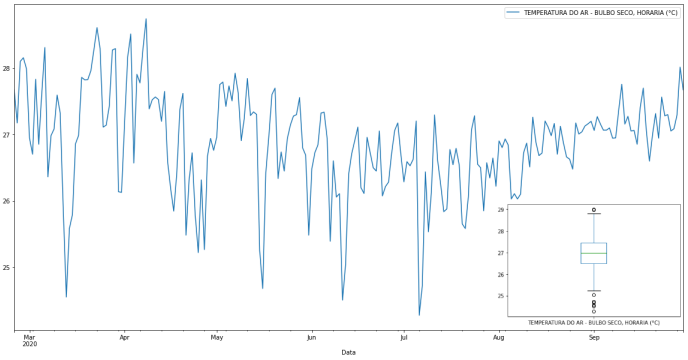

Fig. 9. Temperatura Média na Cidade de Fortaleza-CE.

B. *Dataset Covid-19 (Brasil.IO)*

Decidimos usar o dataset de Covid-19 do site Brasil.IO, o dataset possui os casos confirmados e óbitos obtidos dos boletins das Secretarias Estaduais de Saúde (SES). Os dados foram enriquecidos, de forma que a partir do momento em que um município confirma um caso, ele sempre aparecerá nessa tabela (mesmo que para uma determinada data a SES não tenha liberado o boletim - nesse caso é repetido o dado do dia anterior). A coleta dos dados iniciaram do dia 25 de Fevereiro de 2020. Temos nele 27 estados e 5.294 cidades. Na imagem abaixo podemos observar as informações das suas colunas e os tipos dos seus dados.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 925921 entries, 0 to 925920
Data columns (total 18 columns):
 #   Column                                  Non-Null Count   Dtype
---  ------                                  --------------   -----
 0   city                                    919869 non-null  object
 1   city_ibge_code                          922174 non-null  float64
 2   date                                    925921 non-null  object
 3   epidemiological_week                    925921 non-null  int64
 4   estimated_population                    922174 non-null  float64
 5   estimated_population_2019               922174 non-null  float64
 6   is_last                                 925921 non-null  bool
 7   is_repeated                             925921 non-null  bool
 8   last_available_confirmed                925921 non-null  int64
 9   last_available_confirmed_per_100k_inhabitants  908366 non-null  float64
 10  last_available_date                     925921 non-null  object
 11  last_available_death_rate               925921 non-null  float64
 12  last_available_deaths                   925921 non-null  int64
 13  order_for_place                         925921 non-null  int64
 14  place_type                              925921 non-null  object
 15  state                                   925921 non-null  object
 16  new_confirmed                           925921 non-null  int64
 17  new_deaths                              925921 non-null  int64
dtypes: bool(2), float64(5), int64(6), object(5)
memory usage: 114.8+ MB
```

Fig. 10. Informações do Dataset Covid19

Nas observações iniciais, podemos ver que no dataset há dados inconsistentes, que em uma próxima etapa deverão ser tratados e balanceados para melhorar os resultados. A figura

abaixo mostra os dados ausentes no dataset. Também foi observado a presença de valores negativos no dataset, como por exemplo, o valor -1 na coluna de registro de novas mortes.

```
city                                          6052
city_ibge_code                                3747
date                                             0
epidemiological_week                             0
estimated_population                          3747
estimated_population_2019                     3747
is_last                                          0
is_repeated                                      0
last_available_confirmed                         0
last_available_confirmed_per_100k_inhabitants    17555
last_available_date                              0
last_available_death_rate                        0
last_available_deaths                            0
order_for_place                                  0
place_type                                       0
state                                            0
new_confirmed                                    0
new_deaths                                       0
dtype: int64
```

Fig. 11. Dados ausentes/nulos contidos no dataset

Na figura 12 foi feita uma analise em uma semana epidemiológica especifica referente ao feriado prolongado que tivemos no mês de Junho. Investigamos os casos confirmados 7 dias antes do feriadão e 7 e 15 dias após o feriadão Corpus Christi. Perante o gráfico, é notório um comportamento acentuado nos dias 19/06, data essa que coincide 8 dias após o feriadão. Os dias 23, 24, 25 e 26 também apresentaram alto crescimento de casos, essas datas são referentes a segunda semana após o feriadão.

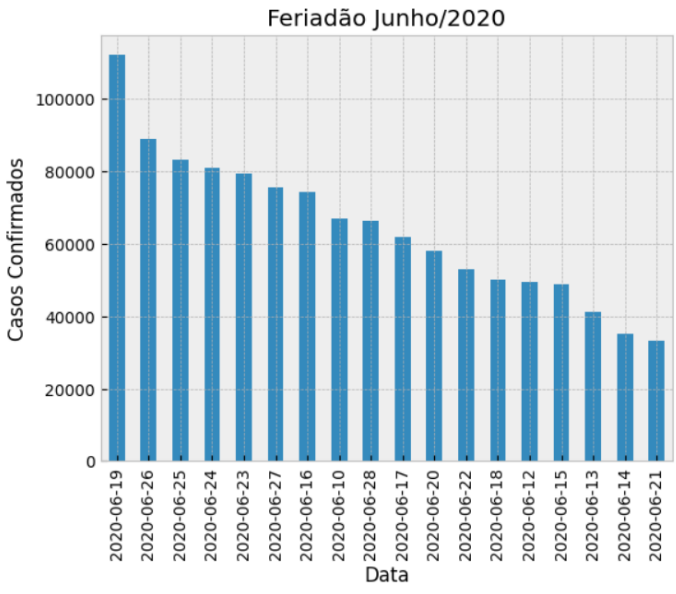

Fig. 12. Novos casos confirmados após o feriadão de Corpus Christi

É possível observar um crescimento acentuado de novos casos confirmados no Brasil, na data 19/06/2020, exatos 8 dias após o inicio do feriadão de Corpus Christi, seguido do segundo maior crescimento no dia 26/06/2020, 15 dias após o feriado prolongado.

Investigamos os dados dos casos confirmados nas 5 principais cidades do estado de São Paulo, após a reabertura do comercio. Como mostra na figura 13.

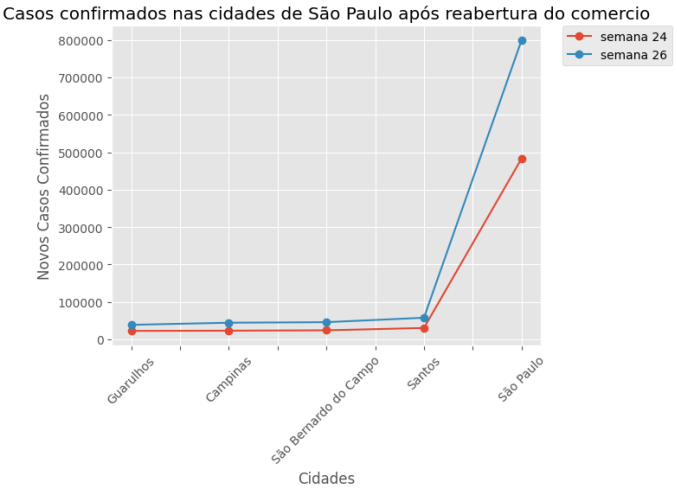

Fig. 13. Casos confirmados após uma e duas semanas da reabertura do comercio no estado de São Paulo

Média global atual dos novos casos confirmados nas regiões do Brasil. Podemos observar que as regiões Centro Oeste e Norte são as que contem os maiores registros.

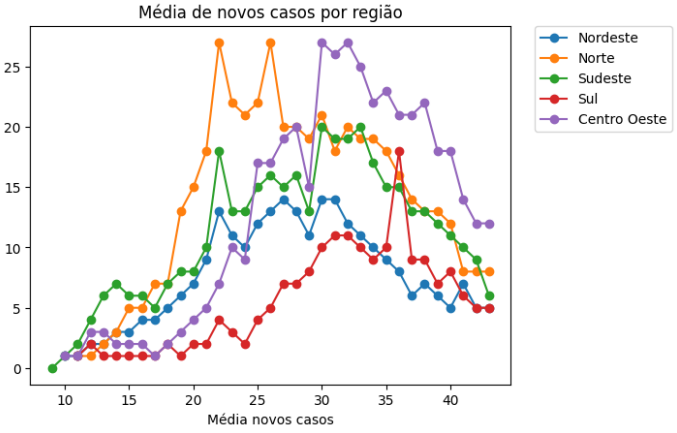

Fig. 14. Média de crescimento atual de novos casos nas regiões do brasil

Estados de cada região que mais teve mortes pelo covid-19.

Na figura 20 podemos visualizar o atual top5 estados brasileiros que mais tem casos confirmados de covid-19 até a data de hoje. E o top 5 estados brasileiros com mais mortes por covid-19 na figura 22

Na figura abaixo podemos ver um grafico

C. *Dataset Google Mobility (Brasil.IO)*

Como forma de identificar possíveis tendências de comportamento como a relação entre o isolamento social e a quantidade de casos optamos por analisar dados de mobilidade.

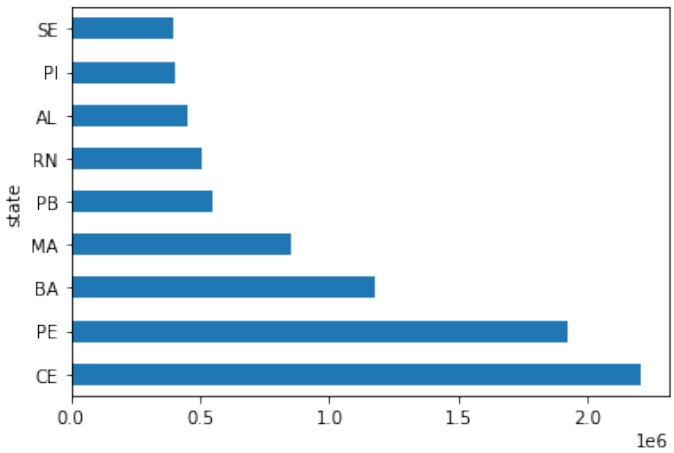

Fig. 15. Mortes por Covid nos estados da região Nordeste

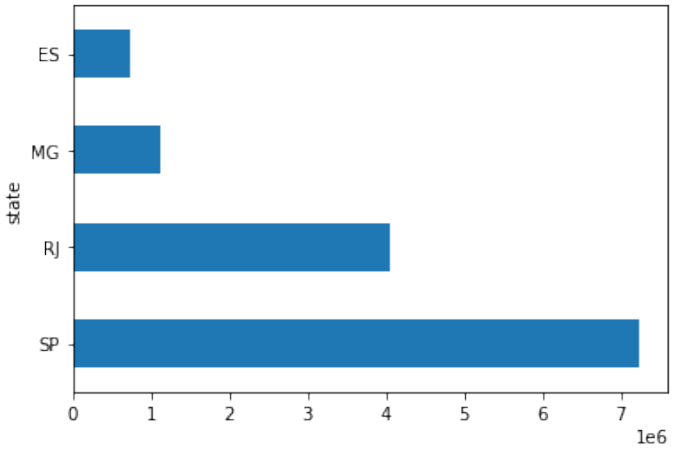

Fig. 18. Mortes por Covid nos estados da região Sudeste

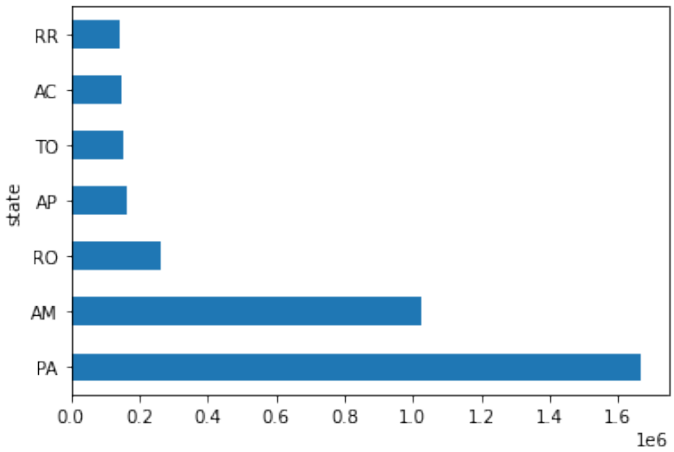

Fig. 16. Mortes por Covid nos estados da região Norte

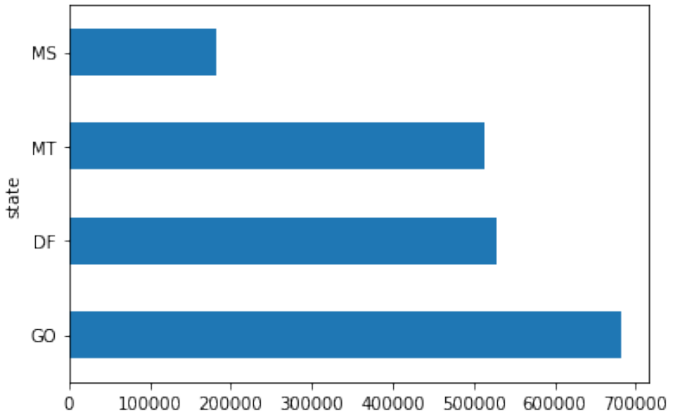

Fig. 19. Mortes por Covid nos estados da região Centro Oeste

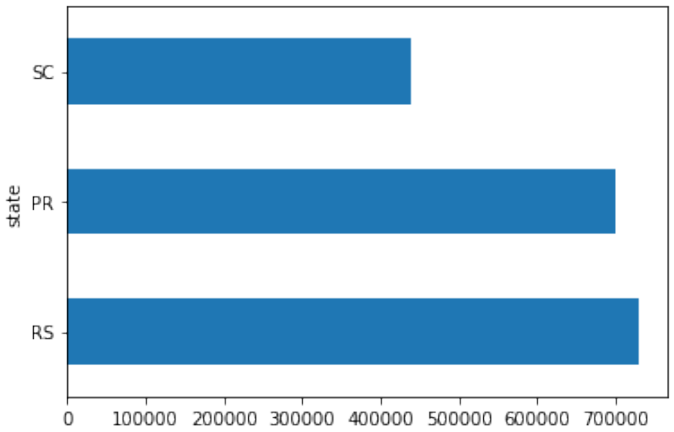

Fig. 17. Mortes por Covid nos estados da região Sul

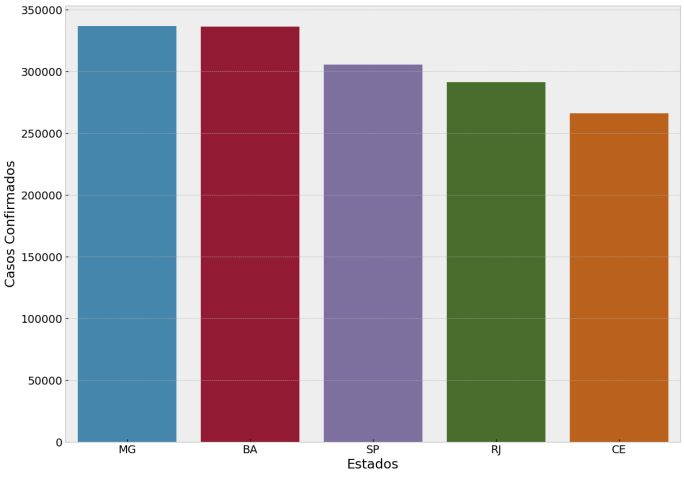

Fig. 20. Top5 dos Estados com mais casos confirmados de Covid-19

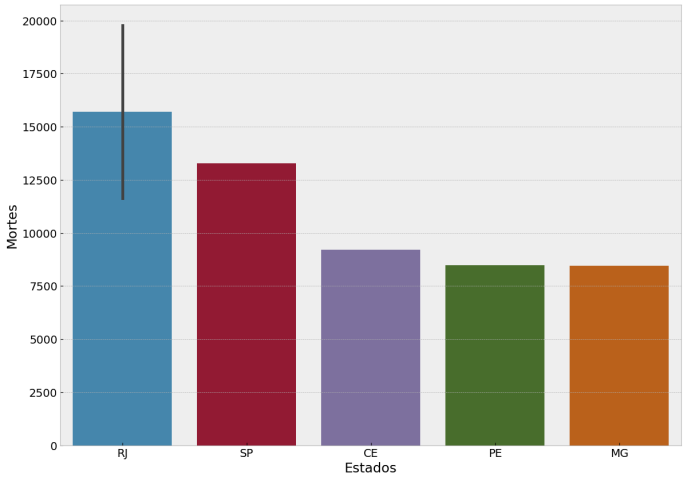

Fig. 21. Top5 dos Estados com mais mortes confirmadas por Covid-19

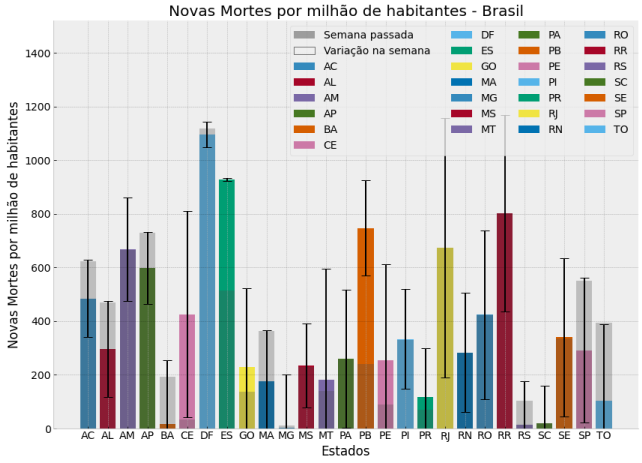

Fig. 22. Novas mortes no Brasil por cada Milhão de Habitantes

Para isso escolhamos utilizar a base Google Mobility[1]. A Figura 24 apresenta as características que a base fornece.

O Google mobility contêm dados que reportam um valor de referência com base na mediana do período pré pandemia. Para entender essa base temos a Figura [**?**], no eixo X temos os dias e no eixo Y temos um valor em porcentagem refente ao quanto o valor está distante do valor de referência. Neste mesmo exemplo podemos observar a tendência seguida pelo Brasil no período em que foi declarado o estado pandêmico, é possível observar que é neste período em que as pessoas estiveram mais tempo em suas casas. Também é possível observar que este índice vem decaindo e está tendendo a se tornar próximo aos valores em períodos normais. Contundo é importante analisar mais informações para tirar conclusões do comportamento dessas curvas.

[1]https://www.google.com/covid19/mobility/

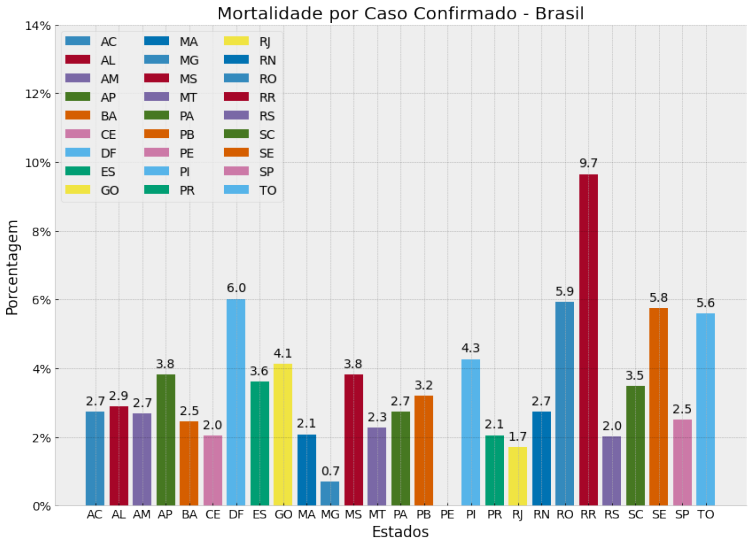

Fig. 23. Porcentagem da taxa de mortalidade nos estados brasileiros

```
RangeIndex: 430432 entries, 0 to 430431
Data columns (total 14 columns):
 #   Column                                                Non-Null Count   Dtype
---  ------                                                --------------   -----
 0   country_region_code                                   430432 non-null  object
 1   country_region                                        430432 non-null  object
 2   sub_region_1                                          430185 non-null  object
 3   sub_region_2                                          423516 non-null  object
 4   metro_area                                            0 non-null       float64
 5   iso_3166_2_code                                       6669 non-null    object
 6   census_fips_code                                      0 non-null       float64
 7   date                                                  430432 non-null  object
 8   retail_and_recreation_percent_change_from_baseline    182567 non-null  float64
 9   grocery_and_pharmacy_percent_change_from_baseline     177081 non-null  float64
 10  parks_percent_change_from_baseline                    156331 non-null  float64
 11  transit_stations_percent_change_from_baseline         132303 non-null  float64
 12  workplaces_percent_change_from_baseline               404785 non-null  float64
 13  residential_percent_change_from_baseline              179990 non-null  float64
```

Fig. 24. Informações do Dataset Google Mobility

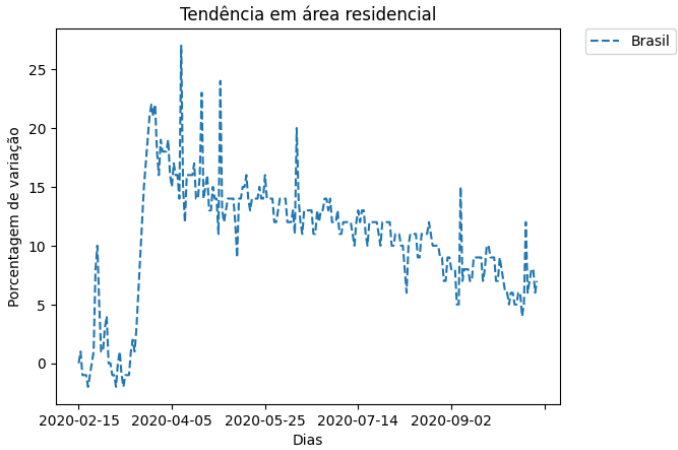

Fig. 25. Tendência em porcentagem que os brasileiros tem ficado em casa

Em nossa primeira análise buscamos filtrar à base em busca de identificar quais os períodos onde as pessoas menos respeitaram o isolamento. O resultado obtido pode ser visto na Figura [**?**]. Vemos que no período próximo ao dia 5 de abril o isolamento foi cumprido com maior rigor. Contudo devemos também analisar o que estava ocorrendo no período reportado. Uma outra informação relevante é identificar as datas onde ocorreram os valores de mínimo e máximo de cada estado e investigar quais os eventos que podem ter influenciado nesses valores.

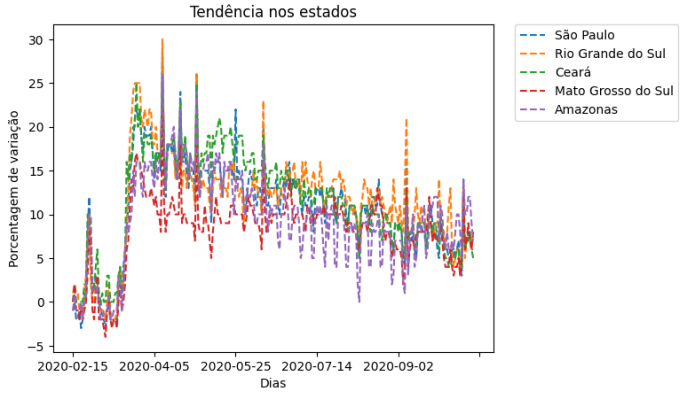

Fig. 26. Tendência em porcentagem que os estados escolhido estão em casa

Como próximos passos pretendemos identificar as datas onde ocorreu maior isolamento não somente baseado no período em que as pessoas estiveram em casa mas também quando estiveram se deslocando para outros estabelecimentos ou utilizando transportes públicos. Junto a uma busca por informações do que estava ocorrendo no Brasil e nestes estados analisados no período reportado.

III. Conclusão

Em nossa análise exploratória conseguir ter o primeiro contato com os dados que pretendemos utilizar ao longo da realização do projeto. A principio eles parecem conter uma quantidade importante de informações úteis e que podem nos ajudar a explorar nosso objetivo final que é obter uma resposta para a pergunta 'O clima pode ter relação com a propagação do coronavírus?'. Esperamos na próxima etapa iniciar uma fase de correlação entre as bases exploradas e assim obter nossas primeiras inferências.