Evolução do número de casos de contágios no Brasil, de acordo com diferentes variáveis demográficas

Análise Exploratória

Germán Darío Buitrago Salazar *
Diego Braga Santi †
Christianne Orrico Dalforno *
Guilherme Aguilar de Oliveira †

1 Descrição do projeto

O objetivo deste trabalho é realizar uma análise considerando diferentes variáveis demográficas como idade, etnia, sexo e escolaridade, correlacionando com a taxa de contágios de COVID-19 no Brasil. Esta análise irá detectar qual a frequência da ocorrência do COVID-19 em função de diferentes fatores demográficos e sua evolução desde o começo da pandemia. Ao finalizar a análise, pretende-se criar um modelo de predição de novos casos que relacione estes fatores.

2 Equipe

O nome da equipe é **Caçadores Demográficos**. O time está composto pelos membros a seguir.

Membro 1

Nome	Germán Darío Buitrago Salazar
RA	164321
Usuário no Gitlab	gedabusa
e-mail	gedabusa@gmail.com
Afiliação	Aluno especial de Doutorado em Ciência da Computação, UNICAMP

Membro 2

Nome	Diego Braga Santi
RA	2149613
Usuário no Gitlab	diegosanti
e-mail	diegobsanti@hotmail.com
Afiliação	Aluno de Doutorado - CPGEI, UTFPR

Membro 3

Nome	Christianne Orrico Dalforno
RA	233556
Usuário no Gitlab	codalforno
e-mail	profcod@gmail.com
Afiliação	Aluna especial, UNICAMP

*Universidade Estadual de Campinas, UNICAMP
†Universidade Tecnológica Federal do Paraná, UTFPR

Membro 4

Nome	Guilherme Aguilar de Oliveira
RA	2127954
Usuário no Gitlab	guilhermith
e-mail	guilhermeoliveira.2019@alunos.utfpr.edu.br
Afiliação	Sistemas de Informação, UTFPR

3 Obtenção e processamento de dados

Para análise dos dados foram utilizadas duas bases de dados públicas, SIVEP-gripe e e-SUS. Na base SIVEP-gripe os dados estão consistentes, porém há muitos registros incompletos. Na análise inicial foram considerados apenas os registros de casos confirmados de COVID-19 e variáveis relevantes para o propósito da pesquisa. Nessa base foram selecionadas variáveis relacionadas a idade, sexo, escolaridade e etnia de pessoas infectadas em diferentes períodos de tempo. Os dados nulos foram desconsiderados.

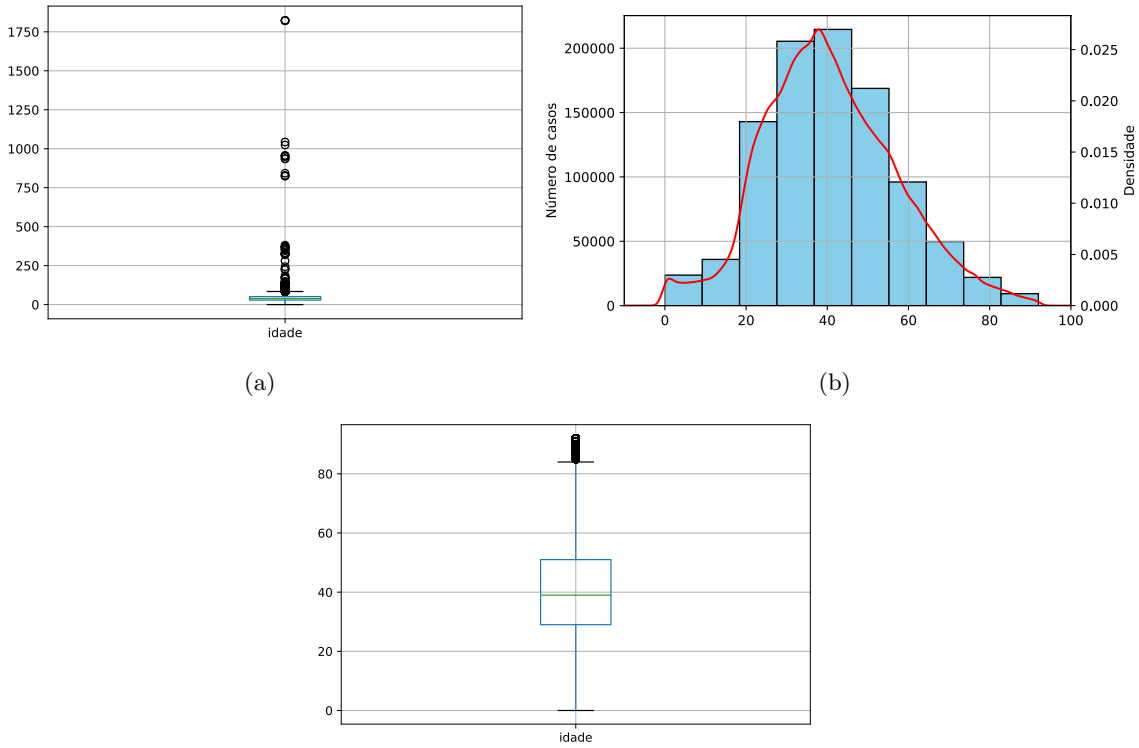
Na base do e-SUS foram inicialmente analisados os dados do estado de São Paulo, devido ao grande volume de informações. Nas próximas análises serão considerados os dados de todos os estados do Brasil. Em termos da qualidade da informação, a base possui dados incompletos e inconsistentes que devem ser pré-processados. Dentre as inconsistências encontradas convém destacar datas que não foram preenchidas ou com formatação não-uniforme, códigos IBGE que não coincidem com o nome do município ou estado e dados com outliers.

Para a limpeza da informação, implementamos expressões regulares que possibilitaram a extração e formatação de datas e valores numéricos no formato correto. Os dados que não estavam preenchidos ou que apresentavam inconsistências e não eram relacionados ao foco da pesquisa foram desconsiderados. Os nomes de municipios e seus códigos foram normalizados deletando os acentos e cruzando com uma base de codificação da IBGE. Finalmente, dados com outliers foram desconsiderados usando o escore padrão (Z-score)

4 Cobertura e distribuição dos dados

A base e-SUS é rica em informação, porém possui dados inconsistentes e incompletos. Um caso específico é a idade que, como mostrado na figura 1a, contém valores superiores a 250 anos. A figura 1a representa uma melhor visão da dispersão existente entre os dados e valores discrepantes dentro do conjunto. As anomalias observadas podem ser geradas por fatores humanos no momento de criar os registros ou sua manipulação para gerar as bases.

A figura 1b mostra o histograma resultante após remover os outliers com maior divergência do centro de distribuição. Pode-se observar que os dados são mais próximos das faixas etárias de vida de uma pessoa. Finalmente, percebe-se que ainda existem valores de outliers que estão afastados do centro de distribuição, como mostrado na figura 1c.

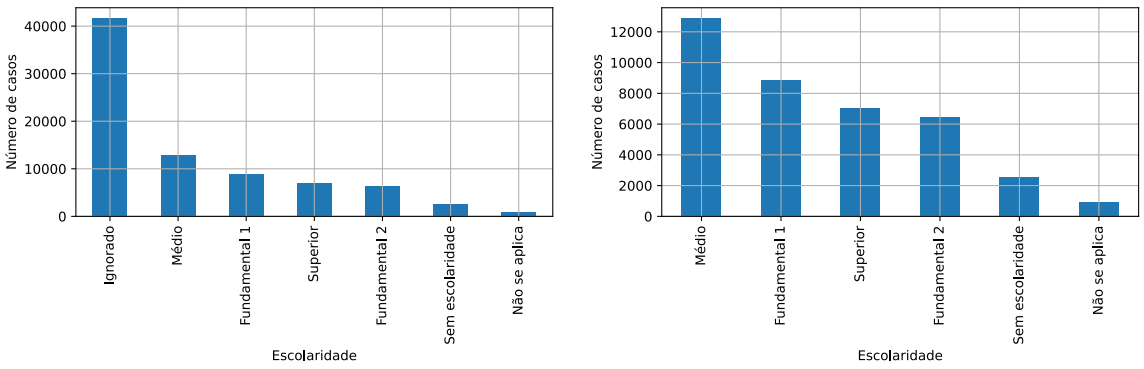


(a) (b)

(c)

Figura 1: (a) Boxplot da idade da base e-SUS. (b) Histograma e KDE da base e-SUS após remover os outliers mais afastados do centro de distribuição. (c) Boxplot da idade após remover os outliers mais afastados.

Por outro lado, a base SIVEP-gripe, mesmo sendo uma base de menor tamanho, comparada com a base e-SUS, contém dados demográficos como idade, grau de escolaridade dos pacientes, distribuição étnica e gênero. A figura 2a representa os graus de escolaridade registrados para todas as amostras. O número de casos rotulados como ignorados é a que representa a maior frequência. Para as análises a serem realizadas no projeto, considera-se este valor como um dado inconsistente que pode derivar em resultados errados. Portanto, é desconsiderado da base na etapa de pré-processamento, como observado no histograma da figura 2b.



(a) (b)

Figura 2: (a) Histograma dos graus de escolaridade dos pacientes resgistrados na base SIVEP-gripe. (b) Histograma dos graus de escolaridade após remover valores catalogados como ignorados.

No caso da distribuição étnica e de género, a maioria dos registros estão completos e são

promissores. Exemplo disso, são o histograma das etnias encontradas na base da figura 3 e o gráfico de porcetagem de gênero da figura 4

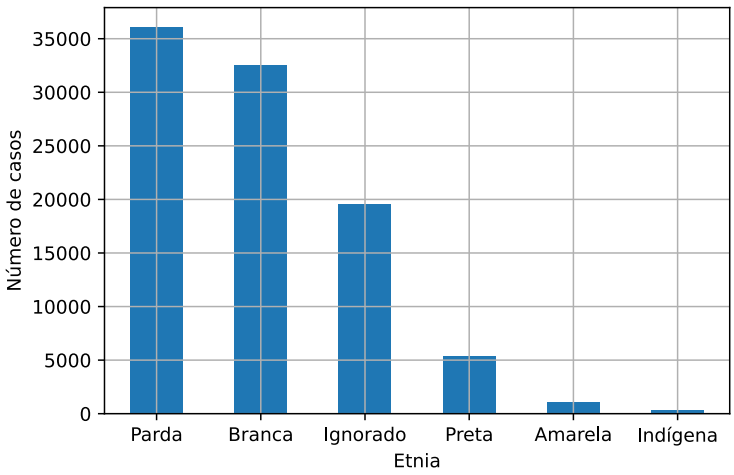


Figura 3: Histograma das etnias encontradas dentro da base de dados SIVEP-gripe.

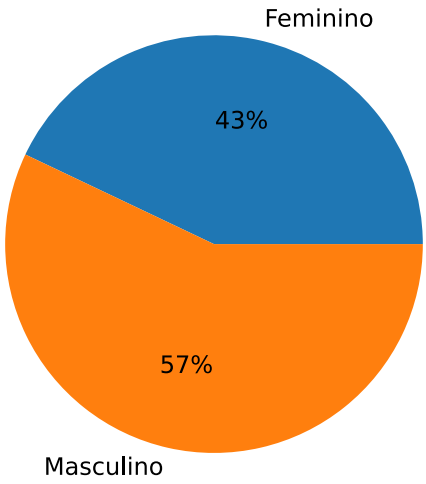


Figura 4: Diagrama de porcentagem dos gêneros encontrados dentro da base SIVEP-gripe.

5 Perguntas de pesquisa e explorações iniciais

O direcionamento do projeto e da análise exploratória foi estabelecido a partir da formulação de perguntas e hipóteses focadas no estudo das variáveis demográficas encontradas nas bases. A elaboração e listagem de perguntas são demonstradas a seguir.

5.1 Existe relação da escolaridade dos pacientes e a evoluação de novos casos de COVID-19 ao longo do tempo?

O objetivo dessa pergunta é estabelecer se existe um padrão que relacione como o nível de escolaridade poderia influenciar no aumento de casos de coronavírus. A figura 5 descreve a frequência de novos casos registrados agrupados pelo grau de escolaridade dos pacientes. Pode-se observar que no início da epidemia, a maior parte dos casos estavam entre pessoas com curso superior e, com o decorrer do tempo, aumentou a predominância em pacientes que possuim ensino médio completo.

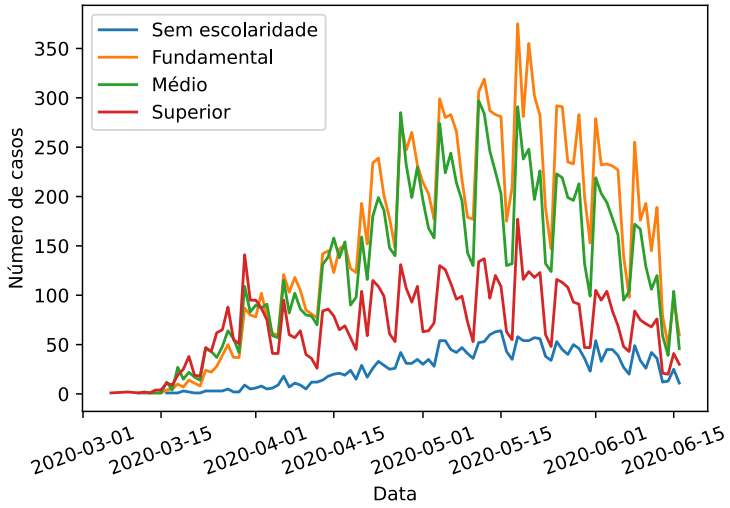


Figura 5: Evolução de novos casos de coronavírus classificados pelo grau de escolaridade dos pacientes.

5.2 Há diferença na quantidade de casos em pessoas de diferentes etnias?

A formulação dessa pergunta visa entender se existe uma correlação entre a etnia do paciente e o contágio pelo virus. Uma primeira análise desta pergunta é mostrada na figura 6 que detalha o número de novos casos diários registrados. É observável que a COVID-19 teve maior incidência nas etnias parda e branca. Porém, deve-se considerar que essas etnias têm uma maior participação na composição da população brasileira. A normalização desses dados em função das proporções étnicas do Brasil poderia otimizar os resultados obtidos no gráfico.

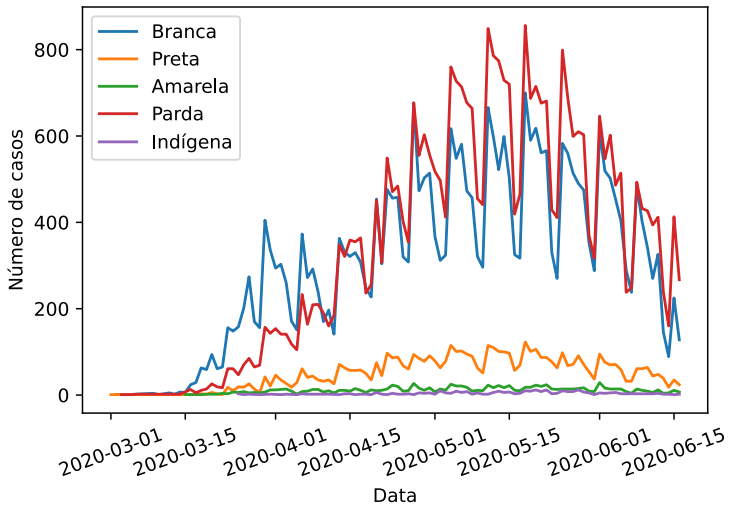


Figura 6: Evolução de novos casos de coronavírus categorizados pela etnia a que pertencem.

5.3 Qual a relação entre o gênero e a incidência de infecção por COVID-19?

Esta pergunta visa entender a relação de casos de COVID com o gênero dos pacientes. Desde o início da pandemia existem maior número de casos registrados em pacientes de sexo masculino do que de sexo feminino como observado na figura 7.Além disso, o resultado da análise poderá responder se esta diferença é significativa.

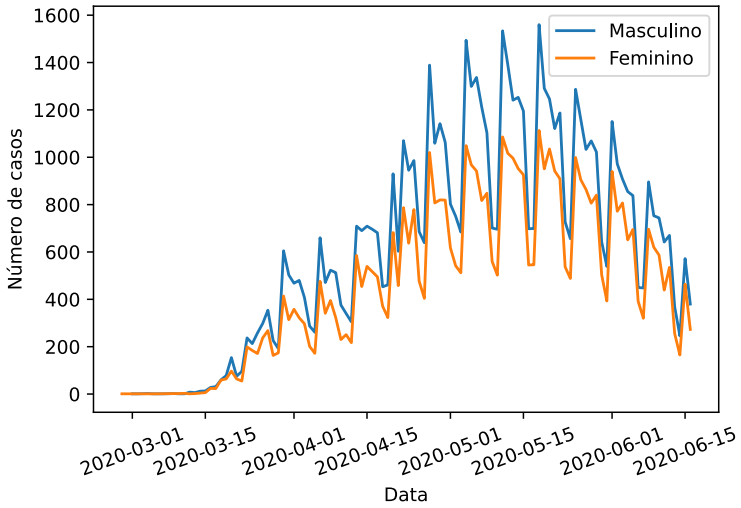


Figura 7: Evolução de novos casos de coronavírus divididos em masculino e feminino.

5.4 Quantitativamente, como a COVID-19 tem afetado as diversas faixas etárias?

Pretende-se analisar o impacto de infecções de COVID-19 em pessoas com diferentes idades em períodos distintos. Segmentando os casos por faixas etárias, observa-se na figura 8 que no período 2020-07 foram registrados o maior número de casos, afetando principalmente as pessoas com idades entre 30-40 anos.

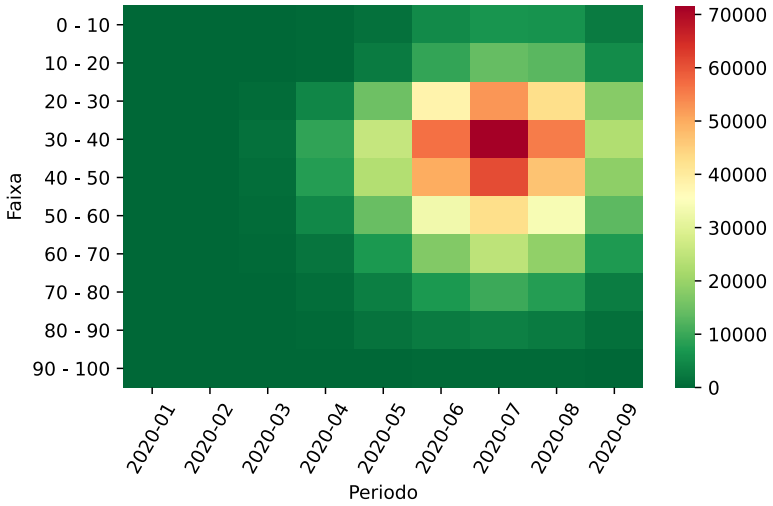


Figura 8: Heatmap da evolução de novos casos de coronavírus classificados em faixas etárias

5.5 Como o COVID-19 foi se espalhando geograficamente ao longo do tempo? Qual seria sua relação com a densidade populacional da região?

Na figura 9 é possível observar a concentração de casos de COVID-19 no estado de São Paulo. Os locais com maior densidade demográfica apresentam o maior número de casos, enquanto os lugares com menor densidade têm menos casos. Baseado nisso, esta análise visa entender como foi a evolução dos casos e a disseminação para zonas adjacentes.

Figura 9: Distribuição geográfica do total de casos de COVID-19 no estado de São Paulo. O tamanho dos circulos vermelho determina a concentração de casos em uma zona, tendo circulos maiores em lugares com mais casos registrados.

5.6 Como é possível relacionar os casos de COVID-19 com as variáveis descritas anteriormente

Após analisar individualmente cada variável, pretende-se correlacioná-las para construir um modelo que demonstre a evolução de casos e sua incidência na população.

6 Atribuições dos membros júnior

Não haverá diferenciação de atribuições entre membros sênior e júnior, visto que o membro júnior tem conhecimento em ferramentas para análise dos dados.

7 Discussão e próximos passos

Na análise exploratória foi possível estabelecer os limites da pesquisa e determinar as variáveis demográficas que serão exploradas. A análise nesta primeira fase foi realizada especificamente em uma região, mas envolverá todas as regiões a nível nacional. Vale ressaltar que a análise irá considerar o fator evolutivo no tempo. Finalmente, será necessário refinar a análise e a visualização dos dados para demonstrar com mais clareza o comportamento das variáveis.