Análise Exploratória e Visualização de Dados do COVID-19 em Manaus/AM

David Cardoso Yonekura, Lucas da Silva Lima, Rafael Barbosa de Carvalho

[1]Universidade do Estado do Amazonas – Manaus – Am – Brazil

{dcy.eng17,ldsllm.eng,rbc.eng}@uea.edu.br

Abstract. *This technical report describes how the analysis and visualization of COVID-19 data were carried out in the city of Manaus-AM using the dataset provided by the Manaus City Hall. The database went through a filtering process so that certain questions are answered and the visual elaboration of the data, in the form of graphs, was more consistent. In addition, future classification and regression tasks are carried out, such as forecasting the number of people recovered from the disease in the coming months, the locations that will have the highest incidence of cases, among others.*

Resumo. *Este relatório técnico descreve como foram realizadas as análises e visualizações de dados de COVID-19 na cidade de Manaus-AM utilizando o dataset disponibilizado pela Prefeitura de Manaus. A base de dados passou por um processo de filtragem para que determinados questionamentos sejam respondidos e a elobaração visual dos dados, na forma de gráficos, fossem mais consistente. Alem disso, futuras tarefas de classificação e regressão sejam realizadas, como por exemplo a previsão na quantidade de pessoas recuperadas da doença nos proximos meses, os locais que terão maior incidencia de casos, entre outros.*

1. Introdução

O cenário de pandemia mundial de COVID-19 [OMS 2020] trouxe questões que estão, pouco a pouco, sendo respondidas pela ciência através de diversas técnicas e métodos. Uma estratégia empregada para identificar características da doença como grupos de risco, incidência por idade e letalidade, além da distribuição geográfica e acompanhamento da curva de crescimento de contágio é a utilização de análise exploratória e visualização de dados. O objetivo deste trabalho é identificar e responder questões importantes acerca da manifestação e comportamento do COVID-19 na cidade de Manaus/AM, Brasil, utilizando a linguagem de programação **Python** e bibliotecas de apoio para sumarizar informações.

2. Visão geral dos casos confirmados

O *dataset* utilizado neste trabalho foi disponibilizado pela Prefeitura de Manaus [de Manaus 2020] e acessado no dia 6 de agosto de 2020. A princípio, foram feitas análises a respeito da quantidade e descrição dos atributos do *dataset*. Com o intuito de visualizar e identificar relações somente com os casos confirmados, optamos por considerar somente estes casos para nossas análises. A quantidade de registros, que anteriormente era de 108351 registros, passou a ser de 36947 registros como mostra a Figura 1. Foi

constatado que o dataset contém 36 atributos que classificam os casos sendo alguns deles: Idade, faixa etária, sexo, bairro, classificacao, diabetes, tipos de comorbidades, conclusão, data de notifição, taxa, data de evolução, raça, data dos sintomas, critério, tipo de teste, sintomas, etnia, profissional da saude, srag, evolução. O *dataset* se refere ao período de tempo de 30/01/2020 à 06/08/2020.

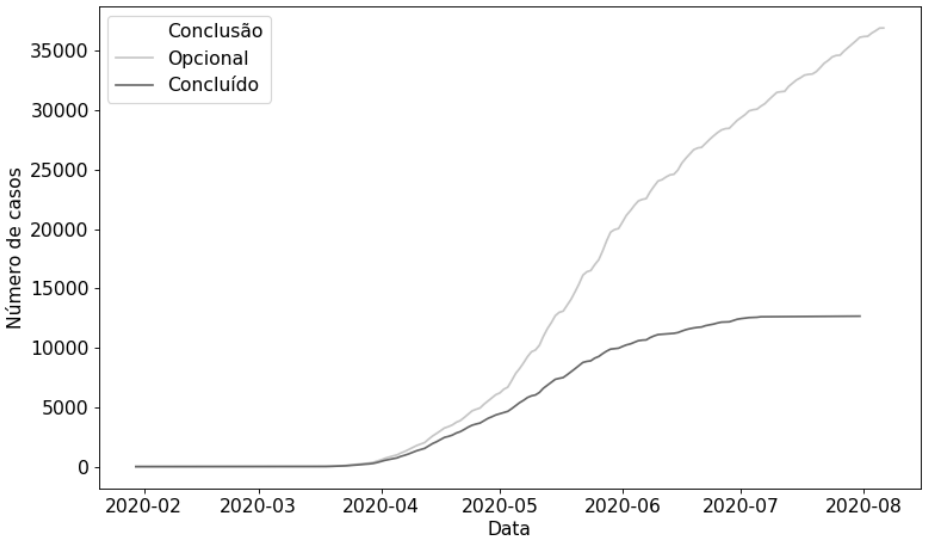
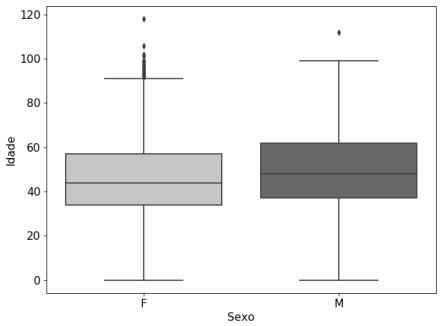
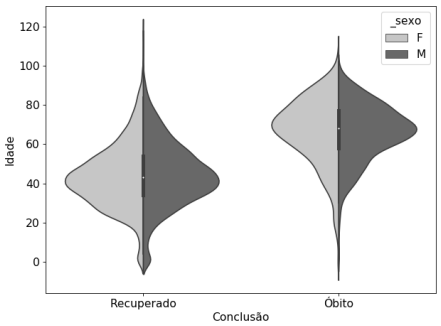


Figura 1. Gráfico cumulativo do número de casos, a linha Opcional representa todos os casos confirmados do dataset, e a linha Concluído representa todos os casos que foram concluídos (Recuperado ou Óbito)

Continuando a análise deste projeto, alguns atributos foram removidos para facilitar o direcionamento deste projeto. Os atributos removidos foram: todos os referentes a comorbidades, sintomas, etnia, profissão, datas não referentes a notificação, origem, raça, critério, evolução, tipo de teste, srag e todos os registros com os quais não se tinha informação completa.



(a) Boxplot relacionando sexo e idade



(b) ViolinPlot relacionando sexo, idade e conclusão

Figura 2. Gráficos referentes à distribuição das idades em relação ao sexo e conclusão.

Após a filtragem, o dataset continha 12672 registros e 11 atributos, com isso

tornou-se possível calcular a porcentagem de indivíduos recuperados a qual é $83,917\%$, a quantidade de casos do sexo masculino é de 6367 e do sexo feminino é de 6305 mostrando que os homens são marginalmente mais afetados numéricamente, porém, nota-se que a diferença de idade entre os sexos não é tao relevante como ilustrado na Figura 2a, inclusive nos indivíduos que se recuperaram ou vieram à óbito como demonstra a Figura 2b. Em seguida, calculando-se a média da idade obtemos o valor de $47,65$ e o desvio padrão de $18,14$.

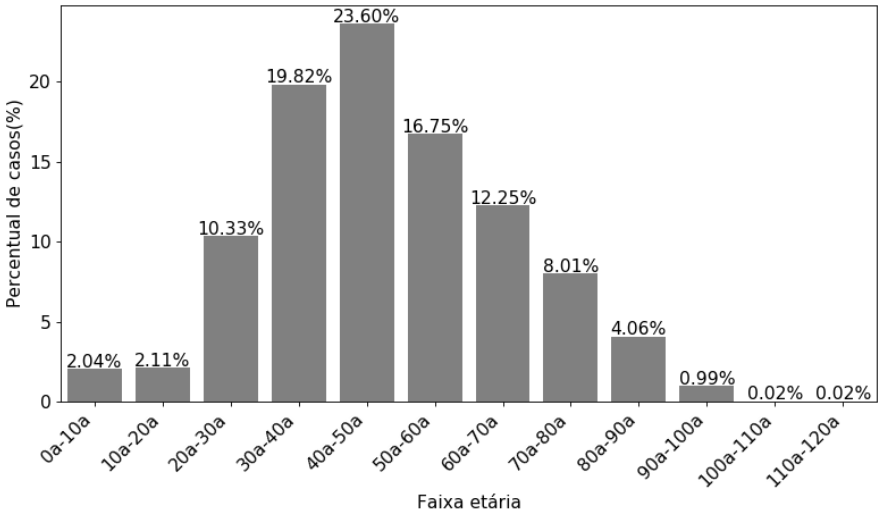


Figura 3. Gráfico do percentual de casos em relação a faixa etária.

Após o cálculo da menor e maior idade de todo o registro, obteve-se 0 anos e 118 anos como mostrado na Figura 3 indicando que existe a faixa etária a qual essas idades pertencem. Analisando a Figura 4 observa-se uma quantidade elevada de casos em indivíduos com 0 anos, é possível que isso seja um erro de entrada, onde casos sem idade definidas foram considerados como 0, outros fatores podem contribuir para essa distribuição concentrada em 40 anos como a distribuição populacional, maior frequência de testes em população de risco e se o indivídiuo faz parte da classe trabalhadora.

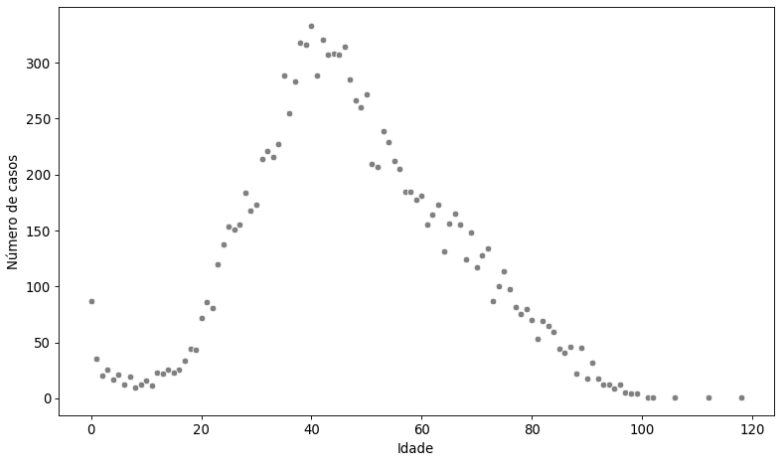


Figura 4. Scatterplot da Idade versus Número de casos

No tocante aos bairros, o bairro com a maior incidência de casos é a Cidade Nova, como mostrado na figura 5. Além disso, nota-se que este bairro também teve a maior incidência de casos recuperados. Sendo assim, os três bairros com a maior quantidade de casos recuperados são: Cidade Nova 720, Flores com 438 e Alvorada com 430 casos recuperados.

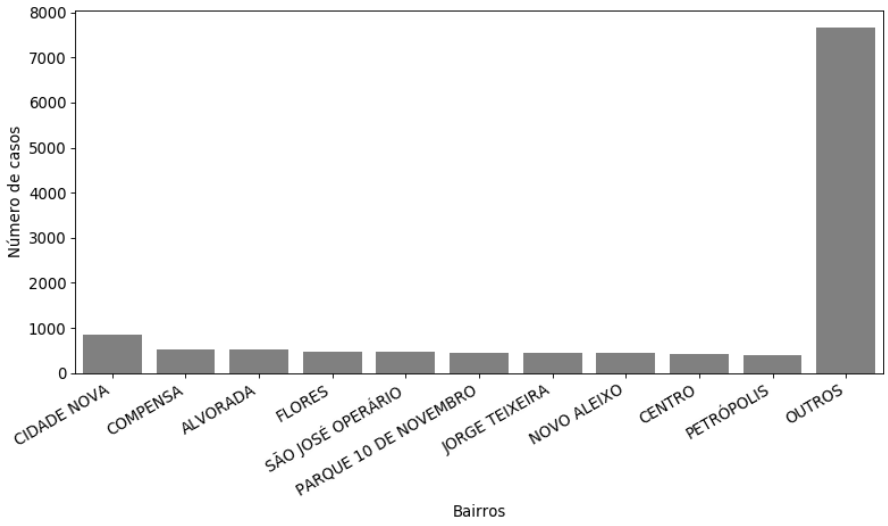


Figura 5. Gráfico ilustrando a quantidade de todos os casos nos top-10 bairros

Seguindo com os tipos de testes realizados, foi necessário repor o atributo de tipo de teste que havia sido filtrado anteriormente para podermos inferir que maior parte dos testes é referente ao Teste Rápido - Anticorpo, representando mais da metade dos testes realizados. Em seguida tal atributo foi filtrado novamente.

Tabela 1. Tipos de teste realizados nos casos confirmados de COVID-19.

Tipo de teste	Quantidade	Percentual
Teste Rápido - Anticorpo	3602	57.9659
RT-PCR	1501	24.1551
Teste Rápido - Antígeno	1101	17.7180
ELISA IgM	6	0.0965
ECLIA IgG	4	0.0643

A taxa de letalidade tem, aproximadamente, o valor de 0.16 e foi calculada utilizando a seguinte equação:

$$L = \frac{Óbitos}{TotalDeCasos} \tag{1}$$

Agrupamos então os casos confirmados por idade e através do método de Correlação de Pearson (Equação 2), que é um grau de relação entre duas variáveis quantitativas que exprime o grau de correlação através de valores situados entre -1 e 1 [Oliveira 2020], inferimos o valor de relação de -0.179869. Este valor denota uma

correlação fraca entre as variáveis, sendo que um valor de magnititude próximo de 1 implica numa alta correlação [Atoum].

$$\frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}} \tag{2}$$

3. Aplicações

Um dos possíveis usos dessa base de dados seria determinar se um indivíduo está infectado levando em conta seus sintomas, localização, idade, comorbidades.

Os sintomas são interessantes visto que boa parte dos casos de COVID-19 apresentam febre, tosse seca e dificuldade respiratória, sendo o último incomum em doenças como a febre e gripe, a localização é útil pois existem áreas com maior incidência de casos, comorbidades e idades que podem afetar o sistema imunológico, facilitando o contagio. O atributo-alvo seria se o caso foi confirmado ou não, configurando uma tarefa de classificação binária, possibilitando o uso de métricas como acurácia, precisão, revocação, curvas ROC e outras métricas relacionadas a erros tipo 1 e 2.

Por se tratar de uma tarefa relacionada à Medicina, uma validação mais robusta seria necessária, então uma validação cruzada seria o recomendado, dado a quantidade de entradas (100K) uma validação *leave-one-out* poderia ser computacionalmente inviável. Essa classificação não seria usada como um diagnóstico e sim como uma recomendação para procurar ajuda ou manter o isolamento.

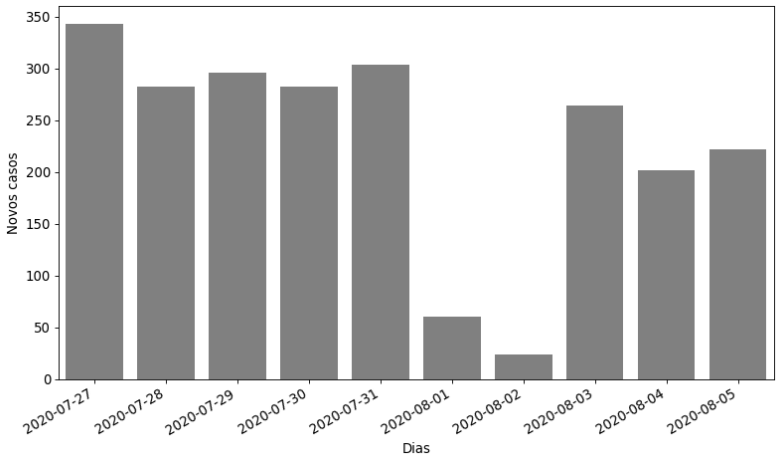


Figura 6. Quantidade de novos casos no período dos últimos 10 dias

Outra aplicação seria um modelo para a previsão do número de novos casos, visto que é possível gerar uma série temporal a partir da base de dados como mostra a Figura 6, onde temos a quantidade de novos casos de um determinado dia. Para aferir o desempenho desse modelo, utiliza-se métricas comuns de regressão como média do erro quadrático ou média do erro absoluto, Além disso, uma validação cruzada própria de séries temporais, na qual a partição de treino sempre engloba os dados mais antigos, e a de validação é adjacente à partição de treino.

Uma possível aplicação não-supervisionada seria a detecção de anomalias nas mortes por COVID-19. Podemos considerar como anomalia quando uma pessoa saudável ou fora do grupo de risco vem a óbito, o sistema poderia detectar esses casos e uma análise mais profunda poderia ser feita em relação à esses casos específicos, seja averiguar o histórico médico ou contato com a família em busca de outros fatores que poderiam agravar um caso.

A detecção poderia ser feita utilizando um autoencoder, também vale ressaltar que mesmo usando o rótulo de conclusão para definir quais casos representavam óbitos, a tarefa ainda é não-supervisionado, visto que não temos o rótulo dizendo se um caso é uma anomalia ou não.

Referências

Atoum, I. *Scaled Pearson's Correlation Coefficient for Evaluating Text Similarity Measures*. Infinite Study.

de Manaus, P. M. (2020 (accessed August 16, 2020)). Dataset covid-19 manaus.

Oliveira, B. (2020 (accessed August 16, 2020)). Coeficientes de correlação.

OMS (2020 (accessed August 16, 2020)). Folha informativa – covid-19 (doença causada pelo novo coronavírus).