Relatório - Lista 2 - MAE0399

Prof. Fábio Prates Machado

Gustavo Gomes, NUSP 12557438

Objetivos

- Fazer estatísticas básicas sobre as variáveis do banco de dados do Citi Bike NYC, respeitando os seus tipos (quantitativa ou qualitativa);
- Fazer gráficos para as variáveis, respeitando os seus tipos (quantitativa ou qualitativa);
- Apresentar observações referentes as estatísticas obtidas que permitam inferir conclusões sobre o conjunto de dados.

Introdução

O Citi Bike é um sistema de compartilhamento de bicicletas administrado por um grupo privado. O sistema atende os bairros de Nova York do Bronx , Brooklyn , Manhattan e Queens , bem como Jersey City, Nova Jersey , e Hoboken, Nova Jersey.

A empresa em seu site oferece arquivos csv que contém informações sobre o histórico de viagens feitas pelos seus usuários. Como os dados possuem quase 1 milhão de linhas, faz-se necessário utilizar ferramentas mais avançadas de análise de dados, neste relatório estarei utilizando a biblioteca Pandas do python., em conjunto do Matplotlib(para geração dos gráficos) e biblioteca Numpy(para estatísticas).

Perfil dos Usuários (Geral)

Sexo

Uma checagem rápida nos mostra que 72.5% dos usuários do mês de fevereiro de 2019 foram homems que eram assinantes do citi bike e 20.8% eram mulheres também assinantes. Os outros 6.7%¨são frequentadores casuais e assinantes que preferiram não identificar seu sexo.

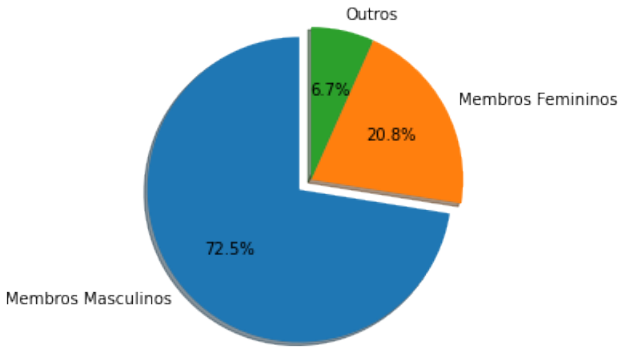

Figure 1: Assinantes ou não por sexo

Idade

Primeiramente, foi preciso criar uma variável de idade no banco de dados a partir da data de nascimento. Plotando um gráfico de barras entre o total de usuários por idade, vemos que há uma predominância de pessoas entre 25-35 anos que usam a plataforma.

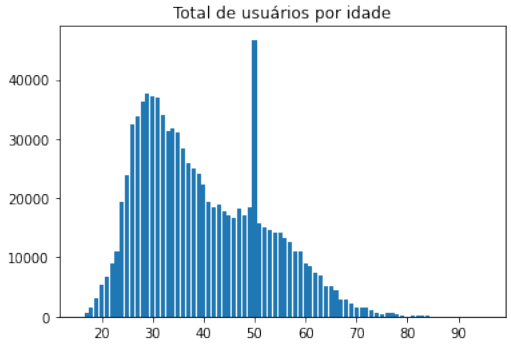

Figure 2: Usuários por idade

Observe que o valor 50 possui um pico muito anormal frente aos seus valores próximos, (49,51,etc...). Por isso analisei os dados referentes as pessoas com essa idade, mas acabei não encontrando normalidades nos valores e nem linhas repetidas...

Distância Percorrida

A plotagem do gráfico de histograma entre a distância entre estações de chegada e destinos dos usuários, mostra que a grande maioria dos usuários utiliza o sistema para viagens curtas (de 1 a 2 km), tanto é que a distância média percorrida pelos usuários é aproximadamente 1.43 km.

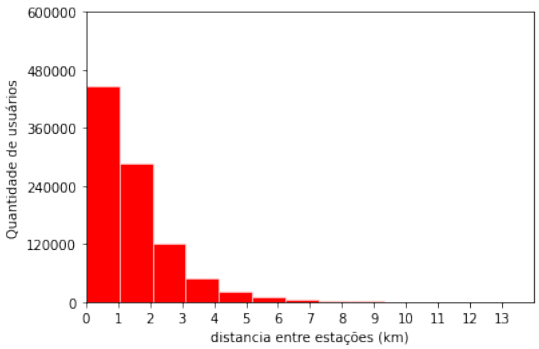

Figure 3: distancia entre estaçoes por usuário

Usuários assinantes vs não assinantes

Analisando a coluna 'usertype' do nosso dataframe pelo comando unique, inferimos que existem dois tipo de usuários: 'Subscriber', 'Customer'. Vamos analisar algumas variáveis e comparar entre esses dois grupos.

Distância Percorrida

A plotagem do gráfico de boxplot, nos mostra que os usuários não assinantes da plataforma em média percorrem uma maior distância entre estações do que os assinantes, provavelmente pelo fato de os assinantes já terem uma certa noção que para distância maiores pode ser mais econômico utilizar o transporte público da cidade, principalmente se o assinante for morador da cidade.

Além disso, a diferença entre o primeiro e terceiro quartis nos mostra que as distâncias percorridas pelos não assinantes possuem uma maior variabilidade quando comparadas com as distâncias percorridas pelos assintantes, o que faz sentido, pois o grupo de assinantes provavelmente usa as bikes para os mesmos

trajetos (de casa ao trabalho,por exemplo) o que reduz a variação da distância percorrida por esse grupo.

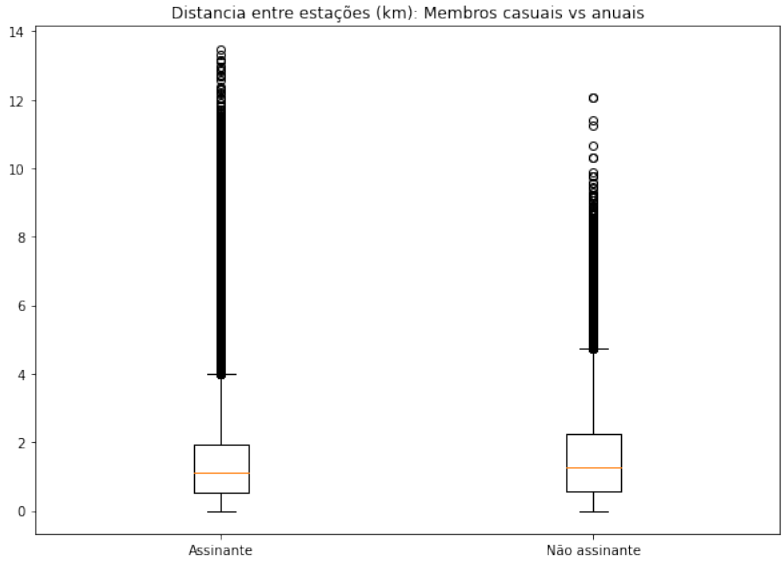

Figure 4: distancia entre estaçoes assinantes e não assinantes

Idade

Antes de fazermos esta análise, é preciso comentar a aparição de alguns valores impossíveis de idade na base de dados numa primeira visualização... Por exemplo, na base temos uma pessoa que possuia em 2019, 140 anos... Para limpar esses absurdos da base de dados, eliminei todas as linhas cujas pessoas tinhas mais do que 95 anos.

```
#exclui usu rios com mais de 95 anos
indexNames_1 = data[data['years old'] > 95].index
data.drop(indexNames_1, inplace = True)
```

Agora estamos prontos para fazer a comparação de idade entre os assinantes e os não assinantes. Novamente gerei um boxplot, que indica que os não assinantes em média (ou mediana) possuem maior idade do que os assinantes e também maior variância na idade.

Também vemos que os pontos "fora da curva" no grupo dos assinantes é bem maior comparado ao grupo dos não assinantes, mas ainda em absoluto é um valor pequeno comparado a uma base que tem um milhão de usuários.

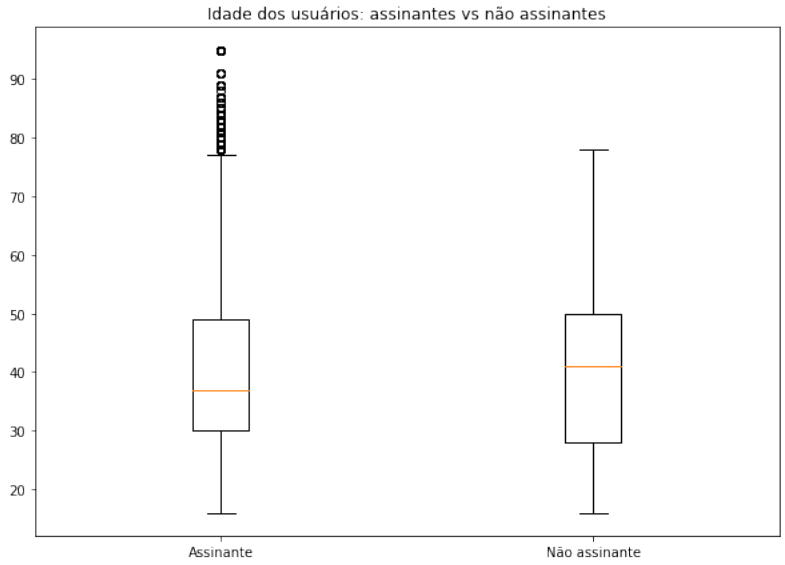

Figure 5: idade: assinantes vs não assinantes

Estação de Partida

Plotando um gráfico de barras das top 6 estações de partidas, vemos que entre os assinantes se destacam estações que ficam perto de áreas residenciais ou de hotéis, enquanto que no grupo dos não assinantes, se destacam estações que ficam perto de estações de trem e metro.

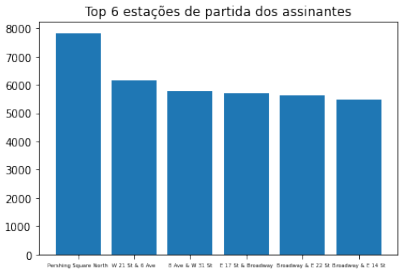

Figure 6: Partida não assinantes

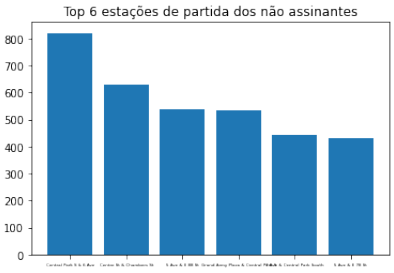

Figure 7: Partida assinantes

Estação de Chegada

Plotando o mesmo tipo de gráfico, mas agora para estações de chegadas, vemos que o top 6 coincide tanto no grupo de assinantes como os dos não assinantes, e são estações que ficam perto de destinos badalados da cidade, como a Broadway,

então faz sentido esta coincidência nos dois grupos.

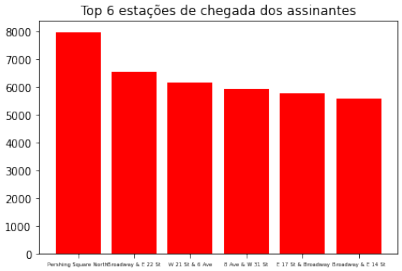

Figure 8: Chegada não assinantes

Figure 9: Chegada assinantes

Idade e Velocidade Média

Por último iremos analisar se há algum tipo de relação entre a velocidade média dos usuários e suas idades. Para isso, irei criar a variável 'average speed' a partir das colunas 'distance' e 'trip duration':

```
#cria a variavel velocidade m dia
data['average speed'] = data['distance']/data['tripduration']
#cria variavel idade do usuario
data['years old'] = 2019 - data['birth year']
```

Um gráfico de dispersão será útil para avaliar a relação entre essas variáveis. A partir do gráfico, vemos que geralmente pessoas mais velhas tem uma menor velocidade média, o que é natural devido as limitações do corpo e também pelas precauções tomadas. Além disso, podemos ver que as variáveis velocidade média máxima e idade tem uma relação aproximadamente linear.

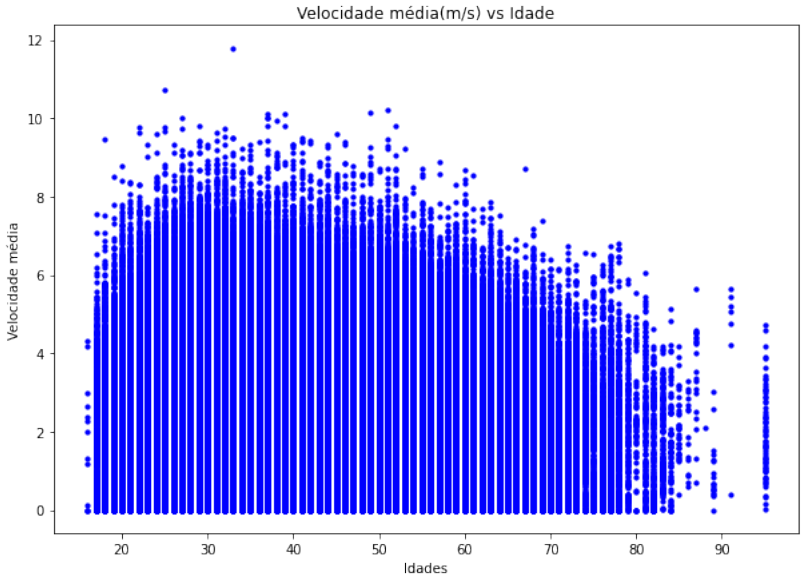

Figure 10: Velocidade média versus idade dos usuários

Conclusão

Podemos tirar de conclusão deste relatório que o uso de softwares apropriados para análise de grandes bases de dados é essencial e pode nos trazer muitas informações e inferências uteis para contribuir para a nossa sociedade.

Referências Bibliográficas

- Estatística Básica, Bussab e Morettin, 8a edição.
- Probabilidade, um curso moderno com aplicações, S. Ross, 8a edição.