Vinicius Batista Mathias da Silva

MELHOR LOCALIZAÇÃO PARA ABRIR O FASTSNACK

1 INTRODUÇÃO

O presente projeto foi elaborado para servir como avaliação final para a certificação de *Data Science* do coursera, curso online da IBM, bem como servir mutuamente como avaliação intermediária do curso da Infnet de *Data Science, Data Analytics anda Machine Learning*. O objetivo do projeto é responder à pergunta: qual a melhor localização para se abrir a lanchonete "FastSnack"? É importante notar que o escopo do projeto limitou a busca da localização ideal aos arredores do bairro de Botafogo, localizado na cidade do Rio de Janeiro, no Brasil.

Botafogo é um bairro localizado na zona nobre de uma das principais cidades da América Latina, o que o torna um local bastante atrativo, no entanto impõe algumas dificuldades. O elevado IDH de 0,952 da região sugere uma boa condição financeira de seus moradores, além disso o bairro possui uma forte área comercial, o que gera um grande volume de circulação de pessoas, ambos os fatores são atrativos para o estabelecimento de negócios. Entretanto, os motivos que impulsionam os interesses de estabelecermos a lanchonete também atrai diversos outros concorrentes.

Dois nichos de clientes de lanchonete estão muito presentes em Botafogo, trabalhadores e estudantes, ambos constantemente em busca de uma comida rápida e prática para consumir entre os seus afazeres. A FastSnack optou pela área justamente para se beneficiar de ambos os nichos, no entanto o foco será direcionado para atender os estudantes, devido a fácil localização dos pontos de concentração dos mesmos bem como a baixa exigência quanto as comidas consumidas. Definido esse direcionamento do negócio fica nítido o foco a ser adotado para alcançar o objetivo do projeto, buscar o local com a maior concentração de escolas.

Entretanto, para embasar melhor a escolha será considerada também o número de concorrentes nos arredores. Essa busca se limitou a um raio pequeno devido à natureza do negócio ser voltada a conveniência de quem se encontra próximo. Assim, a busca do projeto é encontrar na região designada um local com equilíbrio entre um elevado número de escolas e um reduzido número de concorrentes. Essa escolha foi baseada em ferramentas e técnicas da ciência de dados, seguindo a sua metodologia a fim de encontrar o ponto que equilibre as duas variáveis citadas.

A API de suporte a todo a análise foi a Foursuare, que oferece diversas informações confiáveis, entretanto na conta gratuita que será trabalhada o número de dados retornado é limitado. Após a coleta dos dados foi feito um pré-processamento, limpeza, filtragem e formatação dos mesmos a fim de torná-los adequados para serem utilizados no modelo. O escopo pré-definido da utilização de uma machine learning de clustering, especificamente o K-means é adequado, visto que a necessidade do problema é identificar regiões onde houvesse um número elevado de escolas. Por fim a análise do número de concorrentes ocorre inserido nos clusters gerados através da concentração de escolas.

2 MATERIAIS E MÉTODOS

2.1 Aquisição de dados

A escopo do curso define que deve ser usado a API Foursquare para a aquisição dos dados. Embora seja permitida a utilização de outras fontes de dados simultaneamente o processo não foi necessário. A Foursquare API possuí todos os dados levantados como necessários para atender o objetivo. Esses dados são o número de escolas e restaurantes em regiões selecionadas. Estabeleceu-se um raio. Para aumentar o número de requisições diárias possíveis foram utilizadas as credenciais de uma conta de desenvolvedor.

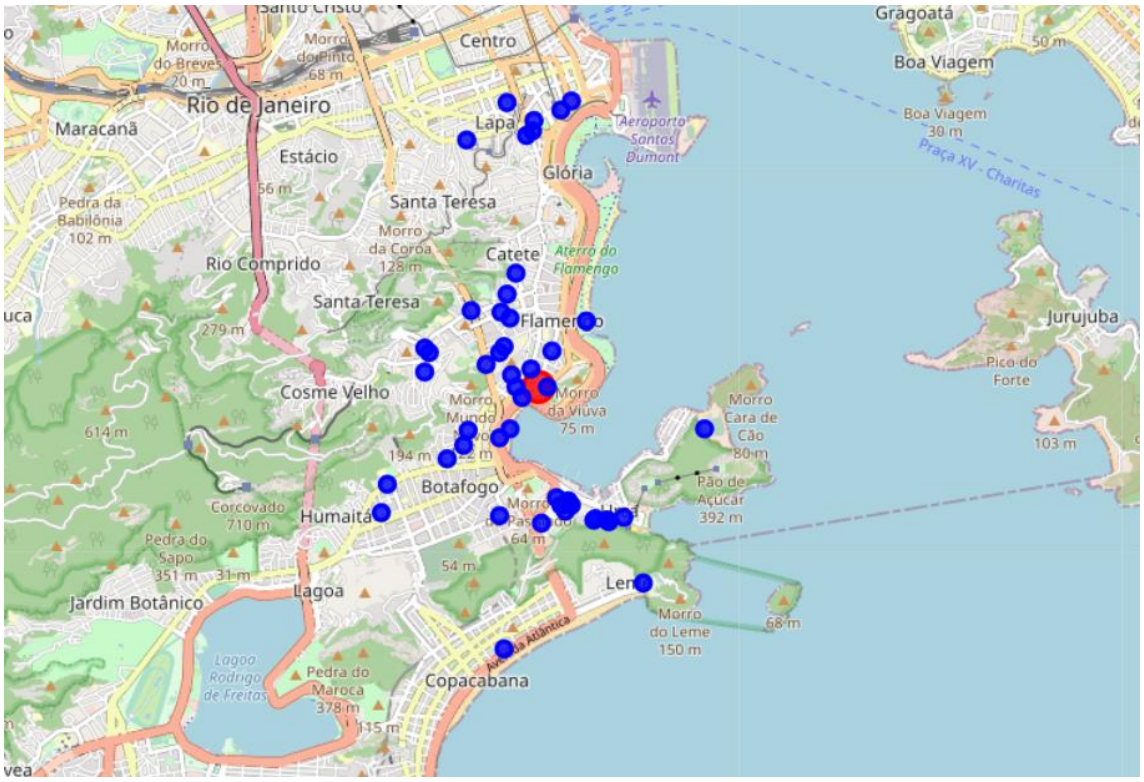
O projeto delimitou a sua área de buscas ao centro de Botafogo e os seus entornos. Foi traçado um raio de 3 quilômetros a partir do endereço rua Praia de Botafogo número 328.

2.2 Metodologia de análise dos dados

Após a aquisição dos dados o primeiro procedimento foi transformá-los de JSON para dataframe, com a finalidade de melhorar a sua manipulação. Ao observar os dados percebeu-se que havia informações que não acrescentavam no andamento do projeto, assim diversas colunas foram descartadas, as que sobraram tiveram os seus nomes ajustados para uma melhor leitura, resultando nas seguintes: id, name, categories, adress, latitude and longitude. Por fim ajustou-se os valores da coluna ‘categorias’ para strings legíveis.

Em seguida foi gerado o mapa a seguir indicando as localidades das escolas na região estabelecida, como apontado na seção ‘aquisição de dados’.

Escolas na região



Os pontos azuis apontam as localizações das escolas enquanto o ponto vermelho aponta o centro de onde se mediu o raio para a busca. Em uma breve análise visual já é supor a existência de um cluster ao norte da imagem, separados do resto das escolas por uma faixa vazia.

Para chegar a resposta da melhor localidade para abrir a lanchonete foi utilizado uma metodologia de clustering. O K-means já era previsto no escopo do projeto, logo obrigatório. Entretanto, atende satisfatoriamente as necessidades do projeto agrupando as escolas baseado em sua posição geográfica de maneira não supervisionada.

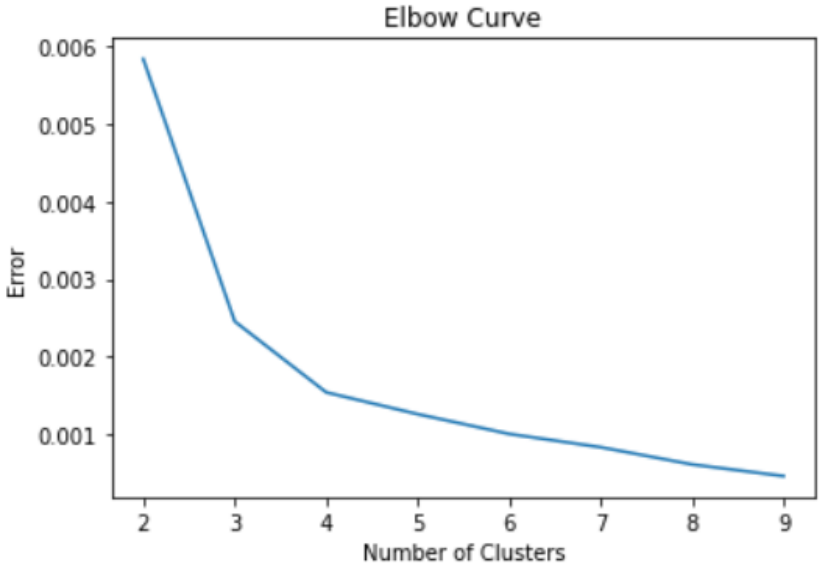
Porém ao visualizar o mapa percebe-se que a escola mais ao sul está afastada das demais fugindo do padrão observado. Isso pode ser um problema para o K-means, visto que esse não identifica outliers. Ao observar a geografia do lugar percebe-se que a mesma também afasta essas escolas das demais pois há um morro entre elas. Devido a esses dois fatores essa escola será retirada da análise para não modificar o centroid da sua posição ideal.

3 ANÁLISE DE DADOS – DESENVOLVIMENTO

3.1 Clustering system evaluate

A escolha do número ‘k’, ou número de clusteres que os dados serão separados, é de extrema importância no K-means. Para descobrir esse número utilizamos duas métricas o Elbow e a Silhouete.

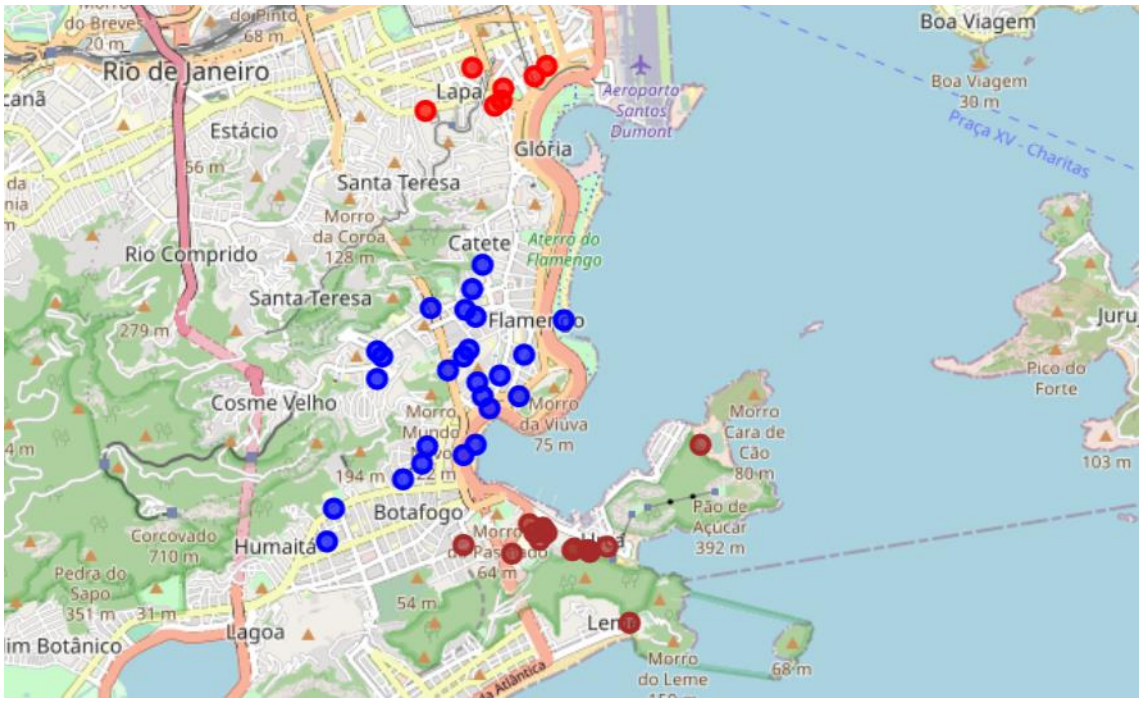
Ao analisar a relação entre o erro, representado no gráfico por score, e a variável ‘k’ temos:



Os dois pontos que se destacam no gráfico é quando k é igual a 3 e quando K é igual a 4, ambos são pontos where the rate of decrease sharply shifts. In this case elbow point is when the k=3, because it’s the first point of changing rate of decrease. Devido a importância do parâmetro k utilizou-se a Silhouete para conferir se o número ideal é 3, e o mesmo foi confirmado. Quando k=3 a Silhouete atinge o seu maior valor de aproximadamente 0.5756.

3.2 Graphical visualization

Utilizando 3 como o número de clusters podemos observar a seguinte distribuição:



O primeiro fator que chama atenção é que como havíamos suposto anteriormente os pontos vermelhos formam um cluster. Percebe-se que a maior parte das escolas se concentram no cluster azul.

Entretanto, uma única variável pode não gerar uma resposta tão confiável, dessa maneira outra variável será considerada, o número de concorrentes. Assim, além de buscar a localização que atenda um grande número de escolas, faremos um equilíbrio dessa métrica com o número de concorrentes na região de cada cluster. Devido a uma limitação no API da Foursquare serão buscados todos os restaurantes e não apenas as lanchonetes. Embora não ofereçam exatamente o mesmo serviço e produto como ambos se tratam de alimentação podem ser considerados concorrentes indiretos. A partir do centroid de cada cluster buscou-se os restaurantes em um raio de 400 metros.

4 DISCUSSÃO DOS RESULTADOS E CONCLUSÃO

4.1 Apresentar os resultados

Uma questão interessante a ser respondida é o quanto a adição concorrência na análise influencia na escolha do melhor local para abrir a FastSnack. A tabela a seguir representa a distribuição das escolas em cada cluster.

Cluster color	Number of schools	Percentage of schools
Red	7	14,3%
Blue	25	51,0%
Brown	17	34,7%

Considerando apenas o número de escolas fica claro qual região devere-se escolher. O cluster azul possui aproximadamente metade do total de escolas analisadas, assim nessa primeira análise a localização ideal da FastSnack seria na centroid do cluster azul, essa é a região aonde conseguiria atender da melhor maneira as suas 25 escolas. Além disso, um fator que não havia sido esperado é que o cluster azul está no centro de Botafogo, embora a proximidade do centro do bairro não fosse uma variável que se pretendia atingir previamente, esse fato acaba sendo uma coincidência satisfatória. Diferentemente o cluster vermelho possui uma baixa concentração de escolas, 7 no total. Já o cluster marrom possui 35% das escolas aproximadamente, um número inferior comparado ao cluster número aul, aproximadamente uma vez e meia menor. No entanto, ainda possui um volume razoável de escolas no seu entorno.

Assim, para encontrar a resposta final outro dado é importante, o número de restaurantes concorrentes em cada cluster.

Cluster number	Number of restaurants	Percentage of restaurants
Red	50	70,4%
Blue	15	21,1%
Brown	6	8,5%

Novamente o cluster red ficou na pior colocação possuindo mais concorrentes nas proximidades. No entanto, nessa análise o melhor colocado foi o cluster Brown com duas vezes e meia menos concorrentes do que o cluster azul. Por fim podemos observar a proporção entre o número de escolas e o número de concorrentes.

Cluster number	Schools / Restaurants
Red	0,14
Blue	1,67
Brown	2,83

A tabela acima aponta o cluster marrom como a localidade com a melhor relação School/Restaurants. Assim o centroid do cluster marrom é o ponto que melhor equilibra a relação entre consumidores e concorrentes.

4.2 Resultados atingiram o objetivo proposto?

Dentro dos dados obtidos o resultado foi satisfatório. No entanto, é importante considerar uma possível enviesamento da resposta visto a limitação no número de dados da conta gratuita na Foursquare. Esse projeto alerta que os números de restaurantes capturados não representam a realidade baseado no conhecimento prévio da região. Assim, se tratando de uma aplicação realista o resultado pode ser considerado deficitário, principalmente quanto ao número de restaurantes em cada cluster. Assim, sugere-se que replique o procedimento do projeto com uma base de dados mais confiável. Caso não seja possível, para uma aplicação real os dados mais confiáveis devem prevalecer. Assim o cluster azul pode ser escolhido por atingir um maior número de escolas.

5 REFERÊNCIAS

https://pt.wikipedia.org/wiki/Botafogo

https://www.coursera.org/learn/machine-learning-with-python/lecture/Ky5Wf/intro-to-k-means

https://www.coursera.org/learn/machine-learning-with-python/lecture/rLcgP/more-on-k-means

https://github.com/CharlesPrado23/Notebooks/blob/main/ProjetoFinal(Coursera)/Relat%C3%B3rio%20T%C3%A9cnico.pdf

https://github.com/CharlesPrado23/Notebooks/blob/main/ProjetoFinal(Coursera)/Guide_Notebook.ipynb