Capstone Project - Dry Bars/ Beauty Express

Applied Data Science Capstone by IBM/Coursera



**Introdução**

Com o objetivo de estudar possíveis locais para a abertura de um negócio ainda não explorado no Brasil, este projeto traz o conceito do *DryBar*, e tem como proposta identificar locais mais adequados para a localização deste tipo de estabelecimento. O resultado pode permitir extrapolar o perfil das localizações mais exitosas para outras cidades com características similares.

**O conceito**

O *DryBar* pode ser definido como espaços expressos de beleza, por preços mais convenientes, se comparado aos serviços tradicionais. Surgiu na Califórnia, em 2010 e se tornaram presença comum em cidades como Nova York e Los Angeles, pela agilidade, facilidades e preços convidativos.

Para o desenvolvimento do projeto, a opção foi mapear uma região da cidade de Nova York, buscando compreender elementos de geolocalização e sua relevância, para a instalação do *DryBar*. Considera-se que este é um tipo de serviço com elevada dependência do fluxo de pessoas do sexo feminino, com pouco tempo disponível e um perfil que valoriza a boa apresentação no ambiente de trabalho. Esses são fatores considerados determinantes para o sucesso deste modelo de negócio.

**Objetivo do projeto**

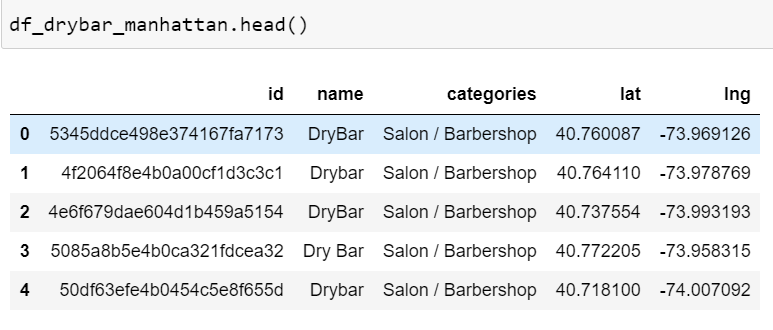
Encontrar *DryBar* na região de observação, analisando seu entorno a partir do estudo de clusters, complementado com avaliações dos clientes, considerando que, além da qualidade do serviço, está implícita a facilidade de acesso, elemento relacionado à localização.

Foi feito o mapeamento da região de Midtown, na Ilha de Manhattan, um dos principais hubs de grandes empresas da Cidade de Nova York. Devido a essas características, a região concentra público com perfil para esse tipo de negócio; mulheres com muitas atividades e pouco tempo para os tradicionais salões de beleza

Com base nesses critérios, o Foursquare (<https://developer.foursquare.com/docs>) foi usado como a única fonte de dados. O algoritmo de aprendizado não supervisionado ‘k-means clustering’ foi o meio para identificar as vantagens de cada área, para que seja definido o perfil da melhor localização, para futuros encaminhamentos aos interessados em implantar este negócio em grandes capitais brasileiras.

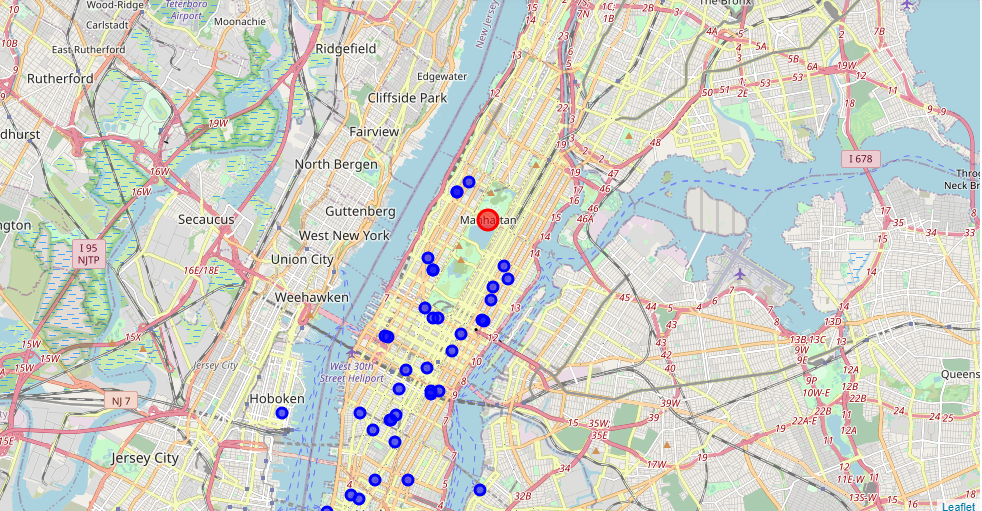
**Desenvolvimento**

Na API do Foursquare foi feita uma busca com um raio de 9,5km do centro do distrito de Manhattan, para encontrar os estabelecimentos de *drybar* na região e após, manipular e organizar os dados em uma data frame.



Parte data frame com a localização dos Drybars

Retornaram 41 Drybars, no gráfico abaixo é possível ver como estão distribuídos na ilha de Manhattan.



Localização gráfica dos Drybars

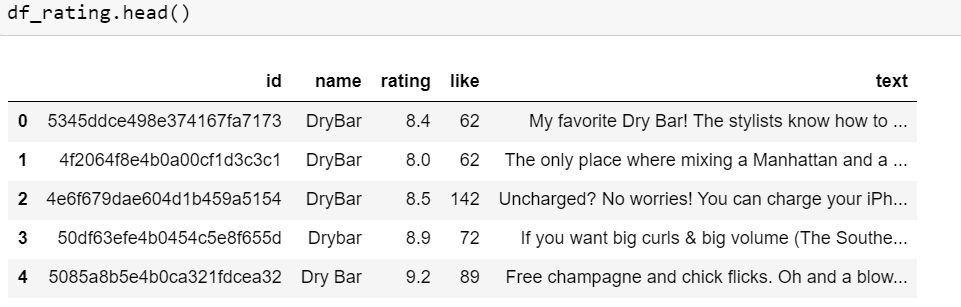
É possível observar que os D*rybars* estão concentrados em uma parte da ilha de Manhattan e não em toda sua extensão. Os pontos estão nas regiões de Midtown, Upper e Lower, três das cinco que dividem a ilha de Manhattan. São nessas regiões que estão concentradas grandes empresas, vários órgãos do governo e diferentes comércios, de fato a proximidade com outros estabelecimentos parece ter uma influência na escolha do local para a abertura desse modelo de negócio, já que o outro lado da ilha são áreas consideradas residenciais.

<https://www.rodei.com.br/amp/para-entender-melhor-os-bairros-de-manhattan-parte-ii/>

**Notas dos *Drybars***

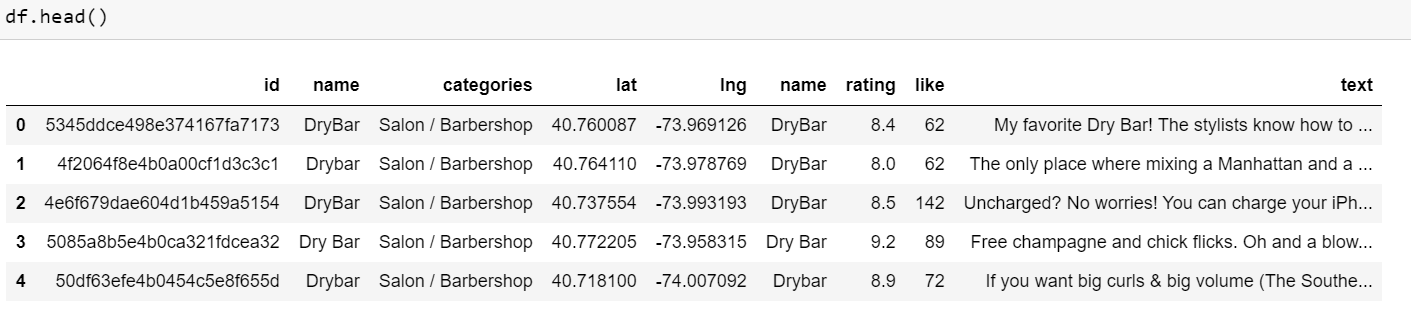
O escopo do projeto está em verificar se existe uma relação do sucesso do negócio com sua localização, neste caso, conhecer quais tipos de categorias, serviços que permeiam estes estabelecimentos e para mensurar o sucesso do negócio que está associado a qualidade de serviço será utilizado as notas dos clientes, que de forma implícita, está a localização, parte-se da hipótese de que estabelecimentos com boas pontuações costumam estar em **pontos estratégicos**.

Novamente é feita uma busca no Foursquare com ID de cada *Drybar* afim de obter as avaliações dadas pelos clientes (detalhes da busca no notebook), após manipulado é organizado um dataframe, df\_rating, conforme abaixo.



Parte do dataframe com as notas dos clientes

O passo agora é concatenar a primeira data frame que contém a categorias e a localização de cada *drybar,* com o segundo data frame, que contém notas.



Parte do dataframe após com a concatenação

Com os dados de busca do Fousquare previamente organizados, próxima etapa será usar o algoritmo de Machine learning para extrair informações que serão transformadas em posteriormente em conhecimento.

**Machine Learning**

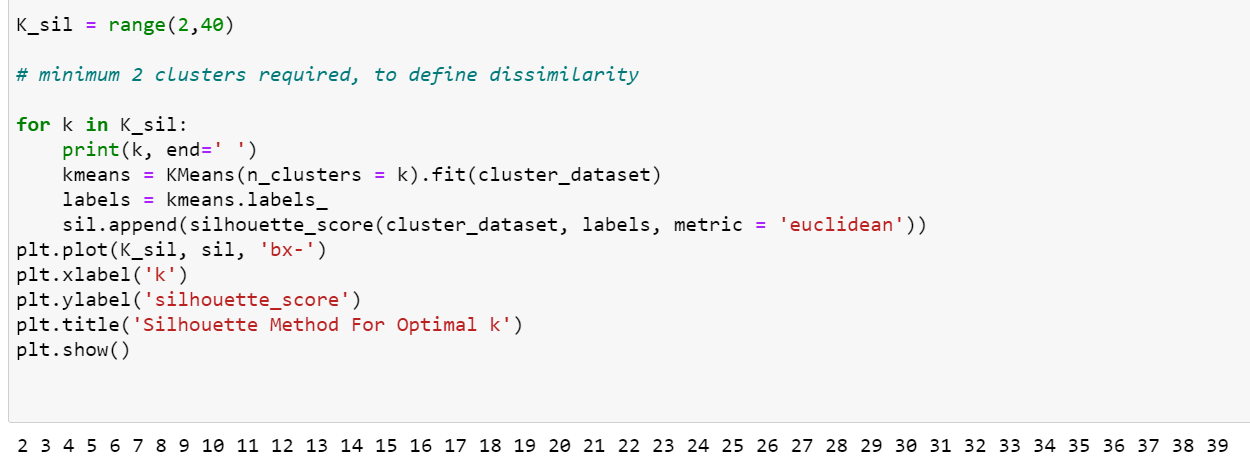
O algoritmoMachine Learning não supervisionado “K-means” agrupa dados por semelhanças entre si em torno de um ponto central, o centróide. Para o processamento são dadas as variáveis de entrada, neste caso serão as latitudes e longitudes e ele retorna como saída grupos, chamados de clusters.

Uma vez que se tem os clusters o passo seguinte é calcular a média das notas avaliadas pelos clientes de cada grupo. Os dois clusters que tiverem a maior média de notas, fará uma nova busca no Foursquare a partir dos centróides destes dois clusters para descobrir as categorias dos estabelecimentos vizinhos e tirar conclusões a partir disso.

**K-means**

Para iniciar o algoritmo é necessário definir o melhor valor de k, como não se sabe qual é o melhor valor para se escolher então se usa o método da *silhouette*, que calcula a média entre o quão coeso estão os pontos dentro do cluster e o quão distantes estão dos outros clusters assim quanto mais próximo de 1 melhor é o valor de k.

Para realizar este processo é preciso fornecer valores para o algoritmo para que ele teste e descubra qual é o melhor valor para k, não há uma regra para a escolha destes números, fica a critério de cada um, no caso do projeto a escolha é de números que vai de 2 até 39 representado por *range (2,40)* conforme imagem abaixo.



Execução para descobrir qual melhor valor para k

O código acima retornou o gráfico para avaliar qual k será escolhido.

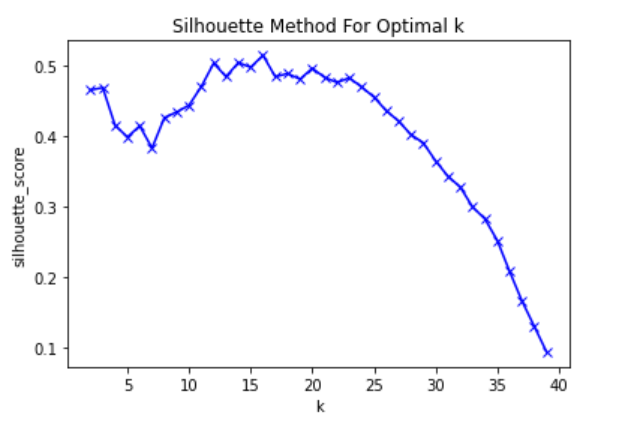
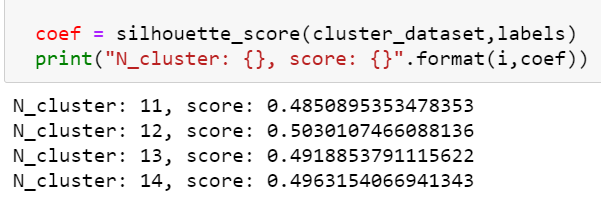


Gráfico da Silhouette

É possível ver no gráfico dois picos, o primeiro aproximadamente entre 11 e 14 e segundo entre 15 e 17, para o cientista de dados a decisão entre escolher quais deles tem o melhor para k é pessoal, mas no caso do projeto com uma amostra pequena de dados não é muito interessante ter muitos clusters.

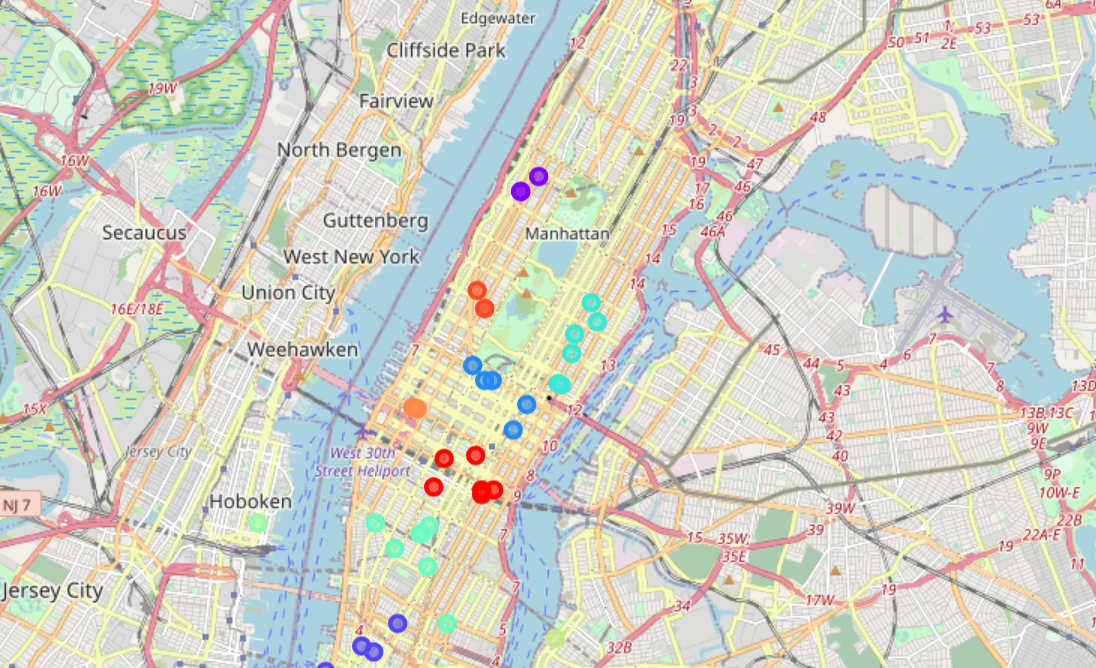
Ao escolher o primeiro pico do gráfico ainda não se sabe qual é melhor valor par k, pode ser 11,12 ,13 ou 14, qual é o melhor entre eles?

Nem sempre é possível tomar esta decisão apenas olhando para os gráficos, então é importante verificar através dos valores dos coeficientes da *silhouette,* conforme abaixo.



Valores dos coeficientes cos claustres 11,12,13 e 14

Este processou gerou a informação de que o melhor valor para k é 12, significa dizer que o algoritmo não supervisionado *k-means* baseado nas variáveis de entrada: latitude e longitude, retornará 12 cluster semelhantes entre si, conforme pode ser visto no gráfico abaixo.



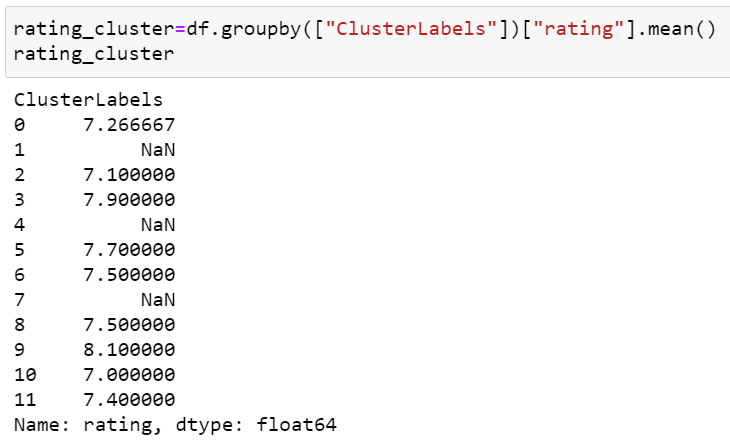
Os clusters

**Conhecendo os clusters**

Como parte do projeto não é explorar todos os clusters, mas sim os estabelecimentos considerados como casos de sucesso a partir das avaliações dos clientes. Uma forma de obter estas informações, é pela média das notas de cada cluster.

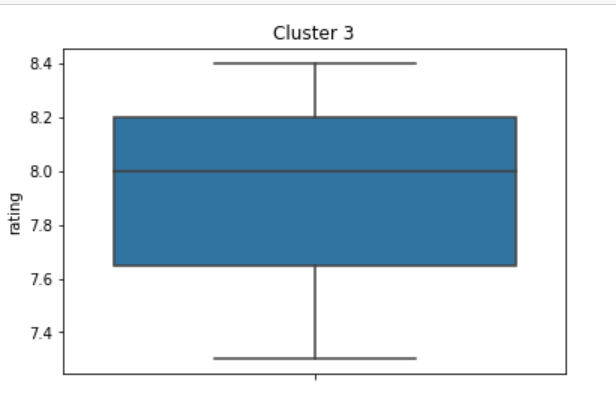
Como a ideia é saber a influência dos estabelecimentos vizinhos, parece ser considerável a comparação entre dois clusters de maior média.

Para saber as médias de cada cluster é preciso usar a função groupby, conforme abaixo.



Média das avaliações dos clientes de cada cluster

Os clusters 3 e 9 apresentam as maiores média, apenas estes dois grupos servirão de parâmetro para conclusão do projeto, para isso é selecionado todos os estabelecimentos referentes a cada um destes clusters.

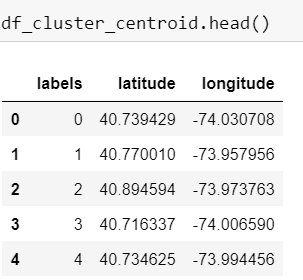


Boxplot do cluster 5

O cluster 3 tem uma boa concentração de notas consideradas boas acima de 7,5 (conforme o gráfico abaixo). Já o cluster 9 tem apenas 1 elemento, ou seja, 1 estabelecimento, por esta razão não é necessário plotar o gráfico, mas como sua nota teve um valor significativo vale a pena conhecer as categorias dos seus estabelecimentos vizinhos. (detalhes no notebook)

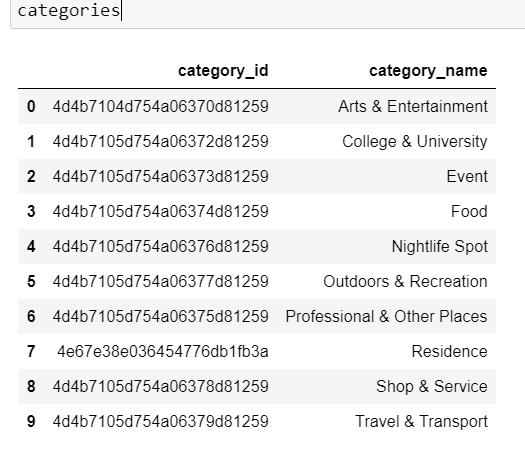
**Em busca das categorias vizinhas dos clusters 3 e 9**

Como dito anteriormente, o algoritmo k-means retorna os clusters e também o ponto central de cada um com a função *“kmeans.cluster\_centers*”, estes centroides são coordenadas. Assim, após manipulados são organizados em uma data frame. (detalhes no notebook)



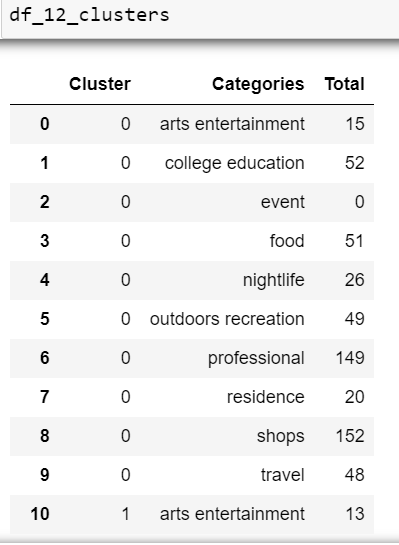
Localização dos 12 centroides

Para buscar as categorias a partir de um ponto de localização primeiro é preciso saber quantas categorias a API do Fourquare disponibiliza, assim com o código e as chaves de cadastro feita anteriormente é possível ter esta informação em um dataframe como este abaixo.



As 10 categorias que a API do Foursquare disponibiliza

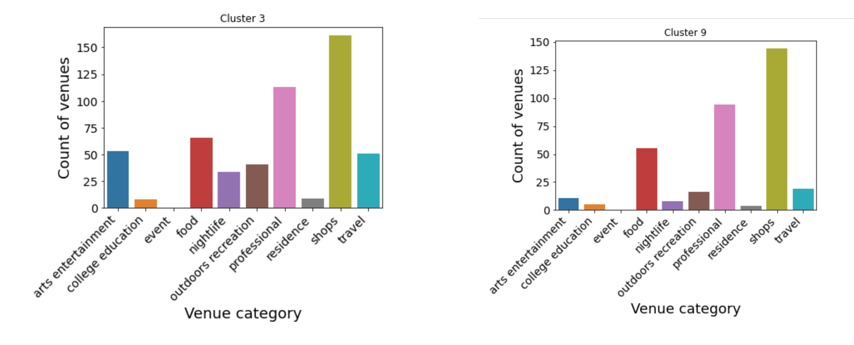
Com a informação das categorias e sabendo a localização dos centroides, já é possível, buscar quantas categorias tem para cada cluster. Após manipulado e organizado em um dataframe o resultado é esse. (Detalhes deste processo no notebook)



Parte do daframe com total de categorias em um raio de 300 metros dos centroides dos clusters

**Resultados**

Uma forma de visualizar quais categorias estão em um raio de 300 metros dos centroides dos clusters 3 e 9 e por meio do gráfico.



Categorias em um raio de 300 metros do centroide

Há uma alta concentração de categoriais como shop e professional nos dois casos, estas categorias possuem estabelecimentos como: Business Center, Distribution Center, Town Hall, Industrial Estate, Wedding Hall, etc. Para saber mais detalhes, é preciso especificar ainda mais a busca, o que certamente ficará para um pouco momento, com estas informações á é possível tirar algumas conclusões.

**Conclusão**

Segundo a Associação Brasileira de Franchising (ABF), a definição da melhor localização é um pouco mais complexa do que aparenta, pois envolve variáveis antagônicas, como fluxo de pessoas e custos. A melhor localização não é necessariamente aquele que proporcionará o maior faturamento, e sim, aquele que trará o melhor resultado, sendo visto nas avaliações dadas pelos clientes.

Vários fatores são levados em conta para abertura de um negócio como fatores de demandas, ofertas e custos, com isso não é possível garantir apenas com as informações extraída quais seriam os melhores lugares para abertura do negócio, apenas conjecturar a partir do algoritmo K-means.

Ao fazer a clusterização o k-means retornou grupos com menores distancias entre si, e verificar estas notas, foi observado os cluster com as maiores médias, tinham ao seu redor categorias como shop e professional, ou seja, uma significa presença de comércios e empresas, o que ajuda a justificar a hipótese de que lugares bem avaliados e bem frequentados estão em pontos estratégicos.