RELATÓRIO TÉCNICO

PROJETO:
LOJA DE SOUVENIR EM COPACABANA

Infnet

Márcia Regina Ferreira Batista
2021

SUMÁRIO

INTRODUÇÃO

Copacabana com sua praia mundialmente conhecida atrai uma quantidade grande de turistas o ano inteiro. Além disso, ocorrem eventos diferenciados que contribuem para aumentar ainda mais a lotação dos hotéis em determinadas épocas do ano.

Sendo assim, combinando a relevância turística com todo o ar boêmio que envolve esse bairro, um grupo de artísticas plásticos manifestou interesse em abrir um comércio em Copacabana. Com isto temos a seguinte questão a ser solucionada: Qual é o melhor lugar para abrir uma loja de souvenir em Copacabana?

Durante a etapa de business understanding, ficaram claros as particularidades do negócio e o anseio dos novos empreendedores.

O novo estabelecimento irá comercializar produtos confeccionados pelos proprietários da loja. Essas peças apresentam um design diferenciado o que acaba acarretando o aumento do preço dos produtos. Dessa maneira, não é necessário levar em consideração a quantidade de lojas já existentes no mesmo ramo pois não serão concorrentes.

Percebemos uma grande concentração de turistas nas proximidades dos hotéis, com isso optamos pela machine learning k-means para agrupar os hotéis e assim, verificarmos as redondezas onde os hotéis estão mais concentrados.

Os dados dos hotéis serão coletados através do Foursquare que é um serviço web com base em geo-localização. Este serviço disponibiliza várias informações do local pesquisado. Após a coleta, limpeza e filtragem dos dados percebemos que existe a atribuição de nota para cada hotel. Este ranking de hotéis deve atrair um maior número de turistas para o hotel que seja melhor avaliado.

O objetivo do projeto é definir a redondeza onde existe uma maior concentração de hotéis, acrescido a isto, iremos avaliar as médias das notas atribuídas em cada agrupamento de hotéis.

MATERIAIS E MÉTODOS

Os dados foram coletados através do serviço web Foursquare que se baseia na geo-localização. É possível acessar dados de lugares através das API disponibilizadas.

O projeto necessita definir uma área de alcance a partir de uma localização central no bairro de Copacabana. O endereço escolhido foi R Domingos Ferreira 6 e o raio de alcance 2.000 metros.

Como o Foursquare necessita das coordenadas de latitude e longitude, foi utilizado o geolocalizador Nominatim para este fim.

A API do Foursquare retorna os dados em formato JSON.


```
{'meta': {'code': 200, 'requestId': '6085570b93d76957ffab6b92'},
 'notifications': [{'item': {'unreadCount': 0}, 'type': 'notificationTray'}],
 'response': {'venues': [{'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/travel/hotel_',
       'suffix': '.png'},
      'id': '4bf58dd8d48988d1fa931735',
      'name': 'Hotel',
      'pluralName': 'Hotels',
      'primary': True,
      'shortName': 'Hotel'}],
    'hasPerk': False,
    'id': '4b058720f964a520128122e3',
    'location': {'address': 'Av. Atlântica, 2600',
     'cc': 'BR',
     'city': 'Rio de Janeiro',
     'country': 'Brasil',
     'distance': 76,
     'formattedAddress': ['Av. Atlântica, 2600',
      'Rio de Janeiro, RJ',
      '22041-001'],
     'labeledLatLngs': [{'label': 'display',
       'lat': -22.9722307,
       'lng': -43.1858418}],
     'lat': -22.9722307,
     'lng': -43.1858418,
     'postalCode': '22041-001',
     'state': 'RJ'},
    'name': 'JW Marriott Hotel Rio de Janeiro',
    'referralId': 'v-1619351307',
    'venuePage': {'id': '182987381'}},
```


Formato Json

Na etapa de pré-processamento, os dados foram transformados para um formato de dataframe com 50 linhas e 19 colunas. Após a análise das informações foram deletadas colunas com informações desnecessárias para o projeto. Abaixo o dataframe com as informações pertinentes aos hotéis:

	id	name	categories	address	hotelLatitude	hotelLongitude
0	4b058720f964a520128122e3	JW Marriott Hotel Rio de Janeiro	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	Av. Atlântica, 2600	-22.972231	-43.185842
1	4bf3fa3dcad2c928b0589b99	Grande Hotel Canadá	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	Av. N.Sa. de Copacabana, 687	-22.971394	-43.187073
2	4f084d40e4b0a2229aaa3317	Olinda Rio Hotel	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	Av. Atlântica, 2230	-22.970240	-43.182959
3	4b058720f964a520048122e3	Majestic Rio Palace Hotel	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	R. Cinco de Julho, 195	-22.971066	-43.190163
4	4b0fe5faf964a520026623e3	Hotel Astoria Palace	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	Av. Atlantica, 1866	-22.968334	-43.180009

Tabela: Dataframe de hotéis

Através da visualização dos dados no mapa abaixo foi possível verificar um hotel localizado no bairro de Botafogo. Como o escopo do projeto inclui somente o bairro de Copacabana, os dados referentes a este hotel foram deletados da base de dados.

Mapa dos hotéis em Copacabana contidos no dataframe deste projeto

O projeto necessita de informações relativas as notas atribuídas a cada hotel. Para obter esta informação foi necessária uma nova busca na API do Foursquare. Neste momento, houve uma limitação de uso do serviço Foursquare pois o acesso está sendo através do cadastro de usuário gratuito, impossibilitando a coleta dos dados. Como a base de dados apresenta agora 49 hotéis, para buscar as notas faz-se necessário realizar 49 consultas. Após repetidas consultas não era mais possível realizar as buscas. Como solução, o resultado da API foi armazenado em um arquivo txt.

Cabe ressaltar que cinco notas de avaliação de hotéis estão ausentes na API. A fim de não descartar dados, as notas faltantes foram incluídas na base de dados após consultas manuais a sites de avaliações de hotéis na web.

O dataframe apresenta os seguintes dados armazenados após as etapas anteriores de limpeza, filtragem e formatação dos dados:

	id	name	categories	address	hotelLatitude	hotelLongitude	rating
0	4b058720f964a520128122e3	JW Marriott Hotel Rio de Janeiro	Hotel	Av. Atlântica, 2600	-22.972231	-43.185842	8.0
1	4bf3fa3dcad2c928b0589b99	Grande Hotel Canadá	Hotel	Av. N.Sa. de Copacabana, 687	-22.971394	-43.187073	6.3
2	4f084d40e4b0a2229aaa3317	Olinda Rio Hotel	Hotel	Av. Atlântica, 2230	-22.970240	-43.182959	6.6
3	4b058720f964a520048122e3	Majestic Rio Palace Hotel	Hotel	R. Cinco de Julho, 195	-22.971066	-43.190163	6.6
4	4b0fe5faf964a520026623e3	Hotel Astoria Palace	Hotel	Av. Atlantica, 1866	-22.968334	-43.180009	6.7

Tabela do dataframe

ANÁLISE DE DADOS – DESENVOLVIMENTO

O projeto busca a solução do problema apresentado utilizando machine learning K-means para agrupar os hotéis em Copacabana em clusters.

K-Means é um algoritmo de clusterização e aprendizado não supervisionado que avalia e agrupa os dados de acordo com suas características.

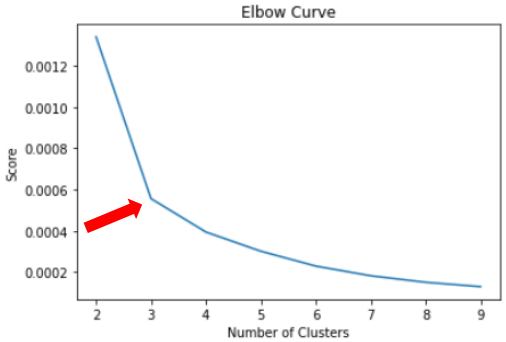
Este projeto utilizou os seguintes parâmetros no K-means:

```
kmeans = KMeans(n_clusters = k, init = 'random', n_init = 100, max_
iter = 300, random_state=123).fit(property_clustering)
```

- n_clusters → número de clusters que serão gerados com os dados
- init → modo de inicialização do algoritmo
    - random → os centróides iniciais serão gerados de forma totalmente aleatória sem um critério para seleção
- max_inter → número máximo de vezes que o algoritmo será executado
- random_state → parâmetro que possibilita a repetição da execução do algoritmo partindo de um mesmo ponto

O valor de K foi definido através de duas métricas: Elbow e Silhouette.

O gráfico Elbow permite visualizar o melhor valor de K como 3 pois é o valor do eixo x onde o gráfico apresenta o “cotovelo”.



O maior valor da métrica Silhouette indica o melhor valor de K correspondente: K = 3

Número de Cluster	Silhouette
2	0,5490
3	0,5643
4	0,4730
5	0,4554
6	0,4532
7	0,4439
8	0,4494
9	0,4716

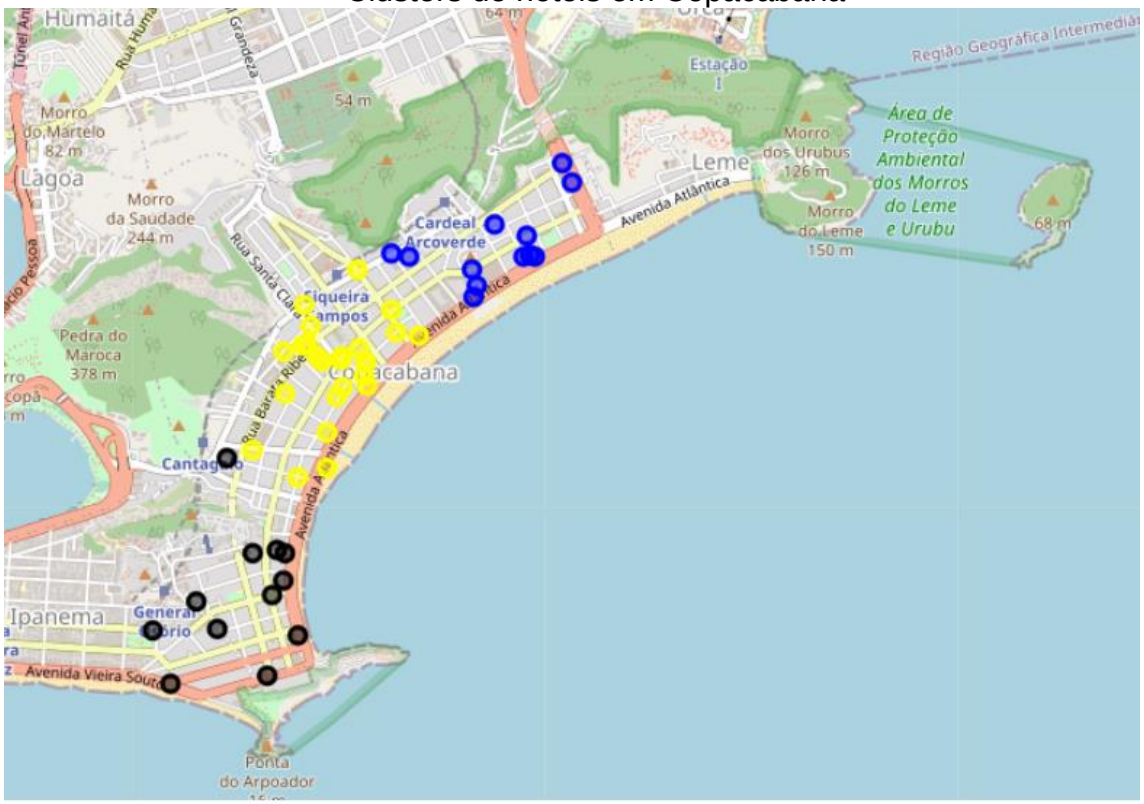
Tabela: valores métrica Silhouette

A tabela abaixo apresenta os valores relativos aos dados dos clusters:

	Contagem de hotéis	Média de notas dos hotéis
Cluster 0	12	6,43
Cluster 1	12	7,40
Cluster 2	25	6,93

Tabela dados dos clusters

Clusters de hotéis em Copacabana



Cluster 0
Cluster 1
Cluster 2

CONCLUSÃO

O projeto visa determinar o melhor local para a abertura da loja de souvenir. Através do método de clusterização k-means, os hotéis foram agrupados a fim de definir uma maior concentração de hotéis e avaliar as médias das notas atribuídas aso hotéis de cada cluster.

Os hotéis de Copacabana foram agrupados em três clusters. Dois desses clusters tinham uma concentração semelhante de hotéis. Outro cluster, representado em amarelo no mapa acima, representa mais do que o dobro do número de hotéis em comparação com os outros clusters. Evidenciando uma alta probabilidade da melhor escolha para a instalação do novo comércio neste cluster (cor amarela no mapa).

Analisando as médias das pontuações dos hotéis por cluster, percebemos que o cluster 1 (cor preta no mapa) apresenta uma concentração de hotéis mais

bem avaliados que os demais. Uma diferença maior é percebida se compararmos os resultados do cluster 1 (preto) e do cluster 0 (azul). No entanto, essa diferença não atinge números que possam impactar profundamente essa análise. A diferença da média de avaliação entre o melhor (preto) e o pior (azul) cluster é de apenas 1 ponto.

A grande diferença do número de hotéis no cluster 2 é mais significativa nessa análise, além da própria variável número de hotéis inicialmente já ser a primeira e principal métrica. Assim a escolha da localização da nova loja é sem dúvida no cluster 2 pela sua maior concentração de hotéis em seus arredores.

A média das notas acabou não sendo um dado relevante para a escolha da localização da loja justamente pela diferença baixa entre os seus valores quando comparado a diferença de concentração de hotéis.

REFERÊNCIAS

https://www.booking.com/index.pt-br.html

https://developer.foursquare.com/docs/places-api/endpoints/