RECOMENDAÇÃO DE HOTÉIS COM BASE NA SIMILARIDADE DE LOCAIS PRÓXIMOS - ABORDAGEM DE CLUSTERIZAÇÃO

A. INTRODUÇÃO

A Yellow Woodpecker (empresa fictícia) é uma agência de viagens brasileira que já atendeu muitos clientes. Em busca de melhorar seus serviços e aumentar a taxa de retenção de seus clientes, a empresa buscou junto a seu time de Data Science uma possível estratégia para melhorar o grau de satisfação de seus clientes. Durante a análise percebeu-se que a satisfação de um mesmo cliente com relação a acomodação das viagens possuía muita variação, e que nem sempre alcançava a satisfação esperada. Diante disso o time de Data Science propôs um projeto de segmentação de todos os hotéis de acomodação de destino da viagem, e então com base na maior satisfação de acomodação de uma viagem anterior, criar um sistema para sugerir Hotéis com base na semelhança de estabelecimentos próximos, buscando assim, aumentar a probabilidade de alta satisfação do serviço de acomodação prestado.

Os dados utilizados neste projeto foram obtidos a partir de requisição à API Foursquare realizada em 3 etapas.

Primeiro realizou-se uma busca em toda a área de uma cidade buscando todos os hotéis existentes nela, e então salvando as seguintes informações:

- - Cidade
- - Latitude da Cidade
- - Longitude da Cidade
- - Nome do Hotel
- - Latitude do Hotel
- - Longitude do Hotel
- - Tipo de Hotel

Em seguida para cada hotel encontrado, realizou-se uma requisição a API para obter a nota de avaliação de cada estabelecimento.

- - Nota de Avaliação

Por fim, para cada hotel, realizou-se uma nova requisição para obter todos os estabelecimentos próximos em um raio de 500 metros e salvar as seguintes informações:

- - Nome do Estabelecimento
- - Latitude do Estabelecimento
- - Longitude do Estabelecimento
- - Tipo do Estabelecimento

B. METODOLOGIA

Para Iniciar o trabalho, utilizou-se a biblioteca geopy para obter as coordenadas lat-long da cidade de Paris - França, que será nossa referência de desenvolvimento. As coordenadas obtidas foram:

- Latitude: 48.8566969
- Longitude: 2.3514616

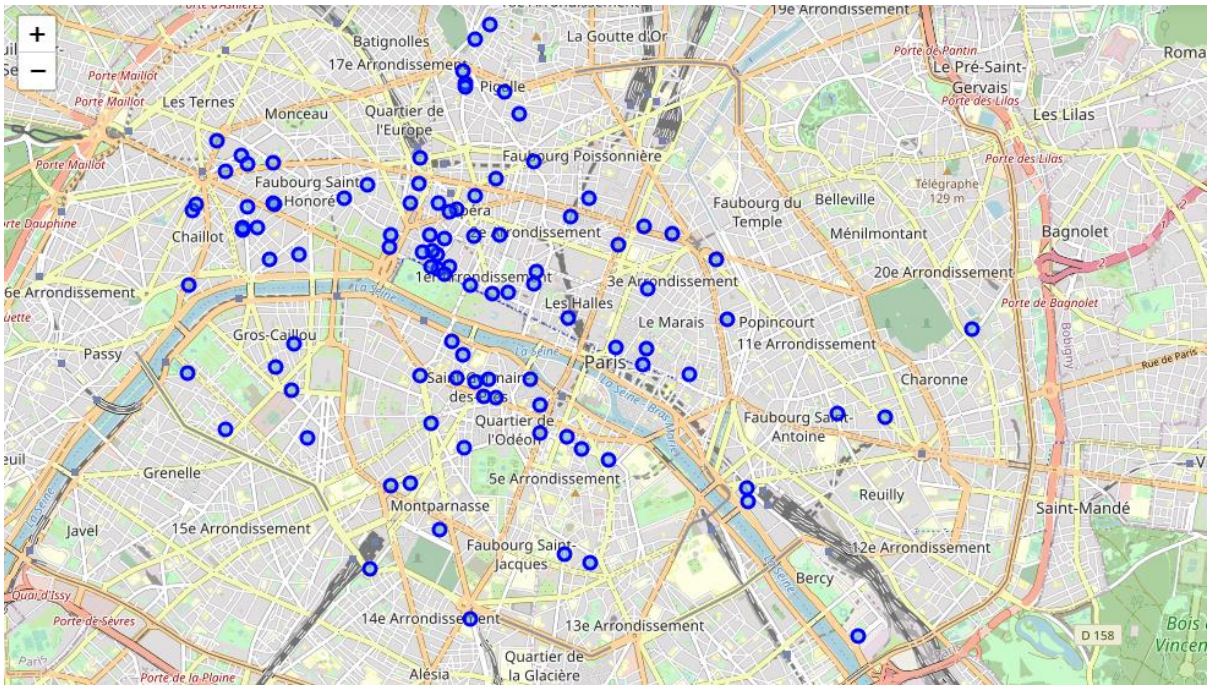
A partir das coordenadas de Paris, utilizou-se a API Foursquare para obter informações de 100 Hotéis em um raio de 10 Km. As informações obtidas foram, nome, número de identificação Foursquare, latitude, longitude e categoria. O cabeçalho do resultado desta consulta é visto na TABELA 1.

TABELA 1 – HOTÉIS EM PARIS

Venue	Venue Id	Venue Latitude	Venue Longitude	Venue Category
Hôtel Le Meurice	4b8b0cbef964a5203b9032e3	48.865333	2.328137	Hotel
Hôtel Bel Ami	4b6ae8dbf964a52066e62be3	48.854918	2.333141	Hotel
InterContinental Paris Le Grand Hôtel	4adcda03f964a520d13121e3	48.870836	2.330725	Hotel
Grand Hôtel du Palais Royal	4b50f4d4f964a5207e3a27e3	48.863183	2.337901	Hotel
W Paris – Opéra	4e8b0ce2f9f464ec8732a56d	48.872098	2.333213	Hotel

Com os dados obtidos anteriormente, foi utilizado a biblioteca python Folium para visualização de localização geográfica de cada Hotel, sendo observado na FIGURA 1.

FIGURA 1 – MAPA HOTÉIS EM PARIS



Após visualização dos Hotéis, utilizou-se requisições premium a API Foursquare para obter as notas de avaliação para cada hotel obtido anteriormente. Tais notas serão utilizadas posteriormente para classificarmos os Hotéis a partir das melhores notas. O cabeçalho do resultado da requisição é visto na TABELA 2.

TABELA 2 – NOTA DE AVALIAÇÃO DOS HOTÉIS (RATINGS)

Venue	Venue Id	Ratings	Venue Latitude	Venue Longitude	Venue Category
Hôtel Le Meurice	4b8b0cbef964a5203b9032e3	8.7	48.865333	2.328137	Hotel
Hôtel Bel Ami	4b6ae8dbf964a52066e62be3	8.7	48.854918	2.333141	Hotel
InterContinental Paris Le Grand Hôtel	4adcda03f964a520d13121e3	8.6	48.870836	2.330725	Hotel
Grand Hôtel du Palais Royal	4b50f4d4f964a5207e3a27e3	8.3	48.863183	2.337901	Hotel
Hôtel Les Jardins du Marais	4b7051cdf964a5205b122de3	7.8	48.860699	2.368530	Hotel

Em sequência, utilizou-se novamente requisições a API Foursquare para obter até 100 estabelecimentos próximos em um raio de 300 metros para cada hotel. O resultado desta requisição é uma lista de estabelecimentos próximos a cada hotel analisado salvos na TABELA 3.

TABELA 3 – LOCAIS PRÓXIMOS

	Hotel	Hotel Latitude	Hotel Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hôtel Le Meurice	48.865333	2.328137	Librairie Galignani	48.864989	2.328570	Bookstore
1	Hôtel Le Meurice	48.865333	2.328137	Hôtel Mandarin Oriental	48.866987	2.327178	Hotel
2	Hôtel Le Meurice	48.865333	2.328137	Balagan	48.865432	2.329680	Israeli Restaurant
3	Hôtel Le Meurice	48.865333	2.328137	Le Dalí	48.865333	2.328137	French Restaurant
4	Hôtel Le Meurice	48.865333	2.328137	Ladurée	48.866121	2.328449	Dessert Shop

Nem todos os algoritmos de machine learning são capazes de trabalhar diretamente com variáveis categóricas, pois necessitam realizar cálculos numéricos sobre elas. Portanto é necessário buscar uma forma de representar numericamente os dados obtidos. Neste projeto utilizou-se ‘One Hot Encoding’, que é uma forma de representação numérica para dados categóricos que não possuem uma relação ordinal, onde é atribuído valor 1 quando há observação para um dado grupo e 0 para quando não há nenhuma ocorrência.

Após isso realizou-se o agrupamento dos dados em ‘One Hot Encoding’ para cada Hotel através da frequência média encontrada para cada categoria. As primeiras linhas e colunas da tabela após tal transformação é dada pela TABELA 4.

TABELA 4 – FREQUÊNCIA MÉDIA

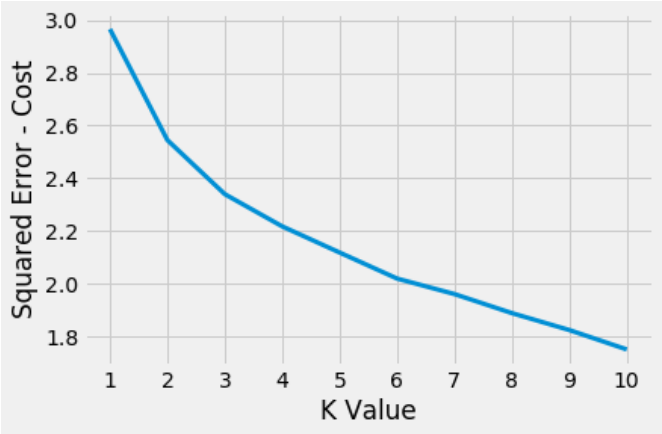
	Hotel	Accessories Store	African Restaurant	Alsatian Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports
0	7 Eiffel Hotel****	0.0	0.0	0.0	0.014085	0.0	0.0	0.000000	0.014085	0.0	0.0	0.014085	0.0
1	Artus Hotel	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.0
2	Champs Elysées Friedland	0.0	0.0	0.0	0.000000	0.0	0.0	0.023810	0.000000	0.0	0.0	0.047619	0.0
3	Citadines Saint-Germain-des-Prés Paris	0.0	0.0	0.0	0.000000	0.0	0.0	0.019608	0.000000	0.0	0.0	0.039216	0.0
4	Cordelia Hotel	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.015152	0.0

Com a tabela de frequências devidamente obtida, iniciou-se a etapa de Aprendizado de Máquina, utilizando a técnica de aprendizado não Supervisionado K-Means, que é um algoritmo que agrupa dados tentando separar amostras em n grupos de igual variância, minimizando o critério conhecido como inércia ou através do erro da soma dos quadrados de um cluster (grupo). Neste algoritmo é necessário especificar previamente o número k de clusters (grupos) na qual os dados serão agrupados. Com isso é necessário buscar encontrar a melhor quantidade k de clusters que resulte na maior semelhança possível entre dados de um mesmo cluster e ao mesmo tempo na menor semelhança possível entre os diferentes cluster.

Para realizar tal tarefa utilizou-se da criação de vários modelos com valores k de 1 a 10, então utilizou-se duas métricas para avaliação dos resultados, sendo elas, Método de Elbow e Método da Silhueta.

O primeiro método utilizado foi o Método de Elbow, que também é conhecido como método do cotovelo. Esse método testa a variância dos dados em relação ao número de clusters, de forma a apresentar o ponto k no qual o aumento do número de clusters não representa um valor significativo de diminuição do custo se comparado com a variância entre os k anteriores. O resultado de tal método é visto na FIGURA 2.

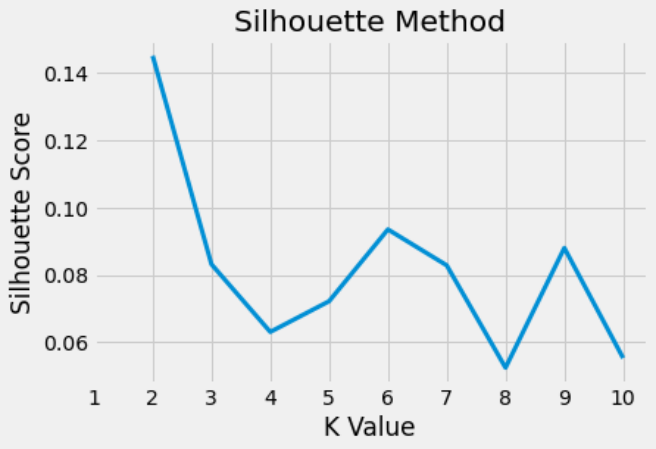
FIGURA 2 – MÉTODO DE ELBOW



Em alguns casos (como o acima), o método de Elbow pode não representar claramente qual o melhor k a ser escolhido, pois este método trata-se de uma inspeção visual dos valores, e que nem sempre é muito claro. O caso acima por exemplo, não deixa muito claro qual k é melhor, apenas observa-se que este valor está entre 2 e 6. Portanto o ideal é ter mais de uma forma de avaliação.

A segunda avaliação da escolha k foi realizada através do método da silhueta, que é utilizado para medir a distância de separação resultante entre dois clusters. O coeficiente de Silhueta próximo de 0 representa que a amostra está em, ou muito próxima da fronteira entre dois clusters vizinhos. Já um valor próximo de 1 indica que a amostra está muito longe de clusters vizinhos. O resultado deste método é visto na FIGURA 3.

FIGURA 3 – MÉTODO DA SILHUETA



Com o gráfico da FIGURA 3 observamos o maior coeficiente de Silhueta foi alcançado para k = 2, resultando assim em 2 clusters, porém somente 2 clusters para nossa aplicação não teria muita utilidade, pois realizar uma sugestão de hóteis baseado em somente 2 grupos, limita bastante a variedade de escolha entre diferentes cluster. Portanto neste projeto utilizaremos o segundo melhor k (k = 6) segundo o método de Silhouette que também coincide com os possíveis bons valores observados no método de Elbow. Com isso treinamos o nosso modelo K-Means com 6 clusters.

C. RESULTADOS

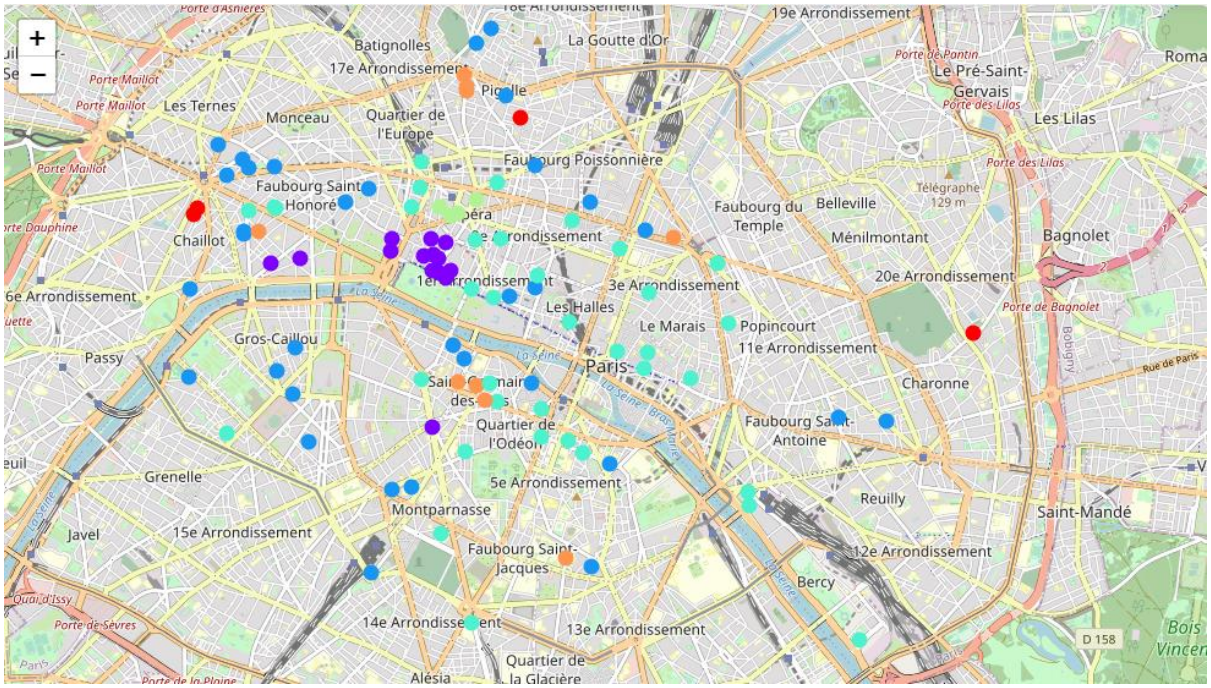
Com o resultado do agrupamento K-Means, conseguimos obter os rótulos de cluster pertencente para cada hotel, e então mesclou-se tal resultado com a tabela dos 10 locais mais frequentes para cada hotel, resultando na TABELA 5.

TABELA 5 – CLUSTER LABELS

Hotel	Hotel Latitude	Hotel Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Hôtel Le Meurice	48.865333	2.328137	1	French Restaurant	Café	Hotel Bar	Chocolate Shop	Japanese Restaurant	Pastry Shop	Beer Garden	Clothing Store	Bookstore	Dessert Shop
Hôtel Bel Ami	48.854918	2.333141	5	Italian Restaurant	French Restaurant	Plaza	Café	Japanese Restaurant	Cosmetics Shop	Clothing Store	Boutique	Sandwich Place	Steakhouse
InterContinental Paris Le Grand Hôtel	48.870836	2.330725	4	Chocolate Shop	Coffee Shop	Clothing Store	Men's Store	Electronics Store	Sandwich Place	Candy Store	French Restaurant	Furniture / Home Store	Roof Deck
Grand Hôtel du Palais Royal	48.863183	2.337901	2	French Restaurant	Plaza	Café	Historic Site	Coffee Shop	Theater	Shoe Store	Spa	Sculpture Garden	Garden
Hôtel Les Jardins du Marais	48.860699	2.368530	3	Bar	Café	French Restaurant	Restaurant	Art Gallery	Clothing Store	Pizza Place	Coffee Shop	Tea Room	Bistro

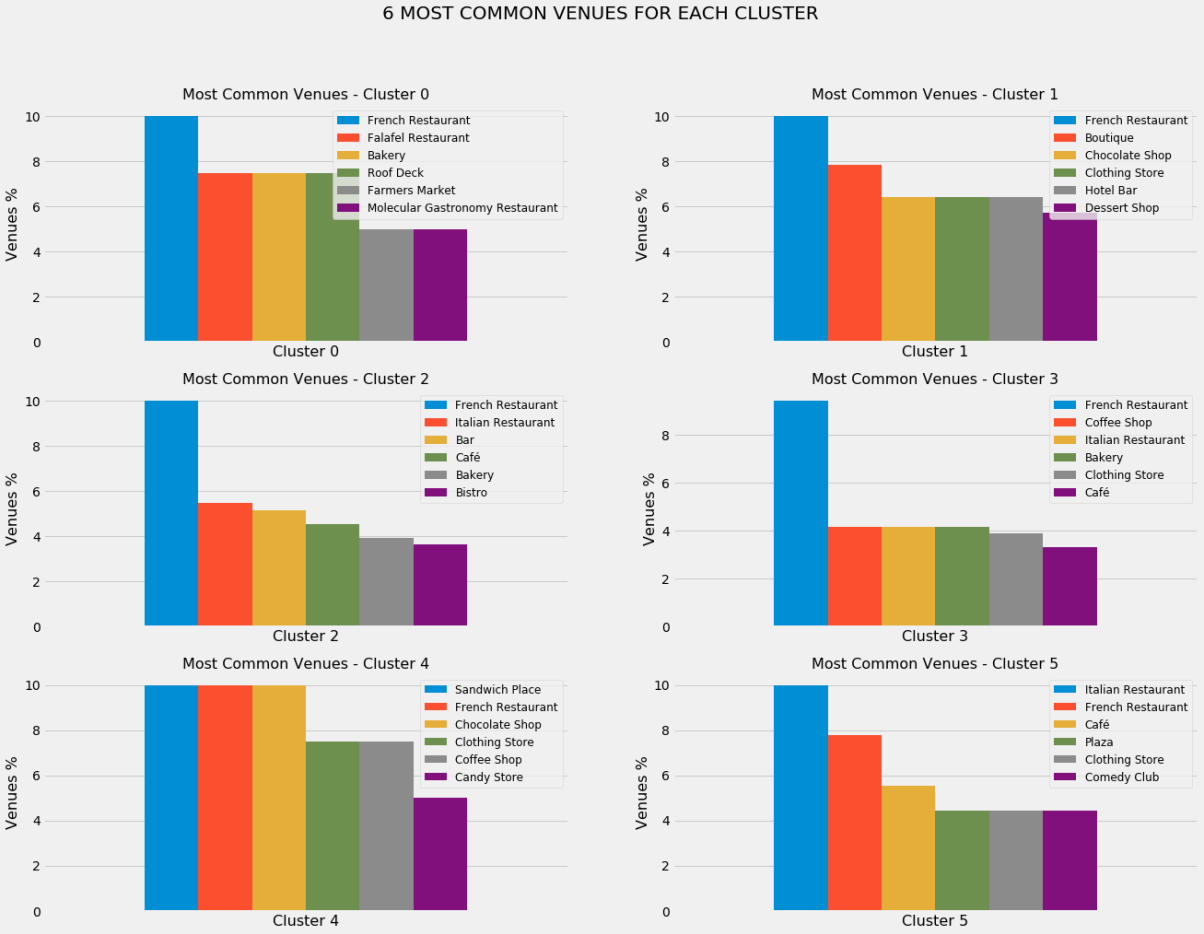
A partir do pacote python Folium, podemos visualizar o resultado dos agrupamentos dos hotéis na FIGURA 4, onde cada cor representa um dos 6 clusters.

FIGURA 4 – MAPA HÓTEIS EM PARIS AGRUPADOS



Para extrair um pouco mais de informações que representam cada cluster, realizou-se um agrupamento de locais por cada cluster e plotou-se o gráfico da FIGURA 5, que representa os 6 locais mais comuns para cada cluster.

FIGURA 5 – 6 LOCAIS MAIS COMUNS PARA CADA CLUSTER



Observando o gráfico da FIGURA 5, observou-se que a categoria 'French Restaurant' aparece entre os mais comuns em todos os cluster, sendo o 1º ou 2º local mais comum. Essa informação é interessante e um tanto inesperada, pois há uma categoria específica para restaurantes franceses em uma cidade Francesa, o que não foi observado semelhantemente em Toronto, pois não se notou nenhum local classificado como Restaurante Canadense.

Como existe tal classificação específica para restaurante franceses, e tal categoria sugestivamente está entre os mais comuns dentro da cidade de Paris na França, talvez seja interessante retirá-los das informações utilizadas para realização do agrupamento, pois talvez não representem informação útil, já que é esperado que se tenha restaurantes franceses em volta da maioria dos hotéis na França.

D. DISCUSSÃO

Agora que se tem todos os hotéis devidamente agrupados, pode-se realizar a aplicação proposta neste projeto. Então, digamos que um cliente já tenha visitado Toronto, e ficou extremamente satisfeito com o seu local de hospedagem atribuindo uma nota 9.7, podemos agora utilizar os dados de sua antiga viagem para sugerir um hotel na cidade de Paris que seja semelhante ao de Toronto.

As informações de localização do hotel em Toronto, são mostradas na TABELA 6.

TABELA 6 – HOTEL EM TORONTO

	City	City Latitude	City Longitude	Venue	Venue id	Venue Latitude	Venue Longitude	Venue Category
0	Toronto	43.653482	-79.383935	Marriott Downtown at CF Toronto Eaton Centre	4b0563c0f964a5200e5822e3	43.654728	-79.382422	Hotel

A partir da localização do hotel em Toronto, realizou-se o mesmo processo de preparação e coleta de dados aplicados aos hotéis de Paris.

Com os dados do hotel de Toronto devidamente preparados, através do modelo K-Means, classificou-se o hotel a um cluster com a maior semelhança encontrada dentre os hotéis em Paris, o resultado foi a classificação do hotel ao cluster 3.

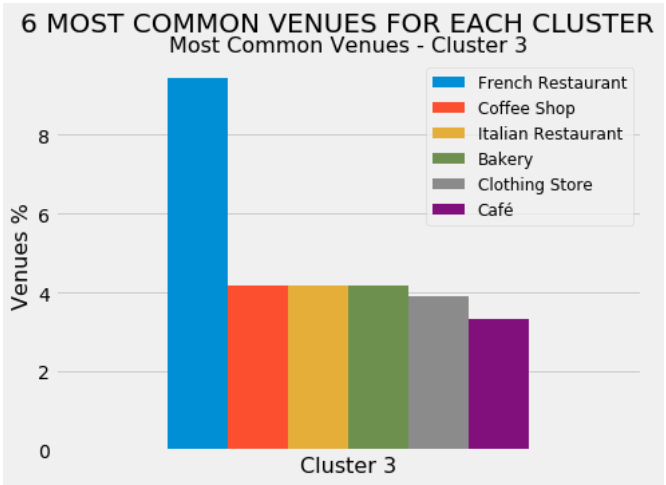
Diante desta informação pode-se sugerir os 5 melhores hotéis em Paris que tenham locais próximos semelhantes ao de Toronto. A sugestão é vista na TABELA 7.

TABELA 7 – HOTÉIS RECOMENDADOS

	Hotel	Ratings	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
27	The Hoxton Paris	9.1	3	Bar	Cocktail Bar	Wine Bar	French Restaurant	Italian Restaurant	Chinese Restaurant	Theater	Salad Place	Gym / Fitness Center	Indian Restaurant
11	Hôtel Barrière Le Fouquet's	8.9	3	French Restaurant	Asian Restaurant	Café	Pastry Shop	Cosmetics Shop	Clothing Store	Electronics Store	Bakery	Hotel Bar	Halal Restaurant
22	Hotel Atmospheres	8.8	3	French Restaurant	Italian Restaurant	Bakery	Coffee Shop	Tapas Restaurant	Portuguese Restaurant	Seafood Restaurant	Ethiopian Restaurant	Flower Shop	Market
23	Hôtel Jules & Jim	8.8	3	French Restaurant	Chinese Restaurant	Café	Art Gallery	Wine Bar	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Restaurant	Museum
20	Hôtel Caron de Beaumarchais	8.7	3	French Restaurant	Clothing Store	Falafel Restaurant	Pastry Shop	Wine Bar	Plaza	Italian Restaurant	Ice Cream Shop	Bistro	Furniture / Home Store

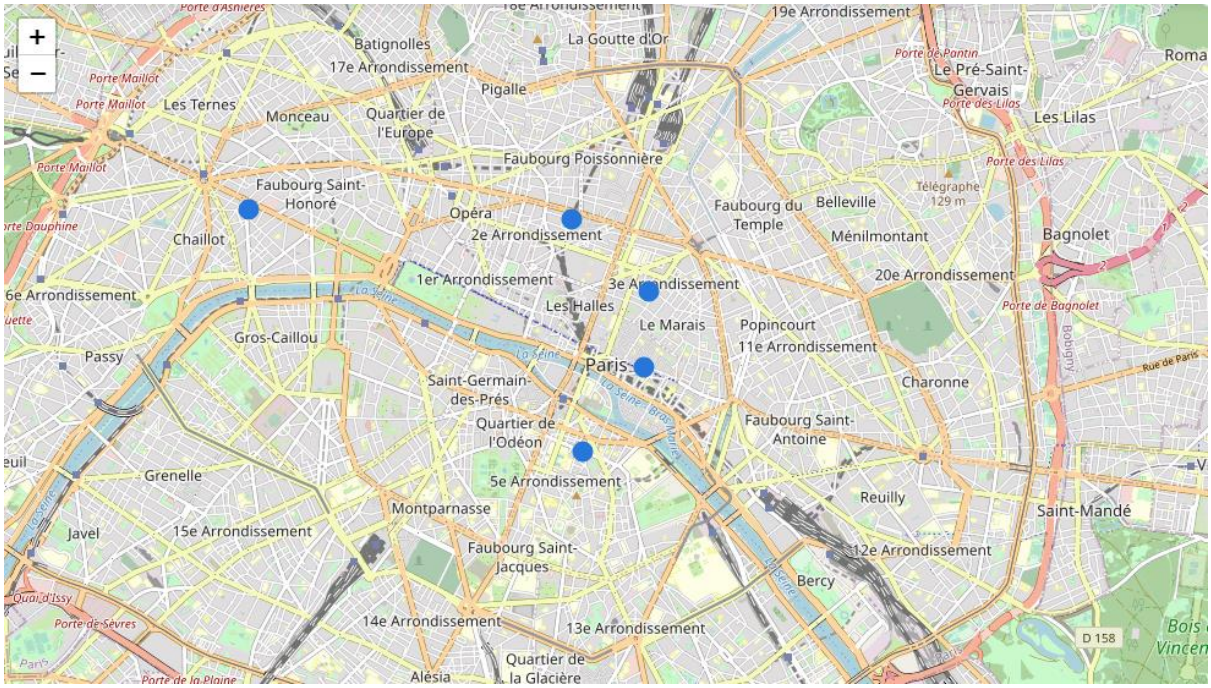
Na FIGURA 6, observa-se os locais mais frequentes para o cluster previsto.

FIGURA 6 – 6 LOCAIS MAIS FREQUENTES PARA O CLUSTER 3



E por último na FIGURA 7, pode-se visualizar os Hotéis sugeridos.

FIGURA 7 – MAPA HOTÉIS RECOMENDADOS



E. CONCLUSÃO

Com os dados obtidos via requisições ao Foursquare conseguimos coletar e agrupar vários hotéis da cidade de Paris, e então com dados históricos sobre um dado cliente, conseguimos realizar a sugestão dos melhores hotéis para hospedagem a partir de tais dados.

Apesar do resultado satisfatório para uma primeira versão deste projeto, ainda há várias e várias coisas que podem ser melhoradas em trabalhos futuros, como, a utilização de mais de uma fonte de dados, conseguir acesso a uma maior quantidade de requisições a API do Foursquare, utilização de outros tipos de métricas para avaliarmos o nosso modelo e escolha do melhor número de clusters, realizar a segmentação para um conjunto que contenha mais de uma cidade, enfim, as possibilidades são várias, portanto resta agora continuar os estudos e ir melhorando os projetos desenvolvidos com os conhecimentos adquiridos ao longo do tempo.

Obrigado por dedicar seu tempo lendo este trabalho :)

F. AGRADECIMENTOS

Meus agradecimentos a plataforma de cursos Coursera, a IBM e a todos os instrutores dos cursos da Certificação Profissional em Ciência de Dados, por disponibilizar tal curso e materiais que possibilitaram o desenvolvimento deste projeto.

G. REFERÊNCIAS

- IBM Data Science Professional Certificate

- One Hot Encode
- An Introduction to Clustering and different methods of clustering
- K-Means - Scikit-learn
- Entenda o algoritmo k-means
- Selecting the number of clusters with silhouette analysis on KMeans clustering
- Forusquare - Documentation for Developers
- Pandas - User Guide
- Folium - Documentation
- A lot of things at StackOverflow :)