IBM Data Science Capstone Project

Final Report

Fadel Victor Shanaa

22.01.2021

fvshanaa@gmail.com

Determining the Optimal Location for Building a Fine Dining Restaurant in Moscow, Russia

Introduction

Fine dining restaurants are a very high class of restaurants which feature prominent décor made of high-quality materials, high prices on the food, and an etiquette that must be followed by staff and visitors alike. Fine dining has been around since the late 18th century and is very popular in Western Europe and parts of North America. The popularity is due to not only the exquisite tastes and décor, but also the geographical positioning of the restaurant. They are typically located in upscale areas of a city. This project will be about a restaurateur trying to open a fine dining restaurant in the city of Moscow. For them to maximize profits, the restaurant must be positioned in the best possible district.

Business Problem

As mentioned in the introduction, this project will focus on a restaurateur trying to open a high-end fine dining restaurant in Moscow, Russia. Fine dining is not as popular in Russia as it is in the West. The cuisine will be a mix of Italian and French. The goal will be to try and find the best district for the restaurant to open given its price level and type of cuisine.

The restaurant must attract the right clientele which will primarily consist of wealthy businessmen, politicians, and tourists. From a technical perspective, this project will have us use a clustering algorithm that will cluster the different districts of Moscow into different groups. For the purposes of this project, we will cluster the districts of Moscow into several groups. The best group should be one that has many gastronomical venues and plenty of sources of fresh ingredients, thus maximizing revenue and minimizing costs on food (which is our raw material in this case).

Theoretically, this restaurant should be located either as close to the Kremlin as possible or in the business district of Moscow. Our Target Audience is the clientele since they will be dining there, the investors since we will need start-up capital, and finally other data scientists who will be interested in seeing how the problem of opening a restaurant might be solved using a data driven approach.

Data

The data we will use will come from Wikipedia. Specifically, it will come from the following link:

https://en.wikipedia.org/wiki/Category:Districts_of_Moscow

The above link contains all the districts of Moscow. We will use pandas to parse and store the content of that page. We will use the geocoder library to obtain the latitude and longitude values of those districts. Finally, we will use the Foursquare API to get the data about the different venues located in those districts.

Methodology

Part A: Exploratory Data Analysis

For the purposes of this task, I simply pasted the districts of Moscow into a text file as follows

```
Aeroport District
Akademichesky District
Alexeyevsky District, Moscow
Altufyevsky District
Babushkinsky District, Moscow
Basmanny District
Begovoy District
Beskudnikovsky District
Bibirevo District
Biryulyovo Vostochnoye District
Biryulyovo Zapadnoye District
Bogorodskoye District
Businovo District
Butyrsky District
Chertanovo Severnoye District
Chertanovo Tsentralnoye District
Chertanovo Yuzhnoye District
Cheryomushki District
Danilovsky District, Moscow
Dmitrovsky District, Moscow
Donskoy District
Dorogomilovo District
Fili-Davydkovo District
Filyovsky Park District
Gagarinsky District, Moscow
Golovinsky District
Golyanovo District
```

Then, I read this data into a pandas DataFrame using the following command

```
"""
Moscow district names are stored in district_list.txt file
"""
district_df = pd.read_csv('district_list.txt', header=None)
district_df.columns = ['Districts']
district_df.head()
```

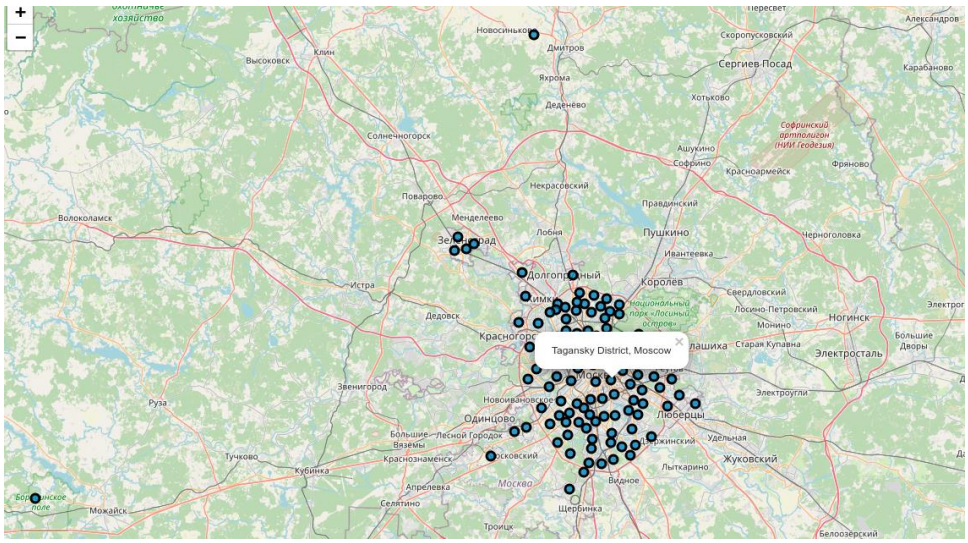
The next stage was to obtain the geocoordinates, latitude and longitude, of the districts. The python library geocoder came in handy.

```
failed_locations = []
for index, row in tqdm(district_df.iterrows()):
    address = row['Districts'] + ', Moscow'
    try:
        location = geolocator.geocode(address)
        latitude = location.latitude
        longitude = location.longitude
        print('The geograpical coordinate of {} are {}, {}.'.format(address, latitude, longitude))
        district_df['Latitude'].iloc[index] = latitude
        district_df['Longitude'].iloc[index] = longitude
    except AttributeError:
        print("Can't find {}. Position {} will be empty".format(address, index))
        failed_locations.append(index)
```

```
district_df
```

	Districts	Latitude	Longitude
0	Aeroport District	55.8026	37.5403
1	Akademichesky District	55.6894	37.578
2	Alexeyevsky District	55.8149	37.6507
3	Altufyevsky District	55.8803	37.5816
4	Babushkinsky District	55.866	37.6639
5	Basmanny District	55.7673	37.6698
6	Begovoy District	55.7819	37.5663
7	Beskudnikovsky District	55.8671	37.5544
8	Bibirevo District	55.8953	37.613
9	Biryulyovo Vostochnoye District	55.5961	37.6753
10	Biryulyovo Zapadnoye District	55.588	37.6363
11	Bogorodskoye District	55.8199	37.7029
12	Businovo District	55.8807	37.4938
13	Butyrsky District	55.8138	37.5929
14	Chertanovo Severnoye District	55.6331	37.6052
15	Chertanovo Tsentralnoye District	55.6152	37.6033
16	Chertanovo Yuzhnoye District	55.5898	37.5957
17	Cheryomushki District	55.6635	37.5611

The next step was plotting the geospatial data using Folium, thus visualising our results so far



Our interest however lies not only with the districts but the venues which are in those districts. To get the venue data, the Foursquare API came in handy.

The table DataFrame looked like this

```
[296]: print(moscow_venues.shape)
       moscow_venues.head(50)
```

```
(2227, 7)
```

[296]:

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aeroport District	55.802618	37.540297	Аптека на Старом Зыковском	55.801950	37.545025	Pharmacy
1	Aeroport District	55.802618	37.540297	ВкусВилл	55.804333	37.535205	Health Food Store
2	Aeroport District	55.802618	37.540297	Отдохни	55.805900	37.536928	Liquor Store
3	Aeroport District	55.802618	37.540297	Cosmic Latte	55.799551	37.534983	Coffee Shop
4	Aeroport District	55.802618	37.540297	Круглосутка	55.802615	37.533316	Convenience Store
5	Aeroport District	55.802618	37.540297	Динозаврик	55.801778	37.536579	Pet Store
6	Aeroport District	55.802618	37.540297	Dali (Дали)	55.802846	37.537420	Salon / Barbershop
7	Aeroport District	55.802618	37.540297	Busido	55.799815	37.543609	Martial Arts School
8	Aeroport District	55.802618	37.540297	ПРАВДА КОФЕ	55.803584	37.534087	Coffee Shop
9	Aeroport District	55.802618	37.540297	Kuzina	55.804354	37.534308	Dessert Shop
10	Aeroport District	55.802618	37.540297	Кулинарная лавка братьев Караваевых	55.801790	37.532737	Deli / Bodega
11	Aeroport District	55.802618	37.540297	Подружка	55.804445	37.535377	Cosmetics Shop
12	Aeroport District	55.802618	37.540297	Фитнес Практика	55.799728	37.543701	Gym / Fitness Center
13	Aeroport District	55.802618	37.540297	Coffee Bean	55.798437	37.537411	Coffee Shop
14	Aeroport District	55.802618	37.540297	Булочная Фокина №12	55.802957	37.533295	Bakery
15	Aeroport District	55.802618	37.540297	Горздрав	55.804181	37.534188	Pharmacy
16	Aeroport District	55.802618	37.540297	Ароматный мир	55.803879	37.534106	Wine Shop
17	Aeroport District	55.802618	37.540297	Остановка ул. Черняховского	55.806251	37.537388	Bus Stop
18	Aeroport District	55.802618	37.540297	137	55.804714	37.533368	Basketball Court
19	Akademichesky District	55.689359	37.577971	Вкусвилл	55.686904	37.575196	Health Food Store
20	Akademichesky District	55.689359	37.577971	Винный буфетъ	55.691166	37.575811	Wine Bar
21	Akademichesky District	55.689359	37.577971	Billy McDaniel	55.688104	37.571608	Pub
22	Akademichesky District	55.689359	37.577971	Здоров.ру	55.687911	37.571558	Pharmacy
23	Akademichesky District	55.689359	37.577971	Сквер «200 лет А. С. Пушкину»	55.687814	37.575538	Park

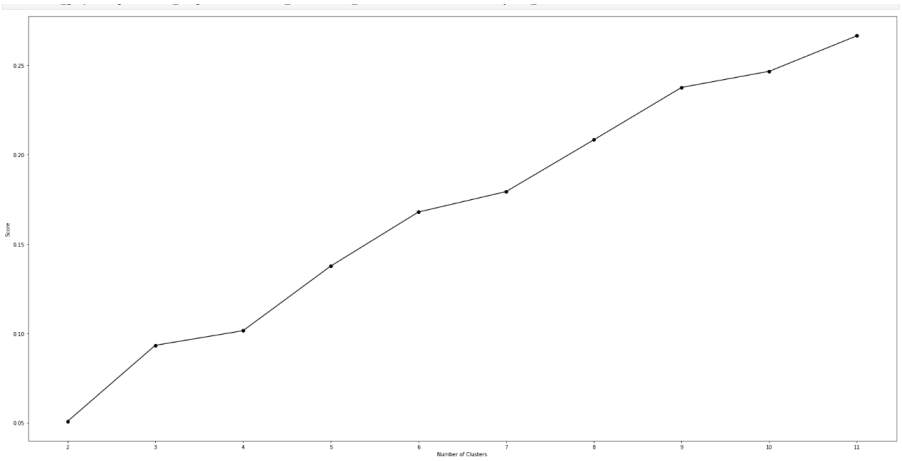
\

This DataFrame contained critically important data such as venue type, coordinates, and the district it was located in. There was a total of 174 unique venue categories we managed to extract. Our main interest lies with the top 15 venues in each district. After using one hot encoding, which encodes each categorial variable (in our case the venue type) and some additional DataFrame manipulation, the resulting DataFrame was as follows:

	District	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue
0	Aeroport District	Pharmacy	Film Studio	Event Space	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Zoo Exhibit	Escape Room	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain
1	Akademichesky District	Health Food Store	Zoo Exhibit	Film Studio	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Fish Market	Escape Room	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain
2	Alexeyevsky District	Liquor Store	Film Studio	Event Space	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Fish Market	Frozen Yogurt Shop	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain
3	Altufyevsky District	Coffee Shop	Zoo Exhibit	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Fish Market	Escape Room	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain
4	Babushkinsky District	Convenience Store	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain

Part B: Clustering

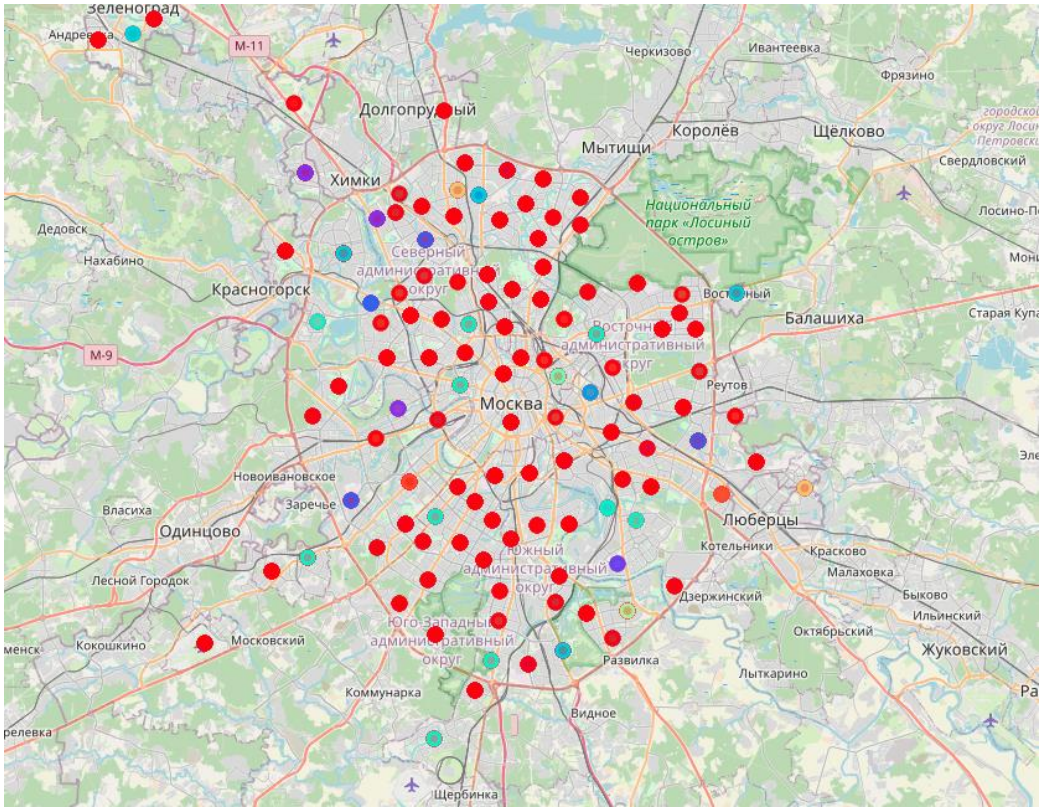
For the purposes of clustering our districts into distinct groups, the k-means algorithm was utilized. K-Means has an evaluation metric known as silhouette score. The higher the silhouette score, the better.



We can see that 11 clusters yields the highest silhouette score

In our case, the best silhouette score was obtained when the number of clusters is set to 11.

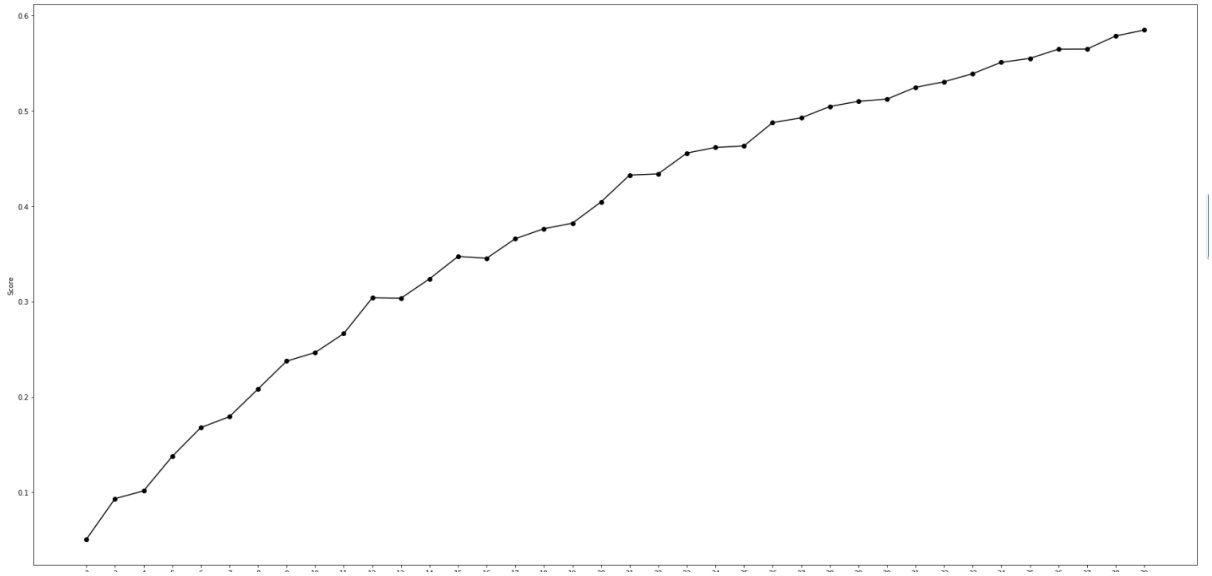
After training the k-means algorithm on our data and splitting it into clusters, we once again used Folium to visualize the different clusters:



Important Note About K-Means:

This algorithm has limitations. Our data has outliers which k-means is not very well suited to deal with. On top of that, the silhouette score does not plateau when n_clusters=11, it keeps rising as evidenced from the following graph

```
silhouette_graph(Range=cluster_range, Scores=score_list, xaxis_lab="Number of Clusters", yaxis_lab="Score")
```



Due to lack of computational power, we had to limit the number of clusters we could work with. From the above graph, we can see that the best silhouette score would be obtained at an n_clusters value of at least 40.

Results

The following are some samples of the obtained results. We filtered the final DataFrame based on cluster values and obtained the following.

Cluster 1

```
[413]: # Cluster 1
cluster_0 = moscow_final_df.loc[moscow_final_df['label'] == 0, moscow_final_df.columns[list(range(2, moscow_final_df.shape[1]))]]
cluster_0[0:5]
```

[413]:

	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	label	District
1	Health Food Store	Zoo Exhibit	Film Studio	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Fish Market	Escape Room	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain	0	Akademichesky District
2	Liquor Store	Film Studio	Event Space	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Fish Market	Frozen Yogurt Shop	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain	0	Alexeyevsky District
4	Convenience Store	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	0	Babushkinsky District
5	Pet Store	Film Studio	Event Space	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Fish Market	Czech Restaurant	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain	0	Basmanny District
6	Salon / Barbershop	Zoo Exhibit	Fast Food Restaurant	Event Space	Exhibit	Fabric Shop	Farm	Farmers Market	Film Studio	Electronics Store	Fish Market	Flower Shop	Food & Drink Shop	Food Court	Forest	0	Begovoy District

Cluster 1 districts have no clear dominant venue; however, zoos and film studios seem to be quite common

Cluster 2

[401]:
```
# Cluster 2
cluster_1 = moscow_final_df.loc[moscow_final_df['label'] == 1, moscow_final_df.columns[list(range(2, moscow_final_df.shape[1]))]]
cluster_1[0:20]
```

[401]:

	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	label	District
26	Sushi Restaurant	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	1	Golyanovo District
37	Sushi Restaurant	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	1	Krasnoselsky District
110	Sushi Restaurant	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	1	Vykhino-Zhulebino District
284	Sushi Restaurant	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	1	NaN
286	Sushi Restaurant	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	1	NaN
478	Sushi Restaurant	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	1	NaN
524	Sushi Restaurant	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	1	NaN
571	Sushi Restaurant	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	1	NaN

Cluster 2 has a high concentration of expensive sushi restaurants as well as fish markets, farms, and fast-food restaurants. Wealthy clientele and short distance to food sources makes for an ideal restaurant location.

Cluster 8

[407]:
```
# Cluster 8
cluster_8 = moscow_final_df.loc[moscow_final_df['label'] == 7, moscow_final_df.columns[list(range(2, moscow_final_df.shape[1]))]]
cluster_8[0:20]
```

[407]:

	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	label	District
11	Cosmetics Shop	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	7	Bogorodskoye District
31	Cosmetics Shop	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	7	Khoroshyovsky District
63	Cosmetics Shop	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	7	Novo-Peredelkino District
113	Cosmetics Shop	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	7	Yasenevo District
118	Cosmetics Shop	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Escape Room	Food & Drink Shop	Food Court	Forest	Fountain	7	Zapadnoye Degunino District

Cluster 8 districts are heavily focused on cosmetics and recreational activities. A fine dining restaurant built here is unlikely to get a lot of attention.

Cluster 11

	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	label	District
14	Bakery	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain	French Restaurant	10	Chertanovo Severnoye District
27	Bakery	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain	French Restaurant	10	Ivanovskoye District
30	Bakery	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain	French Restaurant	10	Khoroshyovo-Mnyovniki District
53	Bakery	Zoo Exhibit	Fish Market	Exhibit	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Flower Shop	Food & Drink Shop	Food Court	Forest	Fountain	French Restaurant	10	Metrogorodok District

Cluster 11 has a high concentration of food sources, bakeries, and cafes thus making it the second-best spot for our restaurant.

Discussion and Recommendation

From the data we obtained, the top two most suitable locations for the construction of our European Fine Dining restaurant would be the districts in either Cluster 2 or Cluster 11. Cluster 2 districts have a lot of sushi restaurants which are generally quite expensive thus guaranteeing us wealthy clientele, while Cluster 11 has a high concentration of food places such as bakeries, farms, fish markets, and fast-food restaurants. Between these two clusters, Cluster 2 would be most optimal since the proximity of farms and fish markets would make access to fresh ingredients easier as well, on top of the large revenue generated thanks to the wealthy clientele that likely resides there or visits it.

Conclusion

This project was done through extensive use of Jupyter notebooks, the Python programming language, and key libraries such as folium, pandas, numpy, and geocoder. The Foursquare API was also instrumental in giving us the venue location that we needed. This was a relatively challenging task and the use of API's and real-world data made it that much more realistic and interesting.

References

Pandas Documentation: https://pandas.pydata.org/pandas-docs/stable/index.html

Scikit Learn Documentation: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette Analysis: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html