Assignment

The Battle of Neighbourhoods

Moscow

The Battle of multicultural gastronomic variety

Report

Introduction

Discuss the business problem and who would be interested in this project.

Moscow Metro has more than 250 stations and due to constant road problems and massive traffic jams is the best way of transportation so if you decide to go out metro is often the best way to get to a place. But which metro station is the best if you want to have a nice restaurant experience with diverse chose of restaurants with many different national and international cuisines?

The main purpose of this project is to determine which neighbourhoods surrounding which metro stations in Moscow offers the biggest number of restaurants with the largest variety of national and international cuisines. In other words we are looking for a metro station that can offer us the most multicultural gastronomic diversity experience and choice when deciding on a restaurant.

This information can be used by both tourists and locals to chose a metro station from where they will have the biggest number of choices for a restaurant.

Another purpose of this project is on the contrary to determine Moscow neighbourhoods with the smallest diversity and/or lack of certain ethic foods and restaurants representation.

This information can assist business people in opening a new restaurant and deciding which national or international cuisine to go for and where to open it.

We will use Foursquare location data with basic descriptive statistics as well as some unsupervised machine learning with cluster analysis to classify stations and derive the information and conclusions we need.

All restaurants must be located within just 500 meters walk from the metro station so it could be an easy walk.

Data

Describe the data that you will be using to solve the problem or execute your idea and the source of the data. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea.

The base list of metro stations and their geographical coordinates are scraped from this Wikipedia page: https://en.wikipedia.org/wiki/List_of_Moscow_Metro_stations

	L	English transcription	Russian Cyrillic	Transfer	Opened	Elev.	Type	Coordinates	Pic.
0	NaN	Bulvar Rokossovskogo	Бульвар Рокоссовского	<>	1990-08-01	−8 m	column, triple-span	55°48′53″N 37°44′03″E / 55.8148°N 37.7342°E	NaN
1	NaN	Cherkizovskaya	Черкизовская	{}	1990-08-01	−9 m	single-vault, shallow	55°48′14″N 37°44′41″E / 55.8038°N 37.7448°E	NaN
2	NaN	Preobrazhenskaya Ploshchad	Преображенская площадь	NaN	1965-12-31	−8 m	column, triple-span	55°47′47″N 37°42′54″E / 55.7963°N 37.7151°E	NaN
3	NaN	Sokolniki	Сокольники	NaN	1935-05-15	−9 m	column, triple-span	55°47′20″N 37°40′49″E / 55.7888°N 37.6802°E	NaN
4	NaN	Krasnoselskaya	Красносельская	NaN	1935-05-15	−8 m	column, double-span	55°46′48″N 37°40′02″E / 55.7801°N 37.6673°E	NaN

Then using Foursquare API the base list of venue types is received:

```
4d4b7104d754a06370d81259 Arts & Entertainment
4d4b7105d754a06372d81259 College & University
4d4b7105d754a06373d81259 Event
4d4b7105d754a06374d81259 Food
4d4b7105d754a06376d81259 Nightlife Spot
4d4b7105d754a06377d81259 Outdoors & Recreation
4d4b7105d754a06375d81259 Professional & Other Places
4e67e38e036454776db1fb3a Residence
4d4b7105d754a06378d81259 Shop & Service
4d4b7105d754a06379d81259 Travel & Transport
```

We can see that `4d4b7104d754a06370d81259` represents Food and that is what we are going to be using to filter only food venues data from Foursquare.

We will use that category ID to first get all food places per each metro station:

```
radius = 500 # define radius
categoryId = '4d4b7105d754a06374d81259' # category ID for "Food"

url = 'https://api.foursquare.com/v2/venues/search?
&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&categoryId={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    moscow_metro_station.Latitude,
    moscow_metro_station.Longitude,
    radius,
    categoryId,
    LIMIT)
```

This Foursquare request will result in something like this:

```
{'id': '5baca878f427de002cfcbc17',
  'name': 'Мистер Круассан',
  'location': {'address': 'Большая Черёмушкинская улица, 1',
   'lat': 55.690278,
   'lng': 37.601879,
   'labeledLatLngs': [{'label': 'display',
     'lat': 55.690278,
     'lng': 37.601879}],
   'distance': 198,
   'postalCode': '117447',
   'cc': 'RU',
   'city': 'Москва',
   'state': 'Москва',
   'country': 'Россия',
   'formattedAddress': ['Большая Черёмушкинская улица, 1',
    '117447, Москва',
    'Россия']},
  'categories': [{'id': '4bf58dd8d48988d1e0931735',
```

```
    'name': 'Coffee Shop',
    'pluralName': 'Coffee Shops',
    'shortName': 'Coffee Shop',
    'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/coffeeshop_',
     'suffix': '.png'},
    'primary': True}],
  'referralId': 'v-1582192970',
  'hasPerk': False},... JSON of all food venues for each metro station;
```

We will later clean-up this data to only have a list of ethnic type restaurants with clear national/international cuisine.

We will need only food_venue_item['categories'][0]['name'] attributes values for our filter.

The goal is then to merge the two data sources and to get them into a pandas DataFrame which would show all metro stations with their latitude and longitude from the Wikipedia page and for each metro station show all associated food venues like that:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bulvar Rokossovskogo	55.8148	37.7342	Burger King	55.814026	37.733659	Fast Food Restaurant
1	Bulvar Rokossovskogo	55.8148	37.7342	Broker Coffee	55.813895	37.733526	Coffee Shop
2	Bulvar Rokossovskogo	55.8148	37.7342	шаурма	55.816166	37.730582	Shawarma Place
3	Bulvar Rokossovskogo	55.8148	37.7342	Китайская Кухня "Лотос"	55.813580	37.732730	Chinese Restaurant
4	Bulvar Rokossovskogo	55.8148	37.7342	Подсолнухи Art&Food	55.816065	37.736457	Food Court

That DataFrame would then, as already mentioned, be filtered down to only ethnic type restaurants with clear national/international cuisine and will exclude things like Fast food places, Coffee Shops, Food courts etc. like so:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bulvar Rokossovskogo	55.8148	37.7342	Китайская Кухня "Лотос"	55.813580	37.732730	Chinese Restaurant
Bulvar Rokossovskogo	55.8148	37.7342	Фо & Ролл	55.815955	37.736421	Vietnamese Restaurant
Bulvar Rokossovskogo	55.8148	37.7342	Суши Wok	55.814660	37.731430	Asian Restaurant
Bulvar Rokossovskogo	55.8148	37.7342	El Taco	55.815503	37.737244	Tex-Mex Restaurant
Bulvar Rokossovskogo	55.8148	37.7342	СушиStore	55.814618	37.730911	Sushi Restaurant
Bulvar Rokossovskogo	55.8148	37.7342	академия плова	55.816015	37.736523	Asian Restaurant
Bulvar Rokossovskogo	55.8148	37.7342	Churros House	55.816125	37.736488	Mexican Restaurant
Bulvar Rokossovskogo	55.8148	37.7342	Ksusha Kitchen	55.816019	37.736674	Italian Restaurant

This resulting DataFrame would allow us to achieve the goad which we set in the introduction section, which is to quantify and classify ethnic cuisine diversity available within 500 meters reach of each Moscow metro station.

Methodology

Discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

After combining the two data sources: the Wikipedia page table with Moscow metro stations and their Lat/Lon coordinates and the food venues query to Foursquare API to get venues for each metro station we had to do a bit of data cleaning.

That involved combing the data sources in to single pandas DataFrame which is already shown above in the Data section as final table. This DataFrame contains food venues that are only restaurants and only those that have clear national/international specialisation such as Italian Restaurant, Chinese Restaurant or Mexican Restaurant etc.

This DataFrame would allow as to calculate how many unique international cuisine places exist within 500 meters reach of each metro station firstly identifying different levels of food diversity in each neighbourhood and secondly identifying if there is any specific food domination or lack of in that neighbourhood.

Initially we will use just basic statistical summary analysis with max , min counts of venues per each station neighbourhood. We will display this resulting analysis apart from simple summary table by overlaying the results on the map using Folium library.

Each circle will be mapped to metro station coordinates and then we will use 'colour' library to create a gradient list of colours from blue to red. These gradient colours will correspond to smaller number of unique venues – more blue to high number of unique venues – more red, with green and orange colours in the middle corresponding to various degree of medium numbers of venues.

We will then use Kmean cluster analysis from 'sklearn.cluster' library with onehot encoding transformation to try to identify if there are any meaningful and explainable clusters among venues in relation to different national/international cuisine mixes.

Each cluster like in the previous min/max analysis will be mapped using Folium maps to metro station coordinates and then we will use another 'colour' library from 'matplotlib.colors' to just create a random unique colour for each cluster so we can clearly see them and visually identify.

Results

Discuss the results.

Firstly lets have a look at a summary statistics show which restaurant food is the the most popular and which is the list popular. Please note that we are only looking at specialist nationality cuisine restaurants and in the data preparation stage we removed all places like Fast food, Burger places, Food courts, Coffee shops etc. which have no clear national identity. Full list of restaurants can be seen in the submitted code.

The following two tables demonstrate top 10 restaurant venues in Moscow and the bottom 10

venues. As we can see Sushi Restaurant following by Caucasian and Italian venues are in the clear lead. Russian, Chinese and Middle Eastern among others are also in to 10.

Venue Category	
Sushi Restaurant	304
Caucasian Restaurant	284
Italian Restaurant	283
Asian Restaurant	179
Middle Eastern Restaurant	177
Eastern European Restaurant	173
Vietnamese Restaurant	138
Japanese Restaurant	135
Chinese Restaurant	81
Russian Restaurant	73

Venue Category	
New American Restaurant	2
Tex-Mex Restaurant	2
Swiss Restaurant	2
English Restaurant	2
Mongolian Restaurant	2
Hawaiian Restaurant	2
Tapas Restaurant	1
Tatar Restaurant	1
Bulgarian Restaurant	1
Belarusian Restaurant	1

The bottom part of the list includes restaurant venues like Bulgarian Hawaiian and among some other Swiss restaurants.

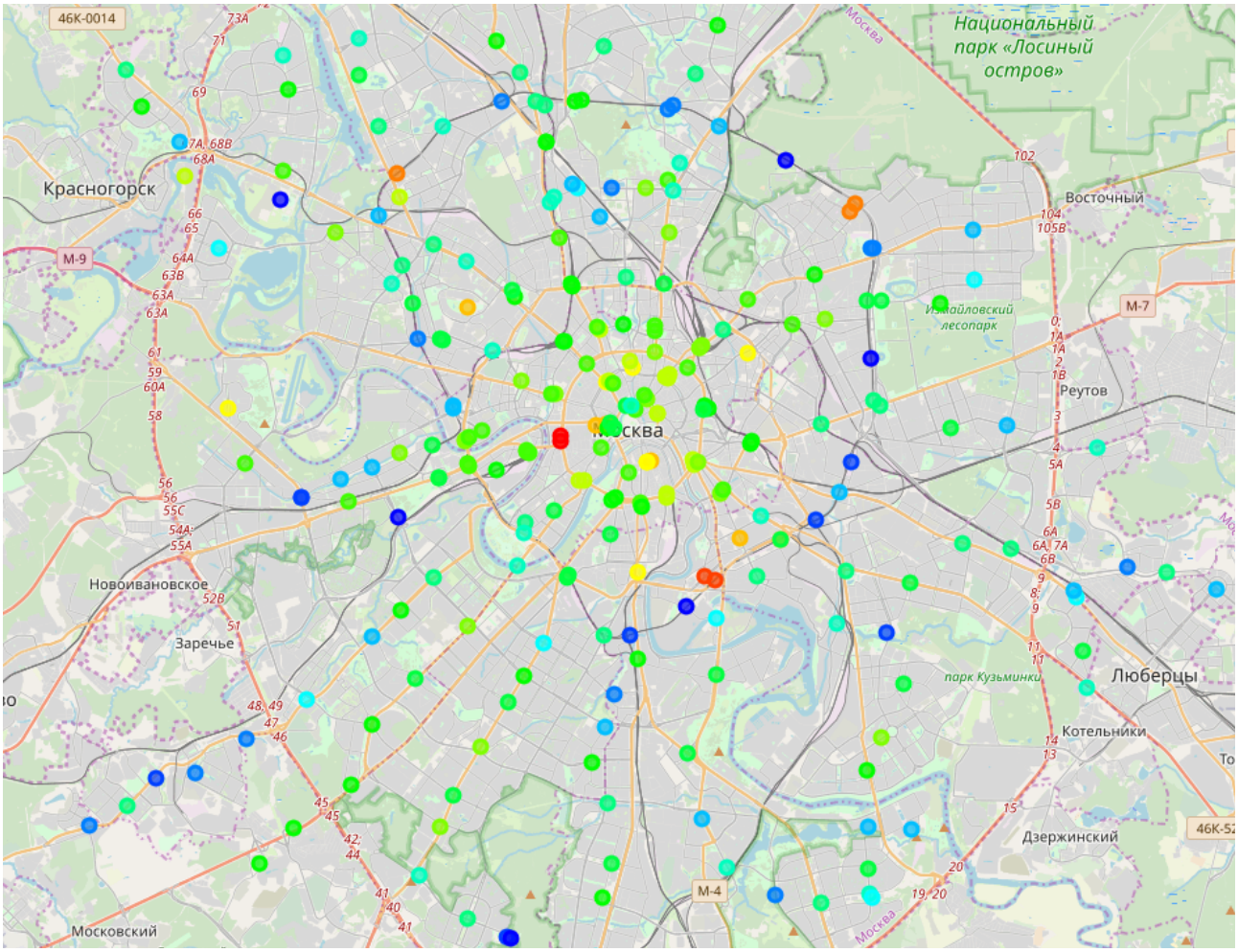
This simple analysis shows the preference and popularity of certain venues as well as it points out to some potential business opportunities.

The next two tables demonstrate maximum frequencies of occurrence of a venue per metro station. As we can see Caucasian, Italian Vietnamese and Russian restaurants among some others (picture on the left) are the most likely to occur as several offers at some metro stations where as again as before Bulgarian, Moroccan, Swiss among others (picture on the right) are very unlikely to be offered at most metro stations.

	count	mean	std	min	25%	50%	75%	max
Caucasian Restaurant	226.0	1.256637	1.556513	0.0	0.0	1.0	2.0	10.0
Italian Restaurant	226.0	1.252212	1.440099	0.0	0.0	1.0	2.0	9.0
Vietnamese Restaurant	226.0	0.610619	0.974644	0.0	0.0	0.0	1.0	8.0
Asian Restaurant	226.0	0.792035	1.073366	0.0	0.0	0.0	1.0	7.0
Middle Eastern Restaurant	226.0	0.783186	1.016259	0.0	0.0	1.0	1.0	7.0
Sushi Restaurant	226.0	1.345133	1.345081	0.0	0.0	1.0	2.0	7.0
Modern European Restaurant	226.0	0.247788	0.705924	0.0	0.0	0.0	0.0	6.0
Russian Restaurant	226.0	0.323009	0.764258	0.0	0.0	0.0	0.0	5.0
Eastern European Restaurant	226.0	0.765487	0.994587	0.0	0.0	0.0	1.0	5.0
Japanese Restaurant	226.0	0.597345	0.822890	0.0	0.0	0.0	1.0	4.0

	count	mean	std	min	25%	50%	75%	max
Kebab Restaurant	226.0	0.026549	0.161117	0.0	0.0	0.0	0.0	1.0
English Restaurant	226.0	0.008850	0.093863	0.0	0.0	0.0	0.0	1.0
Lebanese Restaurant	226.0	0.026549	0.161117	0.0	0.0	0.0	0.0	1.0
Latin American Restaurant	226.0	0.022124	0.147413	0.0	0.0	0.0	0.0	1.0
Dim Sum Restaurant	226.0	0.013274	0.114701	0.0	0.0	0.0	0.0	1.0
Spanish Restaurant	226.0	0.022124	0.147413	0.0	0.0	0.0	0.0	1.0
Swiss Restaurant	226.0	0.008850	0.093863	0.0	0.0	0.0	0.0	1.0
Moroccan Restaurant	226.0	0.017699	0.132148	0.0	0.0	0.0	0.0	1.0
Tapas Restaurant	226.0	0.004425	0.066519	0.0	0.0	0.0	0.0	1.0
Bulgarian Restaurant	226.0	0.004425	0.066519	0.0	0.0	0.0	0.0	1.0

The next image demonstrates counts of unique national/international offers showing ethnic diversity which was discussed in the introduction section. As discussed in Methodology section the image shows gradient colours from Red = maximum diversity to Blue = minimum diversity, with Green-Yellow colours representing different levels of medium diversity.



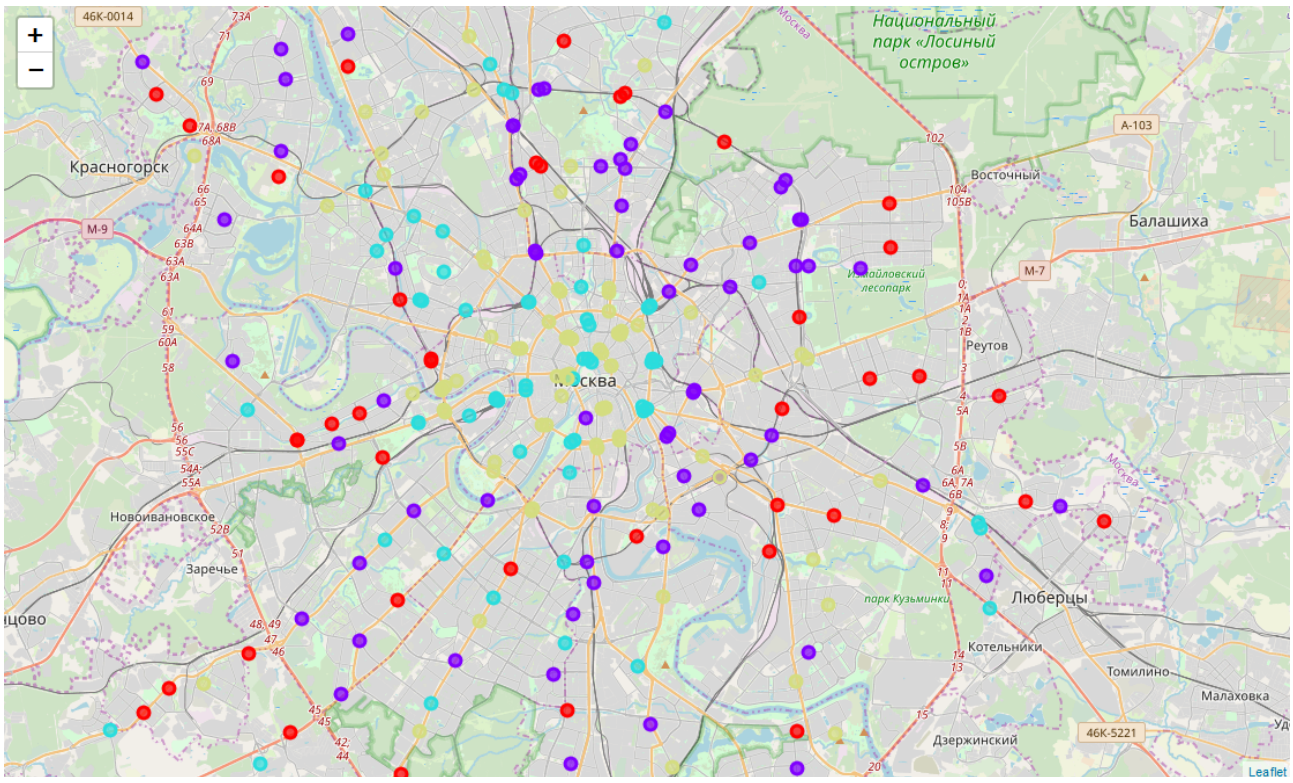
As we can see the centre of Moscow mostly represented by high to medium diversity offers where as on the peripheries we see more 'Blue' stations indicating low cultural cuisine diversity. This again can be used as a business opportunity for finding a new area for a specific ethnic/national cuisine.

For further more in-depth review interactive map can be downloaded here:

https://htmlpreview.github.io/?https://github.com/Serge-ds/Coursera_Capstone/blob/master/GradientMap1.html

Moving on to Kmeans cluster analysis, as mentioned in the Methodology section, we decided to run it using 'sklearn.cluster' library with onehot encoding, selecting k=4 clusters. Four clusters were selected using common sense, trial and error and empirical knowledge as the most suitable number of clusters that can have meaningful interpretation.

The results are displayed in the image below and we can immediately see that centre of Moscow is dominated by two clusters: Pale Yellow and Light Blue. The peripheries of Moscow are dominated by two other clusters Red and Purple.



It turns out that that Light Blue cluster offers the most diverse combination of restaurants with the biggest choice of all nationality groups; Pale Yellow on the other hand offers a slightly smaller selection still large enough but with a slight focus on Asian foods.

As for the other two clusters which dominate the peripheries of the city one of them – Red – offers small selection of venues mostly including Sushi restaurants and Japanese restaurants where as Purple cluster offers slightly larger selection that the Red but focus is more on Middle Eastern food.

For further more in-depth review interactive map can be downloaded here:

https://htmlpreview.github.io/?https://github.com/Serge-ds/Coursera_Capstone/blob/master/ClusterMap1.html

Discussion

Discuss any observations you noted and any recommendations you can make based on the results.

Both maps and the corresponding analysis clearly demonstrate that the centre of Moscow is well services with a great variety of international food available at a walking distance from most central metro stations. However, the peripheries are serviced not as well as the centre. This might be expected of course but it gives us an opportunity to see where to go if we want a great food diversity when we go out and where to look for business opportunities when looking to open a new restaurant venue.

Conclusion

Although Foursquare data is limited it can still provide some insights into food venues location and combining this data with other data sources we were able to conduct meaningful analysis which resulted in more or less clear classification of Moscow neighbourhoods in terms of cultural diversity and business opportunities when looking to open a new restaurant venue.