1. Introduction

* 1. Background

Starting any business is hard. Starting a business in a major city, such as Moscow, is even harder. Moscow is the biggest city in Russia, and it’s economical and cultural center. 12 million people live here permanently, and even more come here looking for job. Competition among small and medium business is extremely high, and any competitive advantage is necessary.

* 1. Problem

This project aims to cluster different districts of Moscow and give some insight on where competition is at its lowest and what kinds of venues are present and what are lacking.  
It is geared towards businessmen who want to open a food related venue, such as a restaurant, a café or a bar. Just providing good services, or serving great food, are often not good enough for survival in a highly competitive environment. That is why selecting a proper place for your future venue is almost equally important.

* 1. Interest

As stated above, this research will beneficial to people who are willing to start a food-related business in Moscow.

2. Data acquisition and cleaning

* 1. Data sources

As sources for our analysis, we will use Foursquare, via Foursquare REST API, and a CSV file which contains all of Moscow’s districts as of 2016 (no changes were made to their names since then).

<https://gis-lab.info/qa/moscow-atd.html#.D0.A1.D0.BA.D0.B0.D1.87.D0.B0.D1.82.D1.8C_.D0.B4.D0.B0.D0.BD.D0.BD.D1.8B.D0.B5>

* 1. Data cleaning

Several measures were taking to insure data tidiness. Due to the nature of our research, and the fact that we’re using unsupervised clustering algorithm, K-Means, we don’t have to check our data for outliers. After assigning geospatial coordinates to each district and combining the two tables, the next step is to ensure that all of the coordinates are correct. To make sure that it is the case, we draw a map and add the coordinates as markers, which allow us to see if we retrieved all of the data correctly. After the markers are drawn, we can see that a few dots are not placed in Moscow due to the fact that some of the districts’’ names are shared between multiple cities. This data will be omitted later since the density of the markers is good enough. After removing those incorrect entries we retrieve data from Foursquare, which is then merged into the districts names. It did not need any cleaning, apart from removing rows with NaNs in them.

* 1. Feature selection

When looking at the very first table we can see that are not relevant to our research - '№', 'Административный округ', "Тип мунобразования", "Код ОКАТО", "Код ОКТМО". We omit this data, and only leave the name of the district.

After getting the data from Foursquare, we only need the following data – Venue, Venue Category.

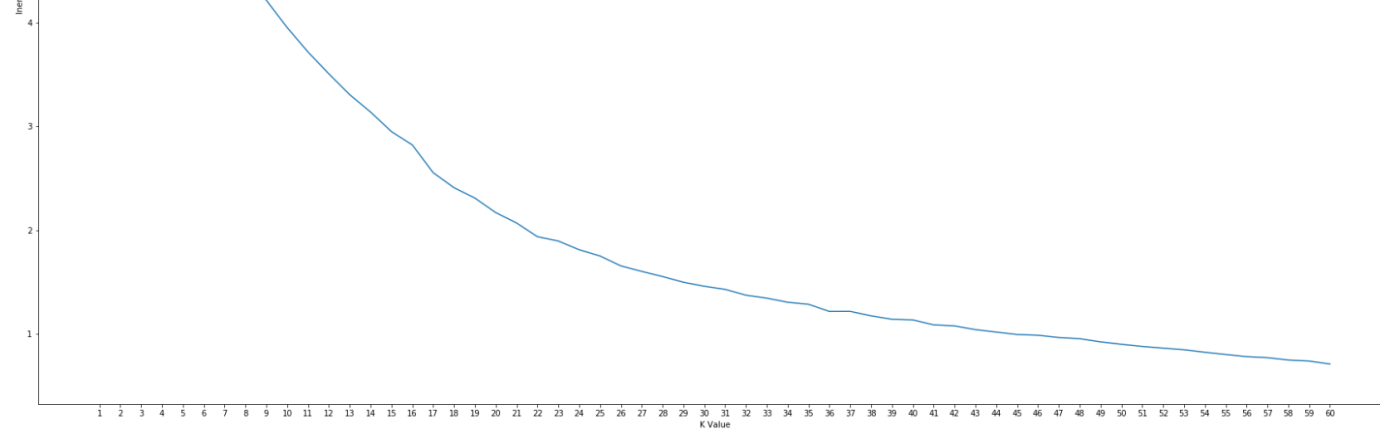
3. Exploratory data analysis

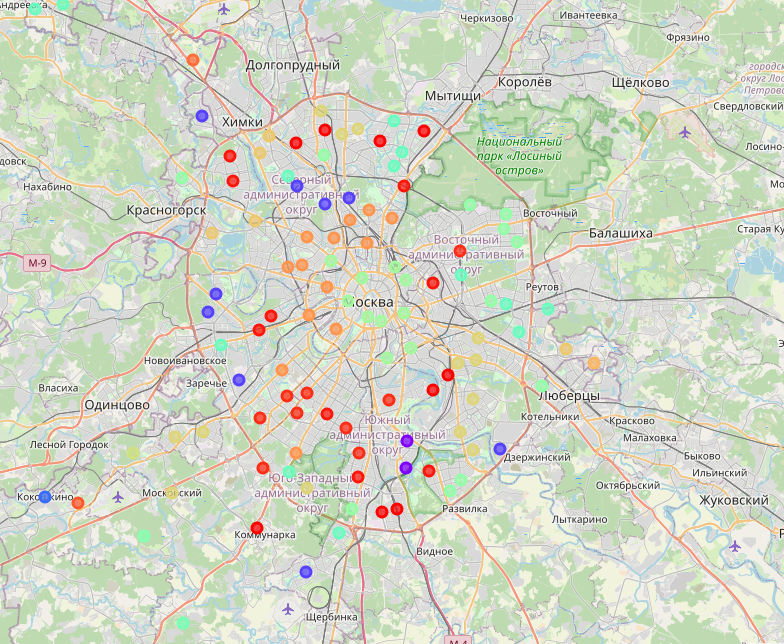
* 1. One hot encoding and top 10 venues

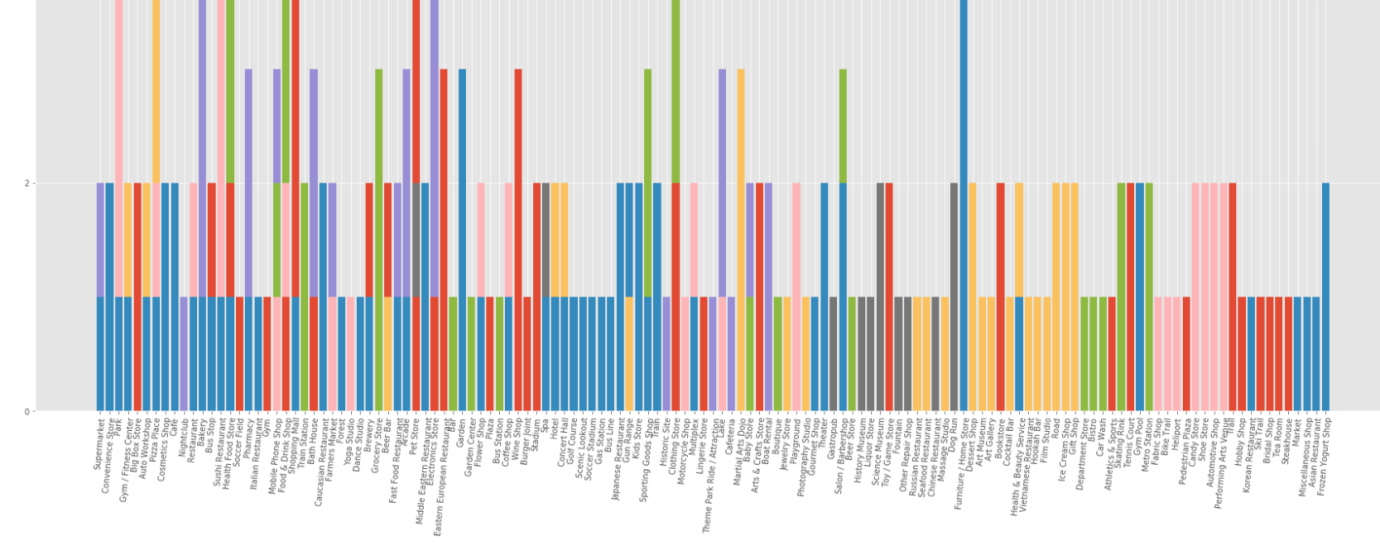
After importing our map and cleaning it, as well as selecting the required features, we perform one hot encoding on our data to get a “district - venue” table which can be then used for clustering. Performing a shape function shows that we have 8789 venues in our list, and 390 possible variants of those venues. After that we convert our 0/1 values into their means, and group the data by district. This allows us to have one row per district, and see how often each type of venue appears in that district. We then create a function which displays top 10 venues for each of the districts; this table will be used later.

* 1. Clustering

Then we start our clustering based on the table with means which we created earlier. To determine the best amount of clusters we create a function which plots a graph for each cluster values from 1 to 60, where X is the amount of clusters and Y is inertia, i.e. distance between data points and their cluster centers.

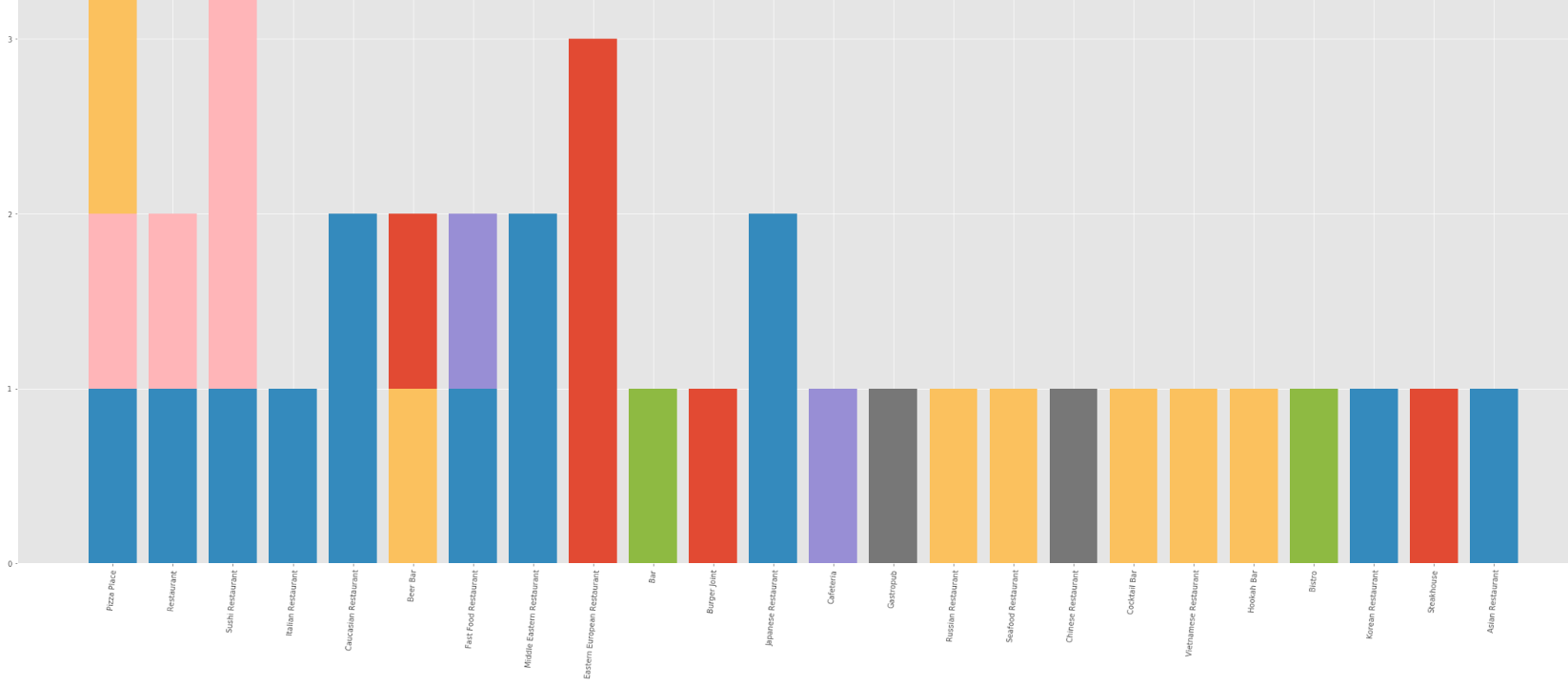


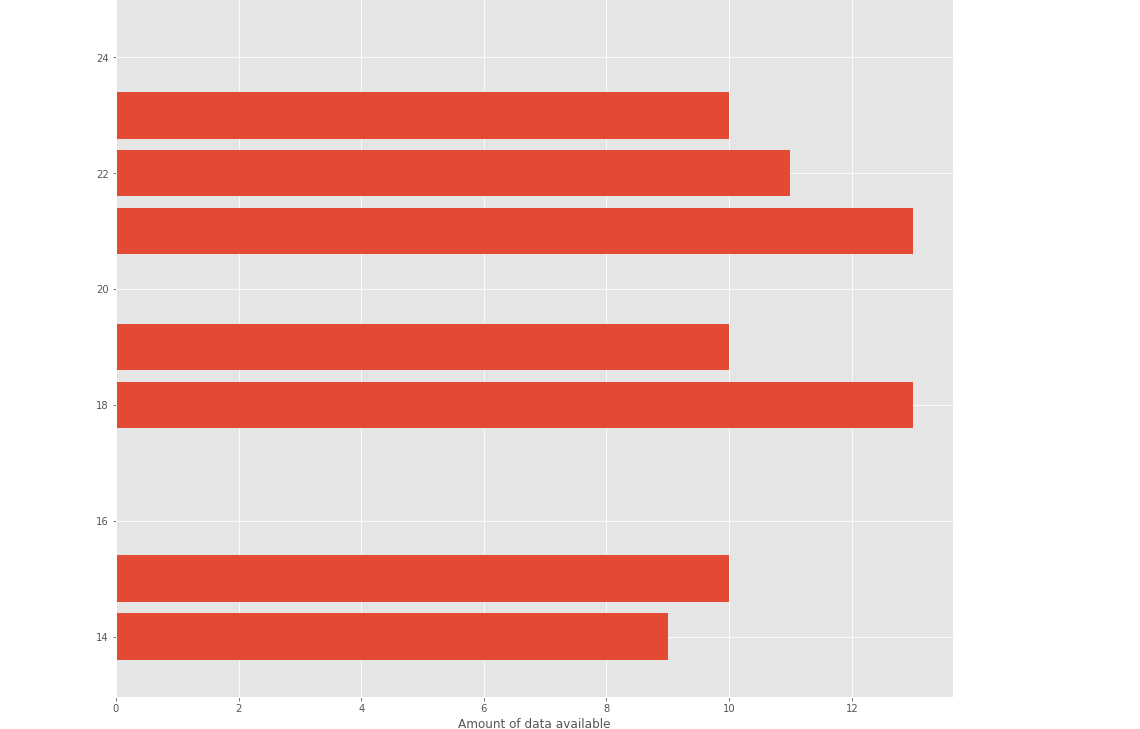
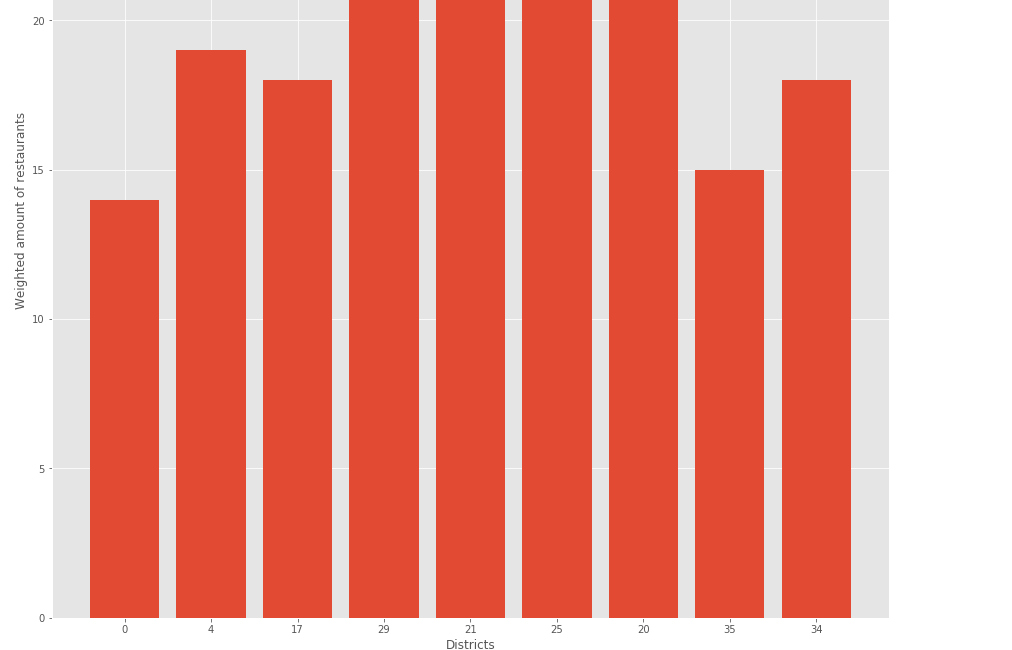
Using the elbow method we can see that the optimal value of K is between 22 and 40. We will pick 36 for more precision. It is worth noting that the model was tested with multiple values, and 36 seems to be the best K for this task. We then train our model, and assign cluster labels to each district in our top 10 table. We then create a map and color each of the markers with their own colors to visually represent each cluster.

As we can see on this map, there’re around 8 different clusters within the circle road which conventionally separates Moscow from Moscow Oblast. Moscow’s shape is not a perfect circle – it’s borders bleed out into several directions, but for this task we will just work with the clusters that are present inside the road circle. These are clusters number 0, 4, 17, 29, 21, 25, 20, 35 and 34. We use those clusters to build a stacked bar plot with all the venues inside the circle.

Due to the vast amount of data (we’re still talking about top 10 venues here!) and this plot’s irrelevance to our main goal it is cropped on the top, but should provide a general idea of venues distribution.

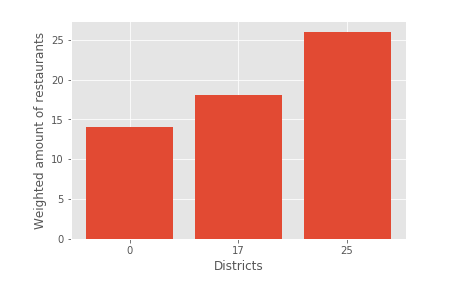
After analyzing what food-related venues are listed in the top 10 list we draw another bar plot which combines all of the food-related venues in those districts.



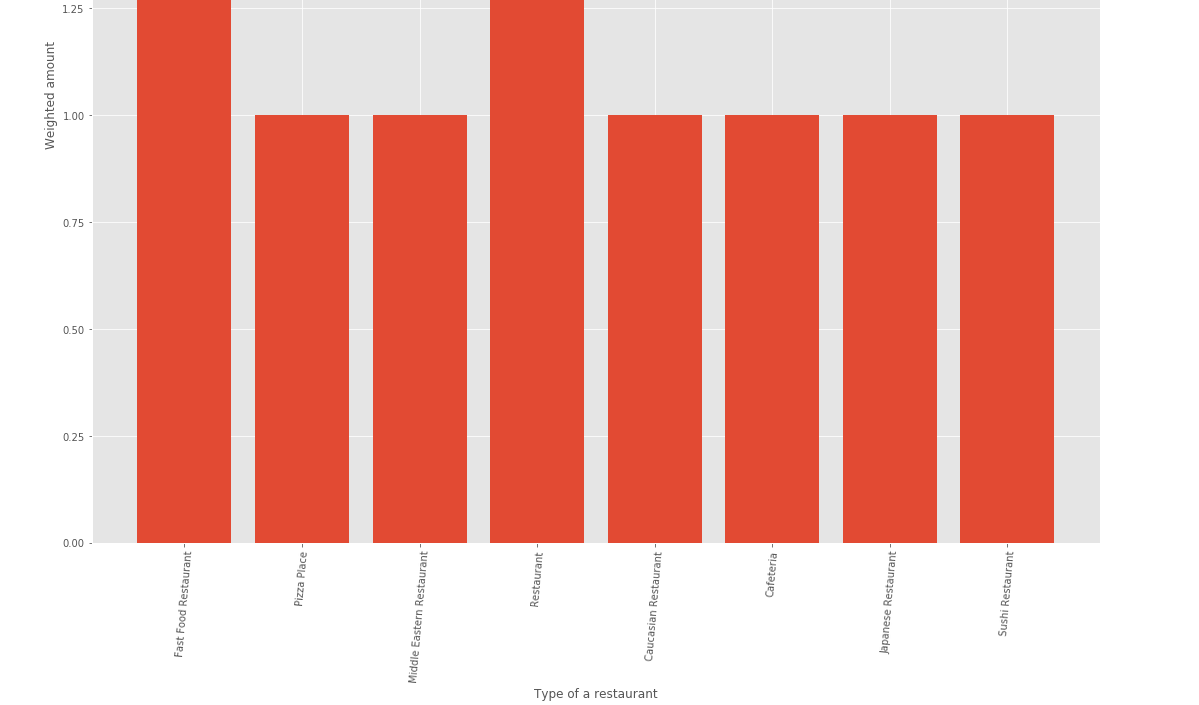
This is still not representative, but allows for a deeper look into the distribution of food venues. After that, we build two new plots – one of them shows weighted amount of food-related venues of each district, and the other one shows how much data we have for that district.

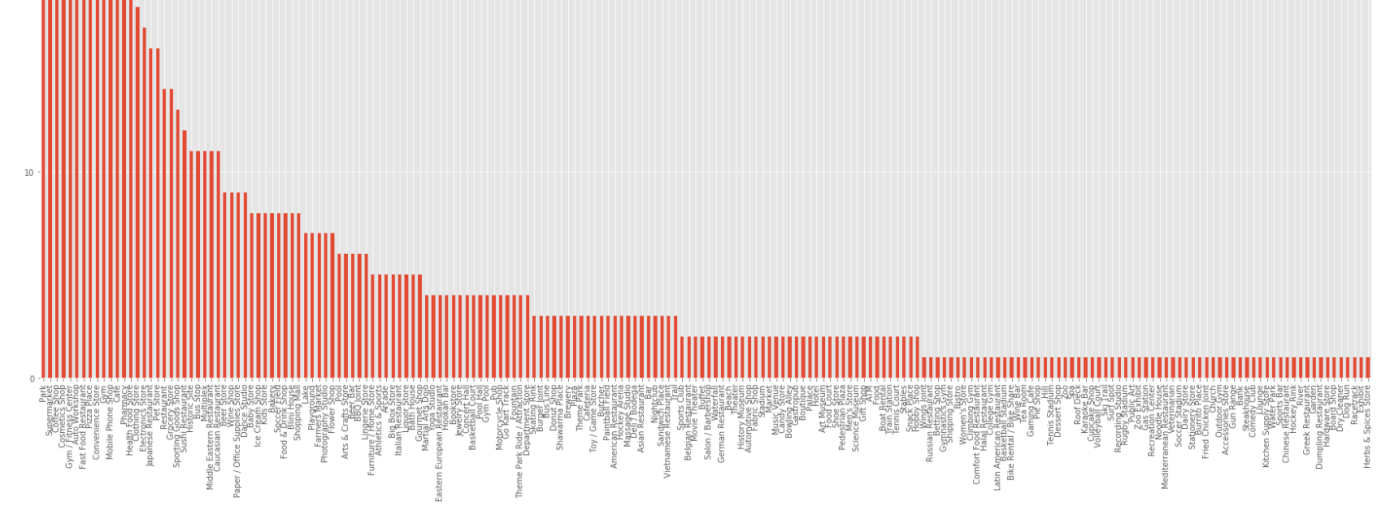
Here we can see what districts are of interest – those that have the least weighted amount of food venues AND sufficient data at the same time.

We select clusters 0, 17 and 25, and reexamine them, building a new plot which shows weighted amount of restaurants for that cluster.

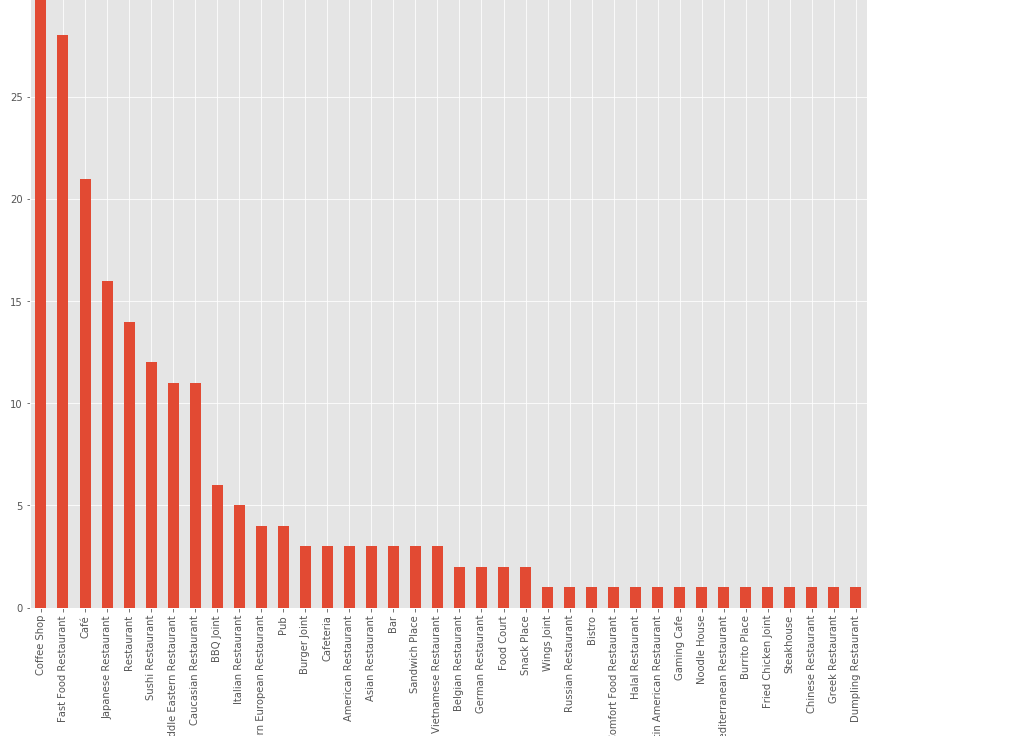


We also get the amount of data for them – 0 has 9 rows, 17 has 13 rows, and 25 has 13 rows. We will use cluster 17 for our further analysis since it combines sufficient data and the lowest amount of restaurants.

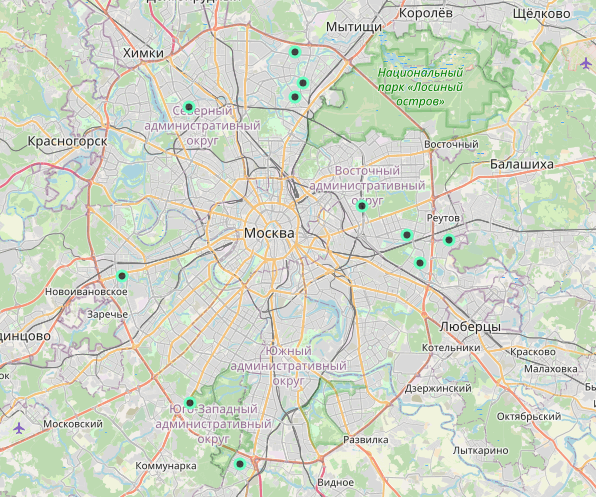
We build a new plot which shows top 10 food venues distribution for this cluster.  


Weighted amount is not good enough for our purposes, which is why we retrieve all of the venues and their real quantity for this specific cluster. 

This gives us insight on what is the contingent of people who live in that cluster, and what venues are present there. Now let’s dive into the food related numbers, and graph the most important plot in our analysis.



This is a graph of real numbers of restaurants and their categories. This graph allows us to examine what kinds of restaurants are present, where the competition is the highest, and what types of restaurants are lacking in the area. We also know that this cluster has the lowest amount of food-related venues compared to the other ones.

To help with understanding we also build a map with this specific cluster.

Here we can see the areas which are the most favorable for opening a new restaurant.

4. Results

After analyzing all of the data we can come to a conclusion that opening a restaurant, specifically a restaurant of a type that is not overwhelmingly present in those clusters, should be done in this specific cluster, which consists of several districts of Moscow. Upon further inspection we can determine the position even more precisely by looking at each point and using our common knowledge to determine which points are a good fit and which are not.

More precisely, all of the points to the north are situated in the so-called “sleeping” districts, which means that people don’t tend to spend time in them and only use them to sleep during the night, only to go to work towards the center of Moscow.

Eastern point which is situated near the center would be a good fit, since this is the place where people tend to spend their time. The same can be said about the western point.

Southern points are also considered “sleeping” districts, and opening a restaurant in that area might prove being hard.

5. Discussion

After examining the results, we can come to a conclusion that most of the cluster points are located in the “sleeping” districts. Not all of such districts are created equal, and while some of them can be nothing more but a place to sleep, others tend to create infrastructure and convenience for the people who live there, which is why opening a restaurant in such a district should be considered for each of them separately.

We also have one point which seems to be a perfect fit – the Sokolinaya Gora, which is not far away from the center and is located near a huge Izmailovsky park, which is an attraction point on it’s own.

Based on the data we have, the best recommendation would be the eastern part of the city, near Izmailovsky park.

6. Conclusion

In this research we have analyzed the city of Moscow using Foursquare data and districts data. After clustering the data, we separated all of the districts into categories based on their similarity, which allowed us to determine what clusters had what venues as the most popular ones, see the distribution of the venues, and determine the best clusters for opening a restaurant, eventually narrowing it down to just one cluster.