Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Gym / Fitness Center in Moscow, Russia

By: Sergei Dyachenko

February 2020

Introduction

Many people lead a healthy lifestyle and that's why it's important to keep fit for them. And I'm among them.

Most of the visitors are adult people and they prefer a comfortable place for trainings with good conditions where they can relax or exercise after the end of the working day or on holiday.

So the location of the fitness center is one of the most important decisions that will determine whether the gym will be a success or a failure.

I chose Moscow, the city where I live, so I could use my first-hand experience.

Business Problem

Of course, as with any business decision, opening a new fitness center requires serious consideration and is a lot more complicated than it seems.

The objective of this capstone project is to analyze and select the best locations in the city of Moscow, Russia to open a Fitness Center. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: in the city of Moscow, Russia, if a property developer is looking to open a new Fitness Center, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in the new fitness center in the capital city of Russia i.e. Moscow.

Data

To solve the problem, we will need the following data:

- List of neighborhoods in Moscow. This defines the scope of this project which is confined to the city of Moscow.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to fitness centers. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

Geojson file which contains a list of neighborhoods in Moscow (with a total of 146 districts) and their coordinates.

We will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers.

Foursquare API will provide many categories of the venue data; we are particularly interested in the Gym/Fitness Center category in order to help us to solve the business problem put forward.

This is a project that will make use of many data science skills: working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

Firstly, need to get the list of neighborhoods in the city of Moscow. Fortunately, the list is available in the geojson file which also contains geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API.

After reading the data, need to populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. It allows to perform a sanity check to make sure that the geographical coordinates data are correctly plotted in the city of Moscow.

Next, use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. But before it, need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. After it make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format where can extract the venue name, venue category, venue latitude and longitude. With the data, it's possible to check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, analyzing each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, the data are preparing for using in clustering. Then data need to be filtered for analyzing the "Gym/Fitness Center" venue category for the neighborhoods.

Lastly, perform clustering on the data with k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. In this project data were clustered into 3 clusters based on their frequency of occurrence for "Gym/Fitness Center". The results allows to identify which neighborhoods have higher concentration of fitness centers while which neighborhoods have fewer number of fitness centers.

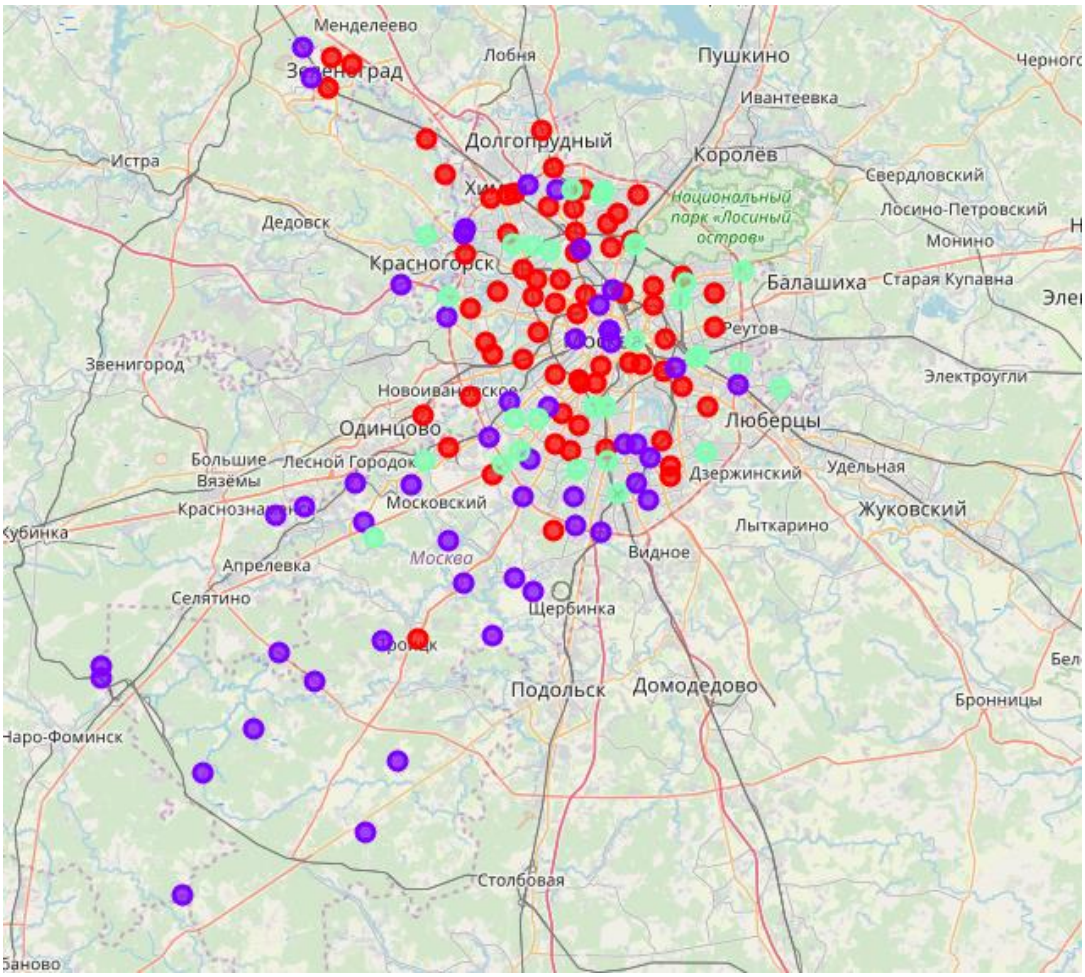
Based on the occurrence of fitness centers in different neighborhoods, it helps to answer the question as to which neighborhoods are most suitable to open new fitness centers.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Gym/Fitness Center”:

- Cluster 0: Neighborhoods with moderate number of fitness centers;
- Cluster 1: Neighborhoods with low number or no fitness centers;
- Cluster 2: Neighborhoods with high concentration of fitness centers.

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in green color.



Discussion

The highest number of gyms in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to totally no fitness centers in the neighborhoods. It represents a great opportunity and high potential areas to open new fitness center as there is very little to no competition from existing fitness centers. Meanwhile, fitness centers in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of fitness centers. This also shows that the oversupply of fitness centers mostly happened on the south of the city with the central area and west of the city still have very few fitness centers. Therefore, this project recommends property developers to capitalize on these findings to open new fitness centers in neighborhoods in cluster 1 with little to no competition. Property developers with good offer on season/year subscription to stand out from the competition can also open new fitness centers in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of fitness centers and suffering from intense competition.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers the best locations to open a new fitness center. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new fitness center. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas (cluster 2) in their decisions to open a new fitness center.

References

Administrative divisions of Moscow Geodata. Retrieved from https://gis-lab.info/qa/moscow-atd.html

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs