Capstone Project - The best place for a beer shop

Applied Data Science Capstone by IBM/Coursera

Roman Makarov

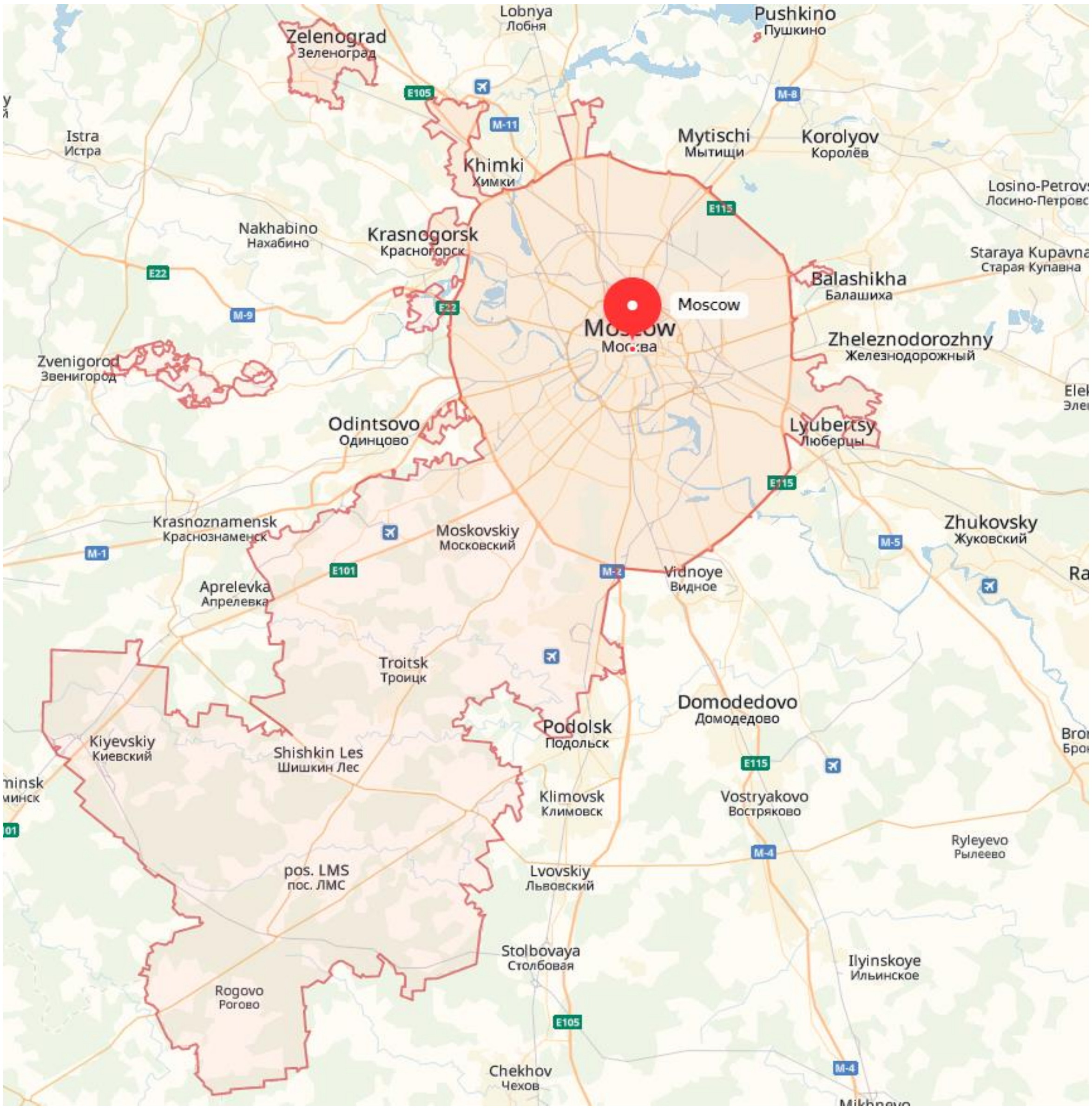
May. 2019

Table of contents

Introduction

Last weekend I met my old friend from Moscow (capital of Russian Federation). That evening was amazing we haven't met for three years. We talked some hours about our jobs, personal file and ideas about our future plans. Eventually he told me that he wants to start his own business. His choose is some strange for me because his is an engineer and wants to open a beer shop.

Moscow is a big city. It is approximate 13000000 people. There are many business opportunities in big cities.



Moscow map

Moscow has been great extended over the last few years (some Moscow regions became the part of the city).

My friend has some amount of money that has saved over the past few years. This amount is enough to open a beer store, but not enough to advertise it, so he decided to choose a franchise shop. There are a lot of "craft beer" franchises in Moscow city. It depends on kind of beers, initial financial investments, monthly income and so on.

I decided to help him make a choice. It's a good practice for me and helps to complete my Data Science specialization. So I decided to help my old friend.

He has already selected a franchise shops:

1. Pivoteka 465.
   http://pivoteka465.ru/franshiza_pivoteka_465.html

   ПИВОТЕКА
   465

2. Kalinkino.
   http://kalinkino.com/franchise.html

   КАЛИНКИНО
   ИНДИ-ПИВОВАРНЯ

3. Pinta.
   http://pinta-company.ru/franshiza.

   ХМЕЛЬНАЯ
   ПИНТА

4. Piv&ko
   https://franchise-pivko.ru

   ПИВ&Ко®

5. BeerMag.
   https://www.russtartup.ru/franchising/vsefranshiza/franchajzing-pivnogo-magazina-birmag.html

All these trademarks are really good. They are work very long time in Moscow but they are so different.

A good reason is to look up other trademarks.

Customer requirements

My customer (my old friend) has these requirements:

1. My shop has to be located in the appropriate district.
2. My shop has to be located near my house.
3. I want to have back my investments as soon as possible.
4. What’s the best price/quality trademark

Business problems

All the trademarks have different customers. That mainly depends on a menu, prices etc.

My customer doesn’t have so much money so he can’t choose really expensive franchise.

All the districts in Moscow city have different residents (according to monthly earnings). So a really expensive shop is not a good choice for all the districts.

My customer wants to have a shop near his house. He wants to get to the shop by car every day.

There are a lot of beer shops in the city. So there is no point in opening the shops near the same trademark shops. The place for the shop would be crowded.

The main business problems:

1. What is the best trademark (according to my friend’s investments etc.)?
2. What is the best place of the city to open a beer shop according to standard of living?
3. What is the most visited place for the shop (place with other entertainment)?
4. What is the best place for the shop depending on city’s traffic jam?

Data requirements and selection

To solve the main business problems a data scientist needs additional information. A good practice is to use official public databases.

1. The main database for these business problems is Foursquare. It's convenient for exploring the city neighborhoods. We may collect all the beer bars, shops according to them locations. Also we may collect data about another public places (for example: public parks etc.), but there are some problems to use this database. This database does not have enough information about Moscow city, so I decided to extend Foursquare database with additional queries to Yandex GEO API https://tech.yandex.com/maps/.
   This API has a free version. In the free version of the API, the total number of requests must not exceed 25,000 requests per day so all the data will be collected to local files in some days. Also the collected data may be corrected according to Open Street Map https://www.openstreetmap.org/ (if necessary).

   In this part I'll extract all common features of the public palaces and save them to "JSON" format (it has approximately similar format for these geo services).
   Possible features are:
   1. Total count of bear shops per district (or part of district)
   2. Total count of public places (parks etc.)
   3. Total count of elementary schools (may be helpful because of prohibition law to sell alcohol near the schools) (if necessary)

2. The Administrative division of Moscow is available on Wikipedia web site.
   https://en.wikipedia.org/wiki/Administrative_divisions_of_Moscow

   The final table:

District Name	Latitude	longitude

3. To indicate the standard of living may be helpful to collect data about total value of real estate in the city by districts. All the data of this year represented by Moscow Marker Indicator.
   https://www.irn.ru/rating/moscow/

   The final table:

District Name	Total value per square meter

   From this site is useful to collect average rental price:

District Name	Average Rental Price

4. For additional standards of living exploring the main database is official Federal State of statistics. Moscow city represents in amount of several excel and word files. http://moscow.gks.ru/wps/wcm/connect/rosstat_ts/moscow/ru/statistics/standards_of_life/

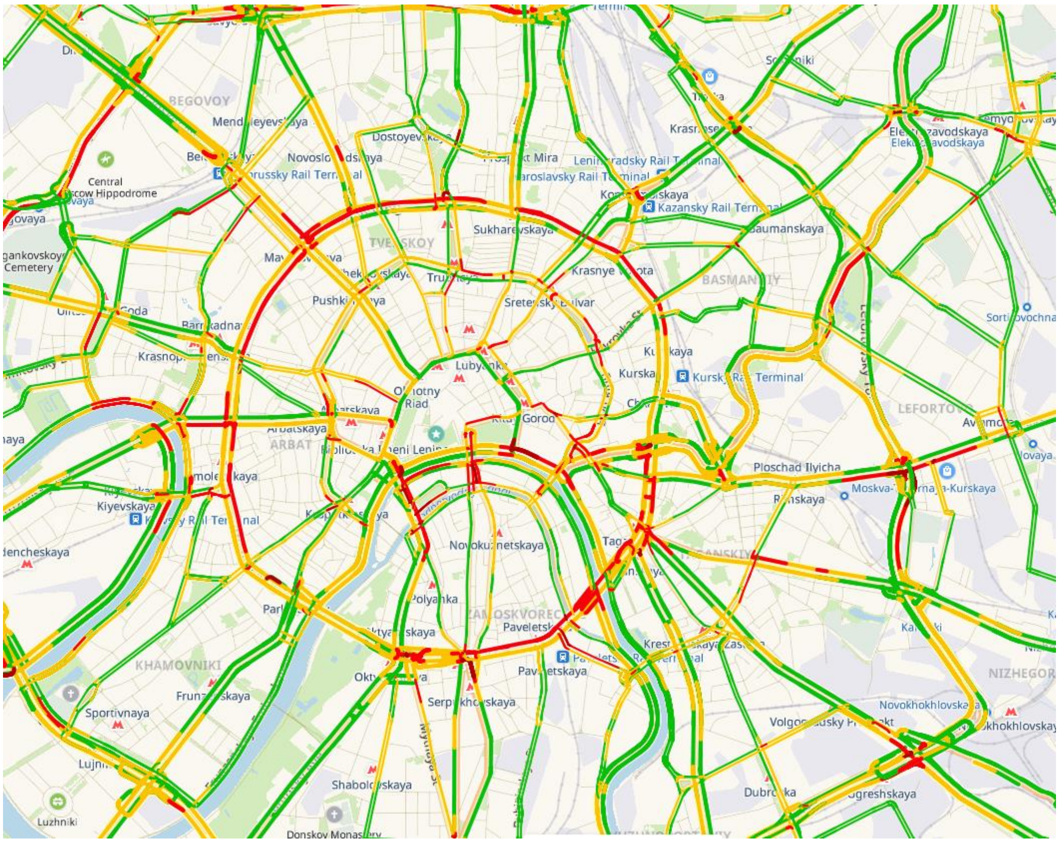
   Possible features in this section:
   1. Total approximate amount of month salary (grouped by district). It may be collected only approximately by the whole site statistics.
   2. Structure of residents
      a. Sex (male/female) (if necessary)
      b. Average age of residents (if necessary)
      c. Total resident count per district.

5. For the calculation of distance between some city points in Moscow the best choice is Yandex maps. https://yandex.com/maps/213/moscow/?l=trf%2Ctrfe&ll=37.633950%2C55.754925&z=13. It also may be correct by additional database (if necessary). For example https://maps.openrouteservice.org/

   The final table:

Probably place Name	Total distance from home (km)

6. To explore the Moscow traffic jam the best database is Yandex. https://yandex.com/support/maps/concept/stoppers.html
   Yandex stores all the traffic jam statistics by all days and hours. This service is not free but I can collect this data manually (not by API) if necessary.



Average traffic jam map on Wednesdays at 18:05

Is this section may be useful to collect data by time. But I have to connect my customer and ask him.

7. To explore additional beer shop franchises (if necessary) may be used:

    1. Russian Business Consulting web site
       http://biztorg.ru/franchises
       (All the data in Russian language only)
    2. Top Franchises Russia aggregator
       https://topfranchise.ru/catalog/
       (All the data in Russian language only).

The final table in this section:

Trademark Name	Capital investment	Estimated Monthly Profit	Approximate payback period

Feature engendering

Most of the data is well structured but some of that should be collected by a grabber or manually.

In this section the best practice is to apply feature engendering.

For example:

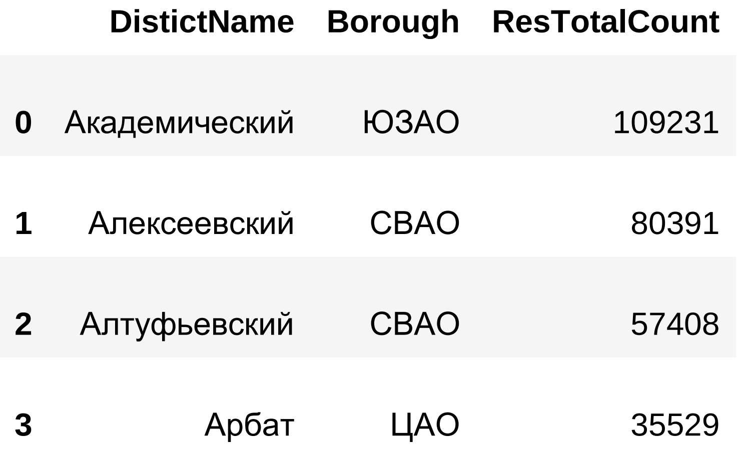
1. Distance rate (speed) = Total distance from home / Total time to distance (according traffic jam).
2. “Crowded” rate = Total count of resident (in area) / total count of public places near the point.
3. “FillnessRate” = Total count of beer shops in area / max beer shop count per districts

This section might be corrected a few times when the main algorithm is adjusted.

Data collecting and clearing

We apply these steps:

1. Parse Wiki page and extract Moscow districts. It’s a base table of Moscow districts



	DistictName	Borough	ResTotalCount
0	Академический	ЮЗАО	109231
1	Алексеевский	СВАО	80391
2	Алтуфьевский	СВАО	57408
3	Арбат	ЦАО	35529

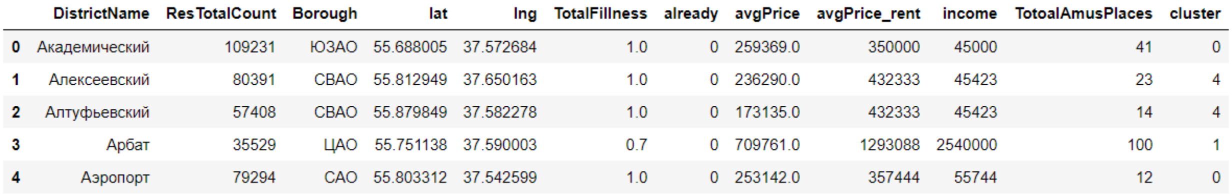
2. Get Districts GEO points (Russian map service Yandex)
3. Load and clear real estate price table (by square meter)
4. Get real estate price (by square meter)
5. Get average rent price by borough
6. Get average monthly income
7. Get nearby places by Foursquare
8. Get the nearest beer shops (by Yandex)
9. Get total places (exclude beer shops).
10. Get total trademarks by places (customer decided to choose only these trademarks)
11. Get districts where trademarks is already present

12. Set NEW Feature (count of beer shops / max beer shop per district)
13. Merge all the data in one DataFrame

Result of data collection

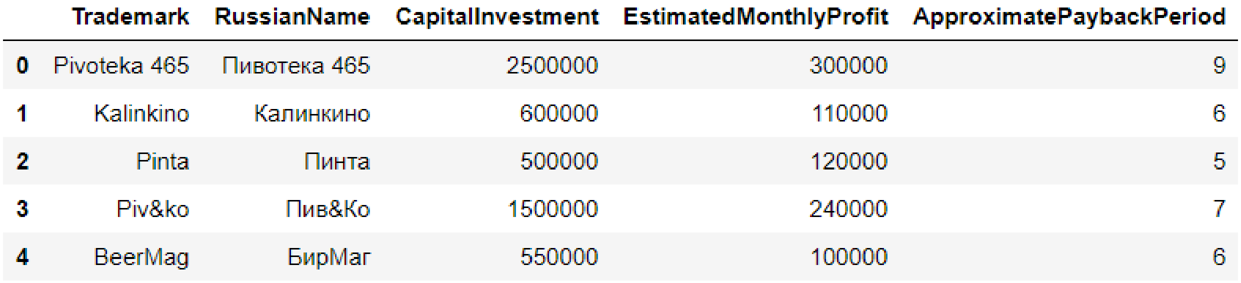
The final result of this section:

1. All the data collected about the city districts



	DistrictName	ResTotalCount	Borough	lat	lng	TotalFillness	already	avgPrice	avgPrice_rent	income	TotoalAmusPlaces	cluster
0	Академический	109231	ЮЗАО	55.688005	37.572684	1.0	0	259369.0	350000	45000	41	0
1	Алексеевский	80391	СВАО	55.812949	37.650163	1.0	0	236290.0	432333	45423	23	4
2	Алтуфьевский	57408	СВАО	55.879849	37.582278	1.0	0	173135.0	432333	45423	14	4
3	Арбат	35529	ЦАО	55.751138	37.590003	0.7	0	709761.0	1293088	2540000	100	1
4	Аэропорт	79294	САО	55.803312	37.542599	1.0	0	253142.0	357444	55744	12	0

2. All the data collected about the possible franchises is this district



	Trademark	RussianName	CapitalInvestment	EstimatedMonthlyProfit	ApproximatePaybackPeriod
0	Pivoteka 465	Пивотека 465	2500000	300000	9
1	Kalinkino	Калинкино	600000	110000	6
2	Pinta	Пинта	500000	120000	5
3	Piv&ko	Пив&Ко	1500000	240000	7
4	BeerMag	БирМаг	550000	100000	6

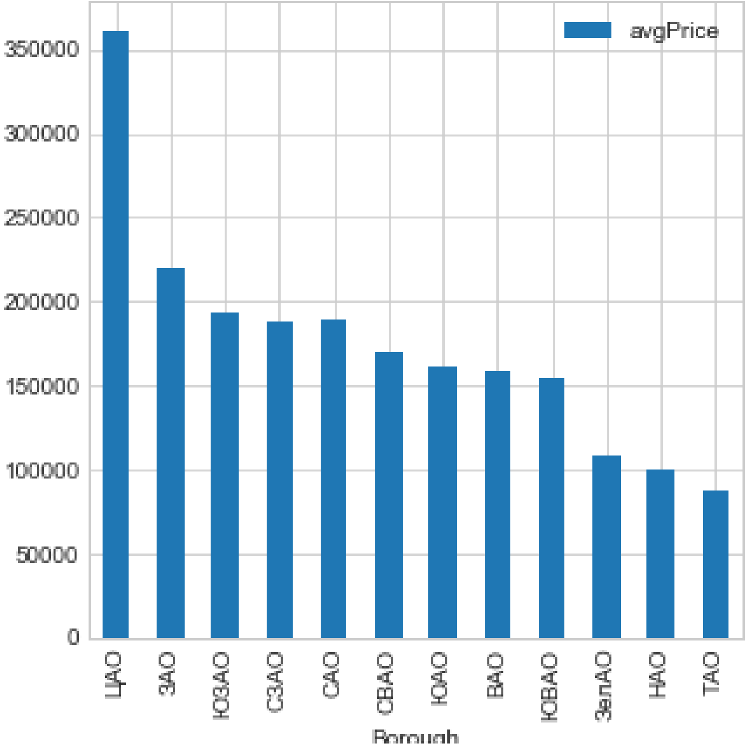
All the data has to be converted in uniform format (for example .csv) and saved to disk to avoid additional access to limit GEO services and etc.

Methodology

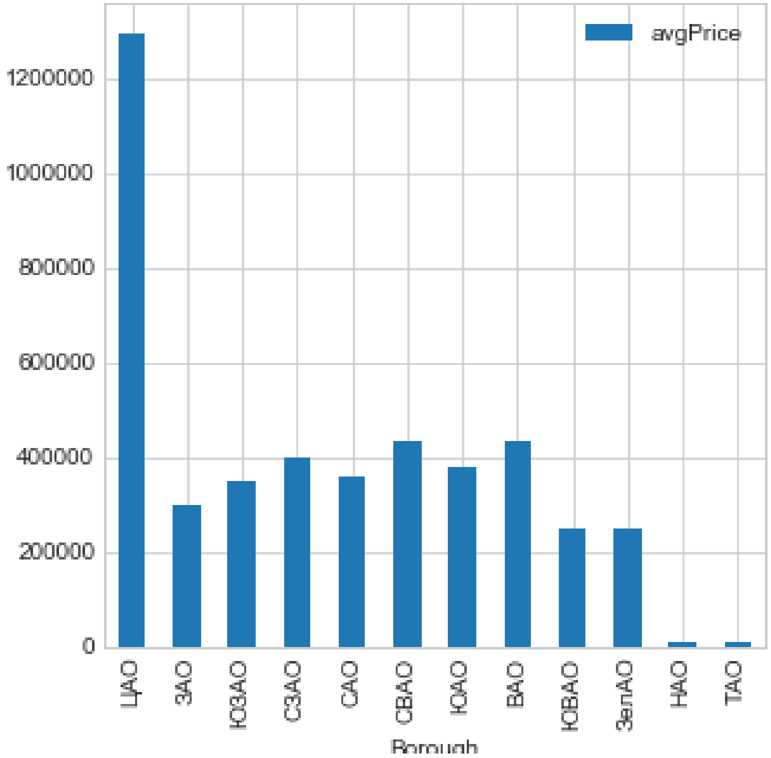
In this section we already have corrected and cleared data.

We have these features:

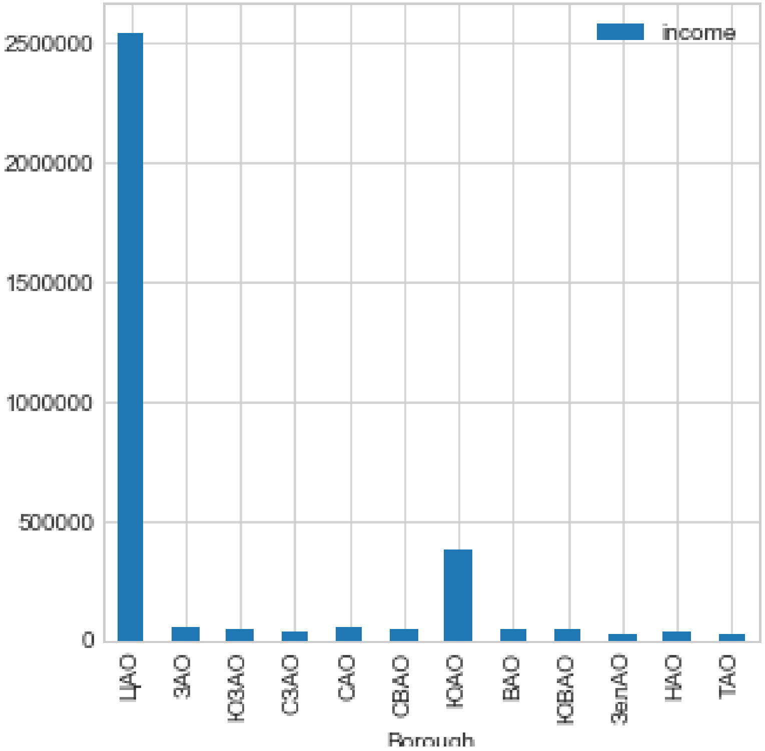
1. Total count of residents in district
2. "Total fillness" - how many beer shops in district
3. 1/0 if district has the same type of beer shop as customer selected
4. Average real estate price (by square meter)



5. Average rent price



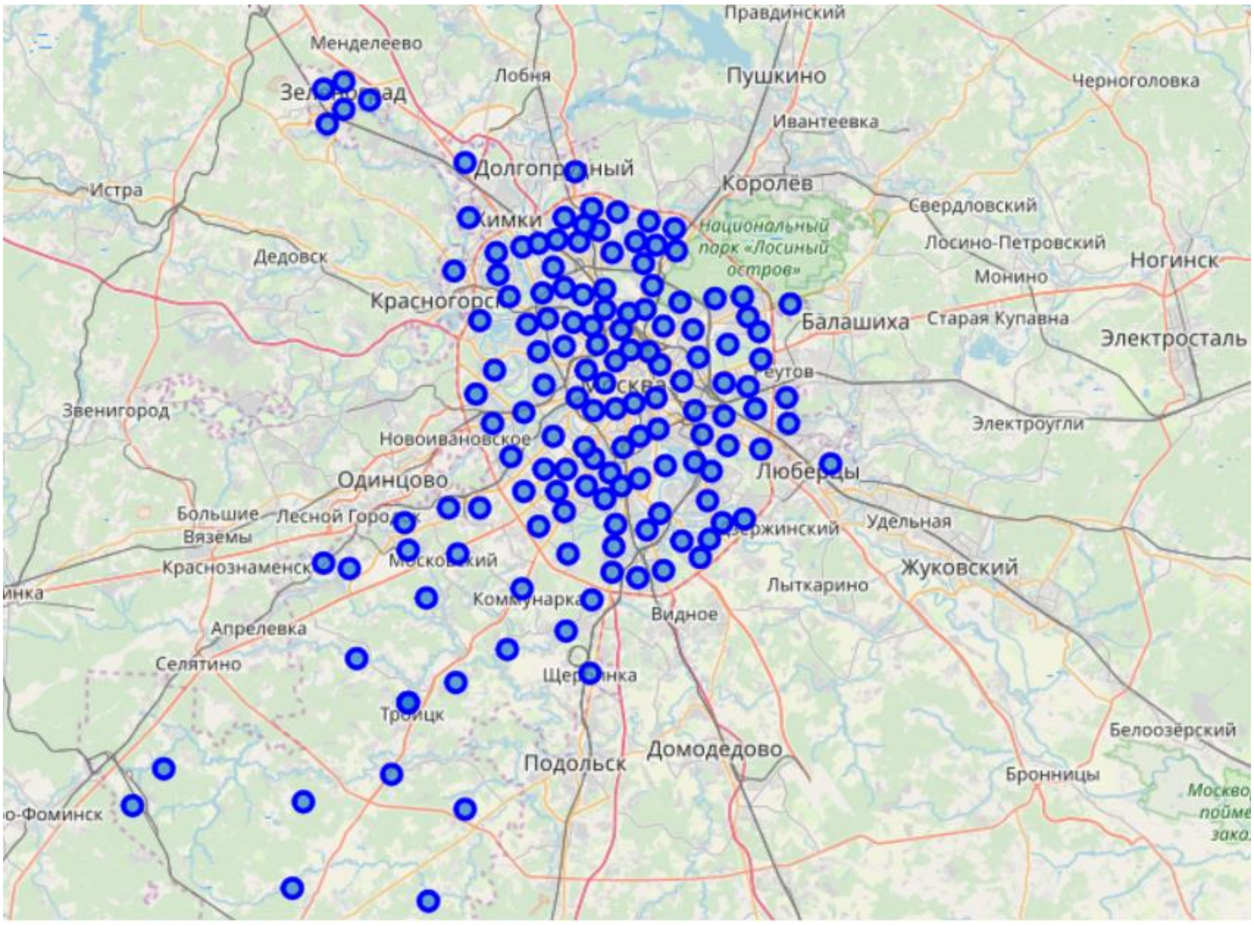
6. Average residents monthly income



7. Total count of interesting places in district

These features are really different in the districts.

Now we can plot districts map.

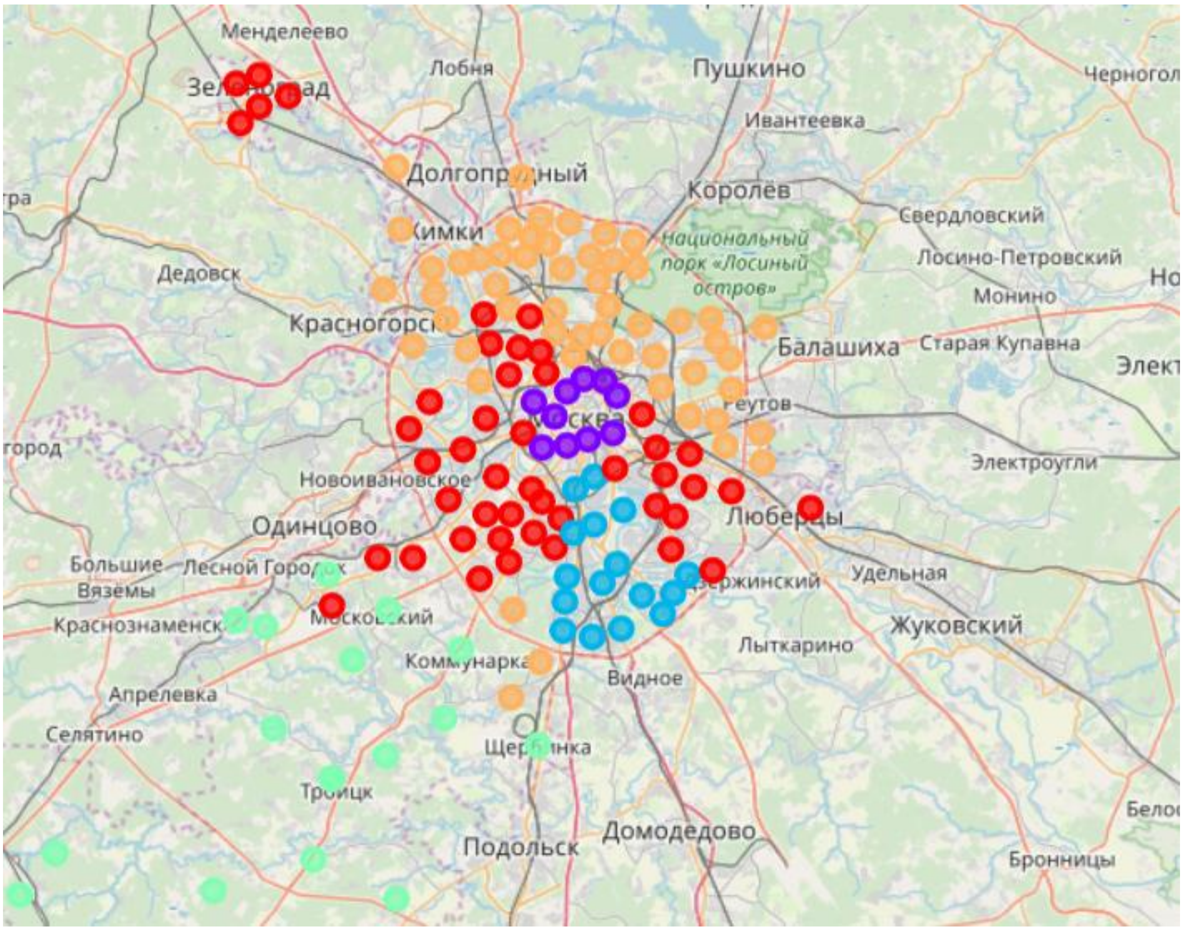


Centroids map.

By this information we can separate the districts by special groups. We will use ***K-means algorithm***.

It's suitable for this task. After splitting our districts by groups we'll explore them. This exploration will help us to detect economic situation in each district. By summary analysis we'll make decision what the best district is.

After detecting the clusters, we plot the cluster map.



After several iterations we detected that the best count of clusters is five.

Analysis

Each cluster represents only one segment of the city districts groped by several parameters. The Analysis gets us some decisions.

1. *Cluster №0 has medium parameters (income, rent price, total count of beer bars/shops etc.)*

This cluster is the most suitable for opening small business.

2. *Cluster №1 has high parameters.*

This cluster is not good for small business. Rich people don't go to shops. They order goods on the internet or some services intended for the rich.

3. *Cluster №2 is good for opening small business but it has worse parameters then cluster №0.*

So we don't consider it.

4. *Cluster №3 represents the poorest people of the city.*

This part of residents doesn't like craft beer. They drink beer in tin cans only. So we drop this cluster too.

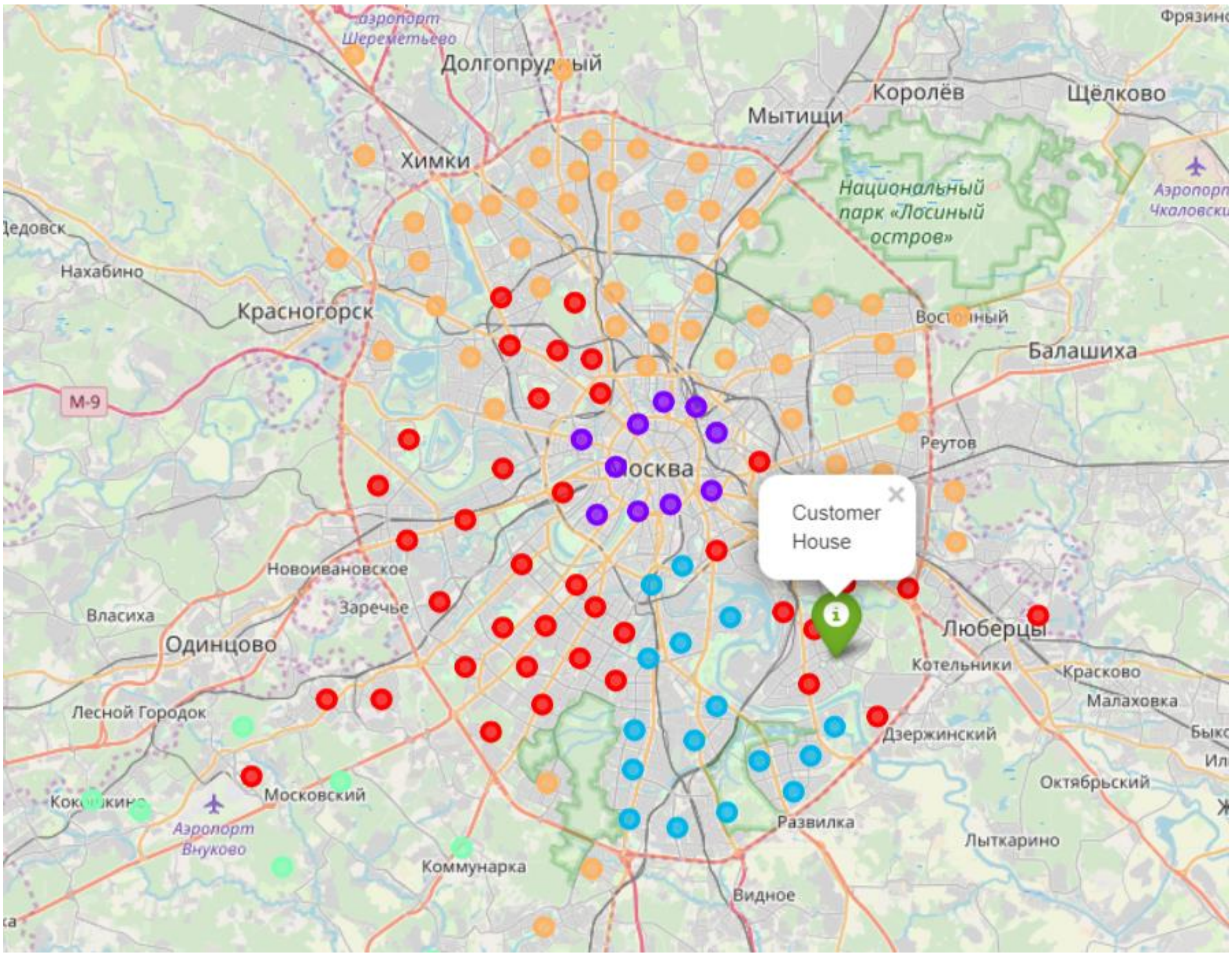
5. *Cluster №4 is the same as Cluster №0 but rent price is higher.*

This cluster is good for small business too but there is high competition in it.

We have two clusters (0,4) for opening a small beer shop. One of the customer's requirements is "Closer to my home". These distances may be calculated by Yandex of Google services, but these services are not free now. Indeed we shouldn't use these services.

Conclusion

By skimming the clustered map, we can see that the customer's house is in cluster №0, so **the best choice is cluster №0**. (We **don't need to use** traffic city web service. That is also **not free**)

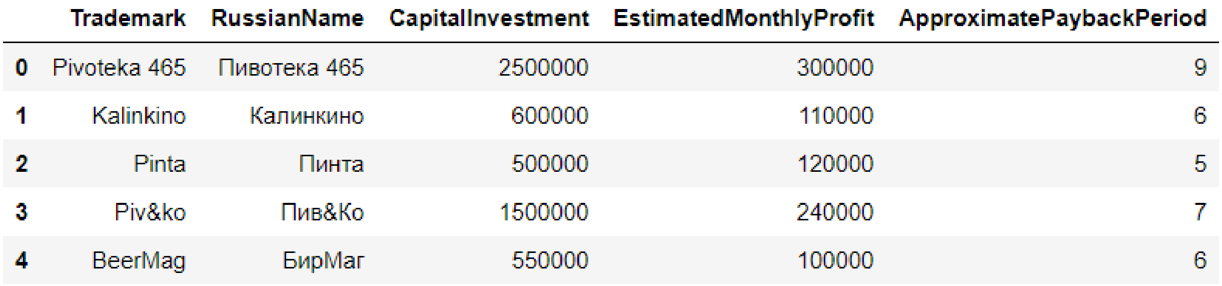


Customer's house location

The best districts (in Russian):

- Южнопортовый
- Нижегородский
- Рязанский
- Текстильщики
- Печатники
- Кузьминки
- Люблино
- Марьино
- Выхино-Жулебино
- Капотня
- Некрасовка
- Лефортово

The customer wants to consider **only** these franchises.



	Trademark	RussianName	CapitalInvestment	EstimatedMonthlyProfit	ApproximatePaybackPeriod
0	Pivoteka 465	Пивотека 465	2500000	300000	9
1	Kalinkino	Калинкино	600000	110000	6
2	Pinta	Пинта	500000	120000	5
3	Piv&ko	Пив&Ко	1500000	240000	7
4	BeerMag	БирМаг	550000	100000	6

It's really right. All these trademarks have been operating in Moscow for many years.

1. Pivoteka 465, Piv&ko are too expensive. So the customer needs to take a bank loan.
2. Kalinkino, Pinta and BeerMag have approximately the same parameters(payback Period, monthly profit etc.).

Kalinkino, Pinta and BeerMag are the most appropriate franchises.

The final decision of optimal business strategy will be completed by the customer only.