Capstone Project (Week 2)

Opening the pub in Moscow, Russia

Introduction

In recent years, the number of pubs in Russia (especially in the Central region) has increased dramatically. Some attribute this not so much to British, Irish and Czech exoticism, but to the increasing culture of beer consumption in the country and the desire to abandon the low-quality product that is offered by a mass producer.
In this study, I want to analyze and determine the **optimal places for the location of an Irish pub in Moscow**.

Moscow is the capital and largest city of Russia. Moscow is among the world's largest cities, being the most populous city entirely within Europe, so owners can easily find potential customers here.
Moscow is divided into twelve administrative okrugs and 123 districts.
In this city, there are people who are happy to go to the pub to have a beer with friends and relax after a hard working week, or just to have a good time together.

Business problem

The owners of the pubs who are the target audience of this study often face the need to expand an existing pub chain or launch a completely new pub.
It is obvious that the owners of pubs are interested in making their establishment bring a noticeable income. To do this, they should choose locations where there are people interested in opening a pub, and there is also a great area where they can open a pub.
This study aims to help owners choose the best pub location based on the following data:

- population density;
- price of apartments in different areas;
- location of competitors (other pubs and bars).
-

Data sources & their description

Based on the problem above, I will need the following data:

- Information about districts and settlements in Moscow (name of the district, administrative district, population density, etc.) from Wikipedia;
- Each district has its own geographical coordinates, this information was obtained using the site Nominatim. Nominatim uses OpenStreetMap data to find locations on Earth by name and address;
- Lists of objects and geodata for all administrative-territorial divisions in Moscow Geo;
- Rating of Moscow districts by apartment price;
- Forsquare API was used to find out the location of competitors (latitude and longitude).
-

Methodology section (data preparation & analysis)

Block 1. Data preparation

I start with importing the necessary libraries and loading the all necessary data such as:

- Moscow Boroughs

```
   Borough_index     Borough_Name District_Name             Borough_Type  \
0            1.0  Академический\n        ЮЗАО\n  Муниципальный округ\n
1            2.0   Алексеевский\n        СВАО\n  Муниципальный округ\n
2            3.0   Алтуфьевский\n        СВАО\n  Муниципальный округ\n
3            4.0          Арбат\n         ЦАО\n  Муниципальный округ\n
4            5.0       Аэропорт\n         САО\n  Муниципальный округ\n

   OKATO_Code  OKTMO_Code
0  45293554.0  45397000.0
1  45280552.0  45349000.0
2  45280554.0  45350000.0
3  45286552.0  45374000.0
4  45277553.0  45333000.0
```

- Coordinates of each borough

	Borough_Name	Latitude	Longitude
0	Академический	55.689738	37.576771
1	Алексеевский	55.814222	37.639196
2	Алтуфьевский	55.902309	37.598674
3	Арбат	55.746223	37.589367
4	Аэропорт	55.800402	37.533156

- Housing Price for each borough

	Borough_Name	Borough_Housing_Price
1	Арбат	438568
2	Хамовники	425741
3	Якиманка	404471
4	Замоскворечье	398544
5	Тверской	386255

- Population Density dataset

	Borough_Name	Borough_Area	Borough_Population	\
0	Академический	5.83	110038	
1	Алексеевский	5.29	80634	
2	Алтуфьевский	3.25	57697	
3	Арбат	2.11	36308	
4	Аэропорт	4.58	79541	

	Borough_Population_Density	Borough_Housing_Area	\
0	18874	2467.0	
1	15242	1607.9	
2	17752	839.3	
3	17207	731.0	
4	17367	1939.7	

	Borough_Housing_Area_Per_Person
0	22.7
1	20.5
2	15.5
3	26.0
4	25.9

At the end, I created the result Moscow Boroughs dataset:

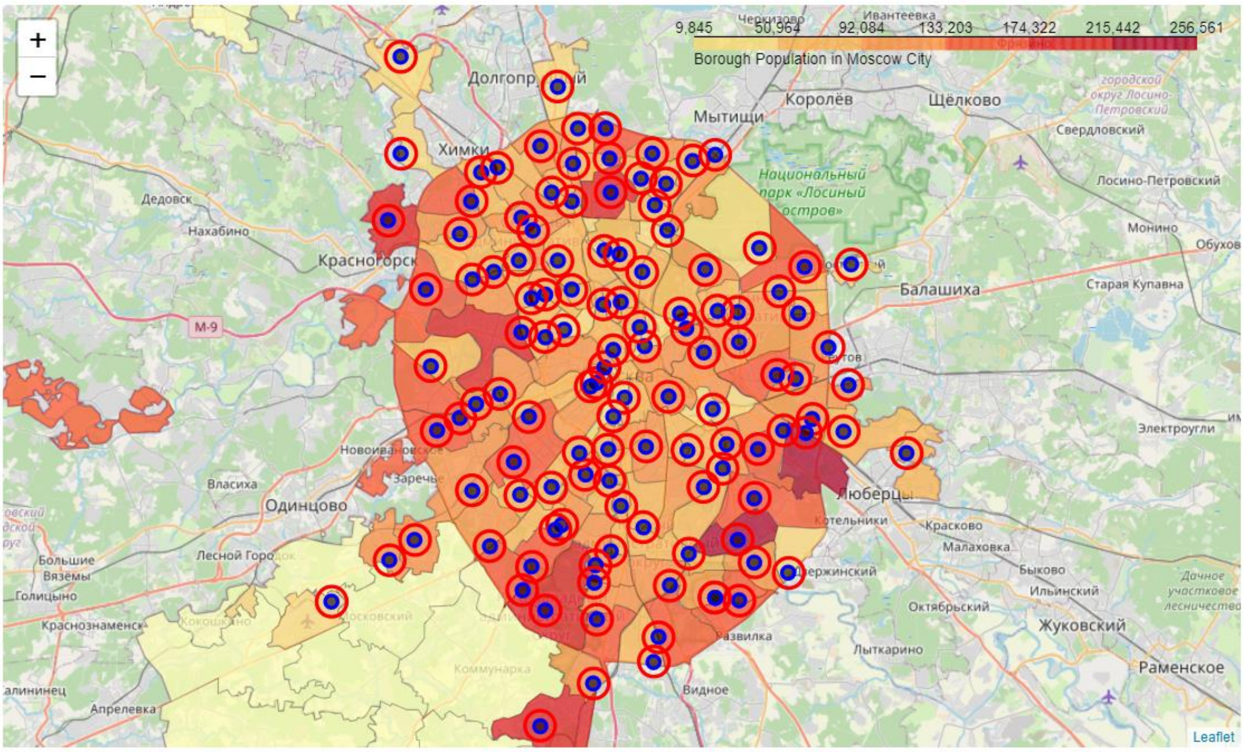
```
    Borough_Name District_Name          Borough_Type  OKATO_Code  OKTMO_Code  \
0   Академический          ЮЗАО  Муниципальный округ    45293554    45397000
1    Алексеевский          СВАО  Муниципальный округ    45280552    45349000
2    Алтуфьевский          СВАО  Муниципальный округ    45280554    45350000
3           Арбат           ЦАО  Муниципальный округ    45286552    45374000
4        Аэропорт           САО  Муниципальный округ    45277553    45333000

   Borough_Area  Borough_Population  Borough_Population_Density  \
0          5.83              110038                       18874
1          5.29               80634                       15242
2          3.25               57697                       17752
3          2.11               36308                       17207
4          4.58               79541                       17367

   Borough_Housing_Area  Borough_Housing_Area_Per_Person   Latitude  \
0                2467.0                             22.7  55.689738
1                1607.9                             20.5  55.814222
2                 839.3                             15.5  55.902309
3                 731.0                             26.0  55.746223
4                1939.7                             25.9  55.800402

   Longitude  Borough_Housing_Price
0  37.576771               199999.0
1  37.639196               199474.0
2  37.598674               138021.0
3  37.589367               438568.0
4  37.533156               234544.0
```

After this, it was necessary to create a map of Moscow Boroughs



Work with Foursquare API allows to get full information about each cell & venue:

```
                              Cell_id  Cell_Latitude  Cell_Longitude  \
0  55.67695842926316,37.18672468925941      55.676958       37.186725
1  55.67695842926316,37.18672468925941      55.676958       37.186725
2  55.67695842926316,37.18672468925941      55.676958       37.186725
3  55.67695842926316,37.18672468925941      55.676958       37.186725
4  55.64659468143242,37.34558577603909      55.646595       37.345586

                   Venue_Id                        Venue_Name  \
0  5850fa8818dc531c1fb939ea                  Мегафон Экспресс
1  4fc10ddae4b08acecb4c714b                        Ретро кафе
2  51cc6b8c498eaabf4b5fba29                 ООО "VL Computers"
3  4f3f5416e4b02d0e6b676a1b                     Рынок за КПП-1
4  4d31b8455017a093fdf3419b  Салон красоты Наталии Волошиной

                              Venue_All_Categories  Venue_Latitude  \
0  [('Mobile Phone Shop', '4f04afc02fb6e1c99f3db0...       55.678210
1             [('Café', '4bf58dd8d48988d16d941735')]       55.678093
2  [('Electronics Store', '4bf58dd8d48988d1229517...       55.679381
3           [('Market', '50be8ee891d4fa8dcc7199a7')]       55.678750
4  [('Salon / Barbershop', '4bf58dd8d48988d110951...       55.643984

   Venue_Longitude           Venue_Location  Venue_Distance  Borough_Name
0        37.187960                  Власиха           159.0           NaN
1        37.187721                  Власиха           140.0           NaN
2        37.189140                   Россия           309.0           NaN
3        37.188707                   Россия           235.0           NaN
4        37.343332  ул. Шолохова, 30, 119634           323.0           NaN
```

GeoJSON file with boroughs allows to create geometry shape and correlate each venue to Moscow Boroughs where they were placed:

```
                        Venue_Name Venue_Category_Name      Borough_Name
0                 Мегафон Экспресс   Mobile Phone Shop               NaN
1                       Ретро кафе                Café               NaN
2               ООО "VL Computers"   Electronics Store               NaN
3                   Рынок за КПП-1              Market               NaN
4  Салон красоты Наталии Волошиной  Salon / Barbershop  Ново-Переделкино
5             Спортзал у Дяди Жени  Athletics & Sports  Ново-Переделкино
6                        Del Gusto  Italian Restaurant  Ново-Переделкино
7   Остановка «Лукинская улица, 1»            Bus Stop  Ново-Переделкино
8                  "Сеньор Помидор"   Food & Drink Shop  Ново-Переделкино
9                          Копейка    Department Store  Ново-Переделкино
```

Then I removed the venues that located outside of the Moscow districts.

The first block of loading, processing, and preparing data for further analysis *is completed*.

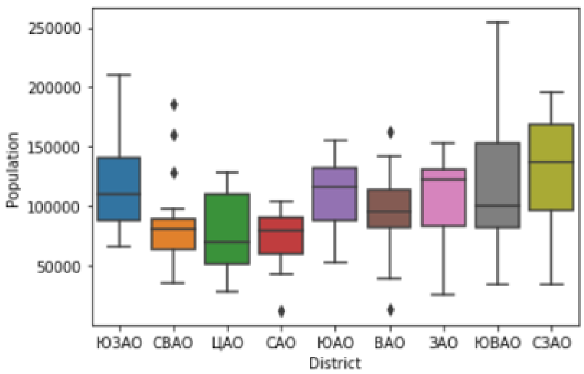
Block 2. Data analysis

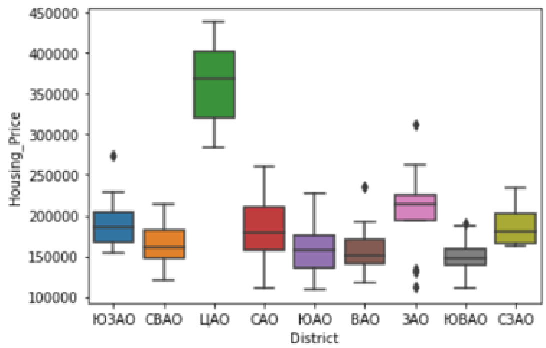
The second block starts with descriptive statistical analysis, where i got basic statistics for all features:

	Area	Population_Density	Housing_Area	Population	Housing_Price
count	120.000000	120.000000	120.000000	120.000000	120.000000
mean	8.706417	13477.566667	1775.684167	100188.700000	190037.316667
std	4.927028	5965.300074	815.978445	44012.960386	66182.885601
min	2.110000	563.000000	69.900000	12264.000000	109421.000000
25%	5.395000	9770.750000	1244.450000	72498.500000	147339.000000
50%	7.680000	13543.500000	1709.450000	94166.000000	168172.500000
75%	10.282500	17217.500000	2206.600000	127001.000000	210978.000000
max	27.570000	30479.000000	4523.000000	254142.000000	438568.000000

Moscow Boroughs has non-uniform population from 12 264 to 254 142 people. The housing price varies from 109 421 to 438 568 rubles/m².

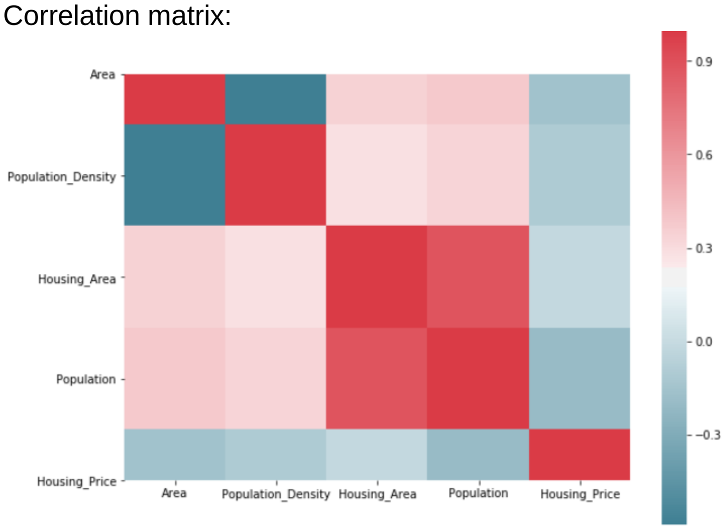
Then i created boxplots:





Right boxplot demonstrates that feature ‘District’ can be a good predictor for ‘Housing Price’.

Correlation matrix:



I used the Pearson Correlation Coefficient and P-value to determinate features with significant correlation & strong relationship:

```
The Pearson Correlation Coefficient 'Area' to 'Population' is 0.3801548969695431  with a P-value of P = 1.846169214258945e-05
The Pearson Correlation Coefficient 'Area' to 'Population_Density' is -0.5886402260872453  with a P-value of P = 1.542565268939
8923e-12
The Pearson Correlation Coefficient 'Area' to 'Housing_Price' is -0.15499599520906004  with a P-value of P = 0.0909599362567613
1
The Pearson Correlation Coefficient 'Housing_Area' to 'Area' is 0.3441883147278516  with a P-value of P = 0.0001184930655545850
8
The Pearson Correlation Coefficient 'Housing_Area' to 'Population_Density' is 0.2836057432971165  with a P-value of P = 0.00169
56094309013407
The Pearson Correlation Coefficient 'Population_Density' to 'Population' is 0.33569063819980893  with a P-value of P = 0.000178
11926001207908
The Pearson Correlation Coefficient 'Population_Density' to 'Housing_Price' is -0.10414012398413303  with a P-value of P = 0.25
766045291148676
The Pearson Correlation Coefficient 'Housing_Area' to 'Population' is 0.8853448154060466  with a P-value of P = 4.6693758864781
55e-41
The Pearson Correlation Coefficient 'Housing_Area' to 'Housing_Price' is -0.0169708163901411  with a P-value of P = 0.854034357
178659
The Pearson Correlation Coefficient 'Population' to 'Housing_Price' is -0.1983247003732802  with a P-value of P = 0.02989976311
718884
```

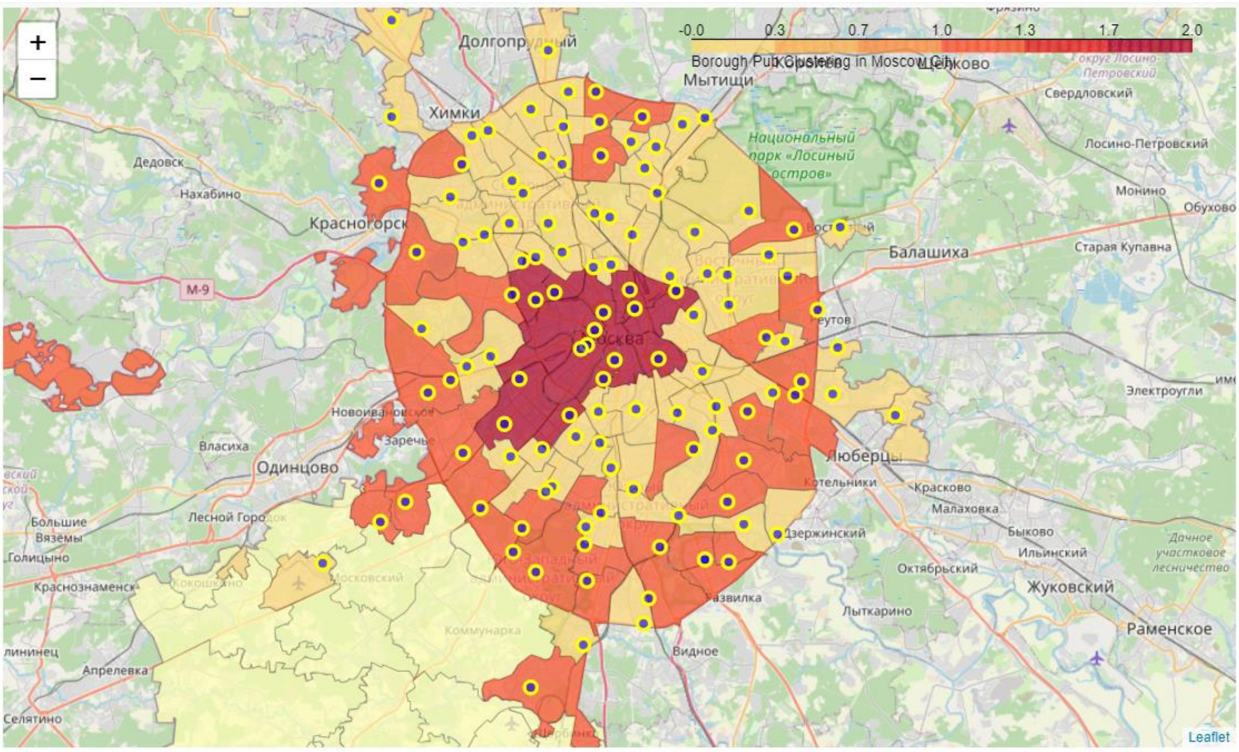
Correlation between 'Area' to 'Population Density' is significant and the linear relationship is strong. The same we can say about 'Housing Area' and 'Population'.

Then I used K-means to get clusters of boroughs:

	Cluster_Labels	Population_Mean	Housing_Price_Mean	Population_Sum	Population_%	Borough_Count	Area_Sum	Area_%	Population_Density
0	0	78939.661972	173695.070423	5604716	46.617999	71	539.87	51.673574	10381.602978
1	1	153465.529412	160741.323529	5217828	43.400004	34	391.25	37.448434	13336.301597
2	2	80006.666667	333794.866667	1200100	9.981997	15	113.65	10.877992	10559.612846

Cluster № 1 has the highest mean population and population density, also it has the smallest mean housing price, so it is perfect for solving the task that I set myself in this project.

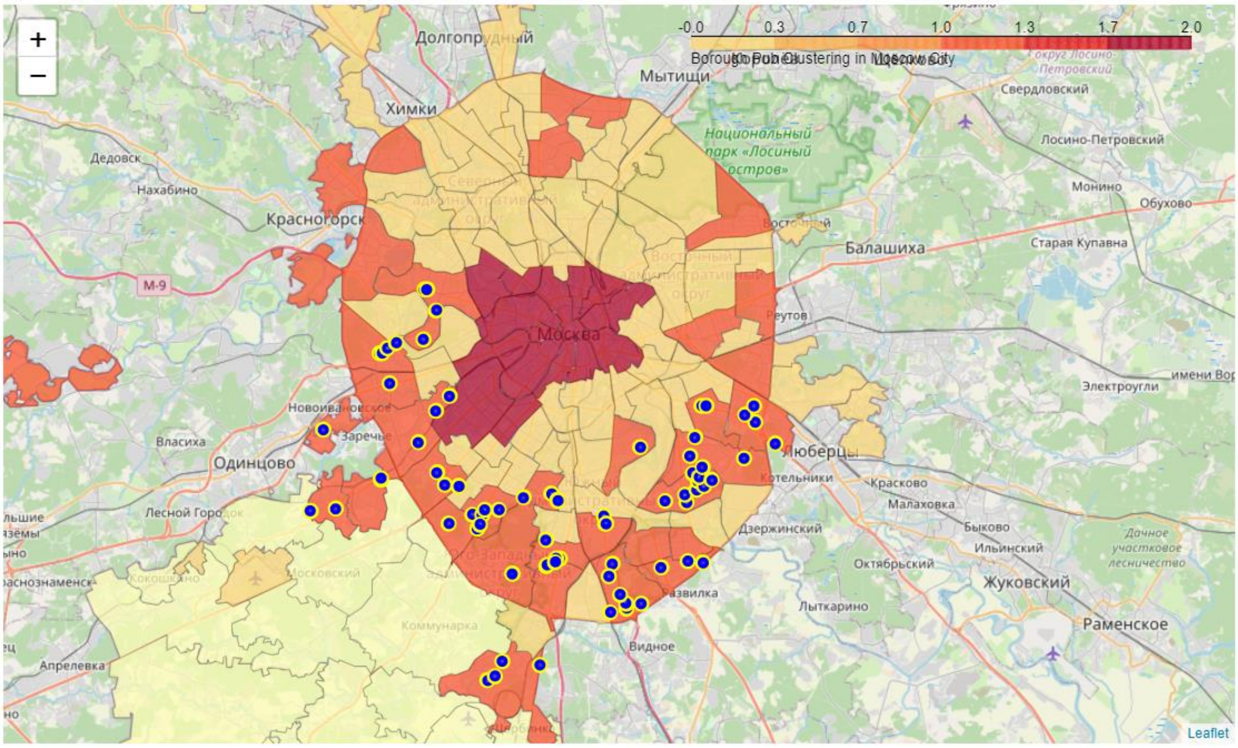
Clusters visualizing:



I used the identification of the dataset that includes our potential competitors in Cluster 1 by using the name of Venue categories:

```
pub_categories = ['Beer Garden', 'Beer Store', 'Beer Bar', 'Irish Pub', 'Pub', 'Gastropub', 'Bar']
```

Visualizing all potential competitors in Cluster 1:



Results

As a result we define the best areas for the location of the pub, according to these criteria:

- high population in the area;
- low cost of apartments in the area.

Also we got list and location of our potential competitors.

Discussion

As a result of this project, a large amount of information was collected about Moscow boroughs and venues, all collected data were listed in the Methodology section.

Since the information will be publicly available, anyone can access it if necessary to use it in their future research.

Perhaps future research can use a different approach to data visualization, delve into segmentation of existing pubs and bars, and take into account all existing cafes and restaurants.

Conclusion

In conclusion, I would like to say that I hope that the results of my report can help owners to find the best place for their new pubs.