Venues Data Analysis of Moscow City

Introduction

Background

Moscow, one of the largest metropolises in the world with a population of more than 12 million people, covers an area of more than 2561.5 km² with an average density of inheritance of 4924.96 people / km² [1].

Moscow is divided into 12 districts (125 boroughs, 2 urban boroughs, 19 settlement boroughs).

Moscow has a very uneven population density from 30429 people / km² for the "Зябликово" borough, to 560 people / km² for the "Молжаниновский" borough [2].

The average cost of real estate varies from 68,768 rubles / m² for the "Кленовское" borough to 438,568 rubles / m² for the "Арбат" borough [3].

Business Problem

Owners of cafes, fitness centers and other social facilities are expected to prefer boroughs with a high population density. Investors will prefer areas with low housing costs and low competitiveness.

On the part of residents, the preference is expected for a boroughs with a low cost of housing and good accessibility of social places.

In my research, I will try to determine the optimal places for the location of fitness centers in Moscow boroughs, taking into account the number of people, the cost of real estate and the density of other fitness facilities.

The key criteria for selecting suitable locations for fitness centers will be:

- High density of the borough population
- Low cost of real estate in the area
- The absence in the immediate vicinity of other fitness facilities of a similar profile

I will use the approaches and methods of machine learning to determine the location of fitness centers in accordance with the specified criteria.

The main stakeholders of my research will be investors interested in opening new fitness centers.

Data acquisition and cleaning

Data requirements

Based on the problem and the established selection criteria, to conduct the research, I will need the following information:

1. main dataset with the list of Moscow Borough, containing the following attributes:
   - name of the each Moscow Borough
   - type of the each Moscow Borough
   - name of the each Moscow District in which Borough is belong to
   - area of the each Moscow Borough in square kilometers
   - the population of the each Moscow Borough
   - housing area of the each Moscow Borough in square meters
   - average housing price of the each Moscow Borough
2. geographical coordinates of the each Moscow Borough
3. shape of the each Moscow Borough in GEOJSON format
4. list of venues placed in the each Moscow Borough with their geographical coordinates and categories

Decribe data sources

Moscow Boroughs dataset

Data for Moscow Boroughs dataset were downloaded from multiple HTTP page combined into one pandas dataframe.

- List of Moscow District and they Boroughs were downloaded from the page Moscow Boroughs
- Information about area of the each Moscow Borough in square kilometers, their population and housing area in square meters were downloaded from the page Moscow Boroughs Population Density
- Information about housing price of the each Moscow Borough were downloaded from the page Moscow Boroughs Housing Price

A special Python function has been developed for HTML table parse. This function help me:

- to find number of rows and columns in a HTML table
- to get cloumns titles, if posible
- to convert string to float, if posible
- return result in form of the Pandas dataframe

Moscow Boroughs geographical coordinates

Geographical coordinates of the each Moscow Borough were queried through Nominatim service. As the Nominatim service are quite unstable it was quite a challenge to request coordinate in several iterations.

Moscow Boroughs shape in GEOJSON format

Shape of the each Moscow Borough in GEOJSON format was downloaded from the page Moscow Boroughs GEOJSON

Moscow Boroughs venues

To determine **venues** the service **Forsquare API** was used.
The API of **Forsquare** service have the restriction of 100 **venues**, which it can return in one request.
To obtain list of all **venues** I used the following approach:

- present Moscow area in the form of a regular grid of circles of quite small diameter, no more than 100 **venues** in each circle
- perform exploration using **Forsquare API** with quite bigger radius than circle of a grid to make sure it overlaps/full coverage to don't miss any venues
- cleaning list of venues from duplicates.

This approach and some of the Python code was taken from the work presented here.
https://cocl.us/coursera_capstone_notebook

Circle of 28 000 meter in radius cover all Moscow Boroughs.
In my research grid of circles contains 7899 cells with radius 300 meter.
Foursquare API have a certain limitation for API call in one day to explore venues.
In my case it was about 2000 calls per day.
So in addition I have to divide grid dataset into subset and call Foursquare API for several days.

Decribe data cleansing

Moscow Boroughs dataset cleansing

As data for Moscow Boroughs dataset were downloaded from multiple HTTP page it was necessary to perform a data cleaning. Such as:

- remove some unused colums
- strip text columns from additional information like ' \n\t'
- replace some Borough_Name as of russian letters "e" and "ё"
- change places of some words in Borough_Name
- clear Borough Name from additional information, such as ', поселение ', ', городской округ '

- replace '\n', ' ↗' and '↘' in some columns
- delete extra spaces in numeric columns
- replace ',' to '.' for float columns
- convert from float to int for integer columns
- convert from string to float for numeric columns

As the result I had a dataset with all 146 Moscow Boroughs. Result dataset contains columns:

- **Borough_Name** - name of the Moscow Borough - is a unique key of the dataset
- **District_Name** - name of the Moscow District in which Borough is belong to
- **Borough_Type** - type of the Moscow Borough
- **OKATO_Borough_Code** - numeric code of the Moscow Borough
- **OKTMO_District_Code** - numeric code of the Moscow District
- **Borough_Area** - area of the Moscow Borough in square kilometers
- **Borough_Population** - population of the Moscow Borough
- **Borough_Population_Density** - population density of the Moscow Borough
- **Borough_Housing_Area** - housing area of the Moscow Borough in square meters
- **Borough_Housing_Area_Per_Person** - housing area per person of the Moscow Borough in square meters
- **Latitude** - geographical Latitude of the Moscow Borough
- **Longitude** - geographical Longitude of the Moscow Borough
- **Borough_Housing_Price** - average housing price of the Moscow Borough

I had a problem to found proper statistics about "housing prices" and "housing area" for some Moscow boroughs, so I had to exclude 26 boroughs from my analysis.
Fortunately, they all had a low population density, which meat criteria of my research and did not reduce it quality.

The result Moscow Boroughs dataset

Index	Borough_Name	District_Name	Borough_Type	ATO_Borough_Co	TMO_District_C	Borough_Area	orough_Populatic	Populatic	rough_Housing_Ar	using_Are	Latitude	Longitude	orough_Housing_Pric
0	Академический	ЮЗАО	Муниципальный округ	45293554	45397000	5.83	109387	18762	2467.00	22.70	55.69	37.58	199999.00
1	Алексеевский	СВАО	Муниципальный округ	45280552	45349000	5.29	80534	15223	1607.90	20.50	55.81	37.65	199474.00
2	Алтуфьевский	СВАО	Муниципальный округ	45280554	45350000	3.25	57596	17721	839.30	15.50	55.88	37.58	138021.00
3	Арбат	ЦАО	Муниципальный округ	45286552	45374000	2.11	36125	17120	731.00	26.00	55.75	37.59	438568.00
4	Аэропорт	САО	Муниципальный округ	45277553	45333000	4.58	79486	17355	1939.70	25.90	55.80	37.53	234544.00
5	Бабушкинский	СВАО	Муниципальный округ	45280556	45351000	5.07	88537	17462	1586.30	18.50	55.87	37.66	164324.00
6	Басманный	ЦАО	Муниципальный округ	45286555	45375000	8.37	110694	13225	1991.80	18.40	55.78	37.69	302021.00
7	Беговой	САО	Муниципальный округ	45277556	45334000	5.56	42781	7694	791.10	18.80	55.78	37.57	261402.00
8	Бескудниковский	САО	Муниципальный округ	45277559	45335000	3.30	79603	24122	1391.70	18.40	55.86	37.56	158398.00
9	Бибирево	СВАО	Муниципальный округ	45280558	45352000	6.45	160163	24831	2521.80	15.80	55.88	37.60	140533.00
10	Бирюлёво Восточное	ЮАО	Муниципальный округ	45296553	45911000	14.77	155863	10552	2122.20	14.70	55.59	37.66	124645.00
11	Бирюлёво Западное	ЮАО	Муниципальный округ	45296555	45912000	8.51	88672	10419	1183.20	13.20	55.59	37.64	109421.00
12	Богородское	ВАО	Муниципальный округ	45263552	45301000	10.24	109324	10676	1744.10	16.90	55.82	37.71	178577.00
13	Братеево	ЮАО	Муниципальный округ	45296557	45913000	7.63	110021	14419	1585.40	15.50	55.64	37.76	136300.00
14	Бутырский	СВАО	Муниципальный округ	45280561	45353000	5.04	71458	14178	1236.20	18.30	55.81	37.59	182641.00
15	Вешняки	ВАО	Муниципальный округ	45263555	45302000	10.72	122285	11407	1976.80	16.20	55.73	37.82	147352.00
16	Внуково	ЗАО	Муниципальный округ	45268552	45317000	17.42	25471	1462	416.60	17.80	55.61	37.30	113399.00
17	Войковский	САО	Муниципальный округ	45277565	45336000	6.61	70729	10700	1531.00	23.10	55.82	37.49	207242.00
18	Восточное Дегунино	САО	Муниципальный округ	45277568	45337000	3.77	98923	26239	1592.50	16.70	55.88	37.56	146300.00

Moscow Boroughs geographical coordinates cleansing

Nominatim service not only quite unstable.
It also have a occasionally problem with russian leter ё. So I have to manyaly obtain coordinates for such boroughs as:

- Десёновское, Поселение, Новомосковский
- Савёлки, Муниципальный округ, ЗелАО
- Клёновское, Поселение, Троицкий
- And some others.

Another problem with Nominatim service is that it return not very accurate coordinate of some Boroughs.
So I needed to adjust they manually in the map.

Moscow Boroughs shape in GEOJSON format cleansing

GEOJSON file downloaded from the page Moscow Boroughs GEOJSON was quite good and not requied any addition clearing.

Boroughs Population in Moscow City
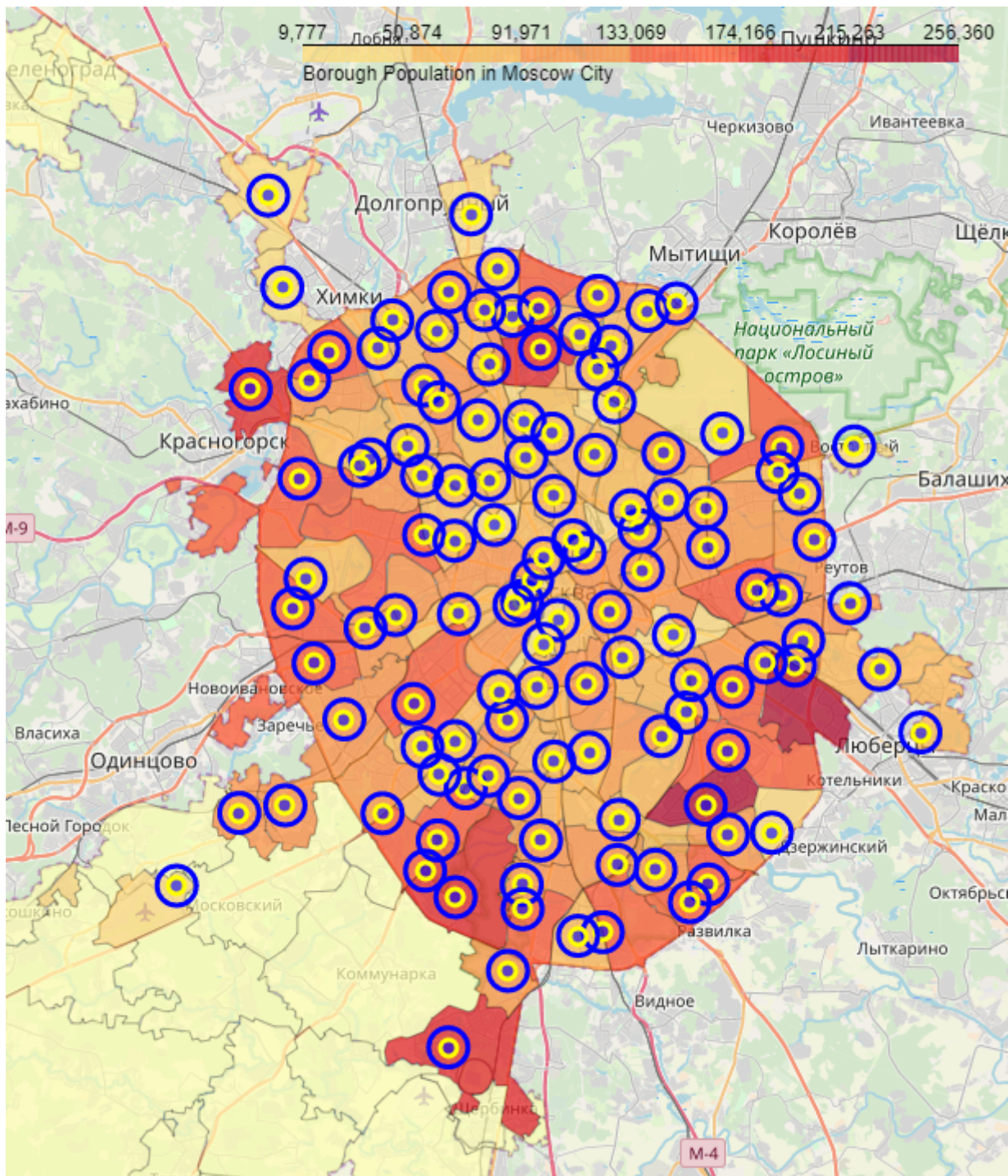
Moscow Boroughs venues cleansing

Usning **Forsquare API** I obtrained 34460 venues in 7899 cells.
As I used a quite bigger radius (350 meters) for venue explorations than circle of a grid (300 meters), there was a need to remove duplicates venus.
After duplicates removal I had 27622 unique venues in the circle radius of 28 000 meters around the Moscow City.

The second task was to bind each venue to Moscow Boroughs in which borders they were placed.
To perform this task I created a polygons for each Moscow Borough from GEOJSON file and found wich venues coordinate included into each polygon.

The third task was to remove all the venues that placed outside of the Moscow boroughs.

The fourth tas was to get main category from the category list for each venue.

As the result I had list of 20864 venues placed in the Moscow Boroughs with their geographical coordinates and categories