Moscow areas classification

For the purpose of apartments purchase

Oleg Adamovich

May 20, 2020

Introduction

Moscow is a capital of Russia and its biggest city with diverse set of environments. The ultimate problem I will be trying to solve is area classification within the city.

Data driven decisions are good decisions. Can i use available data to help myself and other people with making right choice for changing their environment? Selling old and buying a new apartments flat is significant and nervous decision in people's lives. Giving more confidence and guiding decisions can be something which people will be looking for. Decision and confidence increase can be achieved with the reduction of uncertainty of area of future living. What are the schools in the area, are there any theatres? how good the cafes are? Several typical questions in peoples mind before going to new area. As well there are so **many** areas to look for a new flat in a big city. The business question is if the area specific data can be used to shrink down the area of search and help to save time and reduce uncertainty around expectations. This can be achieved through clustering city into areas with similar attributes so people won't waste their time for looking for a place to live in the area which might not fit for their expectations. Rather focus on the areas with expected presence of the priorities of choice.

Data

Metro stations and its coordinates were used as a central point for getting data. List and coordinates can be found on this link [Metro and Coordinates](). Information over apartments can be found on this link [Apartments pricing over the metro](). The final product of this stage is a join dataset containing Metro name, coordinates and pricing associated with the metro location. Metro station coordinates serve as a central point for getting information of the area. Preliminary screening of FourthSquare API suggests that there is limited description of the city regarding different types of venues. Figure 1 show example of school location found on google map and ForthSquare. Google map have 2x times more of venues. Detailed studies suggest that there more ratings which can be used to have more reliable classification.

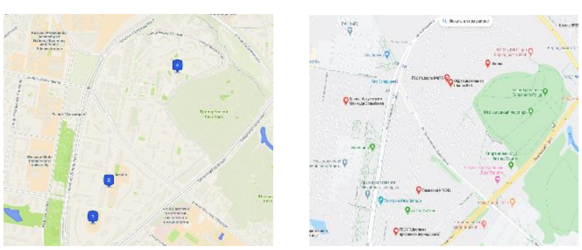

Figure 1. ForthSquare(left) vs Google (Right) comparison

As a result of this screening Google maps API selected as a provider to search for venues. Google provides certain amount of request for free of price basis. Charging policy can be found on the web. Besides this was of some interest to compare different data provider, the flexibility, interface and other components of usability.

Google map provide around 100 different types for request like schools, universities, cafes. For the purpose of reducing number of requests around 20 key were selected for further:

[Aquarium, art_gallery, bakery, bank, bar, bus_station, café, car_repair, clothing_store, doctor, electronics_store, gym, home_goods_store, hospital, laundry, library, movie_theater, museum, park, pharmacy, primary_school, restaurant, school, secondary_school, spa, stadium, store, super market, tourist_attraction, university, zoo].

List 1 – Data types used to get venue information around Metros.

Number of venues acquired from the Google API was from NaN upto 60 for each type. Along with venue type average rating and number of rating were acquired as well. This was pretty long exercise to get all the selected data and took around 8 hours to get it. The Final dataset include average rating, number of venues and number of reviews associated with the venue.

Methodology

First step was data analysis. The starting point was to look the costs and how they are located over the city. Figure 2 shows histogram of the data of this type. Four types/clusters can be easily picked from the visual inspection of the histogram. Because of simple, "1D" nature of cost data I decided no to implement any data clustering techniques but select clusters based on histogram inspection. Four main classes of areas were selected based on price per sq meter: 0-173000 Rub, 173000-222223 Rub, 222223-288220 Rub, 228220-end.

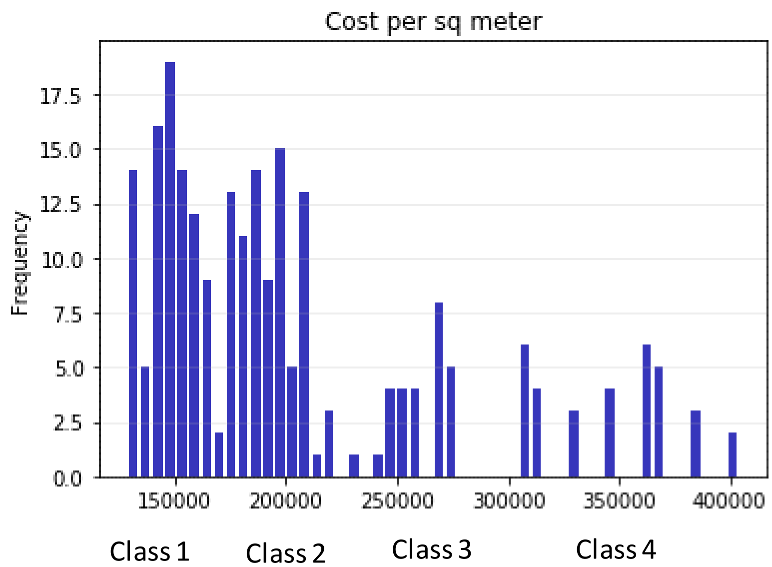

Figure 2. Pricing per sq. m related to metro histogram

Figure 3 show spatial map of the selected classes. Quite obviously the more expensive areas are in the center of the city with the trend for prices to drop down towards the edges of the city. The observation which can be made is the North-East - South-West trend of increased pricing.

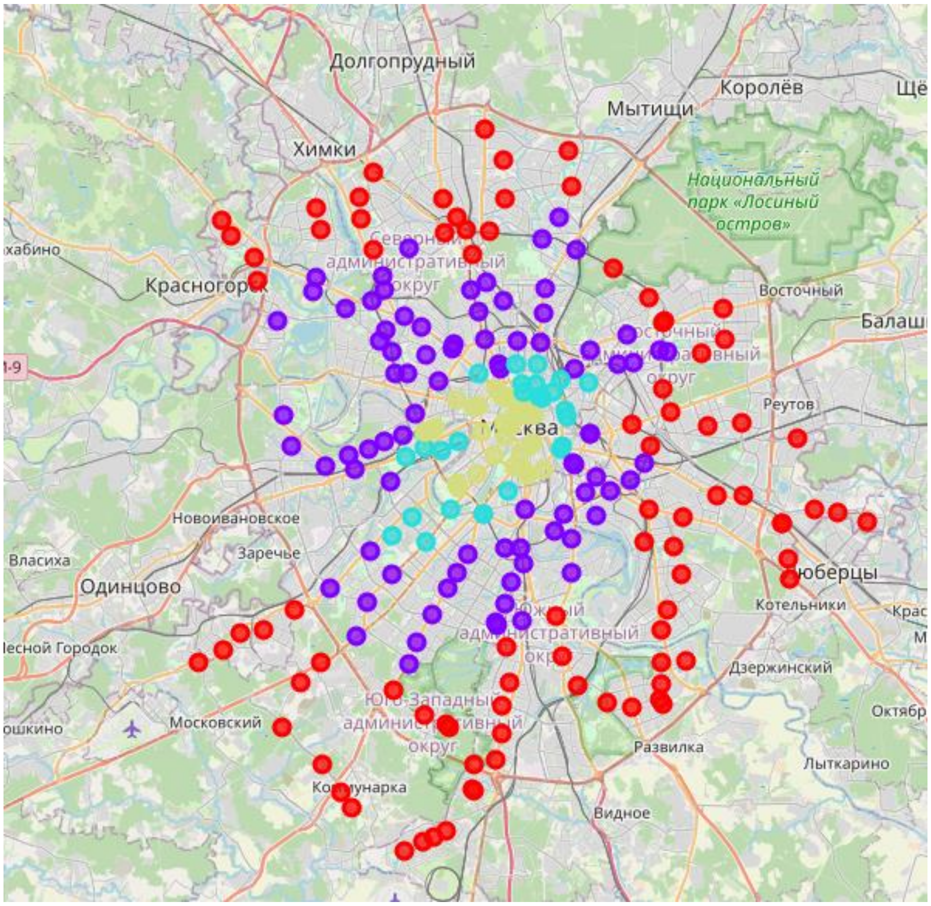

Figure 3. City map with price classes overlaid.

Following manipulation were applied to the data before final data clustering exercise:

- Average rating per venue type was calculated using following equation $\frac{\sum_0^{num} AvRating * NumRatings}{\sum_0^{num} NumRatings}$
- For number of venues and number of ratings - NaN were replaced with zeros.
- For average rating - NaN numbers were replaced with mean numbers.
- Data was normalized on column basis using mini-mx approached

Kmeans algorithm was selected to group data into different clusters. Data for clustering consisted of a very different data types with significant variation of statistics. Kmeans algorithm was tested on a raw data as well on processed and normalized data. Produced maps were compared with the clustering based on price only. This clustering exercise was repeated for each price class independently to provide detailed information fit for the business needs.

K mean clustering was implemented using Kmean function from sklearn library. Key testing parameter was a number of clusters. This was tested by the visual inspection of the maps. As a result, the number of 6 clusters was used for most of classifications.

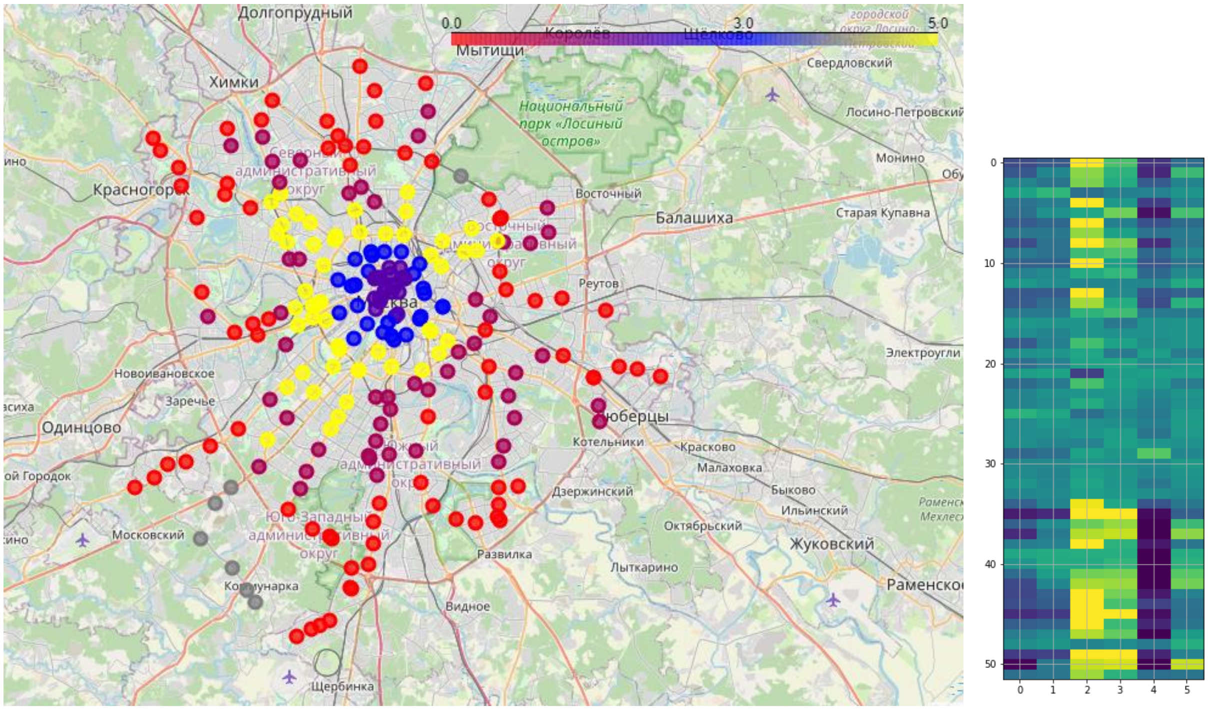

Figure 4. City map with Kmeans clusters overlaid (left). MD matrix(Right)

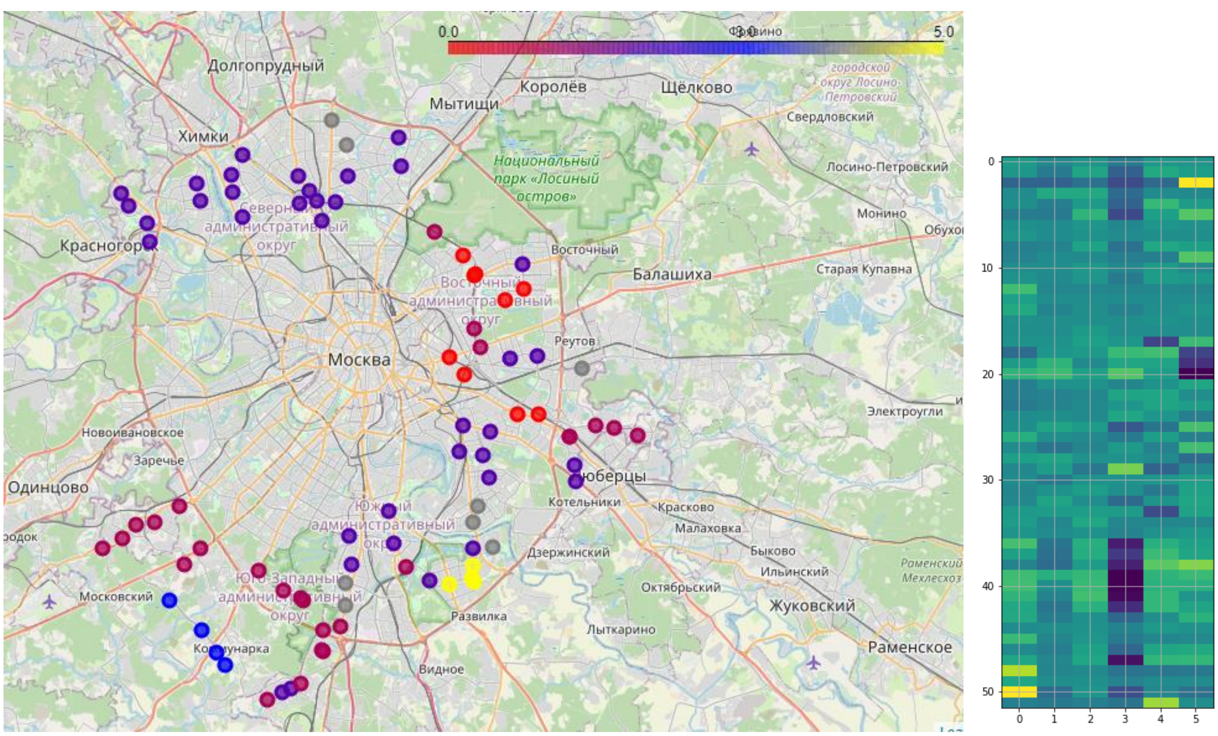

Figure 5. City map with Kmeans clusters overlaid for **first** price class (left). MD matrix (Right)

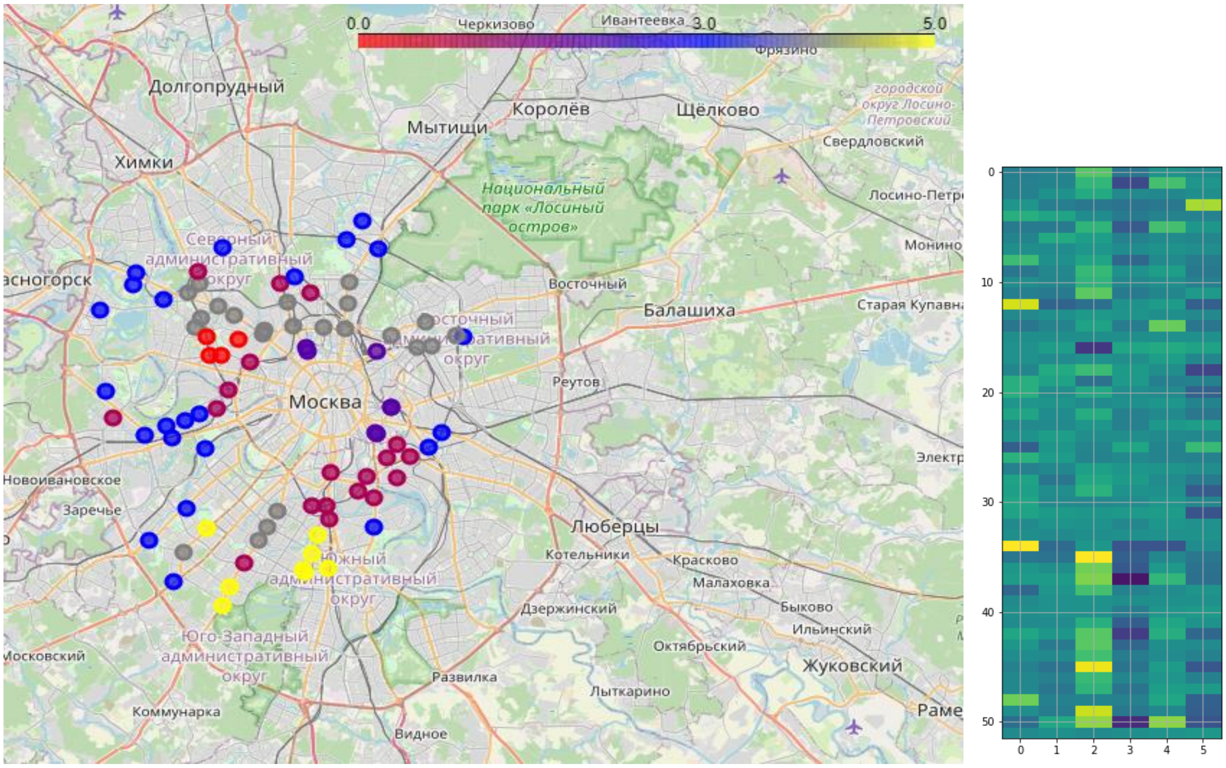

Figure 6. City map with Kmeans clusters overlaid for **second** price class (left). MD matrix (Right)

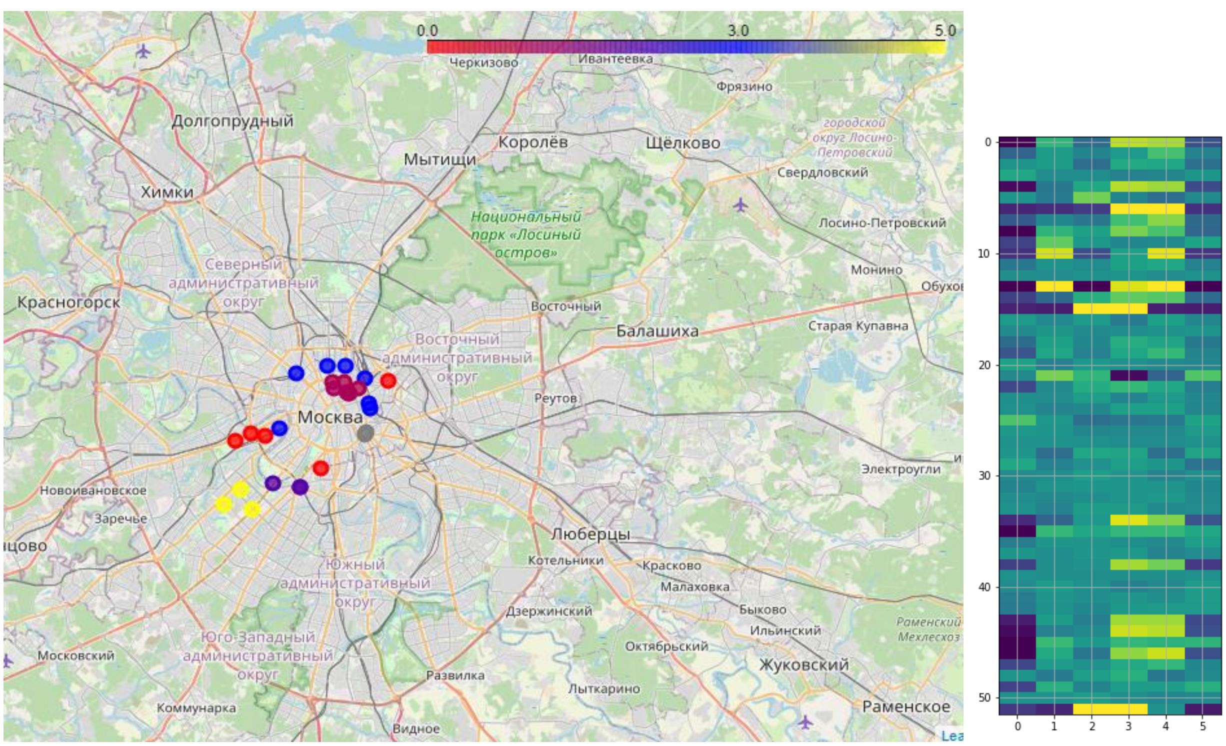

Figure 7. City map with Kmeans clusters overlaid for **third** price class (left). MD matrix (Right)

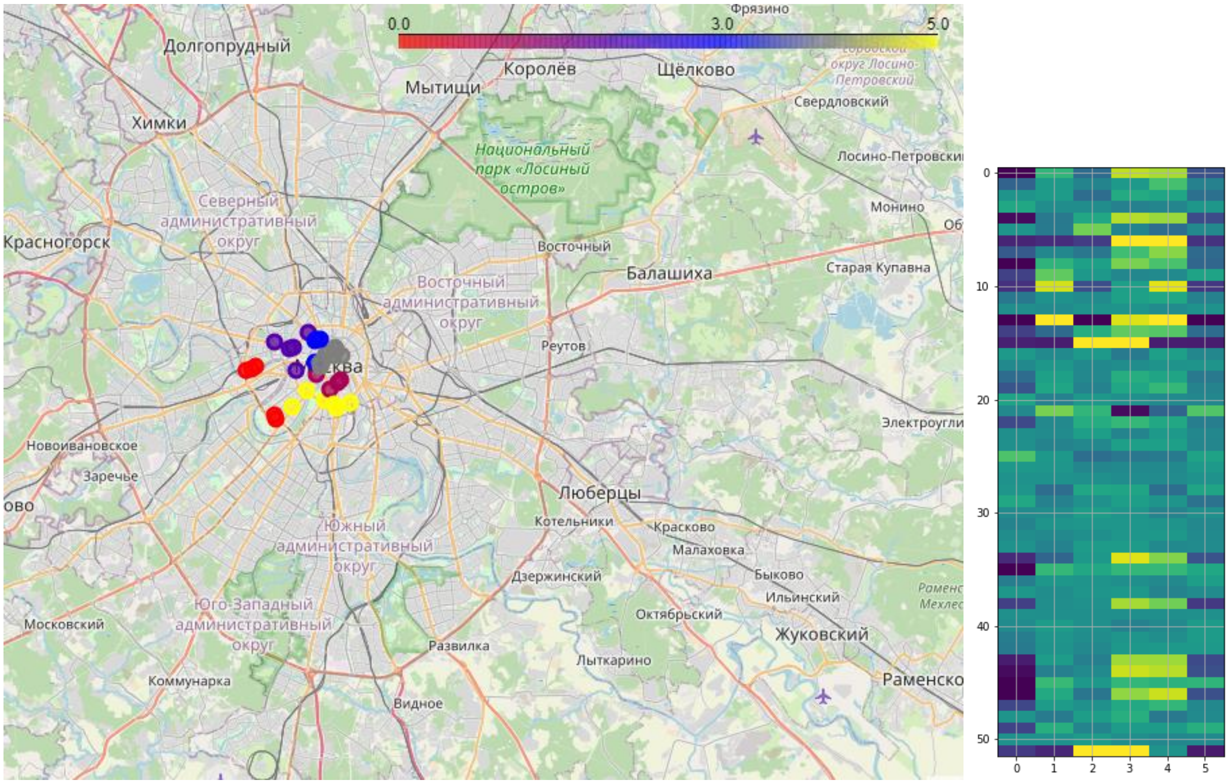

Figure 8. City map with Kmeans clusters overlaid for **forth** price class (left). MD matrix (Right)

Results

Results were evaluated by visual inspections of the Kmean clusters on the city map and mean-difference matrix (MDM). MDM is calculated by subtracting actual value of attribute out of a class from mean number for all 6 clusters and later used for providing clusters characteristics.

Figure 4 shows the results of the area clustering when all the dataset was used. As mentioned before there are good visual correlation with the area classes based on pricing criteria and good indicate that data is somehow fit-for modelling. Unfortunately, this tells nothing else except that to use pricing judgment for selecting the area of living.

Usually one is relatively familiar the budget and potential expenses on new place of leaving. Thus, one can easily associate himself with the cost class on early stages of selecting a new place of living. This logic suggested to repeat data clustering exercise but separately within each price class

Results of Kmeans clustering of each price class are shown on figures 5 to 8. Maps show surprisingly good spatial clusters of the areas. For all four classes. For example, Class 1 have a good spatial zone distribution - violet on north of the city, red on east, cherry on south-west etc.

Class 1

```
Key characteristics of 0 cluster are:
Most outstanding attributes are :
39        aquarium_tot  0.488168
17         stadium_usr  0.348106
53         stadium_tot  0.222728
13    movie_theater_usr  0.147650
31    movie_theater_avr  0.110320
24            bank_avr  0.104317
9            store_usr  0.101675
25     bus_station_avr  0.100078
47             spa_tot  0.071705
35         stadium_avr  0.069695
Most outstanding attributes are :
32        museum_avr -0.052243
16        school_usr -0.055222
5           café_usr -0.058382
10       medical_usr -0.071180
14        museum_usr -0.085431
19    university_usr -0.086752
40   art_gallery_tot -0.086878
55    university_tot -0.111144
43   bus_station_tot -0.133609
30       library_avr -0.156844
```

Class 2

```
Key characteristics of 1 cluster are:
Most outstanding attributes are :
23                café_avr  0.089293
11                 spa_usr  0.087900
55          university_tot  0.075040
9                store_usr  0.066100
24                bank_avr  0.062326
29                 spa_avr  0.047425
21            aquarium_avr  0.044924
31       movie_theater_avr  0.043530
42                bank_tot  0.041890
36   tourist_attraction_avr  0.035446
Most outstanding attributes are :
54   tourist_attraction_tot -0.063894
48              library_tot -0.066476
10              medical_usr -0.069611
16               school_usr -0.070455
52               school_tot -0.085424
33                 park_avr -0.087466
40          art_gallery_tot -0.093896
50               museum_tot -0.096113
39             aquarium_tot -0.121481
17              stadium_usr -0.140722
```

Class 3

```
Key characteristics of 2 cluster are:
Most outstanding attributes are :
40          art_gallery_tot  0.448836
50               museum_tot  0.379355
54   tourist_attraction_tot  0.336273
55           university_tot  0.265642
42                 bank_tot  0.246172
41                 café_tot  0.226827
16               school_usr  0.199185
48              library_tot  0.195930
47                  spa_tot  0.191118
5                  café_usr  0.189992
Most outstanding attributes are :
26             car_avr -0.012992
56             zoo_tot -0.017115
35         stadium_avr -0.027575
30         library_avr -0.037238
53         stadium_tot -0.039176
31   movie_theater_avr -0.042640
8              car_usr -0.079917
24            bank_avr -0.097967
17         stadium_usr -0.103448
21        aquarium_avr -0.271667
```

Class 4

```
Key characteristics of 3 cluster are:
Most outstanding attributes are :
30              library_avr  0.071521
21             aquarium_avr  0.069585
23                 café_avr  0.035571
33                 park_avr  0.035108
36   tourist_attraction_avr  0.033670
24                 bank_avr  0.032711
32               museum_avr  0.019390
18   tourist_attraction_usr  0.015374
17              stadium_usr  0.011833
34               school_avr  0.009066
Most outstanding attributes are :
48      library_tot -0.135126
40  art_gallery_tot -0.155719
46      medical_tot -0.159835
10      medical_usr -0.168784
41         café_tot -0.180495
39     aquarium_tot -0.189368
6          bank_usr -0.221271
47          spa_tot -0.255249
55   university_tot -0.299369
42         bank_tot -0.352206
```

Class 5

```
Key characteristics of 4 cluster are:
Most outstanding attributes are :
55  university_tot  0.262225
19  university_usr  0.213604
6         bank_usr  0.164581
10     medical_usr  0.162781
42        bank_tot  0.129002
47         spa_tot  0.097645
21    aquarium_avr  0.090335
46     medical_tot  0.074684
53     stadium_tot  0.070555
52      school_tot  0.049263
Most outstanding attributes are :
7          bus_station_usr -0.025723
54  tourist_attraction_tot -0.028806
40         art_gallery_tot -0.029632
16              school_usr -0.031984
30             library_avr -0.037531
9                store_usr -0.040680
31       movie_theater_avr -0.043512
24                bank_avr -0.054182
8                  car_usr -0.080065
39            aquarium_tot -0.189368
```

Class 6

```
Key characteristics of 5 cluster are:
Most outstanding attributes are :
8         car_usr  0.296657
30    library_avr  0.137947
21   aquarium_avr  0.060760
28    medical_avr  0.045947
52     school_tot  0.037480
45      store_tot  0.035222
16     school_usr  0.031954
56        zoo_tot  0.024552
10    medical_usr  0.023658
6        bank_usr  0.015763
Most outstanding attributes are :
47                  spa_tot -0.129215
25          bus_station_avr -0.149373
53              stadium_tot -0.152272
17              stadium_usr -0.154138
54   tourist_attraction_tot -0.157278
34               school_avr -0.166455
50               museum_tot -0.184204
36   tourist_attraction_avr -0.191868
55           university_tot -0.192394
23                 café_avr -0.231673
```

Figure 9. Outstanding attributes for six different clusters for price class two.

Attributes which have the highest difference from the average for each cluster out of price class 2 (fig. 6) are shown on Figure 9. The interpretation can as follow:

Class 1(Red)- Many aquariums, well attended stadiums, movie theaters with good ratings, not enough universities, schools and good museums. Class is more on entertainment rather than on cultural aspects.

Class 2 (Cherry) – Good cafes and banks, many spas on opposite – poor parks, fewer schools, museums libraries. Class more of calm life with people preferred calm not very active life, with cafés spas, maybe more elderly people.

Class 3 (Violet) – Many different venues including cultural, commercial, education, but with rating lower than average. Class may reflect presence of different venues and high demand for the quality of those or low quality of the venues

Class 4 (Blue) - class with higher than average venues ratings but severely fewer number of venues. This seems to reflect the fact that there are not enough venues in the areas but those which are available deliver good quality products.

Class 5 (Gray) - Class with high number of universities, stadiums, schools on the opposite the lack of stores, movie theatres, car services and some unsatisfactory with movie theater, banks and libraries, but the absolute value of difference from the average is quite low. The Class is with the focus on number universities, banks and some others while the negative attributes are less pronounced

Class 6 (Yellow) - Class lacks touristic attraction, good cafes, universities but provides lots of car services reviews, surprisingly good library and medical. This can be interpreted as more industrial, heavy urban areas.

Same interpretation can be done for 3x6 other clusters. This part is not provided here due to time/effort limitations.

Discussion

This work reflects high level implementation of machine learning clustering approach to assist the selection of the new area to live to narrow down the search. Accuracy of the model can be potentially increased by incorporating more data like average car traffic maps, pollution/air quality, crime/fire statistics.

The other area for improvement might be to understanding better the available or new potential data and of the impact of different pre-processing/normalization techniques.

The analysis provided in this report is based on K-means clustering technique. Potentially more accurate clustering algorithms like DBSCAN.

Conclusion

I'm personally from price class two and currently consider changing place of living. Based on the report outcome I plan to focus on cluster 2 or cluster 5 from the same pricing class.