# MOSCOW AREAS CLASSIFICATION
## FOR THE PURPOSE OF APARTMENTS PURCHASE

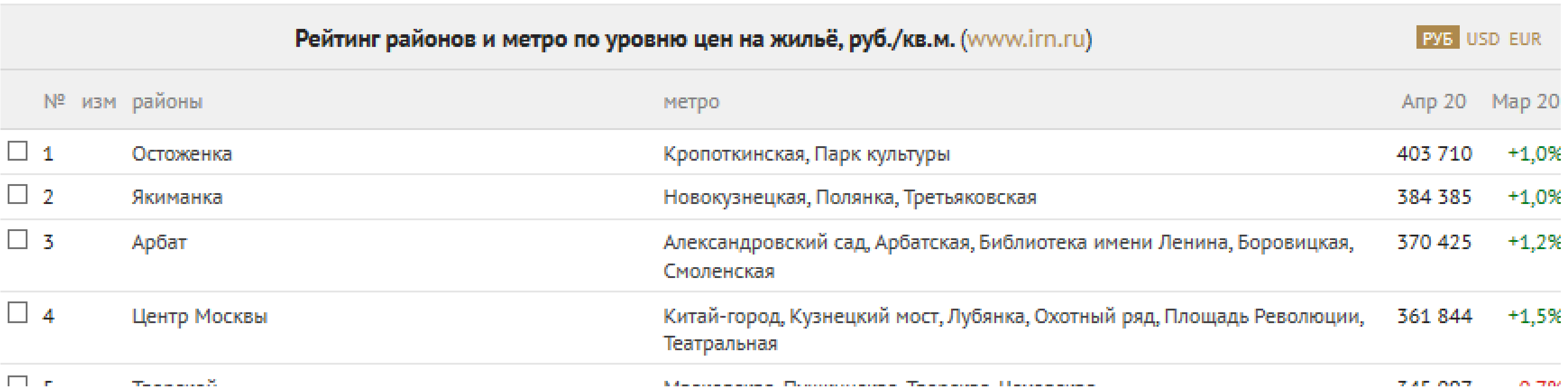
Oleg Adamovich
May 20, 2020

# INTRODUCTION

- Moscow is a biggest city in Russia, officially with 8+ million people providing tons of information
- Data has a high variety  like Ratings and Reviews, venue types
- The data has a potential to cluster the areas to select neighborhoods which have similar features
- Clustering can assist ones decision as in selection as well in narrowing down the area of search

https://www.irn.ru/rating/moscow/

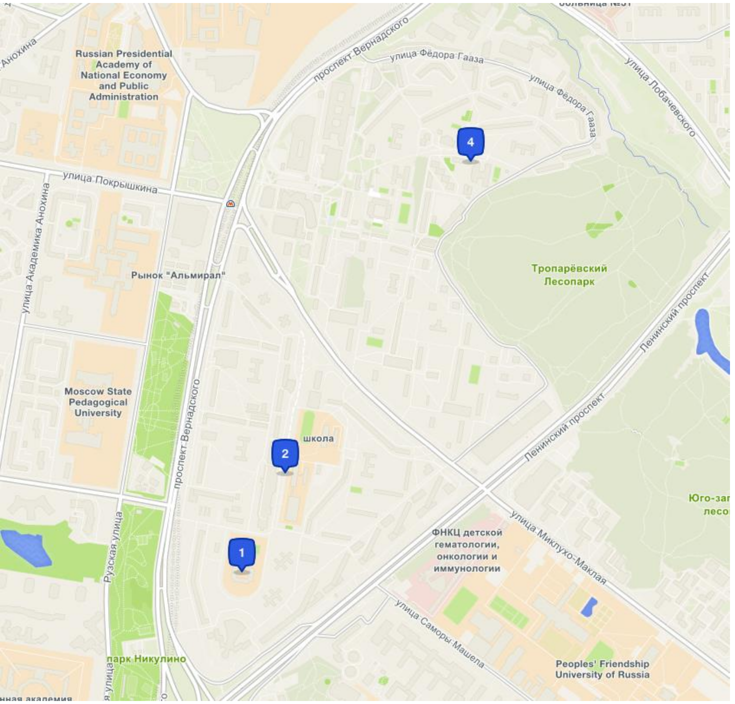
Metro names vs cost per sq. m information

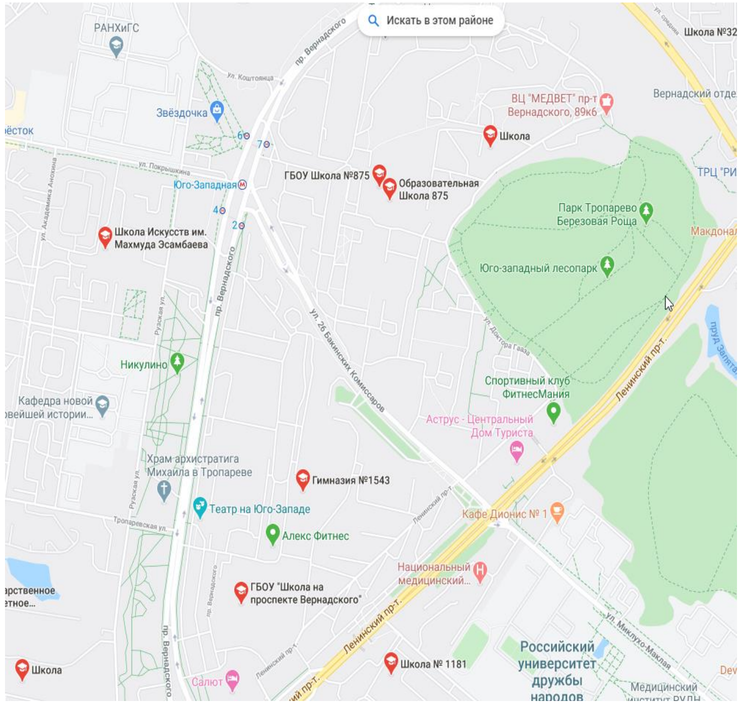


**Рейтинг районов и метро по уровню цен на жильё, руб./кв.м.** (www.irn.ru) РУБ USD EUR

| № | изм | районы | метро | Апр 20 | Мар 20 |
|---|---|---|---|---|---|
| 1 | | Остоженка | Кропоткинская, Парк культуры | 403 710 | +1,0% |
| 2 | | Якиманка | Новокузнецкая, Полянка, Третьяковская | 384 385 | +1,0% |
| 3 | | Арбат | Александровский сад, Арбатская, Библиотека имени Ленина, Боровицкая, Смоленская | 370 425 | +1,2% |
| 4 | | Центр Москвы | Китай-город, Кузнецкий мост, Лубянка, Охотный ряд, Площадь Революции, Театральная | 361 844 | +1,5% |

Data example from different providers




ForthSquare – 3 Schools

Google Maps – 7 Schools

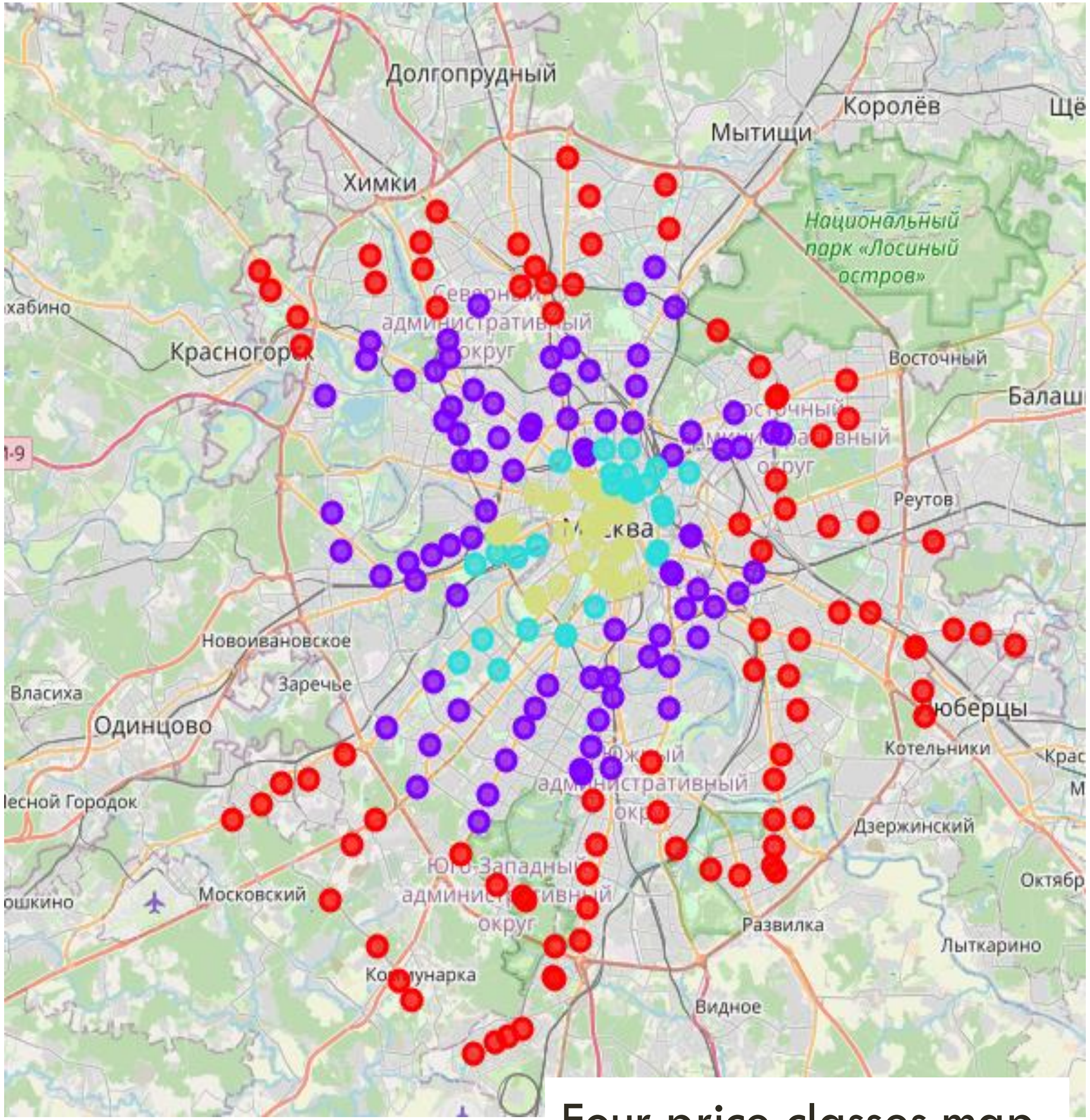
**Google Maps** have more data available – Source for the data

https://api.hh.ru/metro/1

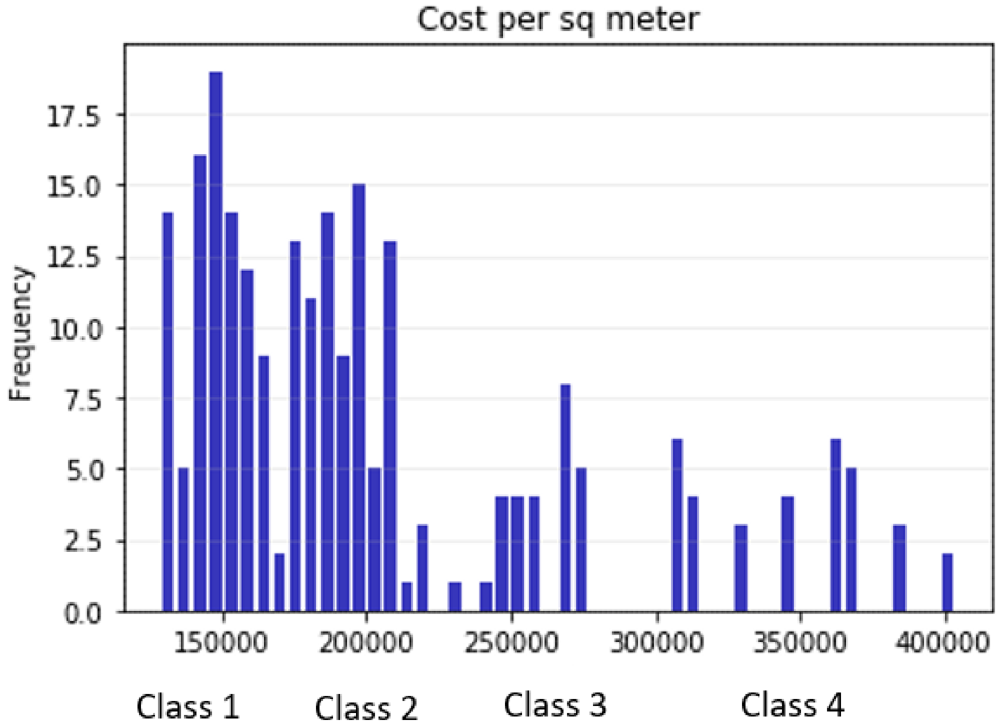

{"id":"1","name":"Москва","lines":[{"id":"8","hex_color":"FFCD1C","name":"Калининская"
{"id":"8.107","name":"Перово","lat":55.75098,"lng":37.78422,"order":2},{"id":"8.158","
Ильича","lat":55.747115,"lng":37.680726,"order":5},{"id":"8.78","name":"Марксистская",
[{"id":"2.558","name":"Ховрино","lat":55.8777,"lng":37.4877,"order":0},{"id":"2.674","
стадион","lat":55.838978,"lng":37.487515,"order":3},{"id":"2.30","name":"Войковская","


Metro names vs coordinates

Metro station coordinates – central points to get the data from Google Maps API




Four price classes map




Metro station can be divided into 4 different classes based on cost

| _usr | park_usr | school_usr | stadium_usr | tourist_attraction_usr | university_usr | zoo_usr | amusement_park_avr | aquarium_avr | art_gallery_avr | café_avr | bank_avr |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 4355 | 325 | 745 | 46 | 80 | 71 | NaN | NaN | 4 | 4.34438 | 3.67742 |
| 241 | 8501 | 362 | NaN | 10052 | 63 | 203 | NaN | NaN | 5 | 4.19598 | 3.16828 |
| 547 | 5028 | 330 | 225 | 11 | 39 | NaN | NaN | NaN | NaN | 4.33205 | 3.2375 |
| 122 | 11789 | 145 | NaN | 27 | 40 | NaN | NaN | NaN | NaN | 4.37893 | 3.96078 |
| 690 | 740 | 159 | 744 | 174 | 416 | NaN | NaN | 4 | NaN | 4.29594 | 3.32552 |
| 292 | 13696 | 2104 | 179 | 5033 | 415 | NaN | NaN | 1 | 4.53878 | 4.39695 | 3.02553 |
| 317 | 9934 | 2277 | 4 | 14098 | 459 | NaN | NaN | 2.5 | 4.54591 | 4.40789 | 2.7409 |
| 638 | 107776 | 451 | NaN | 190257 | 262 | NaN | NaN | 3.5 | 4.735 | 4.43518 | 3.40393 |
| 15 | 5321 | 246 | NaN | 46 | 56 | NaN | NaN | 4.8 | NaN | 4.39742 | 3.46538 |
| 358 | 23654 | 220 | 1 | 147 | 19 | NaN | NaN | 4.8 | NaN | 4.38683 | 2.88772 |
| 388 | 20101 | 288 | 23 | 116 | 67 | NaN | NaN | 4.8 | NaN | 4.39474 | 2.95676 |
| 27 | 17981 | 158 | 11 | 116 | 143 | NaN | NaN | NaN | NaN | 4.39526 | 2.97304 |
| 960 | 10403 | 304 | 0 | 1966 | 1334 | NaN | NaN | NaN | 4.84 | 4.37147 | 2.90905 |
| 330 | 8393 | 469 | 5611 | 1902 | 1354 | NaN | NaN | 4.7 | 4.6 | 4.38894 | 3.23805 |
| 707 | 15445 | 443 | 5611 | 2071 | 972 | NaN | NaN | 4.7 | 4.59452 | 4.3772 | 3.39266 |
| 480 | 7076 | 348 | 6884 | 339 | 249 | NaN | NaN | 4.7 | 4.76818 | 4.36598 | 2.53163 |
| 328 | 33825 | 323 | 247 | 8517 | 480 | 7628 | NaN | NaN | 4.54935 | 4.43442 | 2.97943 |
| 339 | 37540 | 534 | 569 | 22401 | 1041 | 7628 | NaN | NaN | 4.58591 | 4.41554 | 3.21917 |
| 383 | 55138 | 585 | 569 | 159514 | 1089 | 7642 | NaN | 3.5 | 4.59072 | 4.43533 | 3.38034 |
| 235 | 115057 | 591 | NaN | 191807 | 1114 | 17 | NaN | 4.19818 | 4.69392 | 4.42229 | 3.41232 |

| r | park_usr | school_usr | stadium_usr | tourist_attraction_usr | university_usr | zoo_usr | aquarium_avr | art_gallery_avr | café_avr | bank_avr | bus_station_avr | car_av |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 0.031480 | 0.142732 | 0.051450 | 0.000237 | 0.045767 | 0.009291 | 0.759340 | 0.714286 | 0.512160 | 0.605429 | 0.686096 | 0.6770 |
| 1 | 0.061449 | 0.158981 | 0.000000 | 0.051733 | 0.036041 | 0.026564 | 0.759340 | 1.000000 | 0.011969 | 0.391888 | 0.557257 | 0.5826 |
| 2 | 0.036345 | 0.144928 | 0.015539 | 0.000057 | 0.022311 | 0.000000 | 0.759340 | 0.877419 | 0.470615 | 0.420920 | 0.638718 | 0.7145 |
| 4 | 0.085216 | 0.063680 | 0.000000 | 0.000139 | 0.022883 | 0.000000 | 0.759340 | 0.877419 | 0.628630 | 0.724277 | 0.821144 | 0.7040 |
| 8 | 0.005349 | 0.069829 | 0.051381 | 0.000895 | 0.237986 | 0.000000 | 0.789474 | 0.877419 | 0.348911 | 0.457836 | 0.824076 | 0.5137 |
| 4 | 0.099001 | 0.924023 | 0.012362 | 0.025902 | 0.237414 | 0.000000 | 0.000000 | 0.868222 | 0.689385 | 0.332017 | 0.768366 | 0.6559 |
| 1 | 0.071808 | 1.000000 | 0.000276 | 0.072556 | 0.262586 | 0.000000 | 0.394737 | 0.870261 | 0.726242 | 0.212637 | 0.723324 | 0.6688 |
| 8 | 0.779055 | 0.198068 | 0.000000 | 0.979162 | 0.149886 | 0.000000 | 0.657895 | 0.924285 | 0.818221 | 0.490725 | 0.731820 | 0.7235 |
| 2 | 0.038463 | 0.108037 | 0.000000 | 0.000237 | 0.032037 | 0.000000 | 1.000000 | 0.877419 | 0.690946 | 0.516499 | 0.308450 | 0.6498 |
| 7 | 0.170982 | 0.096618 | 0.000069 | 0.000757 | 0.010870 | 0.000000 | 1.000000 | 0.877419 | 0.655255 | 0.274217 | 0.398600 | 0.5911 |
| 1 | 0.145299 | 0.126482 | 0.001588 | 0.000597 | 0.038330 | 0.000000 | 1.000000 | 0.877419 | 0.681936 | 0.303172 | 0.742853 | 0.7488 |
| 3 | 0.129975 | 0.069390 | 0.000760 | 0.000597 | 0.081808 | 0.000000 | 0.759340 | 0.877419 | 0.683663 | 0.310001 | 0.696620 | 0.7306 |
| 6 | 0.075198 | 0.133509 | 0.000000 | 0.010118 | 0.763158 | 0.000000 | 0.759340 | 0.954286 | 0.603470 | 0.283161 | 0.618831 | 0.7089 |
| 8 | 0.060668 | 0.205973 | 0.387500 | 0.009789 | 0.774600 | 0.000000 | 0.973684 | 0.885714 | 0.662378 | 0.421152 | 0.670206 | 0.7124 |
| 1 | 0.111644 | 0.194554 | 0.387500 | 0.010658 | 0.556064 | 0.000000 | 0.973684 | 0.884149 | 0.622794 | 0.485997 | 0.677912 | 0.4465 |
| 4 | 0.051149 | 0.152833 | 0.475414 | 0.001745 | 0.142449 | 0.000000 | 0.973684 | 0.933766 | 0.584979 | 0.124865 | 0.809415 | 0.7823 |
| 5 | 0.244503 | 0.141853 | 0.017058 | 0.043833 | 0.274600 | 0.998168 | 0.759340 | 0.871242 | 0.815662 | 0.312683 | 0.872689 | 0.5335 |
| 4 | 0.271356 | 0.234519 | 0.039296 | 0.115287 | 0.595538 | 0.998168 | 0.759340 | 0.881688 | 0.752032 | 0.413234 | 0.860372 | 0.6629 |
| 8 | 0.398563 | 0.256917 | 0.039296 | 0.820942 | 0.622998 | 1.000000 | 0.657895 | 0.883063 | 0.818748 | 0.480829 | 0.798900 | 0.1312 |
| 1 | 0.831685 | 0.259552 | 0.000000 | 0.987139 | 0.637300 | 0.002225 | 0.841627 | 0.912548 | 0.774774 | 0.494241 | 0.774143 | 0.0516 |

- Data type requests from Google API

[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, supermarket, tourist_attraction, university, zoo].

Data for analysis:

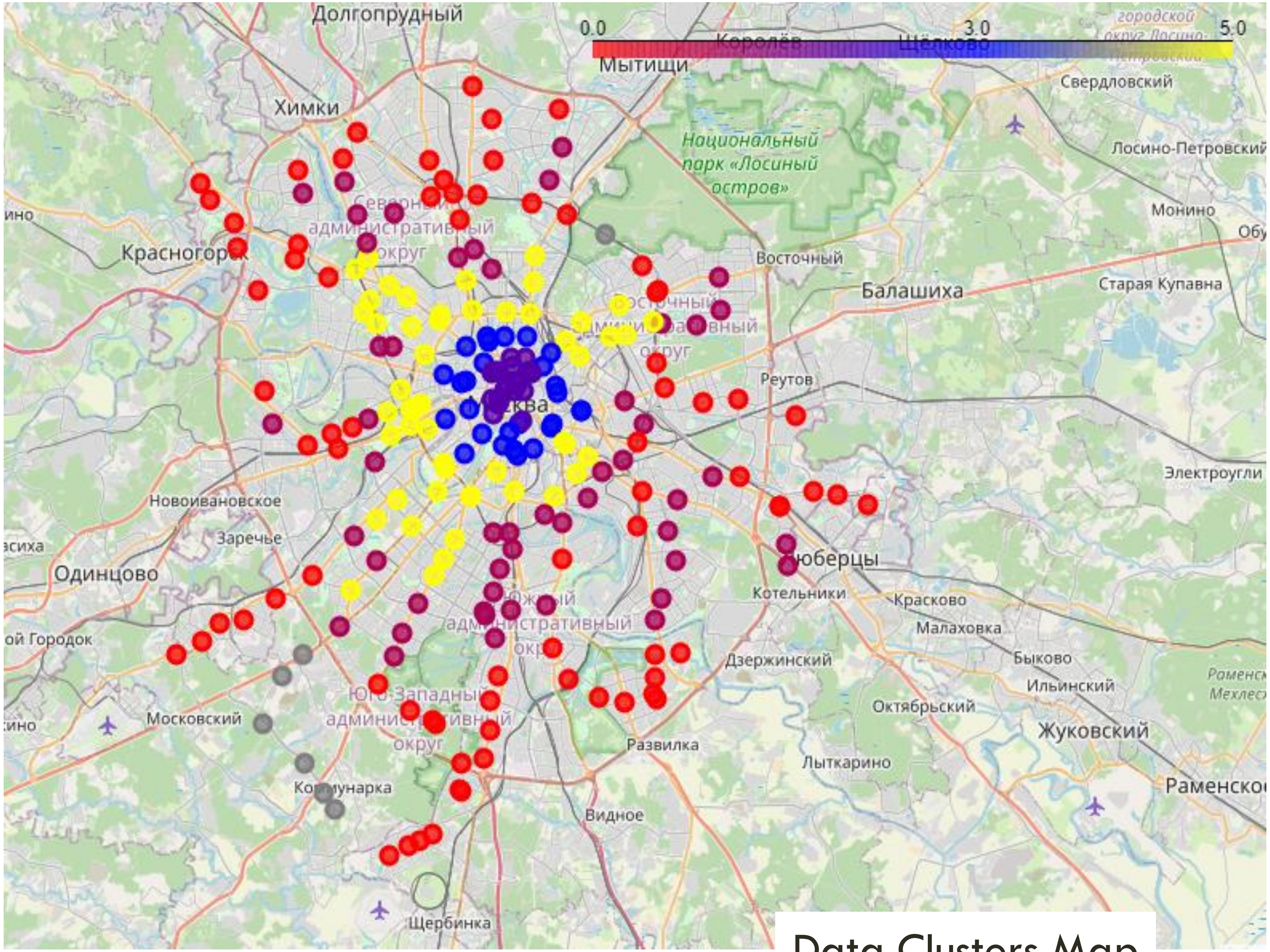
Num of venues of a type per Metro

Num of reviews of venues for a selected type per Metro

Average rating for venue type per Metro

# Data Preparation Workflow:

- Read all the data for Metro Station and merge into single DataFrame. Each Metro has a set of attributes Totally 235 Station and 58 attributes
- Fill in NaN with zeros or mean numbers
- Normalize data




Data Clusters Map

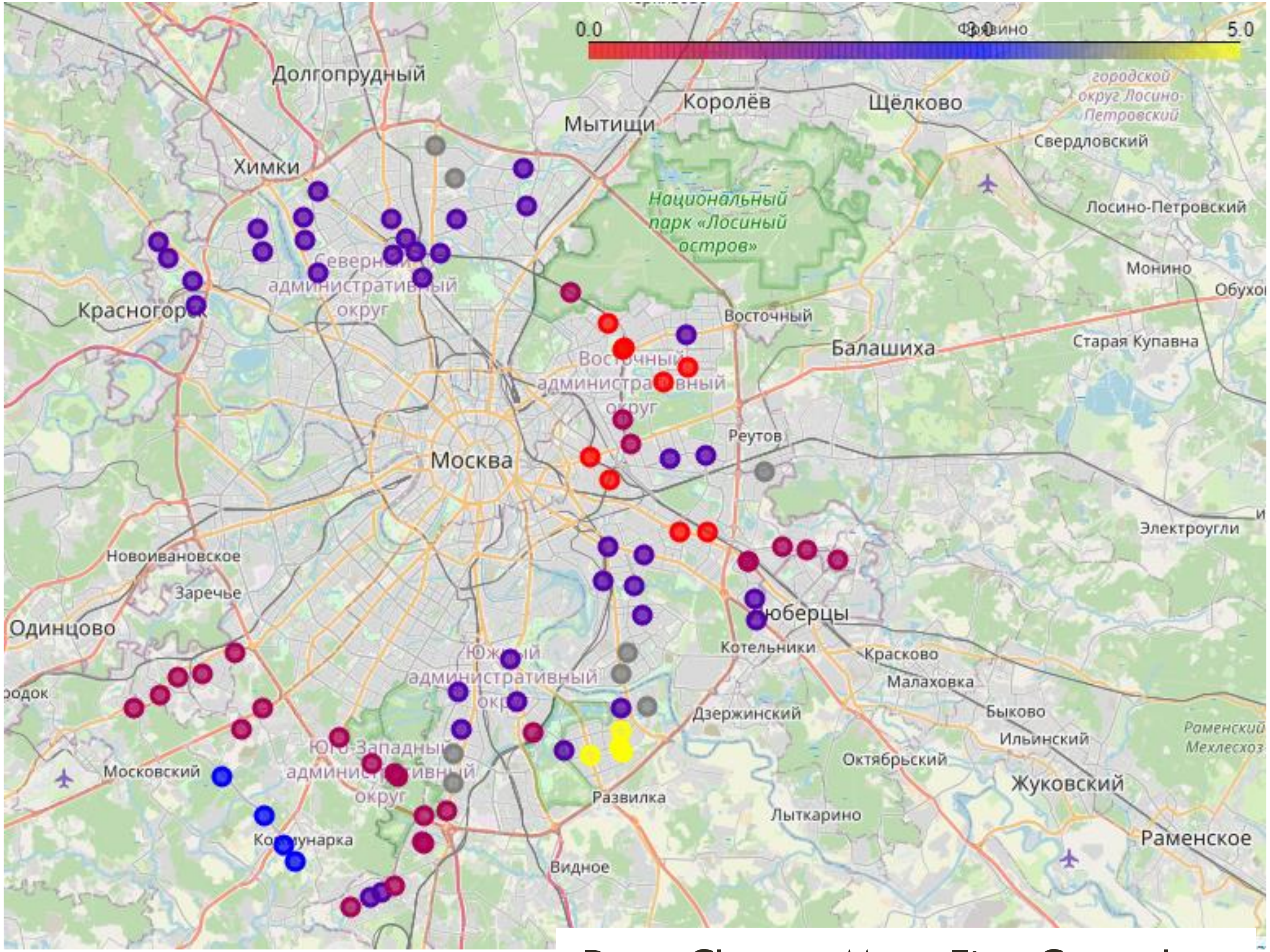
K-means data clusterization based on pre-processed data

Good visual correlation with the cost-based prices (slide 4)

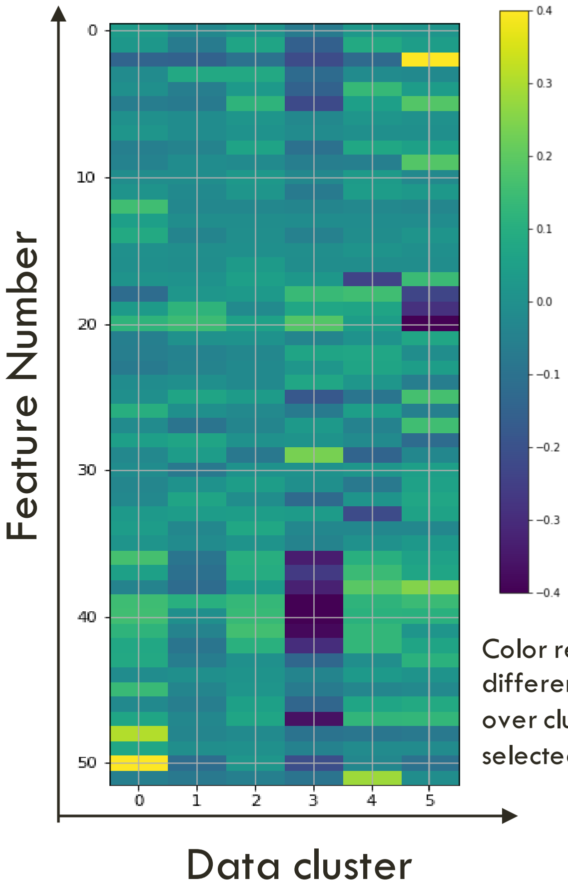
K-means data clusterization was repeated separately for each cost class (Slides 7-10)

K-means data clusterization based on pre-processed data

First cost data class **only**



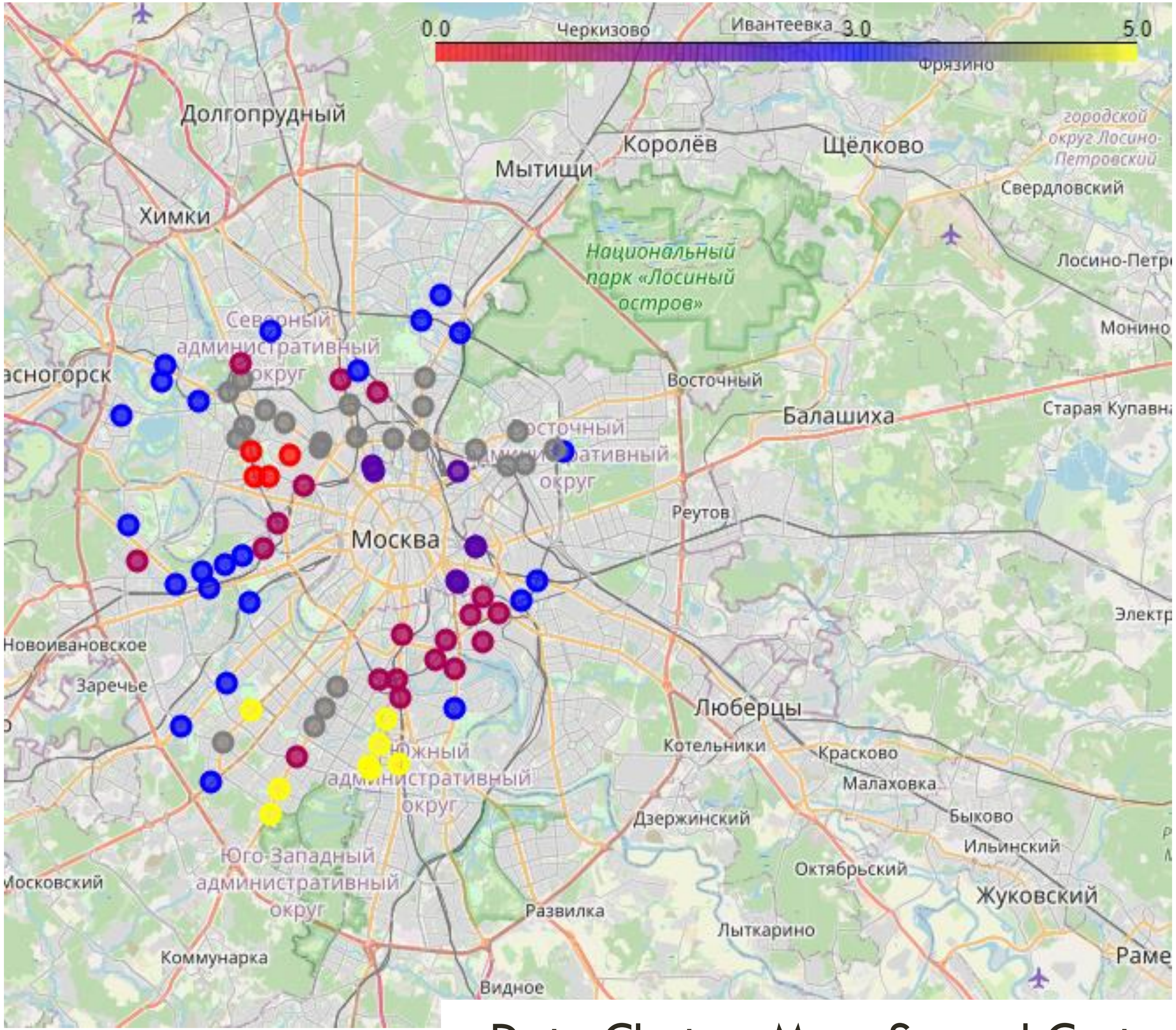

Data Clusters Map. First Costs class



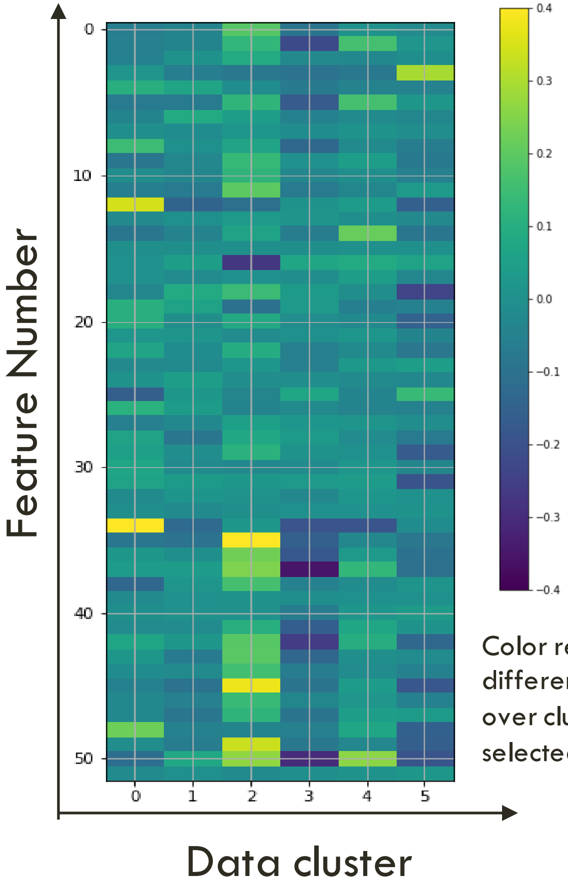

Color represent difference from mean over clusters from selected attributes

K-means data clusterization based on pre-processed data

Second cost data class **only**



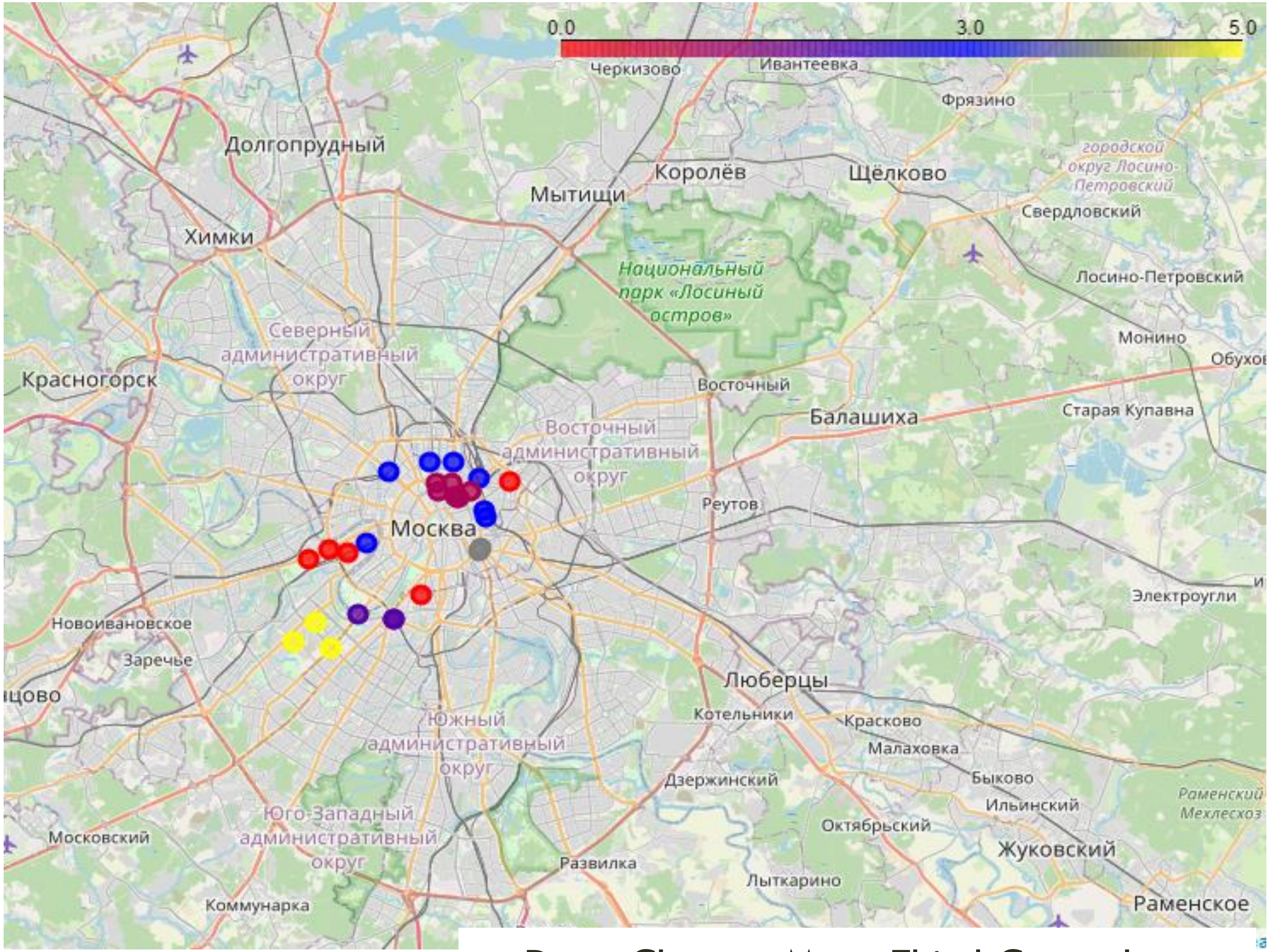

Data Clusters Map. Second Costs class



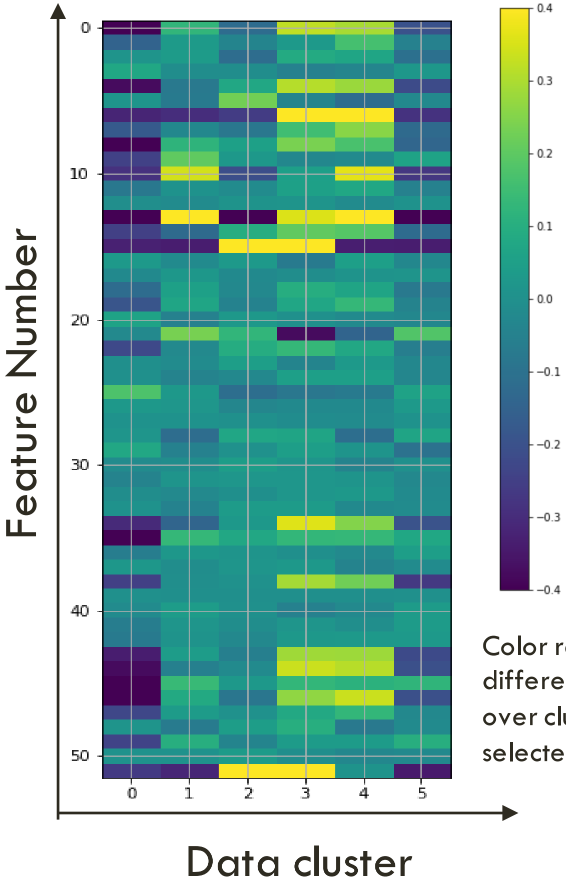

Color represent difference from mean over clusters from selected attributes

K-means data clusterization based on pre-processed data

Third cost data class **only**



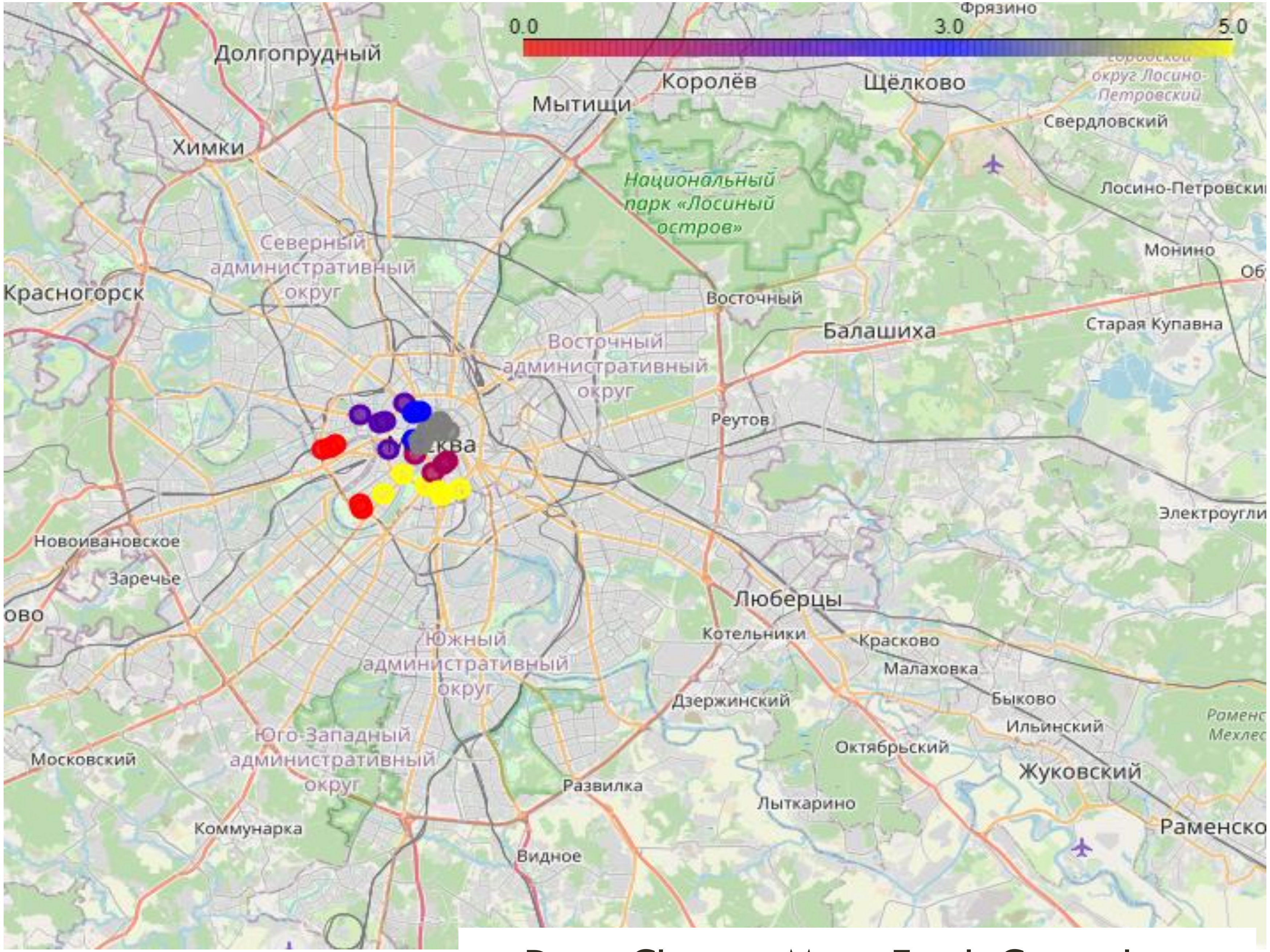

Data Clusters Map. Third Costs class



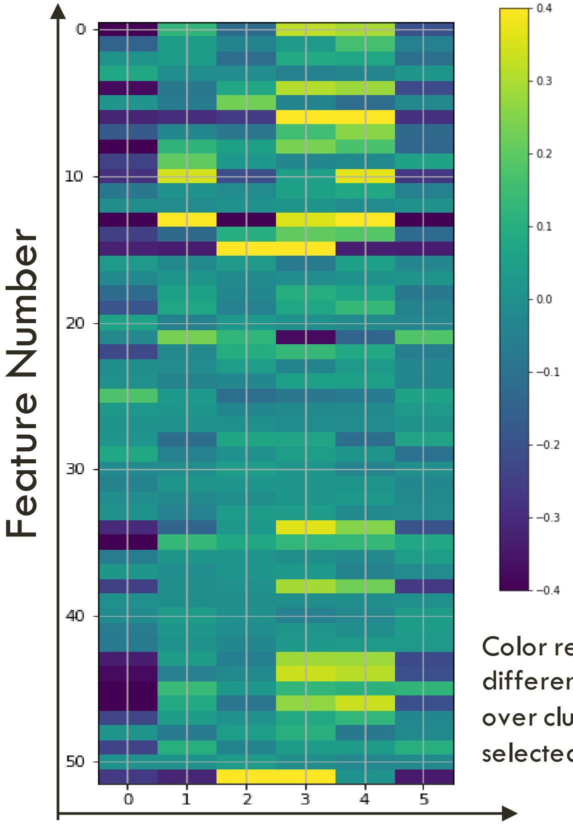

Color represent difference from mean over clusters from selected attributes

K-means data clusterization based on pre-processed data

Forth cost data class **only**




Data Clusters Map. Forth Costs class




Color represent difference from mean over clusters from selected attributes

Cluster 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
```

Cluster 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
```

Cluster 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
```

Cluster 4

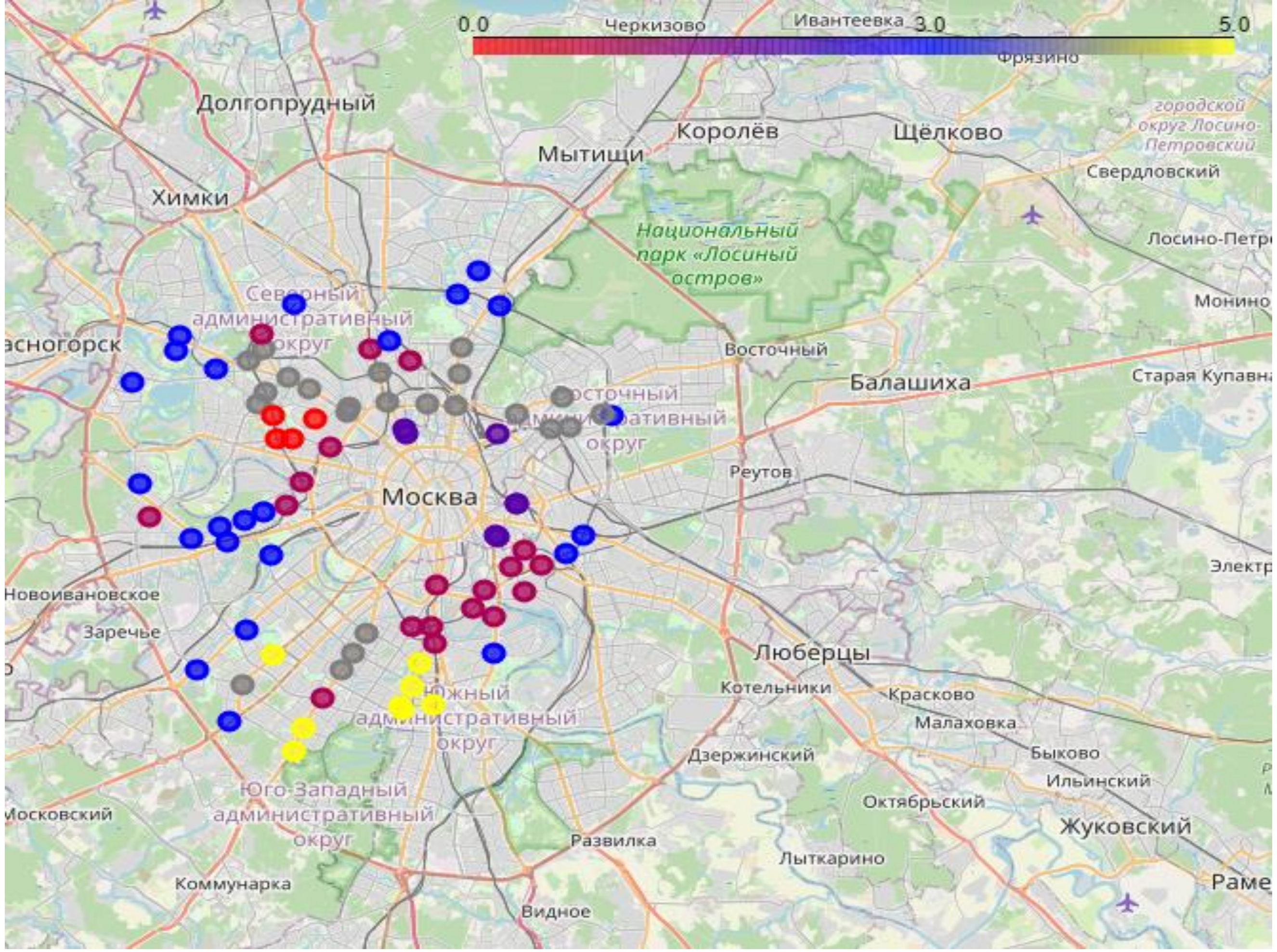
```
Key characteristics of 3 cluster are:
Most outstanding attributes are :
30                library_avr  0.071521
21               aquarium_avr  0.069585
23                   café_avr  0.035571
33                   park_avr  0.035108
36    tourist_attraction_avr  0.033670
24                   bank_avr  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
```

Cluster 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
```

Cluster 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
```




## Clustering interpretation example (Class2)

**<u>Cluster 1</u>**(Red)- More on entertainment rather than on cultural aspects.

**<u>Cluster 2</u>** (Cherry) –More of calm life with people preferred calm not very active life, with cafés spas, maybe more elderly people.

**<u>Cluster 3</u>** (Violet) –Class may reflect presence of different venues and high demand for the quality of those or low quality of the venues

**<u>Cluster 4</u>** (Blue) - Not enough venues in the areas but those which are available deliver good quality products.

**<u>Cluster 5</u>** (Gray) - The Class is with the focus on number universities, banks and some others while the negative attributes are less pronounced

**<u>Cluster 6</u>** (Yellow) - more industrial, heavy urban areas.

# CONCLUSION AND FUTURE DIRECTIONS

- Approach has a potential to predict whether and how much a player will improve
- 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 of existing model and data might be to understanding better the available 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.