Capstone Project - The Battle of the Neighborhoods (Week 2)

Foursquare Users Segmentation

Applied Data Science Capstone by IBM/Coursera

Introduction: Business Problem

Let's assume we are going to establish new retail business venue in a city, but we still don't know what it should be. To make our choice we fist need to answer several questions, that can make our intend more specific:

- What type of venue would be popular in this particular city ?
- Who would be our clients ? What they prefer ? What are their share of the total customers ?

Definitely it's not all questions one needs to set up our own business. Still it is a good start for creating business plan based on data, instead of assumptions.

Data

One of the possible approaches of solving the problem we described is to carry out customer segmentation analysis. To do that we are going to use Foursquare API (venues location and user's profile) in following steps:

1. Find one of the most popular venues in the city and its users.
2. Find friend of these users, who live in the same city. Find friends of their friends and so on until we get all Foursquare users of the city.
3. Get "tastes" for each user, based on his Foursquare profile.

For steps 1-2 we will use standard Foursquare API (explore and friends Endpoints). Step №2 continues until the number of obtained user ids after iteration does not change.

For getting data needed at step №3 using standard Foursquare API is not enough. Foursquare describes tastes Endpoint in its the documentation, however its available only for part of authorized users (the user himself and maybe his friends). Any other person will get en authorization error in response. We could use list Endpoint to get "venueslikes" (a list of venues, which was likes be the user) but via "personal" account it allows you to get only 2 venues per users list, which is not enough to build the relevant tastes profile for clustering. On the other hand all lists including venueslikes are available at web page of each users profile, so we can get them directly from web pages.

I'd like to demonstrate the whole process on my native city Rostov-on-Don, Russia because It has just few thousands of active Foursquare's users, so obtaining the data would be much faster comparing to big megapolis. In spite of that, all of the code is suitable for any other place in the world.

Methodology

When the data is obtained we can achieve user's segmentation by following steps:

1. Transform the list of liked venues for each user into their taste profile.
2. Use clustering method to divide users into separate segments based on their taste profile.
3. Analyze the segments to discover differences between them.

For clustering my choice is using the simplest method of clustering - KMeans, because:

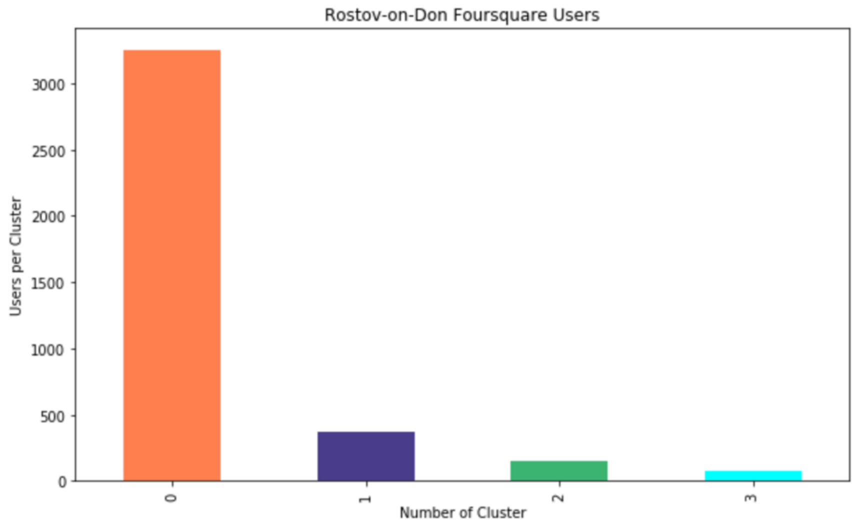
1. It is fast
2. There is no arbitrary shaped clusters for using DBSCAN
3. At this moment we don't need hierarchical structure, as we don't know how to interpret it

The only problem we have is choosing the number of clusters.
My decision is to get maximum users diversity, keeping the number users within the smallest cluster commensurate with other clusters. If we use k=4 we get 81 user in smallest cluster, next we have clusters with 153 and 374 users which is about twice bigger.

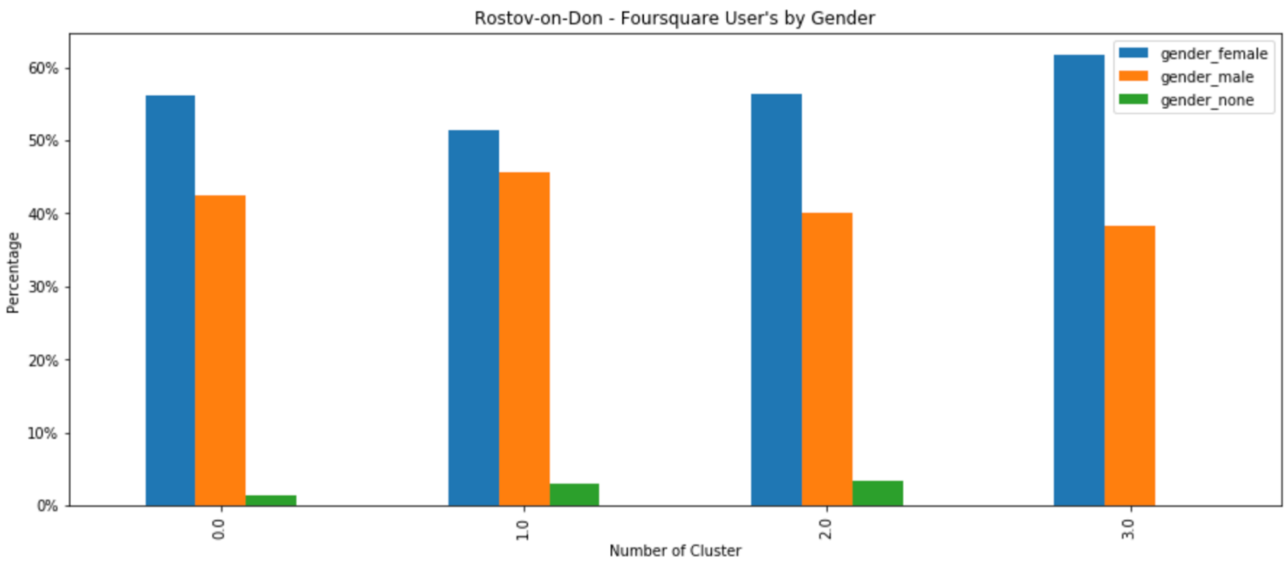
Analysis

Let's build several tables and plots that help us to understand the difference between clusters and and some more details about them.

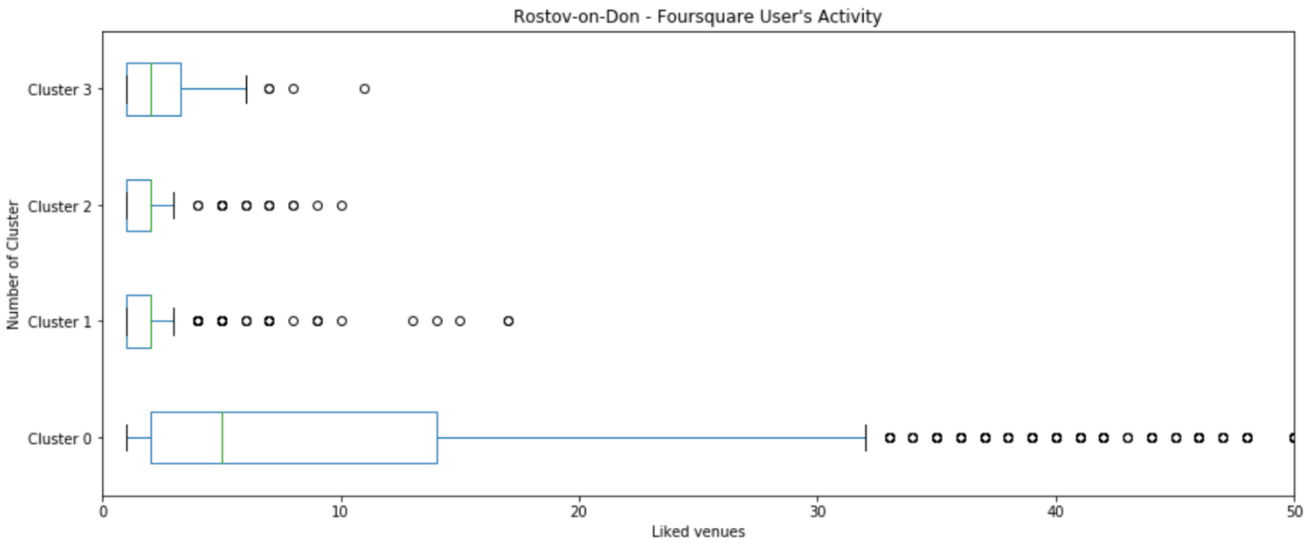
	Cluster	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category	8th Most Common Category	9th Most Common Category	10th Most Common Category	Users In Cluster
0	0	Shopping Mall	Home (private)	Restaurant	Café	Coffee Shop	Salon \/ Barbershop	Italian Restaurant	Asian Restaurant	Hotel	Multiplex	3252.0
1	1	Home (private)	Asian Restaurant	Shopping Mall	Café	Salon \/ Barbershop	Nightclub	Office	Plaza	Sushi Restaurant	Italian Restaurant	368.0
2	2	Restaurant	Home (private)	New American Restaurant	Gym \/ Fitness Center	Hotel	Fast Food Restaurant	Office	Gastropub	Coffee Shop	Asian Restaurant	153.0
3	3	Coffee Shop	Home (private)	Office	Shopping Mall	Italian Restaurant	Café	Gym \/ Fitness Center	Sushi Restaurant	Pelmeni House	Hotel	81.0



As the result of clustering we have 4 segments where the first one is bigger than 3 others in total. The other 3 reduced by half with each next cluster.



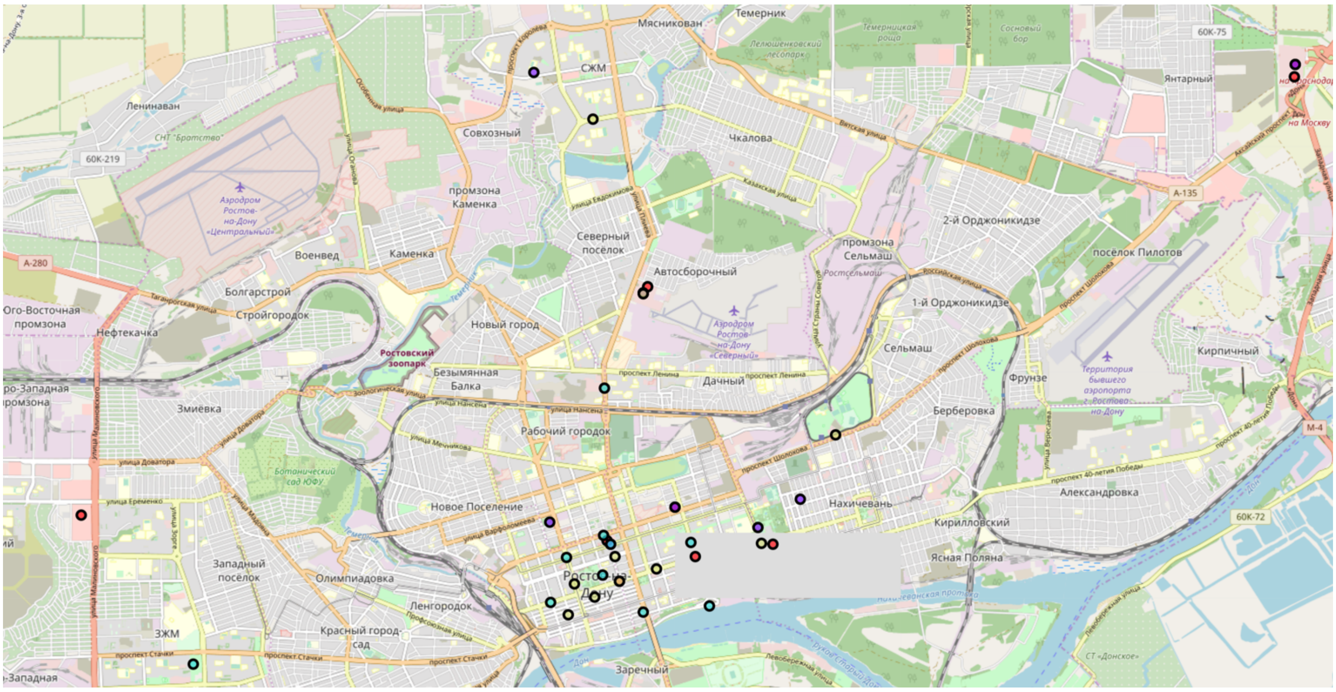
Speaking about gender distribution: In all clusters women have bigger share then men. That means the share is not equal for the whole dataset. (Perhaps - it is the feature of local users) But in cluster №3 the distinguish is maximum (61% - women and only 38% - men)



Comparing the activity of user in different clusters: Cluster №0 not only the biggest, but also most active (median is 5 likes per user, maximum is 500).

	id	name	categories	price.message	price.tier	location.lat	location.lng	Cluster
0	4c3afe4e1a1cd13a4691b50d	Мегацентр «Горизонт»	Shopping Mall	NaN	NaN	47.259950	39.720364	0
1	4ef5f1a5f790731250ba0833	Киноцентр «Большой»	Multiplex	NaN	NaN	47.231582	39.726621	0
2	4c3c51bc7d00d13af0ae3850	МЕГА Ростов-на-Дону / MEGA Mall	Shopping Mall	NaN	NaN	47.290027	39.847268	0
3	4c1a5e978b3aa593187f955f	Буковски	Gastropub	Moderate	2.0	47.227175	39.713399	0
4	5006b1e1e4b0a69557c73b79	Сметана	Pelmeni House	Moderate	2.0	47.221732	39.715690	0
5	4c45c614f0bdd13ac264cbcc	Киномакс-Дон	Multiplex	NaN	NaN	47.260676	39.721266	0
6	4c7a46e4bd346dcb9d04f8ef	IKEA	Furniture / Home Store	NaN	NaN	47.288429	39.847004	0
7	509b9cbc498e0d95ea0e2ff3	New York	New American Restaurant	Expensive	3.0	47.225025	39.730500	0
8	4c430e2bff711b8d475a1405	Золотой Вавилон	Shopping Mall	NaN	NaN	47.230570	39.611110	0
9	4cbbd6a5c7228cfa115820ce	Театральная площадь	Plaza	NaN	NaN	47.226694	39.745553	0
10	4c3c51bc7d00d13af0ae3850	МЕГА Ростов-на-Дону / MEGA Mall	Shopping Mall	NaN	NaN	47.290027	39.847268	1
11	4ffec96ae4b00889b08dc830	Фарш	New American Restaurant	Expensive	3.0	47.228931	39.742613	1
12	4f9dced1e4b008dde2ca2225	ул. Островского, 97	Home (private)	NaN	NaN	47.229618	39.702262	1
13	50bf5b64e4b0a749d220fdcf	Мой Дом	Home (private)	NaN	NaN	47.400802	40.099246	1
14	4c3afe4e1a1cd13a4691b50d	Мегацентр «Горизонт»	Shopping Mall	NaN	NaN	47.259950	39.720364	1
15	4ee247f05c5cfd2133a7baec	Публика	Restaurant	Moderate	2.0	47.226723	39.714018	1
16	4ef5f1a5f790731250ba0833	Киноцентр «Большой»	Multiplex	NaN	NaN	47.231582	39.726621	1
17	4fd8fb00e4b01007e98c85f8	Блатхата	Home (private)	NaN	NaN	47.785925	40.129192	1
18	4f350802e4b0b80edebe0801	Лицей № 102	School	NaN	NaN	47.289016	39.699101	1
19	503d47ade4b04cbc9220e12d	Rozi's Apartaments	Home (private)	NaN	NaN	47.232605	39.750902	1
20	4e1b34f5b0fb59954d3e820f	Джон Do	Restaurant	Moderate	2.0	47.226862	39.729577	2
21	4df8cf21c65b6739205d8aaf	Park Культуры	Restaurant	Moderate	2.0	47.222574	39.712409	2
22	4c8f78091992a1cd3df8e5fb	Famous	Restaurant	Moderate	2.0	47.224905	39.705415	2
23	4ee247f05c5cfd2133a7baec	Публика	Restaurant	Moderate	2.0	47.226723	39.714018	2
24	4c4c19975609c9b66eaa0992	Пирс	Restaurant	Moderate	2.0	47.218506	39.733246	2
25	4f36be27e4b02a70e086c0c6	Нескучный сад	Restaurant	Moderate	2.0	47.218955	39.702339	2
26	50352ac2e4b07119fc8a915a	St. Tropez	Restaurant	Moderate	2.0	47.210766	39.632790	2
27	4ede250c46907c1b47047812	Fartyk	Restaurant	Moderate	2.0	47.227809	39.712690	2
28	50f50664e4b05f2e6600cf7c	Деликатеси	Restaurant	Moderate	2.0	47.217755	39.720300	2
29	4f97ebe5e4b0974543a21ff7	Ассорти-Рио	Restaurant	Moderate	2.0	47.247362	39.712775	2
30	4ea4144fb80355a9826abfb7	ПитьКофе Ралли	Coffee Shop	Cheap	1.0	47.241076	39.757767	3
31	51f11b93498ecce2d6305b4c	Starbucks	Coffee Shop	Moderate	2.0	47.259797	39.720336	3
32	4c624e57de1b2d7fa60de270	Питькофе Джаз	Coffee Shop	Cheap	1.0	47.226823	39.743391	3
33	4c4c663d42b4d13a65e00980	Coffee Man	Coffee Shop	Moderate	2.0	47.225082	39.714818	3
34	4bf0338ed4f70f470b9a390f	Питькофе Кино	Coffee Shop	Moderate	2.0	47.221392	39.707059	3
35	5006b1e1e4b0a69557c73b79	Сметана	Pelmeni House	Moderate	2.0	47.221732	39.715690	3
36	50c10ad545b08bcfc03fba19	Good Morning Coffee	Coffee Shop	Cheap	1.0	47.217353	39.705711	3
37	53451856498e4e3255fd156d	Starbucks	Coffee Shop	Cheap	1.0	47.223361	39.722987	3
38	50e046b5e4b070c02eadfc6d	Питькофе Социальные сети	Coffee Shop	Moderate	2.0	47.282866	39.710516	3
39	4cc9665a46b437040d1248e1	Питькофе Почта	Coffee Shop	Cheap	1.0	47.219712	39.710969	3

The list of top-rated venues gives us more understanding about the nature of clusters.
As we can see, users in cluster №0 mostly like shopping centers, malls and multiplexes. So their main interests are: shopping, cinemas and vising public places. Also they are the most active and the biggest group. The most part of liked venues for Cluster №1 is outside of the city. That probably means - most of cluster №1 users are live in Rostov Region, and visit the city for work, leisure and shopping. Users in Cluster №2 is fond of high priced restaurants/night clubs and not interested at shopping as other clusters. While Cluster №3 consists of coffee fans visiting cheap/medium priced venues.



We can use geo data to find good location for our own venue.

Results and Discussion

Our project shows an approach of making users segmentation in one particular city. We got all Foursquare users of Rostov-on-Don city in Russia, obtained their lists of liked venues and used it to separate them into 4 clusters. By analyzing this clusters we formed hypotheses about their nature:

- cluster 0 - visitors of shopping centers, malls and multiplexes
- cluster 1 - users who live outside the city
- cluster 2 - mid/high priced restaurants/night clubs visitors
- cluster 3 - users who likes cheap/medium priced coffee shops

In addition to information about size and preferences of clusters we got raw lists of users id's in each cluster. This dataset can be used for targeted advertising for example.

The clustering could be enhanced in future by including additional info to venues profile such as: level of prices, rating, attribute means that venue is abroad and so on. This could provide more clear and useful segmentation.

Also at the data stage we described tricky technique of obtaining premium Foursquare data directly from Foursquare web pages, without using API. Eventually this approach could be used for other social networks too. The drawbacks of this method are: slow speed and possibility to be blocked by social network.

Conclusion

The purpose of this project was to build Foursquare users segmentation on the example of one not very big city in order to give the potential investor some insights about the category of his future business and it's customers. The results we gained could be very useful in solving this problem. For example we can choose the category for our venue and then analyze the most popular venues of this category in different clusters. We can locate our venue near the popular places for cluster corresponding to our customer segment, or use targeted advertising.
I hope this project will give stakeholders the idea of making their current businesses and startups more effective.