Coursera Capstone Project

IBM Data Science Specialization

Comparsion of Prague and Bratislava trough Foursquere data

January 2020

Introduction and problem background

My idea for this project is to compare two cities which I am related to – Prague and Bratislava. Both of them, are capitals of their countries, multicultural spots, administrative base and the most important financial hubs. I think, that would be interesting to compare their town districts with using of Foursquere data.

I will approach to that with data science methodology and machine learning method called k means clustering. This kind of approach will lead toward solutions and answers for questions like:

- Is there any big difference between Prague and Bratislava based on foursquare location data?
- Which city is "richer" on venue's categories according to foursquare data?
- Can we see some similarities between Prague administrative districts and Bratislava cadastral districts?

For Prague I decided to make analysis on 112 cadastral districts and then make aggregation for superior 22 superior administrative districts. For case of Bratislava I stayed with 20 cadastral districts, because their sizes are big enough and aggregate them into superior administrative districts would not bring any relevant observation.

I hope that results of that will bring some unexpected similarities and differences based on Foursquere location data. For sure, this analysis will have some limitations and won't be enough for some strategical decisions but will be good as a starting point for possible deeper analysis in future.

Data

I used points coordinates of Prague and Bratislava cadastral districts, which I already prepared in **QGis software**. Those data I transformed into csv files (also part of this project) and into pandas dataframes.

Those kinds of data are from city's open source database and are publicly available.[1]

For case of **Prague** I worked with **112 cadastral districts**, which in the end were grouped into **22 Administrative District**.

For case of **Bratislava** I worked with **20 cadastral.** There was not any superior grouping.

Need to notice that administrative dividing in Prague is quite complicated for the first understanding.

All of the districts are represented by own definitional points (mainly centroid of the polygon layer), which should be located nearly to most frequented spot.

Next data went from Foursquere, which I got by using developer API and with web scraping method.

I am sure, that it is not good to compere those cities one to one based just on values, so for that reason I had average founded data.

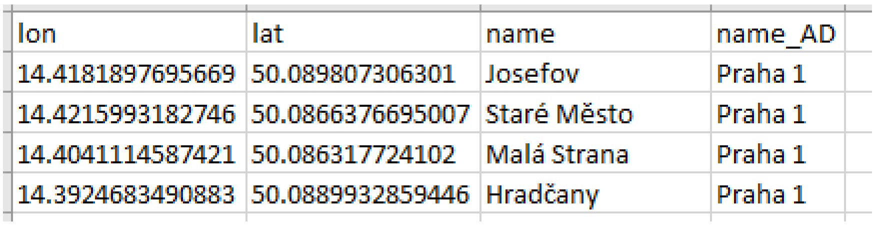
Target audience

- Administrative officials
- Bussiness developers
- Travelers, who planning to go to middle of Europe

[1] https://www.geoportalpraha.cz/en for case of Prague.
https://mapy.bratislava.sk/ for case of Bratislava.

Methodology

At first, I exported just relevant information from polygon layers of Prague and Bratislava cadastral districts. After that I got csv files in format, where I have longitude and latitude columns. “Name” column is a name cadastral districts and “name_AD” is name of superior administrative district. When I had those csv files, I was able to move into jupyter notebook and start to work with Python.



lon	lat	name	name_AD
14.4181897695669	50.089807306301	Josefov	Praha 1
14.4215993182746	50.0866376695007	Staré Město	Praha 1
14.4041114587421	50.086317724102	Malá Strana	Praha 1
14.3924683490883	50.0889932859446	Hradčany	Praha 1

Once I was in jupyter notebook, I imported important libraries for this case and transformed the csv files into pandas datafrmes.

After that, I was able to use custom function with using of Foursquare API which gave me back 100 venues (json format) in defined radius. For that reasons is very necessary to be registered with developer account.

Once I have been done with that, I was working with basic aggregation and pandas codes for case of adjustment and cleaning dataset into form, which was ready for one of last part of the project – clustering of districts based on foursquare venues categories.

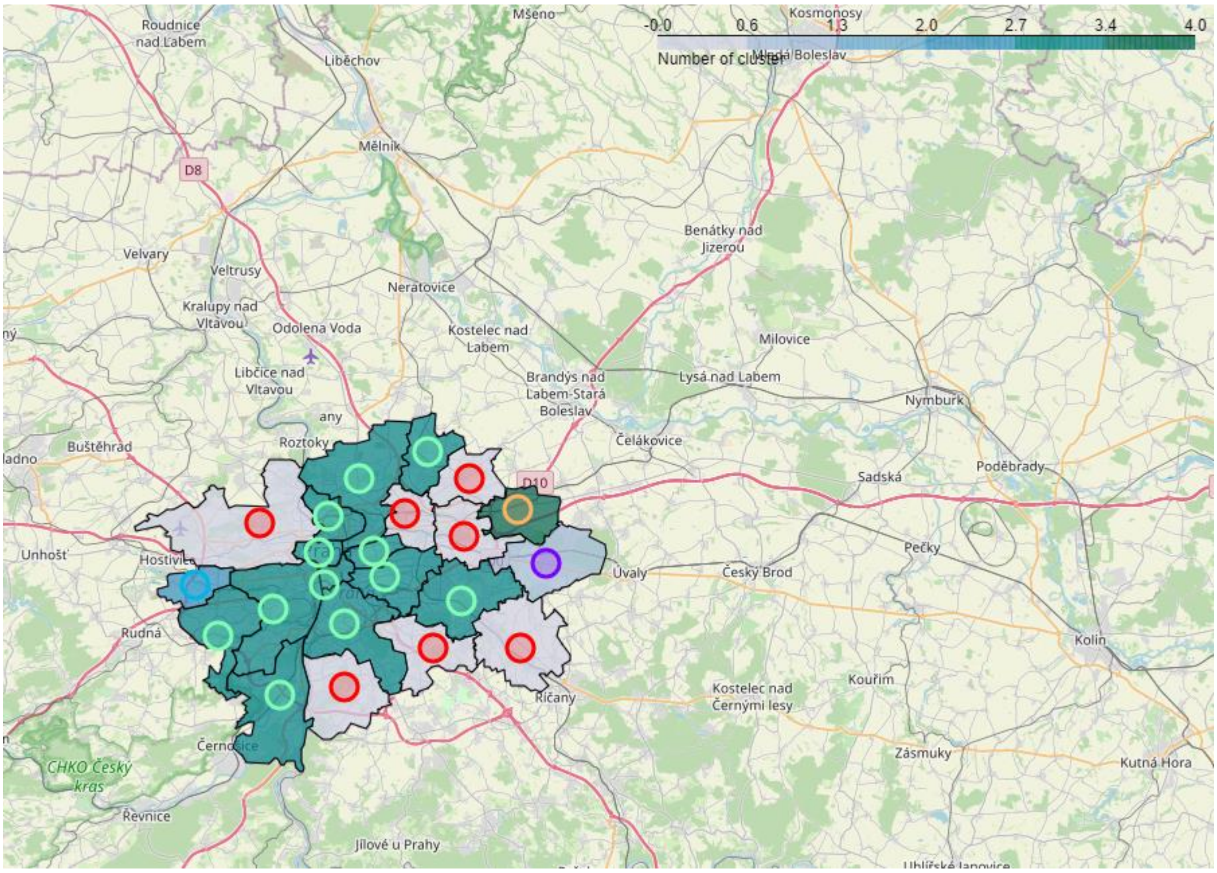
K means clustering is one of popular, unsupervised machine learning method (algorithm) base on grouping of elements with similar attributes. This method identifies k number of centroids and gathers every data point to the nearest centroid. For K means clustering is necessary to define number of clusters into we want to group the results. For case of correct and objective definition of number of clusters is possible to use elbow method.

At the end of project, I visualized the final results in map by using folium library. Also, I used it for that the polygon layers (in geojson format) of administrative and cadastral districts, which I mentioned in Data section.

Results

I have decided to set up 5 clusters for K means clustering and according to that I can comment each cluster.

Prague

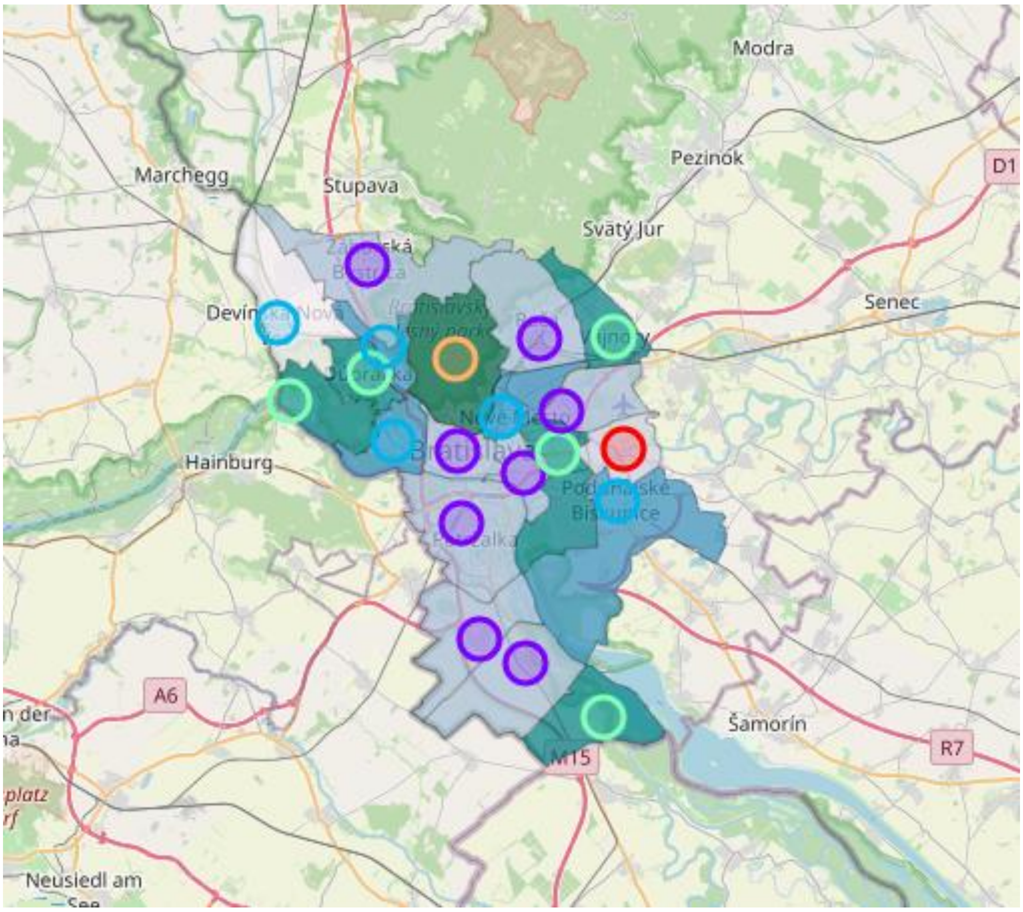


- **Cluster 0 (red circles with light yellow background**) – This cluster counts 7 administrative districts, where 6 of them have at the first place, venue category “restaurant”. Another interesting thing is that at fourth place 4 of districts has venue category “park”. In conclusion, we can define that cluster as district of restaurants in favor.
- **Cluster 1 (purple circle with light (green – yellow) background on the right) –** Just one district is member of that cluster, where the most common venue is soccer field. Also, it is suburban Prague district.
- **Cluster 2 (blue circle with light green background on the left) –** Also just one district is member of that cluster. The most common place is pharmacy. Other venues which can mean differentiation from another districts are reservoir and auto dealership on fifth and sixth place. Also, some food venues (steakhouse in particular) are up to the fourth place.
- **Cluster 3 (green circles with green backgrounds) –** This cluster is the most numerous. There are 12 administrative districts, where most common venue is mostly Café. Very common are Pubs and Bars as well.

- **Cluster 4 (orange circle with dark green background) –** In that cluster is also only one member, where the first six common venues are food venues such as some kind of restaurants or bakery. The train station is in favor there as well.

Bratislava

It is important notice to say, that in Bratislava is very little records on foursquare in comparing to Prague, so this should be taking in account as well. It probably the biggest difference between the cities.



- **Cluster 0 (red circle on light yellow background) –** Just one district is a member of that cluster. Also, that district it is mainly suburban one, that’s maybe why the first common place is garden. Also interesting is that the second one is yoga studio.
- **Cluster 1 (purple circles on light green background) -** This is the most numerous cluster which has 8 districts. Many districts have any kind of restaurants at the first place and for two districts there are Yoga studios at the first place.
- **Cluster 2 (blue circle on green – blue background) -** Paradoxically for all of places is visible Café venue at the first places of table. This cluster has 5 members.
- **Cluster 3 (light green circle with dark green background) –** For this cluster are typical restaurants of all kinds. As well we can see at the third place for Devin cadastral district the iconic monument of whole Bratislava – the Devin castle.
- **Cluster 4 – (orange circle on dark green background) –** There is just one district which is completely different from other. At the first fourth places we can see venues which

are not typical for any other district at all. Campground, Buffet, or Ski Chairlift are the categories which are evidences of this difference.

Discussion of results

The observation of that research and project are enough to answer the research questions. At first the results show big difference in amounts of categories between each city. In case of Prague there are 285 unique categories, but in case of Bratislava, there are just 77 unique categories. So, the answer for the question which country is richer for venues is Prague.

In my opinion, there are visible some similarities between administrative districts in Prague and cadastral districts in Bratislava. But thing which is surprising is a relatively big amount of yoga category in case of Bratislava. I really do not know the reason. I would say that there are not such big differences in case of representation world restaurants between cities too. In both cities we can see large amounts of restaurants from many world countries. In case of Bratislava is visible the border impact of Hungary and Czechia, so many of restaurants are Hungarian or Czech.

Limits of that project

I would say that main limitation of that is an area for the scraping of data in particular space/area. This API offers just use point location with radius and get back just 100 venues. Then, if the point location of district is somewhere, where the radius area is just about family houses, housing estate or is completely different from build-up part of district, results can be absolutely misleading. If would be possible to scrape e.g. 5000 venues from 30 kilometer radius (created from middle of city), which would cover whole Prague and after that separate that places into particular administrative districts, that it would be better and the results information would have better value. But for the purpose of this project it is good. The most important message from that project is to test all the opportunities which are offering

Conclusion

According to this project I went throughout whole process of setting up a concept, identifying business problem, preparing the dataset, use machine learning method and answer the research questions.

That project gave me an opportunity to work with real dataset prepared by myself with using of machine learning method. I tried to lead whole project by my own and figured out all aspects which similar project needs to have.