EUROPEAN CITIES

Capstone final project

24. MAJ 2019

TEJA PERCIC

Introduction

Something about Europe

Europe is a continent located entirely in the Northern Hemisphere and mostly in the Eastern Hemisphere. It is bordered by the Arctic Ocean to the north, the Atlantic Ocean to the west, Asia to the east, and the Mediterranean Sea to the south. European Union (EU) is a political and economic union of 28 member states that are located primarily in Europe. It has an area of 4,475,757 km2 and an estimated population of about 513 million. In 2010, 47.3 million people who lived in the EU were born outside their resident country. This corresponds to 9.4% of the total EU population. Of these, 31.4 million (6.3%) were born outside the EU and 16.0 million (3.2%) were born in another EU member state. Member States have been characterised by an increasing pattern of population concentration, as people move from rural areas towards large cities (and surrounding areas). The high share of young people living in many urban regions of the EU may be linked to lifestyle choices that are linked to education and labour force opportunities
Two of the original core objectives of the European Economic Community were the development of a common market, subsequently becoming a single market, and a customs union between its member states. The single market involves the free circulation of goods, capital, people, and services within the EU.

Bussiness problem

Because it is simple for European citizens to move around the EU, I tried to analyse which cities would be appropriate for young people to live in. Young people don't have a lot of funds available for buying an apartment or house, which is why one of the criteria has to be low average prices of housing. It is important for young people to have a lot of venues in vicinity of their home, because the venues can be their potential employers, educational institution, favourite locations to spend time with friends and family, stores to buy essentials for living etc. So the next criteria is cities with a lot of venues.

Data

Data sources

Based on the definition of the problem, I will need:

- List of EU cities and prices for buying an apartment or house in each city,

The data is available on Eurostat website (csv table downloaded on https://ec.europa.eu/eurostat/web/cities/data/database) for the year 2017 with average prices for buying an apartment or house in EUR per m2.
I will exclude cities that don't have pricing data (clean the table) and edit/clean the data so it will contain only two columns, name of the city and housing price.

- List of coordinates (latitude, longitude) for European cities

A csv file with coordinates of European cities is available on the internet (https://simplemaps.com/data/world-cities).
I will add latitude and longitude to a list of European cities with prices. My data will now have columns: name of the city, housing price, latitude and longitude.

- Number of Venues in each city, within a certain radius from the centre of the city.

With the use of Forsquare API I got data on nearby venues (radius 1500 m) for the list of European cities and decided to look at 100 venues.
My data will now have columns name of the city, housing price, city latitude and city longitude, Venue, Venue Category, Venue Latitude, Venue Longitude.

Cleaning

Because the data on housing prices isn't available for all European cities in the csv file, I decided to drop cities with missing data and analyse only cities with available information.

Some cities had separate information on average prices of housing, one for price of an apartment and the other for a house. I decided to group these information and use median value of average housing prices for one city (df_data_6).

When I created a dataframe containing the data on coordinates (latitude and longitude) of European cities , I needed to drop unnecessary columns (city_ascii, country, iso2, iso3, admin_name, capital, population, id). Then I needed to rename the column City, to match the name of the column in the dataframe cities_grouped (European City).

Because Forsquare didn't return any venue data for the city Eager, I decided to remove it from the data set. The final data set I analysed, contained 57 European cities.

Methodology

After cleaning the data on housing prices for European cities, which were in data frame df_data_6 (european cities with average housing prices) and cities_grouped (European cities with latitude and longitude data), I merged the two dataframes to cities_merged. This dataframe was the basis for gathering of Forsquare venue data.

I created a new dataframe european_venues containing 100 venues per city within the radius of 1500 m of the city coordinates (lat, lng). Because there is no Forsquare data on Eger (data frame european_venues includes only 57 European cities, where the original data cities_merged contain data (name, lat, lng, value) for 58 unique cities), I remove Eger from dataframe cities_merged.

Then I performed clustering by using K-means. K-means identifies allocates every data point to the nearest cluster containing centeroid. I clustered European cities in 4 clusters based on the number of venues they had.

Calculation of target variable

Then I calculated average prices for clusters, to see which one is the cheapest. I also listed 20 most cheap cities to see which cluster they belong to. Clusters were defined based on the number of venues, so the results of the 20 cheapest cities contained cities in different clusters.

Then I analysed clusters to see, which cluster contained cities with most venues. The cluster containing most venues was Cluster 0.

To get the best combination of low price and high number of venues for European cities, I merged data of 20 cheapest cities with 20 cities with most venues.

Clustering

Clustering is an automatic classification of data into a number of groups using a measure of association, so that the data in one group is similar and data belonging to different groups is not similar. To compare the similarity between two observations, distance measure is employed. The clusters were formed with K-means clustering based on number of venues per city. I choose 4 clusters.

I clustered European cities in 4 clusters:

- Cluster 0 (RED)
- Cluster 1 (PURPLE)
- Cluster 2 (BLUE)
- Cluster 3 (GREEN)

Results

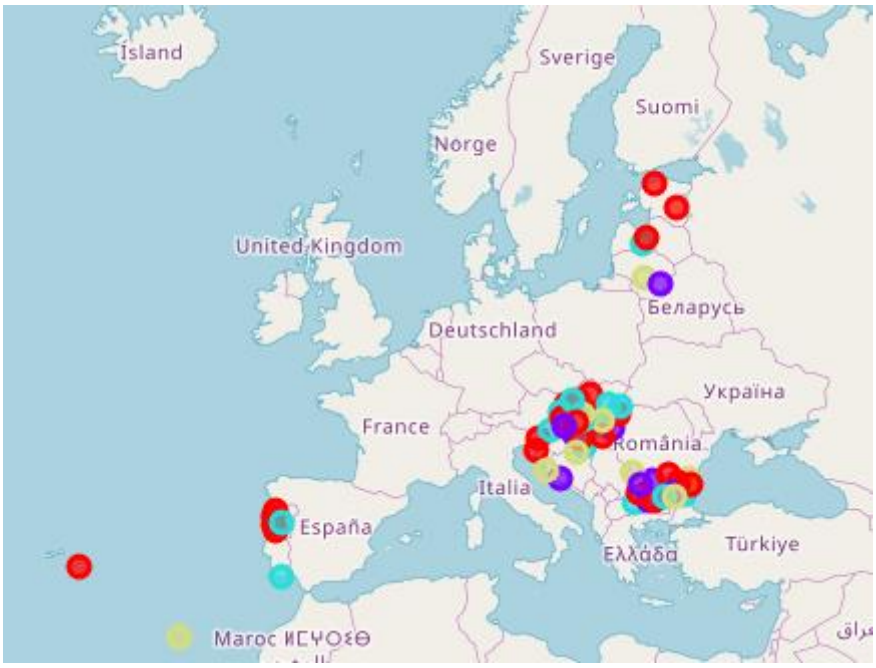

Figure 1: Map of clusters

Cluster 0 (RED)
Average price for buying an apartment/house in this cluster is 21.028,53 EUR per m2.
The city with lowest price in Cluster 0 is **Pécs** with an average price of 502,50 EUR per m2.
Average number of venues per city in the cluster is 96,08 venues.
It contains 24 cities.

Cluster 1 (PURPLE)
Average price for buying an apartment/house in this cluster is 25.866,39 EUR per m2.
The city with lowest price in Cluster 0 is **Békéscsaba** with an average price of 379 EUR per m2.
Average number of venues per city in the cluster is 38,44 venues.
It contains 9 cities.

Cluster 2 (BLUE)
Average price for buying an apartment/house in this cluster is 16.144,88 EUR per m2.
The city with lowest price in Cluster 0 is **Miskolc** with an average price of 373 EUR.
Average number of venues per city in the cluster is 65,78 venues.
It contains 14 cities.

Cluster 3 (GREEN)
Average price for buying an apartment/house in this cluster is 26.554,28 EUR per m2.
The city with lowest price in Cluster 0 is **Szolnok** with an average price of 379,5 EUR per m2.
Average number of venues per city in the cluster is 16,9 venues.
It contains 10 cities.

We can see that the cheapest cities for buying an apartment/house are mostly in Cluster 2, where the average price is the lowest between clusters. But when we look at the list of top 20 cheapest cities, we can see that they are in every cluster. The cities with most venues are in Cluster 0.

Because there is no linear correlation between price and number of venues, we can see that clusters contain cities with different average prices (high and low).

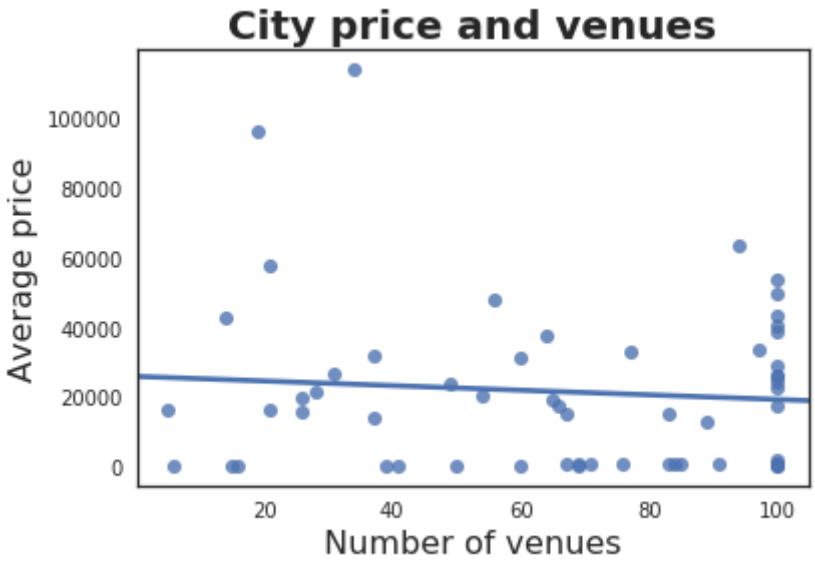

Figure 2: Scatter plot

Discussion

The best choice for me would be to choose a city that is among those in the table "cheapest2" which has the lowest average prices for buying an apartment/house and is in cluster 0 (has most venues).

Table 1: European cities with lowest average apartment/house prices in EUR per m2 and highest number of venues.

European City	Latitude	Longitude	Cluster Label	Numer of Venues	Average price
Budapest	475.000	190.833	0	100	785.0
Debrecen	475.305	216.300	0	100	675.0
Szeged	462.504	201.500	0	100	570.0
Pécs	460.804	182.200	0	100	502.5
Szombathely	472.253	166.287	0	91	634.5

All the cities in Table 1 are located in Hungary, which shows that Hungary is definitely a good choice for young people to live in. Even Hungary's capitol Budapest (the tenth-largest city in the European Union) is among 20 cheapest European cities, offering a lot of job opportunities, educational institutions for young people. It is ranked as the second fastest-developing urban economy in Europe and can offer a lot venues in commerce, finance, media, art, fashion, research, technology, education, and entertainment.

Conclusion

In this final project, I analysed European cities housing prices and geographical data. I identified the most suitable cities for young people to live in, that have affordable housing prices and offer a broad range of activities (venues). I clustered the geographical data using K-means clustering to see which cities have a similar number of venues and can represent a group/cluster. I analysed the clusters to see weather number of venues have any impact on prices of apartments/houses. Because there is no linear correlation, we can see that clusters contain cities with different average prices (high and low).

References

Eurostat
https://ec.europa.eu/eurostat/statistics-explained/index.php/Population_statistics_at_regional_level#Population_density
Wikipedia

https://en.wikipedia.org/wiki/Geography_of_Europe#Population
S. Goswami, Dr. A. Chakrabarti Quartile Clustering: A quartile based technique for Generating Meaningful Clusters
https://arxiv.org/ftp/arxiv/papers/1203/1203.4157.pdf
Zhenfeng Liu: Predicting the Improvement of NBA players
https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DP0701EN/sample_submission/Predicting_the_Improvement_of_NBA_players_Report.pdf
GitHub: limchiahooi
https://github.com/limchiahooi/Coursera_Capstone/blob/master/week5_final_report.pdf