Coursera Capstone Project

The Battle of Neighborhoods - City-to-City Comparison

Alex Bellmann

Contents:

1. Introduction of the business problem and Data description

1.1 Description of the Business Problem

The Importance of Internal Mobility - how to keep good employees once they are hired and how to measure Quality-of-living differentials between cities?

Today, mobility is mainstream, and it's a business strategy that can't be overlooked any longer. The latest research on mobility is also showing that mobility impacts the workforce, leads to better processes and more productivity — and 100% more satisfied employees.

Nothing is more motivating than the feeling that you are in control of your career. Many professionals are willing to explore opportunities with another company, even if they are happy in their current role.

Companies have historically moved operations to new locations to tap into new talent pools and benefit from lower costs, but they should increasingly consider allowing mobile employees a greater flexibility in the choice of where and how they want to work.

Employee Mobility - MOVING PEOPLE TO JOB - should be equally considered with traditional mobility - MOVING JOB TO PEOPLE.

Moving people to jobs is sometimes problematic (family issues, assignment costs, risks in the host locations), thats why we have to find anwsers to following questions:

How to encourage employment mobility?

To encourage employment mobility you need reliable information to help you calculate fair, consistent expatriate allowances and to provide employee with reliable information about the host locations to support his decision.

In this project we will concentrate on the last point -

- provide employee with reliable information about the host locations to support his decision.

What exactly is making a city more attractive to live in and to international business investment?

How to establish quality-of-living differentials between cities and which factors to use for it? Here we need to define the key features (categories) playing the main role for quality of living, which are:

- Schools and education
- Medical and health considerations
- Consumer goods
- Cultural environment
- Natural environment
- Public services and transport
- Recreation
- Socio-cultural environment
- Economic environment

- Political and social environment

How to get clear, objective information on quality of living differences between cities around the world?

To do City-to-City comparison we will need external data sources, which will be described in the next section 1.2

Interested Audience:

International Organisations or organisations with different locations interested in improving of career mobility and keeping good staff. Business leaders can empower their employees with access to opportunities and encourage them to be active participants in the business, rather than setting a passive culture.

Privat persons/ employees looking for opportunities in another city

1.2 Description of the data and how it will be used to solve the problem

1.2.1 Assumption we have taken:

In this project we will concentrate on City-to-City Comparison with assumption of political, social and economic environment to be similar.

1.2.2 Data needed for City-to-City Comparison that summarizes the difference in the quality of natural environment

The following data is required to answer the issues of the problem:

- List of main european cities with their geodata (latitude and longitude) and population
- Foursquare and geopy data to find number of green areas (parks) of european capitals

In [1]:

In this project I will be working only with a subset of the world-cities list, where I selected 25 different popular cities, mainly from Europe and some from Canada, to be compaired.
In [2]:
Here is the first part of requiered data - geodata (latitude and longitude) and population for selection of main european cities

	city	lat	lng	country	iso2	iso3	capital	population
0	Paris	48.8667	2.3333	France	FR	FRA	primary	9904000
1	London	51.5000	-0.1167	United Kingdom	GB	GBR	primary	8567000
2	Toronto	43.7000	-79.4200	Canada	CA	CAN	admin	5213000
3	Montreal	45.5000	-73.5833	Canada	CA	CAN		3678000
4	Berlin	52.5218	13.4015	Germany	DE	DEU	primary	3406000
5	Stuttgart	48.7800	9.2000	Germany	DE	DEU	admin	2944700
6	Frankfurt	50.1000	8.6750	Germany	DE	DEU	minor	2895000
7	Vienna	48.2000	16.3666	Austria	AT	AUT	primary	2400000
8	Mannheim	49.5004	8.4700	Germany	DE	DEU	minor	2362000
9	Vancouver	49.2734	-123.1216	Canada	CA	CAN		2313328
10	Birmingham	52.4750	-1.9200	United Kingdom	GB	GBR	admin	2285000
11	Manchester	53.5004	-2.2480	United Kingdom	GB	GBR	admin	2230000
12	Hamburg	53.5500	10.0000	Germany	DE	DEU	admin	1757000
13	Essen	51.4500	7.0166	Germany	DE	DEU	minor	1742135
14	Leeds	53.8300	-1.5800	United Kingdom	GB	GBR	admin	1529000
15	Lyon	45.7700	4.8300	France	FR	FRA	admin	1423000

Now we will add Foursquare data to our Dataframe to complete the dataset

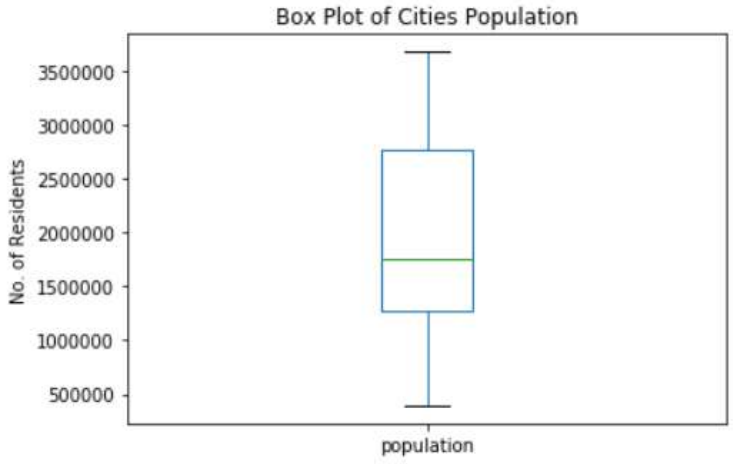
I will investigate on the following Foursquare - Categories:

- Park = '4bf58dd8d48988d163941735' (Natural environment / Recreation)
- Library ='4bf58dd8d48988d12f941735' (Cultural environment)
- Pool = '4bf58dd8d48988d15e941735' (Natural environment / Recreation)
- Playground = '4bf58dd8d48988d1e7941735' (Natural environment / Recreation)
- Cinema = '4bf58dd8d48988d180941735' (Cultural environment)
- Museum = '4bf58dd8d48988d181941735' (Cultural environment)
- Highschool = '4bf58dd8d48988d1ae941735' (Schools and education)
- Kindergarden = '4f4532974b9074f6e4fb0104' (Schools and education)
- Medical Center = '4bf58dd8d48988d196941735' (Medical and health)
- Opera = '4bf58dd8d48988d136941735' (Cultural environment)
- Garden = '4bf58dd8d48988d15a941735' (Natural environment / Recreation)

	city	lat	lng	country	iso2	iso3	capital	population	Park	Library	Pool	Playground	Cinema	Museum	Highschool	Kindergarden
0	Paris	48.8667	2.3333	France	FR	FRA	primary	9904000	50	50	21	50	9	50	48	11
1	London	51.5000	-0.1167	United Kingdom	GB	GBR	primary	8567000	50	46	13	43	12	50	47	4
2	Toronto	43.7000	-79.4200	Canada	CA	CAN	admin	5213000	38	9	8	24	3	0	5	2
3	Montreal	45.5000	-73.5833	Canada	CA	CAN		3678000	50	13	34	23	4	35	50	15
4	Berlin	52.5218	13.4015	Germany	DE	DEU	primary	3406000	49	25	9	47	2	50	26	40
5	Stuttgart	48.7800	9.2000	Germany	DE	DEU	admin	2944700	23	6	5	25	3	22	19	8
6	Frankfurt	50.1000	8.6750	Germany	DE	DEU	minor	2895000	46	10	2	27	3	44	2	10

We will normalize by a Constant, such as the standard deviation of the Population - data

We can see, that 50% of all cities in our Dataset have population between 1.749 m and 2.771 m.

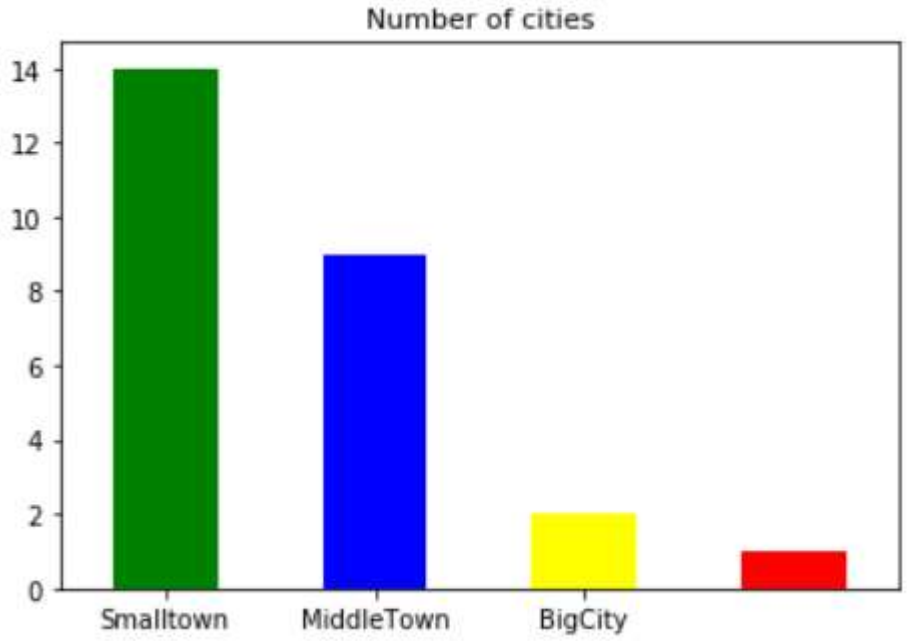


K-means clustering

In this project I will use K-means clustering, the one of the simplest and popular unsupervised machine learning algorithms

Let us try first analyse our dataset deeper based on population and cluster the dataset in 4 Cluster: Mega-city - Big City - Middle town - Small town and see, what k-means clustering made out of it.

```
Number of Mega-Cities: 2
Number of Big-Cities: 1
Number of Middle-Towns: 9
Number of Small-Towns: 14
```



I will investigate on 3 subsets of our data to serve different needs: some people might be more interested in cultural life, some in schools and education, for some of them each kind of venue might be helpful for their desicion making. I assume that natural environment and recreation are important for everyone. So those will be our subsets:

- Subset 1: Cultural life
- Subset 2: Education
- Subset 3: Natural environment/ recreation

As the last step for data preparation I will consider differences in population and normalize the number ov venues in each category by population first.

2.2 Results section

80.76 % of the dataset are cities with very good culture

11.53 % of the dataset are cities with a very good education system

3.84 % of the dataset are cities with very good natural environment and recreation

84.61 % of the dataset are cities with very good/good culture

73.07 % of the dataset are cities with a very good/good education system

30.76 % of the dataset are cities with very good/good natural environment and recreation

Cultural environment
84% are top with cultural offers. This is not surprising, this is what cities are made for.
Here are top 10:

	city	country	population	Total Venues
0	London	United Kingdom	8567000	369
1	Paris	France	9904000	359
2	Vienna	Austria	2400000	342
3	Berlin	Germany	3406000	313
4	Prague	Czechia	1162000	278
5	Vancouver	Canada	2313328	265
6	Montreal	Canada	3678000	254
7	Lyon	France	1423000	235
8	Hamburg	Germany	1757000	211
9	Frankfurt	Germany	2895000	206

Natural environment/ Recreation

Only 30% of selected cities do have a good green environment, definitely something we have to improve in the future, let us see, who is the best:

	city	country	population
0	Tallinn	Estonia	394024

And here are another green cities:

	city	country	population
0	Lyon	France	1423000
1	Munich	Germany	1275000
2	Dusseldorf	Germany	1220000
3	Prague	Czechia	1162000
4	Helsinki	Finland	1115000
5	Copenhagen	Denmark	1085000
6	Cologne	Germany	1004000

Education

73% do have a good education system and here are the best one:

	city	country	population
0	Helsinki	Finland	1115000
1	Copenhagen	Denmark	1085000
2	Tallinn	Estonia	394024

2.3 Noted observations, recommendations based on the results

- There are of course much more venues for analysis to get the proper results in real life, such as air pollution, public transportation, climate, housing, political and economic environment etc
- My Foursquare search was restricted to the 2 km raduis from City - center, as my Foursquare access is restricted, more correct would be to add the city area to the source data and search for venues in adequate radius to get the real number of venues

2.4 Conclusion

I enjoyed working on this final capstone project and learned more about different cities I've never been to yet, my next travel location will be Vienna, looking forward to see this wonderful city! And then? We will see!
Many thanks to Coursera team.