Comparative Analysis between Hamburg (Germany) and Top 10 USA cities

Xuzheng Xiong

24.03.2020

1 Introduction

Immigration is the international movement of people to a destination country of which they are not natives or where they do not possess citizenship in order to settle or reside there. As of 2015, the number of international migrants has reached 244 million worldwide. The reason of that could be financial(e.g. wage rates), personal reasons (e.g. family reunification, transnational marriage).

Barriers to immigration come not only in legal form or political form, but also natural and social form. People when leaving their country also leave their family, friends, social network, and culture. When they arrive in a new country, this is often with many uncertainties including finding work and where to live. To help them ease the migration difficulty, it would be great to help them find a city similar or even better infrastructure and quality of living.

We can use data science nowadays to achieve that goal. Comparing cities around the world by data, we can find out which are similar, which are different.

2 Business problem

We've often believed that more data is better; however, that actually isn't true. The rapid rise in collecting data hasn't been matched by our ability to support, filter and manage the data. Too much data are without enough structure in place to manage and not enough meaningful.

It is difficult to find the truth behind the data. easy to get lost and with so much information it's easy to misunderstand what the data is telling you. It would be great if there was a comparison of the different cities in a country of choice which gives you a general view of each place and it's pro's and contras.

Here the problem will be for a family to decide move from Hamburg, the second biggest city in Germany to United State of America. They want to choose one from the ten biggest cities in USA, which has similar population or more. (Hamburg has a population around 1841179; the biggest city in USA : New York has 8398748, the tenth is San Jose, which has 1025350). Apart from that, they have their factor list for choosing place to live.

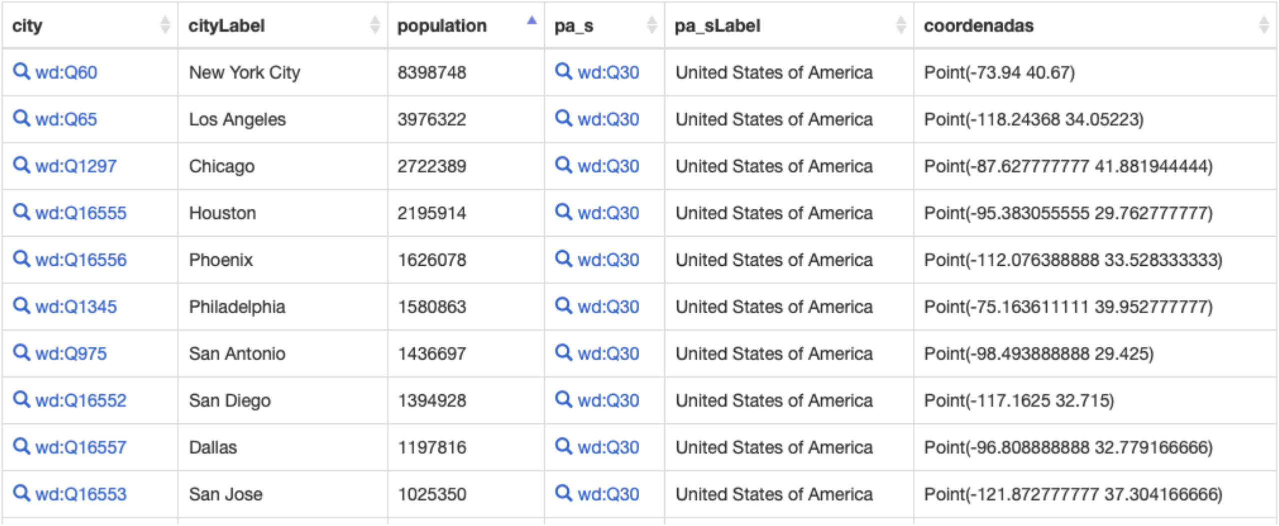
The requirements are:

- *Schools , University*
- *Hospitals*
- *Playground*
- *Shops*
- *Restaurant / Coffee*
- *Entertainment*
- *Nightlife*
- *Lodging*

The findings of this project would be of interest for families moving from Hamburg to USA, but also for companies choose place to explore new opportunities.

3. Data

A list of the cities and they coordinates in Wikidata is used. The data is filtered with country name of United State of America and sorted with population. The results is download and saved as .csv file (see pic 3.1).



city	cityLabel	population	pa_s	pa_sLabel	coordenadas
wd:Q60	New York City	8398748	wd:Q30	United States of America	Point(-73.94 40.67)
wd:Q65	Los Angeles	3976322	wd:Q30	United States of America	Point(-118.24368 34.05223)
wd:Q1297	Chicago	2722389	wd:Q30	United States of America	Point(-87.627777777 41.881944444)
wd:Q16555	Houston	2195914	wd:Q30	United States of America	Point(-95.383055555 29.762777777)
wd:Q16556	Phoenix	1626078	wd:Q30	United States of America	Point(-112.076388888 33.528333333)
wd:Q1345	Philadelphia	1580863	wd:Q30	United States of America	Point(-75.163611111 39.952777777)
wd:Q975	San Antonio	1436697	wd:Q30	United States of America	Point(-98.493888888 29.425)
wd:Q16552	San Diego	1394928	wd:Q30	United States of America	Point(-117.1625 32.715)
wd:Q16557	Dallas	1197816	wd:Q30	United States of America	Point(-96.808888888 32.779166666)
wd:Q16553	San Jose	1025350	wd:Q30	United States of America	Point(-121.872777777 37.304166666)

Pic 3.1 Top Ten Population Cities in USA

It had to be cleaned by splitting the coordinates column into Longitude and Latitude.

The venues for each cities can be queried by using the Foursquare API.

The two set of data can be merged into pandas data frame.

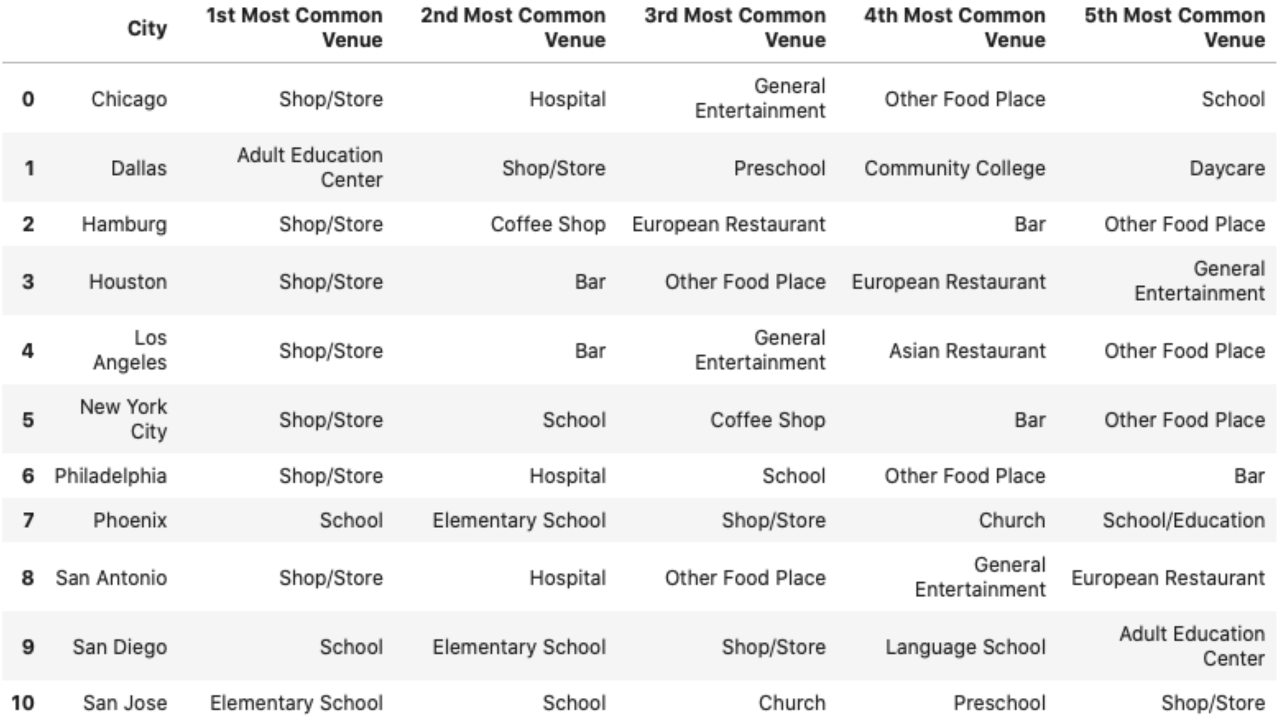
After creating the new dataset with cities, they are plotted on a map using Folium to check if the coordinates were ok.

A new dataframe with the most common venues for each city will be created.

K-Means and Hierarchical Cluster will be used to analyse. The results will be plotted on a Folium Map with colors for each cluster.

4. Methodology

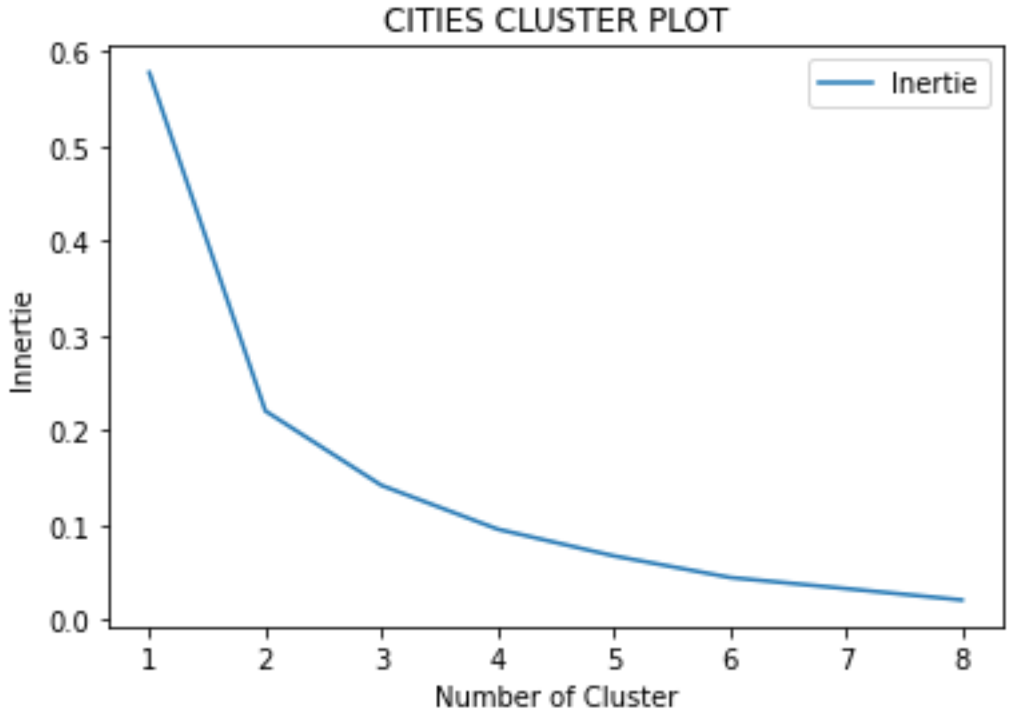
Here we firstly focus on the five most common venue categories of each cities , which is the main characteristics of the city (see Pic 4.1).



	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Chicago	Shop/Store	Hospital	General Entertainment	Other Food Place	School
1	Dallas	Adult Education Center	Shop/Store	Preschool	Community College	Daycare
2	Hamburg	Shop/Store	Coffee Shop	European Restaurant	Bar	Other Food Place
3	Houston	Shop/Store	Bar	Other Food Place	European Restaurant	General Entertainment
4	Los Angeles	Shop/Store	Bar	General Entertainment	Asian Restaurant	Other Food Place
5	New York City	Shop/Store	School	Coffee Shop	Bar	Other Food Place
6	Philadelphia	Shop/Store	Hospital	School	Other Food Place	Bar
7	Phoenix	School	Elementary School	Shop/Store	Church	School/Education
8	San Antonio	Shop/Store	Hospital	Other Food Place	General Entertainment	European Restaurant
9	San Diego	School	Elementary School	Shop/Store	Language School	Adult Education Center
10	San Jose	Elementary School	School	Church	Preschool	Shop/Store

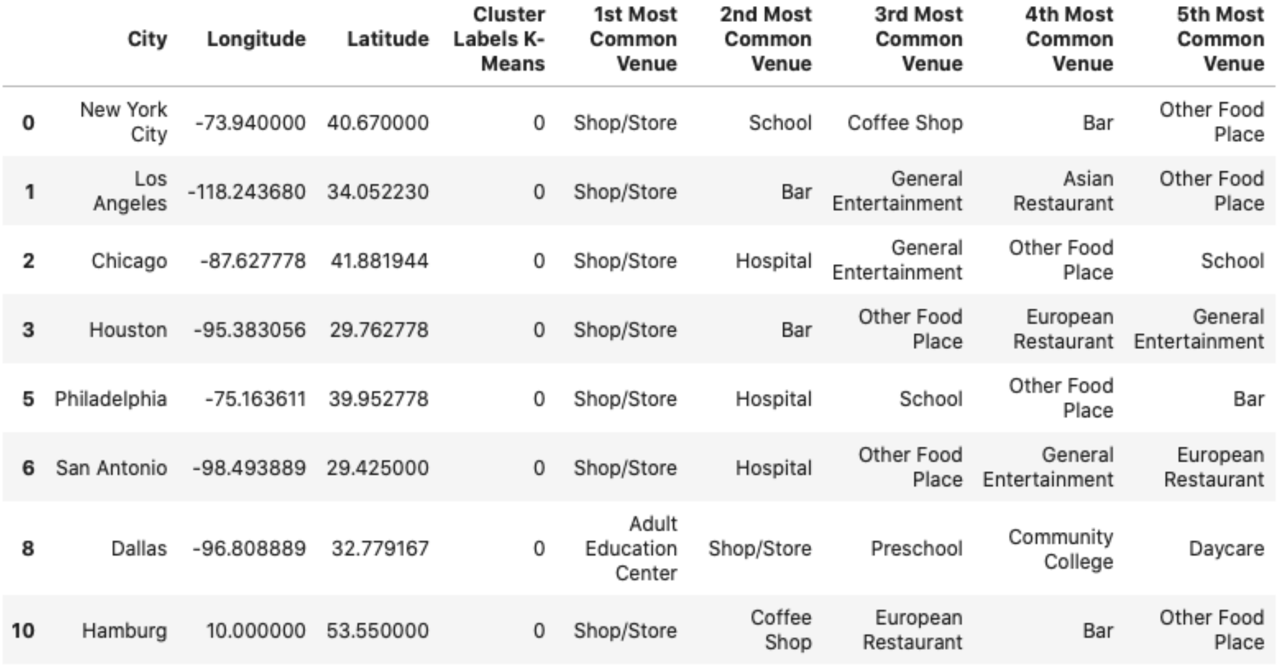
Pic 4.1 Top 5 most common venues of city

Then we use K-Mean clustering method to analyse the diversity. To choose the optimal K value, we will calculate and plot intra-cluster inertia to determine a "elbow" point, which is 2 as shown in Pic 4.2.



Pic 4.2 Inertie Plot for K value choosing

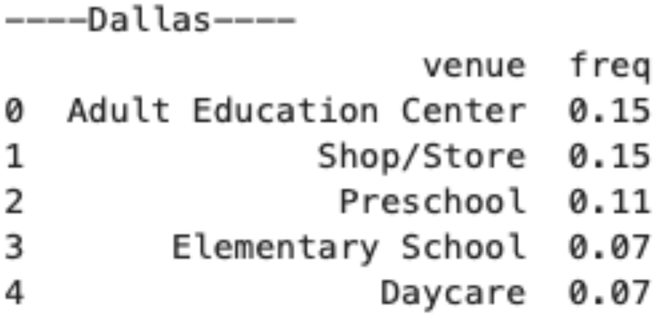
As result, the 8 cities in the first cluster are grouped. The main reason is their first most common venue is shop (see Pic 4.3).



	City	Longitude	Latitude	Cluster Labels K-Means	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	New York City	-73.940000	40.670000	0	Shop/Store	School	Coffee Shop	Bar	Other Food Place
1	Los Angeles	-118.243680	34.052230	0	Shop/Store	Bar	General Entertainment	Asian Restaurant	Other Food Place
2	Chicago	-87.627778	41.881944	0	Shop/Store	Hospital	General Entertainment	Other Food Place	School
3	Houston	-95.383056	29.762778	0	Shop/Store	Bar	Other Food Place	European Restaurant	General Entertainment
5	Philadelphia	-75.163611	39.952778	0	Shop/Store	Hospital	School	Other Food Place	Bar
6	San Antonio	-98.493889	29.425000	0	Shop/Store	Hospital	Other Food Place	General Entertainment	European Restaurant
8	Dallas	-96.808889	32.779167	0	Adult Education Center	Shop/Store	Preschool	Community College	Daycare
10	Hamburg	10.000000	53.550000	0	Shop/Store	Coffee Shop	European Restaurant	Bar	Other Food Place

Pic 4.3 Kluster 1 - Shop TYPE

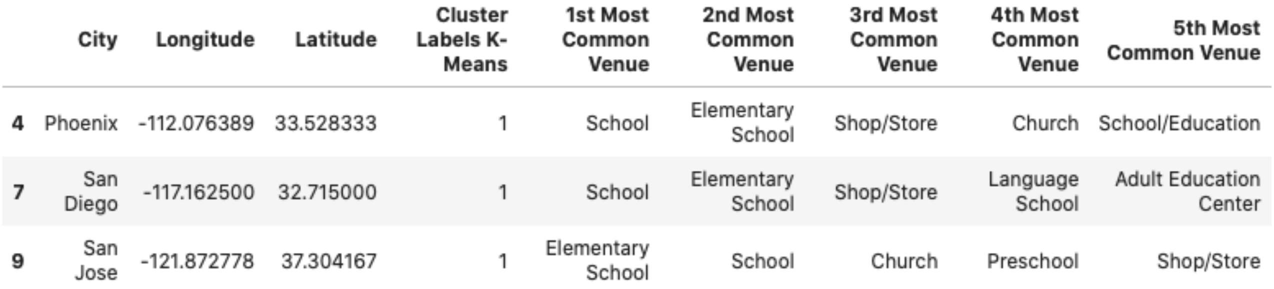
In Pic 4.3 Dallas' first most common and second common (Shop/Store) venues has the same frequency 0.15. Thus it belongs in cluster 1 here. However, we see three of five most common venues are schools (see Pic 4.4), thus, it could belong to cluster 2. We use another category describe data set do the same clustering process, proved the theory.



```
----Dallas----
                     venue  freq
0  Adult Education Center  0.15
1               Shop/Store  0.15
2                Preschool  0.11
3        Elementary School  0.07
4                  Daycare  0.07
```

Pic 4.4 Dallas Top 5 venues frequency

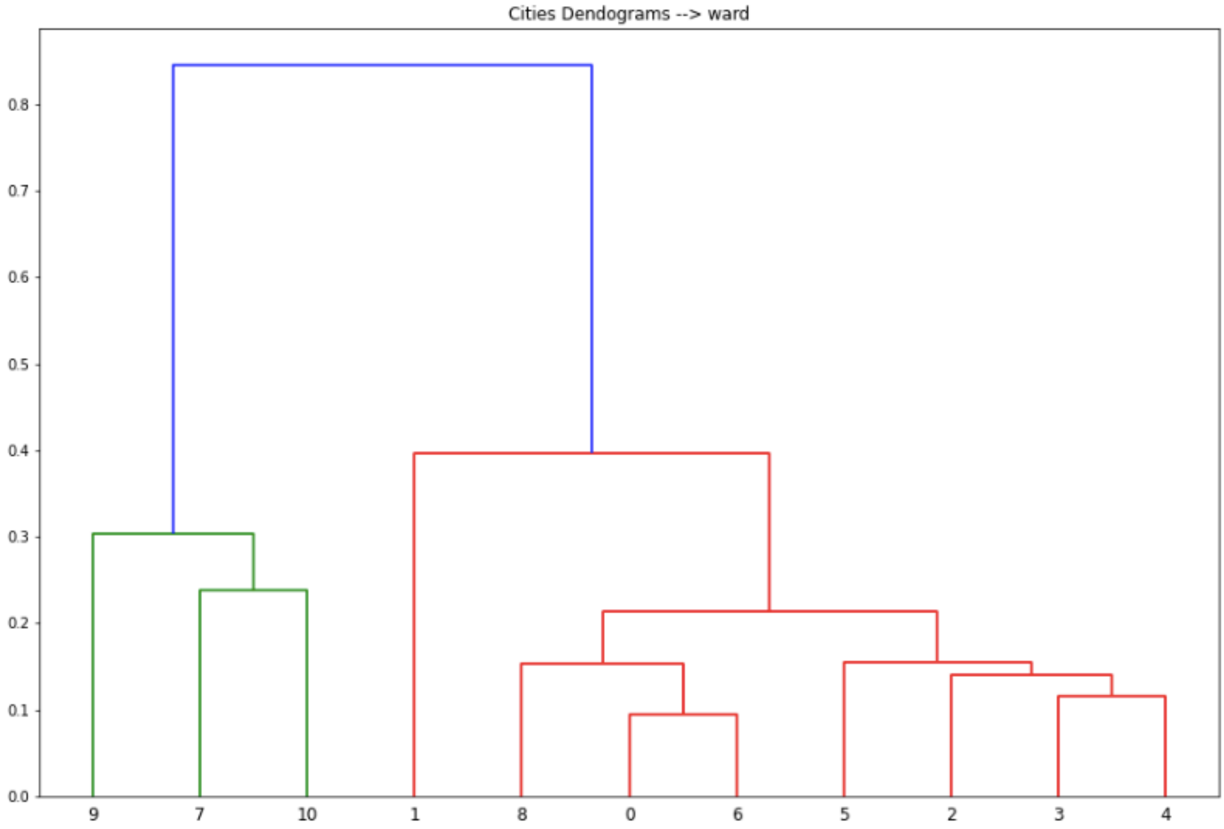
The left 3 cities are in cluster 2, which the first most common venue is school (see Pic 4.5).



	City	Longitude	Latitude	Cluster Labels K-Means	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Phoenix	-112.076389	33.528333	1	School	Elementary School	Shop/Store	Church	School/Education
7	San Diego	-117.162500	32.715000	1	School	Elementary School	Shop/Store	Language School	Adult Education Center
9	San Jose	-121.872778	37.304167	1	Elementary School	School	Church	Preschool	Shop/Store

Pic 4.5 Kluster 2 - School TYPE

We also use hierarchal clustering method to find the most similar city (see Pic 4.6).



Pic 4.6 Hierarchy Clustering

5. Conclusion

Combine both method (K-mean and Hierarchical cluster) , which give us precise and same result . That is the most similar as city 2 (Hamburg) is city 3 (Houston) and 4 (Los Angeles). In respect population, Houston has almost the same as Hamburg, LA has more than double than Hamburg. It could be also a reason for people to choose. People could also choose another city type (school type) in another cluster group.

Furthermore, we can analyse the difference between Houston, Los Angeles, Hamburg more in detail, which can be studied in the future.