Vacations to go: Closeness of German Main Cities



Did you ever visit a city which honestly give a familiar feeling? One day my girlfriend and me visited a new city. While she searched for a parking slot, I surveyed the charming cityscape. After parking we checked-in and discovered the city. Beyond words! There was a tiny funfair. A habit from us is to by a [Paradies Apfel](https://www.google.com/url?sa=i&url=https%3A%2F%2Fheidi-schade-fotografie.de%2Fparadiesapfel%2F&psig=AOvVaw38X2hVlQdYNOkeoJmupCJ0&ust=1594369337898000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCID_7ujev-oCFQAAAAAdAAAAABAE).

While ambling with our luscious apple I marvelled loudly how well known the whole fair was in about 10 meters. My girlfriend argued laughing “a funfair is always the same”. Dauting and mind-blown in the same time. Long time ago I started a Coursera Specialization: Data Science and my claim was to finish it with a classy Blog post.

The aim was theoretical designed.

**Approach:**

“A funfair is always the same”, don’t you wonder why we believe in this and still eat an apple at a new funfair? I glanced at all stores, and ask myself: Is this always the same? The same structure and same shops?

Try to abstract the funfair with big cities and you got your Aim for your Post. Are we able to compare the [biggest cities](https://www.dw.com/en/top-10-germanys-largest-cities/g-52632011) based on the structure and shops with open source tools?

**Approximation:**

In each big cites there are venues, in these venues you will shops or sometime a venue is a shop as well. Can we unsupervised cluster cities based on their shops and categories?

Could we identify to travel to cities which are close too each other? Are we able to cluster Cities based on their shops? Let us dive deeper!

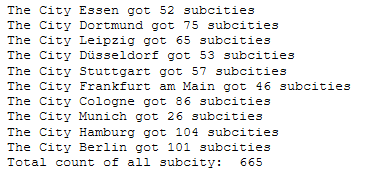
* I used **Forsquare API** to get the most common venues of given Borough of the biggest cities in Germany.
* Web Scraping from site: <https://www.immowelt.de> for getting the districts of cities
* Web scraping form site: <https://de.wikipedia.org/> for getting the population

**Methodology:**

As a first step, we retrieved data from the sites and the foursquare API.

We extract venue information from the districts up to 500 meters

After scraping immowelt the result of all districts was like this:

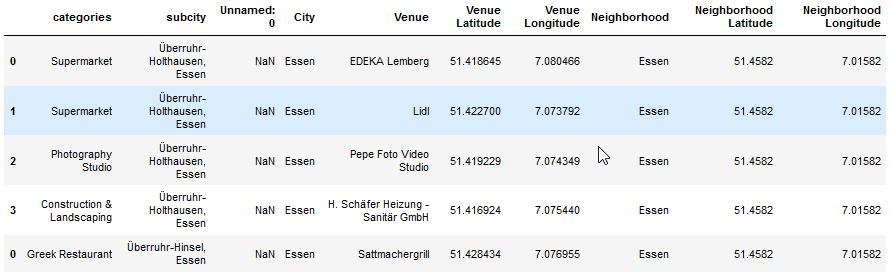


Next, we analyse the data we created on categories and look at a relation between population and shops for the city.

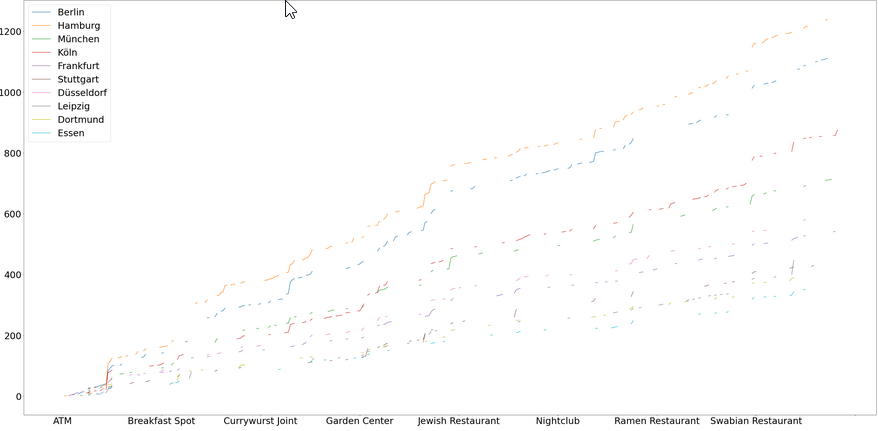
We also explore city areas that shows similarity based on their shops. Finally, we cluster the cities based on the information of each shop category.

The whole process allows us to clearly identify which cities can be recommended for persons which likes similarity to his own next city.

As a database, I used GitHub repository in my study. My master data which has the main components Borough, Latitude and Longitude information of the city.



**Categorization:**



We can see that Germany got many different shop Categories and most of them are overlapping if you are comparing different cities. Do not get confused about the x-axis, I just want to give some examples of our shop categories.

In summary the biggest Germany got 392 different shops, in any case for the foursquare API.

The table below shows a list of the top 10 venue category for each city in below table:



Actually we invest much time and money on our food, same for me with the Paradies Apfel….

**Cluster:**

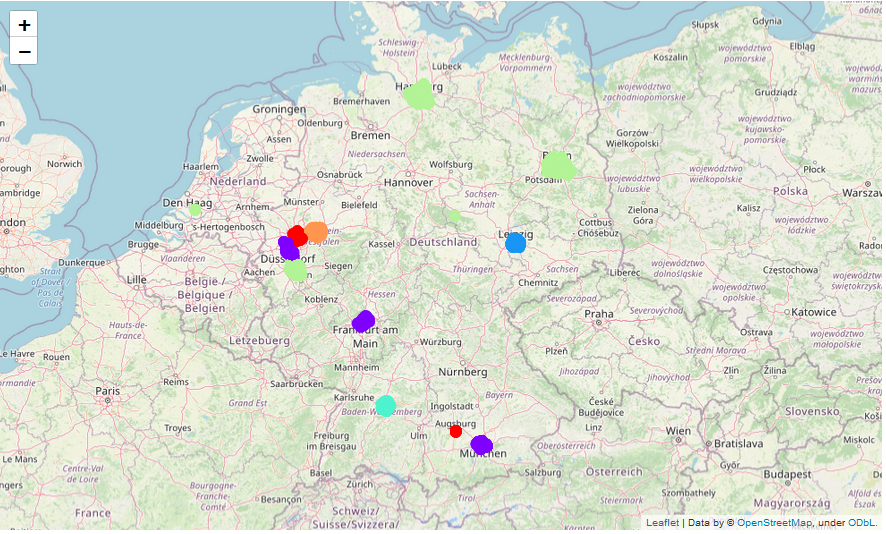
We have some common venue categories in boroughs. In this reason I used unsupervised learning Method K-means to cluster the boroughs.

K-means works in iterative steps to perform two tasks:

* Assign each data point to the closest corresponding centroid.
* For each centroid, calculate the mean of the values of all the points belonging to it. The mean value becomes the new value of the centroid

To visualize the geographic details clusters of all shops and districts I used folium.

Folium helped me to create a map of Germany with boroughs superimposed on top. I used the latitude and longitude from df1 to get the visual as below:



I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as **100 venue** and the radius **500 meter** for each borough from their given latitude and longitude informations. Here is a head of the list Venues name, category, latitude and longitude informations from Forsquare API.

Conclusion

As a result, people are often thinking about which city might be a travel worth.

With this project you can identify which city has some overlapping shops to your hometown. With this approach you may able to identify which city has the most overlapping interests based on the categories from the shops. So take a look at it when you want to travel to a new City. Give it a try and do not hesitate to contact me.

Hope you enjoyed this Article.

To the future,

Lucas Gärtner