# Coursera Capstone Project

Dominik Lechner

May 2021

## 1 Introduction

For the final Project of the Coursera IBM Data Science Course, you have to go through the whole Data Science Circle. First you define yourself a question that you want to answer with data, then you think about which data to collect. After cleaning/ preprocessing the data, you try to use statistical or machine learning methods to extract knowledge from your data. Then you ask yourself if your model solves your question, to refine your process you can apply this cycle literately.

## 2 Business Problem

In my case, I thought about opening a restaurant in London, and the question was, where could be a good place for that. As well I wanted to gain some knowledge about other restaurants, which type of food do they sell, what are the most popular neighborhoods for restaurants. Furthermore I wanted to get some insights in probable customers, so I had a look into demographics data.
This kind of location analysis could be interesting for every entrepreneur, looking to start a restaurant/venue as understanding and finding strategically good positions for any kind of business can be crucial for success.

## 3 Data

- From Wikipedia, I scraped the tabular data about the different neighborhoods in London. `https://en.wikipedia.org/wiki/List_of_areas_of_London`
- On Kaggle I found London borough census data from 2016. `https://www.kaggle.com/marshald/london-boroughs/data`
- Using the foursquare API, I got data about all the different venues in London. `https://developer.foursquare.com/`
- On Github I found a Geojson-file of the different London Boroughs needed to plot them with folium `https://skgrange.github.io/data.html`

# 4 Methodology

All the analysis was done using the Python-programming language mostly with the Data Science package pandas. After scraping the neighborhood table from Wikipedia and cleaning it (e.g. removing citation numbers, transforming OS grid ref to longitude and latitude coordiantes), I merged it with the demographics data. As the census data is given for the different boroughs, I assumed the neighborhoods are more or less homogeneous. For boundary neighborhoods laying in more than one borough, I choose to take the data of the first borough. (Although taking mean values should yield better results. ) After that I took a look at outliers, produced some box plots, and a correlation table using the seaborn library.

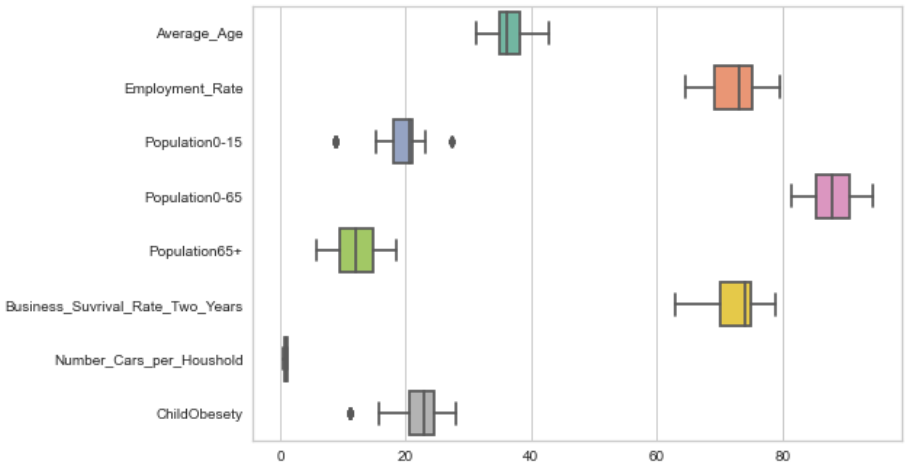


Figure 1: A boxplot

With the Folium package I created some Chorophlethmaps for different demographic facts. From the different Venues I created a subset of Restaurants via the pd.str.contain method. At last a one-hot-encoding was performed to transform the categorical data to numerical one, such that the k-means Algorithm can be applied. The best k has been found by using the "Elbow"-method.

## 4.1 Limitations

- The census data is from 2016, probably outdated. It's a bit coarse for our neighborhood analysis.
- The census data merging isn't perfect (Boundary Neighborhoods).

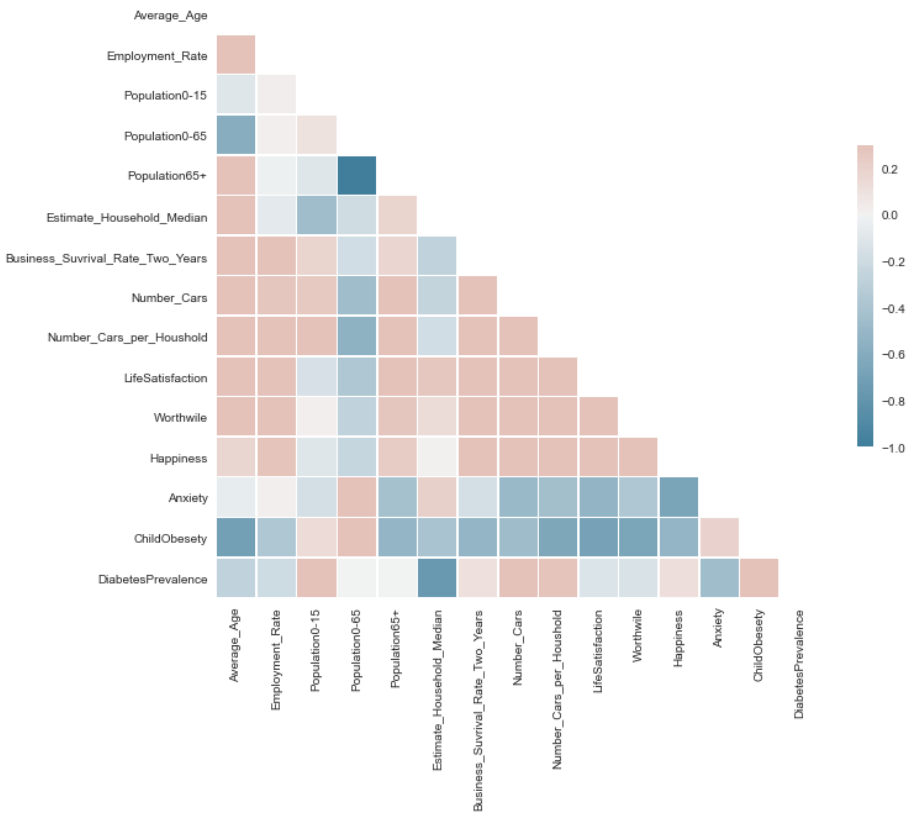


Figure 2: Correlation Plot

- Foursquare didn't deliver restaurants for every neighborhood, one could use different sources (Google maps) to get a more complete Image.

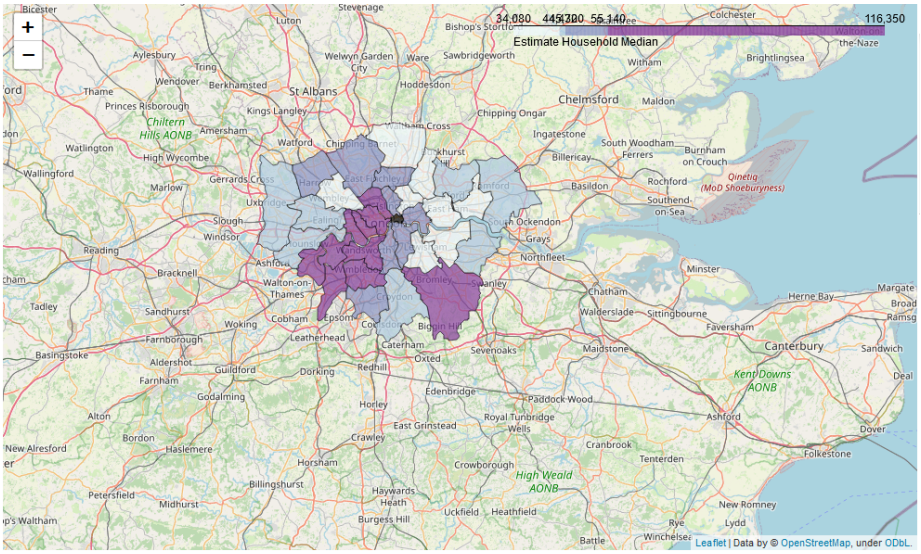


Figure 3: Choroplethmap London Borough Estimated household Median

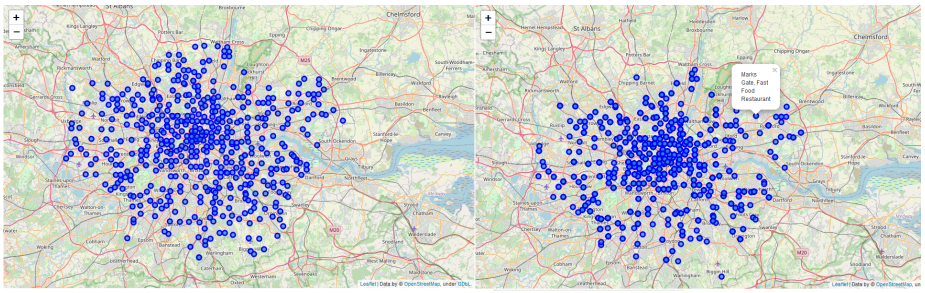

Figure 4: Left image, all the venue data gathered from Foursquare, on the right: the subset of places that serve food

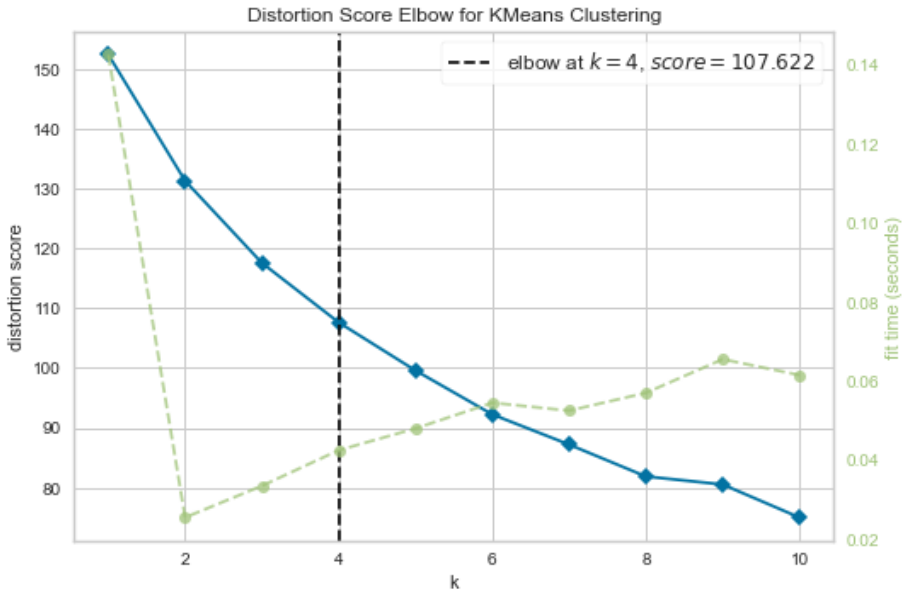


Figure 5: Elbow method to find best k for K-means, as there isn't a kink, the data seems hard to cluster

# 5 Results

We have seen some nice maps, displaying different demographic data from the people of London, one could use that for targeting probable customers. We got some kind of typicality analysis for each neighborhood, giving us the 10 most common restaurant type per neighborhood, which helps to find desire for a specific restaurant. At last by clustering the restaurants in all the neighborhoods,

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Acton | Indian Restaurant | Ethiopian Restaurant | Fast Food Restaurant | Filipino Restaurant | French Restaurant | German Restaurant | Gluten-free Restaurant | Greek Restaurant | Grilled Meat Restaurant | Himalayan Restaurant |
| 1 | Addiscombe | Fast Food Restaurant | Xinjiang Restaurant | Ethiopian Restaurant | Filipino Restaurant | French Restaurant | German Restaurant | Gluten-free Restaurant | Greek Restaurant | Grilled Meat Restaurant | Himalayan Restaurant |
| 2 | Albany Park | Indian Restaurant | Ethiopian Restaurant | Fast Food Restaurant | Filipino Restaurant | French Restaurant | German Restaurant | Gluten-free Restaurant | Greek Restaurant | Grilled Meat Restaurant | Himalayan Restaurant |
| 3 | Aldgate | Italian Restaurant | Salad Place | Middle Eastern Restaurant | Indian Restaurant | Restaurant | Pizza Place | Sushi Restaurant | Sandwich Place | Falafel Restaurant | Vietnamese Restaurant |
| 4 | Aldwych | Restaurant | Italian Restaurant | Sandwich Place | Sushi Restaurant | Seafood Restaurant | Argentinian Restaurant | Ramen Restaurant | Tapas Restaurant | Japanese Restaurant | French Restaurant |

Figure 6: The top 10 most common Restaurant in some Neighborhoods

we see some kind of division between the centre of London and the suburbs.

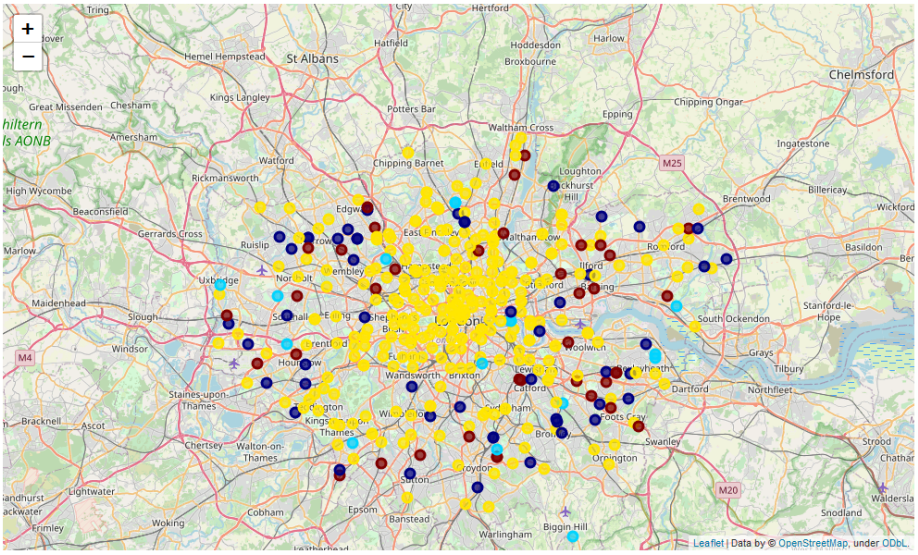


Figure 7: Different clusters created by K-Means

# 6 Discussion

As the questions of where to open a restaurant in London wasn't very concrete, the results also aren't that clear, so one could go much more in depth, trying different clustering algorithms, looking for more Restaurant data. Considering the other venue data, as its highly possible that a shopping mall in walking distance could correlate with a lot of hungry customers, after their shopping trip. Tourists sights/Nice view vs industrial area in the nearby could be interesting. Large offices, implying many workers eating lunch etc..

# 7 Conclusion

As this was a Project to learn to use Data Science tools/libraries, the focus was more on playing around and learning instead of finding the perfect spot for an Italian Restaurant. As the whole Cycle is highly iterable, I would say what was presented here, was some kind of the first iteration. Anyways I had a lot of fun doing that project and learning about different python libraries.

The whole jupyter notebook with all the code, all the graphs/maps can be found on [1]

# References

[1] Dominik Lechner. Jupyter notebook. `https://github.com/DominikLechner1/Coursera_Capstone/blob/main/London%20Neighbourhoods-%20Restaurant%20Analysis.ipynb`, 2021.