1. Introduction/Business Problem

I would like to work with the business problem of where the most optimal place to open an Italien restaurant in Copenhagen is. To achieve this, i will use an analytical approach, where ill use advanced machine learning techniques and data analysis, concretely clustering and data visualization in form of maps.

During the process of analysis, several data transformations will be performed, in order the find the best possible data format for the machine learning model. Once the data is set up and prepared, a modeling process will be carried out, and this statistical analysis will provide the best possible places to locate the Italian restaurant.

The target audience of the report would be either new business owner og exiting looking to branch out there business. They would easily benefit from such an analysis, as it could be used to review where competitors are located

2. Data

A dataset from The copenhagen municipality. https://data.kk.dk/dataset/befolkningen-efter-ar-bydel-alder-og-statsborgerskab The dataset contains information about the inmigrant population per nationality in the discrict of Copenhagen. This data will be used to determine the best location of the restaurant based on peoples nationalities. I will be assumed for this exercise that peoples likes varies according to their nationality, and that people from one specific country will be more attracted to place that matches the environment and culture of their own countries, rather than the ones from foreign countries.

The data set contains the following data: Year - 1992 - 2015 District - Copenhagens 10 districts, with the numbers 1-10 will be replaced with names Age - Age of the population Population - Number Nationality ID - 4 digit code that can be used to find the nationality as described below

I will extract the features from the last year in the dataset "2015" and dismiss the age column and sum the dataset on year, district, natiality ID and pop. For the Nationality ID i will use a webpage for scrabing the name for the ID and then join the it on the
dataset https://www.dst.dk/da/Statistik/dokumentation/Times/forebyggelsesregistret/statkode

3. Methodology

I will use statistical exploration of the data and combine it with map visualizations. I will do the machine learning technique of clustering by using K-Means algorithm. All is implemented using Python and Jupiter notebook.

To find the answer on the business problem the first key is to located the necessary data and scrape the data so it is usefully. Normally when you want to find a optimal restaurant location you will need customer data and there input on when they go out to eat and what they prefer. But for this report I focus on the more simple approx, where I make the assumptions that population from a certain country would prefer / eat at a restaurant that serve national food and atmosphere.

So with this in place I found the two dataset that I describe in the data section. Initial data looks like this:

	Year	Neighbourhood	Nationality	Population
0	2015	1	5100	614
1	2015	1	5104	2
2	2015	1	5106	1
3	2015	1	5110	1
4	2015	1	5120	4

After the data as been cleaned and have been join up it looks like this :

```
(10, 188)
```

Tekst	Neighbourhood	Afghanistan	Afrika uoplyst	Albanien	Algeriet	Angola	Arg
0	Amager Vest	97.0	0.0	1.0	10.0	0.0	21.0
1	Amager Øst	49.0	0.0	5.0	5.0	0.0	9.0
2	Bispebjerg	97.0	2.0	6.0	15.0	0.0	11.0
3	Brønshøj-Husum	142.0	0.0	8.0	10.0	1.0	1.0
4	Indre By	2.0	0.0	3.0	3.0	0.0	23.0

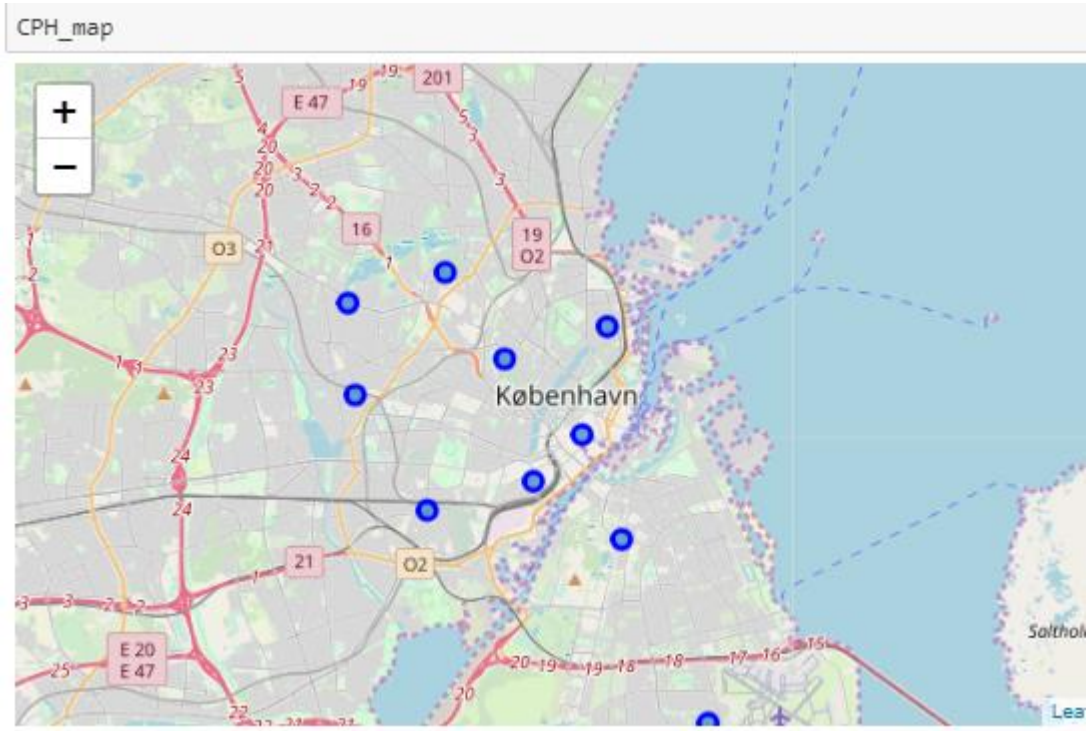
Cleaning data involved:

- Separate 2015
- Add "Neighborhood" text instead of no. So Foursquare can help identify area
- Add "country" from scraping another side with the country name, done by using beautifulSoup.
- Group the data and summed the populations
- Pivot the data, so the 10 neighborhoods are in rows and countries are in columns

Then I found the locations of each neighborhoods

	Latitude	Longitude	Neighborhood
0	55.678316	12.575872	Indre By
1	55.701539	12.585481	Østerbro
2	55.694574	12.546701	Nørrebro
3	55.668443	12.557340	Vesterbro/Kgs. Enghave
4	55.662052	12.516845	Valby
5	55.686833	12.489733	Vanløse
6	55.706569	12.486981	Brønshøj-Husum
7	55.713338	12.523701	Bispebjerg
8	55.616516	12.623910	Amager Øst
9	55.656049	12.591308	Amager Vest

And mapped them using Folium:



Then I obtained venues / restaurant in 1500 meters of the center points for each neighborhood. By using Foursquare API. I was returned in a table:

```
print(CPH_Restaurant.shape)
CPH_Restaurant.head()
```

```
(713, 7)
```

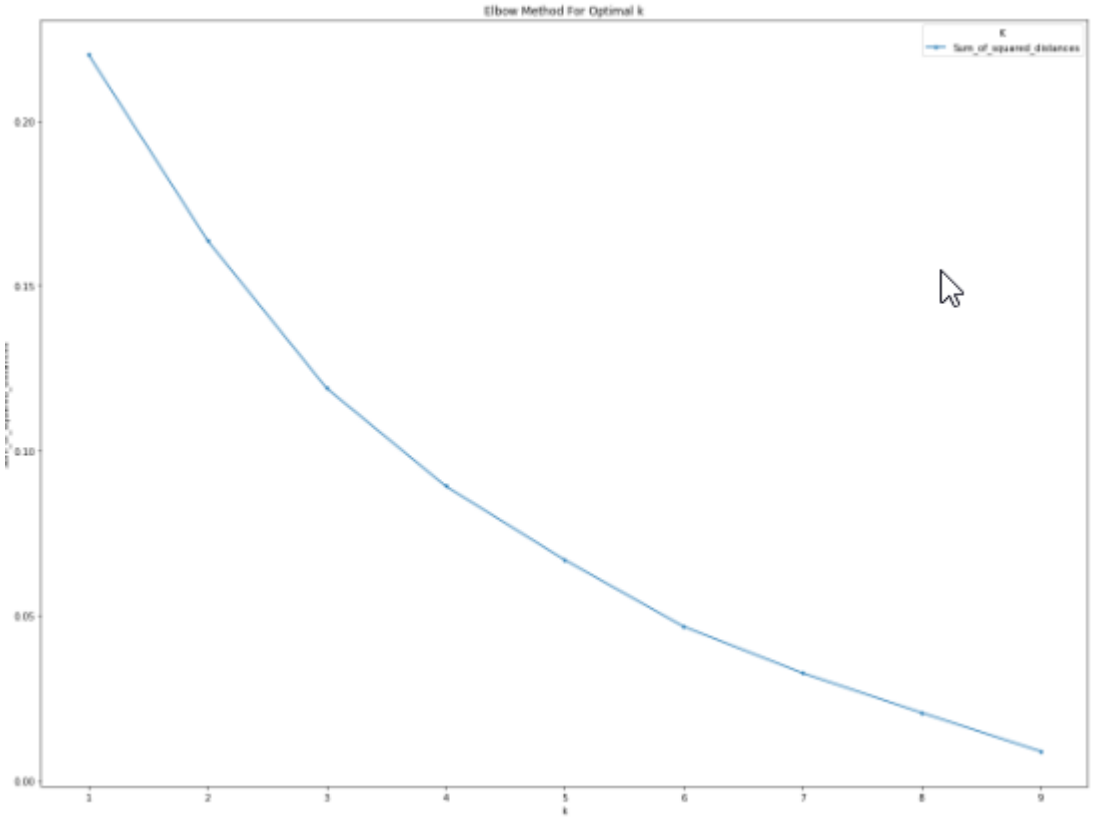
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Indre By	55.678316	12.575872	Ruby	55.676703	12.576727	Cocktail Bar
1	Indre By	55.678316	12.575872	Illums Bolighus	55.678855	12.578590	Furniture / Home Store
2	Indre By	55.678316	12.575872	Bastard Café	55.676483	12.574992	Gaming Cafe
3	Indre By	55.678316	12.575872	Faraos Cigarer	55.679538	12.574541	Comic Shop
4	Indre By	55.678316	12.575872	Bertels Salon	55.677661	12.576550	Dessert Shop

```
CPH_Restaurant.groupby('Neighborhood').count()
```

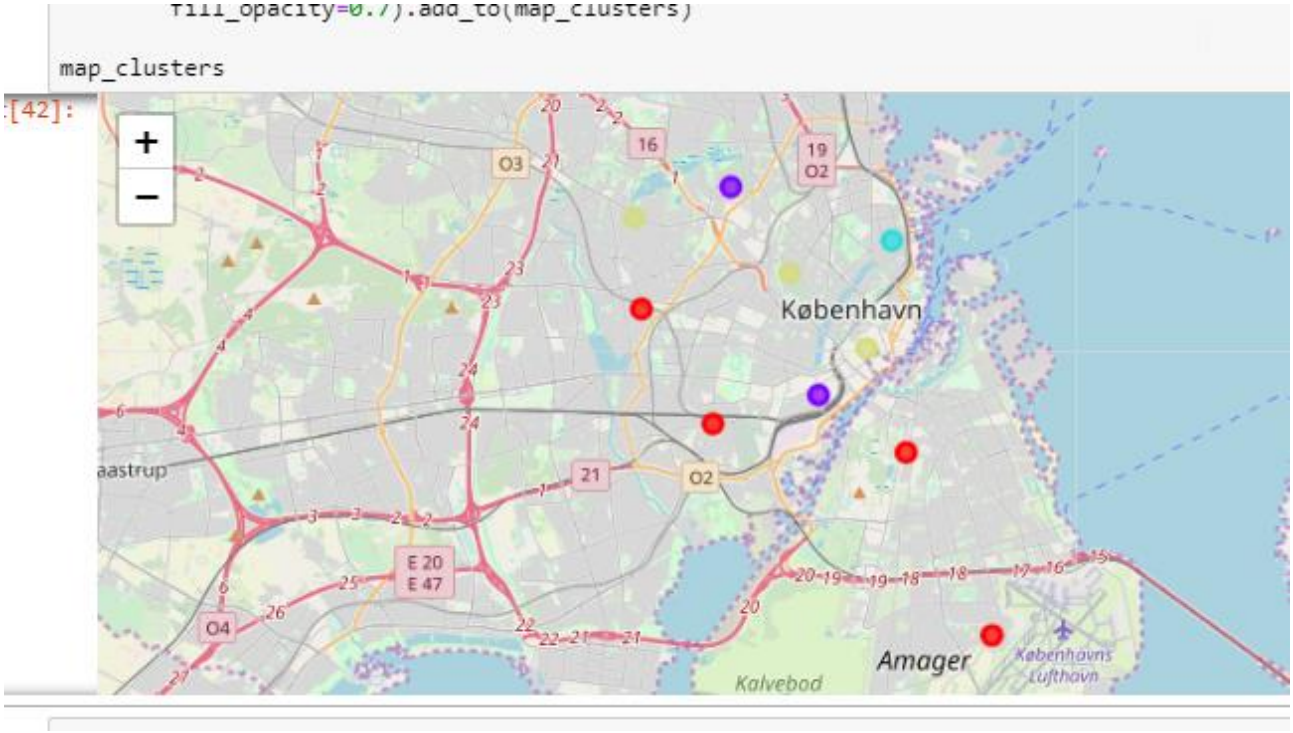
Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Amager Vest	81	81	81	81	81	81
Amager Øst	22	22	22	22	22	22
Bispebjerg	50	50	50	50	50	50
Brønshøj-Husum	44	44	44	44	44	44
Indre By	100	100	100	100	100	100
Nørrebro	100	100	100	100	100	100
Valby	55	55	55	55	55	55
Vanløse	65	65	65	65	65	65
Vesterbro/Kgs. Enghave	100	100	100	100	100	100
Østerbro	96	96	96	96	96	96

On this I did some onehot encording to get data into 0 or 1. So I end up with a dataset with each neighborhood and there top 10 most common venues.

Then I wanted to do some segmentation and clustering, to do this I used the elbow method. This plots our data and return theme as a curve representing the squared distances between each cluster.



Based on this curve the best suited numbers of cluster for my data is around 4. Now this cluster my dataset with neighborhoods into 4 clusters and on that I can do my final analysis to find out the most suited place for my Italian restaurant. Below is a map of my 4 new clusters:



4. Results

The result of all above was 4 cluster that I inspected individually.

- Cluster 1
  This Is the main center of Copenhagen, with lots of turist and so on. Here the top venues are Coffee shops, Cafes and bars. I would suspect this to be the case as rental is high at this locations and driving a lot of the same type of restaurant wouldn't be possible. But that being said the top 10 also includes Pizza and some Italian restaurant witch tells me that maybe this area already as this segment covered.

  The inhibiters are mainly Danish, with France, Germani, India, Italia, Norway & Sweden as the top foreigners

- Cluster 2
  Is the outer area of Copenhagen lots of families and students lives here to do the lower rent. There is a lot of mixture in the venues here, with Pizza and grocery store on top. Not a lot of restaurants in this area.

  The inhibiters are mainly Danish, with Thurkey, Parkistan, Somalia, Polen, Markedonia as the top foreigners. Mainly east europian and middle easten lives here along with the danes.

- Cluster 3
  East of Copenhagen nearer to the airport, main venues are Hotels and fast food restaurants. Mainly Danish ppl living here with Polen, Tyrkie, Germany and USA as the other top countries

- Cluster 4
  Contains areas around Copenhagen abit closer to the town center, here the main inhibiters is similar to the ones in cluster one. And the venues are Pizza & Café

5. Discussion

Its is interesting to see how much each cluster different from each other both in venues and in countries. To make a better predictions on where to open the restaurant we need more data such as income, children, education, job etc. With all this information we could segment more precise.

So for a future project this could be a interesting thing to dive into and see how far you could take it.

6. Conclusion

If we follow the logic of where it is best to open an Italian restaurant based on the inhibiters. Cluster 1 is the very center of the town here you would get lots of competitors and you could argue that in this areal the inhibiters are most likely to go out.

Cluster 2 and 3 does not contains a lot of competitors but the inhibiters are mainly from other countries than Italia. So that leaves us with cluster 4 where you have a good mix of Pizza and Cafes in the top venues and an ok amount of Italian inhibiters, so here would be a place for opening a Italian restaurant that aint a pizzeria to compete against those.