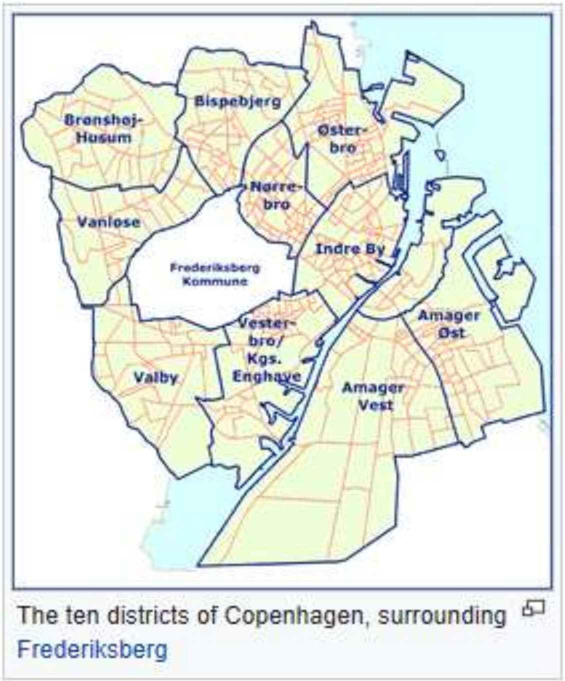
Capstone Final project

1). Introduction

Every person has dreams and ambition of being able to venture off into free lancing or opening up their own business. The comfort of being your own boss also has some insecurity issues because you have to make a lot of decisions that could either lead to success or failure. With that said, if you were to open up a business extensive research should be done to ensure the optimal selection of location and what to open is chosen. As the world is becoming more globalized, many countries are becoming more multicultural especially in western countries. More specifically, Denmark has seen an increase in population in the last years with Copenhagen experience the largest growth. Copenhagen city is the capital city of Denmark with approximately 1,153,615 people, it is the largest city in Denmark with many tourist attractions, shipping harbors, an airport, modern architectural buildings, traditional architectural buildings and many more. Copenhagen can resemble in many ways as New York City minus the vast skyscrapers. There are 10 official administrative districts in Copenhagen: Indre By (central city), Vesterbro/Kongens Enghave (West bridge/Kings garden), Nørrebro (North bridge), Østerbro (East bridge), Amager Øst (Amager east), Valby, Bispebjerg, Vanløse and Brønshøj-Husum. The names are in Danish with some English translation given.



The ten districts of Copenhagen, surrounding Frederiksberg

Figure 1: Map showing the 10 districts found in copenhagen [1]

There is of course a long history involved in the city dating back to the Viking era in the 10th century where the actual establishment of the city being the capital took place in the early 15th century.

However, for this report it is not convenient to provide extensive information, and therefore further detail and history will not be provided.

1.1) Problem definition

An investor has always had an interest of opening of a small business in the Copenhagen area, but is not sure where. Therefore, he has hired my team to analyze the city by seeing what shops/businesses are already established, and in what areas. This will hopefully provide strong evidence in order to help make the decision on what type of business to open, and in what location. The reason for the importance of this is this has always been a dream for him as a young man, and is tired of living the office life that has been a part of him the last 20 years. Since the problem will deal with categorical data, and we really do not know what to label the data would indicate that we could use Clustering as our modelling technique.

1.2) Report structure

The report structure will start off with explaining the data necessary to help drive this investigation. This will be followed by explaining the data processing procedure in order to start the modelling. The next section will be exploring the data. Once the data is explored, the modelling will take place followed by our conclusion

2) Data acquisition and cleaning

2.1) Data sources

To retrieve the names of the districts in the Copenhagen area, the following url https://en.wikipedia.org/wiki/Districts_of_Copenhagen will be scraped to construct a dataframe containing the names. The geocoder library will then used to obtain the coordinates for each name. Lastly, the foursquare api will be used to retrieve the names of each venue in the districts.

2.2) Data cleaning and methodology

Scraping the website was accomplished using pandas and Beautifulsoup libraries. After observing the html code, it was found that the areas of interest started with <ul> and <il>. A list was created by iterating over the html source. The list however contained unwanted strings that needed to be removed, and the first part of the removal process was to shorten the list with the unwanted strings that could easily be removed. Afterwards, a dataframe was constructed followed by further removal of unwanted strings. Another dataframe containing the coordinates was then constructed followed by combining the two dataframes. It was found, after combining the two dataframes, that there were NaN present. Since it was not justifiable to replace the missing values because they were coordinates to a specific location, it was decided to just remove the missing values. The get_dummies function was used to convert categorical features into numbers before applying clustering.

3) Exploratory Data Analysis

The first part of exploring the data was to investigate the venues in the Copenhagen district. It was found that there are a total of 1851 venues in Copenhagen, with 788 being unique and 183 different venue categories. A list of the top 20 venues in the Copenhagen area was plotted just to give an idea of the most popular small businesses, and what interest Danish people have that live in the capital city. The results can be found in the figure below.

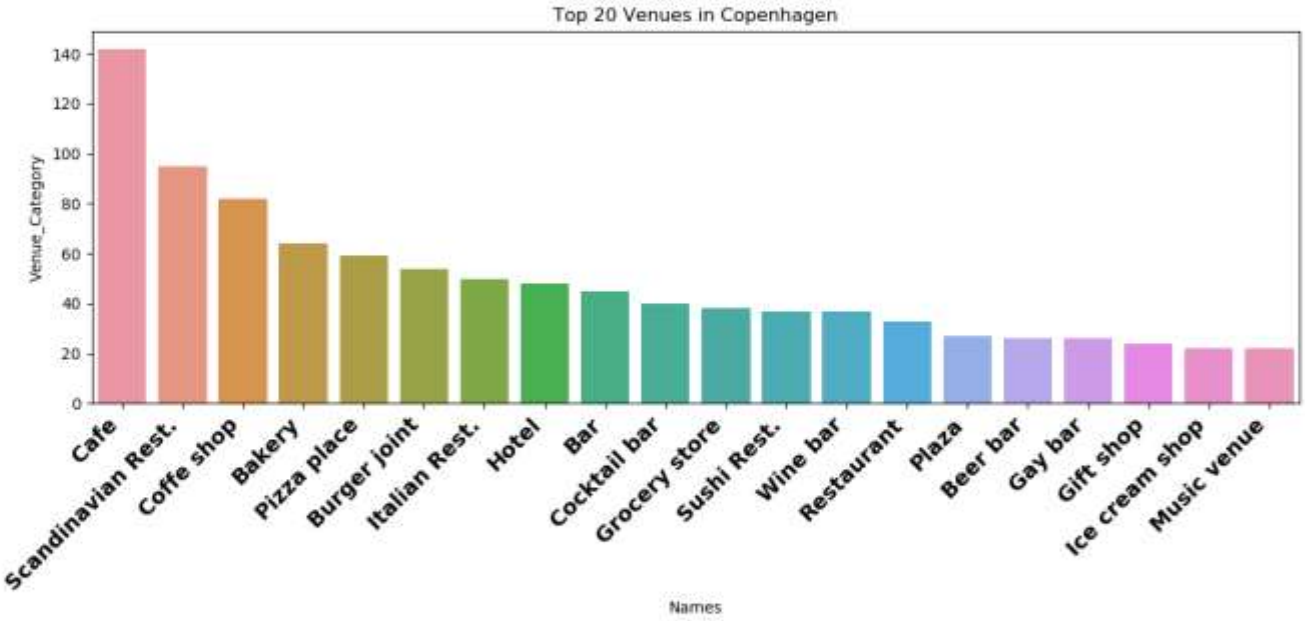


Figure 2: Illustration of each category plotted against the venue names for top 20.

It is observed that café is the most popular venue to open in Copenhagen with 142 shops followed by Scandinavian restaurants with 95 and coffee shops at number three with 82. What we can get out of this is that people living in CPH (short for Copenhagen) likes sitting at a café. Café's in the area contain a variety of things to order from food, cold and warm drinks. Food content varies from place to place, but in general from own experience the majority of cafés have some sort of burger or sandwich. Scandinavian restaurants include casual and fancy eating places that contain traditional food, but also gourmet and international. This is also a popular choice, and there are a lot of Danish people who work long hours and do not have time to cook, so they usually eat out.

The next figure will illustrate each of the districts/neighborhoods that these venues can be found. This is represented by a blue circle, and is constructed using the folium library in python.

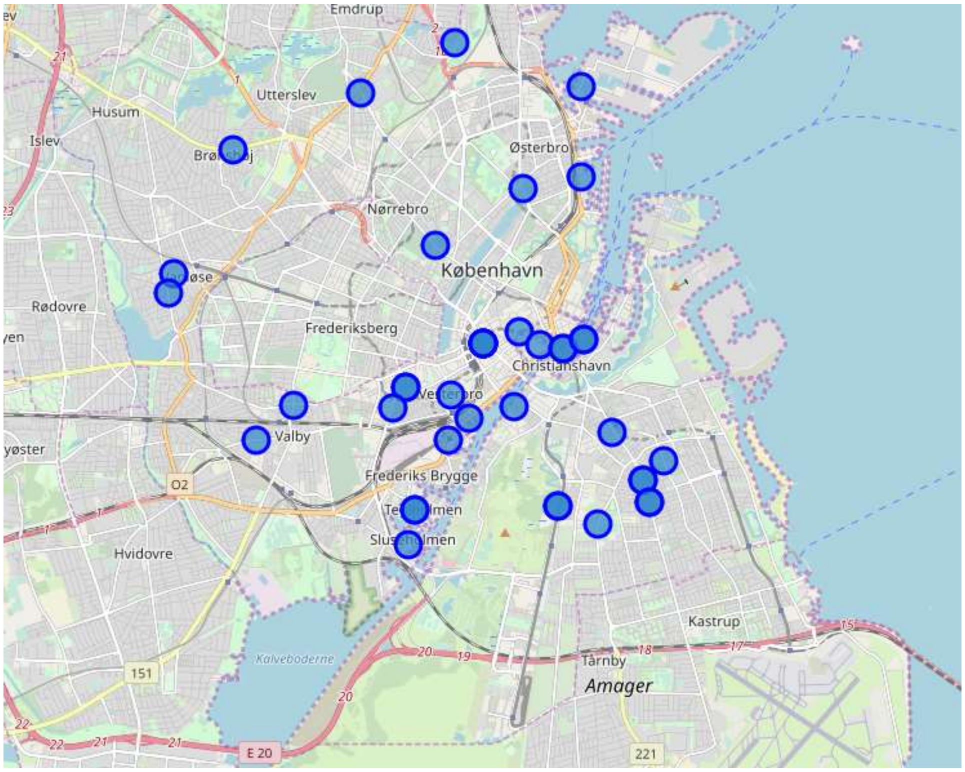


Figure 3: Illustration of the copenhagen area with the districts that are investigated represented by blue circles.

As it can be seen, the majority of locations are in the center of the Copenhagen area (København) with very little on the outskirts of the city. With that said, it will most likely be highly optimal to investigate these locations more closely. Also, at this point it can be hypothesized that the business should be in the center of the city.

4) Modelling

This section will contain the results from applying the machine learning algorithm that will help provide conclusive results in deciding what and where to open a venue. As stated before, the machine learning technique that will be used is clustering from the sickitlearn library. Before applying the technique, it was necessary to convert categorical features into numerical by using the get_dummies function, which was stated in the previous section.

It is not known how many clusters should be used, so it is mandatory that we determine the optimal number of clusters to use. For this, a range of k values starting from 1 and going to 9 were used on the converted dataset. To quantify the choice of the right k value, the sum of square distance to the closest center was used and plotted against the 9 k values. The results can be found in the figure below.

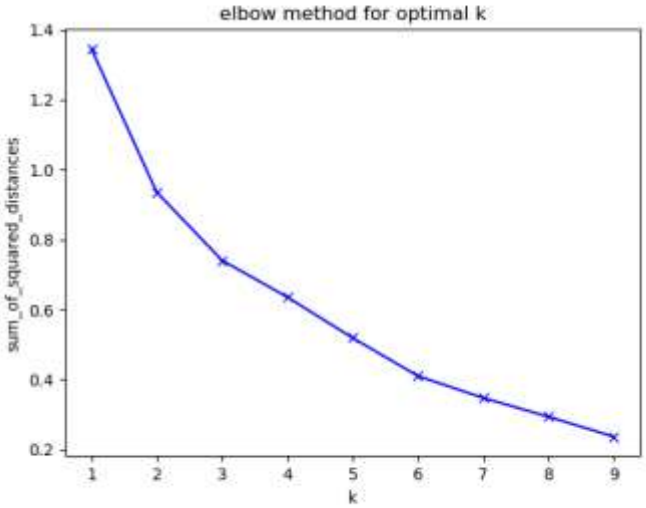


Figure 4: Illustration of sum of square distance plotted against the k values.

The elbow method is defined as the point along the line where it starts to bend. This is not as obvious in the figure above, but it can be argued that either 2 or 3 clusters could be the optimal choice of clusters. It was concluded that 3 clusters will be used for further analysis.

Three clusters were used with the KMeans algorithm on the venue dataset. To visualize this, the folium library was used to see each of the three clusters, and the results can be found in the figure below.

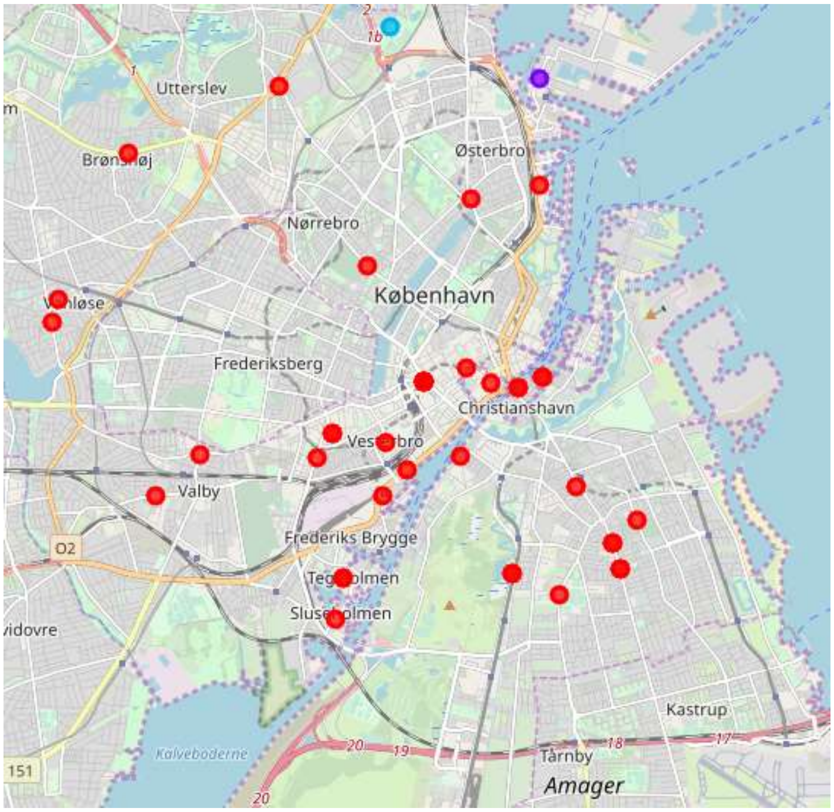
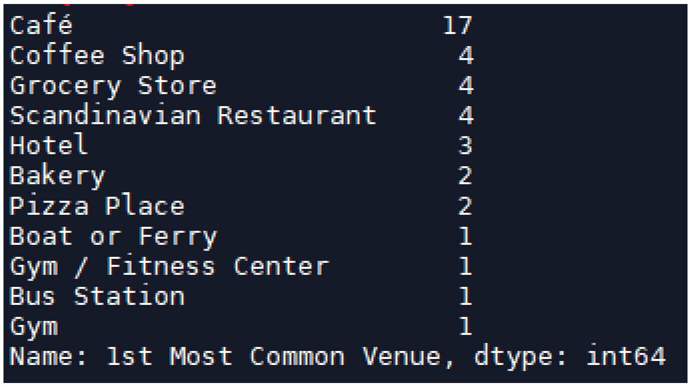


Figure 5:Figure containing the three clusters with cluster 0 (red), cluster 1 (blue) and cluster 2 (purple).

As it can be seen, cluster 0 (indicated in red) has the highest number of venues with cluster 1 and 2 only having one location. The majority of cluster 0 is located in the heart of the city, which is also where a lot of offices, apartments, schools and venues are located. The other two clusters were therefore excluded, and cluster 0 was examined in more detail. Taking cluster 0, it was of interest to investigate the type and number of venues located in this cluster. It was observed that, as expected café being the most popular with a total of 17 venues followed by coffee shops, grocery stores and Scandinavian restaurants all sharing a total of 4 venues. This is interesting since the barplot presented earlier in the report showed that Scandinavian restaurants clearly was second after café venues.



```
Café                          17
Coffee Shop                    4
Grocery Store                  4
Scandinavian Restaurant        4
Hotel                          3
Bakery                         2
Pizza Place                    2
Boat or Ferry                  1
Gym / Fitness Center           1
Bus Station                    1
Gym                            1
Name: 1st Most Common Venue, dtype: int64
```

5) Conclusion

The objective of this paper was to investigate the perfect location for a venue in the Copenhagen area, which is the capital city of Denmark. It was of interest of a private investor that was tired of the office life, and wanted a change in direction. Therefore, I was hired to investigate this problem by using techniques such as web scraping, data analysis and applying a machine learning algorithm to cluster the data. It was first concluded that the optimal number of clusters should be 3 indicated by 0,1 and 2. Cluster 0 had the highest number of venues, and therefore my advice to my client is to consider one of the locations in this area. After further investigation of the popular venues in cluster 0, it was concluded on my behalf that there are many Cafés located in this area, which says a lot about the Danish culture. Therefore, venturing off into a newly idea could be problematic and risky if for example a new gourmet/gastro eating place should be open. Of course, there appears that not many restaurants are located in cluster 0, and therefore could be a good business idea to bring something new to the city. However, a more thorough investigation should be considered which will look at the prices of the establishments, how many people are living in cluster 0 and how many available locations are on the market.

6)References

[1] https://en.wikipedia.org/wiki/Districts_of_Copenhagen