CLUSTERING OF COPENHAGEN STATIONS

Author : Anas Rezk

Submitted : May 27, 2020

TABLE OF CONTENTS

PAGE

LIST OF FIGURES

PAGE

LIST OF TABLES

PAGE

No table of figures entries found.

1 INTRODUCTION

1.1 BACKGROUND

Copenhagen, the Capital of Denmark, has today 125 metro and train stations. Government is discussing building 4 new metro stations and 2 new train stations.

Copenhagen Metro is a 24/7 rapid transit system in Copenhagen, Denmark, serving the municipalities of Copenhagen, Frederiksberg, and Tårnby. The original 20.4-kilometre (12.7 mi) system opened in 2002, serving nine stations on two lines: M1 and M2. In 2003 and 2007, the Metro was extended to Vanløse and Copenhagen Airport (Lufthavnen) respectively, adding an additional six plus five stations to the network. In 2019, seventeen stations on a wholly underground circle line, the M3, was added bringing the number of stations to 37. The driverless light metro supplements the larger S-train rapid transit system, and is integrated with local DSB and regional (Øresundståg) trains and municipal Movia buses. Through the city centre and west to Vanløse, M1 and M2 share a common line. To the southeast, the system serves Amager, with the 13.9-kilometre (8.6 mi) M1 running through the new neighborhood of Ørestad, and the 14.2-kilometre (8.8 mi) M2 serving the eastern neighborhoods and Copenhagen Airport. The M3 is a circle line connecting Copenhagen Central Stationwith Vesterbro, Frederiksberg, Nørrebro, Østerbro and Indre Bydistricts. The metro has 39 stations, 25 of which are underground. In 2019, the metro carried 79 million passengers.

The **Copenhagen S-train** is the S-train of Copenhagen, Denmark. It is a hybrid urban-suburban rail serving the Copenhagen urban area, with the notable exception of Amager. The average distance between stations is 2.0 km, shorter in the city core and inner boroughs, longer at the end of lines that serve suburbs. Of the 86 stations, 32 are located within the central ticket fare zones, 1 and 2. The S-tog is analogous to S-Bahn systems in Germany, and is a separate system from the Copenhagen Metro, which operates in the city centre, Frederiksberg, and Amag

S Train (alone) is used annually by more than 116 million people (2016 figure) whereas 220 thousands residents take Copenhagen metro daily to their work.

1.2 PROBLEM

Following decades of urban development, it is worthwhile to take a snapshot of the current development of areas surrounding metro and train stations and classify them into groups.

Some neighbourhoods are commonly known, especially for Copenhagen residents, as mostly residential, others have more business and commercial spaces surrounding them. The purpose of this work is to let data tell the story.

The venues closest to a station determine how people are using these spaces, e.g. if there are low counts of professional places in a neighbourhood then its residents are likely to commute to other areas for work. This creates daily movement of people.

1.3 INTEREST GROUP

By analysing this data we can classify stations by their primary usage. This data is useful for:

- City planners to tell how people are most likely to move around in the city for work or just to rewind. This also can help plan further extension of the transport network
- Someone looking to decide on their residence area. This analysis provides an insight on the blend of venues around each stations.

moreover,

- It helps businesses determine locations that best suits nature of their business.

- Curious souls like mine....

2 DATA ACQUISITION AND CLEANING

2.1 DATA USED

For this project, I need data on stations (train and metro) and on the venues around them.

- List of transport stations are available on Wikipedia page. The list includes information such as lines, grade, opening year, time, zone and transfer. Also, the page provide a table of future metro and train stations.
  - For metro network, it is [HERE].
  - For train network, it is [HERE].
- Foursquare Venue Category Hierarchy. Foursquare groups venues in high-level categories then those are broken down to sub-categories. This are available [HERE].
- Foursquare API to explore counts of venue categories surrounding each station.

I will be querying the number of venues in each category in a 1000m radius around each station. This radius was chosen because 1000m is a reasonable walking distance.

2.2 DATA PROCESSING

2.2.1 Transport Station

Using the read_html python function from Panda library, I could parse data from the 2 Wikipedia pages. I clean them from:

- Remove Special symbols (# and †) which were in the original tables.
- Replace Special Danish characters (ø, æ, å, é) with English letters.

I use also Google geocode service to acquire the latitude and longitude coordinates of the stations. Metro and train data frames have been cleaned, arranged, saved in .csv files then combined into one data frame.

2.2.2 Venue Category Hierarchy

After passing my foursquare credentials, I use Get Venues Categories request to recover the hierarchy and fill up the categories in a categories_list. There are 10 venue categories:

```
Arts & Entertainment (4d4b7104d754a06370d81259)
College & University (4d4b7105d754a06372d81259)
Event (4d4b7105d754a06373d81259)
Food (4d4b7105d754a06374d81259)
Nightlife Spot (4d4b7105d754a06376d81259)
Outdoors & Recreation (4d4b7105d754a06377d81259)
Professional & Other Places (4d4b7105d754a06375d81259)
Residence (4e67e38e036454776db1fb3a)
Shop & Service (4d4b7105d754a06378d81259)
Travel & Transport (4d4b7105d754a06379d81259)
```

2.2.3 Counts of Venues Categories

Using Get Venues Explore, I pass on the coordinates of the stations and get the counts of venues categories around each station. The work has been done over multiple days in order to respect the max daily qouta of FourSqaure API request.

I normalized the counts in order to enjoy better visualization and easier interpretation. This is done by scaling features to lie between zero and one, or so that the maximum absolute value of each feature is scaled to unit size. This can be achieved using **MinMaxScaler** available in sklearn library.

Data have been cleaned, arranged, saved in .csv files then combined into one data frame (stations_venues_df).

3 EXPLORATORY DATA ANALYSIS

Now that we got the data clean and ready to use, we can start plotting the stations on a map using Folium library.

Red for metro stations, green for future metro stations, blue for train stations, yellow for future train stations

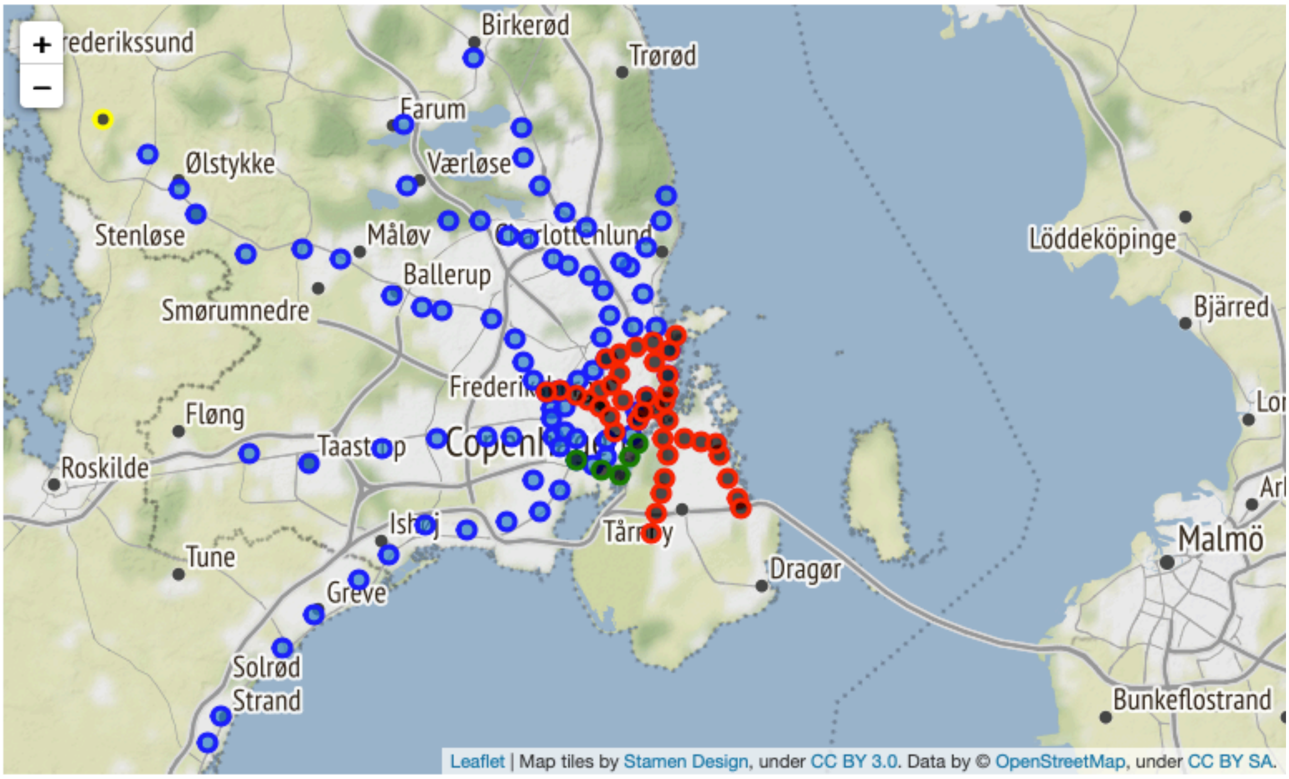


Figure 1 Stations on Copenhagen Map

Also, let's start venues exploring counts and what insights it could provide. I use sns library to visualize data, specifically, the boxplot function. It looks something like Figure 2.

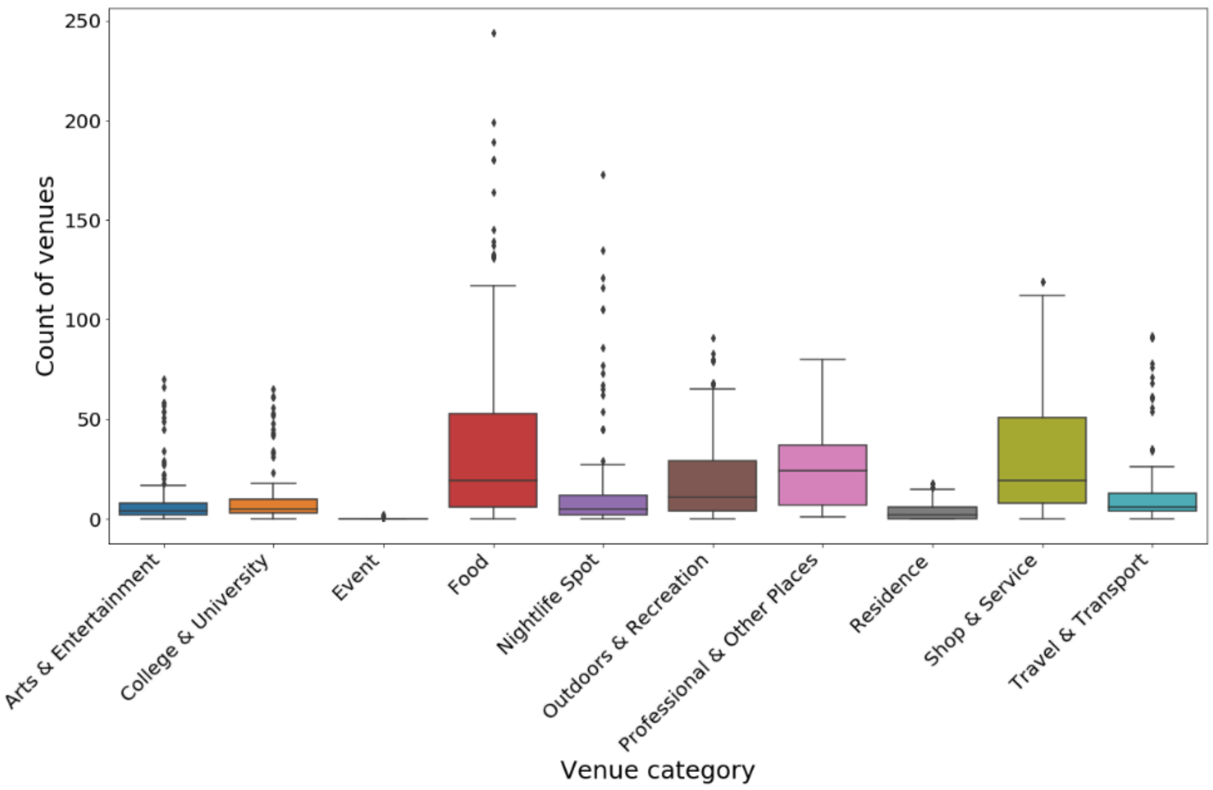


Figure 2 Boxplot of Venues Categories Counts

I observe that **Food** venues have the highest count variability as well as the largest outliers.

Noticeably, median for **Professional & Other Places** is the highest and it sets above medians of **Food and Shop&Service**.

Food, nightlife spot and Shop&Service have the most significant outliers counts among categories in few categories.

Food and Shop&Service's 75th percentile are the highest in the group and set slightly above 50 venues counts per category. The interquartile range (IQR) of these two categories are the widest.

Venues that are categorized as **Event** are comparatively little therefore I delete it from the set.

Let's now dig deeper and try to cluster the stations through their venue categories counts.

3.1 CLUSTERING STATIONS

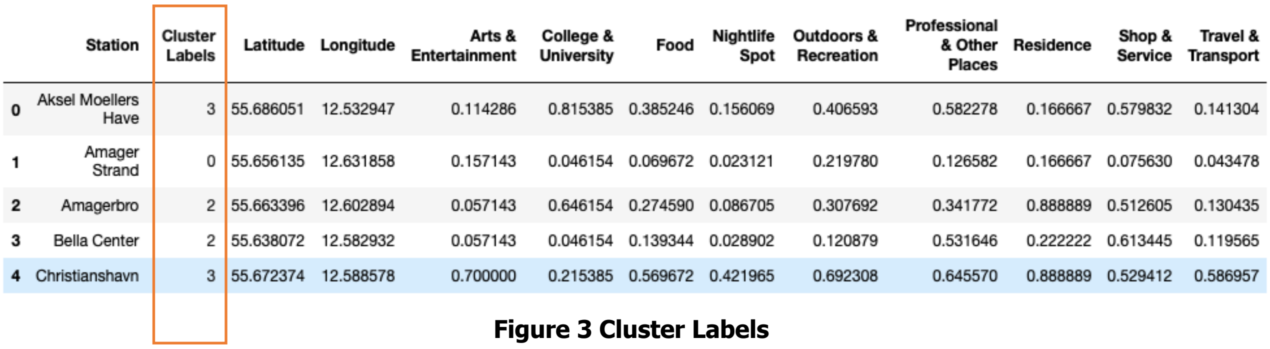
I used the K-means method to run through the now 9 categories (after dropping of the event category).

These were preliminary results with different number of clusters:

- 2 clusters only show the uptown/downtown divide
- clusters add clustering within the downtown
- clusters also identify neighborhoods with very low number of venues
- and more clusters are difficult to interpret

For the final analysis, let's settle on **4 clusters**.

Running the KMeans function available in Sklearn library, data frame is fitted and after normalizing the data. The head of the data frame looks like, Figure 3.



	Station	Cluster Labels	Latitude	Longitude	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Aksel Moellers Have	3	55.686051	12.532947	0.114286	0.815385	0.385246	0.156069	0.406593	0.582278	0.166667	0.579832	0.141304
1	Amager Strand	0	55.656135	12.631858	0.157143	0.046154	0.069672	0.023121	0.219780	0.126582	0.166667	0.075630	0.043478
2	Amagerbro	2	55.663396	12.602894	0.057143	0.646154	0.274590	0.086705	0.307692	0.341772	0.888889	0.512605	0.130435
3	Bella Center	2	55.638072	12.582932	0.057143	0.046154	0.139344	0.028902	0.120879	0.531646	0.222222	0.613445	0.119565
4	Christianshavn	3	55.672374	12.588578	0.700000	0.215385	0.569672	0.421965	0.692308	0.645570	0.888889	0.529412	0.586957

Figure 3 Cluster Labels

The clusters labels are [0, 1, 2, 3] to illustrates the 4 available clusters.

3.2 VISUALIZATION OF RESULTS

Now let’s have make a boxplot for each and stack them in lateral way that we could easily compare them.

Figure 4 provides this information.

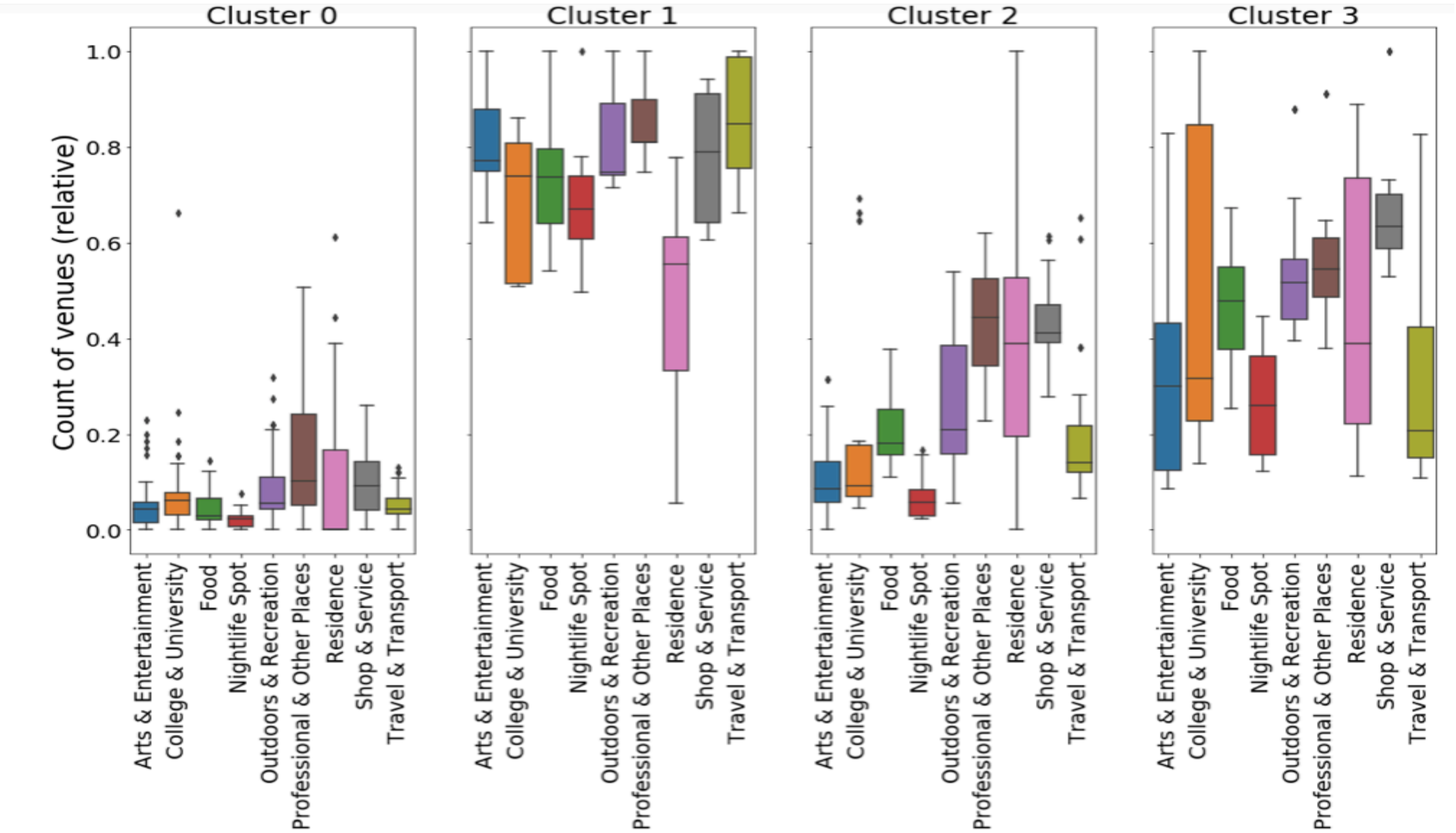
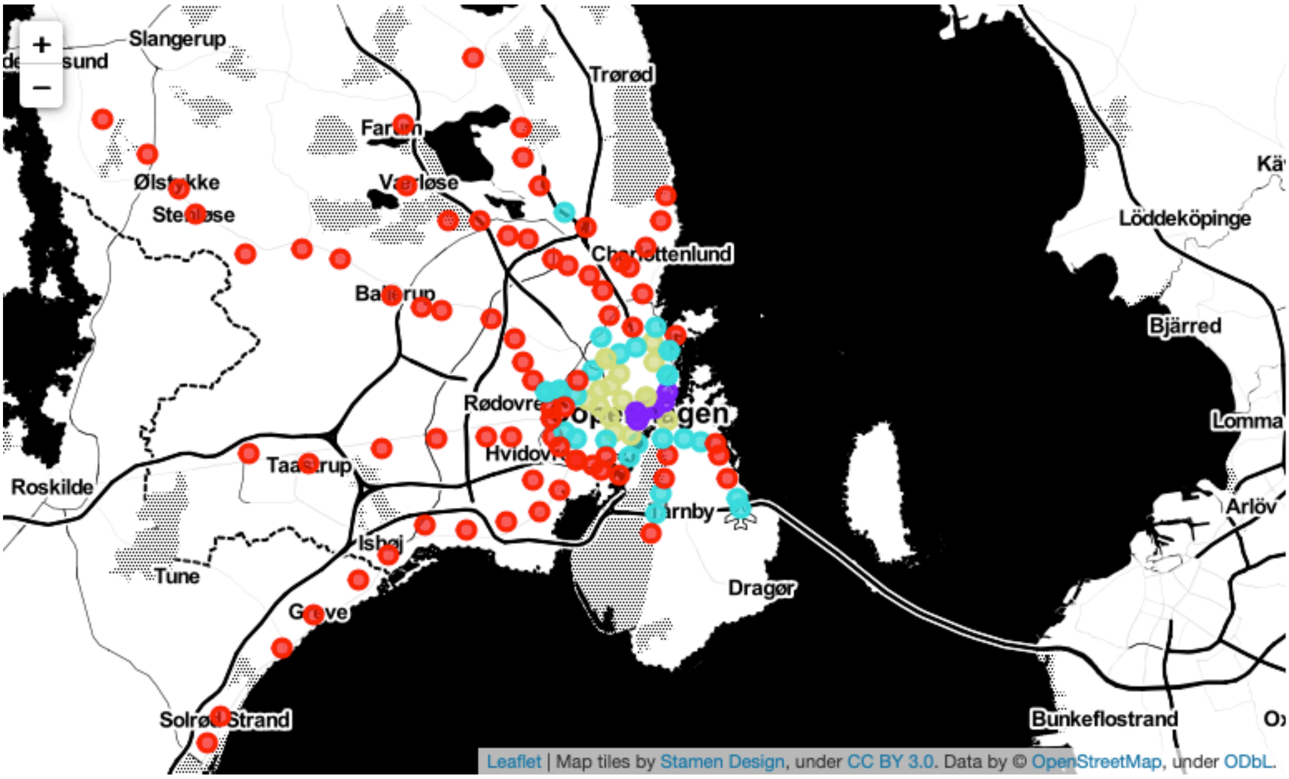


Figure 4 Boxplots of the 4 Clusters

Now we visualize these clusters, also, in a map, see Figure 5.



Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

Figure 5 Clusters Map

4 RESULTS

We can characterize the clusters by looking at venue scores

- Cluster 0 (Red) has the lowest marks across the board. This appears to be underdeveloped areas.
- Cluster 1 (Purple) has the highest scores for all venue categories. Noticeably most of their medians were above 0.6. This is the most diversely developed part of the city.
- Cluster 2 (Blue) has comparatively low counts. Most significant venues counts in this cluster are Professional & Other Places, Residence and Outdoor& Recreation.
- Cluster 3 (Green) has lower score than cluster 1, yet, it has the best scores in Residence and College & University.

Plotting the clusters on a map shows us that

- Cluster 0 areas tend to be at the outskirts of Copenhagen.
- Cluster 1 is the oldest central part of the city
- Cluster 2 and Cluster 3 are intermingled. They are also downtown surrounding cluster 1. Most of these stations have excellent transit accessibility and there are many colleges have their campuses around and small to medium size company.

We notice also, one station "Lyngby", north of the Copenhagen, which is out of the city center, and is part of cluster 2. Many companies choose Lyngby to locate its premises. Lyngby is also the important shopping destination in the northern suburbs.

5 DISCUSSION

I dive through the different neighbourhood surrounding metro and train stations in Copenhagen. Results and data confirms my perceptions of these neighbourhood and what’s commonly known for Copenhagen residents.

The particularity of each clusters and the distribution of venues categories could help me personally choosing a location for residence. At the current time, I have the following selection criteria:

- Residential area with variety of food and outdoor recreation outlets.
- And it provides options for entertainment and art.

Going through the clusters, my preference goes to neighbourhoods around the stations in cluster 3.

Yet, to be objective, Foursquare data isn’t all-encompassing. The highest number of venues are in the Food, Shop & Service categories and outdoor & recreations.

Also, data don’t take into account a venue’s capacity, popularity (e.g. a university building attracts a more people than a hot dog stand – each of them is still one Foursquare “venue”).

6 FUTURE DIRECTIONS

Future directions:

- Results could be combined with other analysis (e.g. demographics, venues popularity/ capacity) to provide more accurate insights
- I excluded the bus station network from my study and this is definitively an area to further develop this work.

7 CONCLUSIONS

Foursquare data is limited but can provide insights into a city’s development. Results could be combined with other analysis (e.g. demographics, venues popularity/capacity) to provide more accurate insights.