SMART INVESTING IN AIRBNB

Coursera Data Science Capstone Project

Patrik Antal

GOALS

The aim of this project is to find the best Airbnb apartment as an investment. For this purpose I want to demonstrate, how to make smart decisions along the way by bringing data into the equation.

HOW WE GET THERE

CITY

Which city to buy the estate that is profitable and fits our purposes?

DISTRICT

What is the idea neighbourhood to buy an estate for Airbnb purposes?

PART1: FINDING THE CITY

PROBLEMS

a) Do I have any personal preferences?

b) What is my budget?

c) What defines a good investment?

d) Where can I find those metrics?

SOLUTIONS

a) Do I have any personal preferences?

Let’s say we have an idea to visit the place in every summer and also check the estate’s condition if it is necessary. Therefore, it would be idea to buy an estate not far away from the home country, Hungary.

b) What is my budget?

We have an initial capital of 450.000 US$.

c) What defines a good investment?

A generally good indicator of whether to buy your own estate or rent one is price-to-rent-ratio. It is calculated as follows:

$$price - to - rent\ ratio = \frac{median\ home\ price}{median\ annual\ rent}$$

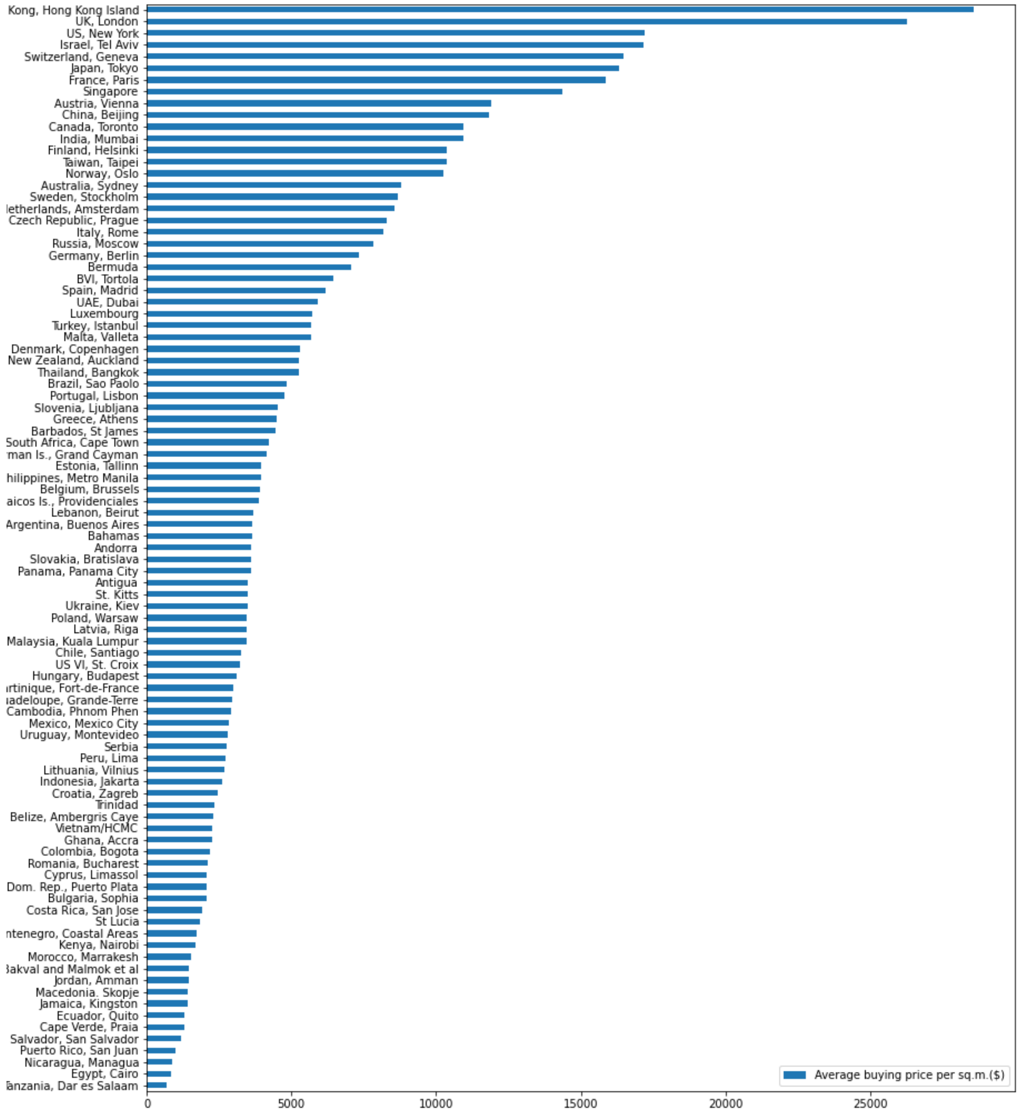
This metric is used for describing the state of the housing market. As growing ratio cloud be a sign that the estates are not fairly valued, consequently the market is like an expanding bubble. As we don’t want to buy an overpriced property, we use investors rule of thumb. Buy if it’s below 21 and rent otherwise [1].

d) What defines a good investment?

Find online available resources for metrics. The GlobalPropertyGuide page gave a good summary of important metrics.

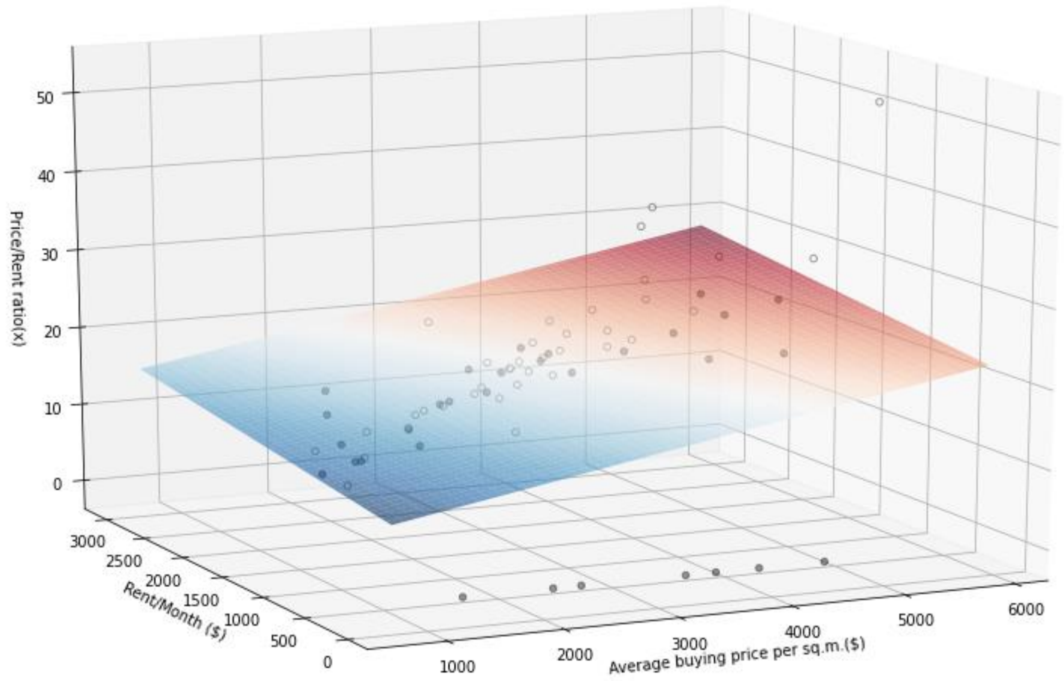
MATERIALS & METHODS

- The housing information from 93 big cities all around the world were available from the previously cited website. I have built a web-scraping algorithm to extract the information from the online table. Some of the important metrics are: average buying price per sq.m, price-to-rent ratio and rent/month.
- As we can see the average estate prices vary on a wide scale. Since we have a limited budget, considering a liveable 75 sq.m apartments the upper limit is 6000 US($)/sq.m.
- Limiting the price range we had 68 cities left.



MATERIALS & METHODS

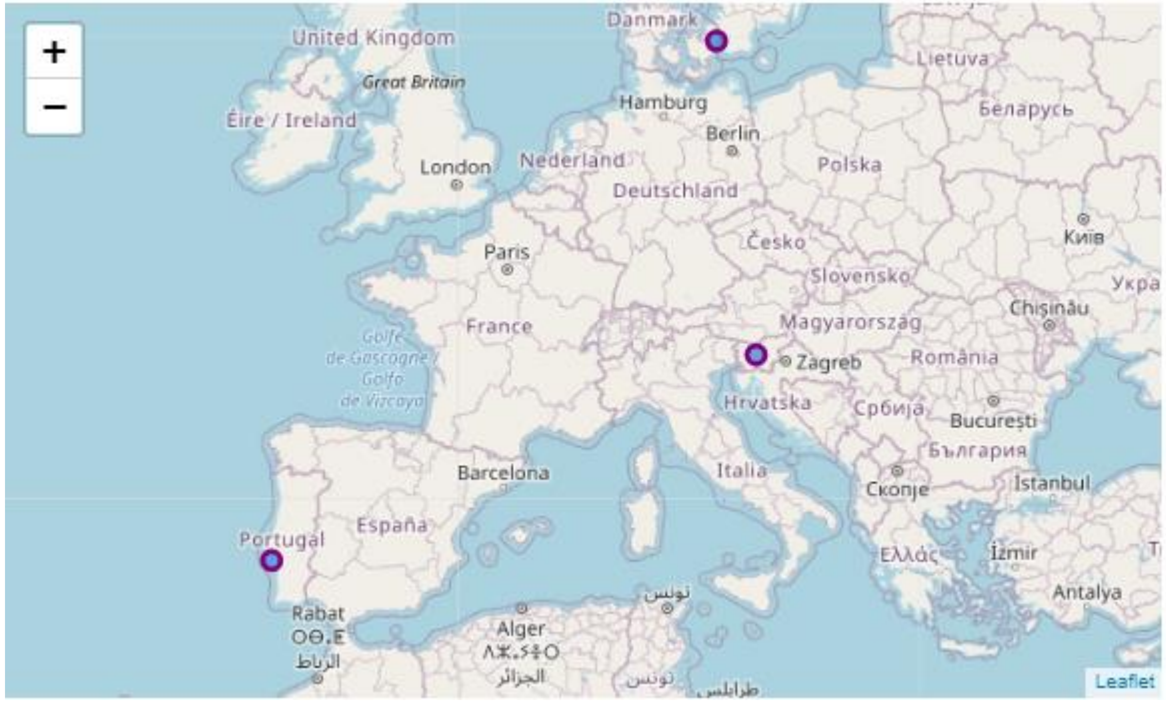
- Next we should consider cities that have lower price/rent ratio than 21 which was the upper limit, till this value it is affordable to invest.
- From the picture we can see that expect some outliers, most values fall into this range.
- We can also discover a linearity between the price and price-to-rent ratio. Meaning the lower the buying price the more stable the market is.
- If we want to make this investment a monthly income, we should be expecting a minimum sum per month. This has been set to the mean, which was 1500 US$.
- Limiting the monthly income and the price-to-rent ratio we had 20 cities left.



MATERIALS & METHODS

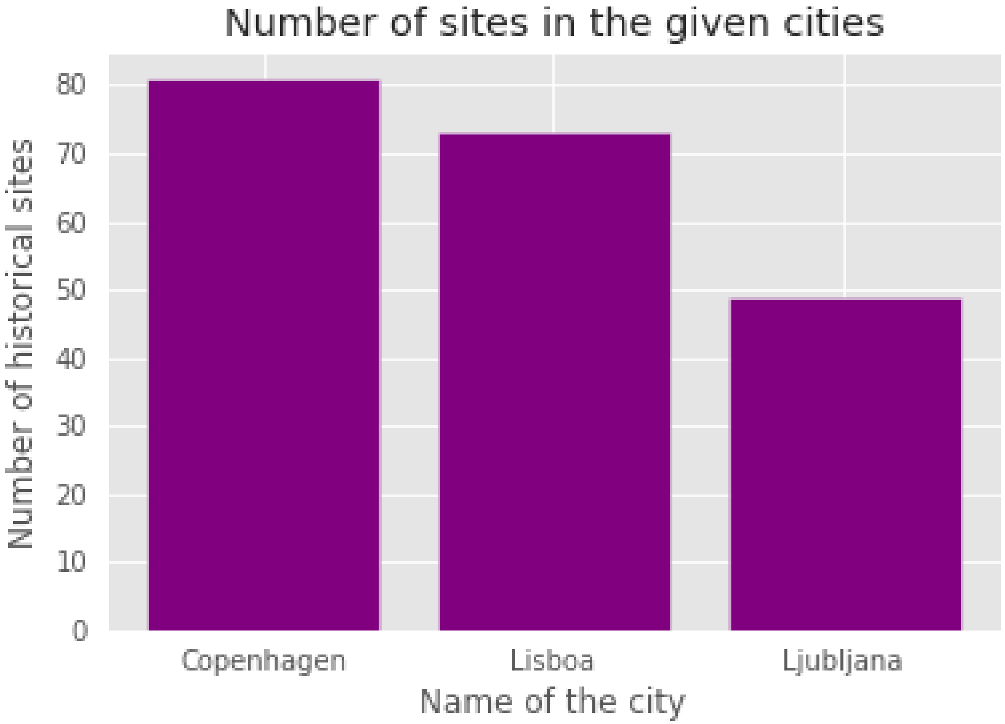
- Now we can concentrate to the last detail, that the estate should be near the home country, namely Hungary. So I have set the boundaries to places in Europe that is further west from Hungary.
- To make this automatically (in case in the future I will have different preferences, or the input data will be more extended), I have set the coordinates of the uttermost places in the boundary box in every directions.
- In the next step I have determined the coordinates of the target cities, and labelled them according to their presence in the boundary box.
- To get the coordinates I have used reverse geocoding implemented by the GeoPy library.
- From the 20 cities only 3 were in the boundary box.
- So we have 3 cities that is in Western Europe, having a good estate market, and fitting in our budget.
- Because our target group is people who are staying for shorter periods mainly for sightseeing, another important aspect is to have historical sites that attracts them.

- Visualize the 3 cities using the Folium library.



MATERIALS & METHODS

- To find the most historical city I have used Forsquare API. This gave me the cultural venues regarding a specified area.
- I’ve chosen this searching area to be around the city center in a 3 km radius.
- According to the evaluation the city having the most sites in the center is **Copenhagen**.
- To sum up:
  - using various techniques, we could find the ideal city which is a good financial choice and fits our personal needs
  - decisions were backed up with data
  - along the way we used web scraping, reverse geocoding, geolocating and location-based search-and-discovery API



PART2: CHOOSING THE RIGHT NEIGHBOURHOOD

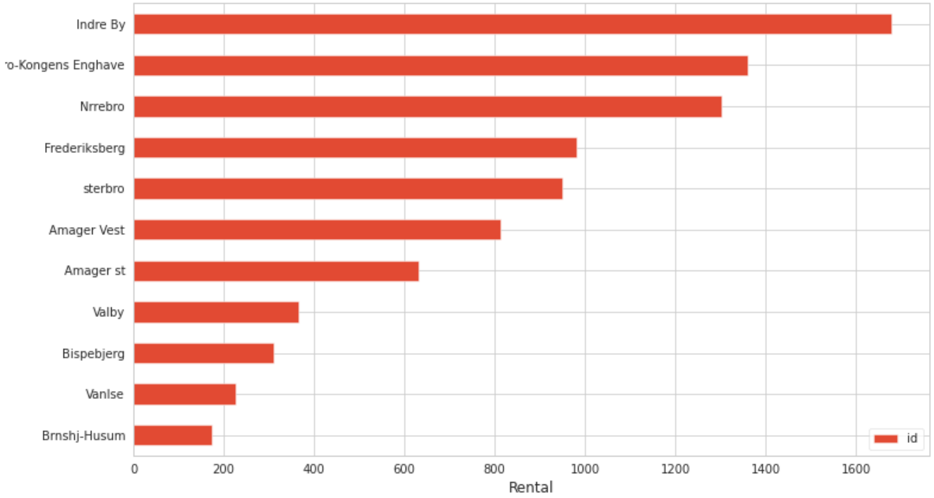
PROBLEMS

a) What defines a good neighbourhood?

b) How can I distinguish between neighbourhoods?

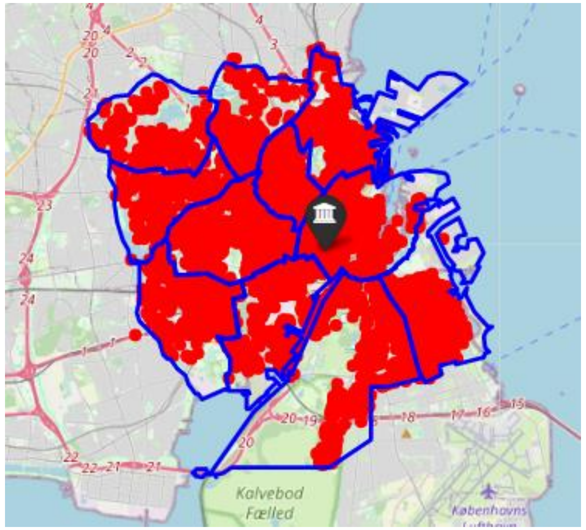
MATERIALS & METHODS

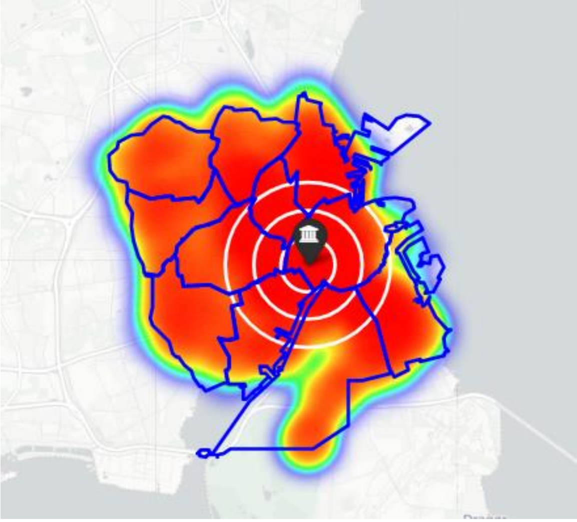
- On the insideAirbnb site there are detailed and fresh datasets related to Airbnb usage in a specific city.
- Obtaining this dataset I had access to all listings in Copenhagen that was available in February of 2021. This meant 8785 unique offers.
- First there was a need for cleaning the data.
- Using this dataset I could easily extract the listing distribution.
- As we can see there are 11 districts in Copenhagen with a great variance in rentals available.
- The highest Airbnb activity can be detected in Indre By, Vesterbro-Kongens Enghave and Nerrbro.



MATERIALS & METHODS

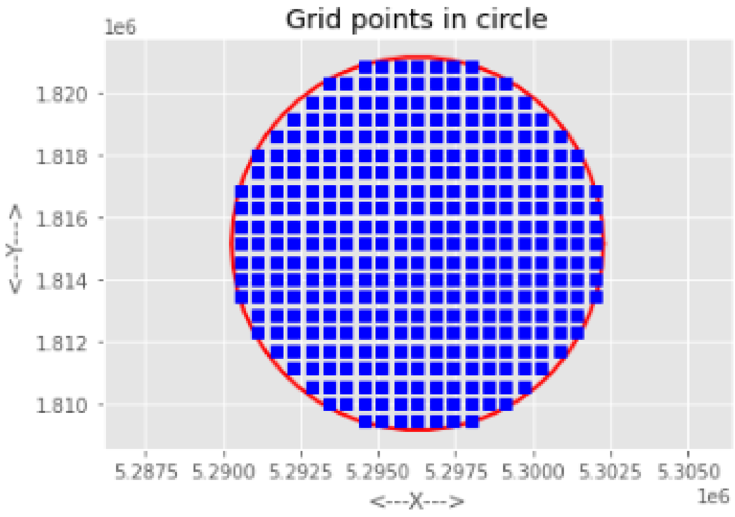
- Using the insideAirbnb I could extract the longitude and latitude coordinates of all listings.
- Picking Copenhagen’s City Hall the city center (marked with the museum sign) I can place all the listings on the map.
- Downloading the GeoJSON file of Copenhagen we can also visualize the neighbourhood borders.
- Using the combination of the 2 maps we can identify potentially good regions.
- It would be idea to find places within the 3 km radius from the center (can be seen on the 2nd map), where there is no great competition.

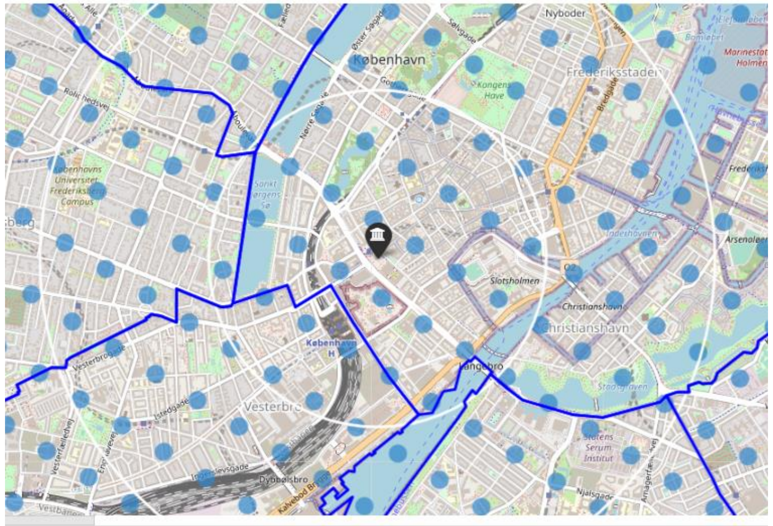




MATERIALS & METHODS

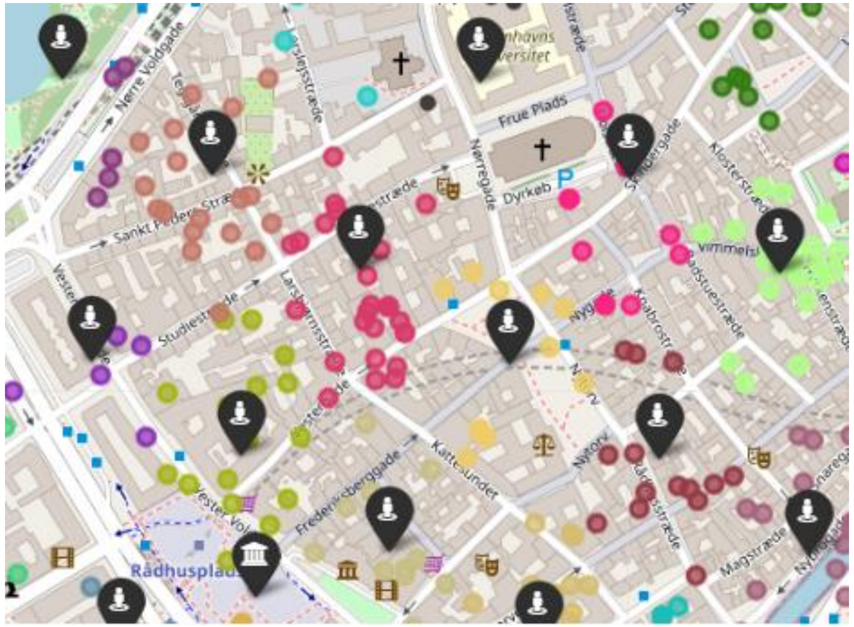
- To better explore this region I have created an equally spaced grid on the center.
- The trick here is to be able to convert locations from WGS84 spherical coordinates (latitude and longitude degrees) to UTM Cartesian coordinate system (X/Y coordinates in meter)
- The equally spaced grid and its overlap with the map can be seen on the images.
- There has been 364 grid points generated within the 3 km radius of the city center.
- Using the Airbnb dataset we can assign each listing to the closest neighbourhood center.
- There were only 2638 listings in the central region.

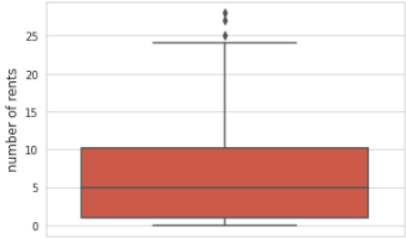


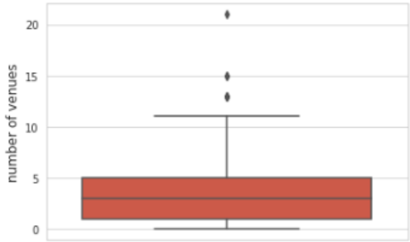


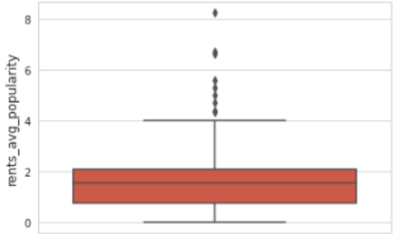
MATERIALS & METHODS

- So now we have 364 neighbourhoods with listings assigned to them.
- The map on the right shows the different neighbourhood centres (with the man icon) and the listings belonging to this center(using different colors)
- Th goal is to find neighbourhoods that are used for Airbnb purposes, but the local market is not full yet.
- Consequently calculate the 90th percentile of all rentals in a neighbourhood and drop anything above it.
- To get to know the Airbnb guest's feedbacks I have assigned a popularity score to every neighbourhood. This is a normalized score from number of reviews per month and the rating of the listing. Based on this metrics I want a neighbourhood that is busy and beloved so limited the range to the top 30 percentile.
- As the original idea was to attract tourists with interest in the historical city, I will measure this aspect by the number of historical venues in every area. The above the median values was kept.
- This 3 aspects has limited the original 364 neighbourhoods to only 61.



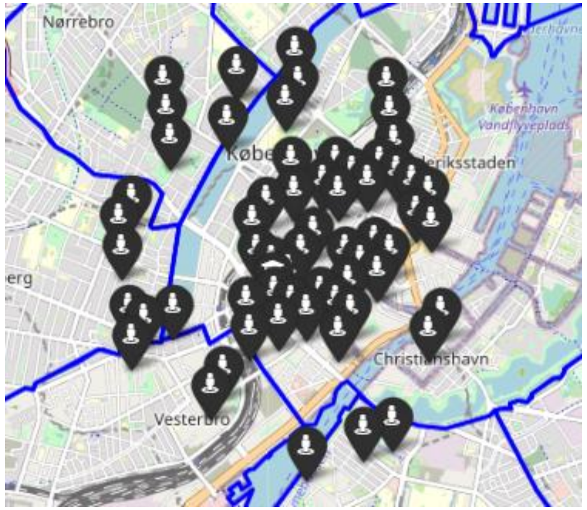


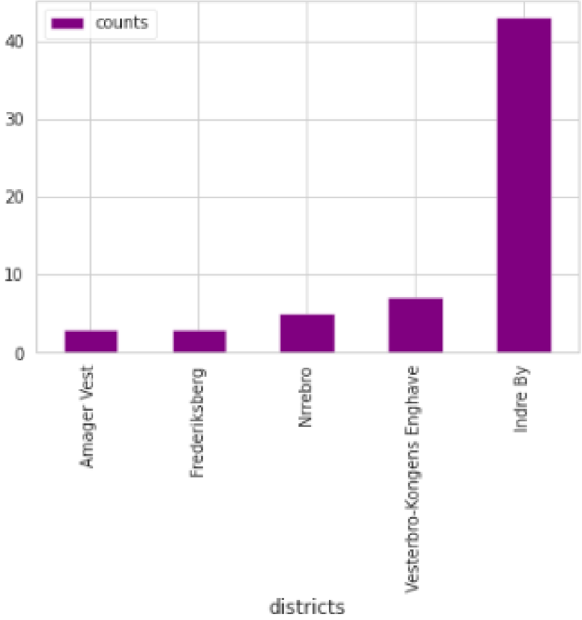




MATERIALS & METHODS

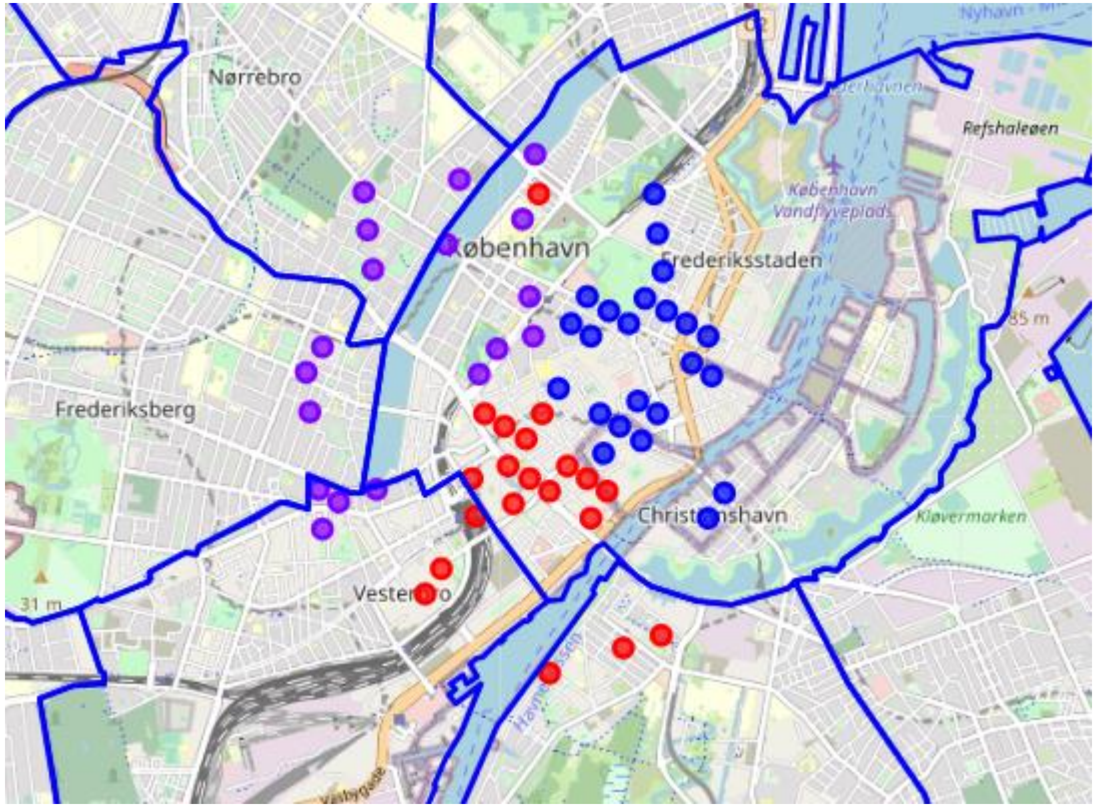
- The popular, historical, but not too competitive district centers can be seen on the map. It is clear that the closer we are to the center the more the area suits our needs.
- As we can see on the histogram, not surprisingly, these regions are mostly in Indre By literally meaning Inner City. But some outer regions are promising as well, like Vesterbro, Nerrbro and Frederiksberg. Which can come handy if we consider prices.





MATERIALS & METHODS

- For further narrowing down the region, I have performed a cluster analysis considering all the venues in those neighbourhoods, using the Forsquare API.
- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).
- For our case, the similarity is based on the venues of the neighbourhood.
- One common algorithm to perform a cluster analysis is K-Means. This starts with random centroid points for a given number of clusters. To state, there are methods for finding the idea number of clusters, but to get regions that are easily decriable I've chosen 3. The algorithm repeatedly adjusting these centroids based on the Euclidean distances between points.
- As we can see expect one point all the others are separated geographically.



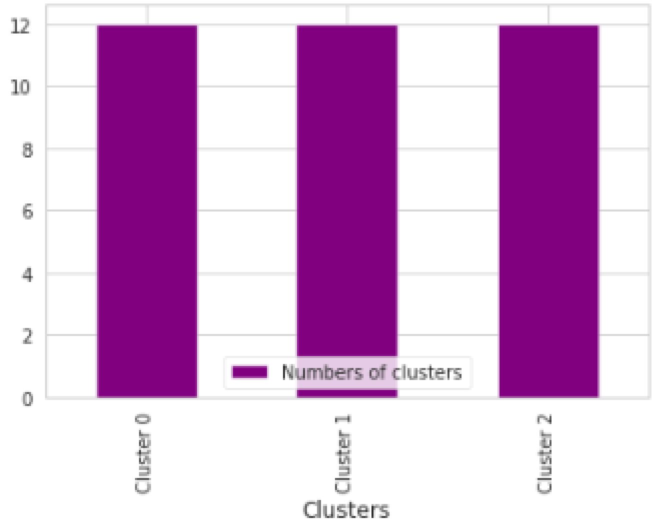
🟪 Cluster 0

🟦 Cluster 1

🟥 Cluster 2

MATERIALS & METHODS

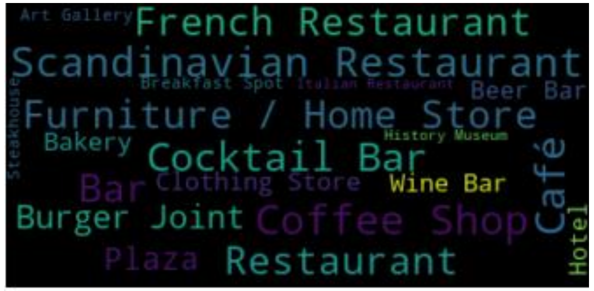
- Each clusters has 12 different neighbourhoods in them.
- For further exploration of the clusters I have used 2 different approaches.
- First I've determined the venues that are most frequent in each region. (kept the top 20 of them)
- Than I have gathered the most frequent venues that are only present in one neighbourhood.



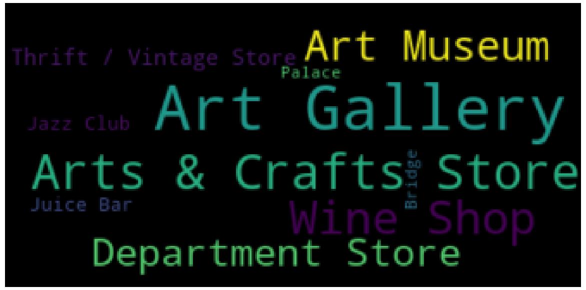
RESULTS

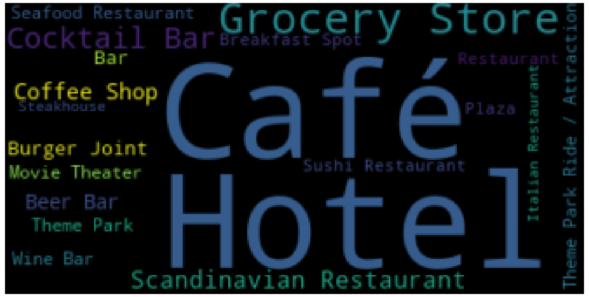
MOST FREQUENT

MOST FREQUENT UNIQUE

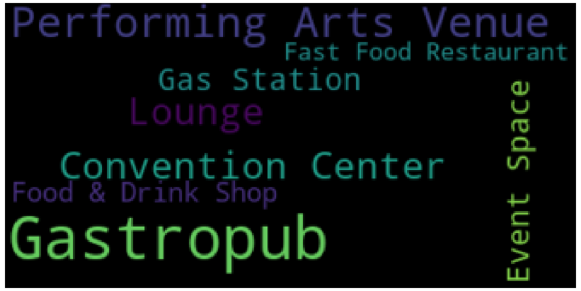


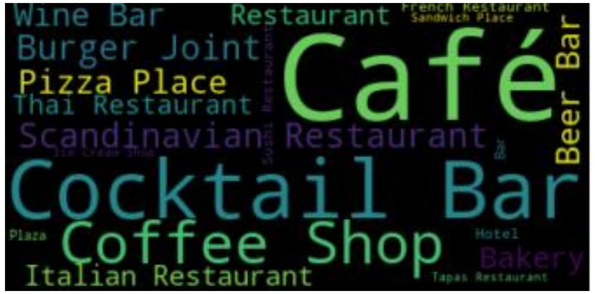
Cluster 0



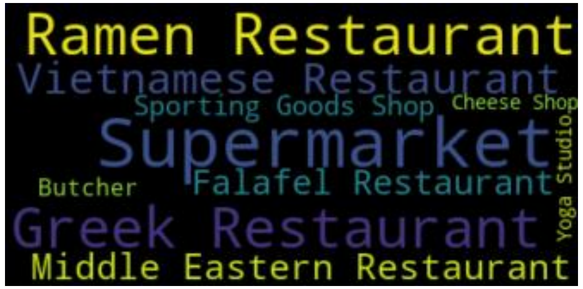


Cluster 1





Cluster 2



EVALUATION

- **Cluster 0:**
  - This area is in the heart of the city, fully placed at Indre By.
  - We can see this is a more traditional part of the city.
  - There are several classical coffee shops and restaurants in this region.
  - Also there is a great variety of shopping possibilities.
  - But the most characteristic in this region are cultural venues (art museums, galleries, etc.)
- **Cluster 1:**
  - This area is full of entertainment opportunities and accommodations.
  - Possible target for night-out programmes(cocktail/wine/beer bar, cinema) and also daylight activities (theme park (Tivoli), coffee shops).
  - It is clear that the target age group for this region is quite low (basically children, youngsters , young adults).
  - Differences to other clusters highlights that this region is for the open-minded Z/Y generation, with gastropubs, event spaces and performing art venues.
- **Cluster 2:**
  - This region is all about the gourmets. Great variety of different restaurants are available in this area (ranging from Vietnamese to Greek).
  - We can see that this is a multicultural region of the city. This is not only just suggested by the different kind of restaurants, but also the other venues (yoga studios, cheese shops).
- **Conclusions:**
  - For our target group there is no doubt that cluster number 0 is the way to go. It is full of cultural and historical venues and provides great shopping possibilities.
  - If we were expecting families or younger adults as guests cluster 1 would be the idea choice with huge variety of entertainment possibilities.
  - Custer 2 is a very vivid multicultural region of the city, that would be suitable for open-minded people who want to sense the true life of a metropolitan.

FINAL CONCLUSION

I have successfully limited the target investment region to **one city with only 12 neighbourhoods**. For this I have used web scraping, geographical data manipulation, several statistic methods and unsupervised machine learning (K-means clustering). Every decision was directed by data and personal preferences. Extending the project it is possible to create a program capable of doing the analysis with arbitrary cities and preferences for which sufficient data is available.

> ”ONE CITY WITH 12 NEIGHBOURHOODS”