Capstone Project - The Battle of Neighborhoods

Exploring where to open a bike shop in Copenhagen

By Michele Deluchi (michele.deluchi@gmail.com)

Table of contents

Project notebook can be publicly accessed at:
https://nbviewer.jupyter.org/github/MicheleDeluchi/CopenhagenBikeshopClustering/blob/master/CPHClusteringFinal2.ipynb

Introduction: Background

Data obtained through publicly accessible statistics provided by the Danish Cycling Embassy, Copenhagen Kommune, Frederiksberg Kommune, Statistics Denmark and various other official sources.

Copenhagen's cycling culture

Cycling in Copenhagen is – as with most cycling in Denmark – an important mean of transportation and a dominating feature of the cityscape, often noticed by visitors. The city offers a variety of favorable cycling conditions — dense urban proximities, short distances and flat terrain — along with an extensive and well-designed system of cycle tracks. This has earned it a reputation as one of the most—possibly the most—bicycle-friendly city in the world.

Every day 1.2 million kilometers (0.75 million miles) are cycled in Copenhagen, with 62% of all citizens commuting to work, school or university by bicycle; in fact, almost as many people commute by bicycle in greater Copenhagen as do those who cycle to work in the entire United States. Cycling is generally perceived as a healthier, more environmentally friendly, cheaper, and often quicker way to get around town than by public transport or car.

The Danish Cycle market is expected to expand

In the private sector there are 289 bicycle shops and wholesale dealers in greater Copenhagen, as well as 20 companies that design and sell bicycles, mainly the city's signature cargo bikes, such as Christiania Bikes (Boxcycles in the U.S.), Nihola and Larry vs Harry, and luxury bike brands as Biomega and Velorbis. These firms generate 650 full-time jobs and a total estimated annual turnover of DKK 1.3 billion (US$222 million).

Also, with the creation of cycle superhighways (cycling routes connecting different cities in the Zealand region with each other) and the advent of e-bikes in the mass market (in just one year the number sold and produced has gone up from 2,300 in 2017 to around 3,000 in 2018 – an increase of around 27 percent), the overall market size of the Danish bicycle industry is expected to grow throughout the next 5 years.

Introduction: Business Problem

Overall, the brief introduction contained in the previous section laid the foundations for the definition of the business problem that this project aims to investigate. In synthesis:

- The cycle market is closely woven into Danish cultural fabric. Cycling indeed represents a crucial resource for Danes to move within urban and rural landscapes.
- The market is forecasted to expand throughout next years, in virtue of A) creation of new cycling infrastructures, and B) a growing demand for e-cycles.
- The competitive landscape for bike shops offering bike sales & repairs within the city of Copenhagen appears to be already densely inhabited.

Thus, these considerations lead us to the overarching problem statement, reportedly:

To capture a (as big as possible) share of the expanding cycle market, where should someone establish a bike shop in Copenhagen?

Data Extraction and Wrangling

Data sources

Based on the business problem definition, the variables that would have driven my conclusions were:

- **number of and distance** to **bike shops** in the neighborhood, if any
- **distance** of neighborhood from the **closest bike trail**
- **distance** of neighborhood from the **city center**

To define neighborhoods, I decided to use all postal codes from the Greater City of Copenhagen.

Following data sources were used to extract/generate the required data:

- postal codes of the Greater Copenhagen Region have been scraped from **state registries** (can be found at: https://www.regionh.dk/english/about-the-capital-region/facts-about-the-region/PublishingImages/PostalcodesEnglish.pdf) and pre-processed to select only the areas included in the 'City of Copenhagen'
- missing addresses were obtained via the **Bing API** reverse geocoding feature.
- upon analysis, candidate areas were algorithmically defined via **gridding**
- centers of candidate areas were generated algorithmically and approximate addresses of centers of those areas were obtained using **ArcGIS and google geocoder APIs**
- number of bikeshops and their type and location in every neighborhood were obtained using **Foursquare API**, along with bike trail coordinates
- coordinate of Copenhagen center were obtained using **ArcGIS and google geocoder APIs** of well-known and central Copenhagen location (Kongens Nytorv square)

Data cleaning

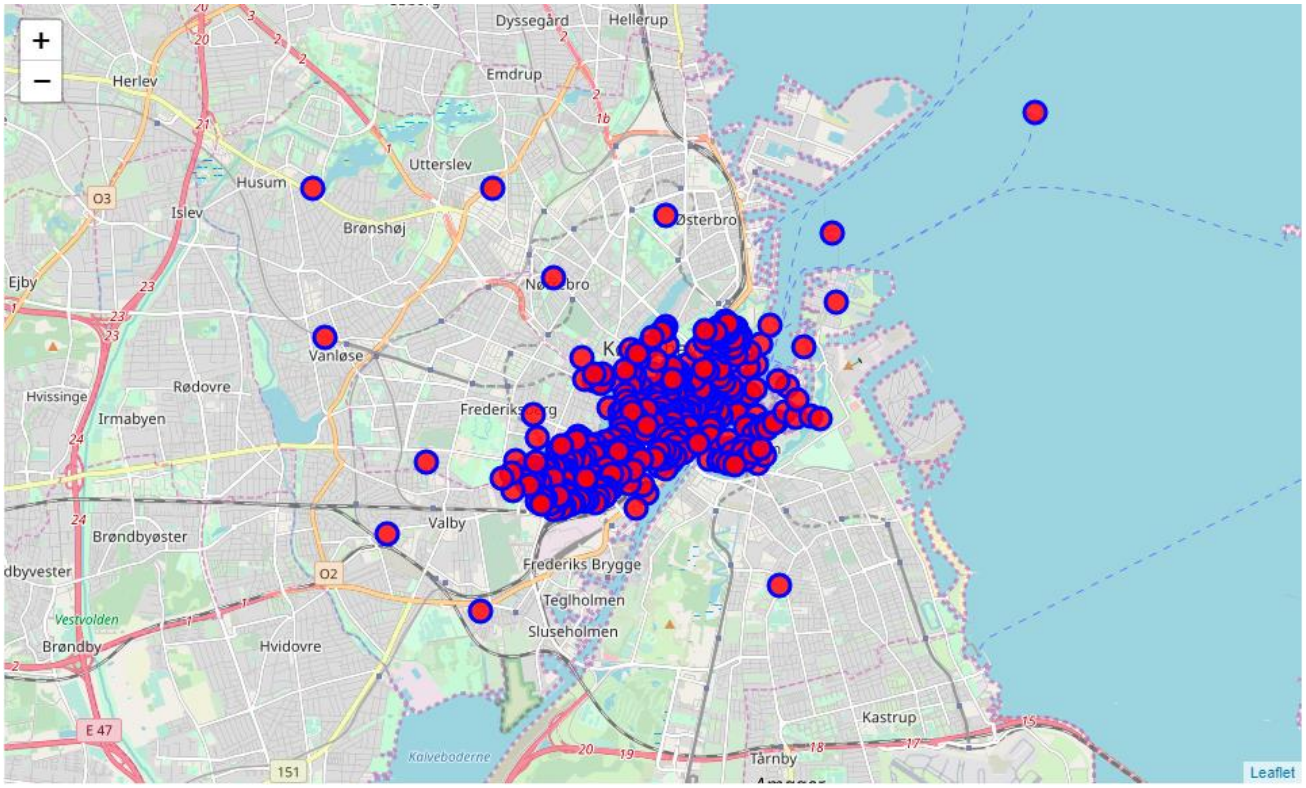
Postal code data scraped from the state registry were stored in a pandas dataframe. However, upon closer inspection, it became evident that:

1. Missing (Nan) entries were common.
2. A series of postal codes were associated to a postbox (literally, a tin box where mail is dropped), rather than to an actual address.
3. A series of postal codes presented duplicate address values.

Thus, to ensure the consistency of source data, the following operations were performed:

1. Rows with postal codes associated to postboxes were dropped.
2. NaN values were replaced with the corresponding address via the ArcGIS and Bing APIs. Respectively, ArcGIS API was used to associate lat/lon coordinates to each postal code, whether NaN or not. Then, Bing API was used to reverse geocode addresses based on the coordinates.
3. Obtained addresses were appended and duplicates were dropped based on the repetition of their respective latitude and longitude.

As the final step of the data cleaning process, the addresses were visualized on a folium map based on their coordinates.



Data selection

From the visualization of the addresses, however, it became evident that postal codes are not evenly distributed within Copenhagen. This could have been quite troublesome to extract venues with Foursquare API from the single dataframe, due to the static radius parameter (either values would have been missing due to a small radius, or there would have been a lot of duplicates due to a large radius). Thus, the dataframe was split per granularity of postcodes per area, leading to a division between inner CPH vs. outer CPH.

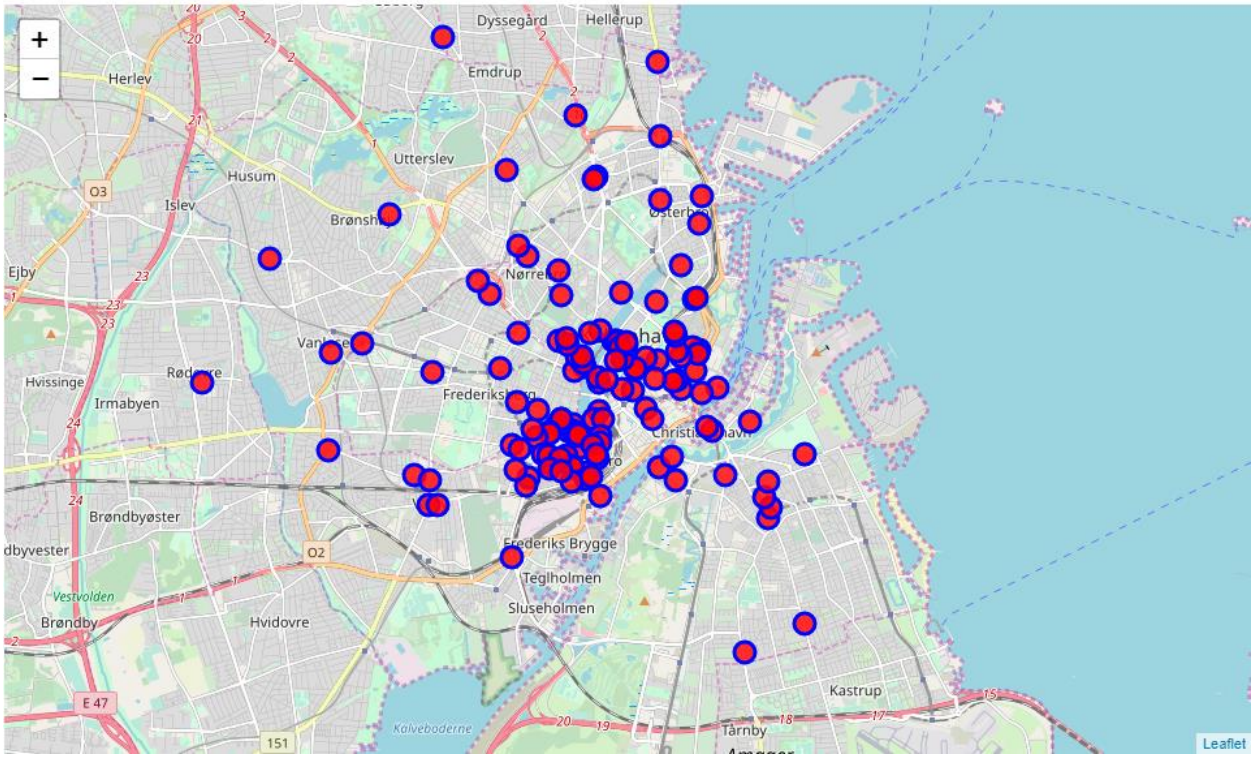
Thus, data was extracted through Foursquare API over two distinct dataframes, and the results were ultimately appended within a single dataframe.

Data extraction

Firstly, inner CPH venues were extracted. 83 unique bikeshops within inner Copenhagen were found. Then, the process was repeated for the Outer Copenhagen area. This time, though, to be sure to capture all locations, radius was increased to 3000m. As a result, 86 unique bikeshop venues were found. Results were then appended, and duplicates dropped, leading to a final selection of 142 unique bikeshop venues.

The process immediately highlighted the technical limitations of the API, as only 142 venues out of the 289 listed on national registries were found. This point will be discussed more in depth in the Limitations section.

Then, geographical distribution of bikeshops was visualized in a folium map as a form of exploratory analysis.



Methodology

Upon completion of the extraction and wrangling of data, the methodology for carrying out the analysis was established.

Firstly, it was decided that the project would direct efforts on detecting areas of Copenhagen with low bikeshop density, as a way to avoid high-competition neighborhoods.

For this purpose, required data was collected in the previous step.

Then, the second step was decided to be the exploration of '**bikeshop density**' across different areas of Copenhagen – **heatmaps were used** to identify few promising areas close to center with a low concentration of bikeshops.

The third and final step was defined as a deep-dive analysis of the most promising areas.

Within promising neighborhoods, **clusters of locations were created** though **k-means algorithm,** and then addresses were skimmed based on a neighborhood's ability **to meet some basic requirements**:

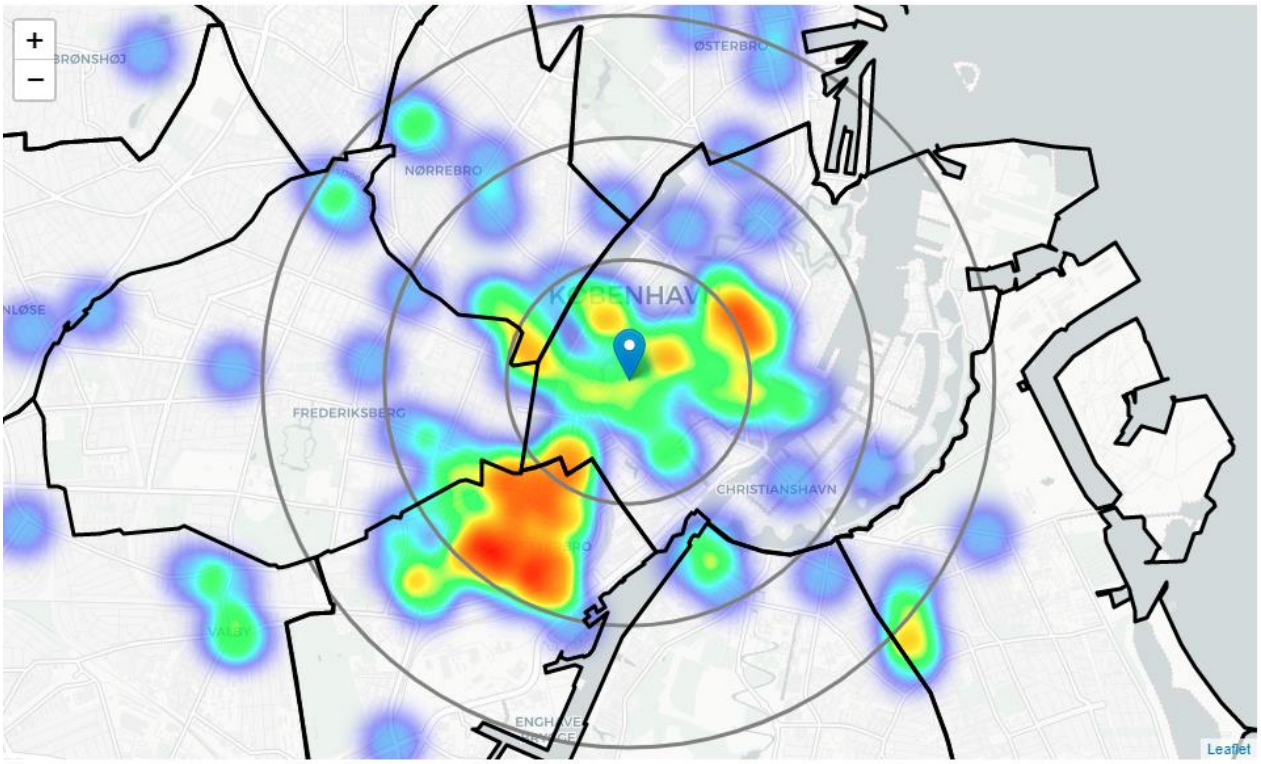
1. **No competing bikeshop within a radius of 250 meters**
2. **Being as close as possible to one of Copenhagen's main bike trails**

In relation to point 2, bike trails coordinates were extracted through Foursquare API.

Analysis

Visualization of bikeshop concentration: heatmaps

To create a visualization of bikeshop concentration, a heatmap was created by superimposing a delimitation of the boroughs of Copenhagen (obtained at https://raw.githubusercontent.com/codeforamerica/click_that_hood/master/public/data/copenhagen.geojson) on a folium heatmap, in which features were added to also visualize circles defining areas within 1, 2 and 3 kms of radius from Copenhagen city center.



From the heatmap, it emerged that most bikeshops in Copenhagen are concentrated in the immediate center of the city (Copenhagen K) and in the upper part of neighboring district in the South-West (Copenhagen V, or Vesterbro). Interestingly, this leaves some competitive space for the opening of new stores within some of the other neighboring districts. In particular, Frederiksberg and Nørrebro have a consistent amount of space comprised within a 2km radius from Kongens Nytorv, and also represent some of the most densely populated areas of the Greater Copenhagen zone. But are such areas actually attractive for opening a bikeshop? To answer this question, desk research was conducted to either confirm or reject the validity of the preliminary findings.

Scope restriction – desk research

Frederiksberg

Frederiksberg is an affluent area, with large parks such as Søndermarken and Frederiksberg Have, as well as a number of educational institutions such as Copenhagen Business School (CBS), Technical Education Copenhagen (TEC), the University of Copenhagen, and the Royal Danish Academy of Music. Furthermore, there are extensive

and vibrant cultural attractions in Frederiksberg, with theatre, events and concert venues such as Aveny-T, Riddersalen, Betty Nansen Teatret, Cisternerne, Forum and KU.BE.

All in all, Frederiksberg is characterized by a vibrant student community, constituting a significant portion of the 70% of the Frederiksberg population that every day commutes to either work or education.

Norrebrø

Nørrebro is a hip, multicultural neighborhood, popular with students and creative types. Kebab joints and indie shops line the main road, Nørrebrogade, and late-night bars are tucked into the side streets. Foodies head to the high-end eateries and trendy coffee spots on Jægersborggade. Nearby, the leafy paths of Assistens Cemetery wind past the graves of such notables as Hans Christian Andersen and Søren Kierkegaard.

In 2016, Copenhagen had 13,100 more bikes than cars and it is said Nørrebrogade is the busiest cycling street in Europe.

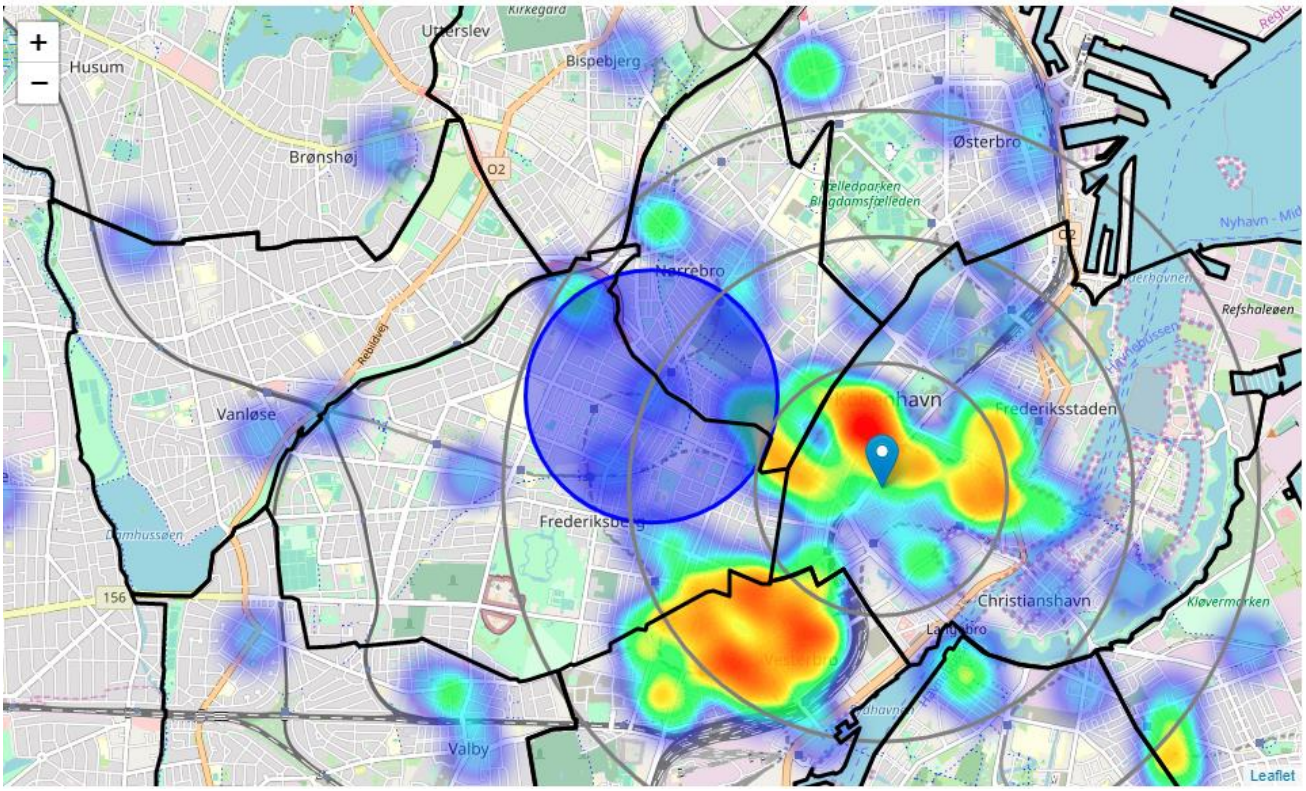
So what?

Popular with tourists, relatively close to city center and well connected for bikers, those boroughs appear to justify further analysis.

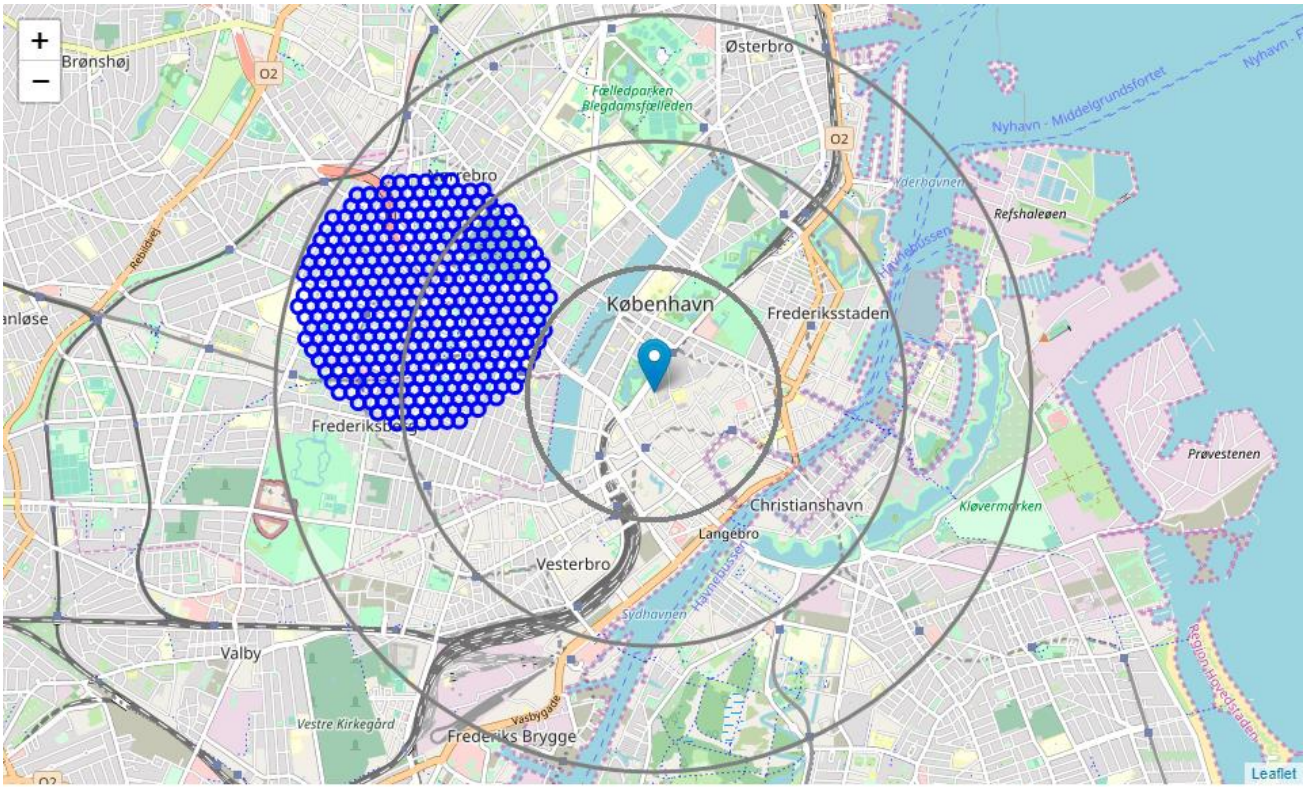
New, narrower region of interest were defined, including eventual low-bikeshop-count parts of Frederiksberg and Norrebrø. To do this, a function for converting lat/lon coordinates to X/Y Cartesian coordinates was defined, as they were going to be required for the creation of the grid segmenting candidate neighborhoods.

Candidate area selection, gridding, clustering and address identification

Restricted scope for neighborhoods of interest was defined and visualized in folium.



Then, the scope was segmented into a grid of candidate areas, nested adjacently in intervals of 100 meters (candidate area radius=100 meters). This led to a further visualization, describing the distribution of the 365 candidate sub-areas

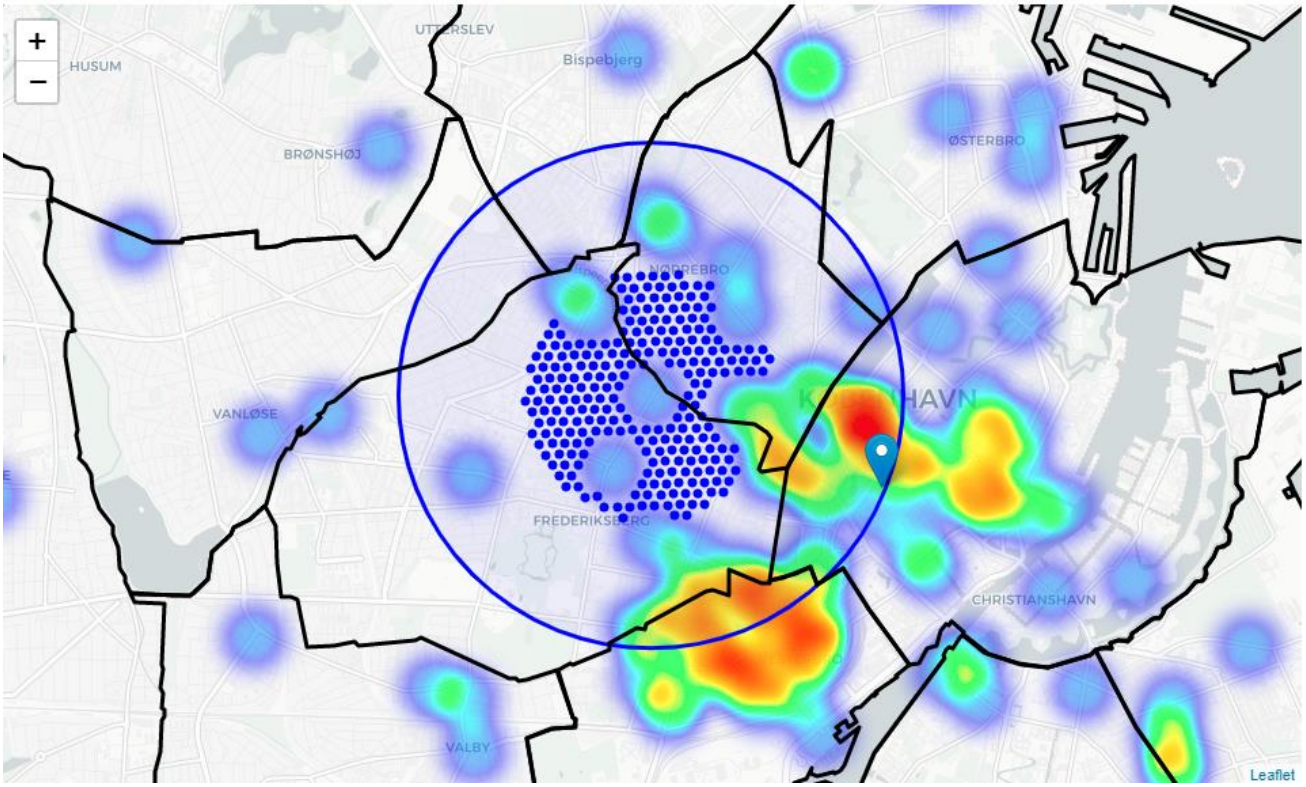


Then, locations were filtered out to exclude those ones with bikeshops present within a 250 metres radius. To do this, I once again converted location coordinates to cartesian, stored them in a dictionary, and used the dictionary to iterate over locations to find those addresses that are free from competition. Preliminary results were based on the count of bikeshops per area of interest, and looked as follows (head):

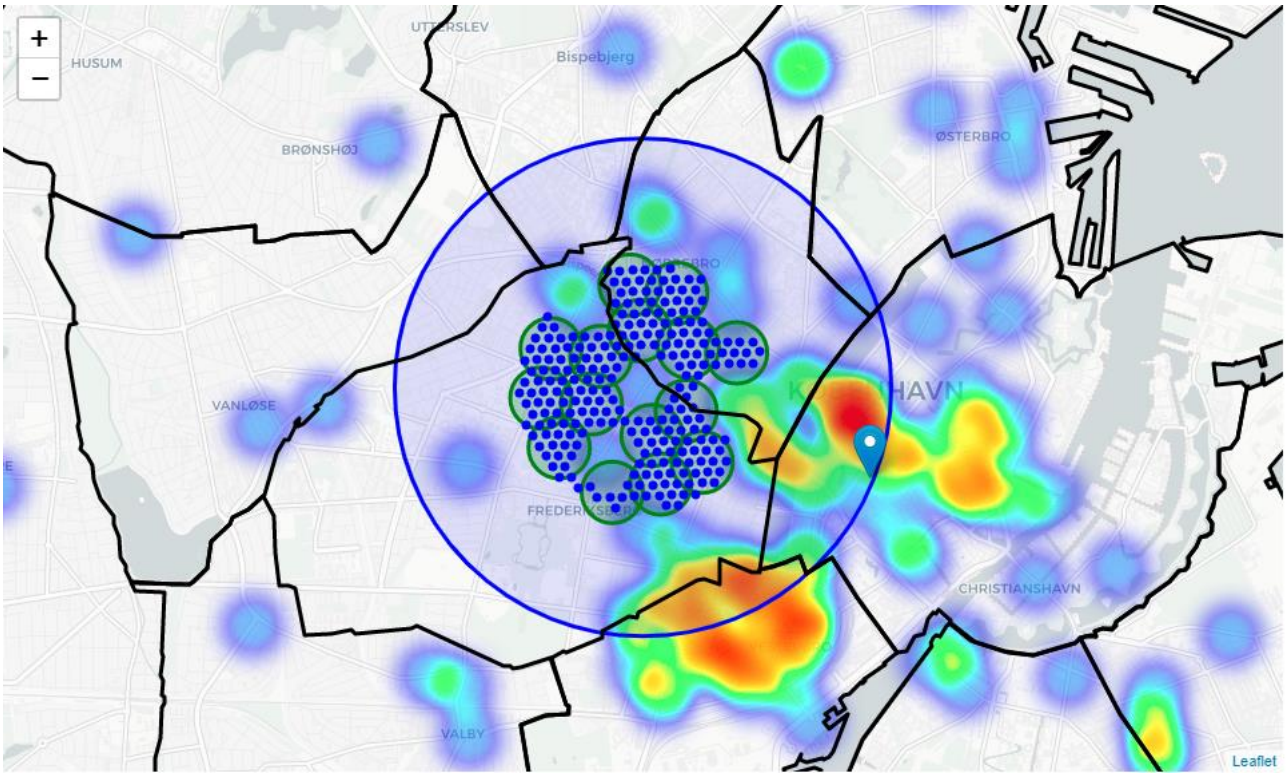
	Latitude	Longitude	X	Y	Bikeshops nearby
0	55.678185	12.536687	345111.741455	6.173014e+06	0
1	55.678217	12.538276	345211.741455	6.173014e+06	1
2	55.678249	12.539864	345311.741455	6.173014e+06	1
3	55.678281	12.541453	345411.741455	6.173014e+06	1
4	55.678313	12.543042	345511.741455	6.173014e+06	0

Then, only locations with no nearby competition (Bikeshops nearby=0) were kept, leading to 247 final candidate addresses.

Results were then visualized in folium.



Then, the candidates were clustered through k-means, leading to the formation of 15 clusters of addresses.

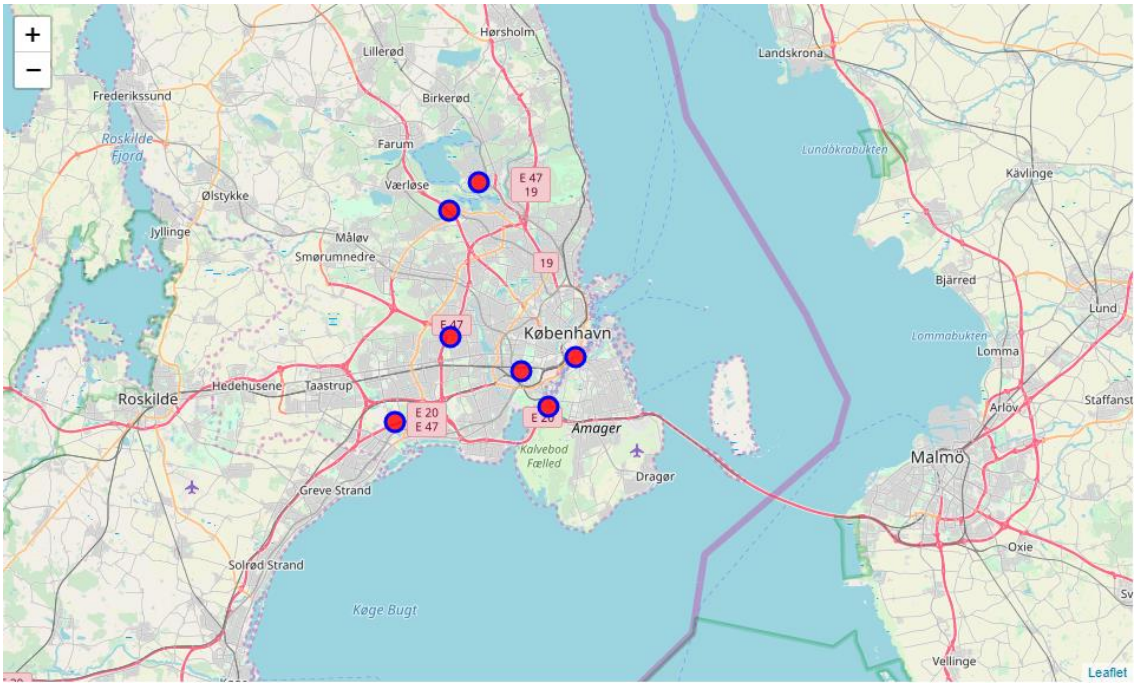


Finally, the Google Places REST API was used to narrow down the list of candidates to the 15 addresses corresponding to the centers of the clusters.

	Addresses	Postcode
0	L.I. Brandes Allé 10	1956 Frederiksberg
1	Lundtoftegade 42	2200 København
2	Guldborgvej 21	2000 Frederiksberg
3	Kapelvej 4	2200 København
4	Nordre Sti 4	1870 Frederiksberg
5	Nyelandsvej 25B	2000 Frederiksberg
6	Dronning Olgas Vej 30	2000 Frederiksberg
7	Husumgade 29	2200 København
8	Duevej 22	2000 Frederiksberg
9	Bille Brahes Vej 10	1963 Frederiksberg
10	Hans Tavsens Gade 40	2200 København
11	Aksel Møllers Have 7	2000 Frederiksberg
12	Rolighedsvej 903	1958 Frederiksberg
13	Stefansgade 73	2200 København
14	Rathsacksvej 14	1862 Frederiksberg

Final skimming

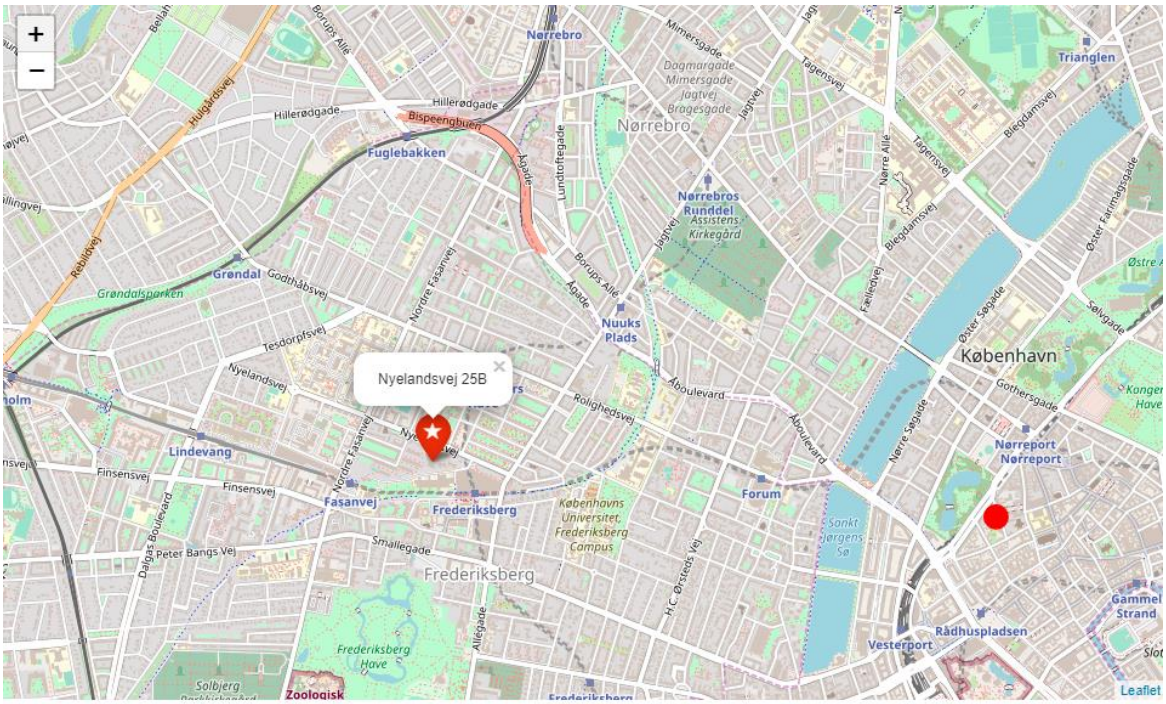
The final step of the analysis involved a further narrowing of the candidate addresses, based on their proximity to a major bike trail. Reportedly, only the closest one was kept. To do this, it was firstly required to extract the coordinates of the bike trails. This was done via the Foursquare API, simply by changing the categoryId by which the API screened for venues. This resulted to a list of 7 major bike trails, that was accordingly visualized in folium.



Then, similarly what was done for the bikeshop grid count, coordinates were converted into cartesian and stored into a dictionary to use the pre-stored function to retrieve the distances of the 15 candidate addresses from the closest bike trail. Results were as follows:

	X	Y	Biketrails_distance
0	5.298028e+06	1.811463e+06	4732.391484
1	5.299611e+06	1.811346e+06	5629.798853
2	5.298781e+06	1.809515e+06	3825.844471
3	5.298569e+06	1.813153e+06	4754.102547
4	5.297511e+06	1.811528e+06	4546.744370
5	5.298222e+06	1.809754e+06	3526.758621
6	5.298996e+06	1.810680e+06	4724.093919
7	5.299371e+06	1.812179e+06	6013.221291
8	5.299242e+06	1.809792e+06	4363.360466
9	5.297589e+06	1.812322e+06	4848.899847
10	5.298788e+06	1.812288e+06	5547.715579
11	5.298551e+06	1.810422e+06	4226.333468
12	5.298158e+06	1.812128e+06	5306.378969
13	5.299122e+06	1.811430e+06	5353.934346
14	5.297592e+06	1.810634e+06	3797.811926

The smallest value for distance was then located, and the coordinates were used to retrieve the address from a previously stored dataframe. This led to the definition of the “winning address”, it being **Nyelandsvej 25B,** hereby visualized in a folium map**.**



Results and discussion

The analysis shows that although there is a consistent number of bikeshops in Copenhagen (142 within the Copenhagen area), there are pockets of low bikeshop density fairly close to city center. Highest concentration of bikeshops was detected north-east and south-west from Kongens Nytorv, so attention was focused on the north-western area (as the south-east area is most isolated from tourist and resident traffic), corresponding to the intersection of the boroughs of Frederiksberg and Norrebrø. Interestingly, these two boroughs not only constitute a relatively low-competition zone, but also present significantly attractive features for the opening of bikeshops (frequency of bike commutes, general bike traffic and favorable demographics).

After directing our attention to this narrower area of interest (covering approx. 2x2km north-west from Kongens Nytorv) we first created a dense grid of location candidates (spaced 100m apart); those locations were then filtered so that those with already established bikeshops were removed.

Those location candidates were then clustered to create zones of interest which contain the greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors, such as the proximity to a major bike trail.

As a result, 15 zones containing the largest number of potential new bikeshop addresses were identified. This, of course, does not imply that those zones are actually optimal locations for a bikeshops. Purpose of this analysis was to only provide info on areas close to Copenhagen center but not crowded with existing bikeshops - it is entirely possible that there is a very good reason for small number of bikeshops in any of those areas, reasons which would make them unsuitable for a new bikeshop regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition.

Limitations

The results produced in the project are subjected to limitations in relation to A) methodology, B) data selection, and C) tool selection.

A) Methodology

The way the overall project was structured is the result of a series of deliberately (and non-deliberately) sub-optimal steps through which the code has been organized. For instance, there was no explicit need to extract postal codes, nor to use four different APIs. In relation to the former, for instance, bikeshop locations could have been extracted in result to a broad API request knowing just the lat/lon of Kongens Nytorv. In relation to the latter, the project could have been based on a single API to ensure more consistency of venue labeling and geocoding. However, I wanted to experiment with various APIs and different approaches to geocoding to get familiar with a broader and more versatile data science toolbox.

B) Data Selection

The data selected for the analysis represent just a minor amount of the different drivers that could have been used to filter out locations. For instance, average household income per neighborhood could have been used to further segment areas based on the likelihood of inhabitants to use bike as their main mean of transport. Similarly, historical data on bike traffic could have been used to define high-traffic areas within Copenhagen, so to maximize the number of potential customers eventually viewing the new venue. Again, these are just but few of the many other approaches that could have been considered for data selection in preparation for the analysis.

B) Tools

The reliability of the results is heavily dependent on the reliability of the tools used to extract data. Specifically, individual limitations of the APIs used throughout the project are a major factor in defining how comprehensive the information used for the analysis was. For instance, three APIs were tested for the extraction of bikeshop addresses (Google, Bing and Foursquare), and the one providing the most exhaustive results was selected (reportedly, Foursquare REST API). Nevertheless, to a citizen of Copenhagen it is evident how such results are incomplete - for example, a number of bikeshops in the area of Frederiksberg were not captured via the API, due to Foursquare API not recognizing them as bikeshops (categoryId was inconsistent). This kind of limitations can be overcome only via manual integration of data, an action that was considered to be out of scope due to time constraints.

Conclusions

Purpose of this project was to identify Copenhagen areas close to center with low number of bikeshops in order to aid stakeholders in narrowing down the search for optimal location for a new bikeshop. By calculating bikeshop density distribution from Foursquare data we have first identified general boroughs that justify further analysis (Frederiksberg and Norrebrø), and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby bikeshops. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of cluster centers were created to be used as starting points for final exploration by stakeholders.

Final decision on optimal bikeshop location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location, real estate availability, prices, socio-economic dynamics of every neighborhood etc.