BATTLE OF NEIGHBORHOODS
HOVEDSTADEN PARKS, DENKMARK

Rockie Ssengonzi

Date

21-04-2021

Course title

Applied Data Science Capstone

Coursera MOOC

Table of contents

Contents

1 Introduction

1.1 Background: Copenhagen, Greater Copenhagen & Central Region of Denmark

Copenhagen is a green city well-endowed with open spaces. It has an extensive and well-distributed system of parks that act as venues for a wide array of events and urban life. As a supplement to the regular parks, there are a number of congenial public gardens and some cemeteries doubling as parks. It is official municipal policy in Copenhagen that all citizens by 2015 must be able to reach a park or beach on foot in less than 15 minutes, as narrated on Wikipedia.

Copenhagen is the capital city of Denmark, a strategic harbor and the largest city in Denmark, and thus the most densely populated area in the country. Many people prefer to inhabit the greater Copenhagen areas also known as Hovedstaden or the capital region of Denmark which will be our Borough on Zealand (Sjaelland) island; Denmark's largest island. There are many parks on Zealand, which we will view in our quest for a clustering model to find similar areas based on neighborhoods.

1.2 Business Problem

The local authorities & Danish government as greener environment stakeholders would like to identify the most ideal way to segment Denmark's Central region parks and allocate them in an efficient way to outdoor recreational service providers and entrepreneurs who are interested in bidding for service provision in various sub-areas in the region. The stakeholders have assigned a data scientist at ugandane with a task to offer them a final deliverable addressing their business problem.

1.3 Interest

This project suffices as a reference for greener and environmentally aware stakeholders interested in identifying and optimal locale segmentation of parks for outdoor recreational service providers or entrepreneurs in parks in the central region of Denmark.

I used data science tools to fetch the raw data, visualize it then generate clusters of parks in the region based on above criteria. I analyzed the best ways to segment the parks, so that stakeholders can make the final decision based on the analysis.

2 Data Acquisition & Cleaning

2.1Data sources

Location and the number of existing parks in the neighborhood based on the definition of our problem, factors that may impact our decision. Park data for every neighborhood was be obtained using Foursquare API and the coordinates of Hovedstaden were obtained using Nominatim Geocoder module in python.
The Hovedstaden region data as our borough of choice data could be gotten readily at the simplemaps website. The data was downloaded as an excel sheet and read into a variable in the notebook for further processing.

2.2 Data Cleaning & Feature Selection

The location data was further cleaned to only consider the latitudes and longitudes, borough and neighborhood columns for the central region of Denmark, as shown in figure 2 below. I had to rename the columns and drop unnecessary columns, whose data I could not obtain, given the time and budget. This data can however be further leveraged upon in other data science projects, given the resources and where it may be essential.

[4]:

	Neighborhood	Latitude	Longitude	Borough
0	Copenhagen	55.6786	12.5635	Hovedstaden
1	Hillerød	55.9333	12.3167	Hovedstaden
2	Frederiksberg	55.6785	12.5221	Hovedstaden
3	Søborg	55.7302	12.5098	Hovedstaden
4	Hvidovre	55.6503	12.4758	Hovedstaden
5	Rødovre	55.6827	12.4644	Hovedstaden
6	Charlottenlund	55.7537	12.5918	Hovedstaden
7	Helsingør	56.0294	12.5863	Hovedstaden
8	Herlev	55.7235	12.4404	Hovedstaden
9	Kongens Lyngby	55.7718	12.5060	Hovedstaden
10	Ballerup	55.7198	12.3520	Hovedstaden
11	Taastrup	55.6475	12.3120	Hovedstaden
12	Vallensbæk Strand	55.6139	12.3895	Hovedstaden
13	Glostrup	55.6666	12.4038	Hovedstaden
14	Brøndby	55.6541	12.4215	Hovedstaden
15	Frederiksværk	55.9677	12.0215	Hovedstaden
16	Dragør	55.5946	12.6690	Hovedstaden

Fig 1: The 29 Neighborhoods of Hovedstaden

2.4 Data Exploration & Visualization

I used the Geolocator to locate Copenhagen's latitudes and longitudes so as to visualize the central region of Denmark (Hovedstaden). The 29 neighborhoods in the central regions were then plotted and visualized as shown in figure 2, so we could get well acquainted with their location as per the latitude and longitude information derived from the table

[6]:

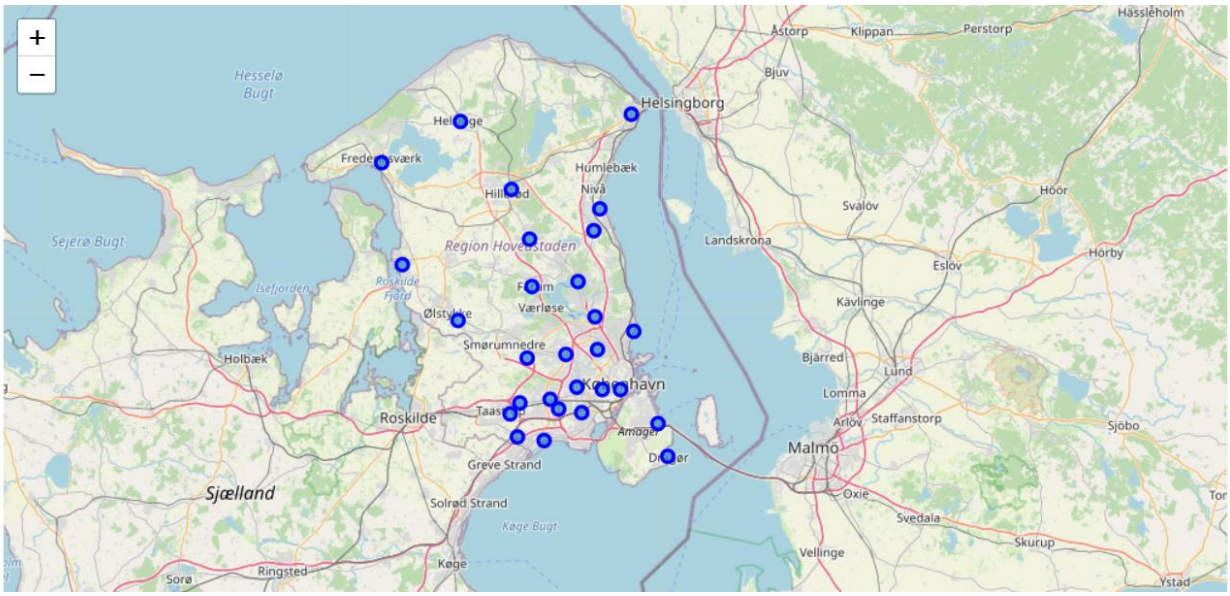


Fig 2: Hovedstaden Neighborhoods

I defined at python function that could leverage on the foursquare api to return data regarding venue categories, venues their latitude and longitudes from the various neighbourhoods. On execution, I parsed over 1411 venues and 202 unique venue categories. Preliminary exploration of the venue data shows that we have over 36 parks. The figure below only shows the few rows of the many venues we extracted from the function.

```
[10]: drk_venues.head()
```

[10]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Copenhagen	55.6786	12.5635	Ørstedsparken	55.680670	12.566367	Park
1	Copenhagen	55.6786	12.5635	Ørsted Ølbar	55.681217	12.564578	Beer Bar
2	Copenhagen	55.6786	12.5635	Boghallen	55.676788	12.568395	Bookstore
3	Copenhagen	55.6786	12.5635	Imperial	55.675365	12.561041	Movie Theater
4	Copenhagen	55.6786	12.5635	Høst	55.683279	12.566076	Scandinavian Restaurant

Fig 3: Neighborhoods

3 Methodology

The business purpose of this project was to find the best way to segment the green parks in the central region of Denmark (Hovedstaden) as a fundamental step towards allocating park recreation & maintenance services by the local authorities and Government to entrepreneurs and service providers. The CRISP DM methodology was adopted throughout the project as the data science standard and best in class approach.

Thus far I had retrieved the following data:

1. All location data of all the neighborhoods in the central region of Denmark
2. All Venues data in the Central region of Denmark

Assumptions and Limitations

1. The radius was limited to 2000 to avoid duplication of venues and limit to 25000 are wide so we can cover as many venues as possible per neighborhood
2. Due to time and budget constraints, other demographics were not taken into consideration, but can be used to further improve on the clustering model to segment allocations

Park distribution is subject to the Danish government and Local authorities' planning as Denmark prioritizes green cities
Parks are by default to be developed within 15 minutes of each block in populated areas like hubs,cities or town

In the final step, I focused on the best way to segment the parks and we will also present the candidate clusters in the map view for stakeholders to make the final decision based on our deliverable.

4 Model Development, Evaluation

4.1 Model Development

I had to feature engineer the venues feature dataset using the one hot encoding method to leverage on machine learning algorithm; K-means to be precise, to cluster the Parks as per neighborhood, and analyze each neighborhood

[17]:

	Neighborhood	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	B
0	Copenhagen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Copenhagen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Copenhagen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Copenhagen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Copenhagen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Fig 4: one hot encoding for K-means

I grouped the venues by neighborhood and got the average of their frequencies of occurrence/ mode and created a new data frame to represent the average number of parks in each neighborhood as shown in the figure below.

[18]:

	Neighborhood	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	BBQ Joint	Bag Sh
0	Albertslund	0.0	0.0	0.0	0.0	0.0	0.020408	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.0000
1	Allerød	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.0000
2	Ballerup	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.0000
3	Brøndby	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.0000
4	Charlottenlund	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.010309	0.0	0.0	0.010309	0.010309	0.010309	0.0103

Grouping the parks for each neighborhood

[19]:
```
# Lets look at the a dataframe with average frequency of hotels grouped by neighbourhood
parks = drk_group[["Neighborhood",'Park']]
parks.head() #.head() shows the first five rows only
```

[19]:

	Neighborhood	Park
0	Albertslund	0.020408
1	Allerød	0.000000
2	Ballerup	0.000000
3	Brøndby	0.021739
4	Charlottenlund	0.051546

Fig 5: Averaging the frequencies of venues per neighborhood and isolating average parks per neighborhood in a data frame

The resulting data frame was used as input in the k-means algorithm with K (number of clusters) instantiated with 3. I scored the model and proceeded to define a k-means Inertia function which we could use as input to evaluate the model using the elbow method and with the help of the yellow brick-stone python module.

4.2 Data Model Evaluation

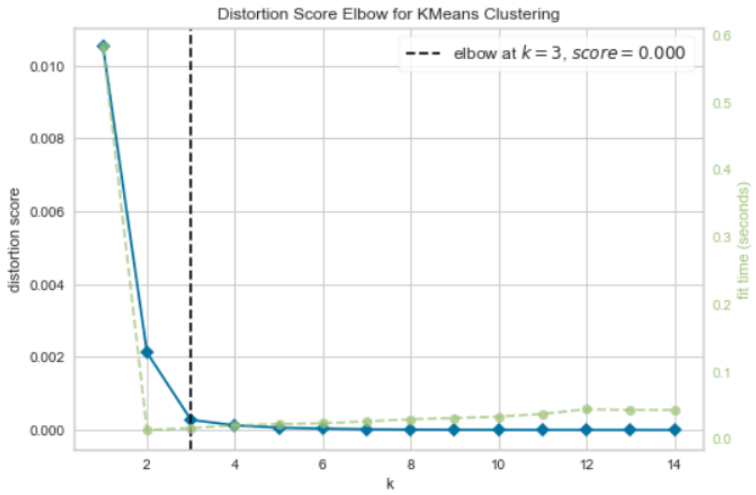
The parks dataframe was first adjusted to only take into account only numerical values, hence I droped the neighborhood coulm and renamed the dataframe X.

```
X = parks.drop(['Neighborhood'], axis=1)
```

X was the used as an input in the K-Elbow visualizer. The range of K in the k-means inertia function was 1 to 15 and the elbow suggested K=3 as the ideal number of clusters with the least distortion and error, shown in figure below.

Visualize the best *K-value* to choose

```
[27]: from yellowbrick.cluster import KElbowVisualizer
      # Instantiate the clustering model and visualizer
      model = KMeans()
      visualizer = KElbowVisualizer(model, k=(1,15))
      visualizer.fit(X)        # Fit the data to the visualizer
      visualizer.show()
```




```
[27]: <AxesSubplot:title={'center':'Distortion Score Elbow for KMeans Clustering'}, xlabel='k', ylabel='distortion score'>
```

Fig 6: K elbow visualizer to choose the optimum value of K

I merged the cluster labels from the scores of the model with the neighborhood, latitude, longitude and venues data given in the data frame shown in figure 3 above. The resulting data was a complete data frame was ideal for me to visualize the clusters using the python folium library model.

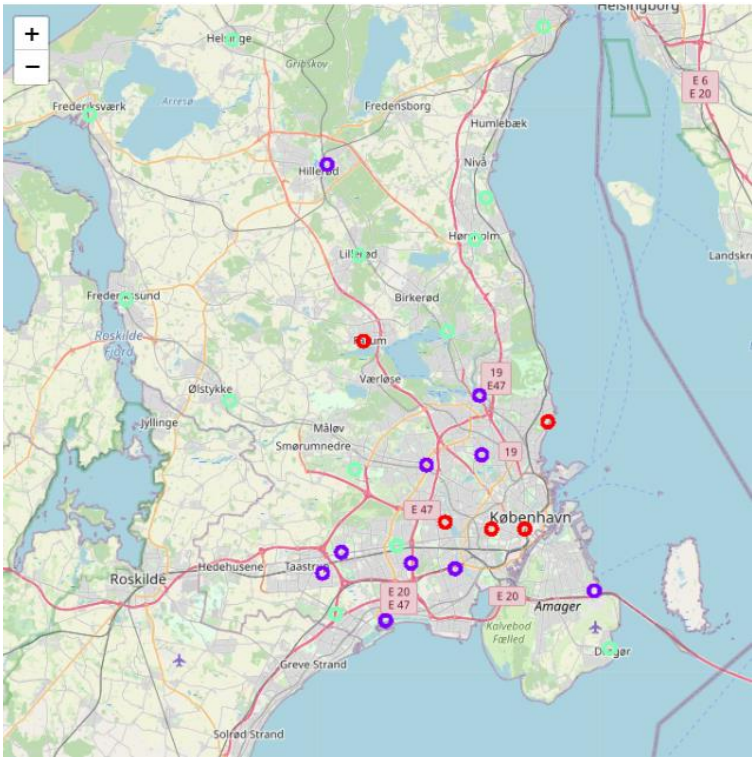


Fig 7: The 3 Clusters generated by the folium code

5 Results, Observations and Discussion

The 36 parks in the 29 neighborhoods seem to be concentrated in clusters with lesser neighborhoods.

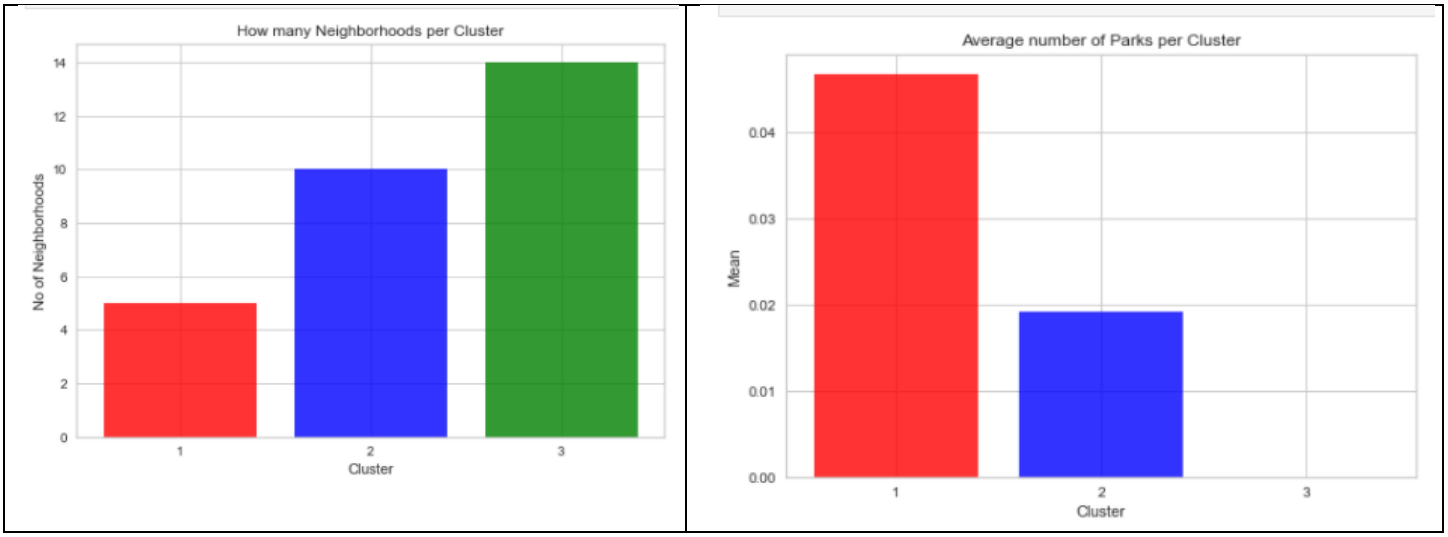


Fig 8: Comparison of clusters

There 39 parks over the 29 neighborhoods seem to be concentrated in clusters with lesser neighborhoods. This could be partially attributed to the fact that less densely populated areas have more areas to reserve as parks. Cluster 1 has 5 neighborhoods and the greatest number of parks. The Venues in cluster 1 should be targeted to tourists, entrepreneurs, service providers and locals who are into outdoor camping, hiking etc and alike activities. Cluster 2 has 10 neighborhood and the perfect balance of parks and neighborhoods hence ideal for both those into indoor recreational services and outdoor activities/ service provision, recreational or not. Cluster 3 has over 15 neighbors and yet little or no parks. Cluster 3 venues are magnet for extreme dwellers and those into indoor activities for the most part.

6 Conclusion & Future Direction

6.1 Conclusion

The purpose of this project was to find the best way to segment the green parks in the central region of Denmark (Hovedstaden) as a fundamental step towards allocating park recreation & maintenance services by the local authorities and Government.

After fetching venues data from Foursquare APIs, we used the K-Means clustering algorithm to group the existing parks into clusters and analyzed areas based on these fundamental data. Cluster 3 with the most neighbors, has the least parks partially due to venues aimed for indoor recreational activities, cluster 2 has the perfect balance of neighborhood and parks and cluster 1 has the most parks and thus ideal for outdoor recreational services which may include outdoor camping in the central region. We suggest the stakeholder to assume the 3-cluster model as an ideal solution for segmenting the parks and inviting service providers and business entrepreneurs to bid based on the segmented park clusters.

6.2 Future Direction

The final decision on optimum allocation of parks to service providers based on the segmentation of these location will be made by stakeholders (government & local authorities) on specific location characteristics of neighborhoods, taking into consideration additional factors like:

Government & Local Authorities Regulation regarding bidding

Population Density & People Traffic
Tourist attractions & camping or hiking opportunities
Frequency of Events and gathering to be held in the parks etc

7 References

- **Towards Data Science Theatre**
- **Towards Data Science Italian Restaurant**
- **Wikipedia Denmark Parks**
- **Folium**
- **Four Square API**
- **Simplemaps**

Code
https://github.com/rockiessengonzi/Coursera_Capstone/tree/master/Final%20Assignment