Battle of the Neighborhoods – Artisanal Coffee Shop Edition

Jonas Warming – 16 July 2020

1 Introduction

This report will solve a fictional business problem anchored in a real-world scenario exclusively via data analysis of openly available data-sets.

1.1 Background

A local Danish businessman wants to open a hip coffee shop to take advantage of the artisanal coffee trend. As coffee shops are plentiful and have low startup costs the businessman wants to open the coffee shop in an area that doesn't have a high density of coffee shops. The businessman would also prefer to place the shop in a neighborhood with a growing population of adults in the age bracket 20 - 44 as the businessman believes this group is most likely to have a higher percentage of spending money and more interested in exploring high quality coffee. The businessman would ideally like to have a foothold in the market by 2022.

1.2 Problem

This project will explore public population data sets of Copenhagen and Foursquare API using exploratory data analysis and K-Means clustering to recommend the best neighborhood(s) to pursue further.

1.3 Interest

Committing time, startup capital and marketing money into opening a new coffee shop is expensive and time consuming. Being able to avoid spending time researching areas, marketing to the wrong client bases and potential losses by later relocating to a new area if the target audience has moved can save a potential store owner significant money and help be better localized when the target age population ramps up.

2 Data Acquisition and Cleaning

2.1 Data Sources

We will be using a series of openly available datasets to solve the problem. The first 3 are curated by the Municipality of Copenhagen. Bing API is free to use for multiple geocoding per day. The last is owned by Foursquare and restricted to the free tier level for this report.

1. **Danish Neighborhoods GeoJson:** Used to segregate the neighborhoods of Copenhagen
2. **Spending Income per household:** To find the neighborhoods with the highest available spending money and highest growing spending money.
3. **Population Data:** To explore which neighborhoods has the highest growing population of our target audience
4. **Bing API:** To find the latitude & longitude of the center of our chosen neighborhoods
5. **Foursquare API:** To catalog existing number of coffee shops to decide where we have the weakest competition

2.2 Data Scraping, Cleaning & Feature Selection

First, I scrape the official neighborhoods of Copenhagen from the geojson file and their ID numbers to later join it with the population data. I have chosen not to use the multi-polygon coordinates and instead have utilized the Bing maps API to find the center of the neighborhoods.

The next piece of the puzzle is to find the population of our target age range (20-44) per neighborhood in 2022, the expected growth of that population and their disposable income. The data sets available from the Municipality of Copenhagen have these values per 5-year age ranges per neighborhood per gender. The population totals have been summed up in the target age range and all genders included. The population data is available up until 2035 however we are only interested in the values until 2022. I have used the percentage growth from 2021 to 2022 as expected population growth. The disposable income is only available up until 2017, but I have accepted that as generally representative and averaged the disposable income between genders.

Finally, I have queried the Foursquare API to create a custom data set of venues in the category coffee shops within 750 meters of the neighborhood center. 750 meters was chosen as the overall total diameter of Copenhagen is ~10km. With 10 different neighborhoods that would leave ~1 km per neighborhood, but to avoid overlaps of venues I decreased that by 25%. With the venue data set I transformed it into a count of venue types per neighborhood.

All these data sources have been joined back into the main dataframe using neighborhoods as the join column to prepare for exploratory data analysis. Using the joined dataframe a final feature of number of coffee shops per population was added.

	id	area	Lat	Long	Coffee Shop	Café	Tea Room	2021	2022	popgrowth	shoppop
0	9	Amager Øst	55.665629	12.613326	3.0	2	0.0	31326.108664	31476.852800	0.004812	0.000095
1	1	Indre By	55.680000	12.580000	28.0	69	3.0	25730.439714	25754.266376	0.000926	0.001087
2	2	Østerbro	55.709209	12.577404	3.0	13	0.0	37098.265154	37320.935596	0.006002	0.000080
3	3	Nørrebro	55.694340	12.548649	14.0	14	0.0	46427.668216	46024.322302	-0.008688	0.000304
4	7	Brønshøj	55.704823	12.496385	3.0	2	0.0	16584.450203	16430.912557	-0.009258	0.000183
5	8	Bispebjerg	55.716881	12.533979	0.0	3	0.0	30373.500777	30457.652863	0.002771	0.000000
6	6	Vanløse	55.690086	12.489993	1.0	3	0.0	17622.023099	17504.647390	-0.006661	0.000057
7	10	Amager Vest	55.641670	12.578060	1.0	1	0.0	40509.784328	42055.232397	0.038150	0.000024
8	5	Valby	55.666290	12.514337	5.0	6	0.0	30582.955249	31545.886441	0.031486	0.000158
9	4	Vesterbro	55.672199	12.555000	27.0	46	1.0	40881.295230	42222.442244	0.032806	0.000639

Fig 2.2 Final Dataframe

3 Exploratory Data Analysis

As I have chosen to approach this from a clustering perspective the exploratory analysis will be done post clustering.

3.1 K – Means Clustering

K-Means clustering is an iterative unsupervised algorithm that clusters data points together based on their feature-set similarity. My expectation is that by running the algorithm on our chosen feature-set (# coffee shops, population of target ages, growth of target ages, shops/population, disposable income) the algorithm will cluster our neighborhoods together and hopefully find a neighborhood with high population growth, high income and few coffee shops to compete with.

After normalizing the feature-set via a standard scaler to avoid over-weighting features with larger nominal numbers I identify the correct number of total clusters. The methodology for this requires running the algorithm multiple times across the relevant range of clusters. The maximum clusters for the algorithm are *n* samples – 1.

I will use two metrics to determine which number of clusters is the most accurate clustering of our data, the silhouette score and the standard score measure from the k-means package. The standard score measures the average distance of the data points to its cluster center. Ideally, we want to minimize this number to ensure the cluster is as similar as possible. As number of clusters increase the distance minimizes but the generalness of the cluster diminishes, and we can draw fewer conclusions. When only using the standard score the “elbow method” is used to choose a cluster number based on when the entropy or distance gained by increasing cluster numbers increases less rapidly between clusters. This method has the drawback of being difficult to choose which cluster number is the one where entropy decreases enough to choose it.

I have therefore supplemented with the silhouette score which measures how similar the objects in a cluster are to its own cluster compared with the other clusters. A higher score here is better.

Let’s look at the scores for the battle of the neighborhoods feature-set.

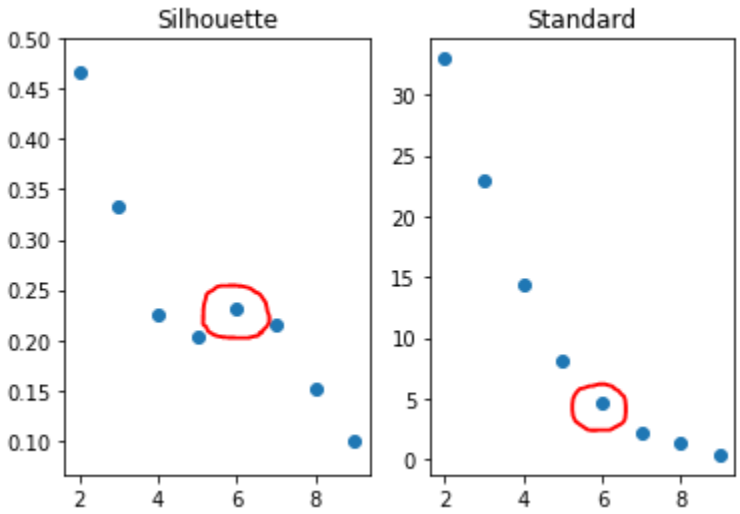

Fig 3.1: Silhouette and Standard score charts for different number of total clusters

Only looking at the standard score method it would be difficult to see whether we should choose 5, 6 or 7 total clusters. By supplementing the silhouette score it makes it easier to see that there is a stronger similarity by choosing 6 total clusters.

3.2 Examining the clusters

Since K-Means is unsupervised there is no upfront explanation of the decision to allocate certain neighborhoods to certain clusters. I will examine the clusters and explain what the defining characteristics of each cluster are.

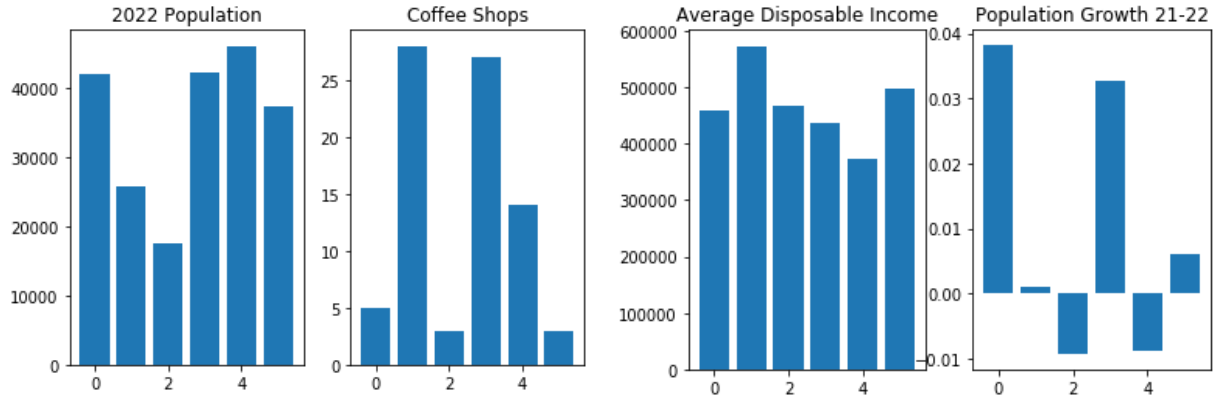

Fig 3.2: Examining the characteristics of the different clusters.

Cluster 5 doesn't have very many coffee shops but has a mid-large population with the second highest disposable income. This is a strong sign. The detracting factor is that cluster 5 is not growing very much.

Cluster 4 has negative population growth, a middling number of existing coffee shops and a relatively low disposable income.

Cluster 3 has a lot of existing shops, high population and middling income but with the second strongest growth rate.

Cluster 2 has few existing shops, middling disposable income, low population and a negative population growth.

Cluster 1 has the highest disposable income a breakeven growth rate and the highest number of existing shops.

Cluster 0 has high population, very few shops, middling income and the highest population growth.

4 Conclusion

Based on our data analysis we should dig further into Cluster 0 which has the strongest blend of our 4 characteristics.

	id	area	Lat	Long	Coffee Shop	Café	Tea Room	2021	2022	popgrowth	shoppop	disposable	Labels
7	10	Amager Vest	55.64167	12.578060	1.0	1	0.0	40509.784328	42055.232397	0.038150	0.000024	459276	0
8	5	Valby	55.66629	12.514337	5.0	6	0.0	30582.955249	31545.886441	0.031486	0.000158	434942	0

Fig 4.1: Our optimal neighborhoods.

Amager Vest and Valby are the neighborhoods that are most similar as having a high disposable income, high population growth of the targeted age range and few existing coffee shops. The fictional businessman should focus on these two neighborhoods when looking for real estate to open the shop in. If those prove difficult the final top 5 ranking of neighborhoods is:

1. Amager Vest
2. Valby
3. Vesterbro
4. Amager Øst
5. Østerbro

5 Improvements

The conclusion could be improved by spending more time on a couple of things. One element would be to add more data to the feature-set. For example, real estate pricing and average rating of the existing venues could add another decision layer to further cluster and select our target neighborhood. A one number comparison could simplify the ranking process instead of eye testing the features that best match up with the requirements.