BATTLE OF THE NEIGHBOURHOODS – AN EXPLORATION OF AMSTERDAM CITY

Samuel Ileoma

June 2020

1. Introduction

1.1 Background

Amsterdam is an amazing city – small in size, but large with diversity. It is only 219.3 square kilometers (Google Maps) which is a miniscule size in comparison to other cities around the world. Compared to New York, it is thrice as small, but still just as diverse. Being the capital of the Netherlands, Amsterdam is the heart of its nation, steeped very deeply in Dutch Culture that is very much alive with all sorts of historical landmarks, European food, and art. Yet, it is also home to people of diverse cultures from its temporary tourists to its permanent residents, transforming it to a vast hub of business ventures trying to satisfy the plethora of needs and wants of its inhabitants. I Amsterdam, one of its most popular city-marketing websites, claims Amsterdam to have 180 different nationalities of which 45 percent are minorities, making it to be amongst the most diverse cities in Europe just under London and Paris.

1.2 Problem

With these observations, there is a need to explore the city to see which types of businesses are most common and where exactly they can be found within such a diverse city still deeply rich in European culture. This exploration would provide insights into relationships between geographical coordinates of the city's neighbourhoods and the popularity of similar businesses existent in them. The results of this analysis can therefore be used by start-up entrepreneurs to determine what food/retail/hospitality businesses are best to set up and which populous locations are best to establish these businesses in Amsterdam.

2. Data Acquisition and Cleaning

2.1 Data Sources

The data used in this project is taken from ClairCity Data Portal of Districts and Neighbourhoods in the Netherlands. The dataset used in the project can be found with this link: https://claircitydata.cbs.nl/dataset/districts-and-neighbourhoods-amsterdam/resource/d02c5f12-1cfa-4d7c-91d3-41af8e4ed634?view_id=7b50de62-7ec8-4228-b19a-8d32702aa363.

Its variables include Neighbourhoods and Districts in Amsterdam each having their own population of inhabitants and geographical coordinates etc. Its last update was in 2016, which is a limitation of research, but still deemed relevant for the explorative analysis enacted in this report.

2.2 Data Cleaning

Downloading the dataset as a csv file was easy, but its content was not properly reading into a DataFrame as none of the values seemed to show properly with pandas. This required me to reformat the dataset in excel and splice it down to these relevant features: subject, region_name, region_ type, region_code, lat, long and ninhabitants. This new dataset was then easy to read into a dataframe. The description of its variables can be found in the list below:

1. Subject – Name of neighbourhood or district.
2. Region_name – Name of city that the neighbourhood belongs to.
3. Region_code – District Code of the Neighbourhood.
4. Lat & long – Latitude and Longitude geographical coordinates of each Neighbourhood.
5. Ninhabitants – Number of residents

There were still several problems with the dataset that required cleaning. Firstly, there were missing values that had to be removed. Because the missing values were very few in number, and could only be found in the region_code, lat, and long columns, I deleted entire rows containing them as they gave no relevant data to the analysis at hand.

Secondly, not all the subjects belonged to the Neighbourhood of Amsterdam, so they also had their rows entirely removed. This reduced the row count from 579 to 575. Thirdly, the data types of numerical values of geographical coordinates needed to be changed from objects to integers for smooth operations in later part of the analysis.

Lastly, because this analysis is focused on the most populous areas of Amsterdam, the dataset had to be cleaned to only having neighbourhoods of more than 4000 inhabitants. All rows containing neighbourhoods of less than or equal to 4000 inhabitants were therefore removed as well. This reduced the row count of subjects from 575 to 107 neighbourhoods that were explored in this analysis.

2.3 Feature Selection

After cleaning the data, the most relevant features of the 107 samples needed to be analysed were taken and renamed. This included: Subject which was renamed to Neighbourhood; Region_Code and Lat and Long values. With this done, it was now appropriate to explore the current neighbourhoods for their most common venues to eventually determine the businesses that are most popular in Amsterdam's streets according to location. To do this, the Foursquare API system was used.

Foursquare API provides a vast amount of location data, giving valid information on venues around the world such as their addresses, tips and comments of visitors laced with variety of photos. To make use of this API, I signed into its platform as a developer and was allocated a

CLIENT_ID and CLIENT_SECRET code which I used to authenticate myself in retrieving the relevant data I needed for analysis.

After defining a function that fed the 107 neighbourhoods and their geographical coordinates, I was able to extract a total of 294 unique venues in Amsterdam using foursquare API service. This then led to the methodology of using one-hot encoding to change the categorical variables of venues to dummy variables, and then use clustering machine learning to determine the most common venues visited in each neighbourhood of Amsterdam.

3. Methodology

3.1 One-hot Encoding

Finding the most common venues in each neighbourhood requires the use of an unsupervised learning algorithm to group the venues into clusters. To therefore perform this kind of machine learning, the venues which are currently categorical in data have to be changed to numerical data for smooth operations. With one-hot encoding, this was made possible as all categorical data was changed to dummy variables, giving them ability to used like numerical values.

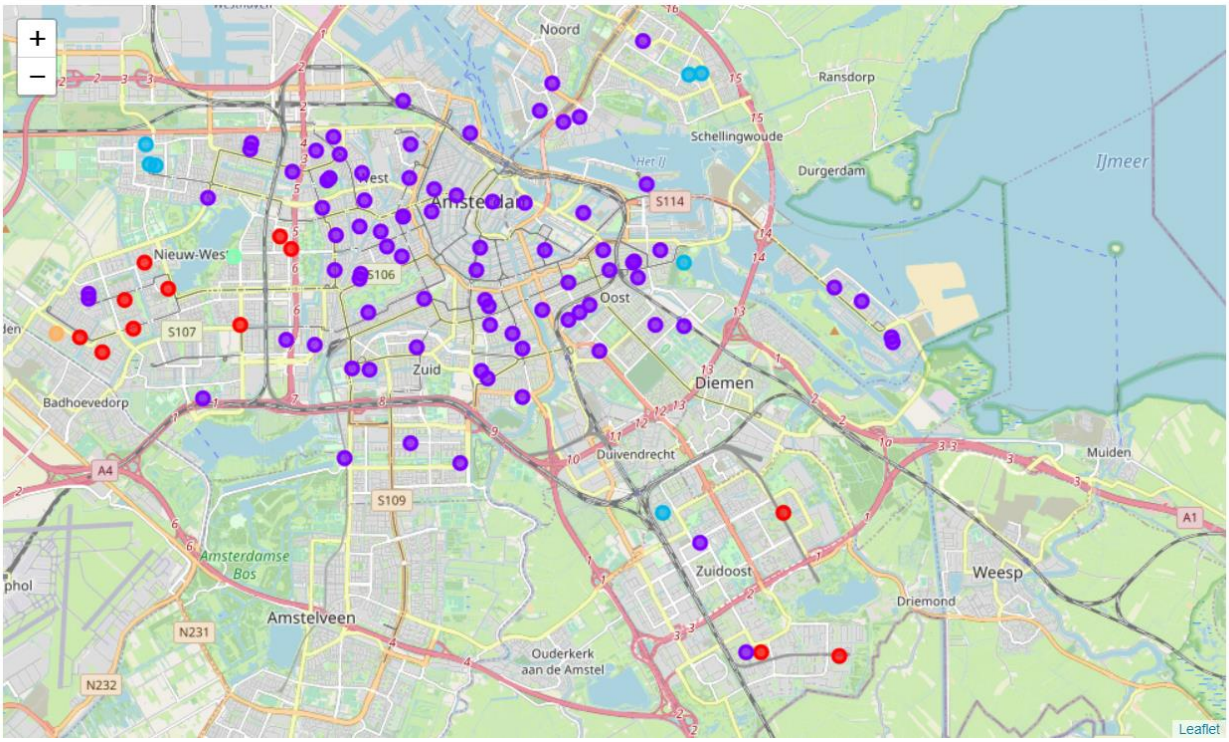
3.2 K-Means Clustering

The unsupervised Machine learning used in this analysis was K-Means Clustering as it is an unsupervised learning algorithm that groups datasets into k set of clusters. The optimal number of clusters chosen for this analysis was a count of 5 clusters. This was then visualized with a map plot showing how the clusters are organized with the use of differentiating colors as seen in the figure below.

4. Results

4.1 Visualization of Clusters

The map below visualizes the results of clusters of neighbourhoods of Amsterdam. This was generated with the use of Folium maps according to the geographical coordinates of the 107 neighbourhoods selected for analysis.



Each Cluster, demarcated with colours, segments the neighbourhoods into groups according to their similarities of the five most common venues of the city. They therefore help to determine which businesses are most popular in the most populous areas in Amsterdam, by grouping the neighbourhoods into similar groups of distinguished venues. The results of this analysis will be explained in the next section below.

4.2 Description of Clusters

Each cluster is named according the most distinguishable type of businesses that can be found in each neighbourhood. With this data, I compiled 5 datasets – one for each cluster – to show the 5 most common venues to which each cluster is similarly comprised of. Here are brief descriptions of each cluster:

1. Retail Cluster: This is the first cluster. It is visualized in the map with red dots. It comprises of a total of 12 neighbourhoods in which the most common venues in the area are that of retail businesses. These venues include the likes of Supermarkets, Shopping malls and pharmacies. A few restaurants can also be found in these places which are mostly of Turkish, Asian, or Mexican descent.

2. Hospitality Cluster: It is the largest cluster of the mix and is visualized with purple dots on the map. It comprises of a total of 83 neighbourhoods in which bars, cafes, and coffee shops are amongst the most common venues in the area. There is also an entourage of hotels and hostels and culturally diverse restaurants in these vicinities. Amongst other business venues that thrive in these areas are night clubs, gyms, and bakeries.

3. Bus Stop Cluster: 7 out of the 9 neighbourhoods in this cluster has a bus stop as its most common venue. It is distinguished by blue dots on the map. There are numbers of retail businesses in these areas with a very few restuarants compared to the first two clusters. Most of the restaurants in this cluster are either Turkish or Indonesian ones.

4. Sports Cluster: Visualized by the one single orange dot on the map, this Cluster has only two neighbourhoods: Oostzanerwerf and Slotervaart Noord. They both have football fields amongst the 5 most common venues. Slotervaart Noord has a few popular restaurants in its vicinity as its most common venues are Asian restaurants, and its second are Moroccan ones.

5. Outlier: This cluster has only one venue, De Aker West, and is represented by the light green dot on the map. It is completely odd from the others as its five most common venues in descending order include: A tram station, harbour, museum, trail, and zoo.

5 Discussion

5.1 Research Insights

It can be seen from the map that retail business thrive in the west and south-east of Amsterdam especially in the areas of Nieuw-West and Zuidoost. We can also assume from the distinguishable types of restaurants set up that these areas contain a significant concentration of minorities. The city-center on the other hand proves itself to be the land of the tourists with the availability and cultural diversity of businesses in the food and hospitality industry. With beer being an incredible staple in Europe, and the Netherlands known for its numerous breweries and beverage brands, it is understandable why bars, and cafes thrive in this areas. Night-life is very active in these areas as well so it is best to have businesses that are open later in the day than earlier.

5.2 Research Limitations

Because this research uses a dataset last updated from 2016, it would be good to use more updated datasets for further research. Also, with this analysis requiring unsupervised machine learning, other techniques other than K-means clustering can be used to bring forth better insights into the exploration in Amsterdam to help determine which businesses popularly thrive in its streets.

Conclusion

In summary, I analyzed a sample of 107 neighbourhoods in Amsterdam to determine the most common businesses in the city and developed a geographical pattern of where these businesses are mostly located. I used the variables of Latitude and Longitude geographical coordinates and population of its inhabitants to select the neighbourhoods analyzed as a sample. I used unsupervised machine learning to generate relationship insights concerning the proposed business problem and developed a cluster map and several datasets (not shared in

this report) to show such insights. These can therefore be used by stakeholders (entrepreneurs of small traditional businesses) as a tool of research to make business decisions regarding where to set up businesses in the future and encourage competition.