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 of Amsterdam to see which types of businesses thrive with customer interest in such a diverse city still deeply rich in European culture. This exploration would provide insights into consumption interests of Amsterdam's inhabitants according to relationships between business venue locations and their popularity with consumers. 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 is this one here. 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 which type of businesses tend to thrive with customers in Amsterdam 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 254 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.