Coursera Capstone project

IBM Applied Data Science

Opening a new bakery in Amsterdam, the Netherlands

By Anastasia Holovchenko

Introduction:

A modern bakery supplies society with first-need daily food, such as bread, as well as special occasion items, such as cakes, cookies, etc. A development and well-being of the neighborhood depends on the existence of local small shops, and bakeries are one of that type.

Business Problem:

A well-known bakery chain wants to open a new unit in Amsterdam, the Netherlands. They requested data scientist to do the analysis and suggest which places/areas of the city would be suitable for opening a new bakery. Using data science methodology and clustering machine learning technique I will answer the question: Where is the best place to open a new bakery in Amsterdam?

Data:

To solve the problem, I will use data from the following sources:

- List of districts in Amsterdam
- Latitude and longitude coordinates of the neighborhoods
- Venue data (location of current bakeries in each neighborhood)

From the sources:

- Boroughs of Amsterdam (https://en.wikipedia.org/wiki/Boroughs_of_Amsterdam)

It contains a list of all 7 districts of Amsterdam. I will use web scrapping technologies to get data from Wiki using beautiful soup, get coordinates of each neighborhood using Python Geocoder package.

- Python Geocoder package for the coordinates

And after that I will access Foursquare API to get the venues available in that particular districts, to determine how many bakeries are there already (in the follow up analysis)

- Foursquare API (https://foursquare.com/) to get the venue data

Methodology:

1. First, I need to get list of neighborhoods from Wiki. For this I scrap the page (https://en.wikipedia.org/wiki/Boroughs_of_Amsterdam) using Python and Beautiful soap and display the results in DataFrame view (table).
2. I get geographical coordinates in the form of latitude and longitude to be able to use it later for Foursquare API. To do so, I use Python Geocoder package which converts addresses into geographical coordinates. And visualize neighborhoods with Folium map package.
3. Next, I will use Forusquare API to get information about venues in a given regions of Amsterdam, in particular I will get the top 50 venues in the radius of 1000meters. This will give an idea on how busy the areas are with certain venues. I make a API call using python code and Foursquare returns me the answer in a json file which is then converted according to the neighborhood.
4. Because we're looking for the best place to open a new bakery, I extract the number of "Bakery" units in the dataframe, and define how many of them are in each district of the city. The frequency of "bakery" appearance allows me to group districts by using k-means clustering method. This algorithm allocates every datapoint to the neighboring cluster (based on previously defined centroids).
5. I clustered neighborhoods in 3 clusters, based on the frequency of "bakery" appearance. The results of this clustering allows me to identify in which areas there are more bakery than in others, which would then help me o answer the business question.

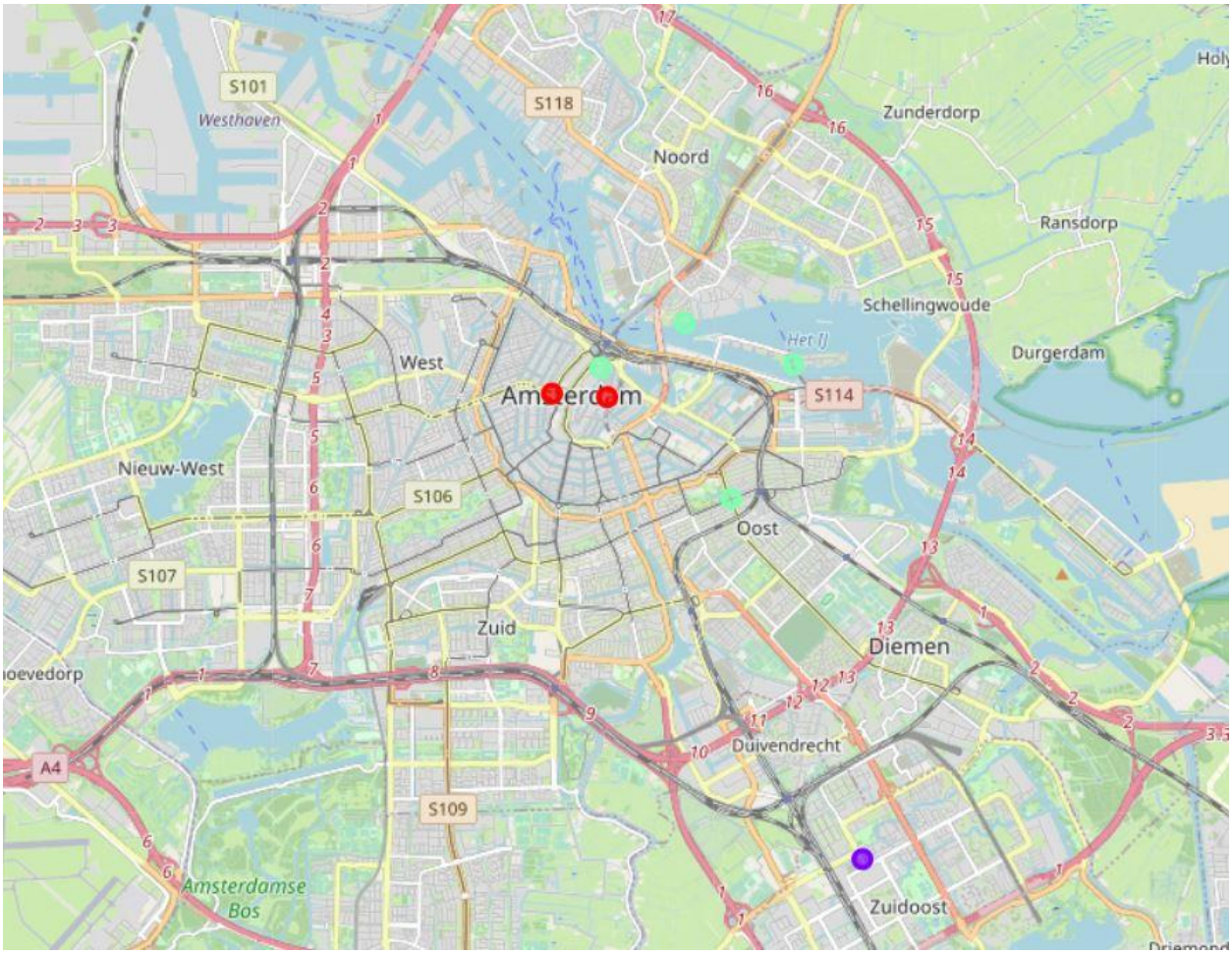
Results:

After k-means clustering technique I obtained the following 3 clusters with different occurrence of "bakery" in it:

- Cluster 0. Neighborhood with low number of bakeries (red color on the map)

- Cluster 1. Neighborhood with medium number of bakeries (purple color on the map)
- Cluster 2. Neighborhood with high number of bakeries (green color on the map)

The results of clustering are shown on the map below:



Discussion:

From the k-means clustering of data we see that the most bakeries are located in Cluster 2, which is geographically located around the city center, in living areas of Amsterdam. On the other hand, Cluster 0 (red on the map) has very low concentration of bakeries, and could be a great opportunity for a new

developer to open a bakery there, because of too little competition. At the same time, the demand for baking products in the city center can be lower, than in any area around it, because mostly tourists are visiting it. But, in Cluster 1 it is found that the concentration of bakeries there is moderate, and thus could be also potentially host as a place to open a new unit.

Conclusion:

In this project I aimed to perform data analysis and answer the business question: Where is the best location to open a new bakery in Amsterdam, the Netherlands. The neighborhoods in Cluster 0 (Niewe West, West) tend to be the most promising areas, as just a few bakeries are currently open there. At the same time, the neighborhoods in Cluster 2 are the worse to open a new bakery, because many of these types of shops exist there already.

References

Boroughs of Amsterdam (https://en.wikipedia.org/wiki/Boroughs_of_Amsterdam)

Foursquare API (https://foursquare.com/)