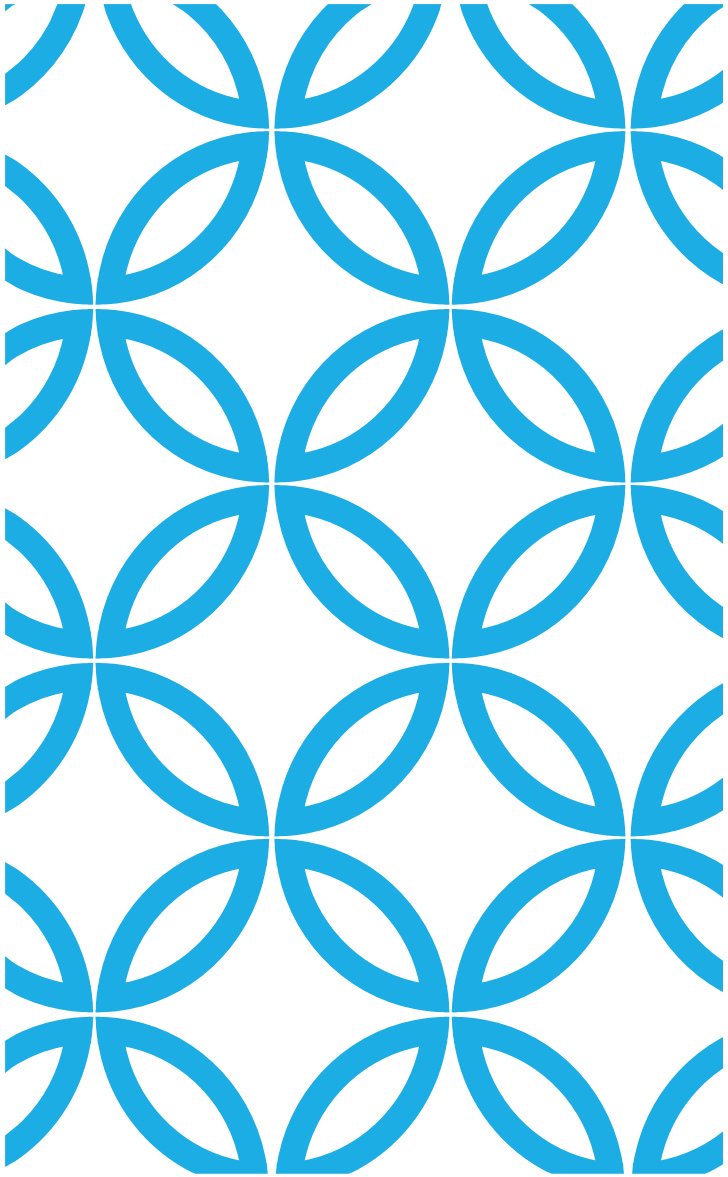


IBM PROFESSIONAL DATA SCIENTIST SPECIALIZATION

Capstone Project

Maria Alejandra Claure Oviedo

CONTENT

INTRODUCTION

A family living in the center of Düsseldorf, Germany would like to move to the outskirts of the City and needs to evaluate which neighborhood will provide them similar venues.

For this they contacted a friend, who is currently studying the Data Science specialization, with the hope she can provide them of some ideas of where to start looking.

Once she understood the problem. She remembered she had a module on Foursquare, where they employed K-means to cluster neighborhoods for Manhathan and Toronto. She started to look at the exercises of for that course and came up with the following results.

DATA REQUIREMENTS

The data requirements to solve this problem will be the neighborhoods from Düsseldorf, latitude and longitude and the zip codes known as "Postleitzahl". For this our data scientist found the required information in this web site: http://postleitzahlen.woxikon.de/plz/duesseldorf

She needed to use the package BeatifulSoup to scrap the information. Then she changed the column names and save it as dataframe.

To get the Latitude and Longitude, the dataframe was uploaded to this service provider: https://csv2geo.com/

Once the csv file was ready with the Latitude and Longitude coordinates. The data was imported to this notebook. Duplicates where removed.

	PostalCode	District	Neighborhood	Latitude	Longitude
0	40210	Düsseldorf Stadtmitte	Oststr.,Steinstr.,Marienstr.,Platz der Deutsch...	51.221643	6.788295
1	40211	Düsseldorf Pempelfort	Malkastenstr.,Louise-Dumont-Str.,Wielandstr.,C...	51.230667	6.791758
2	40211	Düsseldorf Stadtmitte	Liesegangstr.,Leopoldstr.,Kölner Str.,Kurfürst...	51.226600	6.789675
3	40212	Düsseldorf Stadtmitte	Wagnerstr.,Josephinenstr.,Königsallee,Königstr...	51.224387	6.781410
4	40213	Düsseldorf Altstadt	Flinger Str.,Hunsrückenstr.,Altestadt,Andreass...	51.227703	6.773738

METHODOLOGY

1. First all the needed libraries where imported

2. To center the Folium map on Düsseldorf, the coordinates where imported using the geopy package

3. A map of Düsseldorf with Districts superimposed on top was created

4. The credentials of Foursquare where used to get information about venues surrounding the different neighborhoods.

5. A limit of 10 venues per neighborhood and a radios of 500m was set.

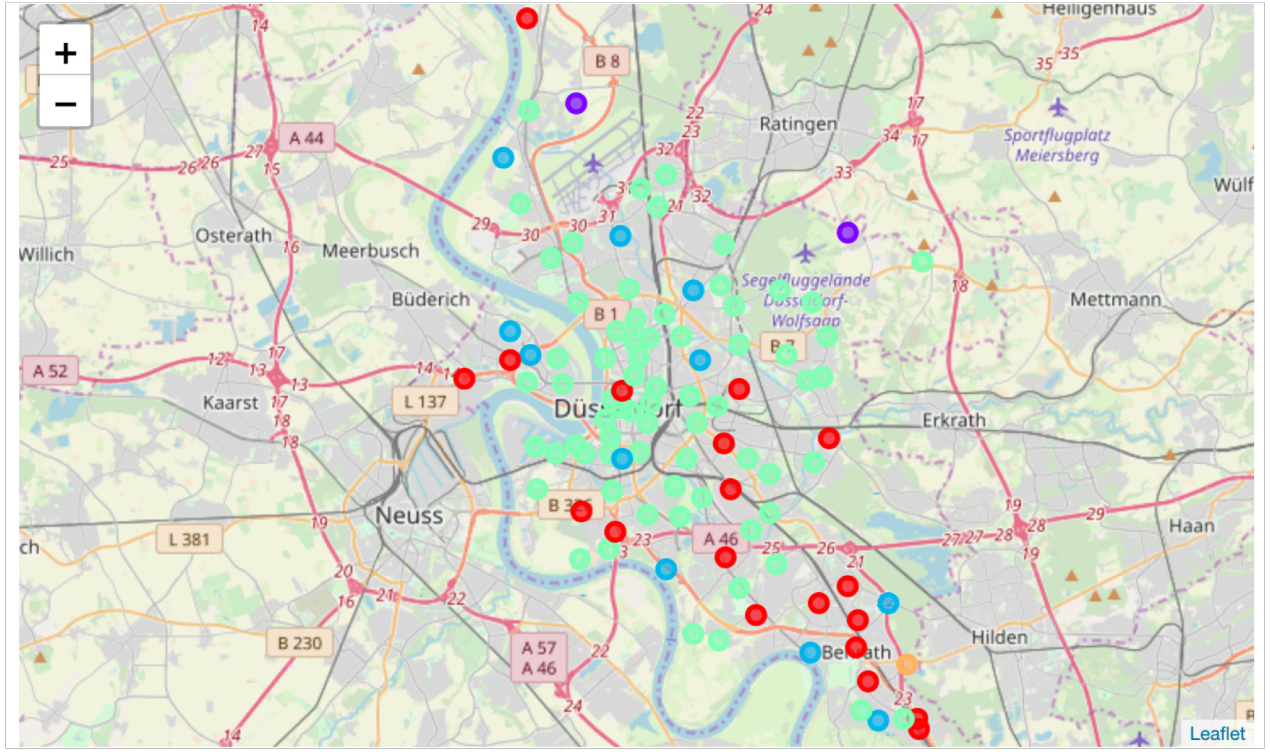
6. One-hot encoding → preparation of the data set before using machine learning

METHODOLOGY (CONT…)

7. K-means was used to cluster the neighborhoods. The value of K was set to 5

8. Results where grouped per neighborhood.

9. The clusters where displayed in map



RESULTS

In the map there was a predominance of points with color green. These points belong to the cluster number 3.

After this analysis the neighborhoods could be classified in three main clusters:

- High density of venues (green) - cluster 3

- Medium density of venues (red) - cluster 0

- Lower density of venues (blue) - cluster 2

DISCUSSION

After running the analysis, it was very interesting to see the algorithm clasify the neighborhoods in a similar way to what the general knowledge is. The areas with a green marker are precisely the most populated, expensive ones and mainly very close to the city center. It was great to see that this family now counts with additional information about other neighborhoods in Düsseldorf. Now they can decide to move e.g. from Bilk to Garath with similar number of venues and types (supermarkets and restaurants).

CONCLUSSION

This kind of analysis proved to be very useful for the decision making process of moving to a new neighborhood. Normally, families employ weeks or maybe months to get to similar results. They first use "analog" techniques like buying a city map and start marking which neighborhoods they kind visit and expending weekends doing so.

Here machine learning proved to be a very useful tool to very easily classify neighborhoods.

A future use of the present study will be to match this information with house prices to create a recommendation engine to optimize the buying of a property.