Coursera Capstone Proposal - Suicide Rates

Manuel Ramsaier
RWU University, Germany

March 27, 2020

Abstract

In the final capstone project of the coursera course "Applied Data Science Capstone", I will research the rural areas around my location: Oberschwaben, Germany. The focus is (randomly) on different gas station brands which are represented in this area. A particular research question is: As a CEO of one of the gas station brands, where would you open your next gas station just based on the competition. Where are low density communities where your brand is not present yet. This and other interesting findings where the outcome of the capstone project

1 Introduction

In the previous examples, we looked at comparing neighbourhoods from Toronto or New York. Big cities with huge skyscrapers and god knows, those are beautiful places to live. But what about the districts in southern Germany, where I come from. Have a look, what I am talking about in Fig. 1 and 2:

Figure 1: Landscape of Oberschwaben, Germany.

I want to apply the lessons learned in a more rural context. Which districts are similar, which districts have most inhabitants and so on. Instead of neighbourhoods I am interested in districts and communities. So area-wise its a larger scale and we have to deal with the lack of data.

Figure 2: Map of Oberschwaben, Germany.

Germany is divided into several states which have state government as well. Those states have districts and those districts then have communities. For four districts around my location and 132 communities I want to analyse the following problem statement

2 Problem Statement

Imagine, you want to invest in a gas station but you are not sure, where to start it. There are lots of different variables to take into account. As it will be a franchise, you will open a gas station of a popular brand, like Aral (https://www.aral.de/) or Esso (https://www.esso.de/). Now, you have the data about the communities (where they are located) and with the foursquare API you can get all the gas stations for each community. You want to find out, where which gas station dominates the community and therefore, where it would be interesting for you, to start your own gas station.

3 Data used

First I will import the geojson file into a dataframe and extract the features I am interested in, namely the descriptive name and location of the district. This information is then used to get venues via foursquare.

Let me visualize this for you, the data used is listed below:

- GEO JSON File for the districts: Fig. 3, small File
- GEO JSON File for the communities: Fig. 4, larger File (but still very compact)

Figure 3: Districts, we will take a look at

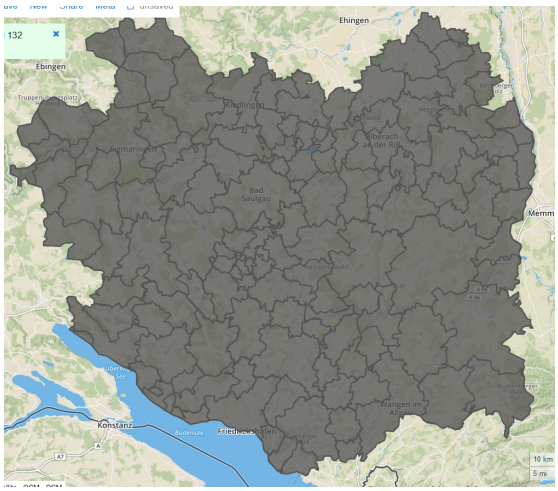

Figure 4: Communities, we will take a look at

4 Data Acquisition, Data Cleaning and Feature Selection

4.1 Data Acquisition and Data Cleaning

The data was rather simple and straight forward. Nevertheless, data cleaning was necessary and two problems where dealt with:

- "," vs. "." - So Germans like to put the decimal point as a "," which is rather unsatisfying, as numpy cannot deal with it straight away. So the "," is replaced by "."
- encoding issues. Also a German thing. Those ä,ö,ü,ß and so on are only recognized correctly, if you enable the utf-8 encoding, which I did.

Not a problem directly, but also required some work: The two jsons have to be merged or better: mapped. There is (thank god) a common identifier, which can be used, called "properties.NUTS". I don't even know what it stands for, but it drove me nuts until I found that pretty function: map(). So it was easy to include the district information in the dataframe from the community json file.

4.2 Feature selection

The seletion of features was easy. I wanted the district's name, the communitie's name, the coordinates (latitude and longitude), the population data (as for future use, did not use it in this study) and the geometry coordinates of the polygon defining the community (also future use).

I checked the data frame for nan values (which of none where found).

5 Exploratory Data Analysis

5.1 Maps of the gas stations

What first comes to mind is a map, so I plotted a folium map of the different communities (Fig. 5):

Using foursquare I was able to get the venues (here: gas stations) around each community coordinate within a specific radius. Unfortunately I found no way to get foursquare to just use all hits within the polygon. There is a polygon selection option with foursquare, but it must have no more than 15 points. So I went with the radius option instead, knowing, that I have to deal with duplicates later on.

Next, as there are a lot of different brands for a gas station, I decided to include only the top brands, which are in this case: "esso", "agip", "ran", "avia", "bft", "shell", "sb","omv","aral","agip". And please don't hit on me, if I forgot any major brand here.

Here, another round of data cleaning became necessary. As the title of the venue may differ and even include the community name like "Aral Bad Wurzach" where Aral is the brand and Bad Wurzach is the community. To overcome this problem, I lowercased the title, and replaced it with one of the major brand if this major brand was found in the title.

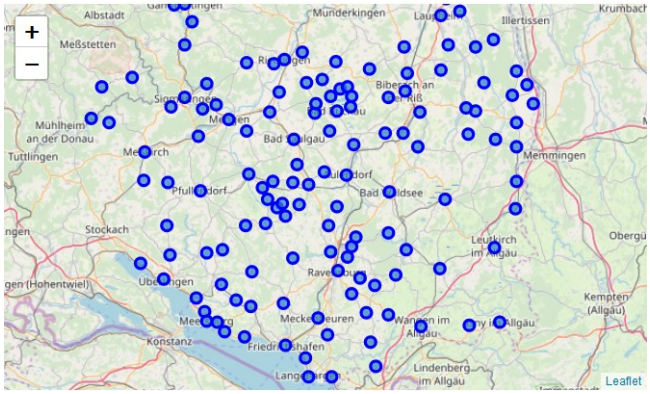

Figure 5: First map plotted in the python notebook containing the community markers which where imported

In total there where 100 gas stations in this area. I also plotted them on a map (Fig. 6):

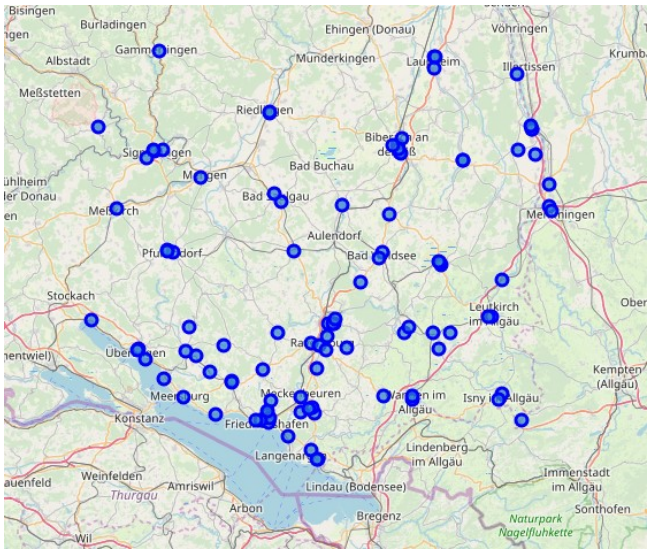

Figure 6: Gas Stations

As you can see, there are less gas stations than communities. So I expect some communities to drop. In fact only 40 communities have gas stations.

5.2 Dominating Gas Station

Via onehot encoding and grouping by community I am able to get the frequency of each gas station brand within each community. As I wanted to plot the top3 gas station brands of each community, I ran into another problem: Many dont even have 3 different gas stations. Welcome to the rural area! So I adapted the functions to deliver NAN values for 0 values. NAN are ignored, so only the top

max(3) existing brands are returned and plotted.

Lets see which brand is represented most within our communities (Fig. 7):

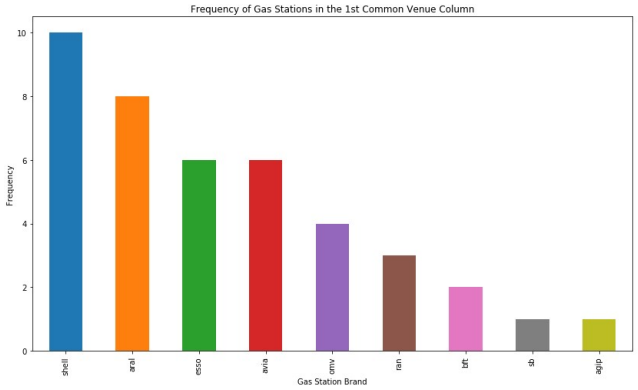

Figure 7: Which brand dominates most communities

In the notebook, there are also the charts for the second most and third most common brand over all communities.

5.3 Clustering based on the frequency table

The frequency of each brand within each community can also be used for a k-means clustering approach. The result is displayed in Fig. 8.

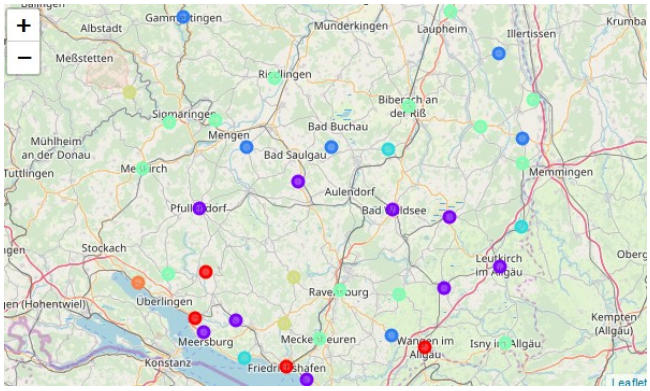

Figure 8: Clustering result with k =7

In order to see if it correlates with the dominating gas station brand, a second map with different colours is plotted in Fig. 9.

6 Conclusion

The battle for gas continued in the suburban areas of Oberschwaben, Germany. We looked at different gas station brands like Aral, Esso and so forth. Using data science, I was able to get insight in where which brand dominated a community.

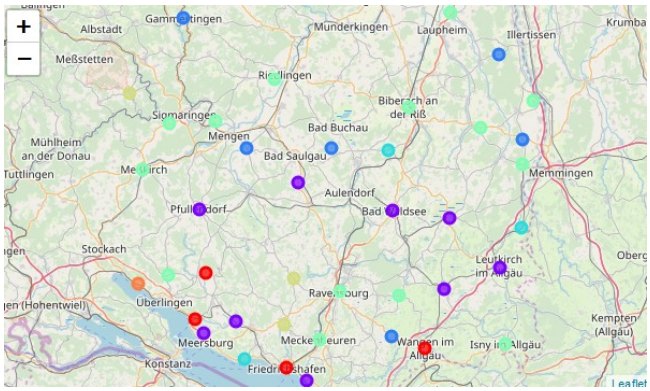

Figure 9: Clustering result with k =7 and dominating gas station brand

But why? First of all let's say you are the Aral CEO and want to find out where to open your next gas station. This approach here can help identify where low density areas are and therefore, where it makes sense to open up a new gas station. Of course this can only be a starting point. You would have to include traffic and population as well. I will maybe enhance this in the future because I really liked working on it.

Now what are the next steps: First, the population data is availabe in this data set. To include this data would be straight forward. Traffic is not so easy and requires further research. As this was just a two week's project, I am quite happy with the results.

Thank you for your time and your interest.

May the force be with you :-D