**Project report – “Finding a good neighborhood for opening a bar in Cologne, Germany”**

1. **Introduction**

For this project I collected data of the city of Cologne in Germany in order to give advice to people, that want to start a business in cologne and want to find an appropriate neighborhood based on specific criteria.

As an example I want to give advice to a client, who wants to open a bar in Cologne, Germany. The target group of the bar is younger people (about age 18 to 35).

My hypothesis is that it would be a good idea to choose a neighborhood with a high population density, a high proportion of young people and many bars in the surrounding.

In order to run a successful bar you first of all need many customers to visit. The rest is up to the good management of the bar. As an estimator for the amount of customer, that visit a bar, I will use the amount of tips, that are listed for each bar in Foursquare. In order to test my hypothesis I want to perform a regression analysis with the proportion of young people, the population density and the bar-counts of each neighborhood as independent variables.

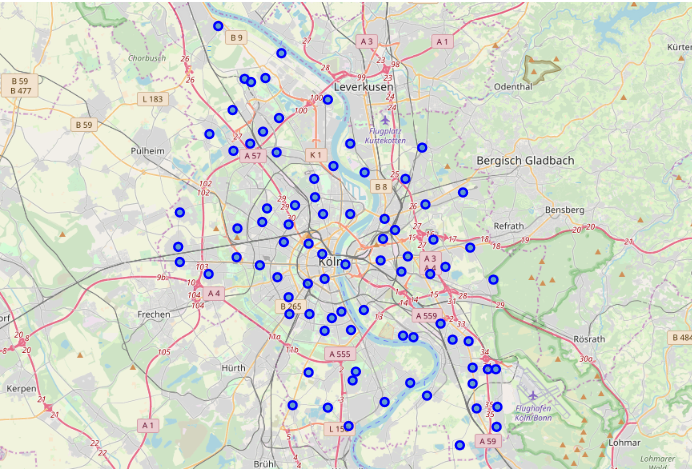
Based on the results of that analysis I want to recommend a neighborhood of cologne that is best suited for opening a bar for young people.

1. **Data Description**

The data I used for this project, I downloaded from the webpage <https://www.offenedaten-koeln.de>. Here I found two tables, that are interesting for my research question. One table contains information about the population in each neighborhood by age (https://www.offenedaten-koeln.de/dataset/fdeb99c8-8640-4539-b342-2eeaf06eefc9/resource/3c2bee9a-fc8b-437f-aa18-727e9b8aa5ef#{}) and one table contains information about the density of the population in each neighborhood (https://www.offenedaten-koeln.de/dataset/28b8e786-edde-4744-b245-13439e9c1ed8/resource/eae7348a-309f-4560-9b93-7f8504a9a40e#{}).

In total there are 86 neighborhoods in cologne. Three neighborhoods were excludes from the analysis, because of missing data for “citizens per square kilometers”. All three neighborhoods had a small population with only a small proportion of young people and would not have been interesting for the analysis. The following map shows the remaining 83 neighborhoods in cologne, which were taken into the analysis.

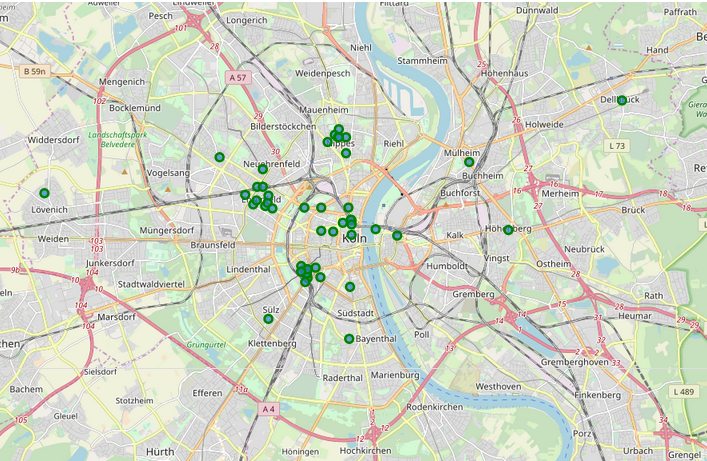
Figure 1: Map of Cologne showing the different Neighborhoods



To get the longitude and latitude for each neighborhood in cologne I used “Nominatim” as a tool. I selected the interesting columns from the above described dataframes and combined them to get my final database.

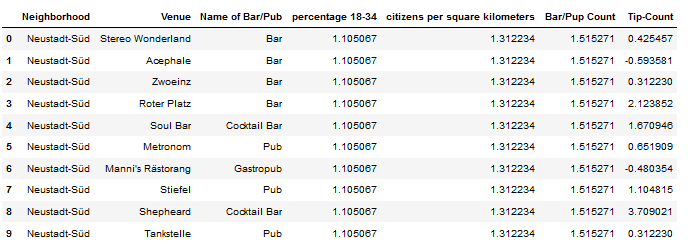
Finally I found the nearby venues of each neighborhood in Cologne with the Foursquare API. I chose a radius of 600 meters for each neighborhood, since increasing the radius led to double hits for found venues. There were 1225 venues for all neighborhoods in cologne. 50 of the venues were listed as bars or pubs. Figure 2 shows a map of all the listed bars in Cologne.

Figure 2: Map of Cologne showing the Bars listed on Foursquare



In a last step the Foursquare API was used to find a visitor-count for each bar. For the regression analysis I built a table with all 50 bars/pups and the dependent and independent variables of interest (Table 1)

Table 1: Dataframe used for Regression Analysis, showing the first 10 entries



1. **Methodology**

**Correlation:**

I first looked at the correlations of the three independent Variables (🡪“percentage 18-34”), the population-density (🡪“citizens per square kilometers”) and the bar-counts per neighborhood (🡪”Bar/Pub”) with the independent (🡪”tip-Count”).

There was a significant positive correlation for the Variables “Percentage 18-34” (r = 0.40, p= 0.00), “Citizens per square kilometres” (r = 0.51, p= 0.03) and “Bar/Pup Count” (r = 0.52, p= 0.00).

**Regression:**

I next performed a regression analysis with “Tip-Count” as the dependent variable. I calculated different polynomial degreed and different combinations of dependent variables to find the best model fit. The best model in terms of fitness and scarsity was provided using only the variables “percentage of young people” and “citizens per square kilometers” with a polynomial degree of 2. As Figure 3 shows the model is significant with a reported R-Square of 0.34.

Figure 3: Polynomial regression (degree 2) with “percentage of young people” and “citizens per square kilometers” as independent variables

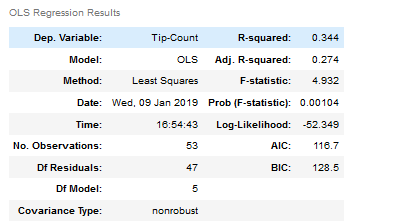
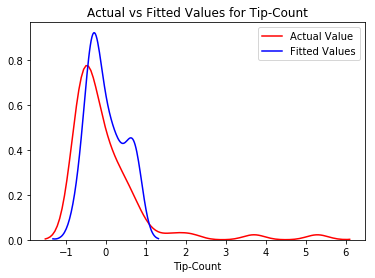


Figure 4 shows the Distribution-Plot of the Model

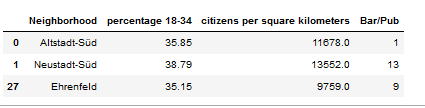
Figure 4: Distribution-Plot



1. **Results**

To answer the question of my customer, which neighborhood would be suited to open a bar in Cologne I selected all the neighborhoods that have a high Value for “percentage of young people” and “citizens per square kilometers” (percentile =0.9). Table 2 shows the result of this selection.

Table 2: Selected neighborhoods in Cologne that are suited for opening a bar



1. **Discussion**

Since the foursquare API does not provide information about the neighborhood each venue belongs to, I used an area of 600 meters of the latitude and longitude for each neighborhood to search for the bars that are in the centers of these neighborhoods. With this method I cannot be sure to find all the bars, that are located in that area, especially in the bigger neighborhoods. Increasing the radius would on the other hand have lead to double hits for the bars (bars belong to the different neighborhoods). Also the Tip-Count Value is only an estimation of the actual visitor-Count. The Foursquare API did not provide a value for that variable. Both problems could be addressed by trying different API’s and look for more specific information.

Nevertheless the collected data can be provide a good base for selecting neighborhoods that are suited for opening a bar. It is also possible to put higher or lower constrains for the selection by increasing or increasing the percentile of the dependent variables.