Capstone project report – Markus Engler

1. Introduction where you discuss the business problem and who would be interested in this project.

Main idea of the project is to compute a tool to search for the optimal geographical location for opening a specific new venue in a German city. The analysis should be valuable to all business owners or public organizations looking for a new location for their projects.

The user should be able to specify the region/city of interest. The tool will then request data of existing venues for this region via the Foursquare API, using a combination of zip codes and geographical information. Afterwards a first analysis will be performed on the received dataset of existing venues in the area of interest. The analysis will enable the user to choose the type of venue easier and to have a first general view on the availability of such venues in the region.

The tool will than filter the data set for the chosen venue category and will calculate a distance matrix of all distances between each venue in the area of interest. The two nearest venues for each datapoint will be used to find the optimal location for a new venue of its kind by using centron calculation and trigonometric methods. The resulting coordinates will indicate the location which has the maximum possible distance to all existing venues of its kind in the area of interest.

For simplification reasons for the final exam, the resulting location will be optimal location, having in mind that other important variables like population, income level or infrastructure are not considered.

2. Data where you describe the data that will be used to solve the problem and the source of the data.

As first data source I will use information on German city location, which can be found on https://raw.githubusercontent.com/TrustChainEG/postal-codes-json-xml-csv/master/data/DE/zipcodes.de.csv. This data includes the zip code of every city/region of Germany, which are used as identifier for the area of interest in this project. Furthermore, the data contains the latitude and longitude data of the centron of the corresponding areas, so the geographical information does not have to be added.

In a second step I will enrich the location data with venue information of the area using the Foursquare API. The enrichment process will be like the New York/Toronto example. The resulting data frame will include the nearby venues in the area of interest with additional venue information available, like name, category or geographical information. Example:

```
[12]: frankfurt_venues.head(10)
```

[12]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	60306	50.1159	8.6702	Opernplatz	50.115399	8.671772	Plaza
1	60306	50.1159	8.6702	Alte Oper	50.115350	8.671428	Opera House
2	60306	50.1159	8.6702	Manufactum brot&butter	50.115958	8.670772	Food & Drink Shop
3	60306	50.1159	8.6702	The Ivory Club	50.114309	8.669109	Indian Restaurant
4	60306	50.1159	8.6702	ZENZAKAN - Pan Asian Supperclub	50.114867	8.669458	Asian Restaurant
5	60306	50.1159	8.6702	Moriki	50.113863	8.669530	Japanese Restaurant
6	60306	50.1159	8.6702	Schneider's Café Snackbar	50.115685	8.669928	Café
7	60306	50.1159	8.6702	Kameha Suite	50.114732	8.670210	Mediterranean Restaurant
8	60306	50.1159	8.6702	MEYER Feinkost Frankfurt	50.114819	8.673398	Gourmet Shop
9	60306	50.1159	8.6702	Sofitel Frankfurt Opera	50.116294	8.673624	Hotel

This dataset will then be used to perform the calculation described above to determine the optimal geographical location for an inauguration of a new venue of the chosen type.

3. Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.

First, German zip-code data is downloaded and filtered for specific city or region. Than the geo-coordinates for the chosen city are requested via geolocator function to generate a map of the city/region and the nearby area, including all city districts.

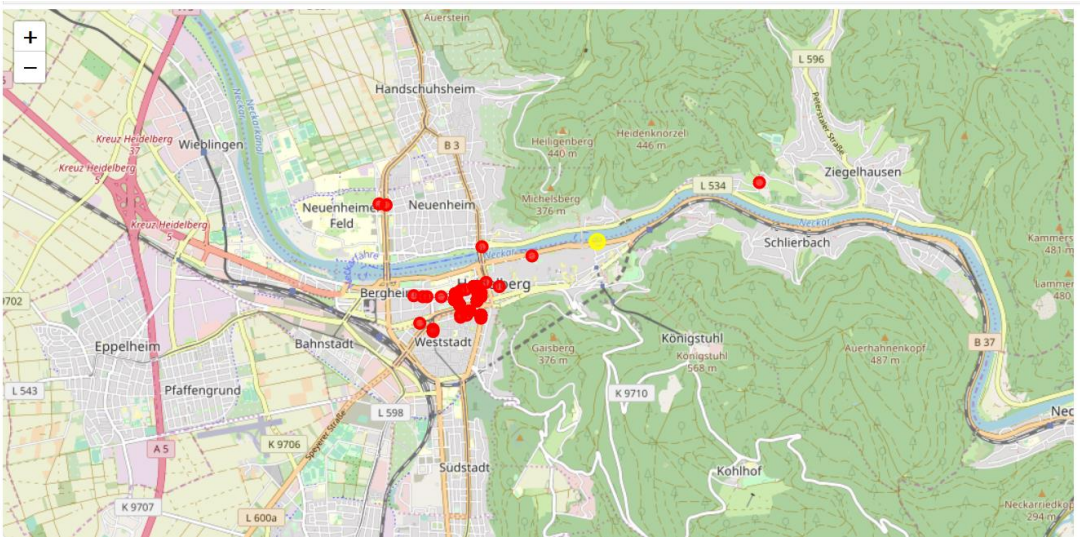
Second, the venue information for all districts is requested using Foursquare API. The 10 most common venues for each district are displayed to enable the user to chose from existing venue categories. The user can also use any other venue category and check presence of venue type in the requested data.

Third, all venues of chosen type are displayed on the map of the city/region of interest and a pairwise calculation of distances between venues of the datasets is executed. The resulting distance matrix is used to assign the two nearest venues to each datapoint (venue). From these three venues a triangular is constructed to calculate the centron of the involved venues. Afterwards the distances of the centron to each of the venues is calculated.

Fourth, to find the coordinates for optimal venue placements, the algorithm looks for the maximum of the sum of distances between the centron and the venues. This calculation provides the coordinates of the centron with the highest possible distance to all venues in the dataset. Centron of triangle is used to find starting points for calculation in first place, but also to restrict the result to coordinates within the city/region boarders.

4. Results section where you discuss the results.

The results will differ with each chosen city/region and with each chosen venue category. For documentation purpose I have chosen the city of Heidelberg in Germany and Restaurant as venue category. The resulting coordinates of latitude 49.41 and longitude 8.71 indicating the coordinates for optimal placement of new venue of type Restaurant in the city of Heidelberg. The yellow circle on indicates the coordinates on the map, with red circles indicating existing venues:



5. Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

There are some possibilities to improve the computed tool for optimal venue placement calculation based on findings which where made during testing and coding:

- Resulting coordinates could lead to venue placement in rivers, lakes or on hill tops. To avoid this issue some more geo information about the sounding area could be included in the dataset.
- Not all venues in the city/area of interest are included in the Foursquare database which could led to missing data or existing venues are not shown during the use of the tool. The missing venue data and/or missing venues could be generated combining other data sources like google map to get more valid results from the tool.
- The optimal venue placement is only based on distances between existing venues (in the Foursquare database). The overall optimal coordinates for venue placement can therefore be located somewhere else in the area of the city/region. To find the overall maximum of distances without a starting point, a more advanced computational approach and more computational time is needed.
- City/region features like population, income level or infrastructure are not included in the calculation of optimal venue placement. To solve this problem a scoring system could be implemented to further evaluate the recommended location. The data and information to calculating the score of each possible venue placement have to be collected from several other databases and be assigned to the venue data.

6. Conclusion section where you conclude the report.

All in all, the presented tool gives an indication for optimal placement of new venue of specific kind in a city/region of interest. The user can choose from all cities/regions in Germany an is able to specify a common venue category based on existing top venues or own choice. The resulting coordinates give the optimal location based on trigonometrical calculations using existing venues of the specified typ. Calculation is based only on available venues in the database and uses distances between venues as optimization criteria only.

Several improvements of the basic version of the presented tool could be made in further work, presented in section 5. This will further improve the result and will ease the decision making procs based on the results, due to more information already included in the optimization process (population, income, geo information, other venue databases).

As for now, based on restricted time and computational resources, the result can provide a first indication of optimal placement or area for optimal placement of new venue. In the next steps the result should be enriched with several other datasets to improve information content and should be checked for accuracy for final decision.