The Battle of Neighborhoods
Capstone Project

Kevin Andoni

February 17, 2021

1 THE PROJECT

1.1 Introduction

The goal of this project is to provide our business partners in Vienna with a clear empirical, visual and descriptive analysis of the best location to open their next Coffee Shop.

1.2 Background

Vienna, the capital city of Austria was recently classified again as the most liveable city in Europe (link). Among the important factors taken into account in this announcement, would be the broad spectrum of cultural and leisure activities. Coffee is one of these activities/passions that bring people together in this city. This is also one of the reasons that this destination has become the top choice of numerous investors for opening their Coffee Shop in.

2 THE DATA

2.1 Data Collection

The data will be collected from different trusted sources that are released by the government to open public, from google API services, from open GitHub resources. For calibrating our geospatial locations with the EPSG dataset we will use the GeoRepository web services which adopt the latest ISO standards and is also open for public use. These resources will help us create our detailed dataframes and geo-dataframes which will include: a highly accurate mapping of the streets that would be of interest to us, the respective districts, the distances from the center and many more.
To get updated information about businesses inside our area of interest, we will use one of the most powerful and trusted resources, the Foursquare data platform. The data will be accessed by using the API Developer tool (offered by the website) and then categorized to a dataframe accordingly.

2.2 Data Preparation

All the dataframes will undergo to continuous appending and filtering processes. As we progress with our analysis, we will benefit from multiple visualizing tools that will help us create relations in and between different sets of data. Among these tools, we could list the variety of the folium's tilesets including CircleMarkers, FastMarkerCluster, GeoJSON-TopoJSON Overlays as well as HeatMaps.
As we make progress in our research, by continuously refining our search parameters and filtering our list of results, we should be able to find our best candidates of places.

3 METHODOLOGY

This project followed a fixed strategical path. It began with the import of all the library of the tools that were used. After a careful territory analysis, it was concluded that the area of interest would include all the streets within a distance of 1.5 km from the city center *Stephansdom* (St. Stephan's Cathedral).

The next step was to retrieve accurate and trusted data about the district divisions and street mapping. This helped in categorizing the streets that were investigated. To make this project safe of possible future changes of the online databases, a copy of this *GeoJSON* file with all the data about the district division was saved adjacent to the project notebook. In the street maps, a simplification method of the *Street Geometry* to *One Point Coordinate* was be applied. The advantages of this process, consisted that these street representing points had the additional function of acting as the centroids, from which the distance of the surrounding venues was measured. This provided a highly detailed density mapping of the venues across the total area of interest.

At this point on the project a connection to the Foursquare database was established and data for all the venues inside the area of interest was retrieved. This was then processed and visualized using tools that included **Folium** and **HeatMap**.

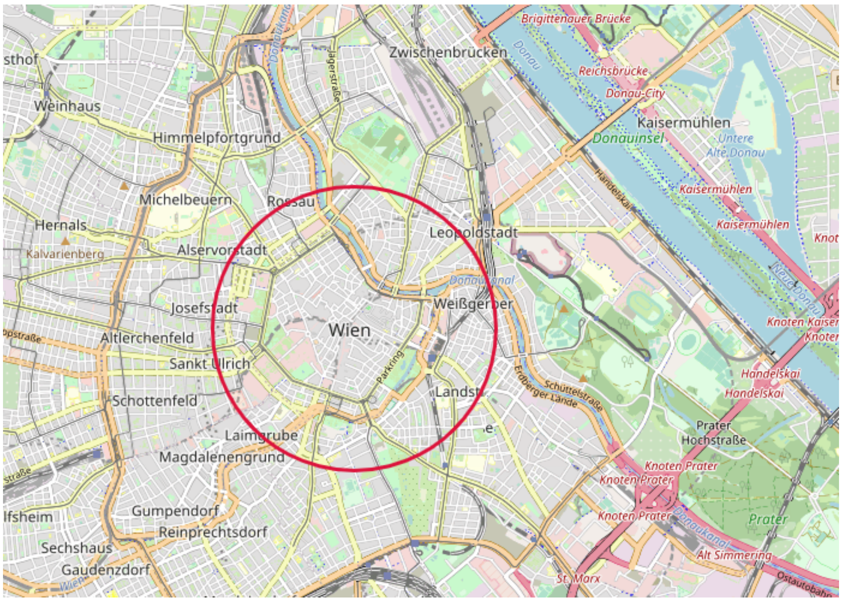
The analysis could finally begin after all the data was retrieved and prepared to meet the research standards. The functions *"coffees in vicinity", "closest coffee" and "distance from center"* that filtered 'the dataframe of results', so that from the total list of locations, only those meeting the three selecting criteria would be retrieved, were defined. The *candidates* dataframe containing the results of the filtering process was built by filtering out the locations with more than two coffee shops in vicinity.

After the values of the created dataframe were normalized, they could finally be ready to be processed by the **k-Means Clustering** machine learning model. To find the optimal number of clusters, the *Elbow method* was chosen. At this point the model was finally deployed and was able to provide four different clusters of place candidates. These were filtered one last time, so that in the end, only one area with the best among all candidates remained.

4 Data Exlporation

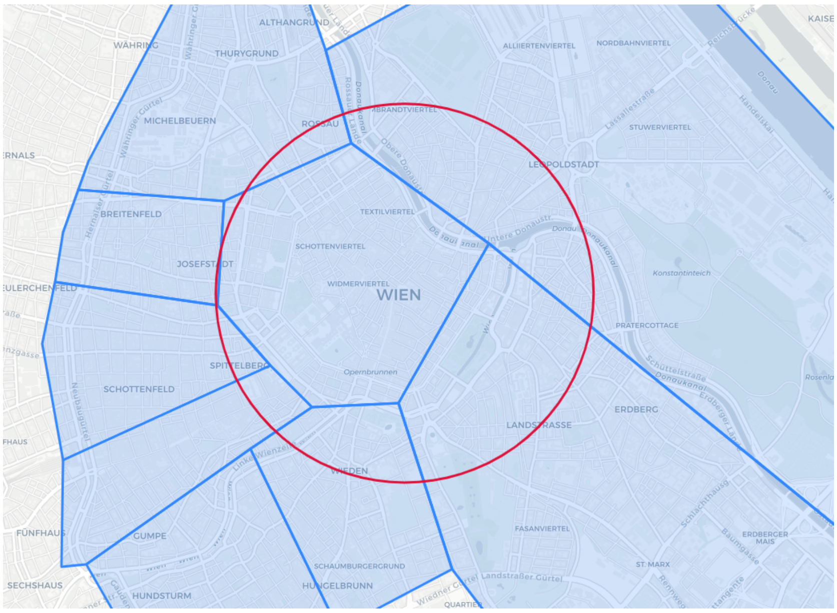
In this part of the report, the methods described in the previous section are going to be put into action and the described path of the project is going to further explained and accompanied with the findings.

The first step taken in this project after the import of all the necessary libraries, was the selection of the area of interest. This was set as the circle with radius of 1.5 km and in the map it looks as follows:



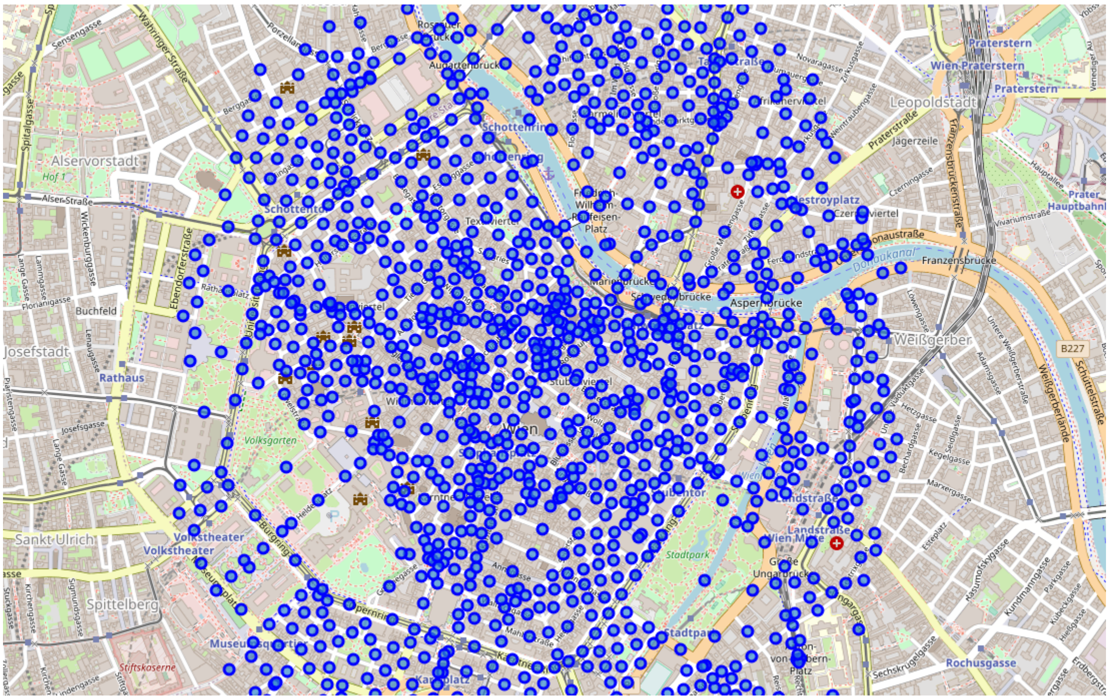
Since one of the criteria that was worked upon, was the calculation of the distance from the center point, it was useful to save this location also as a geopandas 'geometry point'.

The next step consisted in retrieving detailed information about the district division. This helped categorizing the streets that will be investigated. After being converted to a dataframe, the resulting intersection with the area of interest gave the following results:

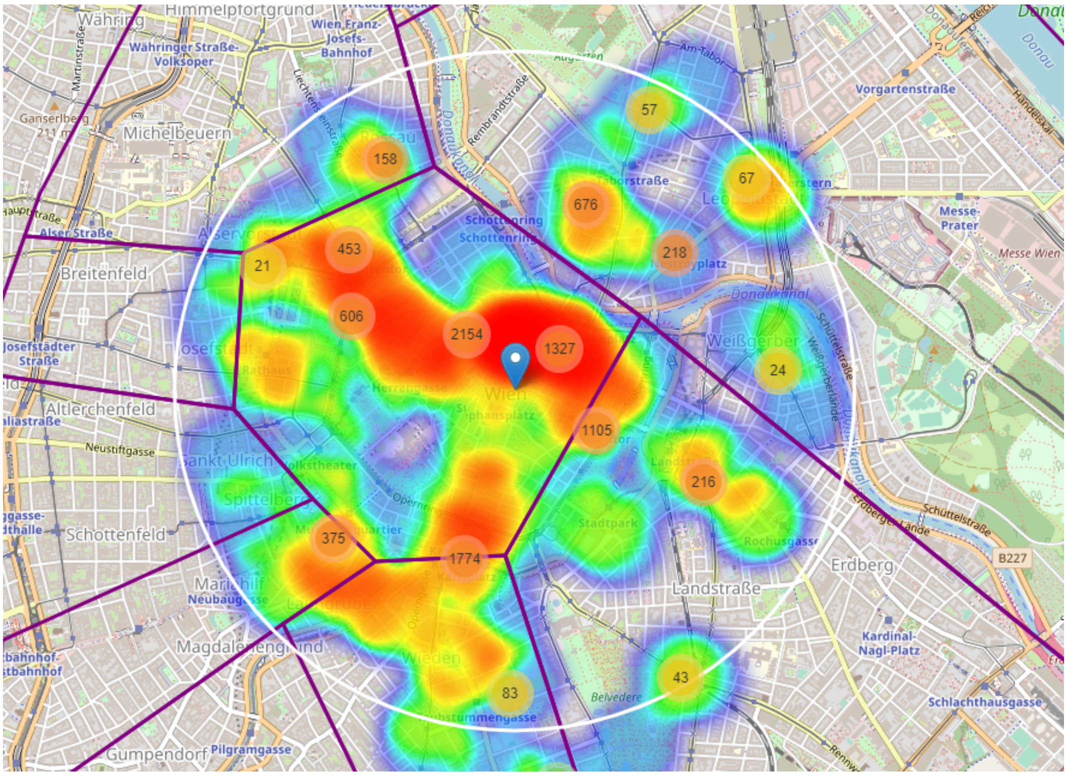


At this point it was observed that the area of interest contained all the streets of the first district and in part those of the sorrounding districts. This helped in considerably reducing the size of the streets database in which it was researched for potentially good locations. By filtering the streets and storing the results to a organized dataframe it was noticed that large proportion of around 57 % of the streets was located in the first district and it would be logcial to assume at this point that the best area of interest would also lie in this district.

The next step was to create a method that would simplify the Street Geometry to one Point Coordinates. The advantages of this process, consisted that these street representing points would have the additional function of acting as the centroids, from which the distance of the surrounding venues was going to be measured. This proved to be helpful in creating a highly detailed density mapping of the venues across the total area of interest.



Now the process of retrieving venues nearby the centroids that were created could begin. To achieve this, a connection to the Forsquare database through the Developer API was established. The first step is creating a function that would have as input parameters a searching radius, the street name and the latitude & longitude values. The function could firstly make an API request to the Forsquare server, then make a GET request, append the retrieve data to an array and finally create & initialize a dataframe with the complete array. From trial and error it was learned that the function takes a considerable amount of time to retrieve all the venues and sometimes gets interrupted. To solve this, after successfully retrieving all the venues of interest and stored to a dataframe, a copy was saved adjacent to this notebook file. This is available along with all the other files on the github repository of this project. The resulting dataframe containing all the coffee shops in the area of interest looks as follows:



5 Conventional Analysis

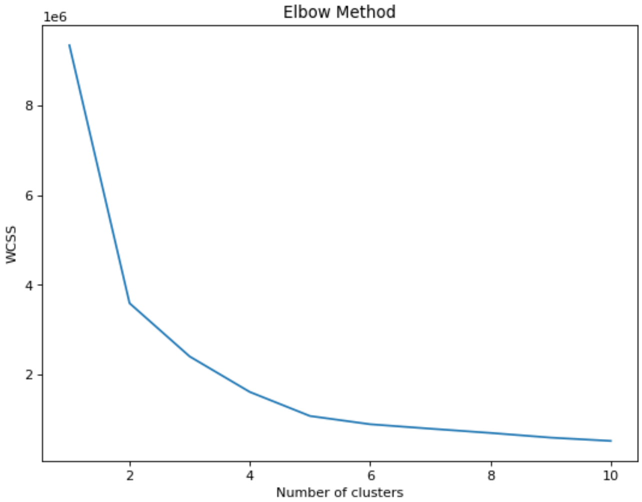
The analysis could finally begin after all the data was retrieved and prepared to meet the research standards. The functions *"coffees in vicinity", "closest coffee" and "distance from center"* that filtered 'the dataframe of results', so that from the total list of locations, only those meeting the three selecting criteria would be retrieved, were defined. The *candidates* dataframe containing the results of the filtering process was built by filtering out the locations with more than two coffee shops in vicinity.

6 MACHINE LEARNING ANALYSIS

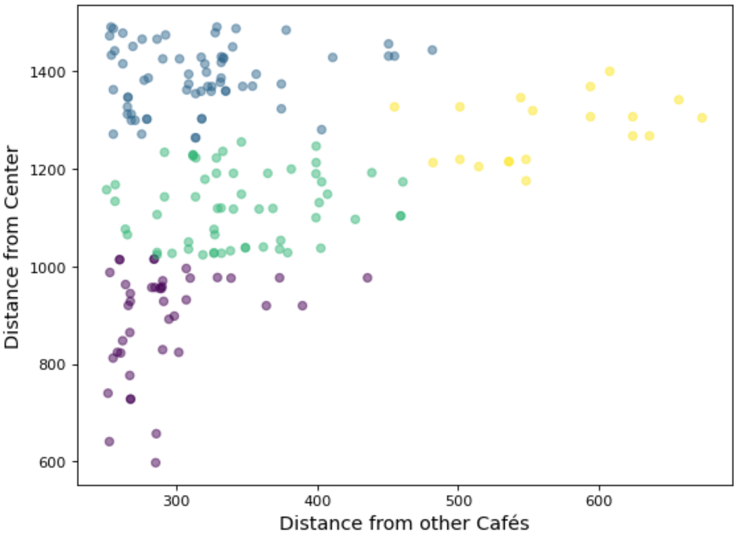
The results dataframe now consisted of valuable information of the three main selection criteria (Closest Café distance,Number of close Cafés,Distance from Center), however it was unlabeled and the best results could be found by segmenting it. One of the best machine learning models for this type problem, would be the *unsupervised learning* approach of **k-Means Clustering**.

An important factor that was additionally considered, was the **normalization** of the values. This guaranteed that the results of the analysis would not be affected by the relative major differences in value between the different columns (e.g. the number close coffee shops was set to range between 0 and 2, on the other hand the distance from the city center can have a range of up to 1500 m).

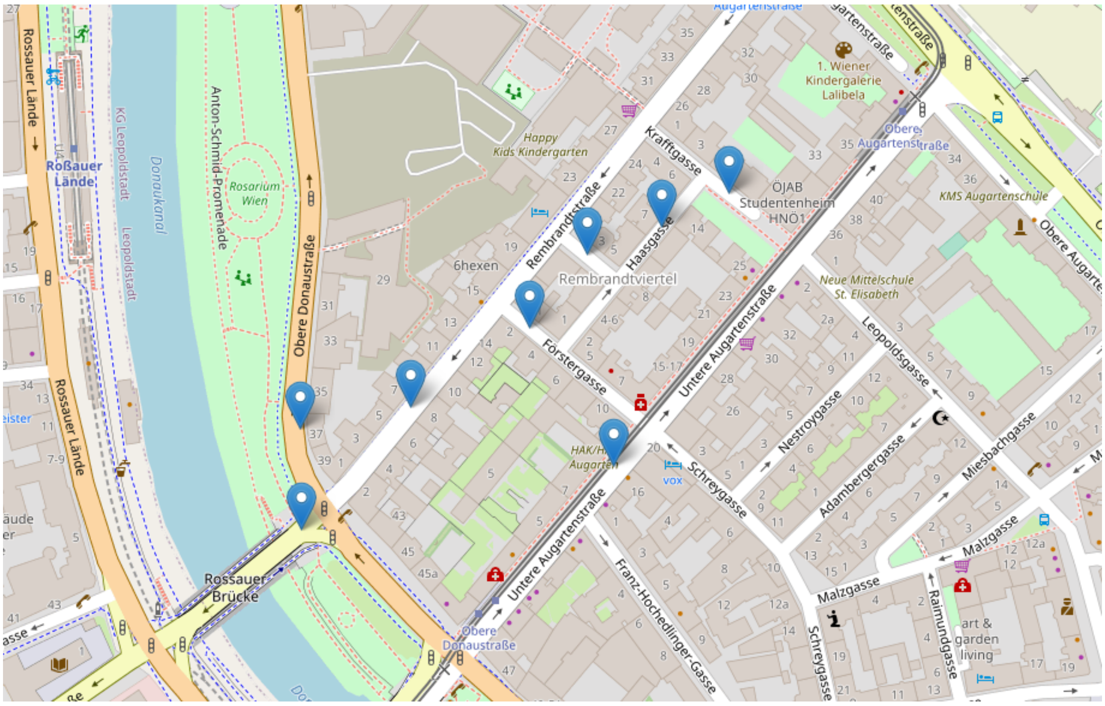
As a possible disadvantage of the k-Means Clustering, could be considered the manual decision of the cluster number. However this was prevented by making use of the *Elbow Method*. This method returned the optimal number of clusters equal to *four*.



The data could then be considered ready to be fitted to the model. Finally after being processed, the following results were retrieved:



As it could be noticed in the plot, the yellow cluster contains locations that are at a relatively small distance from the city center and at the same time, the distance from other Cafés considerably higher for this cluster. By investigating the results closer, the street Förstergasse could be set as center point of this cluster. The first assumption (the area marked with the red circle) is not considered as the best location. The best area to be considered for potential coffee shop locations is located in the North direction (marked with the green circle) of the city center.



7 Results and Discussion

By using the distinctive parameters 'Distance from Center','Distance from other Cafés','Number of close Cafés' for our Machine Learning k-Means Algorithm and the 'Clustering - Elbow Method' it was found that our areas of interest in which we would consider finding potentially good locations for our investment could be divided into four clusters. By then optimally adjusting the distance parameters, our machine learning model recommended one these clusters as favourite for the investment.

8 Conclusion

These findings that we mention however, come as a result of a pure empirical analysis. Should the investors have any non-empirical evaluating preferences, such as the local architecture of the buildings, the local community of the neighbourhood etc., they should feel free to consider the other good clusters of locations that we found in our analysis.