COURSERA CAPSTONE PROJECT

Opening an Italian Restaurant in Berlin, Germany

Website: https://www.coursera.org/professional-certificates/ibm-data-science

PROJECT REPORT

Eduardo Pomar Makthon

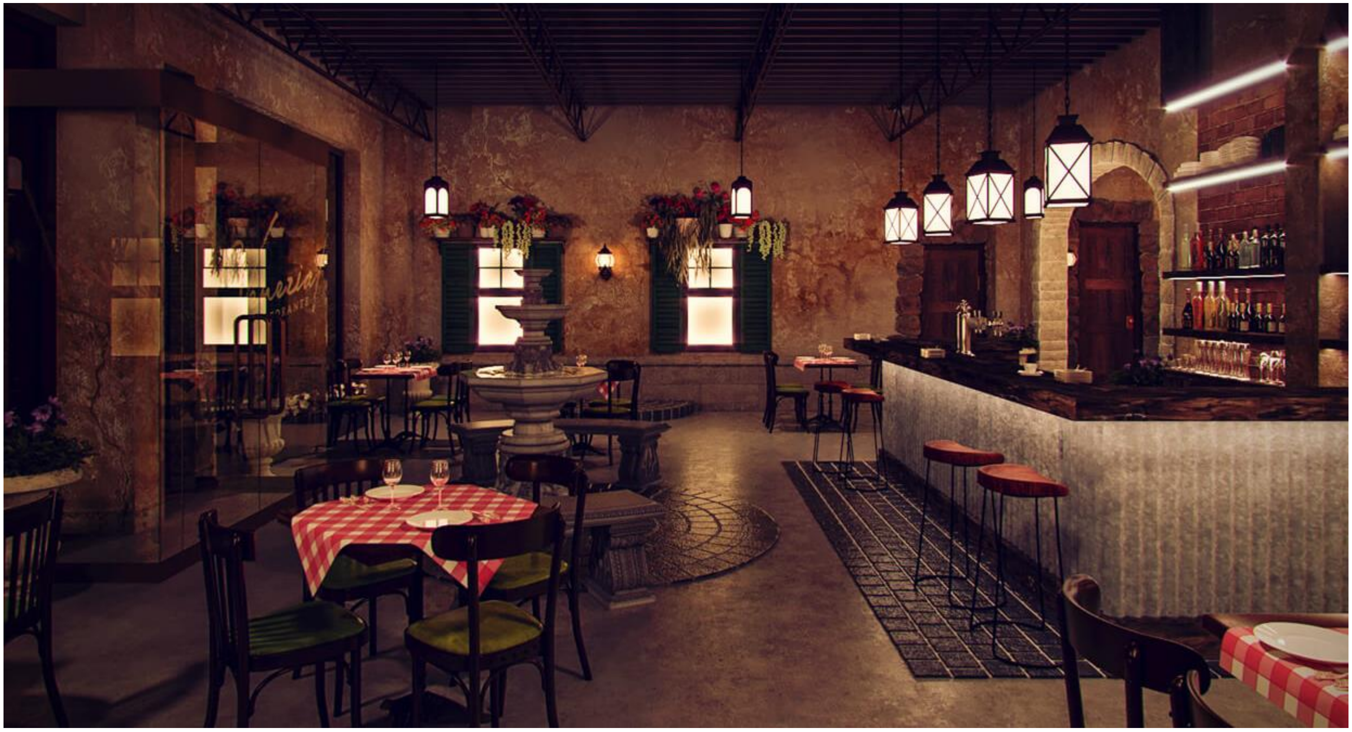
Email: eduardo.pomar.m@gmail.com

CONTENTS

INTRODUCTION

Introduction

Many people like Italian food, as it is an exquisite cuisine. Italian food is one of the few global cuisines that are warmly welcomed in countries worldwide. Italian food regularly features on the dining tables of most urban households, and more often than not, we fall back on pastas, pizzas and risottos to satisfy our cravings for a good meal. There are so many varieties to choose among Italian dishes in veg or non-veg, from when it comes to pasta - penne, lasagna, spaghetti, macaroni, tagliatelle and ravioli among others - that you can toss them in numerous sauces, herbs, vegetables and meats and enjoy a hearty meal. Home-made pizzas are also a favorite option for a quick meal during game nights or family get-togethers.



But as well as it is a really good business idea for a restaurant, it is also very common to have one or two Italian restaurants in your neighborhood already, and that means if one were to open another Italian restaurant, it will probably won't have the same success as if it was open in a neighborhood with no other of these restaurants nearby, as it will have less competition. As in most food business, the location is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

INTRODUCTION

Business Problem

The objective of this capstone project is to analyze and select the optimum locations in the city of Berlin, Germany to open a new Italian restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide insights as to where it will be more convenient to set up a restaurant of this kind.

Target Audience of this Project

This project is made as a request from a fond friend of mine Armando Lingüini, who is moving to Berlin and wants to continue his lifetime job as a pasta cook. However, it can be taken advantage of by any person who is thinking about opening an Italian Restaurant in this city. The only condition to this is that they don't interfere with Lingüini's work.

Data that will be used in this Project

To address the matter at hand, we will be requiring the following data:

- List of neighborhoods in Berlin. This defines the scope of this project which is confined to the city of Berlin, the capital city of Germany.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to Italian restaurants. We will use this data to perform clustering on the neighbourhoods.

Sources of Data and where to find it

In this project, the list of neighborhoods will be extracted from Wikipedia, the exact URL is the following:

https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin

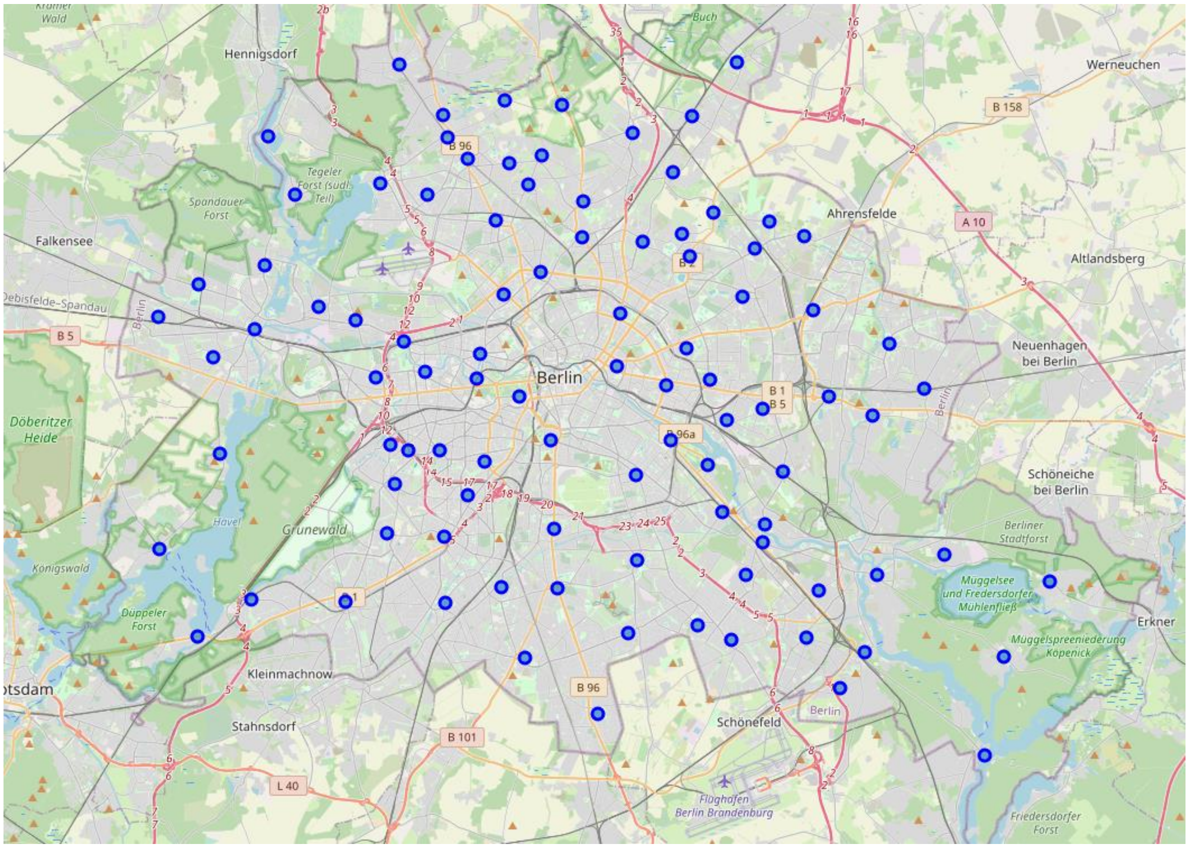
INTRODUCTION

It contains a list of neighborhoods in Berlin, with a total of 96 neighborhoods. Web Scraping techniques will be used to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then, using the Python Geocoder package, the geographical coordinates of each of the neighborhoods will be obtained.

Then, the venue data for those neighborhoods will be obtained using the Foursquare API. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data. In this particular case, the study will be focused around the Italian restaurant category in order to solve the business problem at hand. Also, Folium will be used for map visualization.

Berlin Map

This map represents all the neighborhoods of the city of Berlin.

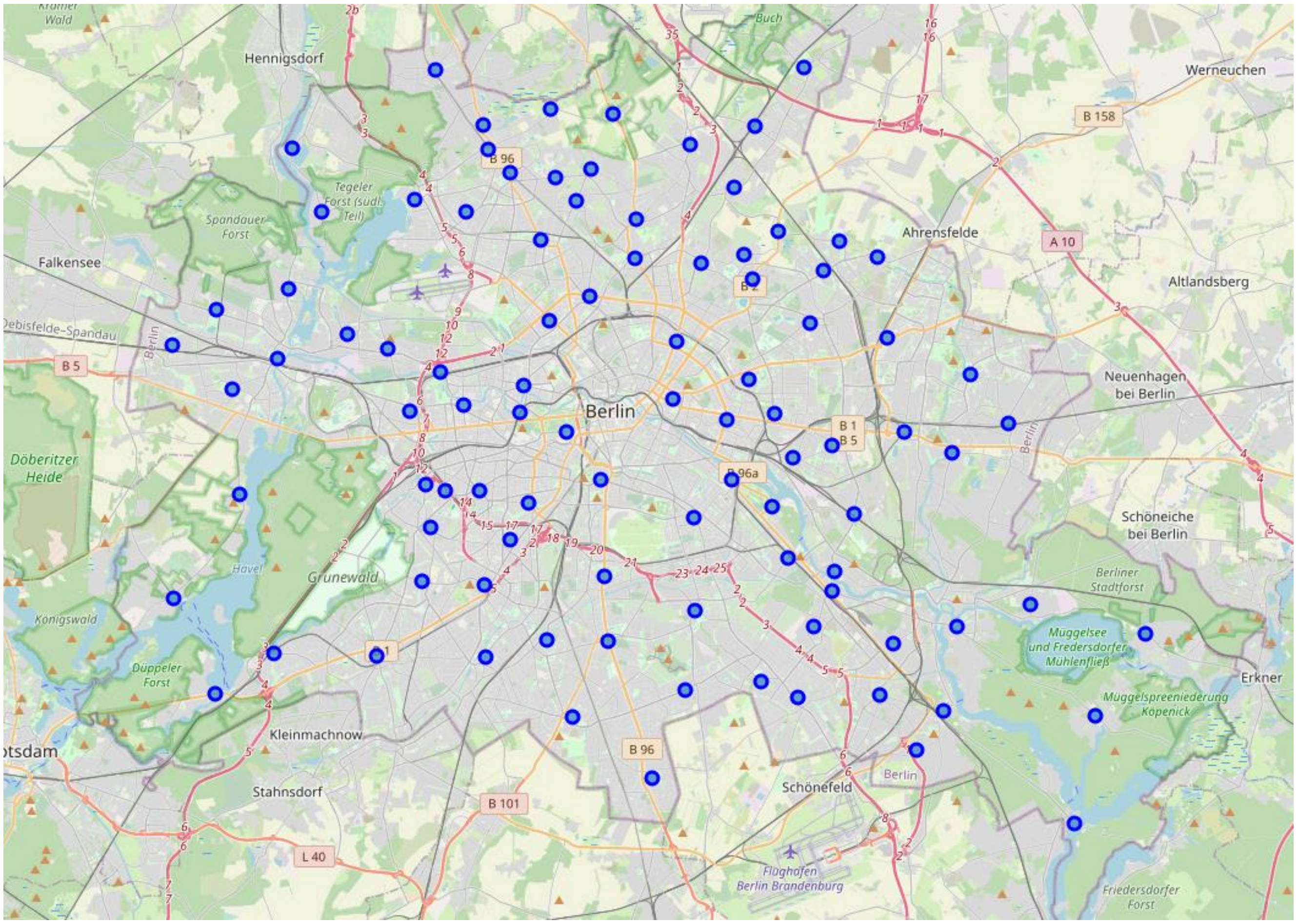


METHODOLOGY

Web Scraping

The first step is to get the list of neighborhoods from Berlin to a dataframe. To do this, web scraping techniques are applied using the BeautifulSoup package.

Having obtained this, geographical coordinates have to be added to the dataframe. The geocoder package is used in order to do this. After doing this, the Folium package is used to show the map of Berlin including the points of the neighborhoods in the dataframe, in order to verify that the geographical coordinates are correct. The result is the following:



METHODOLOGY

Foursquare API

Later, the Foursquare API will be used to get the nearby venues from the neighborhoods in the dataframe. For this, it is required to have a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. Then one must make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and data such as the venue name, venue category, venue latitude and longitude will be extracted.

With the data, it can be known how many venues were returned for each neighborhood and examine how many unique categories (Italian restaurant being one of them), can be extracted from all the returned venues. In this case, there are 262 unique categories.

Data Wrangling

Then, the entries of venue will be grouped by neighborhoods so the total of each venue category by neighborhood can be obtained. This also helps preparing the data for use in clustering. Since the analysis is focused around the Italian restaurant data, only the category of Italian restaurant will be taken into account.

Machine Learning Modeling

After this, a machine learning technique will be used, called K-Means Clustering. What this does is that it creates a number of clusters and assigns each data point to a cluster. It works in a way that it searches to maximize the distance between data points of different clusters, and at the same time, it minimizes the distance between data points of the same cluster.

Metrics

In order to select the optimum number of clusters, a metric called Silhouette Score will be also used. Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters.

The results will allow to identify which neighborhoods have a higher concentration of Italian restaurants while also showing which neighborhoods have a fewer number of them. Based on the occurrence of Italian restaurants in different neighborhoods, it will provide an answer the question as to which neighborhoods are most suitable to open new Italian restaurants.

RESULTS

Silhouette Score

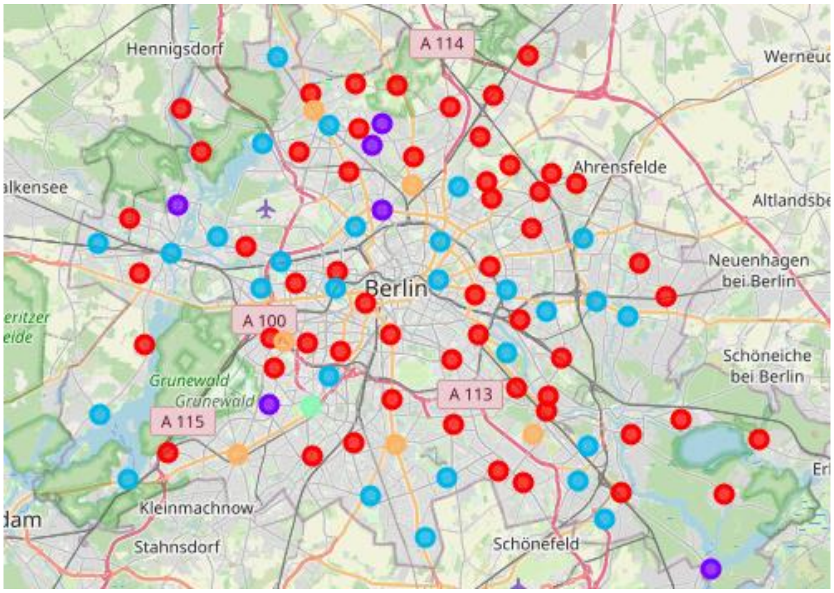
The results from the Silhouette Score for different numbers of clusters are shown below:

- The silhouette score for 2 clusters is 0.7623111393242606.
- The silhouette score for 3 clusters is 0.888850456592392.
- The silhouette score for 4 clusters is 0.9384469696969697.
- The silhouette score for 5 clusters is 0.9895833333333334.

K-Means with 5 clusters

The results of Silhouette Score show that using 5 clusters would be the optimum choice. The results of using 5 clusters are shown next:

- There are 55 neighborhoods with 0 Italian Restaurants, they belong to cluster 0
- There are 28 neighborhoods with 1 Italian Restaurants, they belong to cluster 2
- There are 6 neighborhoods with 2 Italian Restaurants, they belong to cluster 4
- There are 6 neighborhoods with 3 Italian Restaurants, they belong to cluster 1
- There is 1 neighborhood with 5 Italian Restaurants, it belongs to cluster 3



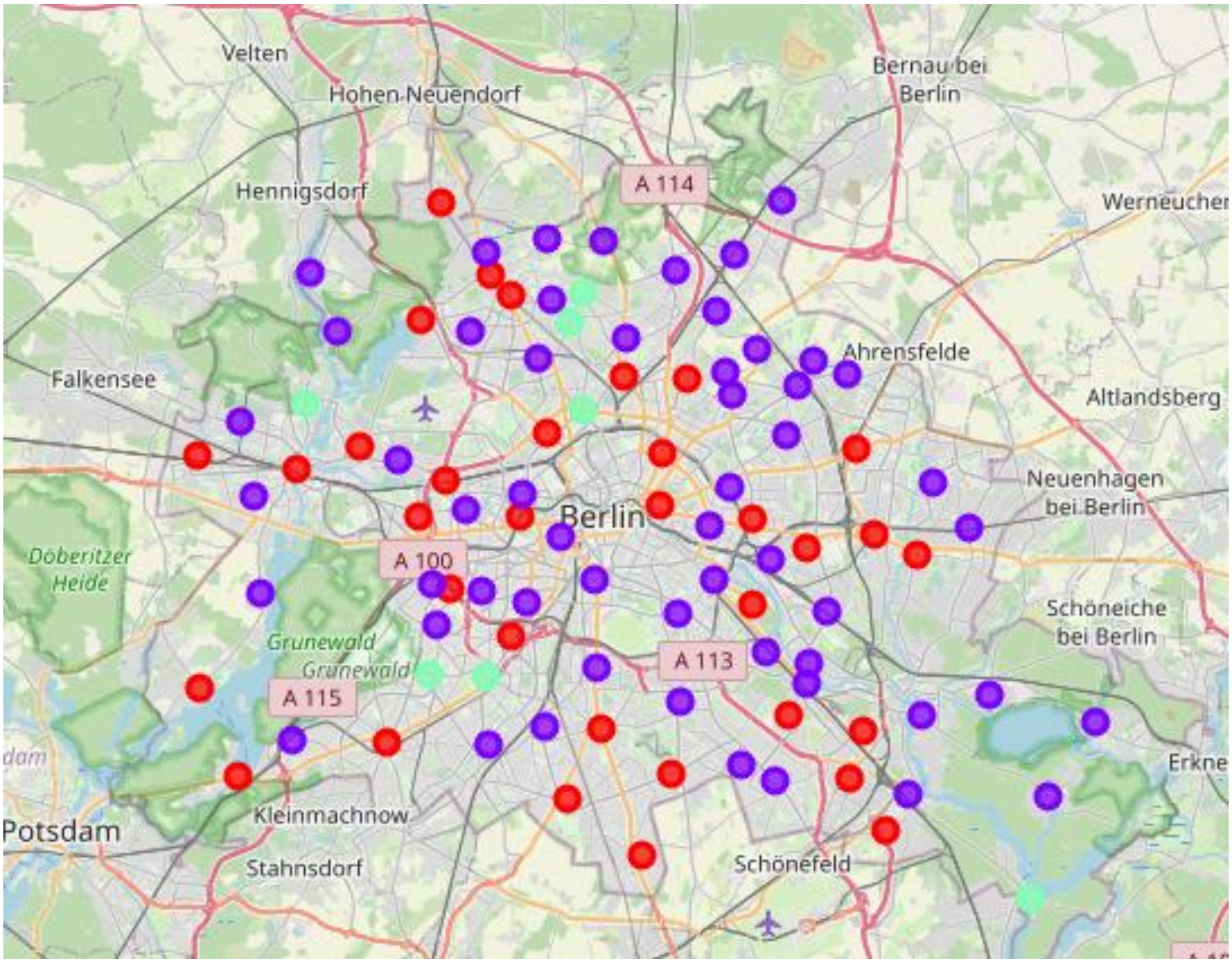
RESULTS

In spite of the results of the Silhouette Score, the clustering result does not provide much help in identifying the best zones in Berlin to open an Italian restaurant, so a second analysis will be made using 3 clusters.

K-Means with 3 clusters

The results of using 3 clusters are shown next:

- There are 55 neighborhoods with 0 Italian Restaurants, they belong to cluster 1
- There are 34 neighborhoods with 1 or 2 Italian Restaurants, they belong to cluster 0
- There are 7 neighborhoods with 3 or 5 Italian Restaurants, they belong to cluster 2



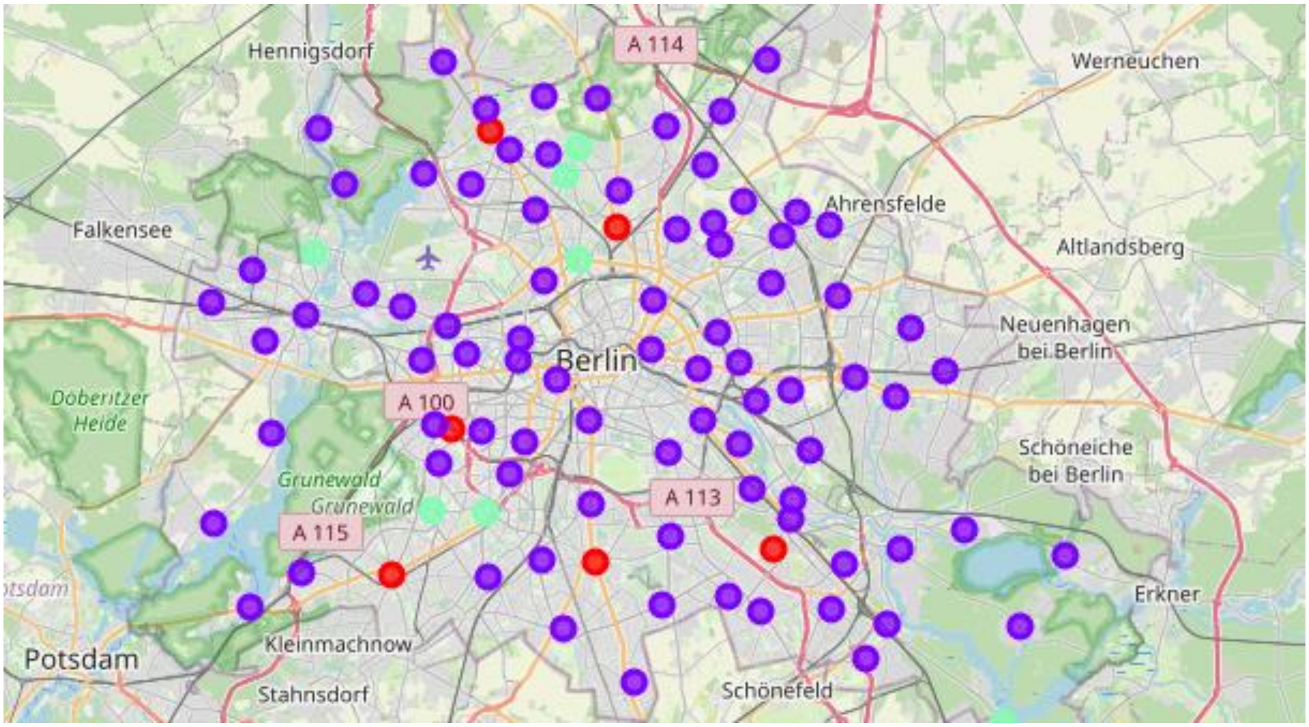
RESULTS

As consideration of the person doing the analysis, the result is better if the neighborhoods are separated in clusters with “few to none”, “medium” and “high” competition, except for one detail.

A neighborhood with just 1 Italian restaurant should be categorized into the “few to none competition” cluster, so the cluster for each neighborhood that has just 1 Italian restaurant, will be changed from 0 (medium competition) to 1 (few to none competition).

The result is the following:

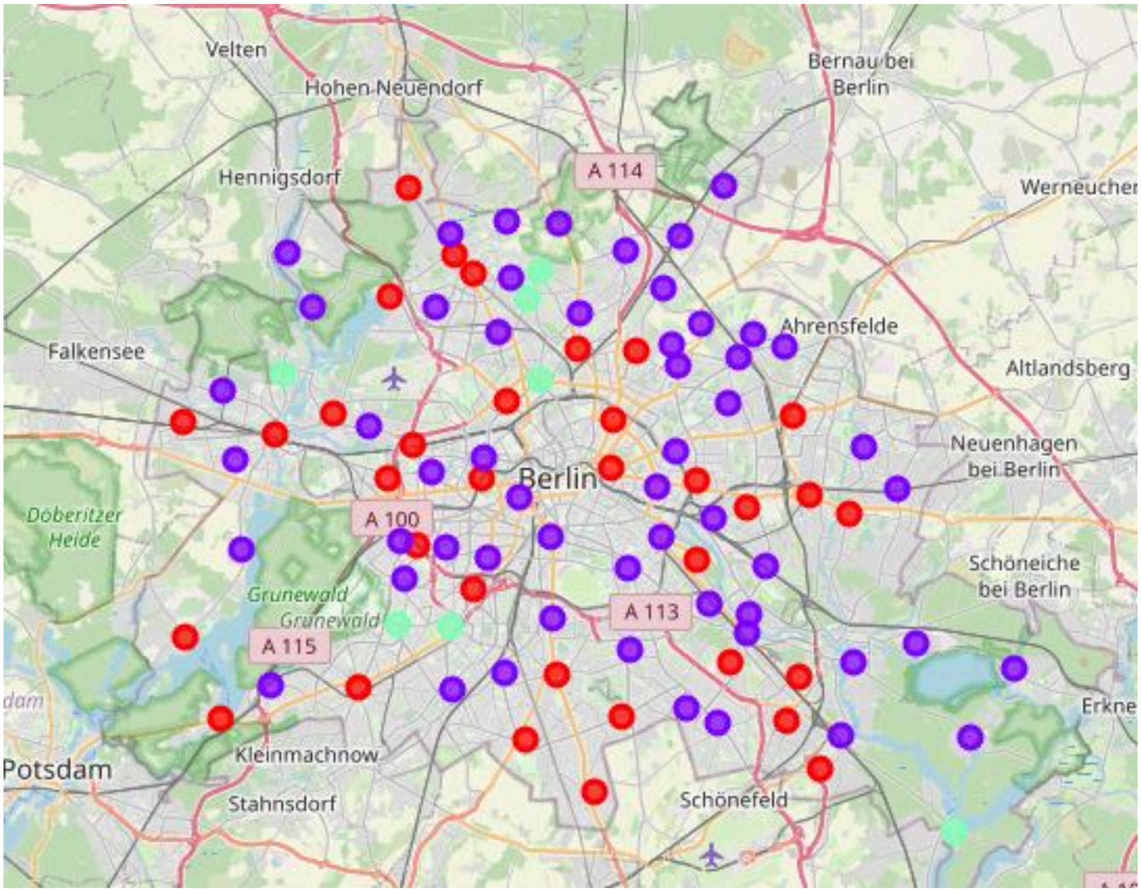
- There are 83 neighborhoods with 0 or 1 Italian Restaurants, they belong to cluster 1
- There are 6 neighborhoods with 2 Italian Restaurants, they belong to cluster 0
- There are 7 neighborhoods with 3 or 5 Italian Restaurants, they belong to cluster 2



DISCUSSIONS

Discussions

First of all, a decision must be made about which clustering selection will be used, the one with 5 clusters, the one 3 clusters, or the one with 3 clusters but modified to include the neighborhoods with one Italian restaurant in the same cluster as the neighborhoods with none. The one with 5 clusters does not prove to be much useful, so it will be discarded. From the remaining 2, whereas it is true that a neighborhood with just one Italian restaurant should be considered as "few to none" competition, the resulting cluster becomes so big that it loses the purpose to segment the neighborhoods, so it will be also discarded and the original selection of 3 clusters will be used.



As It can be seen in the resulting map, the neighborhoods that have "few to none" and "medium" competition are intertwined, and the ones with "high" competition are located to the west side of the city. This would indicate that the best option is opening an Italian restaurant in the east side of the city, while avoiding the neighborhoods that fall in the category of “medium” competition. That would be the best choice according to the analysis made.

Limitations and Suggestions for Future Research

Having finished the analysis, it is important to say that only one factor was considered: number of Italian restaurants per neighborhood. Whereas this is an important factor, there are other ones such as population density and average income of residents that should also be taken into account to decide the location of the new Italian restaurant.

This first analysis could be taken as an starting point and, depending on the results of the analysis of the data of other factors such as the ones mentioned before or other ones, maybe the 3 clustering options could be used, and in that case the original 3 cluster choice would prove not to be the best one, but instead one of the other two, or a fourth one perhaps. Future research should use this analysis, as mentioned, as a starting point to consider other factors for taking the decision of where to open the restaurant.

It should also be noted that a free Sandbox Tier Account of Foursquare API was used, with the limitations that it brings with it, such as the number of API calls and results returned. An improvement could be made if the study was performed using a paid account to get better results without these limitations.

Conclusion

The analysis made in this study has proved that there are certain neighborhoods that are more convenient for opening a new Italian restaurant, taking into consideration only the number of restaurants of this sort into account, and it has made easy to know which ones of them fall into this category. For the relevant stakeholders, this information will be useful when taking this decision. And, as have been mentioned, it should be complemented with further analysis of the data of other factors.