Opening an Italian Restaurant in Berlin

Coursera Capstone Project

“Life is a combination of magic and pasta.”

— Federico Fellini, Director

Table of Contents

Introduction

Discussing the Business Problem & Stakeholders

Business Problem

Objective

Analyze and select the optimum locations in the city of Berlin, Germany to open a new Italian restaurant.

How we did it?

Data Science Methodology

Machine Learning Techniques

Obtained Result's Analysis

Target Audience

Stakeholders

Armando Lingüini and any other person thinking about opening a new Italian restaurant in the city of Berlin.

Data Required

The data that has been used for this study:

- List of neighborhoods in Berlin.
- Latitude and longitude coordinates of those neighborhoods.
- Venue data, particularly data related to Italian restaurants.

Data Sources

The data required was gathered from:

- Wikipedia.
- Geocoder.
- Foursquare API .

02

Methodology

Web Scraping, Data Wrangling & ML

Methodology

1. Get the list of neighborhoods from Berlin. To do this, web scraping techniques are applied using the BeautifulSoup package.

2. Geographical coordinates have to be added to the dataframe using the geocoder package.

3. Foursquare API will be used to get the nearby venues from the neighborhoods and their unique category.

4. Data will be grouped by neighborhoods so the total of each venue category by neighborhood can be obtained.

5. Only take into account the Italian Restaurant Category.

6. Use Silhouete Score to obtain optimum number of clusters.

7. Apply K-Means Clustering to data.

Whoa!

We have analyzed 1,623 venues just from Berlin!
Those sure are a lot of venues...

03

Results

Applying Silhouette Score and K-Means

Silhouette Score

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

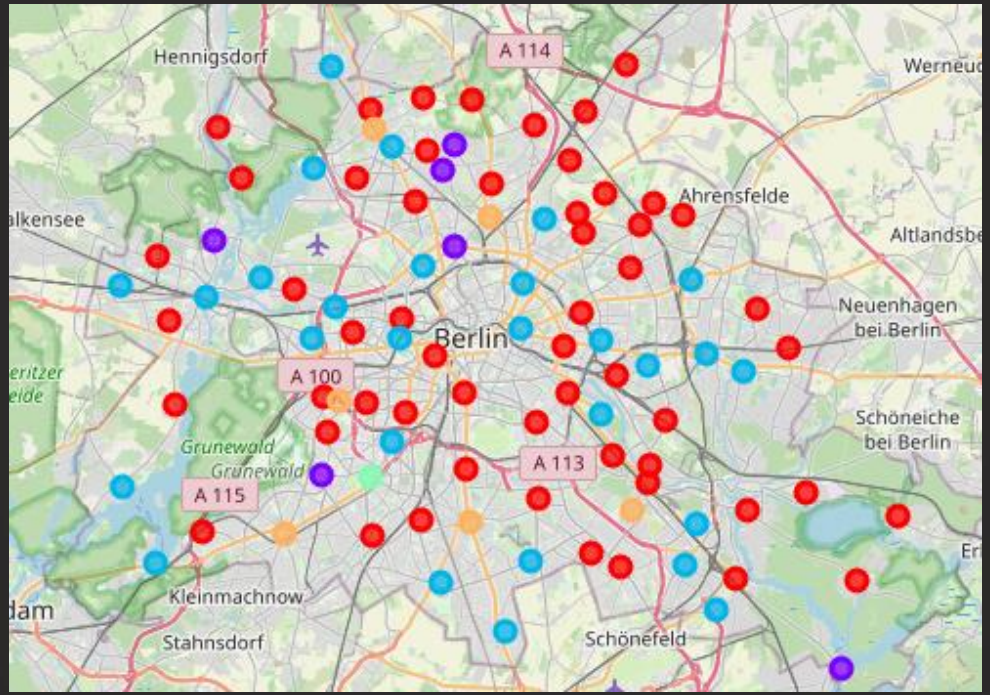
- 2 clusters: 0.7623111393242606.
- 3 clusters: 0.888850456592392.
- 4 clusters: 0.938446969696969697.
- 5 clusters: 0.9895833333333334.

K-Means with 5 Clusters

The results of using 5 clusters, for a total of 96 neighborhoods:

- 55 neighborhoods with no Italian Restaurants (cluster 0)
- 28 neighborhoods with 1 Italian Restaurants (cluster 2)
- 6 neighborhoods with 2 Italian Restaurants (cluster 4)
- 6 neighborhoods with 3 Italian Restaurants (cluster 1)
- 1 neighborhood with 5 Italian Restaurants (cluster 3)

K-Means with 5 Clusters

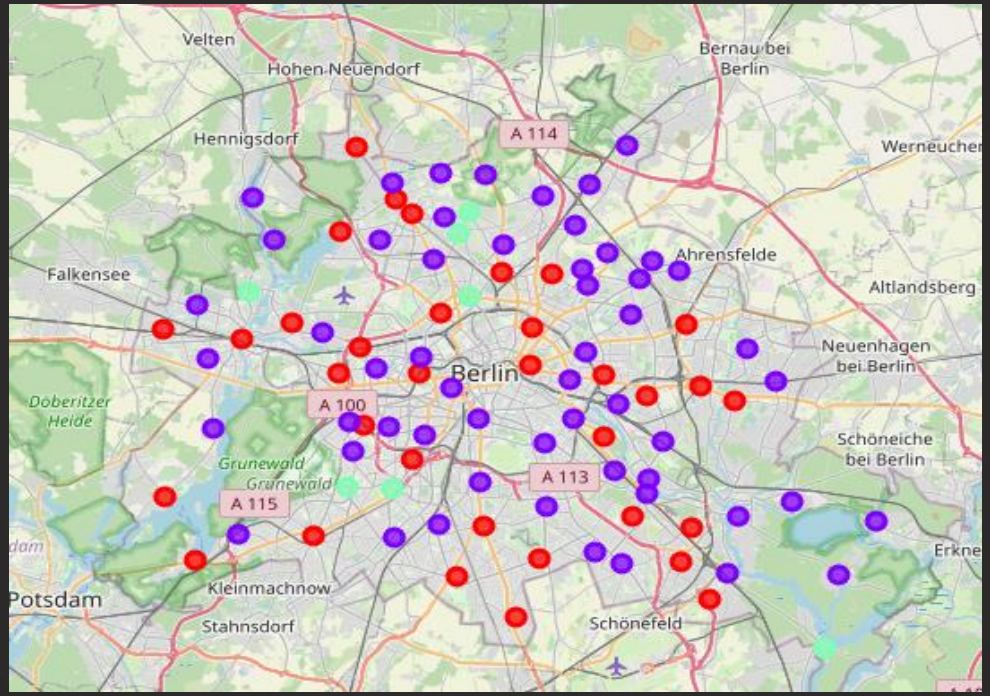


K-Means with 3 Clusters

The results of using 3 clusters, for a total of 96 neighborhoods:

- 55 neighborhoods with no Italian Restaurants (cluster 1)
- 34 neighborhoods with 1 or 2 Italian Restaurants (cluster 0)
- 7 neighborhoods with 3 or 5 Italian Restaurants (cluster 2)

K-Means with 3 Clusters

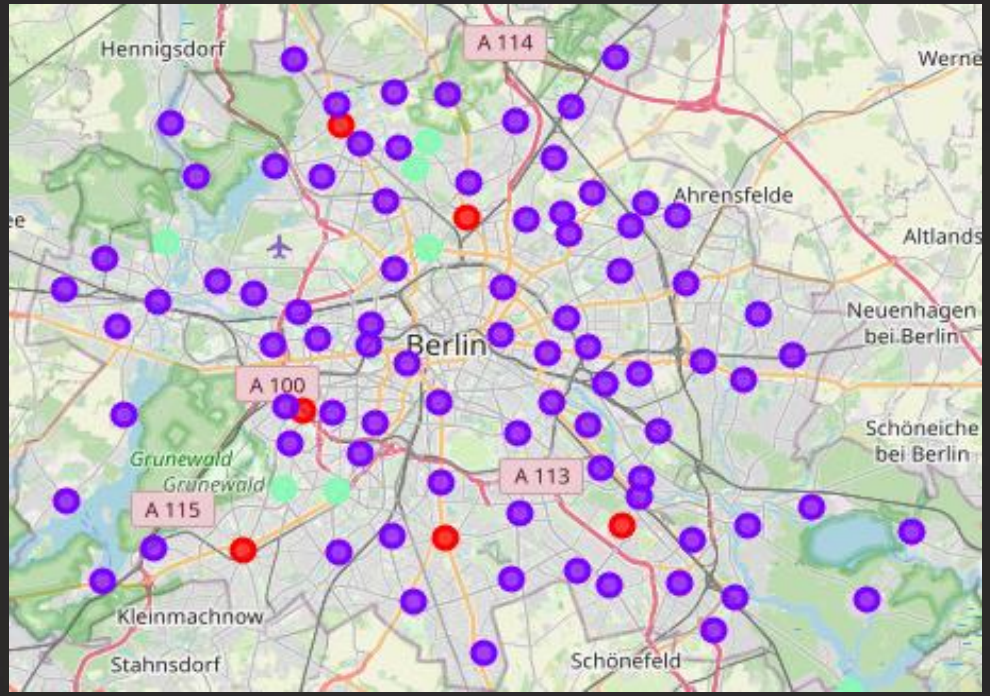


K-Means with 3 Clusters (modified)

The results of using 3 clusters, for a total of 96 neighborhoods:

- 55 neighborhoods with 1 or no Italian Restaurants (cluster 1)
- 6 neighborhoods with 2 Italian Restaurants (cluster 0)
- 7 neighborhoods with 3 or 5 Italian Restaurants (cluster 2)

K-Means with 3 Clusters (modified)



04

Discussions

Picking the best model

Result's Discussions

The following can be acknowledged from the results:

- The model with 5 clusters is not much useful.
- In the 3 clusters (modified) model, the cluster with 0 or 1 occurrences becomes so big that it loses the purpose to segment the neighborhoods.
- With the considerations that have been taken, the original 3 cluster model is the optimum.

05

Conclusion

What we learned and what's next?

Conclusions!

- 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.
- 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.

Limitations and Suggestions for the Future

It is important to keep in mind:

- Only the number of Italian restaurants per neighborhood was considered. In the future other factors should be considered (i.e. population density, average income of residents, etc.)

- When doing analysis considering other factors, the 5 cluster and 3 cluster (modified) models should also be taken into account.

- A free Sandbox Tier account of Foursquare API was used. A paid account should give better results.

“If your mother cooks Italian food, why should you go to a restaurant?”

— Martin Scorsese, Director