Ideal Geographical Locations for New Restaurant Businness in Berlin

Summary

This project aims at serving recommendations for ideal geographical locations for a new restaurant business in Berlin. A restaurant business is a very prospective and profitable business in general, particularly big cities such as Berlin. One of the most important parts of this business is to choose a well-suited place where a restaurant will be established. This project is focusing on recommending a model for profitabe places for restaurant business based on analysing open data especially Foursquare API data and other public internet resources.

Table of Contents

Introduction

This project aims at serving recommendations for ideal geographical locations for a new restaurant business in Berlin. A restaurant business is a very prospective and profitable business in general, particularly big cities such as Berlin. One of the most important parts of this business is to choose a well-suited place where a restaurant will be established. This project is focusing on recommending a model for profitabe places for restaurant business based on analysing open data especially Foursquare API data and other internet resources.

Berlin is the capital and largest city of Germany by both area and population. Its 3,769,495 (2019) inhabitants make it the most populous and crowded city of the EU. The city is one of Germany's 16 federal states. Its economy is based on high-tech firms and the service sector, encompassing a diverse range of creative industries, research facilities, media corporations and convention venues. Berlin serves as a continental hub for air and rail traffic and has a highly complex public transportation network. In addition, Berlin is hosting many universities, museums, movie theaters, and diverse historical and cultural places. From this perspective, a new restaurant business is a real option for this city with this target population group.

Before starting to analysis, we should determine the data which we need for this business. One of the challenges in data analysis is to define the required data which can represent the problem adequately and this should be done with the project partners. As bussiness understanding,let's check the open sources to define the location based requirements for the restaurant business such as https://fitsmallbusiness.com/choose-a-restaurant-location/, and summarize the required demographic and geographic data for this analysis. What we need for this analysis is as follows:

1. Neighborhoods (For clustering and classification purposes)
2. General population and age classification
    1. 15-35 (Fast-food)
    2. 25-45 (Bar-Bistro)
    3. 30-50 (Casual dining)
    4. 35-65 (Fine dining)
    5. 65+
    6. Male/female distribution
    7. Ethnicity (or domestic and foreign population)
3. Locations of parking places
4. Locations of stadiums, theaters, transportation hubs, airports, malls, and universities
5. Cusine type and/or restaurant styles
6. Crime rates

After getting all the required data, our first goal is to classify them according to neighborhoods of Berlin and to make cluster analysis using venue categories. The final target is to be able to define a prediction model to choose or rank the best locations from a given dataset by using machine learning algorithms on the data we obtained from the attributes of previously established restaurants.

Required Data

Let's start with the neighborhoods of Berlin data in the following link: https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin for our analysis.

Table 1. Berlin neighborhoods

	Borough	Neighborhood	Population	Latitude	Longitude
0	Berlin	Charlottenburg-Wilmersdorf	319.628	52.497058	13.296490
1	Berlin	Friedrichshain-Kreuzberg	268.225	52.501500	13.435120
2	Berlin	Lichtenberg	259.881	52.514581	13.498392
3	Berlin	Marzahn-Hellersdorf	248.264	52.539720	13.584280
4	Berlin	Mitte	332.919	52.516740	13.366790
5	Berlin	Neukölln	310.283	52.480200	13.433640
6	Berlin	Pankow	366.441	52.571050	13.404970
7	Berlin	Reinickendorf	240.454	52.567550	13.331650
8	Berlin	Spandau	223.962	52.550090	13.200356
9	Berlin	Steglitz-Zehlendorf	293.989	52.443640	13.229080
10	Berlin	Tempelhof-Schöneberg	335.060	52.447630	13.385350
11	Berlin	Treptow-Köpenick	241.335	52.445817	13.574580

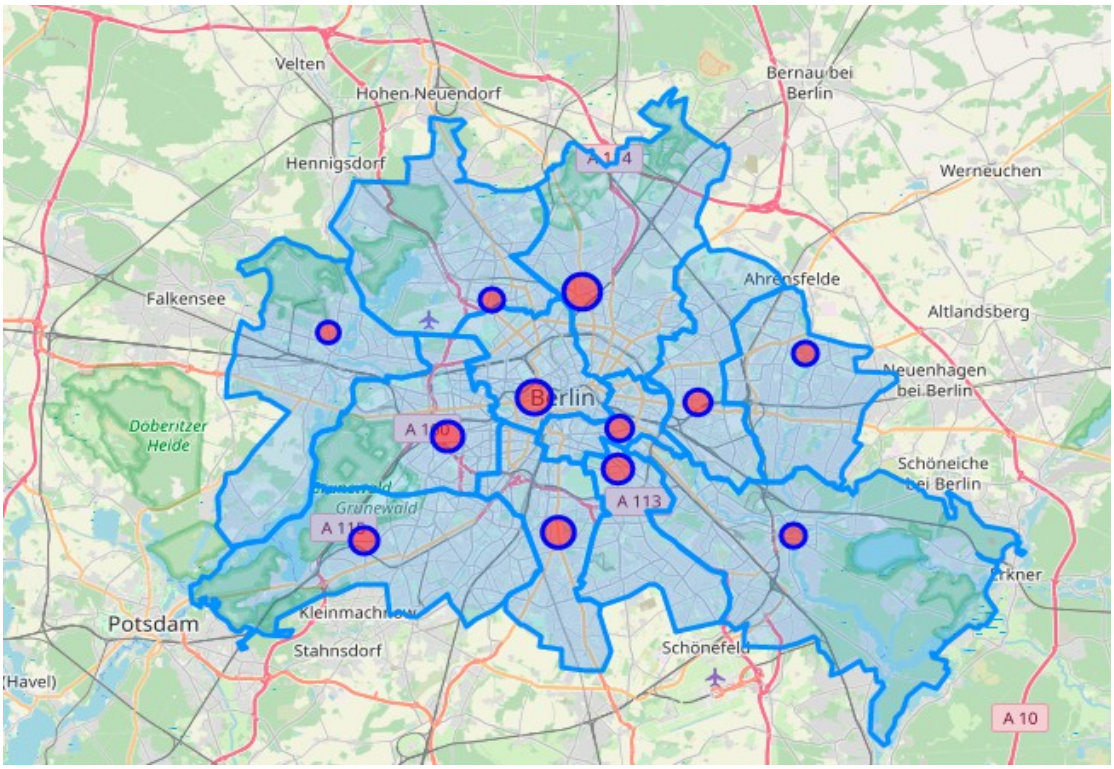


Image 1. Berlin neighborhoods (Sizes of dots show the population of that neighborhood)

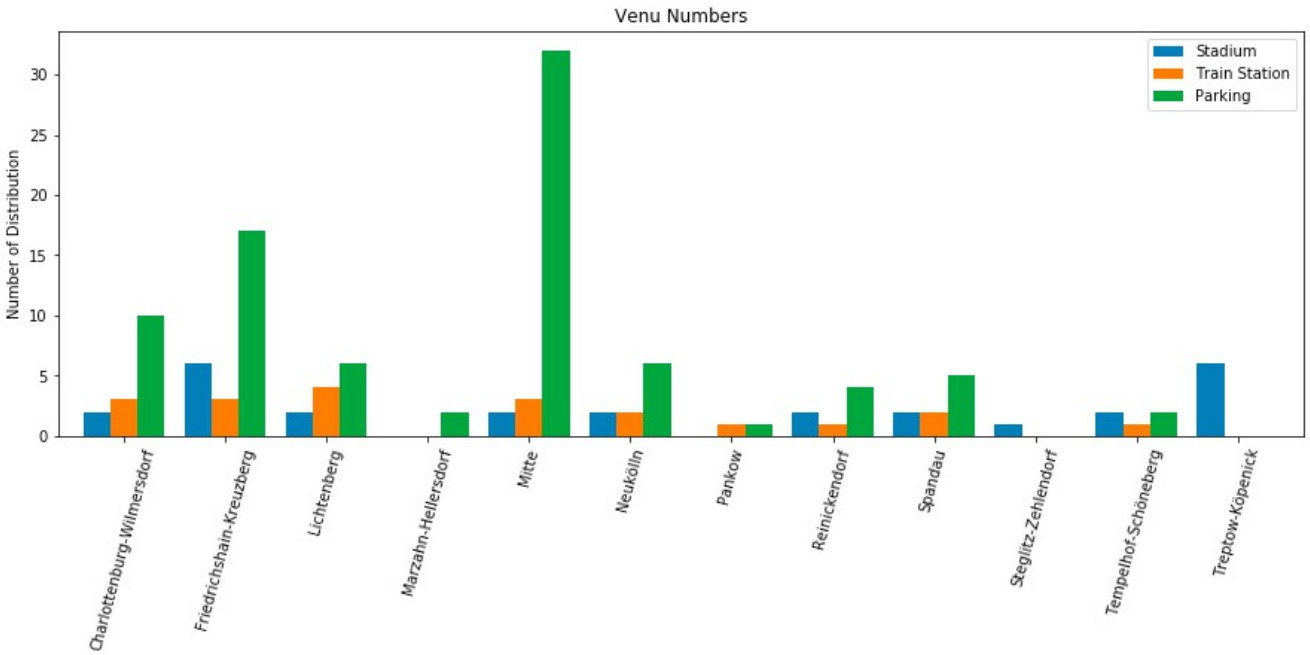
Next with Foursquare data which can be obtained by means of Foursquare API freely. Categories of venues can be found in this link: https://developer.foursquare.com/docs/build-with-foursquare/categories/ . When we examined the category list, we can see the following venues have valuable information for our needs:

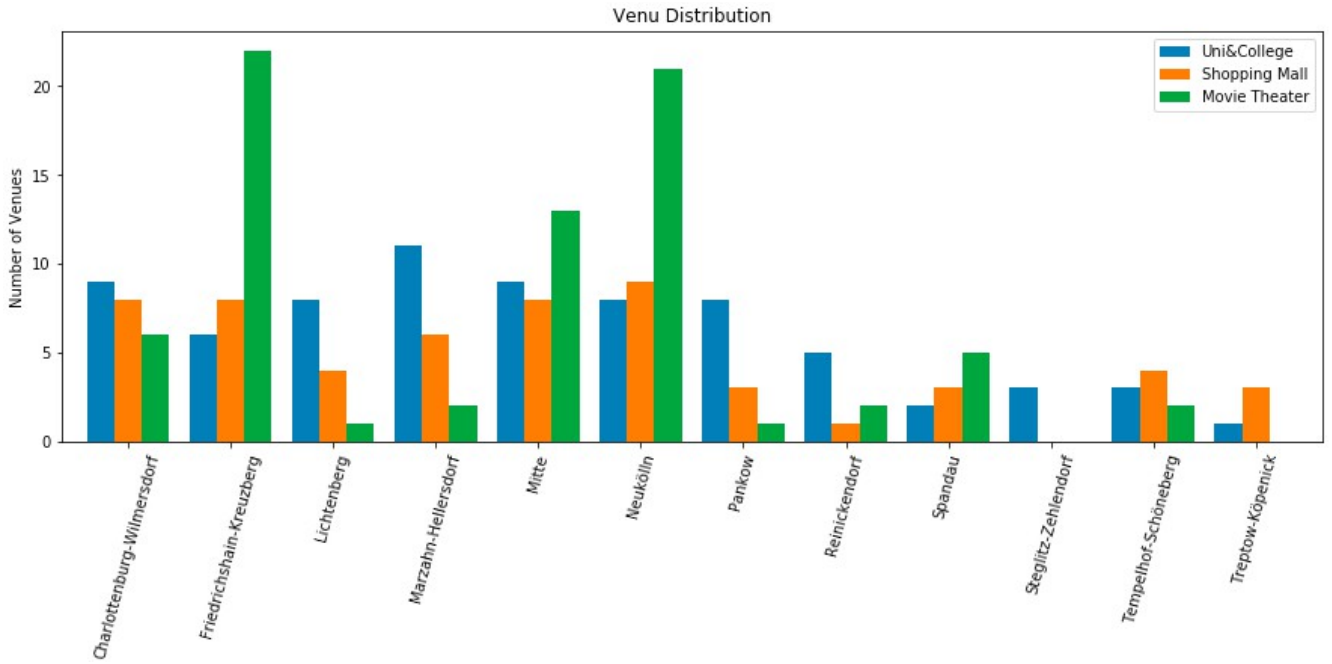
- 'College & University':'4d4b7105d754a06372d81259',
- 'Shopping Mall':'4bf58dd8d48988d1fd941735',
- 'Movie Theater':'4bf58dd8d48988d17f941735',
- 'Stadium':'4bf58dd8d48988d184941735',
- 'Train Station':'4bf58dd8d48988d129951735',
- 'Parking':'4c38df4de52ce0d596b336e1',
- 'Airport':'4bf58dd8d48988d1ed931735'

While using the Foursquare API for this type of analysis 'intend' parameter should be 'browse' or 'global' depending on the purpose.

As a result of the exploratory analysis we obtained Table 2. We can see the newly added columns(ie: the numbers of universities and colleges, shopping malls, movie theaters, stadiums, train stations and parking places).

Figure 1. Number of important gathering places in Berlin

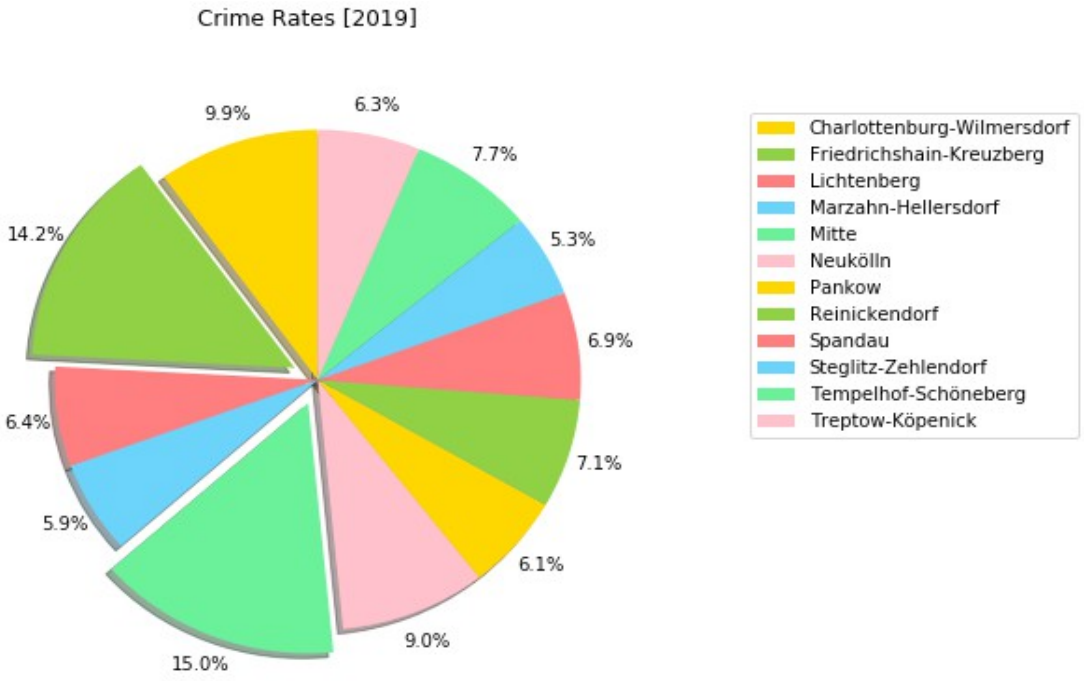




we can get the “Crime statistics of Berlin” from the following link:
https://www.kriminalitaetsatlas.berlin.de/K-Atlas/bezirke/Fallzahlen&HZ%202012-2019.xlsx

Let's examine the crimes data. As we can see data contains the number of events in the neighborhoods of Berlin. On the other hand, the population of neighborhoods are not the same. So, to normalize the data with the population of neighborhoods are much more significant for our case. That is why, we calculate the crime rates by dividing the number of crimes to population of the corresponding neighborhood. Figure 3 shows the crime rates of the neighborhoods.

Figure 3. Crime rates of Neighborhoods of Berlin

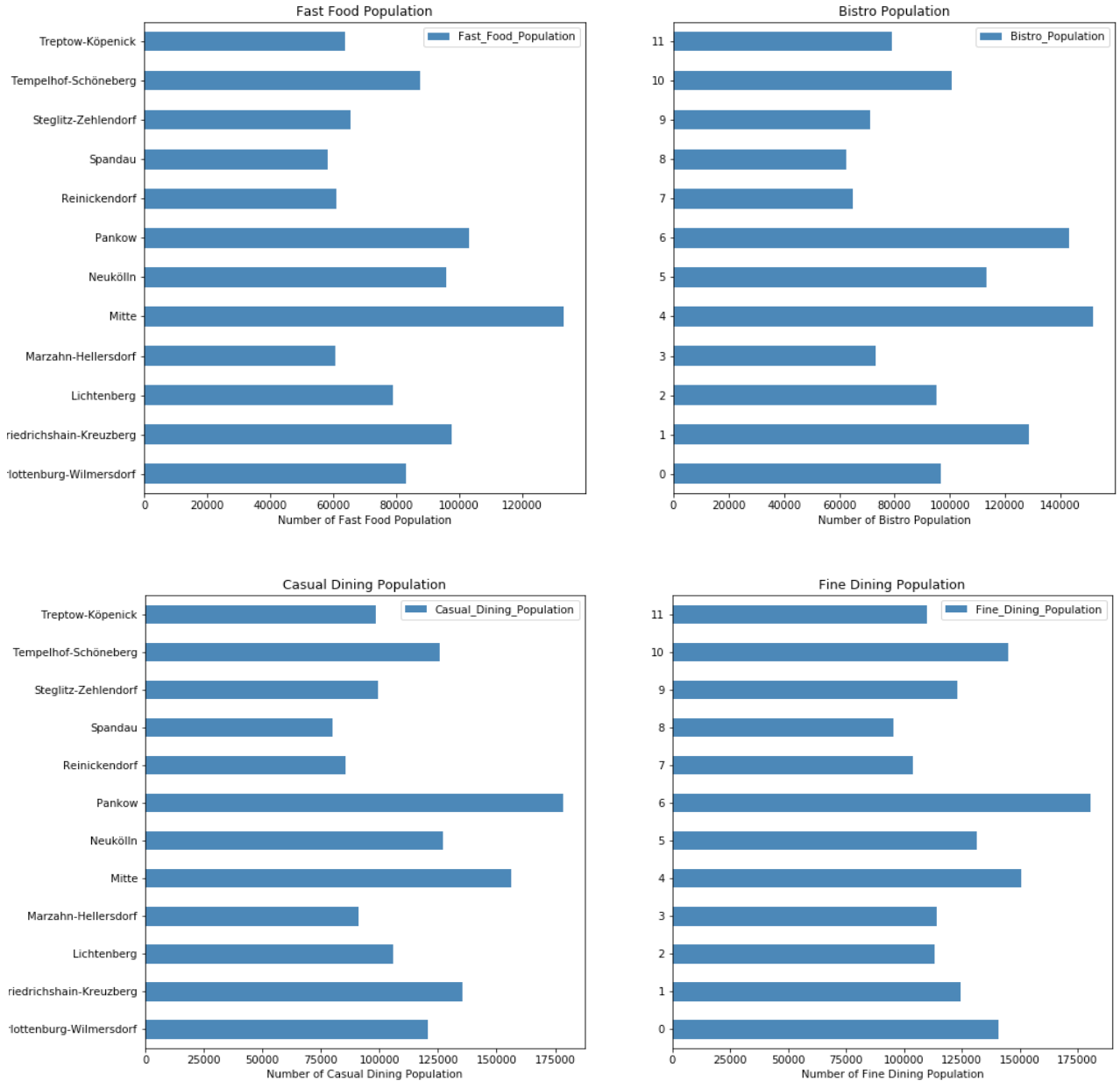


Finally, we can find population-related data which we need from the following link: https://www.statistik-berlin-brandenburg.de/opendata/EWR_Ortsteile_2018-12-31.csv This data named as 'Residents in the districts of Berlin on December 31, 2018' officially and public for research and other purposes.

In addition to these demographic data, we can generate age classification accorging to dining types (In open source there are some age classifications such as between 15 and 35 ages prefer fast-food etc.)

- Ages 15-35 (Fast-food)
- Ages 25-45 (Bar-Bistro)
- Ages 30-50 (Casual dining)
- Ages 35-65 (Fine dining)

Figure.2 Age classification accorging to dining types in the neighborhoods of Berlin



Besides, we can get details of each restaurant venues. This is a restricted part of the data provided by Foursquare API. Restrictions apply based on your account type. On the other hand, we can 500 API calls per day for venue details freely which provides us valuable information such as 'price', 'rating' and 'likes' keys for restaurant categories. These categories are defined in API reference as follows:
price: An object containing the price tier from 1 (least pricey) - 4 (most pricey) and a message describing the price tier.
rating: Numerical rating of the venue (0 through 10). Returned as part of an explore result, excluded in search results. Not all venues will have a rating.
likes: The count of users who have liked this venue, and groups containing any friends and others who have liked it. The groups included are subject to change.

The complete list of venue details can be found in the following link:
https://developer.foursquare.com/docs/api-reference/venues/details

After downloading all venue details and merging into previous obtained dataframe, we finally reached a dataframe which contains 31 columns valuable data for further analysis by machine learning

Data columns (total 31 columns):

Neighborhood	546 non-null object
Neighborhood Latitude	546 non-null float64
Neighborhood Longitude	546 non-null float64
Venue Id	546 non-null object
Venue	546 non-null object
Venue Latitude	546 non-null float64
Venue Longitude	546 non-null float64
Venue Distance	546 non-null int64
Venue Category	546 non-null object
Price	546 non-null int64
Rating	546 non-null float64
Likes	546 non-null int64
Borough	546 non-null object
Population	546 non-null float64
Latitude	546 non-null float64
Longitude	546 non-null float64
Uni&College	546 non-null int64
Shopping Mall	546 non-null float64
Movie Theater	546 non-null float64
Stadium	546 non-null float64
Train Station	546 non-null float64
Parking	546 non-null float64
Crime_numbers	546 non-null int64
Male_Population	546 non-null int64
Female_Population	546 non-null int64
Foreign_Population	546 non-null int64
German_Population	546 non-null int64
FasFood_Population	546 non-null int64
Bistro_Population	546 non-null int64
Casual_Dining_Population	546 non-null int64
Fine_Dining_Population	546 non-null int64