Relocating to Berlin

Maaz Bushra Ali

10-3-2020

Introduction

Berlin is the capital and largest city of Germany by both area and population inhabitants make it the most populous city proper of the European Union. The city is one of Germany's 16 federal states. It is surrounded by the state of Brandenburg, and contiguous with Potsdam, Brandenburg's capital. It is a world city of culture, politics, media and science. Its economy is based on high-tech firms and the service sector, encompassing a diverse range of creative industries. Moving to berlin is one of my childhood dreams.

Business Problem

The objective of this capstone project is to analyse and select the best neighbourhood in the city of Berlin, Germany to relocate. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question:

If you are going to move to Berlin, where is the best place to rent an apartment?

Target Audience of this project

This project is particularly useful to students and immigrants to help them in choosing the right place neighbourhood that meet their different needs. It's also beneficial for travellers and tourists as it gives them general idea about the places in berlin and the cost of living.

Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Berlin and the average rent price of each neighbourhood.
- Latitude and longitude coordinates of those neighbourhoods.
- Venue categories data.

Sources of data and methods to extract them:

A large data 96 columns for Berlin's AirBnB listings-summery are available for free in Kaggle, from this data set we can extract apartments rent price, coordinates and neighbourhood.

This will be the base for rent cost analysis. After that, using Foursquare API we can obtain the venue categories data for clustering.

Links:

Kaggle Data: https://www.kaggle.com/brittabettendorf/berlin-airbnb-data#listings_summary.csv
Foursquare: https://foursquare.com/developers/apps

Data Pre-processing:

Reading AirBnB data, it's obvious that it contains much more fields than needed for this analysis, this for we extract only important columns and drop the rest of it. Moreover, as the data is real life data, it's noticed that some cells are missing, thus it's important to clean or recover the missing cells to avoid misleading in advanced stages of the analysis process

- Data shape before pre-processing: (22552, 96)
- Data shape after pre-processing: (20791, 6)

Methodology

After Installing and pre-processing the AirBnB data frame, two tables are derived from it to satisfy requirements for the two parts of the analysis.

- Part 1: Analysing for average rent cost per day for every neighbourhood:

For this part the data frame is grouped by neighbourhood names and the average of the prices are calculated for each neighbourhood. After that the distribution of the prices is visualized, Information as most expensive and least expensive neighbourhood and average price are elaborated to give a better understanding of the rent cost in Berlin.

- Part 2: Analysing venues of each neighbourhood and clustering:

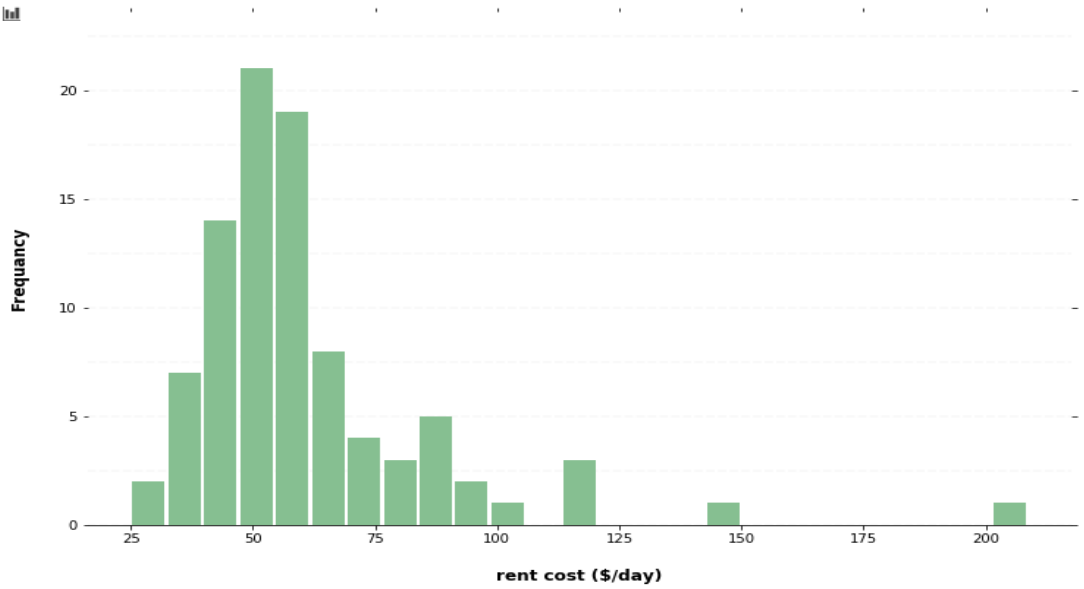
Mid-point for every neighbourhood is calculated and Foursquare API is used to retrieve venues within 500 m for every neighbourhood. These venues are then grouped by category to facilitate segmentation process of neighbourhoods.

For clustering, K-Mean method is used with K-value = 4.

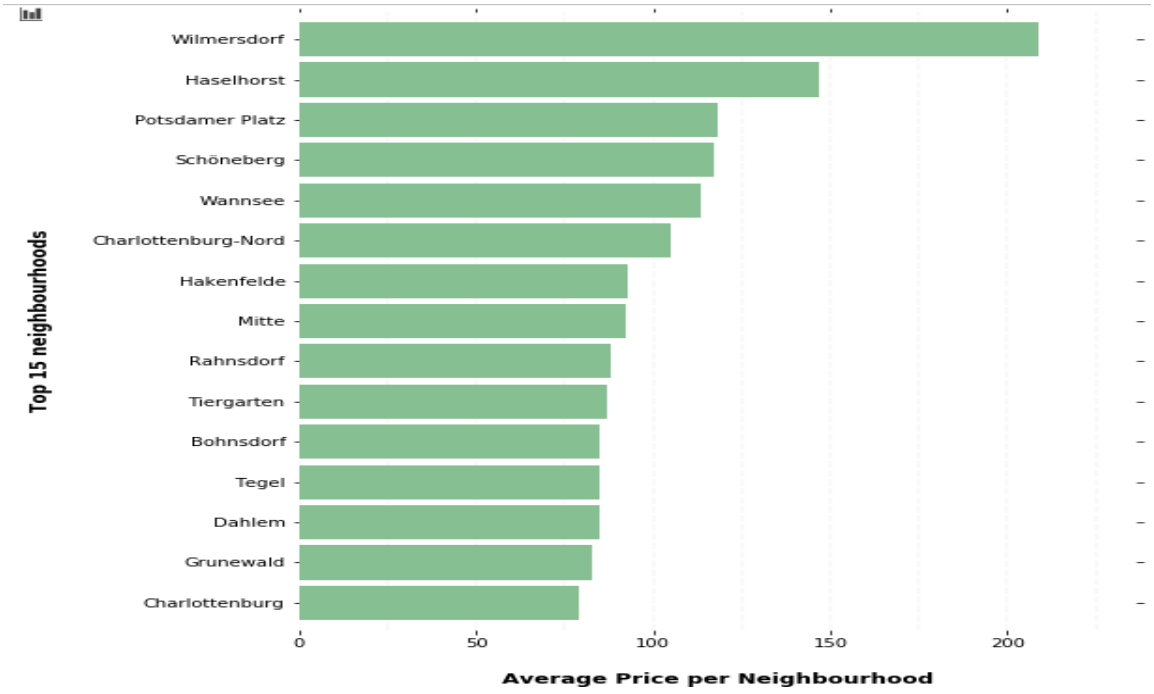
Results & Discussion

Rent cost analysis:

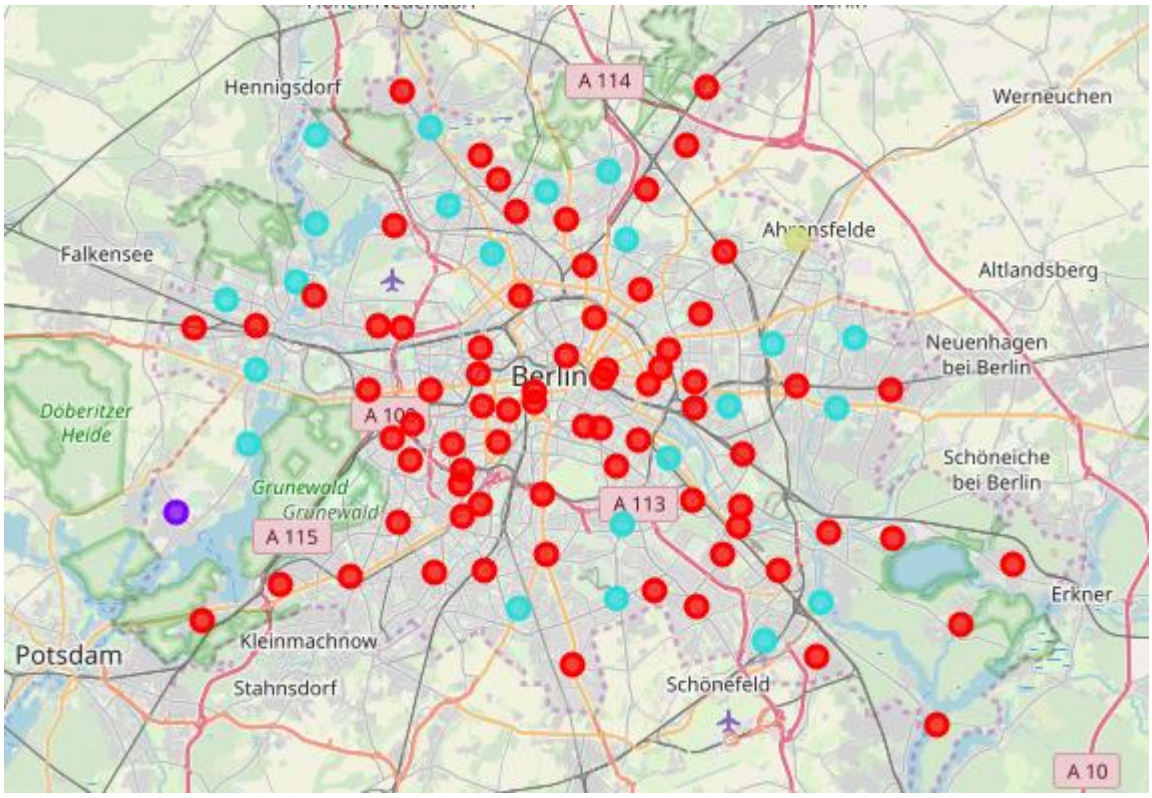
- cost ($/day) distribution is right skewed, with:
  - mean: 60.9
  - std : 26.0
  - min : 25.0
  - max : 208.6



- Lübars is the least expensive neighbourhood and Wilmersdorf is the most expensive one.



Clustering analysis:



Cluster	Name	neighbourhoods	venues
1	Hangout	65	Bars, Café, night clubs, all types of restaurants.
2	Family	1	farm, zoo exhibit, home service, electronics store
3	Shopping	22	supermarket, plaza, shopping stores, home store, circus
4	foreign	1	gas station, farmers market, electronic store, foreign restaurants

Berlin city is showing high quality of living as most of neighbourhoods are categorised between Hangout and shopping clusters.

- The 1st cluster (Hangout cluster) is the major cluster as it contains 73% of neighbourhoods.
- For me as I love hanging out, the 1st cluster is the best choice.
- Luckily the least expensive neighbourhoods happen to be also in the 1st cluster. Thus, Lübars is the choice.