Coursera Capstone

IBM Applied Data Science Capstone

Title: Location for opening a new shopping mall in Berlin

June 2021

1. Introduction

Shopping malls have become more and more popular for people to visit, where people can not only do shopping but also enjoy themselves such as dining, watching movies and so on. This is why shopping malls spread all around the world especially in big cities, from the city center to suburbs. However, the problem of selecting the location for opening a new shopping mall is always difficult. On the one hand, the new shopping mall should be located near the city center so that people can visit it conveniently. On the other hand, it also should be located to keep a distance from the city center to avoid hard competition. Therefore, how to select a suitable location for opening a new shopping mall to balance the above contradiction always confuse the investors.

This project aims at solving the problem of choosing locations for opening a new shopping mall in Berlin, Germany. The problem will be solved by using the data scraped from the internet, and analyzed by using machine learning. Then, some suggestions could be provided to investors to make a better choice about the locations of a new shopping mall in the city of Berlin. Moreover, the method can be can also be applied to other cities for the same kind of problems.

In order to solve the above problem, we firstly need to collect the related data, then suitable methods should be applied to analyze the obtained data. In the end, the conclusions should be made to help. All these steps will be provided in the next sections.

2. Data

In order the solve the proposed problem, two kinds of data are needed including the neighborhoods of Berlin and the corresponding venue data of these neighborhoods.

The data of neighborhoods of Berlin can be scraped form the Wikipedia(https://en.wikipedia.org/wiki/Category:Localities_of_Berlin). For this, we will use the method of requests in Python and the beautifulsoup packages to obtain the data by scraping the above website. A list of neighborhoods can be obtained. Then, the Python Geocoder package will be applied to get the geographical coordinates of the neighborhoods so that a further step could be proceeded.

With the geographical coordinates, the Foursquare API will be used to further get the venue data for these neighborhoods. Foursquare has been widely used by over 125,000 developers with a large dataset of over 150 million places. Foursquare API provide a variety of categories of venue data, among which we are interested in the shopping mall category so that we can do our analysis. Therefore, we will use the Foursquare API to get the venue data of shopping mall in Berlin.

The related data can be collected as introduced above. Now the methods to analyze the corresponding data to help investors make decisions should be selected, that will be introduced in the following section.

3. Methodology

We will use Python as the programming language for this project, which is very good for solving data related problem with many popular packages such as Numpy, Pandas and so on.

For this project, we firstly need to obtain the list of neighborhoods of Berlin. For this purpose, we will scrap the website of Wikipedia (https://en.wikipedia.org/wiki/Category:Localities_of_Berlin), that provides the related information. To this end, we will use the method of requests in Python and the beautifulsoup packages to extract the corresponding information and generate a list for the obtained neighborhoods. Moreover, we also need the geographical coordinates of the above information for the further analysis. Fortunately, this can be easily realized by using the Geocoder package that can get the geographical coordinates in the form of latitude and longitude corresponding the above neighborhoods. In order to show the obtained data, we will put the data into a Dataframe of Pandas. Then, the data can be visualized with a map by using a map based on the Folium package.

With the above data, we will apply the Foursquare API to get the top 100 venues within a radius of 2k meters. For this, we must firstly register a Foursquare Developer Accountant to get the corresponding ID and Key. Then, the Foursquare API calls can be made with the geographical coordinates of the above neighborhoods. After that, a JSON file will be returned with the information of the venue name, venue category, the venue latitude and longitude. By analyzing the

above data, we can get the frequency of the occurrence of each venue category, among which we are interested in the category of shopping mall. Therefore, we will extract the frequency of occurrence of the category of shopping mall for the further analysis.

For the obtained frequency of occurrence of the shopping mall, we will use k-means to perform clustering. K-means is a popular unsupervised machine learning algorithm. In this project, we will use k-means to cluster the neighborhoods into 3 clusters based on the occurrence frequency for shopping mall. The results will clearly show different density of shopping malls in different neighborhoods. Therefore, this will help investors to make decisions about the location for opening a new shopping mall.

4. Results

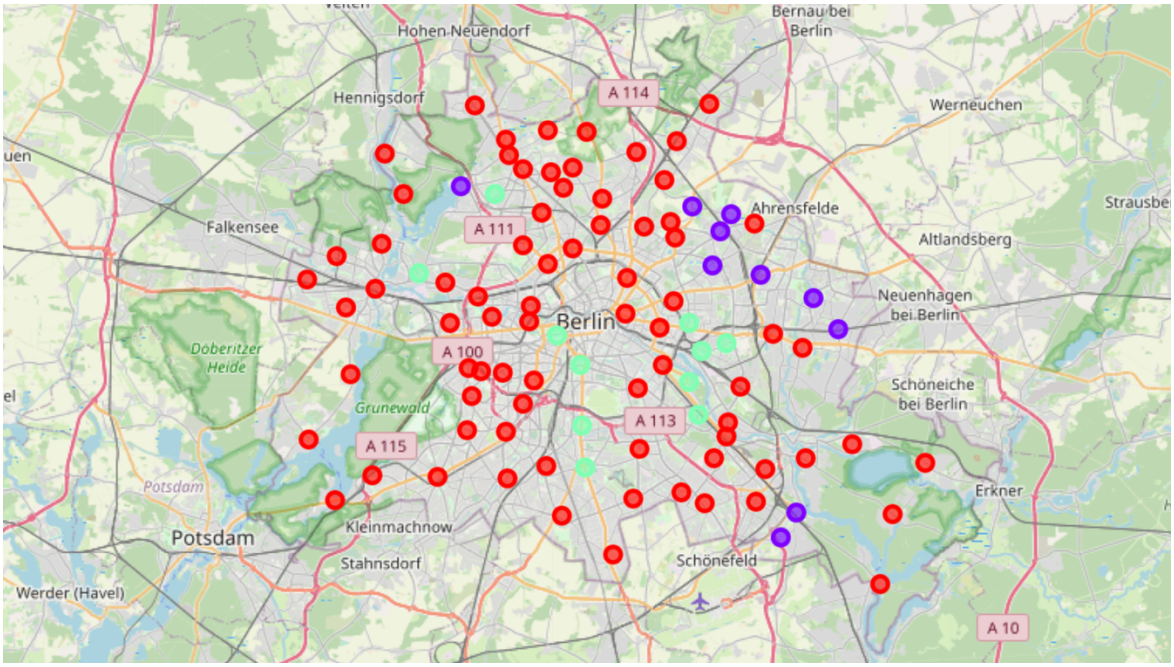
There are 3 clusters based on the occurrence frequency of shopping mall as follows:

Cluster 0: the neighborhoods with no shopping malls.

Cluster 1: the neighborhoods with high number of shopping malls.

Cluster 2: the neighborhoods with medium number of shopping malls.

Moreover, we get the following map for the visualization the above results.



where cluster 0 is in red color, cluster 1 is in green and cluster 2 is in purple.

5. Discussion

The above map gives us a good visual about the distribution of the existing shopping mall in Berlin. We know that most of shopping malls are located in the city center. Some ones are located outside the city center which are in the area of cluster 2. And there are still no shopping malls in the area of cluster 0.

6. Conclusions

With the results obtained in last section, we know that most of shopping malls in Berlin are located in the city center, which is the cluster 1. That means there will be a hard competition for the shopping malls in the areas of cluster 1, which is not a good choice for opening new shopping mall. On the contrary, the suburb areas are still lack of shopping malls. However, the areas of cluster 0 may suffer the problem of lack of visitors so that it is neither a good choice for opening a new shopping. In conclusion, I think it may be a good choice for a new shopping mall in the area of cluster 2 which is a balance of the avoiding hard competition and the lack of visitors.

7. Future work

This project is just a simple try to solve the proposed problem. Actually, there are still more factors that affect this problem such as the population in different areas, the cost and so on. Thus, the problem can be further explored by considering more factors to build a more accurate model to help make better choices.