Coursera Capstone

IBM Data Science Capstone

Opening a new shopping center in Germany

By: Henrik Verolet

August 2020

Introduction

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the country of Germany to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Germany, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in Germany.

Data

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

- List of largest Cities in Germany. This defines the scope of this project which is confined to Germany.

- Latitude and longitude coordinates of those Cities. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the Cities.

Sources of data and methods to extract them

This Wikipedia page of Germany contains a list of Cities, with a total of 79 cities. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the citys using Python Geocoder package which will give us the latitude and longitude coordinates of the citys.

After that, we will use Foursquare API to get the venue data for those citys. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.
Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of cities in Germany. Fortunately, the list is

available in the Wikipedia page
(https://en.wikipedia.org/wiki/List_of_cities_in_Germany_by_population).

We will do web scraping using Python requests and beautifulsoup packages to extract the list of

Cities data. However, this is just a list of names. We need to get the geographical

coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so,

we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the cities in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in Germany.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the citiess in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each City and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each city by grouping the rows by city and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the citys.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the citiess into 3 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which citys have higher concentration of shopping malls while which citys have fewer number of shopping malls. Based on the occurrence of shopping malls in different citys, it will help us to answer the question as to which citys are most suitable to open new shopping malls.

Results

The results from the k-means clustering show that we can categorize the citys into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

- Cluster 0: Citys with moderate number of shopping malls
- Cluster 1: Citys with low number to no existence of shopping malls
- Cluster 2: Citys with high concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.

Conclusion

We can observe that there are several major cities without any Shopping Malls. THese cities are summarized in Cluster 1. Those cities bring the most opportunity for the succession of a new shopping Mall.