Coursera’s Capstone Project

IBM Applied Data Science Capstone

Opening a Hotel in Krakow, Poland

By: Tomasz Cyparski

April 2020

Introduction

Krakow was the official capital of Poland until the end of the XVI century. It has been one of the leading centers of Polish academic, economic, cultural and artistic life. It's one of Europe's most beautiful cities, its Old Town was declared the first UNESCO World Heritage Site in the world.

More than 14 million tourists have visited Krakow in 2019 according to the Malopolska Tourist Organization. From that over 3 million were foreign visitors mostly from Germany, Great Britain, Italy, France and Spain.

Based on Statista website (www.statista.com) there were approximately 14 five-star hotels in the city in 2018, an increase of 27 percent compared to the previous year.

That determine the demand for the new, high quality hotels in Krakow as the number of foreign visitors increases every year.

Business Problem

The objective of this project is to analyze and select the best location in the city of Krakow, Poland to open a new hotel. By using data science and machine learning techniques including clustering and geolocation data and using foursquare API this project will try to answer the the business question whether opening a new hotel in Krakow is still a profitable investment.

All Krakow neighborhoods will be taken into the consideration (and of course the close proximity to the center of the city would be a priority). Knowing that center part of Krakow (also known as the Old Town) is a very attractive business spot and in general very hard to find a space needed for a hotel.

Target audience

The results of this project may be the most useful for the investors, real estate agencies that are looking to do business in Krakow. With the increasing number of visitors coming to the city especially during the summer, opening a new hotel may be a good investment.

It's also worth mentioning that the number of well-known sport and cultural events that take place in Krakow is constantly growing and thus the need for a quality accommodation for guests and VIPs.

Data

In order to solve the mentioned above problem the following data will be collected:

- List of all Krakow neighborhoods
- Latitude and longitude coordinates of those neighborhoods and the city itself. They will be used to plot the map with the existing venues
- Hotels data in order to determine their locations and proximity from the city center and from each other

Data sources and methods used to extract them

The list of Krakow neighborhoods and their corresponding geographical coordinates will be extracted from Polish version of Wikipedia (https://pl.wikipedia.org/wiki/Podzia%C5%82_administracyjny_Krakowa). Web scraping technique will be used to extract the data from the website and import to Python. Similarly, all 18 Krakow neighborhoods geographic locations will be imported and processed to allow for the map creation.

Next, using Foursquare API the list of the Krakow venues will be imported and sorted and in the end those representing hotels will be used to create

clusters and based on the result the final conclusion will be made if the opening of a new hotel (and if yes, where) in Krakow is still profitable.

Methodology

The needed list of the neighborhoods in Krakow can be obtained from Polish version of Wikipedia. The list is then extracted and imported into Python (Panda's data frames). Columns irrelevant to this project like: Area, Population and Population Density are going to be removed.

In similar way geographical coordinations of all 18 neighborhoods were being imported and added to existing data frame with the help of Geocoder package allowing to provide both the latitudes and longitudes.

To visualize its locations on the map Folium package was used. All 18 neighborhoods were represented with blue dots on the map. That step is also a confirmation of the correctly obtained geographical data.

Using Foursquare API will allow us to import the list of the venues for each neighborhood. The limit of the venues was set to 100 within a radius of 2000 meters. Foursquare results are being provided in JSON format which need to be translated into Panda's data frame. In our case 142 unique categories were discovered.

Obtained results were then added into the table containing the list of all neighborhoods and only the ones related to hotel category were picked. The total number received was 12.

In the final step the k-mean clustering algorithm was used with chosen number of clusters set to 3, which will provide data of hotels concentration frequency in the corresponding areas. Based on the results the answer to the main question whether opening a new hotel in Krakow could be a good investment.

Results and Discussion

As mentioned earlier the results from the clusters will provide the answer to the question set for this project. The first cluster containing the lowest concentration of hotels (marked in red color) is located on the outskirts of the city, mostly in the east side. The second cluster contained the moderate number of hotel concentration (marked in purple) located in the center as well around the city center. The last, third cluster with the highest concentration of hotels (marked in mint green color) was located in the city center and in close proximity to the center but mostly in the northern-west part of the city. The reason for that is the proximity to the airport, hence the high number of hotels only in that part of the town.

From the results we can conclude that there's a potential for another hotel in Krakow in the relative vicinity of the city center but rather to the east part of the city. This would result in the smaller number of visitors booking the stay in that hotel coming from the airport.

The final decision would have to made based on other factors that were not part of this exercise. Also, other choices of accommodation were not considered here like: hostels, motels, private room rentals etc.