Coursera: IBM Applied Data Science Capstone

Opening a New Hotel in Mexico City's Central Borough Cuauhtémoc

By: Cynthia Valadez

February 2021

Introduction

Tourism in Mexico City has been booming in recent years. Foreign tourists have flocked to this bustling city in pursuit of more favorable weather, affordable experiences and cultural immersion. With this growth in visitors, there is also a growing need for more accommodation. After all, a pleasant hotel experience is a major component of any successful vacation. In particular, the borough of Cuauhtémoc, which is centrally located and is the home to top tourist attractions such as the Zocalo, Palacio de Bellas Artes, Palacio Nacional, and Torre Latinoamericana could surely benefit from a new hotel. Yet, Cuauhtémoc covers a large area, and the location of a new hotel would have to be strategic as to not take away from bookings at other hotels. Nevertheless, a new hotel in Cuauhtémoc would increase accessibility to popular attractions for hotel guests thus leading to more positive reviews on social media/word of mouth and therefore fueling more tourism to the area. This would be beneficial in terms of revenue and publicity. With that said, the business problem to be solved is where is the ideal location to develop a new hotel in Cuauhtémoc? Utilizing data science methodology and machine learning techniques, the ideal location will be determined so that property developers as well as the tourism industry in Mexico can best take advantage of hotel development opportunities in the city.

Data

In order to determine the most strategic location of a new hotel in Cuauhtémoc various sources of data will be used. First, a complete list of "colonias" or neighborhoods of the Cuauhtémoc borough of Mexico City will be compiled. A list of these neighborhoods will

be scraped from a table of the Cuauhtémoc borough located within a Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Mexico_City#Cuahtemoc) using Python's Beautiful Soup library to extract the necessary data. Next, the geographic coordinates for each of the neighborhoods in the list will be obtained. The latitude and longitude coordinates of these neighborhoods are necessary in order to plot the map and access relevant venue data. To obtain these geographic coordinates, Python's geocoder package will be used. Ultimately, venue data indicating how many hotels there are in each neighborhood will be collected. With this data, clustering will be performed. Foursquare API will be utilized to obtain venue data for each neighborhood and from there the relevant hotel data will be selected. Once all the data is gathered, we will proceed to use various data science/machine learning techniques to determine the ideal location to build a new hotel in Cuauhtémoc based on the existing preponderance of hotels in each neighborhood and their proximity to popular tourist attractions.

Methodology

The first data science technique that will be employed to help solve this business problem in web scraping with Python. Python's Beautiful Soup library will be used to perform the scraping of a Wikipedia page to extract a list of Cuauhtémoc colonias from one of the many tables within the webpage. Needless to say, once this list is obtained there is not much analysis or deductive reasoning that can happen until we obtain further quantitative data. Therefore, the next step is to obtain the latitude and longitude coordinates of each colonia to use with Foursquare API. Python's Geocoder library will help us obtain this data by taking

in the name of each colonia as well as the city and country to then return a comprehensive list of the respective geographic coordinates. Subsequently, these coordinates and their corresponding colonia names will be converted into a mutable Pandas DataFrame. With this DataFrame, we will use the Folium package to populate a map to better visualize the location of each of these neighborhoods across the Cuauhtémoc borough. Judging by the positioning of locations in the map, we can determine whether the coordinates returned by the Geocoder library make sense and if they are properly plotted within the chosen borough.

Furthermore, Foursquare API will be used to obtain the top 100 venues of each colonia within a 3000-meter radius. With a Foursquare developer account, we obtain the necessary ID and Secret Key to make API calls. Using a Python “for” loop we make API calls to Foursquare passing in the colonias and their coordinates. Foursquare will return the venue data for each colonia in JSON format. From there we will extract the relevant information for each nearby venue including the venue name, venue category, and latitude/longitude coordinates. With this data, we can determine how many venues were returned and how many unique venue categories result from the compiled venue data. Next, we will group each neighborhood and determine the mean of the frequency of existence of each distinct venue category. This serves as preparation to perform data clustering. The venue categories will be filtered for “Hotel” in order to only analyze relevant data to our business problem.

Ultimately, through k-means clustering we will cluster/group the data. This type of clustering method is performed by partitioning the observations(n) into k clusters in which each observation corresponds to the cluster with the nearest mean. Essentially, it is a

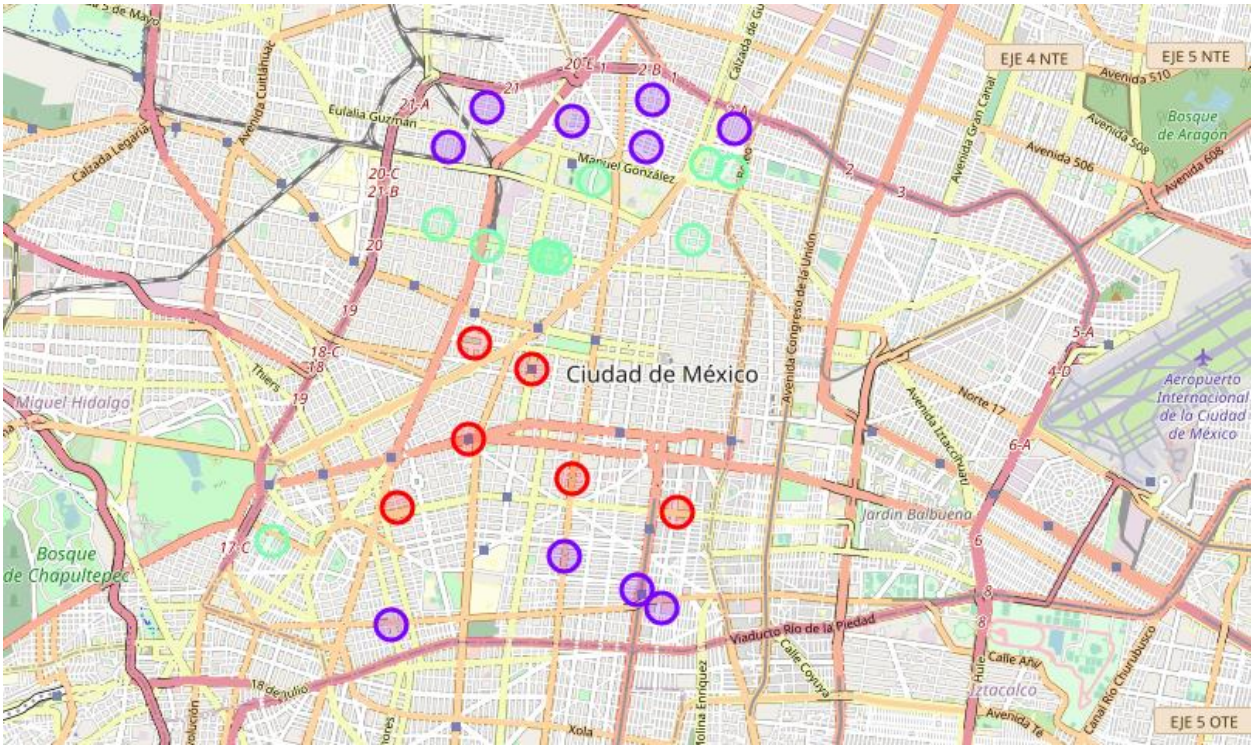
distance-based algorithm where we calculate the distances to assign any given point to a cluster. This simple unsupervised machine learning algorithm will be used to help solve our problem. Three clusters will be generated based on the recurring frequency of “Hotel” in each colonia. The results of this clustering will allow us to determine which colonias have a higher preponderance of hotels and which are on the lower end of the spectrum. By analyzing these results, we can strategically decide on the ideal location for a new hotel in Cuauhtémoc.

Results

Through k-means clustering, we can categorize the colonias into three distinct clusters based on the preponderance of hotels in the corresponding area. The resulting clusters are as follow:

Cluster	Hotel Preponderance	Color on Map
Cluster 0	High	**Red**
Cluster 1	Low	**Purple**
Cluster 2	Moderate	**Green**

** These clusters can be visualized in the map below.*



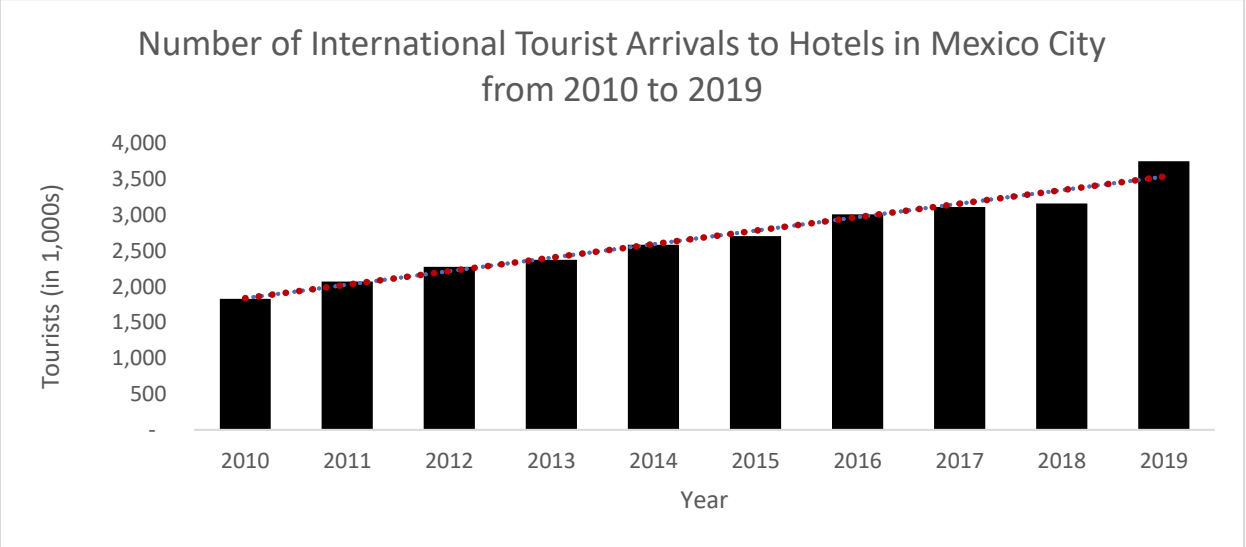
Discussion

As can be noted from the map and chart in the Results section, most of the hotels are located in the central area or "Centro Histórico" of Mexico City's borough Cuauhtémoc. Cluster 0 has the highest frequency of hotels in each colonia followed by a more moderate frequency of hotels in colonias in Cluster 2. This leaves us with a lower preponderance or even nonexistence of hotels in colonias in Cluster 1 which encompasses neighborhoods located more on the outskirts of the Cuauhtémoc borough. These colonias in Cluster 1 present hotel developers an ideal opportunity to build a new hotel in areas that are not currently saddled with significant hotel competition. On the other hand, it is not advised to build new hotels where hotel competition is higher such as in colonias in Cluster 0 or even in Cluster 2. Therefore, by analyzing the results of this clustering, it can be recommended that a new hotel is built in a colonia in Cluster 1. Notably, the colonias in the southern area of the Cuauhtémoc borough would be a more favorable option to build a hotel since those neighborhoods are at a closer proximity to some of the major tourist attractions—such as the Bosque de Chapultepec—than their northern cluster counterparts. Nevertheless, there are still other factors that must be considered before building a new hotel other than just competition. Some of these factors include the populations of each distinct colonia, the income distribution, and available land for development in each neighborhood. Additionally, there would need to be even more research on local offerings and as well as hotel offerings to stand out. More intrinsic qualities such as aesthetic and theme would also have to be considered prior to development.

Conclusion

Now that we have concluded the process of formulating a business problem, gathering and preparing the necessary data and employing appropriate data science and machine learning techniques, we have arrived at a strategic recommendation. The recommendation to the business problem of identifying the ideal location to develop a new hotel in the Mexico City borough of Cuauhtémoc was arrived to specifically by clustering the neighborhoods of Cuauhtémoc based on the frequency of existing hotels in each area. Through this methodology the resulting recommendation is to build a new hotel in a colonia in Cluster 1 which exhibits the lowest frequency of hotels and would present the least amount of competition. The colonias in Cluster 1 are on the northern and southern ends of the borough but are still relatively nearby the center of the city which is home to the most tourist attractions. Nevertheless, with enticing prices and distinguishable accommodations/offerings, interested stakeholders who build a new hotel in one of these colonias could surely generate prominent bookings and thus capitalize on significant revenue.

Appendix



Cluster 0	Cluster 1	Cluster 2
• Tabacalera • Tránsito • Cuauhtémoc • Doctores • Juárez • Roma Norte	• Peralvillo • Obrera • Valle Gomez • San Simón Tolnáhuac • Santa María Insurgentes • Roma Sur • Ex Hipódromo de Peralvillo • Esperanza • Atlampa • Asturias	• Unidad Habitacional Nonoalco-Tlatelolco • San Rafael • Santa María la Ribera • Ampliación Asturias • Morelos • Maza • Hipódromo Condesa • Guerrero • Felipe Pescador • Condesa • Buenavista • Paulino Navarro • Vista Alegre