COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening a new Hotel in Sydney, Australia

By Naci ARSLAN

Graduated 2020 from Istanbul University – School of Business

Contact : https://arslan0007.github.io/

August 2020

Sydney, Australia

Contents

1.INTRODUCTION

1.1 BACKGROUND OF THIS CAPSTONE

Hotel is the place where we can stay with the best amenities without any worries. Generally, most people would like to go to hotels to relax and take a load of their busy lives. Such establishments offer luxury & comfort, along with other benefits that can make a trip enjoyable for everyone that is involved in the proceeding. And people would like to rest during a weekend away in this case hotel could be a good option. If the hotel is close to the airport, some people may use the hotels to avoid missing their flight.

Staying in a hotel is always an exciting occasion, especially for children as they tend to compare it to an adventure. Also Due to the repetitiveness of everyday life, we want a better silently time. In this case, people could like to go hotel. Because in the hotels due to cleanness, peacefulness, changeable day they can feel very nice.

1.2 BUSINESS PROBLEM

The objective of this capstone project is to analyze and select the best locations in the city of Sydney, Australia to open a new hotel. For finding a solution we will use a data science methodology and machine learning techniques like clustering, and this project aims to provide solutions to answer the business question

1.3 TARGET AUDIENCE

This project is useful to the investors and who are interested in looking at a nice location to open or invest in a new hotel in the capital city of the New South Wales Territory, Australia i.e. Sydney.(*List of Australian Capital Cities - Wikipedia*, n.d.) The best area for first-time visitors to Sydney: Darling Harbour in the Central Business District with its many waterside cafés, bars, and restaurants could be the nice option.

2. DATA ACQUISITION AND CLEANING

For solve this problem, we need the do the following steps as shown below:

We need to define the scope of this project which is confined to the city of Sydney, the capital of New South Wales Territory, Australia.(*List of Australian Capital Cities - Wikipedia*, n.d.)

For that, we need a list of neighborhoods in Sydney. (*Category:Suburbs of Sydney - Wikipedia*, n.d.)

Data scraped from sources. Now we need latitude and longitude coordinates of neighborhoods. This is required to get the venue data and in order to plot the map. Venue data, specifically data related to hotels. We will use this data to perform clustering on neighborhoods.

2.1 SOURCES OF THE DATA AND METHODS TO EXTRACT THEM

This data is we scraped from Wikipedia page. contains a list of neighborhoods in Sydney, with a total of 200 neighborhoods. We will use a web scraping technique to extract the data from the Wikipedia page, also we will use the help of Python requests and beautifulsoup packages. Afterward, we will get the geographical coordinates of the neighborhoods using Python geocoder package which will help us to latitude and longitude coordinates of the neighborhoods.

Subsequently, we need to get venue data of those neighborhoods and for that, we can use Foursquare API which has one of the largest databases of 105+ million places and is used by over 125.000 developers. It will provide many categories of the venue data, and in these data, we will focus on the Hotel category, due to help us to solve this business problem put forward.

This project mostly will help us with scraping the data, working with Foursquare API, data cleaning - wrangling, and also we will use machine learning techniques like K-Means clustering and or map visualization we'll get benefits from the Folium. In this next step, we will discuss the Methodology.

3. METHODOLOGY OF CAPSTONE

As we told from the Data acquisition and cleaning section, firstly we need the data to get the list of neighborhoods in the city of Sydney. Fortunately, the list is available on Wikipedia (*Category:Suburbs of Sydney - Wikipedia*, n.d.). So here, we will do scraping the data using Python requests and beautifulsoup packages to get the geographical coordinates as a latitude and longitude due to we will use Foursquare API. In this process, the Geocoder package will allow us to convert an address into the geographical coordinates in the form of latitude and longitude. After that, We will populate the data into a pandas DataFrame with the pandas' library then we will use the Folium package to visualize the neighborhoods on a map. This will allows us to perform a sanity check to make sure that the geographical coordinates data returned by geocoder are correctly plotted in the city of Sydney.

Now time to use API to get the top of 100 venues that are within a radius of 2000 meters. Foursquare API will help us to get this data. For that, we need to register a Foursquare Developer Account due to obtaining the Foursquare ID and Foursquare Secret Key. Then we will able to make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. It will return the venue data as a JSON format and we will extract the venue name, venue category, venue latitude, and venue longitude. In this case, we can check how many venues were returned for each neighborhood and examine how many unique categories can be created from all the returned values. Analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency o occurrence of each venue category. So at the same time, we are also preparing the data for use in clustering. We will filter the "Hotel" as a venue category for neighborhoods.

Terminally, 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 it is particularly suited to solve the problem for this project. In this case, we will cluster the neighborhoods into 3 clusters based on their frequencies of occurrence for "Hotel". Then the results will allow us to identify which neighborhoods have a higher concentration of hotels while which neighborhoods have a fewer number of hotels. Based on this occurrence of hotels, it will help us to answer the question as to which neighborhoods are most suitable to open a new hotel.

4. RESULTS

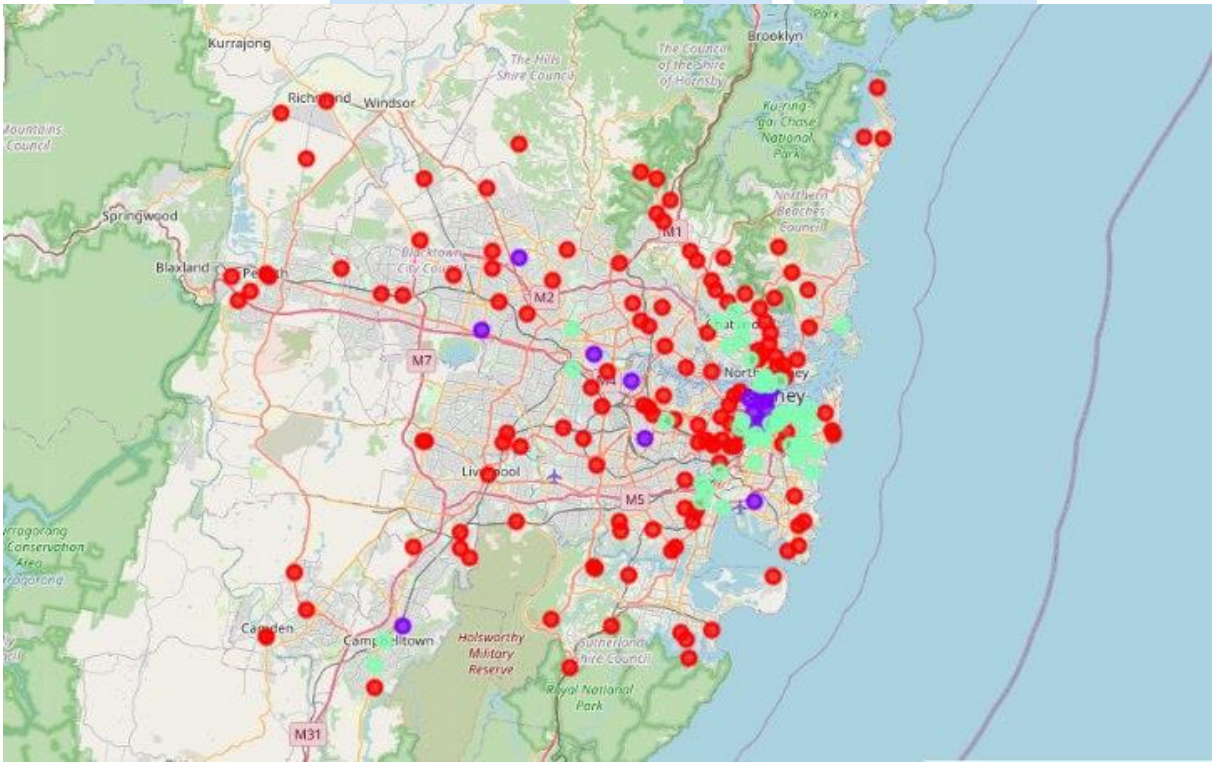
The results of this data from the K-Means clustering show that we can categorize the neighborhoods into 3 clusters based on their frequency of occurrence for "Hotel". We can clarify these 3 clusters like as shown below:

Cluster 0: It is representing neighborhoods with a moderate number of hotels

Cluster 1: It is representing of neighborhoods with a low number to no existence of hotels

Cluster 2: It is representing neighborhoods with a high concentration of hotels.

As we can see from the results of clustering are visualized in the map below with cluster 0 in red color, cluster 1 is in purple, and cluster 2 is in mint green color.



Sydney, Australia I: Clustering with Machine Learning

5. DISCUSSION

As we can notice from the map in the Result section, we can see the highest number in cluster 2 is the most of hotels are concentrated in the central area of Sydney. If we look at other clusters such as cluster 1 has very low numbers to no hotels in the neighborhoods and moderate number in cluster 0. These represent high potential areas to open a new hotel and of course, it will be a competition between other hotels but these areas have less competition chance from existing hotels. Meanwhile, hotels in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of hotels, also in my research of that area for rent a room in a hotel, as I saw in forums; also central city has a parking problem , in this case, these can be effect badly to choose a hotel in this areas. If we look another perspective, the result is also showing that the oversupply of hotels has mostly happened in the central area of the city, but as shown from the map also other areas have hotels but not as often in the central area. Hence, the project will recommend property developers to capitalize on these findings to open a new hotel in cluster 1 without petty competition. Real estate agents with unique selling propositions to stand out from the competition can also open a new hotel in cluster 0 areas which are moderate competition. Eventually, real estate agents or investors are could be advised to avoid neighborhoods in cluster 2 which already have a high concentration of hotels.

6. CONCLUSION

First of all, we were through the process of identifying the business problem and afterward specifying the data as required, extracting as we need it, and also preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. real estate agent or investors regarding the best locations to open a new hotel. As a suggestion for this situation; as shown from the map we seeing the neighborhoods in cluster 0 are the most preferred locations to open a new hotel. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new hotel.

6.1 EXTRA RESEARCH AND COMBINE WITH THE CAPSTONE PROJECT

But if we follow this way and, also make extra searches such as which area is a good option for fun, and also good for parking in Sydney, etc. Then we can combine these areas with the result of this machine learning project. On that occasion, our decision would be more sensible to choose the location of the hotel area. And of course, we shouldn't forget the owner's decision, his decision could be change as depends on his taste. For that, we can separate areas such as Restaurants, Sightseeing, Families, Nightlife, etc. We can start to do an empathy like:

As a tourist, customer, or consumer (have differences), which choice could we have and for these choices which place could perfect match with that?

May we offer to them these areas as shown below:

"Circular Quay and The Rocks" areas; for mostly who is like the sightseeing example; Opera House, Harbour Bridge, shopping of Darling Harbour, and city center.

"King Cross" for who would like to spend their time mostly for nightlife.

But if you prefer to stay a bit far from the city center and also near the water like surfing, swimming, hiking. etc. It could be a good option to choose Bondi and Manly location for families.

Now at the same time if we check our map which we were built-in machine learning project, the areas of what were we talked about it:

Circular Quay and Rocks, it shown in the map as Cluster 1 which means has very low numbers to no hotels in the neighborhoods and these represent high potential areas to open a new hotel.

King Cross on the map it is showing as a Cluster 2, and Cluster 2 is likely suffering from intense competition due to oversupply and high concentration of hotels.

(*Where To Stay & Best Areas in Sydney – Updated for 2020*, n.d.)

Note:

And we can make this project more detail.

7. LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

In this capstone, we only contemplate one factor which is the frequency of occurrence of hotels, also other factors exist like; parking areas, does the location is close to the airport? or shopping centers, an income of residents that could influence the location decision of a new hotel and population. For future research, the researcher can follow a methodology to estimate data to be used in the clustering algorithm to determine privileged locations to open a new hotel. Besides, this capstone letting us use free Sandbox Tier Account of Foursquare API which is have limitations as to the number of API calls, and with this way results are returned. In this case, this API situation could make use of a paid account to bypass these limitations, and then you can obtain more results about it. (*Foursquare Developer*, n.d.)

8. REFERENCES

Category:Suburbs of Sydney - Wikipedia. (n.d.). Retrieved August 26, 2020, from https://en.wikipedia.org/wiki/Category:Suburbs_of_Sydney

Foursquare Developer. (n.d.). Retrieved August 26, 2020, from https://developer.foursquare.com/

List of Australian capital cities - Wikipedia. (n.d.). Retrieved August 26, 2020, from https://en.wikipedia.org/wiki/List_of_Australian_capital_cities

Where To Stay & Best Areas in Sydney – Updated for 2020. (n.d.). Retrieved August 26, 2020, from https://santorinidave.com/best-places-sydney