**Predicting Optimal Hotel Locations in Australia.**

Nathan Wee

February 9th, 2020

Introduction:

**Background:**

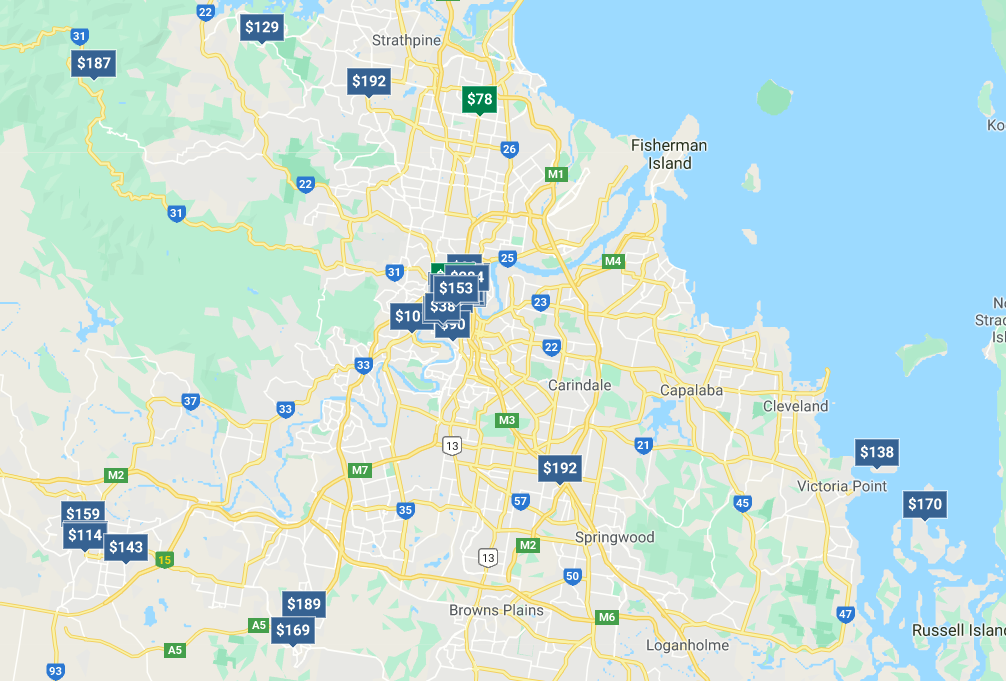
Hotels have always been an integral part of the tourism industry in Australia forming up to nearly a billion dollars in asset value. They are dedicated to providing hospitality to travelling tourists and businessmen from both domestic and foreign areas. Each hotel is judged by these customers by the features and the pricings it offers. However, the ratings of a particular hotel does not just rely on the quality and variety of the services it provides to clients, but rather the location and convenience for tourists or locals to travel to their intended locations. Henceforth, as it is paramount for a potential hotel to flourish properly, one of the many variables its owner needs to consider is its location – by which its optimal place can possibly be predicted by testing, evaluating, and using a model.

**Problem:**

Data provided will show the locations of hotels at their respective locations. In this project, the number and locations of hotels in the Australian cities of Brisbane and Melbourne, and how the following analyses of data will aid in advising relevant groups with where a hotel could ideally be opened at.

**Interests:**

This report will greatly interest entrepreneurs and pioneers with the interest in kickstarting hotels dedicated to certain services as the model therein provides information on the appropriate locale for these sites of tourism to be stationed at. Alternatively, this model works on other potential entrepreneurs of tourist entertainment. This is because as tourists tend to visit landmarks within the vicinity of each other, cluster marketing has become a popular strategy amongst business owners in common tourist locations.



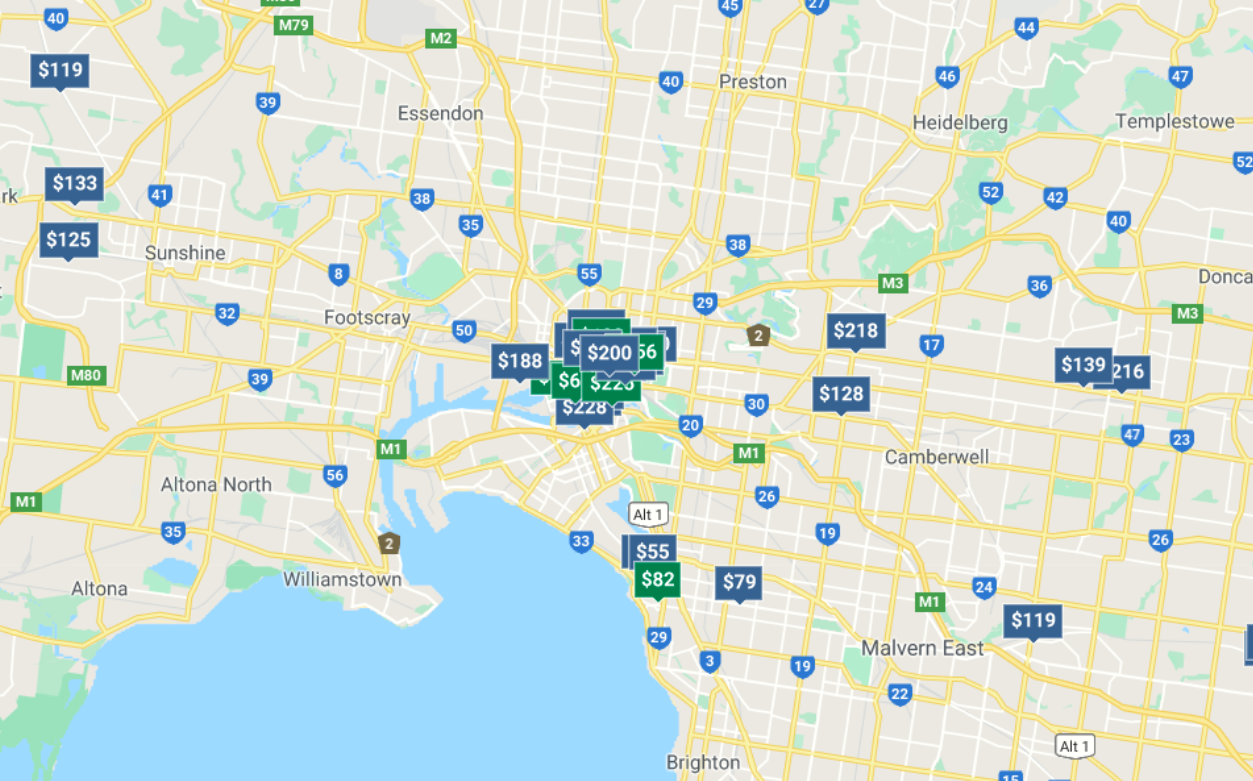


Figure 1: A map of the Australian cities of Brisbane (above) and Melbourne (below) with all surrounding hotels marked.

Data Type:

**Data Sources:**

The following data type that will be used in this report is derived from making regular and premium API calls to Foursquare for obtaining co-ordinates and ratings respectively. The coding on obtaining the following data will be illustrated on the repository. In general, a ‘for’ loop is required for automating continuous calls into a list of data which then gets placed into a dataframe. The following information will be derived from hotels within a 10km radius from the centre of Brisbane and Melbourne.

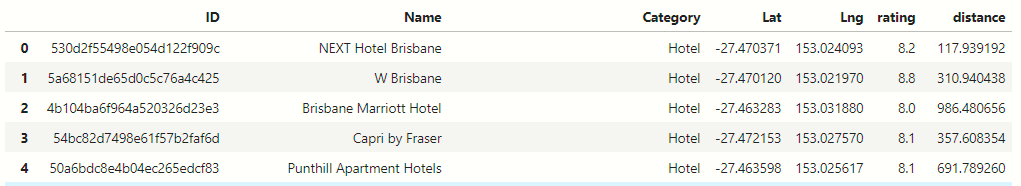
**Data Cleaning:**

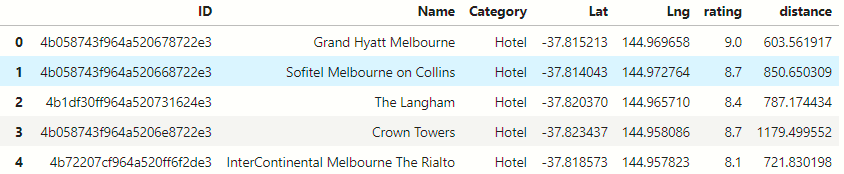
The data cleaning comprises mainly of calling from API and having it return as geojson file containing all the hotels in this specific area, with longitude and latitude co-ordinates. This information is required to be tabulated into a csv file and thus read into the Jupyter notebook through the pandas import.

With regards to the ratings, obtaining them through premium APIs in through Foursquare’s features proved to be troublesome and limiting due to relying on its free feature. Therefore, ratings can be individually obtained through the API call from each city and processed into a csv file. The same method also applies for finding ratings for each museum.

Exploratory Data Analysis:

**Procedure:**



The figures above represent 5 of the 30 hotels present in Brisbane (above) and Melbourne (below). The same examples will be carried and tabulated into a larger csv files for the capital cities of every Australian state for processing.

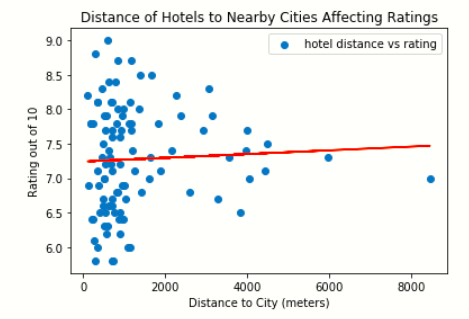
All nearby hotels will have their co-ordinates compared to the co-ordinates of the middle of the nearest city. This is because tourists stay at hotels which tend to be within walking distance from certain popular landmarks in the city. This will show the possible relationship between accessibility to these landmarks and how it affects ratings.

Therefore, data analysis occurs that for each hotel present in Brisbane and Melbourne, a score is created out of utilizing Pythagoras’s theorem from the absolute differences between the locations of the hotel and city centre. This therefore creates a consistent and reliable way that represents the distance the hotel is away from the centre of the nearest city. The ratings are obtained through a technique of calling each venue based on its ID and running a ‘for’ loop for each venue applied.

This score will then be plotted on the x-axis of a graph with the ratings on the y-axis to compare the two. Furthermore, if there is a comparison, a simple linear model can be made, tested and evaluated to be used on the possible locations of a hotel. Conversely, this following developed model can help with predicting ratings if a hotel is started on a specific location. These models are completed by extrapolating the data from a linear equation and determining the value of one variable by the value of the other.

**Results:**

The results of this the data collected shows a visible lack of relationship between the distance of the centre and the hotel to its ratings. This is evidenced by its correlation score of 0.002295, evidencing the null hypothesis that these variables are mutually exclusive. However, the model has returned three predicted values from inputting distances at 500m, 1000m and 2000m, being 7.25478732, 7.26831478, 7.37653443 out of 10.



Conclusion:

**Discussion:**

Although multiple variables influence the ratings of Australian museums, this investigation in particular has constructed a trained model between only two variables, making it of the simple regression kind. As a result, what the model demonstrated was that the independent variable (which is the distance between the museum and the centre of the city) has little effect on the museum’s ratings, and that there are more influencing variables present at large. Which does indicate that as long as a museum is planned for construction from a reasonable distance from a capital city, ratings would not be heavily influenced by where it would be.

**Future Directions:**

Given that those variables are able to be quantified and tabulated into a csv file, they can be placed into a multiple linear regression model for more accurate predictions to be made. However, not only would obtaining more data prove to be troublesome and expensive, the investigation is required to utilize foursquare APIs for the purpose of data processing.

Furthermore, the maximum distance of calling one venue with a specific location is 100km, meaning that a few hotels in rural areas or near non-capital cities are not included in this investigation. Had the data been included, it would make the following model more precise but not accurate.