Data Mining

Project 1 – Group 4

Joshua Ball R11330455 | Dipendra Yadav R11538518

GitHub link: <https://github.com/dipenpratap/Data-Mining/tree/master/Project%201>

# **Project Report**

## **Problem Objective**

Traveling is expensive overall, but one of the biggest costs associated with a trip is often the lodging.

Studies have suggested that the best time to buy a plane ticket for lower-price options is 70 days in advance.

But does the same reasoning apply to hotel rooms?

### **Can this problem be solved using Data Science?**

We are using data from https://www.kaggle.com/ekretsch/hotel-booking-dataset to develop a clustering algorithm

that can find out the best time of the year to book a hotel room in order to get best prices in the market.

## **Problem Understanding Phase:**

* 1. Traveling is expensive overall, but one of the biggest costs associated with a trip is often the lodging. Our dataset consists of several years of hotel booking information, in which we will develop a clustering algorithm that can determine the best time of year to book a hotel room based on the price.
  2. We will clean and prepare the data and then explore the data in order to determine the most probable relationships in order to have the most accurate modeling. We both agreed after carefully looking over the dataset and from life experience that time and cost are the most important factors to focus on.

## **Data Preparation Phase:**

1. In order to properly prepare the data for analysis we must first ensure it is clean and no outliers or unnecessary data exists.
2. Dipendra first looked at the dataset holistically, looking at object data types and their respective attributes. He noticed many variables are carrying many ‘NaN’ values. The main outliers are ‘country’, ‘agent’ and ‘company’. Given none of these variables have any value to our main problem objective, we drop these columns from our dataset.
3. In order to use variables in a K-means clustering model, the variables involved in the modeling must not be categorical. One of the main variables in our dataset in ‘arrival\_date\_month’ which contains the arrival date for each stay by month name. Joshua used a simple dictionary for the ‘arrival\_date\_month’ variable to transform each variable from its respective name to its integer equivalent.

## **Exploratory Data Analysis Phase:**

1. In order to better understand the variable relationships in our dataset, to check the most common travel times, we explore the variable ‘arrival\_date\_month’. From this exploration we see that the most common months of travel are August followed by July, May and October. Dipendra then use linear regression to find the correlation matrix with a heatmap. From this heatmap, there are clear strong relationships between the variable ‘adr’, or average daily rate, and the time the family stays in the hotel.
2. Our exploration focuses on variables based around the cost of the stay at the hotel and the time the stay occurs. Cost is simplified in this dataset to a variable ‘adr’ which is average daily rate. In order to see how much the average family spends, Joshua created a new variable ‘rate\_per\_stay’ by summing the total duration a family stayed and multiplying it by their respective average daily rate. Plotting the relationship of ‘arrival\_date\_month’ versus ‘rate\_per\_stay’ we can see a clear spike in price in the summer months and a strong tapper off into the winter.
3. From further inspection of the data, Dipendra also gathered that the number of guests increases quite dramatically in the summer months. It can be inferred that this is a strong piece of the puzzle in why costs are increasing so much during this time – classic supply versus demand. This strong correlation shows that in the months of April through September when the most people are traveling, the cost of hotels goes up dramatically. This also shows that in the winter months of October through March when the number of travelers decreases, so does the price of the hotel rooms. Therefore, as expected we can see that the price of a hotel room is directly proportional to the number of guests. The price of the hotels increase as demand increases. From our observations, in order to get cheaper prices on hotels, it would be ideal to make trips during the winter months.
4. We will validate this correlation further in the next phases as we move to create a K-means clustering model focused on finding the time of year where hotels are the cheapest.