**Application of Data Science for Developing a Hotel Selection System**

# Motivation:

In today’s data-driven world, several e-commerce businesses have emerged that use data to help customers find their desired items which match to their personal preferences. These businesses provide customers with an online selection system in which they can filter items based upon particular information and features. In fact, selection systems collect classified information about a group of products or services and filter them based on the specific needs of individual users. For example, online travel agencies, such as Expedia and Priceline, help travelers with finding available hotels/properties that meet their expectations regarding price, star rating, number of rooms, amenities, location, etc.

# Introduction to the Project:

In this project, you are going to help with building a “Hotel Selection System”, using two datasets, i.e., Property\_details and Order\_details. All the columns in these two datasets are explained in detail as below:

## Property\_details:

*propertyid*: The ID of the property. *propertyname*: The name of the property. *address*: The address of the property.

*city*: The city where the property is located. *country*: The country where the property is located. *zipcode*: The zipcode of the property.

*propertytype*: The type of the property, i.e., Hotels, Inns, Motels, etc. (8 types in total). *starrating*: The star rating of the property, from 1 to 5 (e.g., 4 indicates a 4-star property). *latitude*: The latitude of the property.

*longitude*: The longitude of the property.

*url*: The website link to book the property (not necessarily the official website).

## Order\_details:

*id*: The id of the order.

*PropertyCode*: The ID of the property.

*dtcollected*: The date of collecting the information of the property.

*reservationdate*: The date that the order was received.

*los*: The length of stay (i.e., the number of nights that a guest stays in the property).

*guests*: Number of guests who stay in the property.

*roomtype*: The room type that is booked.

*onsiteprice*: The price a customer would pay for the order.

*ratedescription*: The booked room description, which may include room size, number of bedrooms and beds, etc.

*ratetype*: Booking information, including cancellation policy, possible pay in hotel option, and extra low-price reminding.

*sourceurl*: The website link to book the property (not necessarily the official website).

*roomamenities*: The amenities available of the booked room.

*maxoccupancy*: The maximum number of guests allowed to stay in the booked room.

*ispromo*: Whether promotion is used when the order is placed (Y means in promotion, N means not in promotion).

*closed*: Whether the hotel is closed when the order is placed (Y means close, N means not close). *discount*: The monetary value of the discount that an order received (already been deducted from the “*onsiteprice*”).

*promoname*: The name of the promotion that the order participated in. *proxyused*: The proxy used to retrieve the order data. *mealinclusiontype*: The type of meal included in the order.

*hotelblock*: Whether the hotel was fully booked (Sold Out means no rooms are available for an order, Blank means rooms are available for an order).

*input\_dtcollected*: The date inputting the collected data of the hotel.

You are required to use the Pandas and Numpy packages in order to read the above csv files and analyze their data.

**QUESTIONS**

1. Part (a): How many properties are there in the region with the zip code of 84100?

Part (b): What is the mean, standard deviation, median, min, and max of “starratings” for all properties in the region with the zip code of 84100?

1. Part (a): Create a new column called "weekday", which is the day of the “reservation date” in one week (for example, if the reservation date is 2021/10/22, the corresponding value in the new column "weekday" should be “Friday”).

Part (b): Which "weekday" does receive the most reservation?

1. Part (a): In “roomamenities”, what are the top 10 common room amenities? What about the 10 least common room amenities? ( for example, Air conditioning is one amenity)

Part (b): What percentage does each type of room amenities occupy of the total number of reservations for all properties?（do not use the total number of amenities as denominator）

1. Part (a): For each property, there are some abnormal values of 0 in the “onsiteprice”. To better organize the data, you would like to create a new column “replaced onsiteprice” in the dataset by retaining the original non-zero “onsiteprice” of one specific property and replacing the zero value with its median of non-zero “onsiteprice”.

Part (b): For each property, calculate the maximum and minimum value of “replaced onsiteprice”, and store these two into corresponding two columns named “Maximum” and “Minimum”. Then create a column named “Normalized Maximum” to store the normalized form of the “Maximum” column. You can use the formula below for the normalization (do not round the result). Store the “hotelcode”, “Maximum”, “Minimum”, “Normalized Maximum” to "Mx\_Min Price.csv".

$ X\_{norm} = \frac{X-X\_{min}}{X\_\max-X\_{min}}$

1. Part (a): A family of three is planning a trip. How many available hotels do offer a room with the “maxoccupancy” of 3 or more? Available hotel are those whose “propertype” are “Hotels”, “close” are “N”, and “hotelblock” are not “sold out” .

Part (b): If this family does not want to pay a room for a “replaced onsiteprice” higher than 150 per night, how many hotels are still available? Use the maximum of “replaced onsiteprice” to compare with 150 due to price fluctuation.

1. Part (a): For each country, find the most expensive property by using “replaced onsiteprice”. Provide id, name, city, country, zip code, address, and average “replaced onsiteprice” of these properties.

Part (b): For each country, find the cheapest property by using “replaced onsiteprice”. Provide id, name, city, country, zip code, address, and average “replaced onsiteprice” of these properties.

Hint: Each country has numbers of hotels, and each hotel has numbers of prices due to price fluctuation. You need to find the average “replaced onsiteprice” for each hotel first, and sort out the cheapest and the most expensive hotels then.

## Business Questions

1. **How do prices compare with max occupancy in switzerland ?**

From analysis on the data provided, the following questions were formulated. The first question that was tackled was if there was a relatioship that existed between max occupancy and the onsite price. The case of switzerland was  selected to be the focus of the question. From the boxplot of max occupancy against replaced onsite price, the replaced onsite price,increased, generally as the maximum number of occupancy increased. The maximum occupancy of one had a lot of outliers and also was  higher than two and three maximum occupancy. This is because of the large variaety of property types that offered one maximum occupacy. The type of properties that offer two or more maximum occupancy, reduce significantly, and those that do, have the same characteristics and therefore, charge the same. A case study was selcetd because, prices change significantly from country to country, and each country has its unique characteristics in terms of price and property types.

1. **Which is are the most expensive cities in Greece ?**

Greece is known is for its rich historical cultural heritage, it therefore makes sense to identify which areas are the most expensive; for customers to plan ahead on the charges they expect to incur or select other areas. In determining the most expensive cities in Greece, the cheapest cities were also identified, therefore the customer was given more option. For one intending to start a business, one can determine how to set the appropriate prices if they are start the business there. The most expensive cities were Mykonos, Naxos Island, Kamares, Kylini, Tsilivi and Corfu Town. The error bars on the plot showed that prices  are not fixed, they have a range, a huge one, therefore, the properties charge prices on their own merit, it is a free market.

1. **Which day has the highest discount generally for the whole dataset and that of Italy ?**

The relationship between discounts and the day of the week was explored. Based on the day of the week, hotels could be offering some kind of discount. There were two approaches in this instance. The whole dataset, that is all countries and the second one was that of Italy. Sundays in both instances offered the largest discount and on Thursday. Sundays because, probably people are travelling to work conferences that take place when the week starts and on Thursday, it is end of week and people are making reservations for weekend entertaininment, that begins at end of Friday's working day. The treend above howeveris not the case for Italy. The discounts do not differ much in the global dataset, but one can note that for Italy, the discounts for different days have significant differences.

1. **Compare the starratings and discounts in Romania and France?**

Discounts vs Star rating: From the plot above comparing French and Romanian hotels, there seems to exist a relation between the amount of discount offered by a hotel and how customers rate the hotel. First, French hotels offer higher discounts, than Romanian Hotels. In both cases, hotels offering higher discounts  are rated higher than hotels that offer lower discounts.

1. **Compare discounts in United Kindom and Germany ?**

United Kingdom vs Germany

Four Star rated hotels and two star rated hotels in the United Kingdom are more expensive than those in Germany. Three star rated hotels in Germany are more expensive than in the United Kingdom. The information will help one determine the prices of the hotels if they were to set in Germany or the United Kingdom, or customers to plan ahead, if they are travelling through the European Union.

## Analysis 2 Business Questions

1. **Which 10 cities are the most frequented cities in the world ?**

The approach taken for this assignment was to look at the best performing cities in the world, and try to benchmark with them to determine which aspects of the recorded data, made them to be reserved in the world. London and Paris had the most orders to reserve hotels when compared to the other hotels. It is important to note that there is a huge gap between the reservations made in the first two cities, when compared to the rest of the cities.

1. **How do prices compare with max occupancy in the 100 most frequented cities ?**

The second step was to look into the price differences of hotels in the top 10 cities that were set aside in the first step. Since prices are also impacted by the max occupancy of the property. From the plot, note that London offers property with occupancy that goes beyond 6. However, the prices for this property, the prices are also very high. For the rest of property with max occupancy less than 6, the prices are generally the same, except for Edinburg amd Barcelona, the two cities are cheaper.

1. **How do discounts compare with max occupancy in the 100 most frequented cities ?**

The next comparison between the 10 cities was that of discounts against max occupancy rooms. London seems to offer the lowest discounts on the property with large max occupancy such as that of 6. London when compared to the rest of the cities generally gives the lowest discounts on the various max occupancy of the property. Manchester and Edinburg offer the highest discounts on room with maximum occupancy from 1 to 3. It is also important to note that hotels in Edinburg, Munich and Berlin, offer  a range of discounts from very discounts to very large discounts.

1. **What is the general distribution of star rating of this cities ?**

London, Paris and Edinburg appear to have ratings that range from 2 to offer, with London and Edinburg having the highest number of hotels in all the cities with the various ratings. This however, might not reflect the actual case because London  has the highest number of hotels in the dataset being considered. So far we can conclude that Edinburg is the best city to make a reservation and or set up a hospitality business.

1. **Which hotels are most frequented in the 10 cities ?**

Note that the hotels with reservations that are over 100 are mostly from the United Kingdom. Out of the top ten hotels only two are from outside the United Kingdomm, one from Spain and the other from France.