Capstone Project Submission

Team Member's Name, Email and Contribution:

1. Puroshotam Kumar Singh: purshotamsingh61@gmail.com
   - Data Cleaning – Handling missing values.
   - Numerical variable analysis – Arrival_date_year, Stays_in_weekend_nights, Stays_in_week_nights, Adults, Babies, Children, Total_of_special_requests.
   - Categorical variable analysis – Hotel, Arrival_date_month, Country, Reserved_room_type.
   - Presentation, Technical documentation.

2. Vikram Pandey: pandey.vicky@yahoo.com
   - Data Cleaning – Imputing null values.
   - Numerical variable analysis – Lead_time, Is_canceled, Is_repeated_guest, ADR.
   - Categorical variable analysis – Reservation_status, Deposit_type, Distribution_channel, Market_segment.
   - Presentation, Technical documentation.

Please paste the GitHub Repo link.

Puroshotam's Github Link:- https://github.com/PuroshotamSingh/hotel-booking-analysis

Vikram's Github Link:- https://github.com/vickypandey07/Hotel-booking-analysis

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

The dataset contains hotel bookings data. One of the hotels is a resort hotel and the other is a city hotel. The dataset have the structure, with 31 variables describing the 40,060 observations of resort hotel and 79,330 observations of city hotel. Each observation represents a hotel booking. The dataset comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Since this is hotel real data, all data elements pertaining hotel or customer identification were deleted.

The problem statement was to identify what impacts booking cancellation, from which country most guests are coming, who did the booking and whether customers repeating their bookings or not.

The first step in the analysis involved taking initial look at the data, looking for any missing values and null values and tackling them.

The second step involved analyzing numerical type features, with the help of different visualization techniques such as heatmap, distplot, bar graphs, boxplots, pie charts, etc. Finding correlation between each variable and also finding the important features that had an impact on cancellation of bookings.

The third step involved analyzing categorical variables such as hotel, arrival_date_month, country, reserved_room_type, reservation_status, deposit_type, distribution_channel, market_segment and finding any underlying pattern that affects the rate of cancellations.

The final step was to point down the insights developed during the analysis of the data. Some observations draw were; increase in lead time increases rate of booking cancellation, increase in ADR also increases rate of booking cancellation, non-refund policy also increases rate of booking cancellation, majority of guests are from Western Europe, mostly couples booked the hotels and majority of customers are not repeating their bookings.