**Instructions:**

a) **The attached PDF has questions on R/Python & Stats. Related dataset (Table A, Table B, Table C and Table D) files are also given.**

b) **Questions can be solved using either R or Python. Output expected: code and csvs generated.**

c) **For the stats question please send solution**

**Dataset Explanation**

There is Brand of hotel chain named X. Attached is the data of the hotel chain, and below is the description

- Table A - Customers who had made booking in Jan 2017
- Table B - Customers who had made booking in Feb 2017
- Table C - Customers who had made booking in March 2017
- Table D – Hotel Mapping with City where hotel is present

Definition of the attributes that are present in the table are as follows –

- Booking Id - Id by which particular booking that has been made is identified
- Customer Id - Unique key for the customer those are making booking
- Source - There are five sources through which customer are making bookings with values from 0 to 4
- Status – Status of the booking.
    - Status 2: customer stayed
    - Status 3: Customer cancelled
    - Status 4: Customer did not turn up
- Checkin - Date when the user checks-in into the hotel
- Checkout - Date when the user checksout from the hotel
- Rooms - # of rooms booked
- Hotel_id - It is the hotel_id in which we had booked the Room
- Amount – Amount paid by the customer.
- Discount – Discount given while booking
- Date - it is the date when the booking has been created.

Abbreviation Definition -

- Revenue ->Total Amount Paid i.e amount
- Room Night -> Rooms*(checkout –checkin)
- Average price per room –> Revenue/Room Night

**Solve using R/Python :**

Please provide well commented codes for the following:

1. Import all 4 datasets provided in the previous question.
2. Extract unique users for each month and calculate total number of bookings made by each, total amount spent in each month, total room nights stayed (status2) for each user for each month.
3. Merge these summarized datasets to create one dataframe such that you can see all these summarized columns for each month side by side. Below is an example of the output:

| Guest_id | No_bookings.jan | Total_room_nights.jan | Total_amt.jan | No_bookings.feb | Total_room_nights.feb | Total_amt.feb | ... |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 8 | 6000 | 1 | 2 | 1800 | |
| 2 | NA | NA | NA | 2 | 4 | 3000 | |

4. Calculate Repeat Rate for the month February (If X customers had made the bookings in the month of Jan 2017 (TableA), how many of them made them in Feb 2017 too. (TableB) too i.e Y) - (Y/X*100)
5. For each city, give the top 3 revenue earning hotels over this time period. (Not se parately for Jan, Feb, Mar)

**********

**Q.3** Consider two random variables x and y. x can assume values 0 and 1 with some unknown probabilities. Also y can assume values 2 and 3 with probabilities 0.4 and 0.6 respectively. We also know that Prob[x=0|y = 2] = 0.5 and Prob[x =0|y = 3] = 0.8, the probabilities are conditional probabilities. Now suppose we observe the value of x to be 0 and based on this observation we predict t he value of y to be 3, what is the probability that our prediction is correct?

**********