Test

We are pleased to invite you to the interview process for our Decision Science Team! This is a practical exercise that will test your programming and analytical skills, please **include your codes as a PDF** in the submission. The programming language that is acceptable is python or R.

Instructions: Please read carefully

- **Submit 1 pdf file with all the answers. The submitted pdf file name should be in '<your_full_name>_<date>.pdf' format.**
- **Your code, comments & output should be present in the pdf. Please make sure that all the output code and text are organized and readable in the submitted PDF.**
- You may not consult with any other person regarding the test.
- You may use internet searches, books, or notes you have on hand.
- The test has 7 parts, **all of which are mandatory**. Failing to complete any one part would result in the rejection of the submission.
- In case of doubts please make thoughtful assumptions.

Part 0: Reading the data

- Please find the data (test_DataScience.xlsx) and take it as the input (as data frame).
- Print all the column names and the data types in each column.
- Print the cities of India from which the page was accessed.
- Write a brief paragraph about what you think about this dataset along the lines of :
    - which geo-location this dataset belongs to?
    - Given that this dataset is for a website like Flipkart, what could be the possible definitions of the columns Level 1, 2, 3, 4 in the given dataset?

Part 1: Data cleaning

- Write a function called data_cleaning() which, when called, would perform the following activity:
    1. Create a new column, called 'Month_Year', using lambda function. The new column should be at the 3rd position from the start in the given dataset & its values should be : '01-01-2020' for January, 2020 and '01-02-2020' for February 2020 and so on. (snippet added)
    2. Replaces the null values with the average of the respective column in the data.
    3. In column 'B' replace Jan with 1, feb with 2, march with 3 and so on.
    4. In column 'E' Replace "Came_From_LinkedIn" with "LinkedIn" and "Landed_Directly" with "Direct_traffic" .

Part 2: Descriptive statistics

- Write a function called descriptive_stats('Year', 'Month' , 'Laptop/Desktop' , 'Type_of_Customers?' , 'Coming from' , 'Place_in_India') which, when called, would perform the following activity:

1. Would filter the dataframe with the given parameters; if any parameter is missed, then consider a default value to that parameter (e.g., default: 'year' – 2020, 'month'-Jan, & so on) . Let's call this new dataframe 'df'.
2. Generates the summary statistics (Mean, Median, Quartile, standard deviation) of all the numerical columns of the new dataframe, df.
3. Produce a list of all the unique values & data types present in the non-numeric columns in df.

Part 3: Prescriptive statistics

- The marketing manager has asked you the following questions, please provide the answers along with summarized data supporting your answer.
    1. What are the top 3 "Place_in_India" on the basis of column "Level 1" for the year 2021 and 2022 separately ?
    Below is a snippet of the data that is requested:

A	B	C	D
Year	Rank by column 'Level 1'	Place_in_India	Sum of Level 1
2021	1	city1	Add all numerical values of column 'Level 1'
2021	2	city2	Add all numerical values of column 'Level 1'
2021	3	city3	Add all numerical values of column 'Level 1'
2022	1	city1	Add all numerical values of column 'Level 1'
2022	2	city2	Add all numerical values of column 'Level 1'
2022	3	city3	Add all numerical values of column 'Level 1'

    2. Please, provide the data for all the cities & for all the years, the following format as shown in the below snippet:

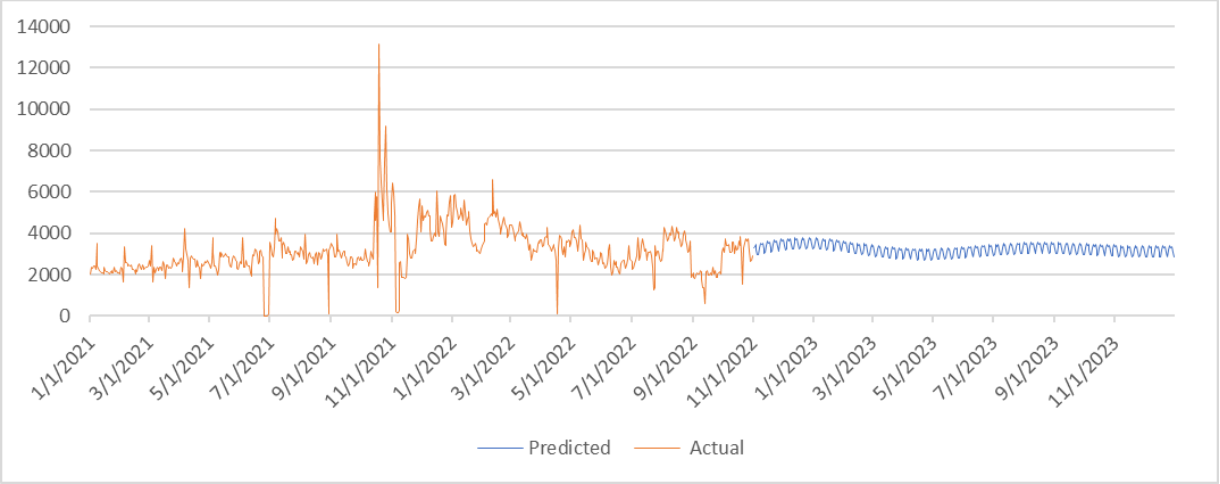
City	Year	(sum of Level 2) / (sum of Level 1)	(sum of Level 3) / (sum of Level 1)	(sum of Level 4) / (sum of Level 1)
city1	2020			
city2	2021			
city3	2022			
city 4				
...				

    3. What are the bottom 3 "Place_in_India" on the basis of column "Level 4" for the year 2021 and 2022 separately ?

Part 4: Simple Machine learning questions

- Write a function called predict_future('Year', 'Month' , 'Laptop/Desktop' , 'Type_of_Customers?' , 'Coming from' , 'Place_in_India') which, when called, would perform the following activity:
    1. Predict "Level 4" for the 12 months of 2023 given the parameters of the function. (Please make sure the parameters have default values in place)
    2. Generates the overall Forecast error, MAPE and RMSE of your prediction of the year 2022, 2021 & 2020 for the given parameters.
    3. Plot a line graph of the level 4 actual numbers from 2020-2022 & in the same graph, there should be the predicted numbers for 2023. The x-axis should be the timeline from 2020 Jan to 2023 Dec and the y-axis should be the value of the level 4 column, The below graph is just an example of

how your plot should look like.



You may use only Huber regression, ARIMA or prophet for forecasting.

Part 5: Visualization

- Please write a code to display :
    1. A line graph for “Level 2” for the different “Place_in_India?” over the months of the year 2020 & 2021.
       (Hint: On x-axis, there should be months for 2020 & 2021 and Y axis should be “Level 2” and there should be different lines depicting different regions of “Place_in_India?”)
       Plot a neat graph.
    2. A line graph for “Level 1” for the different “Laptop/Desktop” over the months of the year 2020 & 2021.
       (Hint : On x axis there should be months from jan- 2020 to dec- 2021 and Y axis should be the sum of “Level 1” and there should be different lines depicting different devices used.)
    3. A line graph for “Level 2” for the different “Coming from” over the months of the year 2021 & 2022.
    4. A line graph for “Level 1” and “Level 2” over the months of the year 2020, 2021 & 2022.
    5. A line graph for “Level 3” foyearslace_in_India” over the months of the year 2020 and 2021.
    6. Please add any insights you could derive from all the graphs above.

Part 6: About the Previous projects

- Please describe any interesting project you did in the Data Science domain in more than 250 words. Attach Github links if possible.

Part 7: Time management

- Can you please share your thoughts, in less than 120 words, on “If you get selected, how will you manage your time for this full-time internship opportunity”

Best of luck!