**Test**

We are pleased to invite you to the interview process for our Data 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 o

**Part 0: Reading the data**

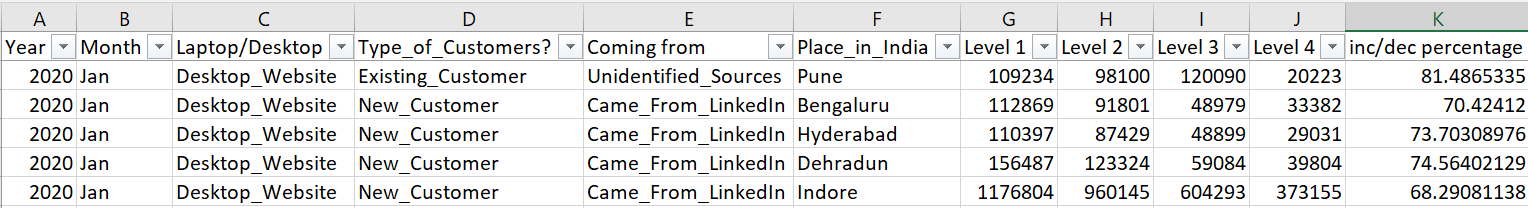
* Please find the data (test\_DataScience.xlsx) and take it as the input ( as data frame ).
* Print all the column names and the data types in each column.
* Print the cities of India from which the page was accessed.
* Which columns are having some values missing in them?
* Write a brief paragraph about what you think about this dataset along the lines of :
  + - What type of company this dataset belongs to?
    - Suppose 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? Do you observe any pattern?
* Give the number of new customers who are from Pune and came from LinkedIn.

**Part 1: Data cleaning**

* Write a function called data\_cleaning() which, when called, would perform the following activity:

1. Create a column at the end, named “inc/dec percentage”, which would give the growth/reduction percentage in numbers of level 1 vs level 4 columns.

Like, (“Level 1” – “Level 4”) / “Level 1” \* 100



1. Replaces the null values (if they exist) with the average of the respective column in the data.
2. In column ‘B’ replace Jan with january, feb with february, march with march, apr with April and so on. Use the lambda function to do the same.
3. In column ‘E’ Replace “Came\_From\_LinkedIn” with “From LinkedIn” and “Landed\_Directly” with “Direct\_traffic” .

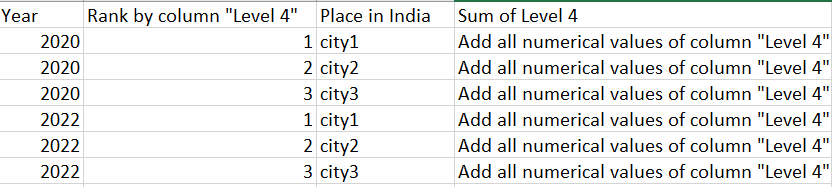
**Part 2: Descriptive statistics**

* Write a function called [descriptive\_stats](https://www.statisticshowto.com/summary-statistics/)(‘Year’, ‘Month’ , ‘Laptop/Desktop’ , ‘Type\_of\_Customers?’ , ‘Coming from’) which, when called, would perform the following activity:
  1. Gives the minimum values present in all the level-columns. (Level 1, 2, 3, 4)
  2. Gives the maximum value of “Level 2” / “Level 1” among those who came directly to the via desktop website.
  3. 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’.
  4. Generates the summary statistics (Mean, Median, Quartile, standard deviation) of all the numerical columns of the new dataframe, df.
  5. Produce a list of all the unique values & data types present in the non-numeric columns in df.

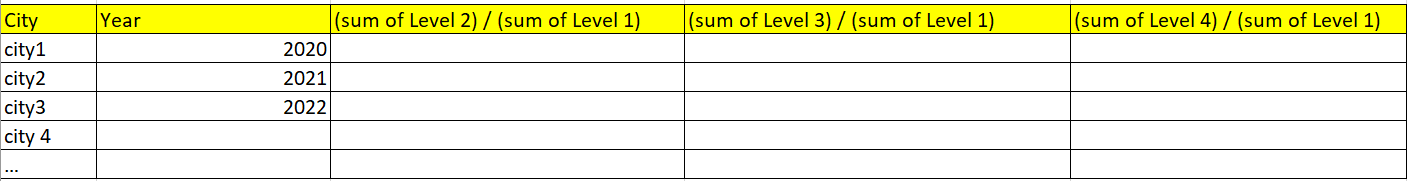
**Part 3: Prescriptive statistics**

* 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 4” for the year 2020 and 2022 separately ?

Below is a snippet of the data that is requested:



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



* 1. What are the bottom 3 “Place\_in\_India” on the basis of column “Level 4”/ “Level 1” for the year 2021 and 2022 separately ?
  2. Which place in India is having “Level 4” value greater than 150000 most of the times?
  3. Which place in India is having least number of existing customers?

Best of luck!