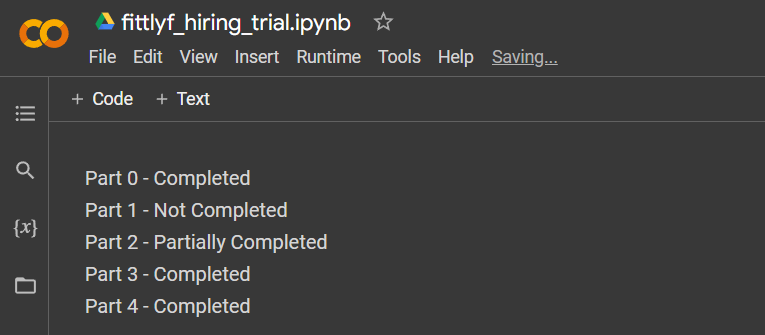
Test

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 **submit your codes as a ‘Google Colab link’** in the submission. The programming language that is acceptable is python.

**Instructions: Please read carefully**

* **Submit 1 colab link with all the answers. The submitted colab notebook’s name should be in ‘<your\_full\_name>\_<date>’ format.**
* **Your code, comments & output should be present in the colab notebook. Please make sure that all the output code and text are organized and readable in the submitted colab notebook.**
* 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 5 parts,** **all of which are mandatory**. Read the questions carefully and answer accordingly. **Each section must to have a visualization**. Failing to complete any one part would result in the rejection of the submission. **Code should be commented properly**.
* In case of doubts please make thoughtful assumptions.

**Start your colab notebook with a checklist mentioning the parts you were able to complete / was not able to complete.**

**Reference:**

**NOTE:**

* Please refer to sheet “Funnel“ in “Intern Hiring Assignment Datat.xlsx“ for Part 0 to Part 3
* Please Refer to sheet “AB\_Test“ in **“**Intern Hiring Assignment Datat.xlsx“ for Part 4

**Part 0: Data Preparation**

* Please find the data (Intern Hiring Assignment Datat.xlsx -> Funnel worksheet) and take it as the input (as data frame).
* Look for the missing values in the given dataset and visualize them using “Missingno” to find out type of missing value the dataset contains. What pattern do you find in these missing values?
* Handle these missing values with appropriate imputation method. To complete this part write a function data\_prep() which takes the whole dataset as the input and returns cleaned data with no null values and print the returned data at the end and a total count of null values.
* Write a brief paragraph explaining why the imputation technique you choose is the most appropriate one to handle these missing values.
* Missingno Documentation :- <https://github.com/ResidentMario/missingno>

**Part 1: Descriptive Analysis**

* 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 an e–commerce funnel analysis, what could be the possible definitions of the columns Level(visitors) 1, 2, 3, 4 and 5 in the given dataset? Do you observe any pattern in these levels?
* Give a suitable visualization and proper pointers to answer the following question:
  + Generate the distribution plot of the Quantitative variable by representing the distribution based on KPI column. (Hint: 5 different distribution plots- one for Level 1 visitors, one for Level 2 visitors…)
  + Choose an appropriate technique to identify outliers in the data. If you find outliers in the data, use your preferred way to handle them. Plot the distribution of the Quantitative variables by representing the distribution based on KPI after the removal of the outliers.
  + What can you analyze from this distribution plot? Considering the whole descriptive summary e.g., min, max, mean, mode etc. Which level has best distribution of visitors?
* Write a function descriptive\_stats which when called, would perform the following tasks.
  + Return a pivot view of the total number of visitors segmented/divided by each level, in each year? (Print the original and pivot view one after the other)
  + Returns the region with maximum no. of Level 5 visitors in each year.

**Part 2: Prescriptive Analysis**

* Create a new feature (Level 5 visitors/Level 1 visitors) and name it as Conversion Rate and use this created feature and proper pointers to answer the below questions.
  + Create a graph that shows how the Conversion rate is spread out across different groups of customers. Look at the graph and share what you observe from it. (Important: Make sure the graph is different from the one you used in Part 1).

Which group of customers has the best Conversion rate?

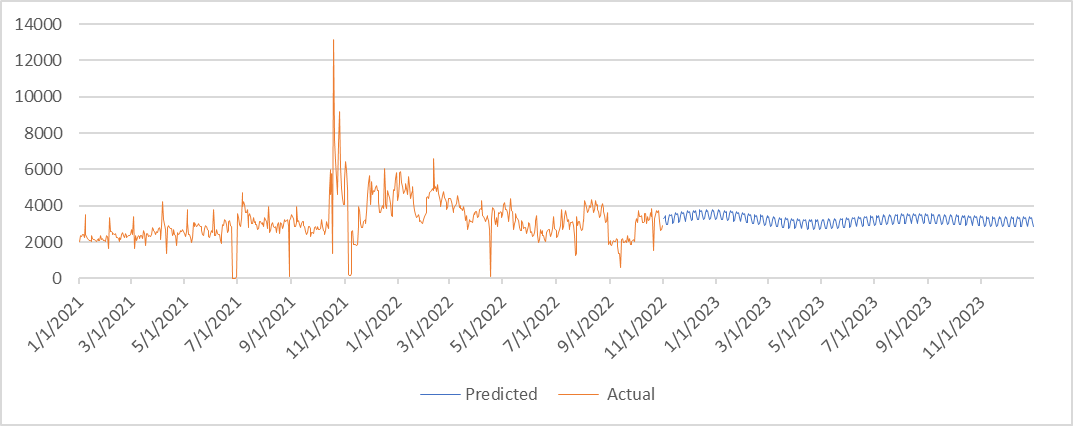
* + Assume you are a Data Analyst at Fittlyf Company, analyze the data for the region performing worst in all the years and prescribe what could be the reason.
  + Identify which region is having a better YearOnYear growth.

**Part 3: Prediction**

* Write a function called predict\_future( ) which, when called, would perform the following activity:
  + Predicts “Level 5” and “Level 1” future values for the next 12 months. Also, plot it.
  + Generates the MAPE and RMSE of your prediction.

Plot a line graph of the level 5 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 5 column and predicted values. Repeat the same for level 1 visitors.

The below graph is just an example of how your plot should look like. You may use Rolling Average and ARIMA for forecasting. (link for reference: <https://youtu.be/jiQM93dmUek>)



**Part 4: A/B testing**

* Using “AB\_TEST” sheet in the shared excel file, what is the possible metric you can create for A/B testing excluding no. of clicks and no. of visitors.   
  (Any derived metric from the given no.of clicks and visitors ) – [link](https://segment.com/growth-center/a-b-testing-definition/metrics/) for reference
* Perform an AB testing to find which variant whether control or treatment is better using the concept of hypothesis testing and statistical significance. Give a detailed step by step explanation of your answer, write down your hypothesis and the whole solution to test that hypothesis.
* Which statistical test did you use to find the result and why? Put brief pointers.
* AB Testing references
  + <https://towardsdatascience.com/a-b-testing-the-basics-86d6d98525c9>
  + <https://vwo.com/ab-testing/>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_