JOBATHON FINAL SOLUTION APPROACH

Data Loading :

- Since the size of VisitorLog dataset file was about 500MB I wasn't able to read the dataset using Pandas from my local machine
- Hence I did some research and found that PySpark is better suited in handling such kinds of data of large size.
- I spun up a Google Colab instance and loaded the data files using a PySpark session.
- Since it was mentioned to only generate target file for registered users I combined both the datasets using inner join on UserID so that I group based on only registered users for further analysis.

Data Imputation :

- There were a lot of null values in most of the columns in the combined dataset and hence these were handled step by step
- Initially removed some irrelevant columns from the combined dataset like Country, City, User Segment etc.
- Then went on to handle the values stored in different cases for the text information. These were converted to a common case.
- The later on analysed each column with a high number of null values and decided to impute them with the mode/mean of occurrences of each column's values based on their data types. For e.g Imputed Activity column with most frequently occurring Activity based on each UserID.
- Then converted the different timestamp formats (both UNIX and datetime formats) to common date format using date formatting functionalities in PySpark.

Data transformation & extraction :

- Some derived features were created for further analysis. Like a Duration feature which can provide the recency of visit of each user.
- These columns were further used to create different dataframes that were used to extract the columns required in the final target dataframe.
- I tried to use groupBy, OrderBy and Window functions mostly to aggregate and extract most of the columns required in the final dataset.