Download all CSV files from here.

Problem Statement:

You are given a raw dataset that contains some missing values, duplicated records, and inconsistent formats. Your task is to clean the data and prepare it for analysis. You will be evaluated based on the accuracy, completeness, and efficiency of your data preprocessing steps.

Dataset:

You are provided with a CSV file **raw_data.csv** that contains the following variables:

- **id:** unique identifier of a customer (string)
- **gender:** gender of a customer (string, values: "Male", "Female") - **dob:** date of birth of a customer (string, format: "dd-mm-yyyy" or "yyyy-mm-dd")
- **income:** annual income of a customer (int)
- **marital_status:** marital status of a customer (string, values: "Married", "Single", "Divorced", "Widowed",)
- **city:** city of residence of a customer (string)
- **score:** credit score of a customer (integer, range: 0-1000)
- **last_purchase_date:** date of the last purchase made by the customer (string, format: "dd-mm-yyyy" or "yyyy-mm-dd")

Tasks:

Your tasks are as follows:

- Load the **raw_data.csv** file into a pandas DataFrame.
- Check the data for missing values and duplicated records.
- Remove any duplicate records from the data.
- Fill in any missing values in the **gender**, **marital_status**, and **city** columns with the mode of the respective columns.
- Create a new column named **age** that contains the age of each customer based on their dob.
- Create a new column named **income_group** that categorizes customers into three groups based on their income values: "Low", "Medium", "High" based on each 33% percentile.
- Create a new column named **score_group** that categorizes customers into three groups based on their score values: "Poor", "Fair", and "Good" based on each 33% percentile.
- Remove any rows where the **last_purchase_date** is before the year 2019. - Save the cleaned data to a new CSV file named **clean_data.csv**.

Evaluation Criteria:

Your test submission will be evaluated based on the following criteria:

- Accuracy of data cleaning steps (40%)
- Completeness of data cleaning steps (40%)
- Efficiency of data cleaning steps (20%)

Submission:

Please submit the following files in a zip folder named data_cleaning_test.zip:

- Jupyter or Google Colab notebook or Python script that contains the code for data cleaning.
- **clean_data.csv** file that contains the cleaned data.

Note:

Please make sure to comment on your code and explain your thought process in the Jupyter notebook or Python script. Additionally, you may be asked to explain your reasoning behind your data-cleaning steps during an interview.

MasterCourse holds the right to disqualify an assignment submission if some moral and ethical issues like plagiarism, hate speech, etc. are found.

Submission Link: *https://forms.gle/PEq7rRQhMDZZxXmK9*

Deadline: 31st March 2023, 11:59 PM.