Data Scientist Intern Assignment

Process Data & Get Insights

NOTE: Use Google Colab Notebook to write all the codes and steps you followed for each task in short for review.

Introduction:

In this assignment, your objective is to process CSV data and create a MySQL relational database. Your work will be evaluated based on your code quality, design decisions, creativity, and the uniqueness of your approach.

Dataset URL :
https://github.com/xscientisttech/detaset/blob/main/india-state-wise-data-analysis.csv

Instructions:

- Each student must submit their own design and code implementation, which should be distinct from others.
- You're allowed to use any external libraries or frameworks such as Pandas, Matplotlib, Plotly, etc.
- Your code should be well-documented with appropriate comments, and you should utilize a Google Colab notebook for this task.
- The assignment completion time will be tracked, and you have 24 hours to finish the assignment, starting from when you begin working on it.

Submission:

Here are the steps to submit your code:
1. Put your code on GitHub. This means creating a repository (a place to store and manage your code) on GitHub's website.
2. Make sure your repository is public. This means that anyone can see the code you've uploaded.
3. Share the link of your repository with us in the following google form.

Upload following files in your repository:
- Your Colab notebook
- SQL file
- An ER-diagram PNG image file

Sub Assignment 1: Processing the Dataset [Level Easy]

Examine the data in each column to see if any preprocessing is required. Preprocessing might involve cleaning up the data, handling missing values, converting data types, or separating combined data into distinct columns. For instance, if a column contains combined data like "first name" and "last name," you should preprocess it to split these into separate columns.

1. Write python functions to pre-process the data and these code should be well commented.

Sub Assignment 2: Exploring Indian States Data [Level Medium]

The provided CSV file is packed with information about various Indian states. Your objective is to extract meaningful insights from this data and visually represent these insights using either the Plotly or Matplotlib library. Your imagination is the limit! For instance, you can pose questions about the dataset such as:

1. "What is the population of each state?" and create a bar graph illustrating the population distribution.
2. You can come up with other questions related to the dataset that intrigue you.
3. You should have at least 10 questions and its visualization on the google colab notebook.
4. Write all your code and comments in each cell.

Note: You can easily read the csv file inside google colab notebook using following code:

```
import pandas as pd
# Load the CSV file into a DataFrame
df = pd.read_csv(
"https://raw.githubusercontent.com/xscientisttech/detaset/main/india-state-wise-data-analysis.csv"
)
```

Remember, your creativity is highly valued, and we encourage you to think outside the box. Your unique perspectives and insights will be greatly appreciated.

Sub Assignment 3: Transforming CSV Data into a MySQL Relational Database [Level Hard]

Your main task is to transform the given CSV data into a MySQL relational database. To do this, follow these steps:

1. Understanding the Columns: Carefully study the columns in the provided CSV data. This step involves identifying the data contained in each column, understanding the relationships between different columns, and recognizing any patterns or potential issues.

3. MySQL Database Creation: Once you've prepared the data, create a MySQL database schema that reflects the relationships between different tables. Create primary keys, foreign keys, and the structure of each table. This will involve designing tables based on the processed data.

4. Populating the Database: Write code or upload csv files directly to populate the MySQL database tables with the preprocessed data from the CSV. This step will involve inserting data into the appropriate tables while maintaining referential integrity.

5. Documentation: Alongside your code, provide comments explaining the decisions you made during the preprocessing, database schema creation, and data population stages. This documentation will help others understand your thought process.

6. Export ER-Diagram: Once the tables are populated export the entity relationship diagram of your database.

Modern product-based startups often assess job candidates by reviewing their open-source contributions and GitHub profiles. This hiring approach emphasizes code quality, documentation, and contributions as key factors, enhancing your chances of making a standout impression.

Good luck