Data Collection, Data Cleaning and EDA Project

This project helps you to apply **Python, Numpy, Pandas, Matplotlib, Regular Expressions, Web Scraping and EDA skill sets.** This project will help you in **business understanding, data cleaning and data visualization in a life cycle of Data Science projects**.

After the completion of the project add it to your profiles such as **LinkedIn, GitHub and CV/Resume which will put more weightage to your Resume and Digital Profiles.**

Case Study Selection

- Select the use case like house price data, vehicle price data, stock market data, etc.. from the list of use cases mentioned below. 👇

- First Hand / Second Hand Cars
- Mobiles/Laptops brands and Prices
- Hotel Accommodations
- Zomato/ Swiggy Food Restaurants and Food items
- Any Sports(Cricket, Football, Basketball) statistics from 2013 - 2022 (Team / Players Performances)
- Covid Analysis with respect to Stock Exchange, Gold prices during the time lines.
- Real Estate (Flat Rate using Area and Zone / Cities)
- First / Second hand Bikes
- IMDB Movies
- Fortune 500 Companies
- Any other with your area of interest

2. Search for Relevant Websites

- Find the websites which are allowing you to scrape the data from the domain/use case you have chosen.

3. Define the problem Statement:

Do some research and try to properly define the problem statement.
Make sure to identify a target feature and other appropriate features relevant to that target feature.

Note 🚩 : Kindly get it reviewed the above steps with the @mentors before proceeding further in the project

4. Extract the Data

- Confirm the features/variables (columns) which are useful according to your problem statement. To do the better analysis and draw the more conclusions about your business problem it would be suggestable to scrap the minimum of 8 features/variables (i.e. columns) and minimum of 400 rows

5. Create a Data Frame

- Convert the scraped data to DataFrame

6. Export into .csv format

- Export the data frame into .csv format

7. Read CSV File

- After exporting the data frame you need to import the data frame for

How many features(Columns) do you have?
How many observations(rows) do you have?
What is the data type of each feature(Columns)?
How many missing values are there?

Note 🚩 : Kindly get it reviewed the above steps with the @mentors before proceeding further in the project

8. Clean the Data :

In this section you have to clean the data like:
- Removing the special characters,
- Incorrect Headers,
- Incorrect Format of the data (Invalid Values, Columns)
- Converting the data types
- Identifying and Treatment of the missing values,
- Identifying and treating the Outliers (Based on the Problem Statement or Data)

Note: Fixing Rows and Columns:

We can merge different columns if it helps in better understand the data
Similarly, we can also split one column into multiple columns based on our requirements or understanding. Add Column names, it is very important to have appropriate column names in your dataframe.

Note 🚩 : Kindly get the above steps reviewed from the @mentors before proceeding further in the project

9. Data Analysis and Visualization (EDA) : # Problem statement : EDA , ratings and gross income : these two features with all other available features : draw some of the conclusions

As we can see there two data types in Statistics

1. **Categorical**
2. **Numerical Data** (Continuous Data and Discrete Data):

Uni-variate Analysis:

- ☐ **Continuous Variables**, calculate Central Tendency and Measures of Dispersion . One can utilize Statistical metrics visualization methods s touch as Box-plot, Histogram/Distribution Plot, Violin Plot, and others.

- ☐ **Categorical Variables**, calculate the frequency distribution/Count and percentage of the Categorical variable. One can use plots like count plot, bar plot, pie chart.

Bi-variate Analysis/Multivariate :

This will help us in understanding the relationship between the variables. Usually the data variables are Categorical and Continuous . We can use this method to find the relationship between any variables

Do the BiVariate/ Multivariate: Groupby, Pivot, Crosstab.

The combinations are mentioned below.

- Continuous and Categorical variables (groupby, pivot table)
- Continuous to Continuous variable (correlation plot)
- Categorical to Categorical variables (crosstab)

Plots we can be drawn based on the data features/columns:

[Box-plot, Bar-plot, count plot, pie chart, scatter-plot, violin-plot,
distribution-plot, heat map, histogram and kde-plot etc., Use all plots for individual variables]

Note 🚩 : Kindly get it reviewed the above steps with the @mentors before proceeding further in the project.

10. Interpretations (Must)
- Write the interpretations/ Comments of each stage in the above steps in the jupyter file for better understanding to all.

11. Conclusion
- At last give the interpretations for what you infer about your problem statement.

Note 🚩 : Kindly get it reviewed the above steps with the @mentors for the Final review before you submit in the LMS and Update in the github or Resume

12. Presentation and Project VIVA:

Based on the project completion you have to give the presentations about what you infer from your business problem. We will schedule a presentation once the Project is submitted. Please contact your Mentors for the Presentation Schedule.

👉**Click here to find the PPT Template for the Project Presentation**

Note 🚩 Kindly get it reviewed the above steps with the @mentors before presenting

Submission:

After completion of the project Zip the web scraping code file, csv file and data analysis and data visualization file upload the zip file with your name and batch number.

Note 🚩 :: For the doubts clarifications you are welcome to join our One to One mentorship session