Assignment #04

Hope to Skills

Free Artificial Intelligence Advance Course

Instructor: Irfan Malik, Dr. Sheraz

Submission:

- Make a Google Collab notebook to implement this assignment.
- In case you face difficulty in creating the Google Collab Notebook Follow these **Steps**
- Submit a **.ipynb** file detailing all the information. No other format will be accepted
- Submission file should be named as **Assignment_04_StudentName.ipynb**
- Deadline for this Assignment is **Sunday 17-03-2024**.
- Strictly follow the submission deadline.
- Make Submission in the **Assignment-04** Google Form and press the submit button.
- Click **here** to submit the Assignment

What you will learn

- How to create Google Colab Notebook from scratch.
- Basic understanding of Kaggle and how to search datasets there.
- How to use dataset in your project.
- How to perform EDA on data
- Data Visualization

Solve the Following Tasks

Question 1: Perform the following tasks: (20 Marks)

- Create a pairplot to visualize relationships between multiple numerical variables.
- Generate a heatmap to display the correlation matrix of these variables.
- Identify and print pairs of variables with the highest and lowest correlation coefficients

Dataset: Dataset is provided or you get from Titanic Dataset

Question 2: Perform the following task: (20 Marks)

Explore the Netflix dataset to understand the distribution of movies vs. TV shows, identify the countries producing the most content, and analyze the trend of releases over the years. What genres are most common, and how does the rating distribution vary across genres?

Dataset: Netflix Shows

Question 3: Perform the following task: (20 Marks)

Conduct an exploratory data analysis on the Melbourne Housing Market dataset. Identify trends in housing prices over time, analyze the impact of location, and explore the relationship between property attributes (such as the number of rooms, type of property, and land size) and price.

Dataset: Melbourne Housing Market

Question 4: Perform the following tasks: (20 Marks)

- Analyze and describe the data by .head(), .shape , .describe() etc
- Visualize the data with different plots (Histogram, line or bar plot, Scatter plot, Box or violin plots) NOTE : Also write the insights you learn from different plots.
- Understand the data, handle missing values if exists
- Apply strategies can be employed to deal with missing values? (e.g., imputation, removal, etc.) Use different imputation methods for different features.
- Identify outliers in the numerical features and handle them with different methods of outlier removal.
- Convert categorical variables by one-hot encoded or treat with other encoding methods?
- Do any numerical features require transformation (e.g., min-max scaling, standard scaling, log transformation) to achieve a more normal distribution?
- Did scaling or normalization improve model performance?
- Are there any inconsistencies or errors in the data (e.g., negative fare amounts, unrealistic values)?
- Address these inconsistencies to ensure data quality?
- Which features are most important for predicting the target variable?
- Can feature selection techniques (e.g., feature importance, recursive feature elimination) help identify the most relevant features?
- What are the relationships between different numerical features? Identify
- relationship between the variable through heatmap or other correlation methods.

Dataset: Use **sns.load_dataset(“taxis”)** to load the dataset.

Question 5: Perform Data Cleaning and Exploration using Pandas: (20 Marks)

- Load the famous Titanic dataset
- Check the first few rows and data types.
- Handle any missing values in the dataset, if present.
- Check for and handle any outliers in the data. (if there)
- Use matplotlib and seaborn to visualize the data distributions
- Create a new column indicating whether a passenger is a child, adult, or elderly based on their age.
- Calculate the average fare paid by passengers in each class.

Dataset: Dataset is provided or you get from Titanic Dataset

Notes for Students:

- Each question is of **20 marks**.
- Please use proper comments and formatting in your code. Bad formatting or no comments will result in marks deduction.
- Any other file format except “.ipynb” will not be accepted and will result in 0 marks.
- The submissions after the deadline will not be considered.