Case Study: Exploratory Data Analysis on Used Cars Dataset

Objective:

The goal of this case study is to conduct a comprehensive **Exploratory Data Analysis (EDA)** on a dataset containing listings of used cars. The objective is to identify key factors influencing the **price of used cars**, discover trends in car features like mileage, power, fuel type, and ownership history, and assess the dataset's quality for potential predictive modeling tasks.

Dataset Description:

The dataset includes real-world data from used car listings across various locations. Each record represents a car and contains features related to the car's specifications, condition, and price.

Attributes:

- `S.No.`: Serial Number
- `Name`: Full name of the car including brand and model
- `Location`: City of sale
- `Year`: Year of manufacture
- `Kilometers_Driven`: Total distance driven
- `Fuel_Type`: Type of fuel used (e.g., Petrol, Diesel, CNG)
- `Transmission`: Gearbox type (Manual/Automatic)
- `Owner_Type`: Ownership history (First, Second, etc.)
- `Mileage`: Fuel efficiency (kmpl or km/kg)
- `Engine`: Engine capacity (cc)
- `Power`: Maximum power (bhp)
- `Seats`: Seating capacity
- `New_Price`: Price when the car was new (may be missing)
- `Price`: Current market price of the used car (target variable)

Tasks:

Data Overview

1. Load the dataset and display the structure (rows, columns, data types).
2. Identify and handle missing values across the dataset.
3. Check for duplicate entries.

Descriptive Analysis

1. Generate summary statistics (mean, median, std. dev., etc.) for numeric features like `Price`, `Mileage`, `Power`, `Engine`, `Kilometers_Driven`.
2. Analyze the distribution of categorical features like `Fuel_Type`, `Transmission`, `Owner_Type`, and `Location`.

Visual Explorations

3. Create histograms and boxplots for numerical features like `Price`, `Engine`, and `Mileage`.
4. Plot bar charts showing the frequency of different `Fuel_Type`, `Transmission`, and `Owner_Type`.
5. Visualize the trend of average car prices over `Year` of manufacture.
6. Compare price distributions across different `Locations` and `Fuel_Types`.

Correlation and Feature Relationships

7. Compute a correlation matrix for numerical columns and visualize it using a heatmap.
8. Use scatter plots to analyze relationships between `Price` and key numerical variables (`Mileage`, `Engine`, `Power`, `Kilometers_Driven`).
9. Analyze multivariate interactions such as how `Transmission` and `Fuel_Type` together affect `Price`.

Outliers and Data Quality

10. Use boxplots to identify outliers in `Price`, `Engine`, and `Power`.
11. Identify features with possible inconsistencies (e.g., non-numeric `Power`, ambiguous `Mileage` units).
12. Analyze the `New_Price` feature and its missing data—what percentage is missing and how might this affect modeling?

Insights and Trends

13. Determine the most and least expensive car models and their characteristics.
14. Compare the average price of cars based on number of previous owners.
15. Evaluate how mileage efficiency correlates with car price and manufacturing year.

Predictive Readiness

16. Assess which features could be strong predictors of the `Price`.
17. Check if `New_Price` could be used to estimate depreciation.
18. Suggest encoding techniques and feature engineering approaches for model training.

Outcome:

- Gain insights into what affects used car prices.
- Identify trends in ownership, fuel types, and manufacturer years.
- Prepare the dataset for predictive modeling, addressing data quality issues and highlighting useful features.

Deliverables:

1. A **Jupyter Notebook** with:
    - Complete EDA code
    - Visualizations
    - Explanatory comments
2. A **Summary Report** including:
    - Key insights from the EDA
    - Notable trends and patterns
    - Suggestions for feature engineering and next steps for predictive modeling