Project Report: Automobile Data Analysis

1. Project Overview

This project involves an end-to-end Exploratory Data Analysis (EDA) of an automobile dataset containing 205 records and 26 attributes. The primary objective was to clean the raw data, perform statistical analysis, and identify the key features that drive vehicle pricing and insurance risk ratings.

2. Tools & Libraries Used

- **Python:** Core programming language.
- **Pandas:** Data manipulation, cleaning, and ingestion.
- **NumPy:** Numerical operations and outlier detection.
- **Matplotlib & Seaborn:** Data visualization and correlation heatmaps.

3. Data Cleaning Strategy (Crucial Step)

Raw data contained missing values denoted by '?' and incorrect data types. The following cleaning steps were performed:

- **Handling Missing Values:** Identified '?' in columns *normalized-losses*, *price*, *horsepower*, *bore*, and *stroke*. Used **Mean Imputation** to replace these with the column average.
- **Data Type Conversion:** Converted object types to integers/floats after cleaning to enable statistical analysis.
- **Outlier Removal:** Detected anomalies in *horsepower* (values > 10,000) and filtered data to keep only values within 3 standard deviations of the mean.
- **Dropping Rows:** Removed rows with missing *num-of-doors* data as the impact was negligible.

4. Key Analysis & Business Insights

A. Market Distribution (Univariate Analysis)

- **Dominant Make:** Toyota is the market leader in this dataset, with >40% more vehicles than the runner-up (Nissan).
- **Fuel Preference:** Gas (Standard) engines comprise >80% of the dataset vs. Diesel/Turbo.
- **Drivetrain:** Front-Wheel Drive (FWD) is the most common configuration.

B. Price Drivers (Bivariate & Correlation Analysis)

- **Engine Size vs. Price:** Strong positive correlation. As engine size increases, price increases linearly.

- **Curb Weight Impact:** Heavier cars (higher curb weight) have lower MPG (City/Highway) and higher prices.
- **Manufacturer Pricing:** Mercedes-Benz, BMW, and Jaguar occupy the premium pricing tier (>20k), while Chevrolet and Dodge dominate the budget tier (<10k).

5. Conclusion

The analysis confirms that physical attributes (**Curb Weight, Engine Size**) are the strongest predictors of a car's price. Furthermore, there is a clear relationship between the insurance risk rating (**Symboling**) and financial losses (**Normalized Losses**). Negative symboling ratings (safer cars) correlate with lower financial losses.

6. Future Scope

1. **Machine Learning:** Implement Linear Regression to predict car prices based on Engine Size and Curb Weight.
2. **Feature Engineering:** Categorize cars into 'Luxury', 'Mid-Range', and 'Budget' for classification tasks