Exploratory Data Analysis Report

Submitted by: Pragya Gupta
PGDM – Research & Business Analytics

Abstract

This project explores a car dataset to perform Exploratory Data Analysis (EDA). The goal is to clean, preprocess, and analyze the dataset to uncover insights into variables affecting car features, pricing, and market patterns. The analysis emphasizes data quality improvements and visual explorations to better understand relationships among numerical and categorical features.

1. Introduction

The automotive industry generates large amounts of structured and unstructured data. Understanding car-related datasets is critical for decision-making by manufacturers, policymakers, and consumers. This project applies EDA techniques to the car dataset to: • Understand the structure of the dataset. • Clean and preprocess the data. • Generate insights into numerical and categorical features. • Highlight trends and relationships useful for further modeling.

2. Dataset Description

• Source: Car dataset (Excel file) • Size: Rows × Columns (depending on data) • Key Features: o Car specifications (make, model, year) o Numerical variables (mileage, price, horsepower, engine size) o Categorical variables (fuel type, transmission, body type) Missing values were treated, duplicates removed, and new derived variables created for deeper analysis.

3. Methodology

The project followed these steps: • Data Loading & Inspection: Read dataset, checked dimensions, datatypes, and missing values. • Data Cleaning: o Removed duplicates. o Handled missing values with imputation or deletion. o Treated outliers in numerical features. • Feature Engineering: o Normalized and scaled numerical variables using StandardScaler. o Encoded categorical variables into dummy variables. • Exploratory Analysis: o Univariate analysis of numerical and categorical variables. o Bivariate analysis to explore relationships. • Visualization: Applied matplotlib and seaborn to visualize distributions, correlations, and trends.

4. Results & Insights

Key insights from the analysis include: 1. Distribution of numerical variables shows skewness requiring normalization. 2. Outlier detection highlighted extreme values in price and mileage. 3. Encoding categorical variables allowed integration into numerical models. 4. Scaling improved comparability of features with different units. 5. Visualization showed meaningful relationships, e.g., mileage vs. price and body type vs. price.

5. Conclusion

The EDA of the car dataset revealed significant patterns: • Data preprocessing ensures better quality for downstream tasks. • Univariate and bivariate analysis highlight market patterns in car features and pricing. • Visualizations provide actionable insights for understanding consumer preferences. Future Work: • Apply machine learning models for price prediction and clustering. • Extend dataset to include multi-year car sales and regional variations. • Build interactive dashboards for decision-making.