****

**DATA SCIENCE TOOLBOX PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

**DIWALI SALES ANALYSIS**

**Submitted by**

**NAME:** K Hema Sri Sai

**REGISTRATION NO:** 12317398

**PROGRAMME AND SECTION:** K23GN40

**COURSE CODE** **:** CSE375

Under the Guidance of

**MRS. AASHIMA (UID: 28968)**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **K.Hema Sri Sai** bearing Registration no. **12317398** has completed **CSE375** project titled, **“Mrs.Aashima”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, **K.Hema Sri Sai** student of CSE (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No: 12317398 Name of the student: **K.Hema Sri Sai**

|  |  |
| --- | --- |
| **INDEX** | **Pg.No:** |
| **1.Introduction** | **5** |
| **2.Sources of Dataset** | **6** |
| **3.About EDA** | **6** |
| **4.Regarding EDA** | **7** |
| **5.Analysis of Dataset** | **8** |
| **6. Correlation Analysis** | **10** |
| **7. Outlier Detection** | **11** |
| **8. Brand-wise Price Analysis** | **12** |
| **9. Transmission Type Distribution** | **13** |
| **10. Price vs Mileage** | **14** |
| **11. Feature Relationships (Pair Plot)** | **15** |
| **12. Price Trends Over the Years** | **16** |
| **13.Conclusion** | **17** |
| **14. Future Scope** | **17** |
| **15.Reference** | **17** |
| **16.Source Code** | **18** |

**Car Price Analysis: An Exploratory Data Analysis (EDA) Approach**

**1. Introduction :**

In today’s data-driven and competitive world, understanding consumer behavior, market trends, and product performance is crucial for business success. The automotive industry, one of the largest and most dynamic sectors globally, relies heavily on data analytics to optimize operations, understand market demand, and enhance customer satisfaction.

Car pricing is influenced by numerous factors such as brand reputation, vehicle age, mileage, fuel type, transmission system, engine capacity, and even macroeconomic conditions. For manufacturers and dealers, analyzing historical price trends helps with inventory planning, setting competitive prices, and targeting the right customer segments. For buyers, this analysis offers clarity on what to expect when investing in a vehicle, ensuring better financial decisions.

In this project, we perform Exploratory Data Analysis (EDA) on a used car dataset to identify important patterns and relationships among various features. EDA not only helps us clean and understand the data but also prepares it for further predictive modeling or business insights.

The primary goals of this analysis are:

To identify how different factors like brand, year, mileage, engine size, and fuel type affect car prices.

To detect trends and anomalies in the dataset.

To visualize the distribution and relationships among key features.

To provide actionable insights for stakeholders in the automobile domain.

By performing a detailed EDA, we aim to transform raw data into meaningful information, which can later support advanced tasks like price prediction or market segmentation.

This report is structured to guide the reader through each step of the EDA process, starting from data loading and cleaning, followed by univariate and multivariate analysis, and concluding with key insights, future possibilities, and references. Through graphs, charts, and descriptive explanations, we aim to make the data speak for itself.

**2. Source of Dataset :**

The dataset used in this Exploratory Data Analysis project was obtained from a publicly available car price database. It contains detailed information about various pre-owned vehicles and their associated attributes, which are essential for determining car value in the second-hand market.

The dataset includes the following key features:

* **Brand:** The manufacturer of the vehicle (e.g., Toyota, BMW, Honda), which can significantly impact the price due to brand value and reliability.
* **Year:** The manufacturing or registration year of the vehicle. Newer vehicles generally have a higher price due to less wear and modern features.
* **Transmission Type:** Indicates whether the vehicle uses a manual or automatic transmission, a key factor in buyer preference and pricing.
* **Fuel Type:** Shows whether the car uses petrol, diesel, or alternative fuels, which can affect long-term running costs and hence price.
* **Mileage:** The total distance traveled by the vehicle in kilometers, which directly impacts the car's resale value.
* **Engine Size:** Refers to the volume of the engine in liters or cubic centimeters (cc), affecting both performance and fuel consumption.
* **Price:** The selling price of the vehicle in the market, which is the target variable for this analysis.

**3. What is EDA (Exploratory Data Analysis)?**

Exploratory Data Analysis (EDA) is a fundamental step in the data science workflow that involves examining and summarizing the main characteristics of a dataset—both visually and statistically—before any modeling is done. The primary goal of EDA is to gain an in-depth understanding of the data, identify its structure, detect anomalies or inconsistencies, and uncover underlying patterns or relationships that could influence outcomes in predictive modeling or business decision-making.

EDA acts as the bridge between raw data and actionable insights. Rather than rushing into building machine learning models, EDA encourages a data-first approach where insights are drawn organically by allowing the data to "speak."

Objectives of EDA

The core objectives of Exploratory Data Analysis include:

* Understanding Data Structure: Recognizing the types of variables (categorical, numerical), their distributions, and the relationships among them.
* Detecting Anomalies and Outliers: Identifying values that deviate significantly from the norm, which might skew the analysis or indicate data entry errors.
* Handling Missing Values: Detecting and managing gaps in the dataset to ensure a consistent and reliable analysis.
* Discovering Patterns and Trends: Recognizing how features interact with each other and influence the target variable.
* Preparing for Modeling: Selecting important features, identifying transformations, and evaluating assumptions before applying statistical or machine learning models.

**Importance of EDA in Data Analysis Projects**

**Performing a thorough EDA is crucial because:**

* **Ensures Data Quality:** Raw datasets often contain missing values, duplicated entries, inconsistencies, or outliers. EDA helps detect and correct these issues before further analysis.
* **Informs Decision-Making:** Data visualizations and descriptive statistics can highlight business-critical insights, such as sales trends or customer preferences.
* **Improves Model Performance:** Clean, well-understood data leads to better-performing predictive models and more reliable results.
* **Enhances Interpretability:** Visualizations make complex data relationships easier to understand, especially for non-technical stakeholders.

**How EDA is Achieved in this Project :**

In this car price analysis project, EDA was conducted using Python and its popular data analysis libraries—Pandas, Matplotlib, and Seaborn. The following techniques and steps were performed:

* **Data Loading and Preview:** Used df.head() and df.info() to understand the shape and structure of the data.
* **Handling Missing Values:** Identified with df.isnull().sum() and removed using df.dropna(inplace=True) to maintain data integrity.
* **Descriptive Statistics:** Summary statistics such as mean, median, and standard deviation were used to assess central tendency and dispersion.
* **Correlation Analysis**: A heatmap was generated to examine relationships between numerical variables, highlighting how features like mileage and engine size relate to car price.
* **Outlier Detection:** Boxplots helped visualize and manage extreme values that could skew results.
* **Univariate Analysis:** Plotted individual variable distributions to understand their spread and frequency.
* **Multivariate Analysis:** Used scatter plots, pair plots, and grouped bar charts to study interactions between multiple variables.
* **Visualization Techniques:** Effective visual tools such as pie charts, line graphs, and bar graphs were used to interpret data more intuitively.

**4. Step-by-Step Analysis of Dataset**

**4.1 Dataset Overview**

The first step in any data analysis process is to become familiar with the dataset. This involves loading the data and performing a basic examination of its structure, size, and content. The dataset in this project contains details about used cars, including various features that influence their market price. These features provide a foundation for evaluating car valuation patterns and extracting useful insights.

To begin the analysis, we used basic Pandas functions to:

* Preview the dataset using .head() to check the first few records.
* Get a concise summary of column names, non-null counts, and data types using .info().
* Identify missing values in the dataset using .isnull().sum().

These steps help in understanding the composition and integrity of the dataset before moving on to more advanced analysis.

These commands are part of the Pandas library and are essential for initial data inspection and cleaning.

* The dataset contains multiple columns such as Brand, Year, Transmission, Fuel\_Type, Mileage, Engine\_Size, and Price.
* Each column was inspected for missing values and data types.
* A few columns contained missing values, which were detected using df.isnull().sum(). To avoid analysis errors or skewed results, rows with missing values were removed using df.dropna(inplace=True).
* All data types were identified correctly: for instance, Price, Mileage, and Engine\_Size are numerical; Brand, Fuel\_Type, and Transmission are categorical.

This foundational step ensures that the dataset is clean and ready for further analysis.

You can insert the following visuals here:

* Table 1: First 5 rows of the dataset (df.head())
* Table 2: Dataset summary from .info() showing column data types and non-null counts
* Table 3: Missing values summary (df.isnull().sum())
* Screenshot or table of df.head() output
* Screenshot of df.info() output
* Screenshot or formatted table of missing value counts

**4.2 Correlation Analysis**

**ii. General Description:**  
Heatmap to visualize correlation coefficients.

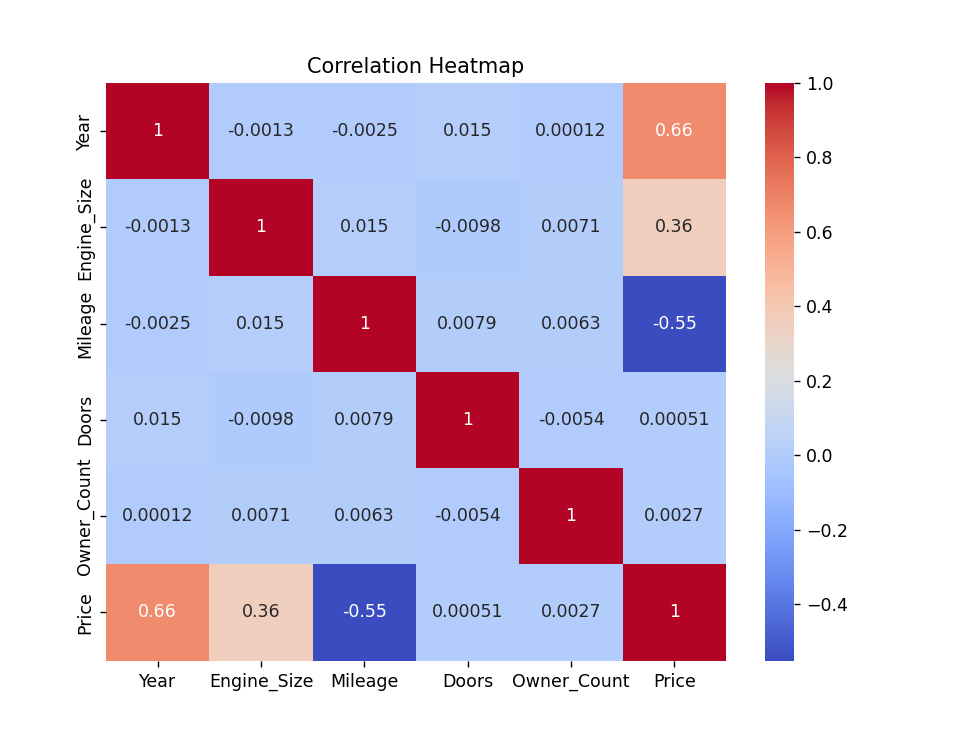
**iii. Function: sns.heatmap(df.corr(numeric\_only=True), annot=True)**

**iv. Results:**

* Strong correlation between engine size and price.
* Negative correlation between mileage and price.

v. Visualization:

***Insert Correlation Heatmap here***



**4.3 Outlier Detection**

**i. Introduction:**  
Identify extreme values affecting distribution.

**ii. Function: sns.boxplot()**

**iv. Results:**

* Some outliers found in price and mileage.

**v. Visualization:**  
 Insert Boxplot of Price, Mileage, and Engine Size here

**A diagram of a box plot

AI-generated content may be incorrect.**

**4.4 Brand-wise Price Analysis**

**i. Introduction:**  
Which brands are priced higher on average?

**ii. Function: df.groupby('Brand')['Price'].mean()**

**iv. Results:**

* Premium brands have higher average prices.

**v. Visualization:**  
+

**A graph of blue bars

AI-generated content may be incorrect.**

**4.5 Transmission Type Distribution**

**i. Introduction:**  
Check popularity of transmission types.

**ii. Function: df['Transmission'].value\_counts().plot(kind='pie')**

**iv. Results:**

* Majority of cars are manual.

**v. Visualization:**  
Insert Pie Chart: Transmission Types here

**A pie chart with numbers and text

AI-generated content may be incorrect.**

**4.6 Price vs Mileage**

**i. Introduction:**  
Analyze price trend with respect to mileage.

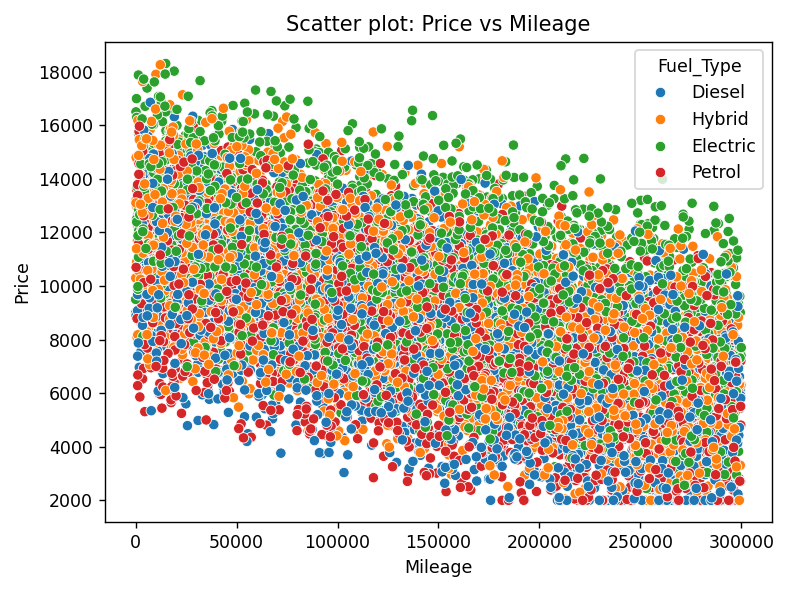
**ii. Function: sns.scatterplot()**

**iv. Results:**

* As mileage increases, price tends to decrease.

**v. Visualization:**

Insert Scatter Plot: Price vs Mileage here

****

**4.7 Feature Relationships (Pair Plot)**

**i. Introduction:**  
Visual inspection of multiple numeric feature combinations.

**ii. Function: sns.pairplot()**

**iv. Results:**

* Linear patterns between price, year, and engine size.

**v. Visualization:**  
Insert Pair Plot here

**A group of blue and black graphics

AI-generated content may be incorrect.**

**4.8 Price Trends Over the Years**

**i. Introduction:**  
How average car prices changed over time.

**ii. Function:**

python

CopyEdit

avg\_price\_by\_year = df.groupby('Year')['Price'].mean()

plt.plot(...)

**iv. Results:**

* Prices fluctuate over years, older cars are cheaper.

**v. Visualization:**  
Insert Line Chart: Average Car Price Over the Years here

**A graph showing a growing trend

AI-generated content may be incorrect.**

**5. Conclusion**

This EDA project helped in uncovering key insights about how different features impact car prices:

* Engine size and year are positively correlated with price.
* Mileage negatively affects price.
* Transmission and brand also play a significant role.

**6. Future Scope**

* Implement regression models to predict car prices.
* Use feature engineering to improve model accuracy.
* Create interactive dashboards using tools like Power BI or Streamlit.
* Expand dataset with real-time data sources (web scraping APIs).

**7. References**

* Seaborn Documentation: https://seaborn.pydata.org/
* Pandas Documentation: https://pandas.pydata.org/
* Dataset Source: Mention if from Kaggle or custom
* Python Matplotlib Docs: <https://matplotlib.org/>
* Exploratory Data Analysis - Wikipedia

**8. Source Code:**

# 1. Import Libraries  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
# 2. Load Dataset  
df = pd.read\_csv("C:\\Users\\Surendra Reddy\\Downloads\\car\_price\_dataset.csv")  
  
# 3. Preview and Basic Info  
print("📄 First 5 rows:")  
print(df.head())  
  
print("\nℹ️ Dataset Info:")  
print(df.info())  
  
print("\n❓ Missing Values:")  
print(df.isnull().sum())  
  
# Drop missing values  
df.dropna(inplace=True)  
  
# 4. Correlation Heatmap  
plt.figure(figsize=(8, 6))  
sns.heatmap(df.corr(numeric\_only=True), annot=True, cmap='coolwarm')  
plt.title("Correlation Heatmap")  
plt.show()  
  
# 5. Boxplot for Outliers  
sns.boxplot(data=df[['Price', 'Mileage', 'Engine\_Size']])  
plt.title("Boxplot of Price, Mileage, and Engine Size")  
plt.show()  
  
# 6. Basic Graphs  
# Bar Graph: Average Price by Brand  
df.groupby('Brand')['Price'].mean().sort\_values(ascending=False).plot(kind='bar', title='Avg Price by Brand', color='skyblue')  
plt.ylabel('Average Price')  
plt.xticks(rotation=45)  
plt.tight\_layout()  
plt.show()  
  
# Pie Chart: Transmission Types  
df['Transmission'].value\_counts().plot(kind='pie', autopct='%1.1f%%', title='Transmission Types')  
plt.ylabel('')  
plt.tight\_layout()  
plt.show()  
  
# Scatter Plot: Price vs Mileage  
  
sns.scatterplot(data=df, x='Mileage', y='Price', hue='Fuel\_Type')  
plt.title('Scatter plot: Price vs Mileage')  
plt.tight\_layout()  
plt.show()  
  
  
  
# Pair Plot  
sns.pairplot(df[['Price', 'Year', 'Mileage', 'Engine\_Size']])  
plt.suptitle("Pair Plot of Key Features", y=1.02)  
plt.show()  
  
  
# Line Chart  
avg\_price\_by\_year = df.groupby('Year')['Price'].mean().sort\_index()  
  
  
plt.figure(figsize=(10, 6))  
plt.plot(avg\_price\_by\_year.index, avg\_price\_by\_year.values, marker='o', linestyle='-', color='green')  
plt.title('Average Car Price Over the Years')  
plt.xlabel('Year')  
plt.ylabel('Average Price')  
plt.grid(True)  
plt.tight\_layout()  
plt.show()