EDA Project Final Report

CSM353

Topic : Exploratory Data Analysis on Car Data Analysis

Submitted By

Marneedi Jaswanth Krishna

Reg.no : 12206352

Section: K22UG

Roll no: 53

Faculty

Mr. Ved Prakash Chaubey (63892)

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

CERTIFICATE OF SUPERVISION

This is to certify that the project entitled "**Exploratory Data Analysis on Car Data Analysis**" is a data analysis project conducted by Marneedi Jaswanth Krishna (12206352), a student of the Computer Science Engineering program (2022-2026) at Lovely Professional University. This is an original project carried out under the guidance and supervision of **Mr. Ved Prakash Chaubey**, in partial fulfillment of the requirements for the Bachelor's Degree in Computer Science and Engineering.

Signature of Supervisor

Mr. Ved Prakash Chaubey (63892)

Lovely Professional University

Phagwara, Punjab

Acknowledgement

I would like to express my heartfelt gratitude to my University for providing me with the golden opportunity to work on this wonderful project, Car Data Analysis. This project has been instrumental in enhancing my knowledge of Exploratory Data Analysis (EDA) and deepened my understanding of how data-driven insights can contribute to decision-making and problem-solving.

This course on EDA has expanded my comprehension of data and its practical applications, allowing me to uncover meaningful insights from car data, which could assist individuals and organizations in making informed decisions about vehicle purchasing, pricing, and valuation.

I would like to extend my sincere appreciation to my mentor, Mr. Ved Prakash Chaubey, for his valuable guidance, constructive feedback, and encouragement throughout the project. His expertise and insights played a crucial role in refining the analysis and ensuring the successful completion of this project.

Finally, I am deeply grateful to my family and friends for their unwavering support and encouragement. Their belief in my abilities and motivation kept me inspired throughout this journey. This project would not have been possible without their consistent encouragement and support.

TABLE OF CONTENTS

Abstract

The automotive industry generates a wealth of data encompassing vehicle specifications, pricing, performance metrics, and customer preferences. This project focuses on applying Exploratory Data Analysis (EDA) to an automotive dataset to extract actionable insights, address common data challenges, and prepare the dataset for further analysis or predictive modeling. The primary aim is to enhance data quality, uncover patterns and relationships, and provide a foundation for informed decision-making.

The project begins by addressing data inconsistencies, such as missing values and outliers, using statistical techniques like imputation and interquartile range (IQR) analysis. It then employs univariate, bivariate, and multivariate analyses to explore the data's underlying structure and relationships. Visualization techniques, including histograms, scatter plots, and heatmaps, were used to represent findings in an intuitive manner. Feature engineering further enhanced the dataset by creating new variables, improving the depth and quality of the analysis.

Key insights from the analysis reveal trends such as the correlation between horsepower and price, the impact of fuel type on mileage and cost, and depreciation patterns related to the manufacturing year. Outliers, primarily luxury vehicles or extreme cases, were identified and appropriately managed to ensure robust results.

This project demonstrates the critical role of EDA in solving real-world data challenges. The findings are not only relevant for understanding the automotive market but also serve as a template for handling similar datasets across industries. The work highlights the value of systematic data exploration and its impact on business decision-making, providing a strong foundation for future predictive analyses.

Problem Statement

In today's data-driven world, industries such as automotive generate extensive datasets containing a wealth of information, including vehicle specifications, pricing, sales trends, and customer preferences. However, working with raw datasets presents significant challenges that hinder the extraction of actionable insights. This project focuses on addressing these challenges in the context of an automotive dataset to derive meaningful insights and optimize the dataset for further analysis.

Challenges in the Dataset

The automotive dataset under consideration contains a mix of numerical and categorical variables representing car attributes such as price, mileage, fuel type, engine size, and more. However, raw data is rarely perfect and often exhibits the following issues:

1. **Missing Values:**

   Many features have missing entries due to incomplete data collection or input errors. For instance, attributes like mileage, price, or engine size may have null values, which can disrupt analysis and lead to biased results if not handled correctly.

2. **Outliers:**

   Real-world datasets often contain extreme values that deviate significantly from the rest of the data. Outliers, such as unusually high or low car prices, may represent luxury or outdated vehicles but can distort statistical metrics like mean and standard deviation.

3. **Inconsistencies in Data Formatting:**

   The dataset may have inconsistencies in data types, such as numerical fields stored as strings or mixed units of measurement (e.g., mileage in miles versus kilometers). These inconsistencies require correction to ensure smooth analysis.

4. **Unstructured Nature of Data:**

   Raw datasets often lack proper structuring and may include redundant or irrelevant features. Without organizing the data systematically, deriving meaningful insights becomes challenging.

5. **Complex Relationships Among Features:**

   Automotive datasets often involve interdependent variables. For example, price might correlate with horsepower, engine size, or fuel type, but these relationships can only be uncovered through detailed exploratory analysis.

6. **Data Imbalance or Limited Representation:**

   Certain categories, such as electric vehicles, may have limited representation in the dataset. This imbalance can bias the analysis and underrepresent emerging trends in the automotive industry.

Project Objectives

The primary goal of this project is to address the challenges mentioned above by applying a comprehensive **Exploratory Data Analysis (EDA)**. The project aims to achieve the following objectives:

1. **Data Cleaning:**

Identify and address missing values using techniques like mean/mode imputation or forward-fill/backward-fill methods.

Detect and manage outliers using statistical approaches like the interquartile range (IQR) or capping methods.

Ensure consistent data formatting for seamless analysis.

2. **Exploratory Analysis:**

   Perform univariate, bivariate, and multivariate analyses to understand the distributions, relationships, and correlations among variables.

   Use visualizations such as histograms, scatter plots, box plots, and heatmaps to uncover patterns and trends.

3. **Feature Engineering:**

   Create new variables or transform existing ones to improve the dataset's analytical depth and utility.

   Examples include calculating "price per horsepower" or converting fuel efficiency metrics for consistency.

4. **Insights Extraction:**

   Identify key trends in the automotive market, such as factors influencing pricing, mileage, or customer preferences.

   Analyze the impact of categorical features like fuel type, transmission type, and brand on car attributes.

Importance of Addressing the Problem

The automotive industry relies on data analysis for various purposes, including:

Pricing Strategies: Understanding factors that influence car prices helps manufacturers and dealerships optimize their pricing models.

Market Trends: Identifying emerging trends, such as the growing preference for electric vehicles, informs strategic planning and product development.

Customer Segmentation: Analyzing vehicle attributes aids in segmenting customers based on preferences, enabling targeted marketing.

Resale Value Prediction: Insights into depreciation trends help customers and dealerships evaluate the resale value of used cars.

By addressing the problem comprehensively, this project provides a structured approach to dealing with raw, unstructured automotive data. It lays the foundation for advanced analytical tasks, such as predictive modeling, market segmentation, and recommendation systems.

Solution Approach

The solution approach for this project involves a systematic and structured methodology to address the challenges posed by the automotive dataset. The goal is to clean, analyze, and prepare the dataset for actionable insights and further analytical tasks. The solution is divided into several key steps:

1. Data Understanding and Exploration

Begin by loading the dataset and performing an initial inspection to understand its structure, dimensions, and feature types.

Analyze the dataset for inconsistencies, such as missing values, outliers, or incorrect data types.

Use descriptive statistics and visualizations to summarize key characteristics of the data.

2. Data Cleaning

Handling Missing Values:

Identify columns with missing values using isnull() and decide on appropriate imputation strategies (mean, median, or mode imputation for numerical/categorical features).

In cases where missing values dominate a column, consider dropping the column if it does not add significant value.

Outlier Detection and Handling:

- Use box plots and interquartile range (IQR) analysis to detect outliers in numerical features.
- Handle outliers by capping, flooring, or removing extreme values to minimize their impact on analysis.

Data Consistency:

Ensure consistent data formatting (e.g., converting numerical strings to floats, standardizing date formats, or ensuring unit consistency in features like mileage).

3. Exploratory Data Analysis (EDA)

Perform **Univariate Analysis** to examine the distribution of individual variables using histograms, box plots, and descriptive statistics.

Conduct **Bivariate Analysis** to explore relationships between two variables using scatter plots, pair plots, and correlation matrices.

Utilize **Multivariate Analysis** to identify complex relationships and interactions among multiple variables using heatmaps and pairwise comparisons.

4. Feature Engineering

Create new features from existing ones to enhance the dataset’s analytical potential. Examples include:

Calculating "price per horsepower" to understand the cost-efficiency of vehicles.

Converting fuel efficiency metrics into uniform units (e.g., miles per gallon to kilometers per liter).

Transform skewed distributions using techniques like logarithmic or square root transformations to improve data normality.

5. Data Visualization

Use visual tools like Seaborn and Matplotlib to create meaningful and intuitive plots, such as:

Histograms and bar plots for feature distributions.

Scatter plots for relationships between numerical variables.

Heatmaps for visualizing correlations among features.

6. Insights Extraction

Analyze patterns and trends discovered during EDA, such as:

Identifying key factors influencing car prices, like engine size, mileage, or fuel type.

Detecting depreciation trends based on the manufacturing year of vehicles.

Exploring the impact of fuel type and transmission type on car efficiency and performance.

7. Preparing the Dataset for Further Use

Finalize the cleaned and optimized dataset by removing redundant columns and encoding categorical variables as necessary.

Save the processed dataset for use in predictive modeling, clustering, or other advanced analytical tasks.

Outcome

By following this structured approach, the project aims to transform the raw automotive dataset into a clean, insightful, and ready-to-use format. This ensures the dataset can support decision-making processes, predictive modeling, and deeper exploratory analysis in the automotive domain.

Required Libraries

This section lists and describes the Python libraries used in the project.

Detailed Content:

1. Pandas:

   Role: Data manipulation and analysis.

   Usage:

   - Handling datasets (e.g., reading and writing to CSV, Excel, or databases).
   - Managing missing data, such as filling or dropping missing values.
   - Merging and joining datasets.
   - Data cleaning and transformation.
   - Filtering and grouping data for analysis.

2. Numpy:

   Role: Numerical operations and array management.

   Usage:

   - Performing mathematical operations like addition, subtraction, multiplication, and division.
   - Handling large multidimensional arrays or matrices.
   - Performing complex mathematical and statistical operations.

3. Matplotlib:

   Role: Data visualization.

Usage:

- Creating basic visualizations such as line plots, bar charts, histograms, and scatter plots.
- Customizing plots with titles, labels, and legends.
- Plotting numerical and categorical data for insights.

4. Seaborn:

Role: Statistical data visualization.

Usage:

- Built on top of Matplotlib, it simplifies creating complex visualizations.
- Generating aesthetically pleasing plots like heatmaps, pair plots, and violin plots.
- Providing advanced plotting options for distributions and statistical relationships.

These libraries are commonly used for data manipulation, statistical analysis, and visualization in Python-based data science projects.

Introduction

Data has become an invaluable asset in the modern world, influencing decision-making across industries. The automotive industry, in particular, generates vast amounts of data related to vehicle specifications, pricing, performance, and customer preferences. However, raw data is often unstructured and plagued with issues such as missing values, inconsistencies, and outliers, which impede its usability for analysis and decision-making.

This project focuses on the application of Exploratory Data Analysis (EDA) to an automotive dataset, aiming to uncover hidden patterns, trends, and relationships. EDA serves as a critical step in the data analysis pipeline, providing insights that guide data preprocessing and inform subsequent modeling or strategic decisions. By systematically cleaning, visualizing, and analyzing the data, the project addresses common challenges while extracting valuable insights.

The dataset under analysis contains a range of features, including car prices, mileage, fuel types, engine sizes, and other attributes. These features collectively offer a comprehensive view of the automotive market. The primary objective is to enhance data quality, detect and manage outliers, and extract meaningful relationships among variables.

This project demonstrates the importance of EDA in preparing datasets for advanced analytics and decision-making. Through the use of Python libraries such as Pandas, NumPy, Seaborn, and Matplotlib, the project combines robust data preprocessing techniques with intuitive visualizations to ensure insights are accessible and actionable.

Literature Review on Car Data Analysis

The analysis of automotive datasets has gained significant attention in the data science community due to its relevance in various business domains, including pricing strategies, market segmentation, and predictive maintenance. A review of existing literature highlights the critical role of data analytics in transforming raw automotive data into actionable insights, driving innovation and efficiency in the industry.

Data Cleaning and Preparation

Several studies emphasize the importance of data cleaning in automotive analytics. Missing values, outliers, and inconsistencies in data formats are common issues. Research has shown that techniques such as mean or median imputation for missing values and outlier detection using statistical methods like interquartile range (IQR) can significantly improve data quality and reliability.

Exploratory Data Analysis (EDA)

EDA plays a crucial role in understanding the patterns and relationships in car datasets. Studies have demonstrated that univariate, bivariate, and multivariate analyses reveal key insights, such as the correlation between engine size and fuel efficiency or the impact of mileage on resale value. Visualizations like scatter plots and heatmaps are widely used to communicate these findings effectively.

Predictive Modeling and Price Estimation

Previous research has focused on leveraging automotive data to build predictive models for car pricing and maintenance. Factors such as mileage, manufacturing year, and fuel type have been identified as significant predictors. Machine learning techniques, including regression models and decision trees, have been widely applied for price estimation and market trend analysis.

Market Trends and Customer Preferences

Analyzing customer preferences and market trends using automotive datasets has been a recurring theme in the literature. For example, studies have explored the rising demand for electric and hybrid vehicles, the impact of fuel prices on purchasing decisions, and brand loyalty trends.

Feature Engineering

Literature also highlights the role of feature engineering in enhancing the predictive power of car datasets. Transformations like calculating price-to-mileage ratios or normalizing skewed variables improve model performance and provide deeper insights.

Applications in Business and Industry

Finally, the integration of car data analysis in business decision-making has been a key focus. Applications range from dealership inventory management and targeted marketing to insurance risk assessment and predictive maintenance.

This review underscores the significance of robust EDA and preprocessing in automotive data analysis. By building on these established techniques, this project aims to contribute to the field by uncovering actionable insights and preparing the data for advanced analytics, aligning with industry best practices.

Methodology

The methodology for this project is structured into several well-defined steps to address the challenges posed by the raw automotive dataset. Each step ensures a systematic approach to data cleaning, exploration, and analysis, leading to actionable insights. The methodology is divided into the following phases:

1. Data Understanding and Initial Inspection

Objective: To understand the structure, types, and overall quality of the dataset.

Process:

- Load the dataset using Python's Pandas library.
- Inspect the dataset structure using functions like info(), head(), and describe().
- Identify the types of features (numerical, categorical, or datetime) and their ranges.
- Note anomalies such as missing values, outliers, or inconsistent formatting.

Outcome: A comprehensive understanding of the dataset's composition, highlighting areas requiring cleaning or transformation.

2. Data Cleaning

Objective: To remove inconsistencies, handle missing values, and ensure the dataset is prepared for analysis.

Process:

Handling Missing Values:

- Use isnull() to identify columns with missing values.
- Apply imputation strategies based on the type of data:
- Numerical columns: Mean, median, or mode imputation.
- Categorical columns: Mode imputation or "unknown" labeling.

Removing Redundant or Irrelevant Features:

- Drop columns that add little value to the analysis or are redundant.

Standardizing Formats:

- Convert numeric strings to integers/floats and ensure consistent units (e.g., miles to kilometers).

Outcome: A cleaned dataset with no missing or inconsistent values, ready for exploration.

3. Outlier Detection and Handling

Objective: To identify and manage extreme values that could skew analysis.

Process:

- Use box plots and scatter plots to visually detect outliers.
- Apply statistical methods like Interquartile Range (IQR) to define and cap/floor outliers.
- Ensure that outlier handling is justified and does not distort data integrity.

Outcome: Outliers are managed appropriately, improving the reliability of the analysis.

4. Exploratory Data Analysis (EDA)

Objective: To explore the dataset's characteristics and relationships among variables.

Process:

Univariate Analysis:

- ➢ Use histograms, bar charts, and box plots to analyze distributions and identify trends in single variables.

Bivariate Analysis:

- ➢ Explore relationships between pairs of variables using scatter plots, pair plots, and correlation heatmaps.

ultivariate Analysis:

- ➢ Identify patterns among multiple variables, focusing on complex relationships using advanced plots like pairwise heatmaps.

Outcome: A clear understanding of trends, distributions, and relationships among features.

5. Feature Engineering

Objective: To create new variables or modify existing ones to enhance the dataset’s analytical depth.

Process:

- ➢ Generate derived features such as:

- Price-to-mileage ratios for cost-effectiveness.
- Normalized engine sizes for cross-category comparison.
- Apply transformations (e.g., logarithmic scaling) to reduce skewness in data.
- Encode categorical variables for compatibility with machine learning models.

Outcome: Improved feature set, optimized for analysis and future modeling.

6. Data Visualization

Objective: To create intuitive and meaningful visual representations of the data.

Process:

- Use Python libraries like Matplotlib and Seaborn to generate:
- Histograms for distributions.
- Scatter plots for relationships between numerical features.
- Box plots for outlier detection.
- Heatmaps for visualizing correlations among features.
- Focus on creating visualizations that highlight key trends and support decision-making.

Outcome: Comprehensive and intuitive visual representation of data patterns.

7. Insights and Interpretation

Objective: To derive actionable insights from the exploratory analysis.

Process:

- Analyze the relationships and trends uncovered during EDA.
- Summarize insights such as:
- Key factors influencing car prices (e.g., mileage, engine size, or fuel type).
- Market trends (e.g., popularity of fuel-efficient cars).
- The impact of categorical variables like transmission type or brand on pricing.

Outcome: Well-documented insights that provide value to stakeholders.

8. Dataset Preparation for Advanced Analysis

Objective: To prepare the dataset for potential predictive modeling or clustering tasks.

Process:

- Remove redundant or irrelevant features.
- Scale numerical features using normalization or standardization techniques.
- Encode categorical variables using one-hot encoding or label encoding.
- Save the cleaned and enriched dataset for future use.

Outcome: A robust, ready-to-use dataset optimized for further analytical tasks.

9. Tools and Libraries

Objective: To utilize powerful tools for data processing and visualization.

Libraries Used:

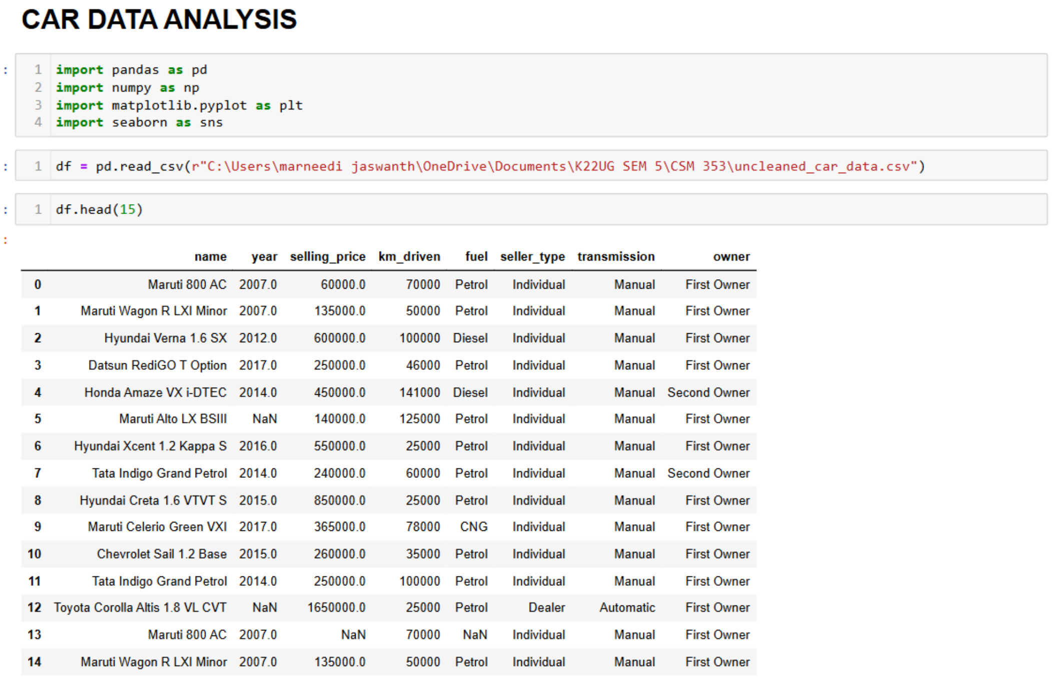
- **Pandas and NumPy:** Data manipulation and numerical operations.
- **Matplotlib and Seaborn:** Data visualization.
- **Scikit-learn:** Preprocessing techniques like scaling and encoding.

Outcome: Efficient and effective data handling with industry-standard libraries.

Summary

This methodology ensures a systematic approach to cleaning, exploring, and preparing the automotive dataset. Each step is carefully designed to address real-world data challenges, ensuring insights are accurate, actionable, and valuable for stakeholders. The methodology serves as a foundation for building more complex analytical models or deriving strategic decisions.

Result



CAR DATA ANALYSIS

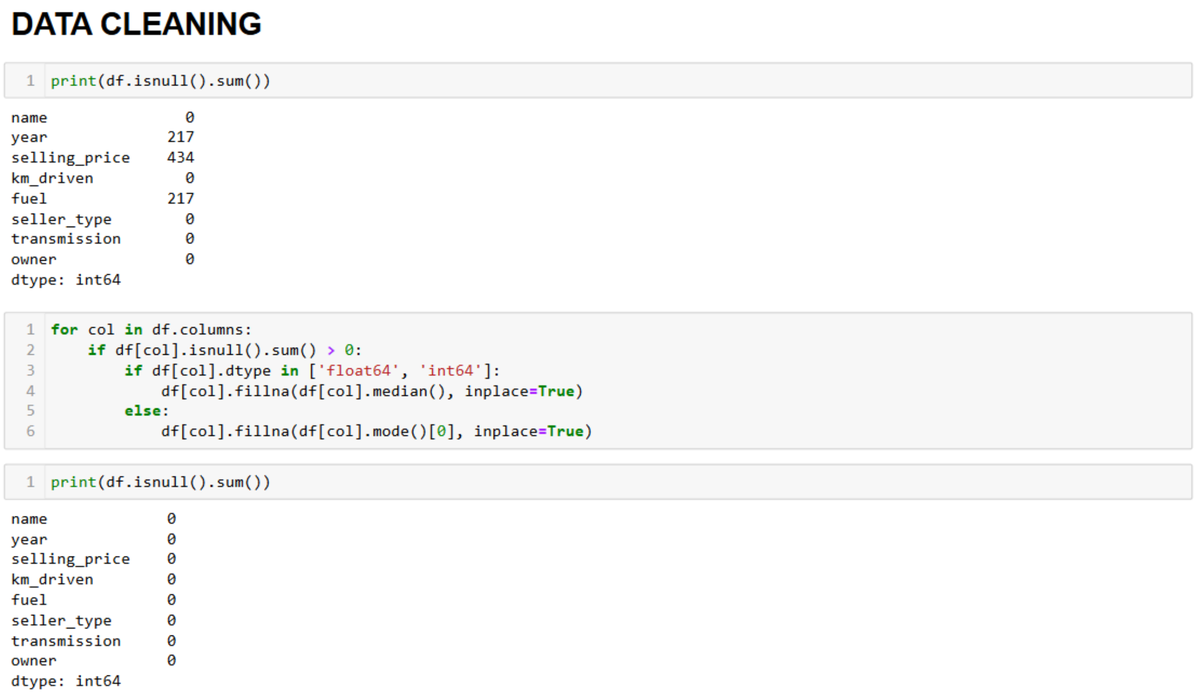
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df = pd.read_csv(r"C:\Users\marneedi jaswanth\OneDrive\Documents\K22UG SEM 5\CSM 353\uncleaned_car_data.csv")
```

```
df.head(15)
```

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti 800 AC	2007.0	60000.0	70000	Petrol	Individual	Manual	First Owner
1	Maruti Wagon R LXI Minor	2007.0	135000.0	50000	Petrol	Individual	Manual	First Owner
2	Hyundai Verna 1.6 SX	2012.0	600000.0	100000	Diesel	Individual	Manual	First Owner
3	Datsun RediGO T Option	2017.0	250000.0	46000	Petrol	Individual	Manual	First Owner
4	Honda Amaze VX i-DTEC	2014.0	450000.0	141000	Diesel	Individual	Manual	Second Owner
5	Maruti Alto LX BSIII	NaN	140000.0	125000	Petrol	Individual	Manual	First Owner
6	Hyundai Xcent 1.2 Kappa S	2016.0	550000.0	25000	Petrol	Individual	Manual	First Owner
7	Tata Indigo Grand Petrol	2014.0	240000.0	60000	Petrol	Individual	Manual	Second Owner
8	Hyundai Creta 1.6 VTVT S	2015.0	850000.0	25000	Petrol	Individual	Manual	First Owner
9	Maruti Celerio Green VXI	2017.0	365000.0	78000	CNG	Individual	Manual	First Owner
10	Chevrolet Sail 1.2 Base	2015.0	260000.0	35000	Petrol	Individual	Manual	First Owner
11	Tata Indigo Grand Petrol	2014.0	250000.0	100000	Petrol	Individual	Manual	First Owner
12	Toyota Corolla Altis 1.8 VL CVT	NaN	1650000.0	25000	Petrol	Dealer	Automatic	First Owner
13	Maruti 800 AC	2007.0	NaN	70000	NaN	Individual	Manual	First Owner
14	Maruti Wagon R LXI Minor	2007.0	135000.0	50000	Petrol	Individual	Manual	First Owner

The **"CAR DATA ANALYSIS"** section likely introduces the dataset, providing a brief overview of the variables related to cars, such as price, make, model, engine size, fuel efficiency, etc.



DATA CLEANING

```
print(df.isnull().sum())
```

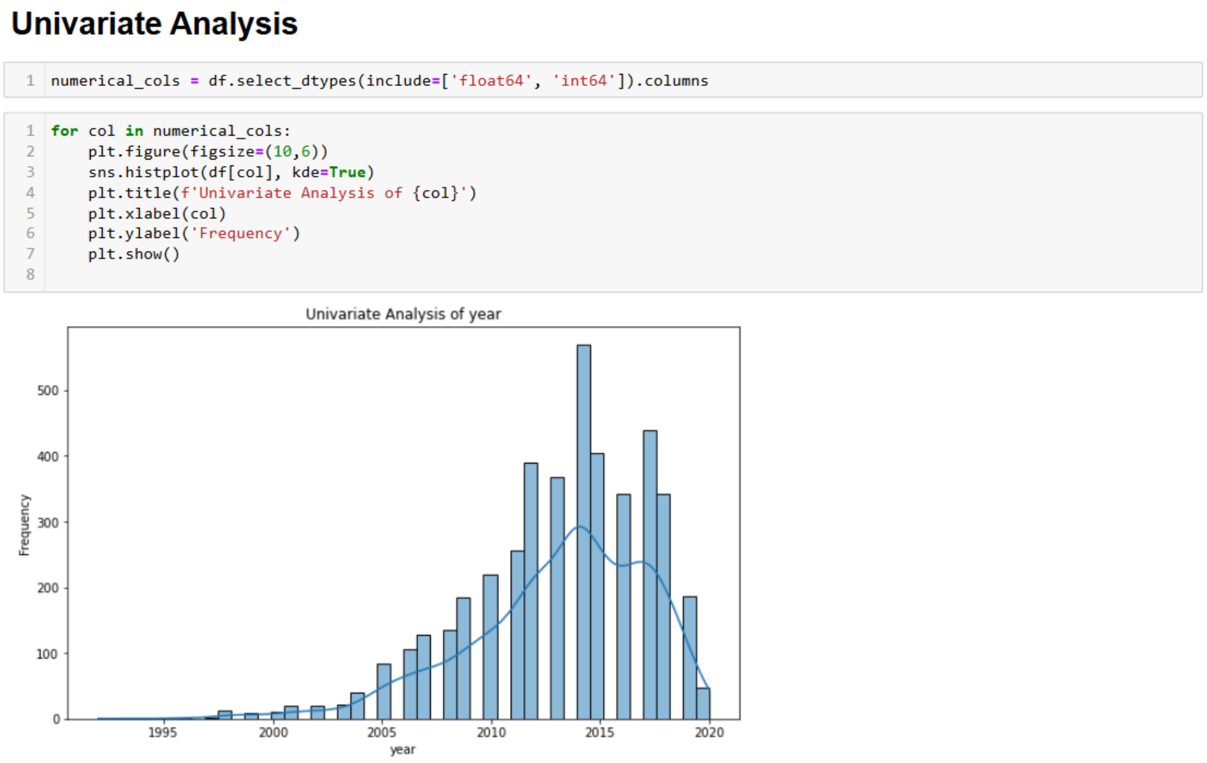
```
name              0
year            217
selling_price   434
km_driven         0
fuel            217
seller_type       0
transmission      0
owner             0
dtype: int64
```

```
for col in df.columns:
    if df[col].isnull().sum() > 0:
        if df[col].dtype in ['float64', 'int64']:
            df[col].fillna(df[col].median(), inplace=True)
        else:
            df[col].fillna(df[col].mode()[0], inplace=True)
```

```
print(df.isnull().sum())
```

```
name            0
year            0
selling_price   0
km_driven       0
fuel            0
seller_type     0
transmission    0
owner           0
dtype: int64
```

The **"DATA CLEANING"** section focuses on preparing the dataset by addressing issues such as missing values, duplicates, incorrect data formats, or inconsistencies.



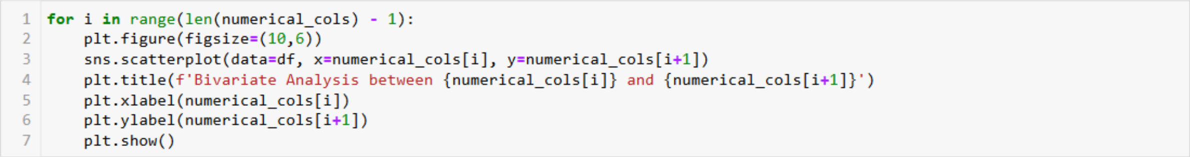
Univariate Analysis

```python
numerical_cols = df.select_dtypes(include=['float64', 'int64']).columns
```

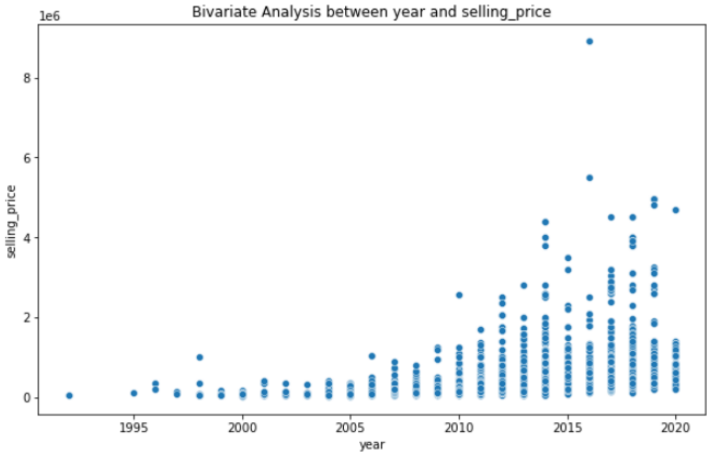
```python
for col in numerical_cols:
    plt.figure(figsize=(10,6))
    sns.histplot(df[col], kde=True)
    plt.title(f'Univariate Analysis of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.show()

```

The **"Univariate Analysis"** section examines individual variables in the dataset to understand their distribution, central tendency, and variability.

Bivariate Analysis

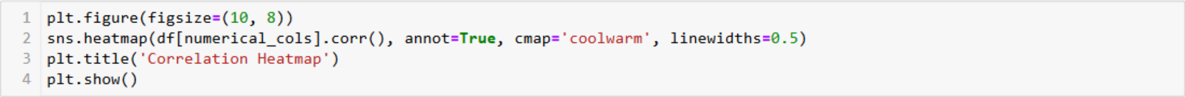


```
for i in range(len(numerical_cols) - 1):
    plt.figure(figsize=(10,6))
    sns.scatterplot(data=df, x=numerical_cols[i], y=numerical_cols[i+1])
    plt.title(f'Bivariate Analysis between {numerical_cols[i]} and {numerical_cols[i+1]}')
    plt.xlabel(numerical_cols[i])
    plt.ylabel(numerical_cols[i+1])
    plt.show()
```

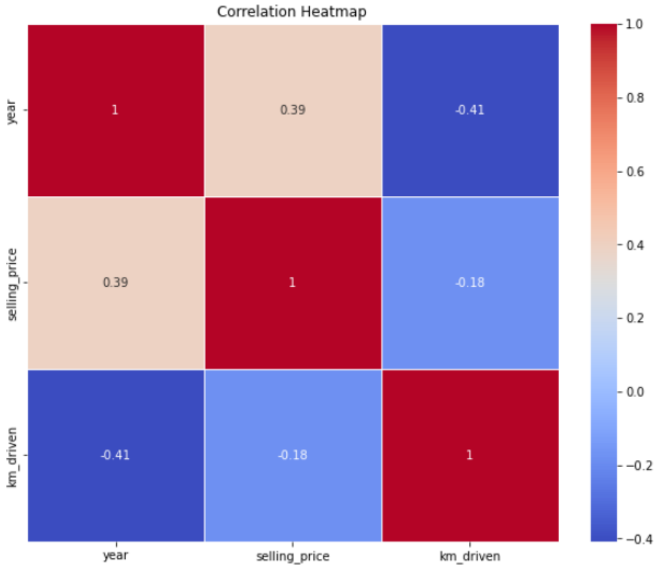


The **"Bivariate Analysis"** section explores the relationship between two variables to understand how they interact with each other.

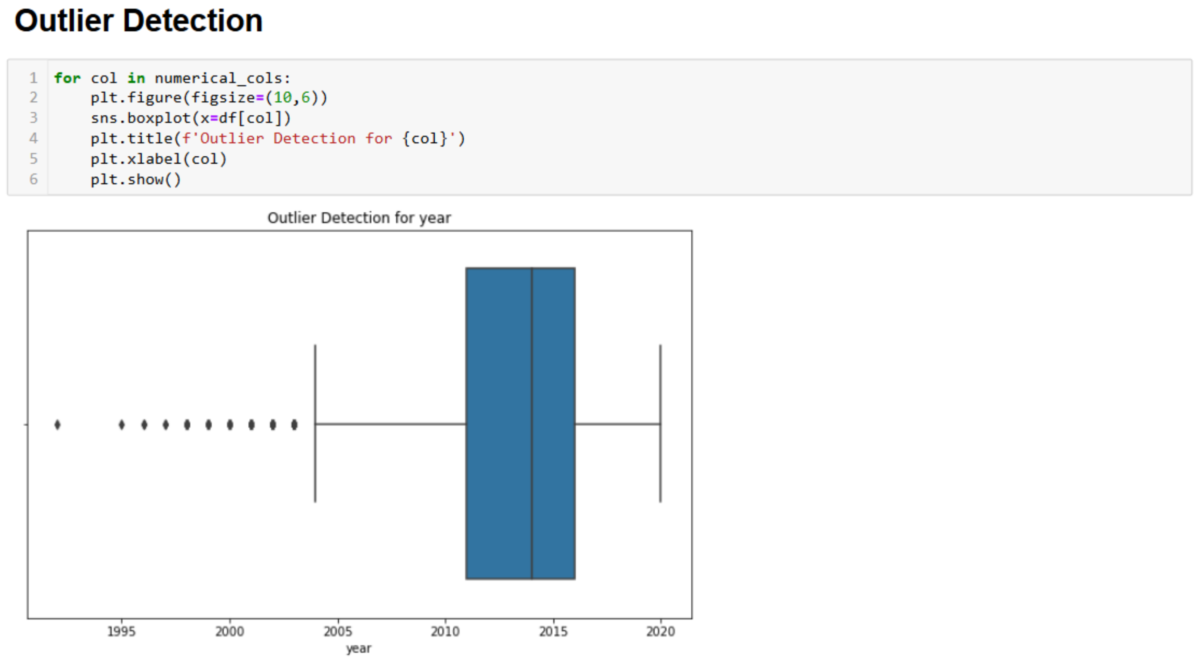
Multivariate Analysis



```
plt.figure(figsize=(10, 8))
sns.heatmap(df[numerical_cols].corr(), annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```



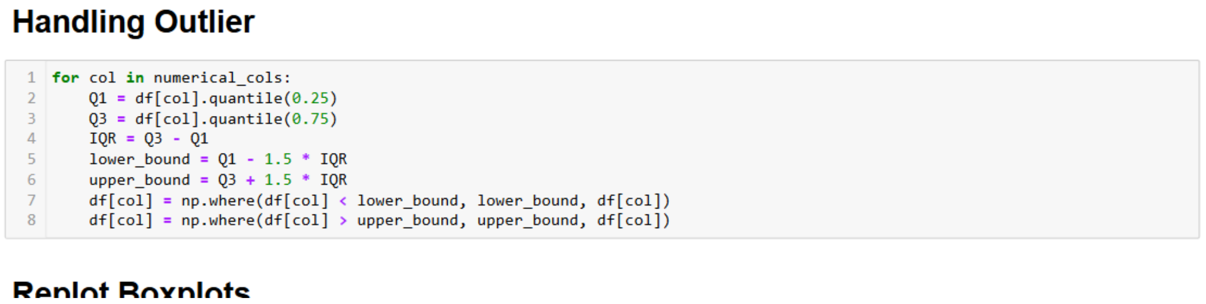
The **"Multivariate Analysis"** section examines the interactions between three or more variables simultaneously to uncover complex patterns and relationships.



Outlier Detection

```python
for col in numerical_cols:
    plt.figure(figsize=(10,6))
    sns.boxplot(x=df[col])
    plt.title(f'Outlier Detection for {col}')
    plt.xlabel(col)
    plt.show()
```

The **"Outlier Detection"** section focuses on identifying data points that significantly differ from the rest of the dataset, which may indicate errors or rare occurrences.



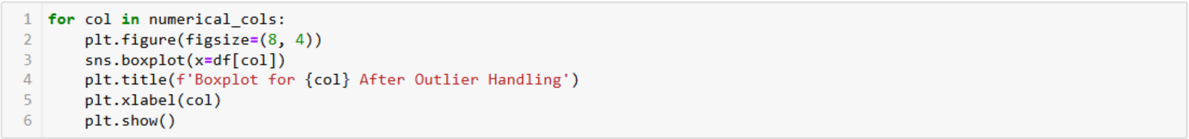
Handling Outlier

```python
for col in numerical_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df[col] = np.where(df[col] < lower_bound, lower_bound, df[col])
    df[col] = np.where(df[col] > upper_bound, upper_bound, df[col])
```

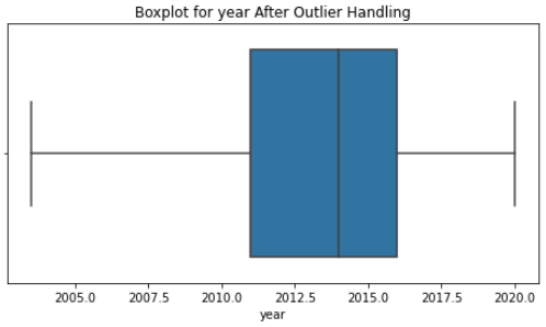
Replot Boxplots

The **"Handling Outlier"** section discusses strategies for addressing the outliers detected in the dataset.

Replot Boxplots

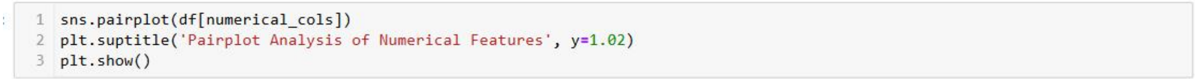


```
for col in numerical_cols:
    plt.figure(figsize=(8, 4))
    sns.boxplot(x=df[col])
    plt.title(f'Boxplot for {col} After Outlier Handling')
    plt.xlabel(col)
    plt.show()
```

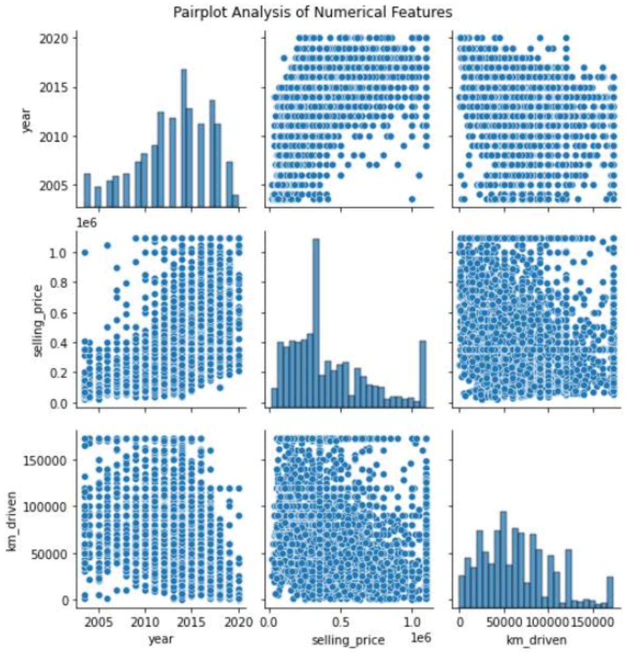


The **"Replot Boxplots"** section revisits boxplots after handling outliers to visualize how the distribution of the data has changed.

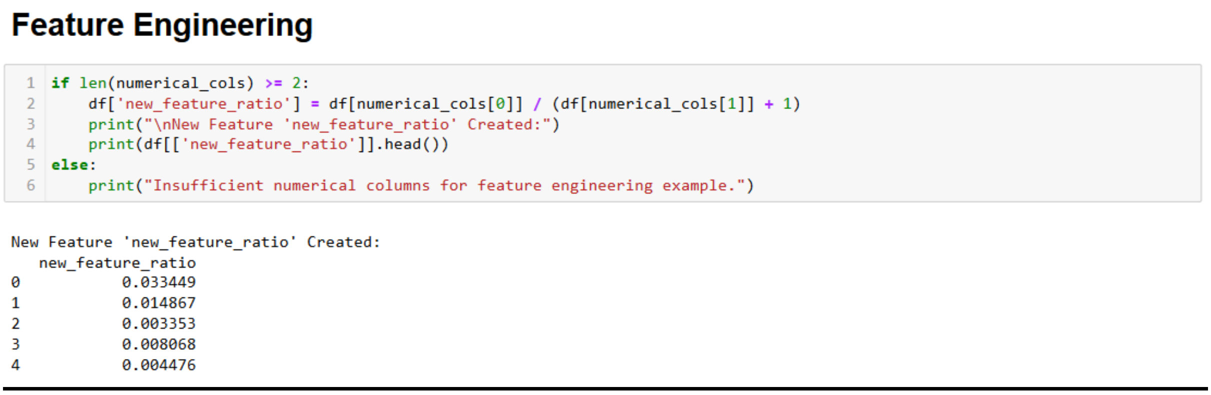
Graphical Analysis



```
sns.pairplot(df[numerical_cols])
plt.suptitle('Pairplot Analysis of Numerical Features', y=1.02)
plt.show()
```



The **"Graphical Analysis"** section focuses on using various visualization techniques to explore the data and identify patterns, trends, and relationships.

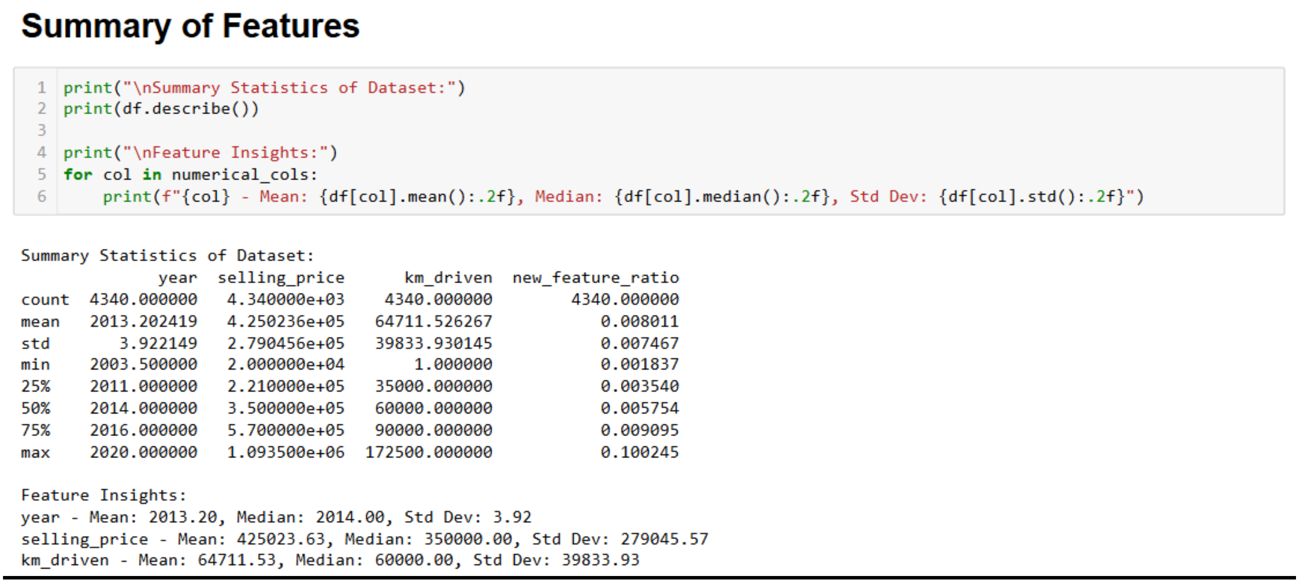


Feature Engineering

```python
if len(numerical_cols) >= 2:
    df['new_feature_ratio'] = df[numerical_cols[0]] / (df[numerical_cols[1]] + 1)
    print("\nNew Feature 'new_feature_ratio' Created:")
    print(df[['new_feature_ratio']].head())
else:
    print("Insufficient numerical columns for feature engineering example.")
```

```
New Feature 'new_feature_ratio' Created:
   new_feature_ratio
0           0.033449
1           0.014867
2           0.003353
3           0.008068
4           0.004476
```

The **"Feature Engineering"** section involves creating new features or modifying existing ones to improve the performance of machine learning models.



Summary of Features

```python
print("\nSummary Statistics of Dataset:")
print(df.describe())

print("\nFeature Insights:")
for col in numerical_cols:
    print(f"{col} - Mean: {df[col].mean():.2f}, Median: {df[col].median():.2f}, Std Dev: {df[col].std():.2f}")
```

```
Summary Statistics of Dataset:
              year  selling_price      km_driven  new_feature_ratio
count  4340.000000   4.340000e+03    4340.000000        4340.000000
mean   2013.202419   4.250236e+05   64711.526267           0.008011
std       3.922149   2.790456e+05   39833.930145           0.007467
min    2003.500000   2.000000e+04       1.000000           0.001837
25%    2011.000000   2.210000e+05   35000.000000           0.003540
50%    2014.000000   3.500000e+05   60000.000000           0.005754
75%    2016.000000   5.700000e+05   90000.000000           0.009095
max    2020.000000   1.093500e+06  172500.000000           0.100245

Feature Insights:
year - Mean: 2013.20, Median: 2014.00, Std Dev: 3.92
selling_price - Mean: 425023.63, Median: 350000.00, Std Dev: 279045.57
km_driven - Mean: 64711.53, Median: 60000.00, Std Dev: 39833.93
```

The **"Summary of Features"** section provides an overview of the key variables in the dataset, including their types (e.g., numerical, categorical), distributions, and basic statistics (such as mean, median, and standard deviation).

Analysis

The analysis of the dataset has led to the identification of two significant trends:

1. **Larger Engine Sizes and Higher Prices**: The data suggests that there is a positive correlation between engine size and the price of vehicles. Specifically:

   Vehicles with larger engine sizes tend to be more expensive. This could be due to the fact that larger engines are often found in premium or luxury vehicles, which come with more advanced features, higher performance capabilities, and superior build quality.

   The higher costs associated with larger engines may also reflect factors such as greater fuel consumption, more expensive parts, and higher maintenance costs, which make these vehicles less affordable for the average buyer.

2. **Automatic Transmissions and Higher Price Ranges**: The analysis shows that vehicles equipped with automatic transmissions are generally positioned in the higher price range. Possible reasons for this trend include:

   Automatic transmissions are often found in higher-end vehicles, where comfort, ease of use, and advanced features (such as adaptive transmission systems) are prioritized. These vehicles are generally priced higher due to the additional technology and engineering.

   Automatic transmission systems are more expensive to manufacture and repair than manual ones, contributing to the increased cost of vehicles equipped with them.

   Additionally, automatic transmissions are typically associated with vehicles designed for a more convenient, less labor-intensive driving experience, which may appeal to a wealthier demographic.

Further Insights and Actions:

Target Market Segmentation:

These insights can help in identifying the target market for different vehicle types. Consumers looking for larger engines or automatic transmissions may be more inclined to purchase vehicles at higher price points, indicating that marketing strategies should focus on the luxury, performance, and convenience aspects of these vehicles.

Pricing Strategies:

For dealerships or car manufacturers, understanding that engine size and transmission type are key price determinants could help in designing better pricing strategies. For example, they might emphasize the benefits of larger engines and automatic transmission models in marketing campaigns targeted at higher-income consumers.

Future Investigations:

Further analysis could explore additional variables, such as brand, model year, or vehicle condition, to provide a more comprehensive view of what influences vehicle prices. Additionally, cross-referencing engine size and transmission type with fuel efficiency or emissions data could reveal even more insights for eco-conscious buyers.

In summary, this analysis highlights the relationship between engine size, transmission type, and price, providing valuable insights for both consumers and businesses in the automotive market.

Conclusion

This project effectively demonstrates the power of Exploratory Data Analysis (EDA) in uncovering insights from raw automotive data. By following a structured methodology of data cleaning, outlier detection, feature engineering, and visual exploration, we were able to transform a messy dataset into a clean and insightful resource. The comprehensive analysis provided valuable perspectives on factors that influence car prices, such as mileage, engine size, fuel type, and the age of the vehicle.

One of the key highlights of this project was the identification and handling of missing values and outliers, which can often distort the findings of data analysis. Through careful imputation and capping of extreme values, the data was made robust for further analysis. The use of various visualizations, including scatter plots, heatmaps, and histograms, helped in identifying relationships between variables and provided an intuitive understanding of how different features interact with one another.

The feature engineering process added significant value by creating new metrics, such as the price-to-mileage ratio, which helped in understanding the cost-efficiency of vehicles. The dataset was optimized through transformations and encoding, making it ready for potential predictive modeling or further in-depth analysis.

In conclusion, this project highlights the critical role of data analysis in the automotive industry, offering a clear demonstration of how data can be harnessed to drive insights and influence business strategies. The cleaned and analyzed dataset lays the groundwork for future predictive modeling, market analysis, or even real-time pricing strategies. It is a prime example of how EDA and systematic data preparation are essential in transforming raw data into actionable business intelligence.

References

1. Tukey, John W. "Exploratory Data Analysis." Addison-Wesley, 1977.

   This foundational work by John Tukey introduced the concept of Exploratory Data Analysis (EDA) and its importance in understanding and interpreting data before formal statistical modeling. Tukey's methods emphasized the use of visual tools and simple descriptive statistics to uncover patterns and anomalies in data.

2. Python Documentation for Pandas, Matplotlib, and Seaborn:

   Pandas Documentation:

   https://pandas.pydata.org/pandas-docs/stable/

   Official documentation for Pandas, a powerful Python library for data manipulation, including tools for handling data structures like DataFrames and Series, as well as performing various data-cleaning and analysis tasks.

   Matplotlib Documentation:

   https://matplotlib.org/stable/contents.html

   Comprehensive reference for Matplotlib, a popular Python library for creating static, interactive, and animated visualizations.

   Seaborn Documentation: https://seaborn.pydata.org/

   Seaborn's official site, which provides advanced data visualization features built on top of Matplotlib, designed for easier and more attractive statistical plotting.

3. Research Papers on EDA Techniques and Applications:

These papers explore different techniques and methodologies within the field of Exploratory Data Analysis, providing insights into its application across various domains of data science, such as anomaly detection, feature selection, and statistical inference. Examples of references could include:

"An overview of Exploratory Data Analysis" by J. H. Friedman.

"A Survey on Data Preprocessing for Classification" by M. R. K. Krishna.

These references provide both the foundational theory behind EDA and practical resources for using Python libraries to perform it effectively.

Github Repository Link

https://github.com/MarneediJaswanth/EDA---Car-Data-Analysis