Project Report: DAI-101 Assignment 1
Tanay Kapadia (23323044)

Introduction

Data cleaning is a crucial step in the data science workflow that involves identifying and rectifying errors or inconsistencies in the dataset. This process includes handling missing values, removing duplicates, correcting data types, and standardizing formats to ensure the dataset is accurate and ready for analysis. In this project, I used NumPy and pandas for data manipulation and cleaning tasks.

Exploratory Data Analysis (EDA) is the process of summarizing and visualizing the main characteristics of a dataset. It helps to uncover patterns, detect anomalies, test hypotheses, and gain insights into the data. I utilized seaborn and matplotlib to create visualizations such as histograms, scatter plots, and correlation matrices, which provided a deeper understanding of the dataset's structure and relationships.

The used cars dataset contains information on pre-owned vehicles, including features such as price, mileage, year of manufacture, fuel type, transmission type, and more. This dataset is sourced from online car listings and reflects real-world data on vehicle attributes and market trends. It serves as an excellent resource for analyzing factors that influence car prices, identifying market patterns, and building predictive models. Throughout this project, I applied various data cleaning techniques and performed EDA to extract meaningful insights from the dataset.

Understanding the Dataset

Upon performing an initial analysis using basic functions, it became evident that the dataset was both extensive and detailed. Specifically, the dataset contained a total of 9,576 entries and included 10 distinct attributes, each providing various pieces of information related to the characteristics of the cars.

During this preliminary examination, it was also observed that the dataset contained a significant number of null (or missing) values. Furthermore, by utilizing the data.describe() function, it became apparent that a considerable portion of the entries for key attributes, such as price and mileage, were recorded as zero. Since these values are unrealistic in real-world scenarios, they should also be treated as missing or invalid data points. This observation suggests that further data cleaning and preprocessing will be necessary to ensure the accuracy and reliability of any subsequent analysis.

Data Cleaning

As an initial step, I examined the dataset for duplicate rows and identified exactly 113 duplicate entries. These duplicates were subsequently removed using the built-in functions provided by the Pandas library. Removing duplicate records is a logical approach, as it is improbable that two cars would have identical attributes across all fields. Moreover, even if such cases exist, retaining only unique entries is preferable for analytical purposes to ensure the dataset accurately reflects distinct observations.

From the previous analysis, it was already established that two columns represented missing values using the number 0, while the engType column used the label "other" to signify null values. To standardize the handling of missing data, I utilized the replace() function in Pandas to convert these placeholders into actual null (NaN) values. This adjustment resulted in a total of three numerical and two categorical columns being treated as containing missing data.

To address these missing values, I employed different imputation strategies for categorical and numerical data. For categorical variables, I replaced missing values with the mode (the most frequent value), while for numerical variables, I used the median. However, the imputation process was performed with careful consideration to preserve the dataset's accuracy and reflect real-world conditions as closely as possible.

For instance, when imputing missing values in the price column, I followed a hierarchical strategy to determine the most appropriate median value. If multiple cars of the same brand and model were available, I used the median price of those specific cars. If only cars of the same brand (but not the same model) were available, I took the median of those records. Only when neither condition was met did I fall back to using the median of the entire dataset. For the mileage column, I included the year attribute in the imputation process, as it is reasonable to assume a strong correlation between a car's mileage and its manufacturing year.

Additionally, based on the earlier descriptive analysis of the dataset, it became apparent that the maximum values of certain metrics—particularly mileage and engine volume (engV)—were unreasonably high. This observation was supported by an examination of the mean and standard deviation, which indicated extreme deviations from typical values. Such outliers are unrealistic and could significantly distort the analysis if left unaddressed. Therefore, I applied the Interquartile Range (IQR) method to detect and remove these outliers, ensuring a cleaner and more reliable dataset for subsequent analyses.

Univariate Data Analysis

The dataset consists of four numerical and six categorical attributes. To summarize the numerical attributes, I used the built-in describe() function, which provides key statistical measures. Additionally, I visualized the distributions by creating histograms and box plots for all four numerical attributes, helping to identify patterns and potential outliers.

For the categorical attributes, summarization required the use of the value_counts() function to determine the five most and least frequent categories. To better visualize these distributions, I created bar plots representing the frequency of each category. In cases where the number of unique categories was high, I limited the visualization to the top 10 most frequent categories for clarity.

Bivariate/Multivariate Data Analysis

The analysis utilized three main techniques. First, pair plots were used to analyze multiple relationships simultaneously between variables such as price, mileage, year of manufacture, and engine size. Second, heatmaps were generated to visualize correlations among multiple variables and identify strong or weak associations. Finally, grouped comparisons were conducted by categorizing data based on features like car brand, fuel type, and transmission to identify combined effects on price and other variables.

The analysis of the used-cars dataset revealed several key insights. There is a clear negative correlation between price and mileage, meaning cars with higher mileage tend to have lower prices. Newer cars, as indicated by their year of manufacture, are generally priced higher due to natural depreciation over time.

Brand influence is also significant. Premium brands like BMW, Mercedes-Benz, and Audi consistently command higher prices compared to mainstream brands like Toyota and Ford. Some brands exhibit slower depreciation, retaining their value more effectively over time.

Fuel type and transmission play a crucial role in pricing. Electric and hybrid cars are priced higher than gasoline and diesel vehicles. Additionally, cars with automatic transmissions tend to have higher prices than manual ones due to consumer preference and increased production costs.

Engine size correlates with price, as larger engines typically command higher prices. However, there are diminishing returns, as extremely large engines do not proportionally increase a car's value.

Conclusion

In conclusion, this project involved thorough data cleaning and exploratory data analysis on a used cars dataset. By addressing missing values, correcting inconsistencies, and visualizing key attributes, I was able to uncover important patterns and insights. The combination of NumPy, pandas, seaborn, and matplotlib enabled efficient data manipulation and visualization.