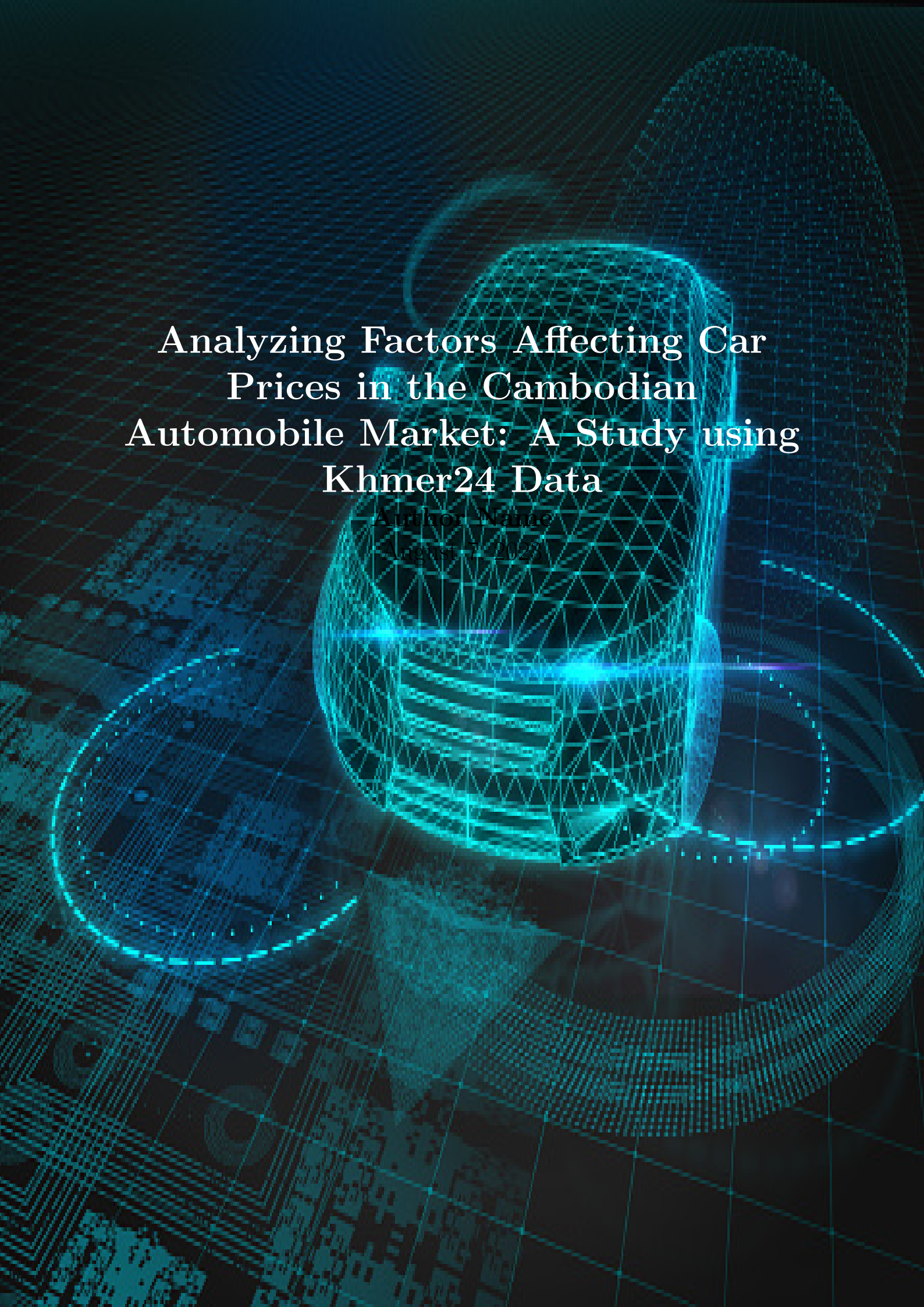

Analyzing Factors Affecting Car Prices in the Cambodian Automobile Market: A Study using Khmer24 Data

Author Name

August 7, 2023

Contents

1 Introduction

The automobile market in Cambodia has been growing rapidly in recent years, with an increasing number of car buyers seeking to purchase new and used cars. As a result, there has been a growing interest in determining the factors that influence car prices in Cambodia. In this study,**we aim to explore the relationship between various car features and their impact on car prices in the Cambodian market.**

To achieve this goal, we collected data on car prices and features from Khmer24 website, one of the largest online marketplaces for cars in Cambodia. We scraped the data and compiled a data set containing information on various car features such as make, model, year, and transmission type, as well as their corresponding prices.

2 Data Description

The data set used for this analysis contains 17,873 rows/records and 16 columns/features and information related to cars listed on the Khmer 24 website for sale in Cambodia. The data set includes the following attributes:

- **Ad ID:** A unique identifier is assigned to each car listing on the Khmer 24 website.
- **Category:** The type of car being advertised for sale, such as sedan, SUV, hatchback, etc.
- **Posted:** The date when the car was posted for sale on the website.
- **Car Makes:** The brand or manufacturer of the car, such as Toyota, Honda, Ford, etc.
- **Car Model:** The specific model of the car, for example, Camry, Civic, Focus, etc.
- **Year:** The year in which the car was manufactured.
- **Tax Type:** The type of tax associated with the car, with two possible categories.
- **Condition:** The condition of the car, which can be either new or used.
- **Body Type:** The type of car body, such as SUV, sports car, sedan, etc.
- **Fuel:** The type of fuel used by the car, like gasoline, diesel, hybrid, etc.
- **Transmission:** The type of transmission system that transfers power from the engine to the wheels, such as automatic or manual.
- **Color:** The color of the car, such as red, blue, black, etc.
- **Link:** A link to the car's detailed information on the Khmer 24 website.
- **Title:** A description or title of the car's listing.
- **Price:** The price of the car, serves as the target variable in this analysis.

3 Data Cleaning and Prepossessing

Data cleaning and prepossessing are essential steps in preparing the data set for analysis. In this section, we describe the various techniques applied to clean and prepossess the raw data.

3.1 Handling Missing Values

Missing values in the data set can adversely affect the analysis. We examined each attribute to identify missing values and decided on an appropriate strategy to handle them [1]. Common techniques used include:

- **Dropping Rows:** If the number of missing values in a row is significant, we considered dropping those rows from the data set.
- **Imputation:** For numerical attributes, we used techniques such as mean, median, or mode imputation to fill in missing values.
- **Categorical Imputation:** For categorical attributes, we imputed missing values using the most frequent category.

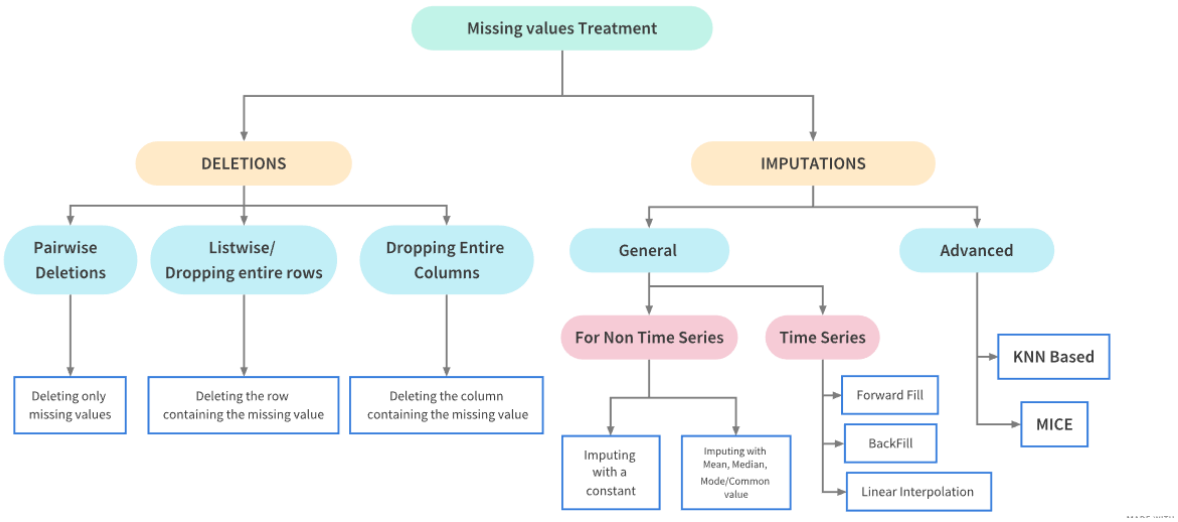

Figure 1: Technique to handle missing values

Real-world data is messy and often contains a lot of missing values. There could be multiple reasons for the missing values, but primarily the reason for missingness can be attributed to the following:

1. Data doesn't exist: Some data points may be genuinely missing because they were never collected or recorded.
2. Data not collected due to human error: Missing values may result from errors or oversights during data collection, where certain data points were not recorded or were recorded incorrectly.

3. Data deleted accidentally: Missing values can also occur if data was accidentally deleted or lost during data handling or processing.

Either way, we need to address this issue before we proceed with the modeling stuff. It is also important to note that some algorithms like XGBoost and LightGBM can treat missing data without any prepossessing.

3.1.1 Detection missing data visually using Missingno library

To graphically analyze the missingness of the data, let's use a library called Missingno It is a package for graphical analysis of missing values.

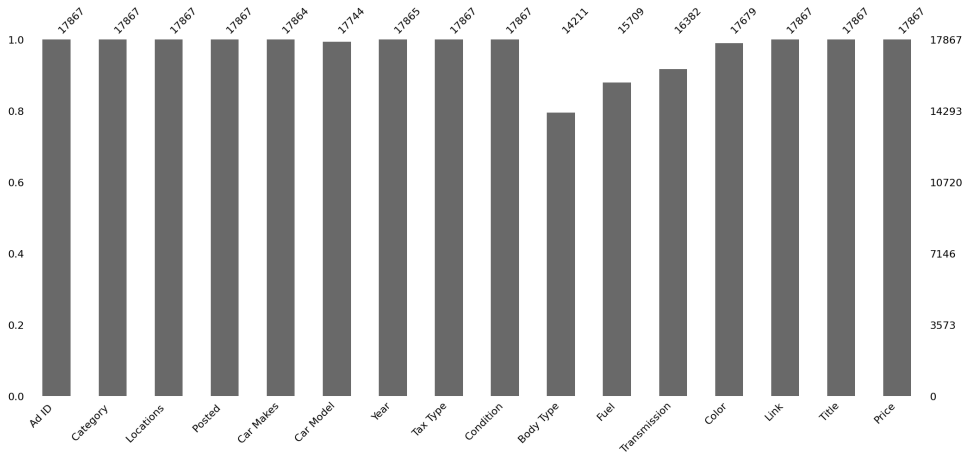

Figure 2: Visualizing Missing values

The bar chart above gives a quick graphical overview of the completeness of the data set. We can see that Body Type, Fuel, and Transmission columns have missing values. Next, it would make sense to find out the locations of the missing data.

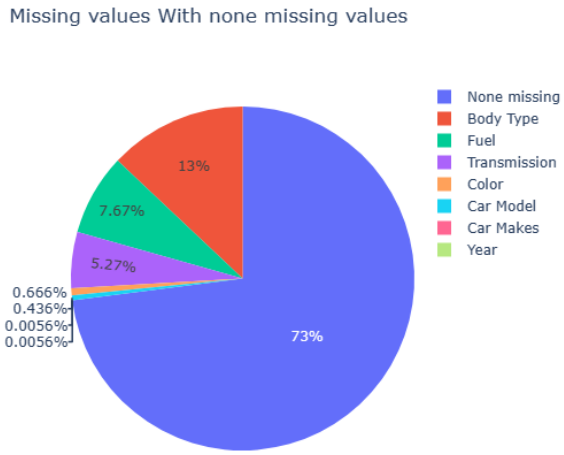

Figure 3: Visualizing percentage of Missing values

3.1.2 Visualizing the location of the missing data

The msno.matrix nullity matrix is a data-dense display that lets you quickly visually pick out patterns in data completion.

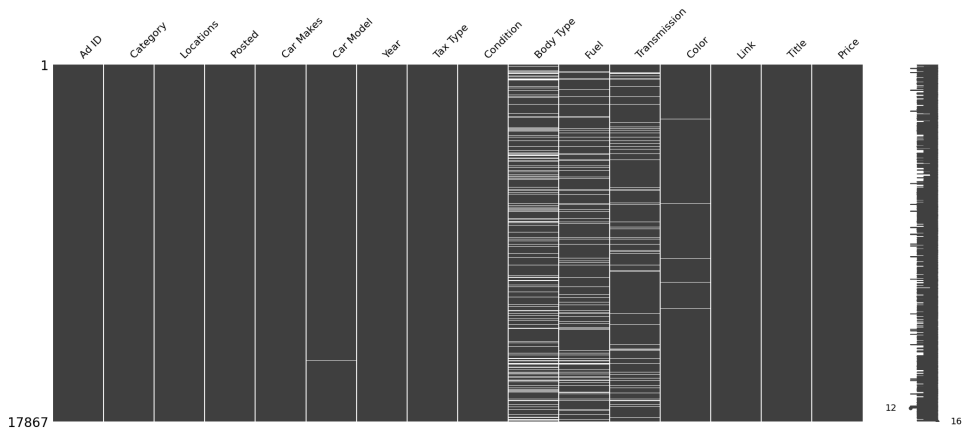

Figure 4: Location missing values

The Color and Car Model columns have very few missing values and do not seem to be correlated with any other column. Hence, the missingness in these columns can be attributed as Missing Completely at Random (MCAR).

Both the Body Type, Fuel, and Transmission columns have a lot of missing values. This could be a case of Missing at Random (MAR) as we cannot directly observe the reason for the missingness of data in these columns.

Before we start treating the missing values, it is important to understand the various reasons for the missingness in data [3]. Broadly speaking, there can be three possible reasons:

1. **Missing Completely at Random (MCAR)**
   The missing values on a given variable (Y) are not associated with other variables in a given data set or with the variable (Y) itself. In other words, there is no particular reason for the missing values.

2. **Missing at Random (MAR)**
   MAR occurs when the missingness is not random, but where missingness can be fully accounted for by variables where there is complete information.

3. **Missing Not at Random (MNAR)**
   Missingness depends on unobserved data or the value of the missing data itself.

3.1.3 Finding reason for missing data using a Heat map

The correlation heat map of missing values revealed interesting patterns in the data set from the Khmer24 website. We observed moderate correlations between certain missing values, indicating potential dependencies among these variables. Specifically, the car body type exhibited a high correlation with fuel and transmission, with correlation coefficients of 0.5 and 0.3, respectively. Additionally, fuel displayed a notable correlation of 0.4 with the transmission variable. These findings suggest that the missingness in the car body type, fuel, and transmission columns might not be entirely random. The high correlations

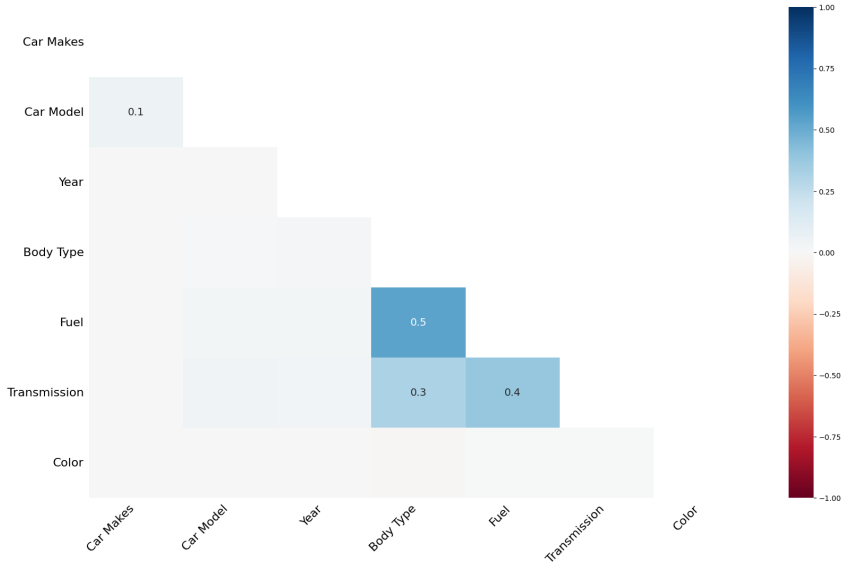

Figure 5: Missing values heat map

could be attributed to specific seller behaviors or input patterns on the website, leading to missing values in these columns.

3.1.4 Finding reason for missing values using Dendrogram

A dendrogram is a tree diagram of missingness. It groups the highly correlated variables together.

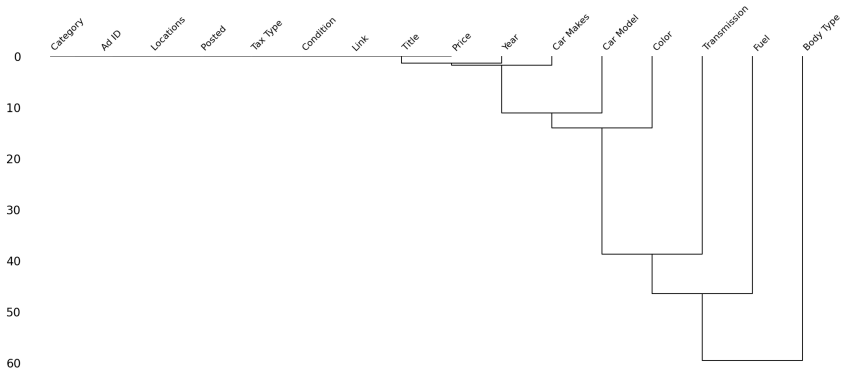

Figure 6: Missing values with dendrogram

Let's read the above dendrogram from a top-down perspective:

- Cluster leaves which are linked together at a distance of zero fully predict one another's presence—one variable might always be empty when another is filled, or they might always both be filled or both empty and so on(missingno documentation)
- The missingness of Transmission tends to be more similar to Fuel than to Body Type and so on. However, in this particular case, the correlation is high since Body Type column has very few missing values.

This data set doesn't have many missing values, to address this issue, we have decided to use list-wise deletion as the method for handling missing values. List-wise deletion

involves excluding entire rows with missing values in any of the variables of interest. By doing so, we can retain the cases with complete information, thereby ensuring the reliability and accuracy of subsequent analyses based on this data set. However, it is important to acknowledge that list-wise deletion may lead to a reduction in the sample size and potential loss of valuable information. Therefore, we will carefully assess the impact of this approach on our analyses and results. Domain expertise and careful consideration of the missing data patterns will be crucial in making informed decisions for handling these missing values effectively.

3.2 Data Transformation

Some attributes may require data transformation to make them suitable for analysis. Common transformations include:

- **Normalization:** We performed normalization on numerical attributes to scale them to a similar range.
- **Encoding Categorical Variables:** Categorical variables were encoded using techniques like one-hot encoding or label encoding to convert them into a numerical format.

3.2.1 Target Distribution

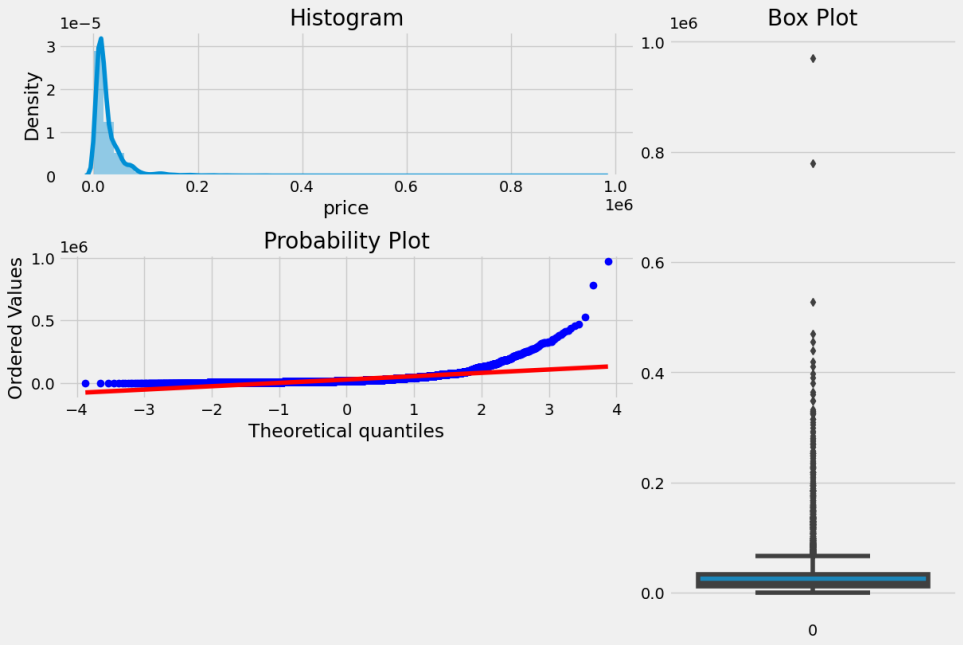

Figure 7: Visualize target variable

Table 1: Descriptive Statistics of Car Price

	Min	Mean	Median	Max	Std	Skew
Price	$500.00	$28,462.82	$18,000.00	$970,000.00	$35,626.04	6.0948

We can clearly observe that our Car Price feature is highly right-skewed, which can be problematic for some machine learning algorithms. Right-skewed data can lead to biased models and inaccurate predictions.

Furthermore, there is a significant difference between the mean and median values, indicating the presence of outliers in the data.

The skewness of the car price is above 1.5, which means that the data points are highly spread out in the positive direction. This further confirms the right-skewness of the distribution.

Addressing the skewness in the Car Price feature is essential before building any machine learning models to ensure better model performance and more reliable predictions.

3.2.2 Handling Skewed Data

Skewed data refers to a distribution where the curve appears distorted either to the left or to the right, with one tail longer than the other. Skewed data can lead to biased analyses and may negatively impact the performance of certain machine-learning algorithms.

There are several techniques to handle skewed data, including:

- **Log Transform:** Taking the logarithm of the data can help reduce the skewness and bring the distribution closer to a normal shape.
- **Box Cox Transform:** The Box-Cox transformation is a family of power transformations that can be applied to stabilize variance and make the data less skewed.
- **Square Root Transform:** Taking the square root of the data can also be used to reduce the skewness and make the distribution more symmetrical.

By applying one of these transformation techniques to the skewed data, we can achieve a more balanced and normalized distribution, which can improve the performance of statistical analyses and machine learning models.

Box Cox Transform

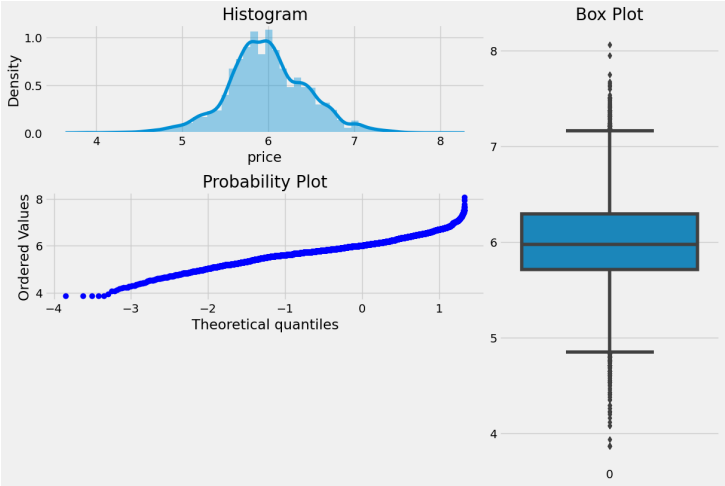

Figure 8: Visualize target variable with box Cox transform

The Box-Cox transforms to the car price data. The resulting skewness value is -0.2. The Box-Cox transformation is a family of power transformations that find the best power value (lambda) to stabilize the variance and make the data more Gaussian-like.

Log transformation

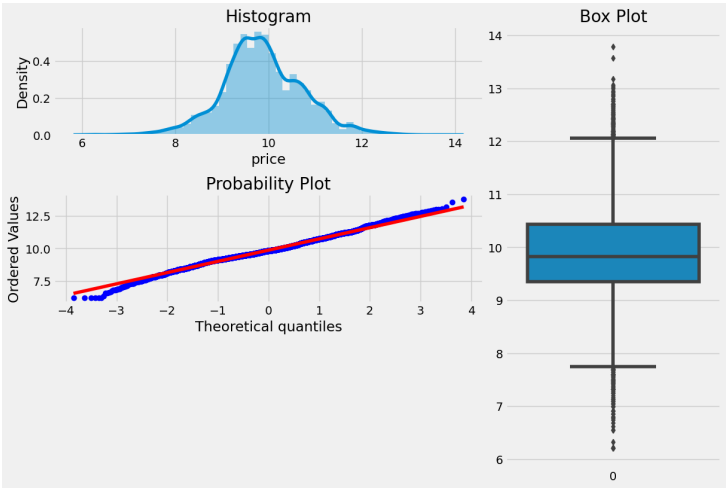

Figure 9: Visualize target variable with logarithm transform

Next, we apply the log transform is applied to the car price data, resulting in a skewness value of 0.11. The log transformation is useful for data with positive values and right skewness. It compresses the data and makes it more symmetric, making it suitable for some statistical analyses.

Square root transformation

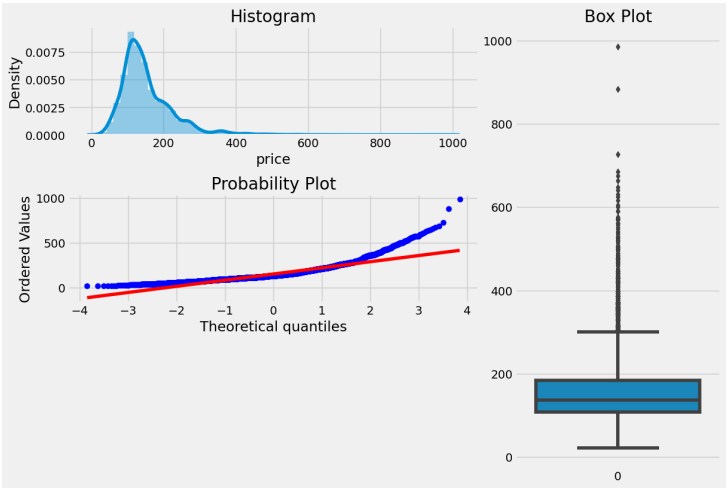

Figure 10: Visualize target variable with square root transform

Finally, we perform the square root transform on the car price data, resulting in a skewness value of 2.04. The square root transformation is useful for data with positive values and right skewness. It reduces the magnitude of large values and can be effective in some cases.

After analyzing the skewness values obtained from each transformation technique, we observe the following:

- The log transform yielded a skewness value of 0.11.
- The Box-Cox transform yielded a skewness value of -0.2.
- The square root transform yielded a skewness value of 2.04.

A skewness value close to 0 indicates that the data is approximately symmetric around its mean. In this regard, the log transform outperforms the other techniques, as it resulted in a skewness value closer to 0 (0.11). Therefore, we recommend using the log-transformed car price data for further analysis and modeling.

The log transformation will make the data more suitable for Gaussian-like assumptions and improve the performance of certain statistical analyses. It helps address the right skewness and ensures more reliable modeling results.

3.3 Dealing with Outliers

An outlier is an observation that is unlike the other observations. It is rare, or distinct, or does not fit in some way. It is also called anomalies. Outliers can have many causes, such as:

- Measurement or input error.
- Data corruption.
- True outlier observation.

Impact Of Outliers:

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA, and other statistical model assumptions.

Outliers in the data can significantly affect the analysis and model performance. We identified outliers using visualization techniques such as box plots and handled them using methods such as:

- **Trimming:** Outliers beyond a certain threshold were trimmed or capped to minimize their impact.
- **Imputation:** In some cases, we imputed outliers with more reasonable values based on the context of the data.

3.3.1 Outlier Detection Techniques

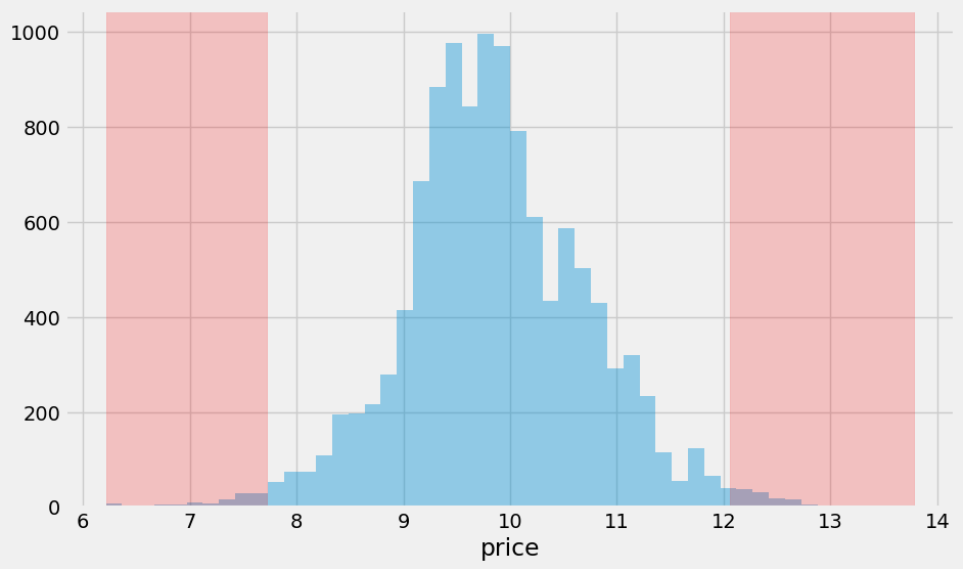

Figure 11: Visualize outliers

The Interquartile Range (IQR) is a statistical measure that represents the range between the first quartile (Q1) and the third quartile (Q3) of the car price data. In our analysis, the calculated IQR for car prices is approximately 2.96.

Using the IQR, we can identify outliers in the car price data. The lower bound for outliers is obtained by subtracting 1.5 times the IQR from Q1, and the upper bound is obtained by adding 1.5 times the IQR to Q3. For our car price data, the lower bound value is approximately 2262.18, and the upper bound value is approximately 172842.50.

We have identified a total of 247 outliers in the car price data. These outliers are observations that fall below the lower bound or above the upper bound. Outliers can indicate potential errors, anomalies, or unusual car prices that deviate significantly from the majority of the data.

After performing outlier detection on the car price data, we identified 247 influential outliers. These outliers have a substantial impact on the analysis and modeling process, and their presence can lead to biased results.

In order to ensure the accuracy and robustness of our analysis, we have decided to remove the identified influential outliers from the car price data set. By doing so, we aim to mitigate the potential distortion they may introduce to our statistical measures and machine learning models.

4 Exploratory Data Analysis

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing [7].

4.1 Exploration of Car company

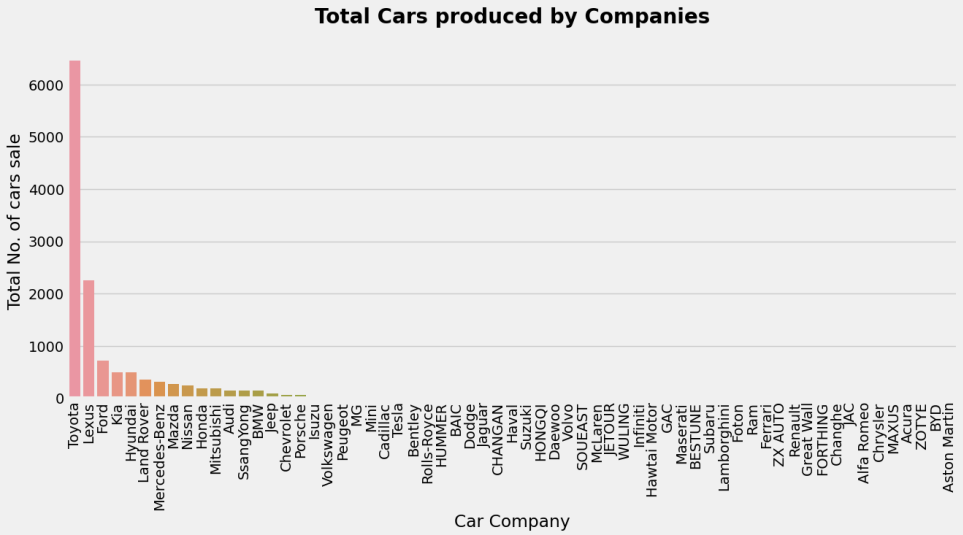

Figure 12: Visualize car companies

Based on the analysis of the car sales data, we have uncovered the following insights:

- Toyota is the leading car company in terms of the number of cars sold. This indicates that Toyota is highly favored by customers and has a strong market presence.
- On the other hand, MAXUS, Chrysler, Alfa Romeo, Acura, ZOTYE, BYD, and Aston Martin have a very low number of data points, which means that there is insufficient data to make any meaningful inferences about the sales of these car companies. Further data collection and analysis are required to gain a better understanding of their sales performance.

These insights provide valuable information about the market dynamics and customer preferences in the car industry. As we move forward with our analysis, it is essential to consider these findings to make informed decisions and draw meaningful conclusions from the data.

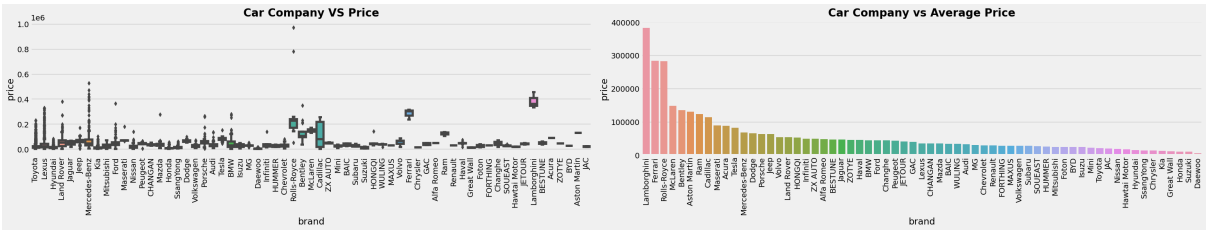

Figure 13: Visualize car companies

After analyzing the car prices, we have identified the following insights:

- Lamborghini, Farrari, and Rolls-Royce are car companies known for producing high-end luxury cars with a wide price range. This indicates that these companies cater to customers looking for premium and luxurious vehicles.

- On the other hand, car companies like Daewoo and Suzuki are underrepresented in our dataset, with only a limited number of data points available for analysis. Due to the small sample size, it is challenging to draw meaningful inferences about the pricing patterns of these car companies.

These insights shed light on the pricing strategies of different car companies and their market positioning. However, it is important to acknowledge the limitations of the data and exercise caution when drawing conclusions for car companies with limited representation in the dataset.

4.2 Exploration of Car fuel type

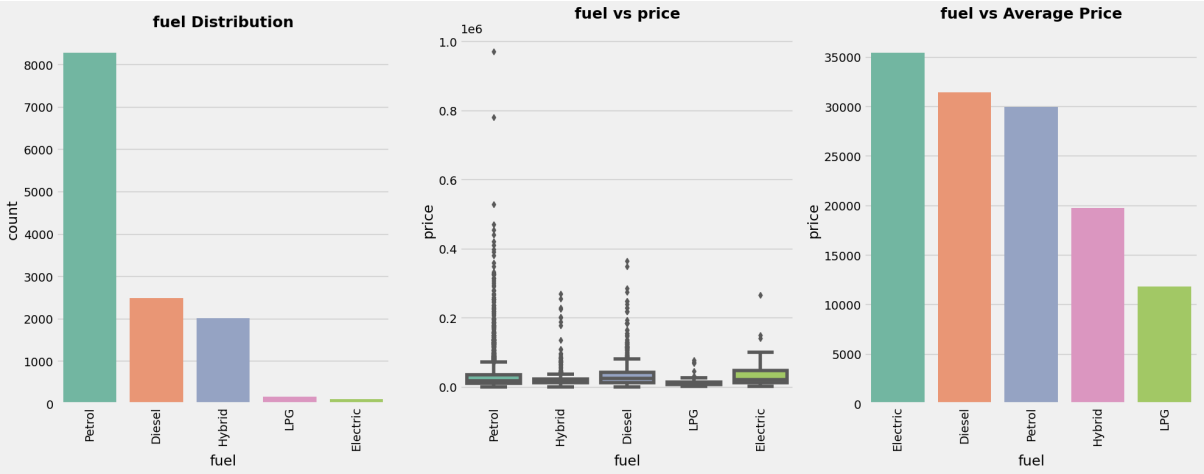

Figure 14: Visualize car fuel types

After analyzing the data related to fuel types and car prices, we have identified the following insights:

- Cars equipped with a petrol fuel system are the most commonly sold in the market. This indicates that petrol-powered cars are preferred by a larger segment of customers compared to other fuel types.

- From the second plot, we can observe that petrol fuel system cars are available within every price range. This implies that customers have various options for petrol-powered vehicles regardless of their budget, making them a versatile choice in the market.

- The third plot reveals that on average, the price of cars with a petrol fuel system is lower than those with a diesel fuel system. This suggests that petrol-powered cars may offer a more affordable option for potential buyers compared to diesel-powered cars.

These insights provide valuable information about the popularity and pricing patterns of cars with different fuel systems. They can assist in understanding customer preferences and market dynamics, which can be essential for car manufacturers and dealers in making informed business decisions.

4.3 Exploration Tax Type

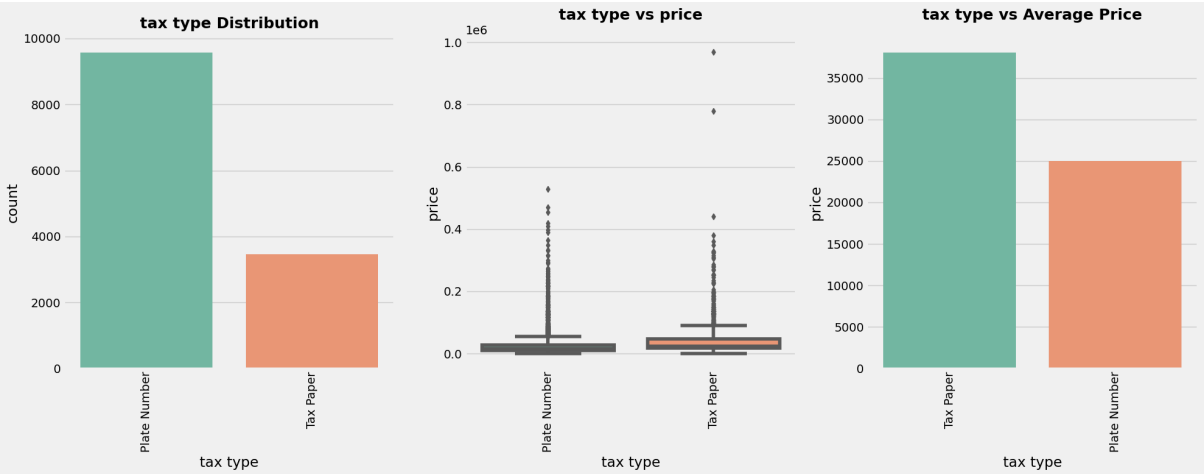

Figure 15: Visualize car tax type

Observation

After analyzing the data, the following observations were made:

- Cars with Plate Numbers are more frequently sold compared to cars with Tax Paper. This indicates that a significant number of car transactions involve cars with Plate Numbers.
- Cars with Tax Paper tend to be more expensive than cars with Plate Numbers. This suggests that cars with Tax papers might have higher taxes or additional fees, making them costlier for buyers.

Insights

Based on the observations, we can derive the following insights:

- The presence of outliers is evident in both the sales of cars with Plate Numbers and cars with Tax papers. Outliers can be influential data points that might impact the overall analysis and conclusions. Careful consideration and handling of these outliers may be necessary during further data analysis.

These insights provide valuable information about the sales and pricing patterns of cars with Plate Numbers and Tax papers. Understanding these trends can be crucial for making informed decisions in the automotive market and related industries.

4.4 Exploration of car body type

Observation

After analyzing the data, the following observations were made regarding the body-type of cars:

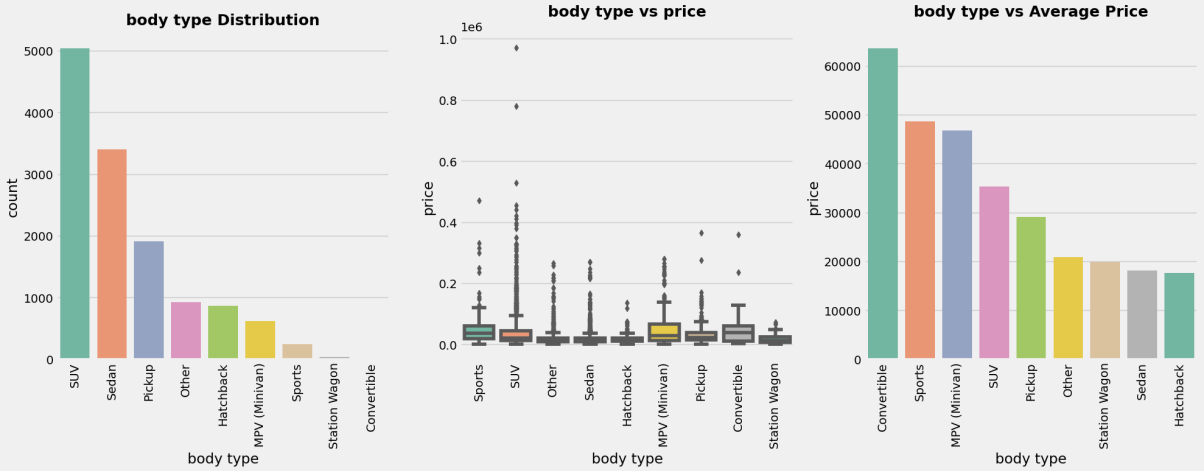

Figure 16: Visualize car body type

- Cars with SUV body-type are the most frequently sold, followed by Sedan.
- Cars with Convertible or Hardtop body-type are relatively less sold in comparison.
- Cars with Convertible body-type are the most expensive cars, followed by Sport.

Insights

Based on the observations, we can derive the following insights:

- The lower sales of cars with Convertible and Sport body-types could be attributed to their high prices. These body-types are often associated with luxury and performance, making them less affordable for a majority of customers. Hence, customers might prefer more budget-friendly options.
- Despite being the fourth most expensive body-type, SUVs have the highest number of car sales. This indicates that customers show a preference for medium-priced cars with SUV body-types, likely due to their versatility and utility.

Note: The insights obtained from this analysis can help car manufacturers and dealers understand customer preferences and make informed decisions about production and marketing strategies.

4.5 Exploration Condition

4.5.1 ANOVA Table

The ANOVA (Analysis of Variance) table provides statistical information about the relationship between the condition variable and the price variable:

Source	sum_sq	df	F	PR($> F$)
condition	366373802411.8281	1.0000	295.1744	0.0000
Residual	16182914796392.3242	13038.0000	NaN	NaN

The ANOVA table shows that there is a significant linear relationship between the condition variable and the price variable.

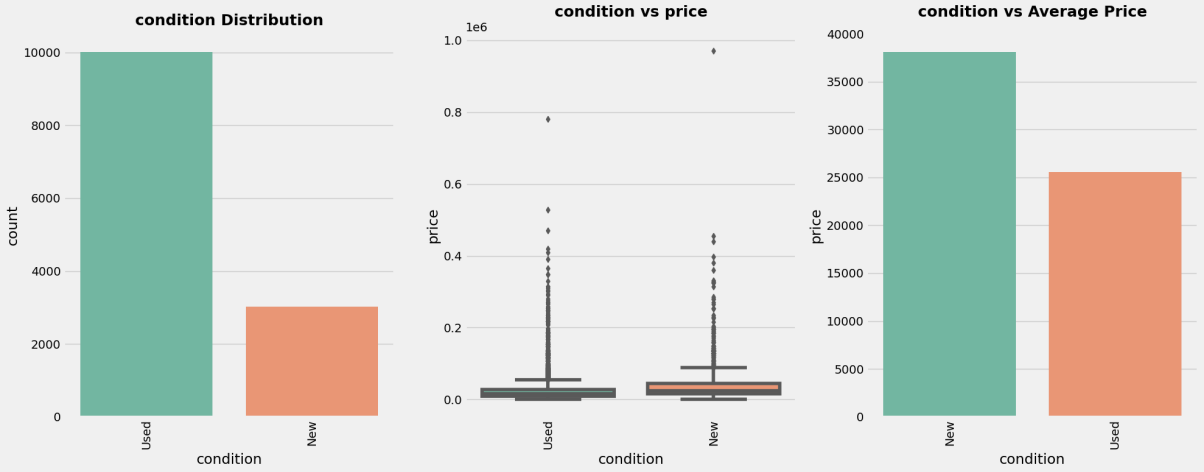

Figure 17: Visualize car condition

Observation

Based on the ANOVA analysis, the following observation can be made:

- The website has a higher number of used cars listed compared to new cars.
- On average, Cars New have a higher price compared to Cars Used, which could be due to factors like depreciation, warranty, and overall condition.

4.6 Exploration of year

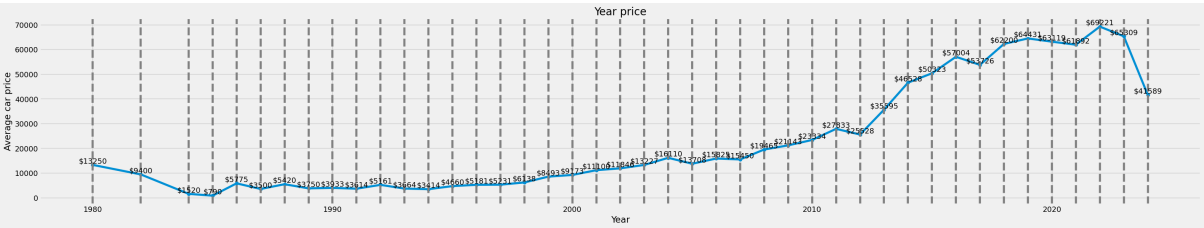

Figure 18: Visualize year

Based on the data analysis, the following observation can be made the average prices of cars seem to have increased from the year 2012 to 2022.

5 Key Observations

Based on the exploratory data analysis (EDA) of the car sales data set, the following key observations have been made:

1. Toyota company has the highest number of car sales, indicating that Toyota is the most favored car company among customers.
2. Some car companies, such as MAXUS, Chrysler, Alfa Romeo, Acura, ZOTYE, BYD, and Aston Martin, have very few data points, making it challenging to draw meaningful inferences about these companies.

3. Luxury car brands like Lamborghini, Ferrari, and Rolls-Royce offer cars with a wide price range, catering to different customer segments.

4. Car companies like Daewoo and Suzuki have a limited presence in our dataset, restricting the insights we can derive about the lowest price range car companies.

5. Cars with petrol fuel systems are the most commonly sold, and they are available within every price range. Additionally, the average price of petrol fuel type cars is lower than diesel fuel type cars.

6 Conclusion

In conclusion, the exploratory data analysis (EDA) has provided valuable insights into the car sales data set. The key findings are as follows:

- Toyota is the most favored car company among customers, with the highest number of sales.
- Some car companies have limited data points and further analysis is required to make meaningful inferences about them.
- Luxury car brands like Lamborghini, Ferrari, and Rolls-Royce offer cars across a wide price range, appealing to different customer segments.
- Cars with petrol fuel systems dominate the market, and they are available at various price points.
- Daewoo and Suzuki have a smaller presence in the data set, and more data would be beneficial for analyzing the lowest price range car companies.

The exploratory data analysis has provided a solid foundation for further analysis and decision-making based on the data. To gain deeper insights, future analysis could involve exploring additional variables and their impact on car sales, investigating customer preferences for specific car models, and conducting market segmentation to target different customer groups effectively.

References

[1] Dealing with Missing Data in Python

[2] How to handle missing values

[3] Reason for Missing values

[4] Algorithm to handle missing values

[5] Box-Cox Transformation and Target Variable: Explained

[6] How to Detect, Handle and Visualize Outliers

[7] Exploratory data analysis