Analyzing the Used Car Market: Insights from Data on Price, Performance, and Preferences

Teslari Maxim

December 17, 2023

Abstract

This research aims to analyze the used car market through a comprehensive data set containing various details about cars available for resale. The study will explore how factors such as the car's age, mileage, fuel type, and engine specifications affect its selling price. It will employ statistical analysis methods to identify trends and patterns in car pricing and consumer preferences.

Additionally, the paper will investigate the impact of different transmission types and owner history on a car's market value. Machine learning techniques will be utilized to predict selling prices based on the car's features. This analysis will provide valuable insights into the used car market dynamics, offering a deeper understanding of the factors influencing car valuation and buyer's choice. The findings could be instrumental for potential buyers, sellers, and market analysts to make informed decisions in the used car market.

1 Introduction

The used car market is a dynamic and complex segment of the automotive industry, offering a diverse range of choices for consumers. Understanding the factors that influence the selling price and demand of used cars is crucial for both buyers and sellers. This research aims to analyze various characteristics of used cars, such as selling price, kilometers driven, fuel type, transmission, and engine specifications, using a comprehensive data set obtained from a popular online marketplace.

The data set comprises detailed information about cars listed for resale, including the year of manufacture, selling price, kilometers driven, type of fuel used, transmission type, and engine specifics. This research employs a variety of data analysis techniques, including descriptive statistics, correlation analysis, and linear regression modeling, to uncover the relationships between these factors and the selling price of the cars.

The initial phase of the analysis involved data pre processing, where the data set was cleansed and transformed for effective analysis. Empty rows were removed, and key variables like engine size, mileage, and maximum power were converted from textual to numeric formats for precise computations. Additionally, categorical variables such as fuel type, transmission type, and ownership were transformed into binary values to facilitate their inclusion in the predictive modeling process.

Further, the study utilized various graphical representations such as histograms, box plots, bar charts, and scatter plots to visualize the distribution and relationship between different variables. This visual analysis provided initial insights into the trends and patterns within the used car market.

The cornerstone of this research is the application of multiple linear regression analysis, which helped in understanding the impact of various factors on the selling price of used cars. The regression model indicated significant relationships between selling price and variables such as kilometers driven, age of the car, fuel type, seller type, transmission type, mileage, and engine capacity. This paper seeks to provide a comprehensive understanding of the used car market dynamics, aiding stakeholders in making informed decisions. The findings from this analysis could be particularly beneficial for potential buyers, sellers, and market analysts by offering a deeper understanding of the factors that influence car valuation and consumer preferences.

The used car market has been a subject of extensive research due to its economic significance and the complex factors influencing consumer behavior and pricing. Several studies have attempted to unravel these complexities, offering insights that form the foundation for this research.

One pivotal area of focus in existing literature is the impact of a car's physical attributes on its resale value. Borthakur (2023)[Bor23] emphasized the significance of factors such as age, mileage, and make and model in determining a used car's price. Their findings suggested a strong negative correlation between a car's age and its resale value, a pattern echoed in the work of Huang (2023)[Hua23], who also highlighted the diminishing value of cars with increased mileage.

Fuel type and engine specifications are other critical factors influencing used car pricing. Knittel (2009)[BKZ09] found that vehicles with diesel engines tended to retain their value better than their petrol counterparts, attributing this to the diesel engines' better fuel economy and longer lifespan. However, this trend has been subject to change due to evolving consumer preferences and environmental regulations, as discussed by Maklari (2023)[Esz23].

The role of transmission type in used car valuation has also been explored. In a comparative study, Gaulier (2000)[GH00] noted that cars with automatic transmission generally have a higher resale value than those with manual transmission, primarily due to the convenience factor and broader market appeal.

Beyond physical attributes, the perception of value and trust plays a crucial role. The study by Lin (2022)[Hua23] on seller types revealed that cars sold by dealerships or trusted sellers fetch higher prices than those sold by individuals, possibly due to perceived reliability and after-sales support.

While these studies provide valuable insights, there remains a gap in understanding the combined effect of these factors in a comprehensive model, particularly in emerging markets. Additionally, the rapid changes in technology and consumer preferences necessitate continuous analysis of trends in the used car market. This paper seeks to address these gaps by offering an integrated analysis of how various factors collectively influence used car prices in the current market context.

2 Methodology

This study aims to analyze the used car market by examining various factors affecting the selling price of used cars. The methodology section outlines the data source, pre processing steps, and analytical techniques employed in this research.

2.1 Data Source and Description

The primary data set for this study was obtained from a publicly available source on Kaggle, titled "car_details_v3.csv". This data set includes comprehensive information about used cars listed for resale, encompassing variables such as make and model, year of manufacture, selling price, kilometers driven, fuel type, seller type, transmission type, owner history, engine specifications, and seating capacity.

2.2 Data Pre processing

Data pre processing was a crucial step to ensure the quality and reliability of the analysis. The initial phase involved data cleaning, where rows with missing or incomplete information were identified and removed, reducing the data set size from 8,121 to 7,907 entries.

The next step was data transformation. Key variables like 'engine size', 'mileage', and 'maximum power', originally recorded as text strings, were converted into numerical values for accurate computations. Categorical variables such as 'fuel type', 'transmission type', 'seller type', and 'ownership' were converted into binary values (0 and 1) to facilitate their inclusion in the regression models.

2.3 Graphical Analysis

Various plots such as histograms, box plots, and scatter plots were generated using the ggplot2 package in R to visualize the data and identify patterns and outliers.

2.4 Regression Analysis

A multiple linear regression model was developed to understand the impact of various factors on the selling price of used cars. This analysis was conducted using the 'lm' function in R, which provided estimates of the regression coefficients and their significance.

2.5 Statistical Software

The entire analysis was conducted using R, a programming language and environment widely used for statistical computing and graphics. The following R packages were used: 'tidyverse' for data manipulation, 'dplyr' for data processing, 'ggplot2' for data visualization, and 'plotrix' for additional graphical tools.

3 Data Analysis

The data analysis section provides an in-depth examination of the data set to uncover trends, patterns, and relationships among the various factors influencing the selling price of used cars.

3.1 Descriptive Statistics

The analysis began with a descriptive statistical overview of the data set. Key variables such as selling price, kilometers driven, engine capacity, and mileage were summarized to understand their distribution. For instance, the average selling price of the cars in the data set was observed to be around 500,000 INR, with a notable range in prices suggesting a diverse market.

3.2 Graphical Analysis

Histograms were plotted for the number of seats in the cars, revealing a predominant preference for 5-seater vehicles in the used car market. Box plots of selling prices against the number of seats indicated a higher median selling price for cars with more seats, suggesting a potential premium on larger vehicles(or for family, business convenience). Bar charts were used to display the distribution of fuel types and seller types, showing a higher proportion of petrol cars and individual sellers in the market.

3.3 Correlation Analysis

Correlation analysis helped in identifying relationships between variables. For example, a negative correlation was observed between the age of the car and its selling price, indicating that older cars tend to sell for less. Conversely, engine capacity showed a positive correlation with the selling price, suggesting that cars with larger engines typically command higher prices.

3.4 Regression Analysis

The model examined the impact of factors like kilometers driven, age, fuel type, seller type, transmission type, mileage, and engine capacity on the selling price. The results indicated that factors such as the age of the car, kilometers driven, and transmission type had significant effects on the selling price. For example, cars with automatic transmission were found to have higher selling prices compared to manual ones. The regression model's R-squared value of 0.5647 suggested that about 56% of the variability in the selling price could be explained by the model.

3.5 Findings

The depreciation of cars with age and increased mileage was evident, aligning with general market expectations. The preference for automatic transmission and larger engine capacities was reflected in their positive impact on selling prices. Surprisingly, the fuel type had a less significant impact than expected, which could be attributed to changing market trends and consumer preferences.

4 Results

After extensive research and work put into the research, we can now present the correlation age, fuel type, mileage and other sections that can affect the selling price of the car. These insightful findings can now be easily shown into graphs for easier explanation of the used car market.

4.1 Age versus Mileage

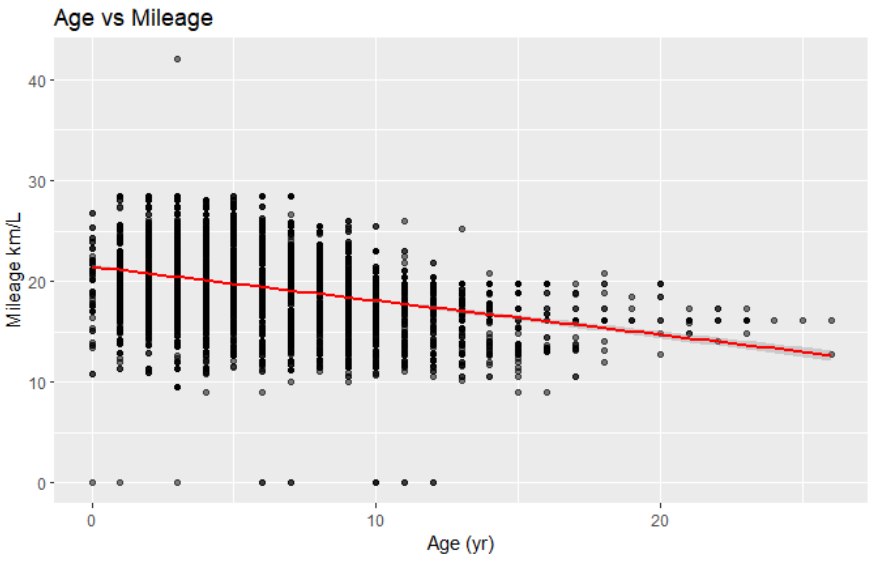

Figure 1: Car Age in comparison to mileage

4.1.1 Relationship Between Age of Car and Mileage Efficiency

The analysis delved into the intricate relationship between the age of a car and its mileage, measured in kilometers per liter. This relationship was visualized through a scatter plot, where each dot distinctly represented an individual car within the extensive dataset. The primary observation from this plot was the noticeable variability in mileage among cars of similar ages, an aspect that warrants a closer examination.

4.1.2 Interpretation of Linear Regression Analysis

A critical component of the analysis was the incorporation of a red line, representing the line of best fit, likely derived from a linear regression model. This line exhibited a downward trajectory, indicative of a negative correlation between the two variables under study. Specifically, as the age of a car increases, there is a discernible decrease in its fuel efficiency. This trend aligns with conventional understandings of vehicle performance degradation over time.

4.1.3 Factors Influencing the Observed Trend

The observed negative correlation between a car's age and its mileage efficiency can be attributed to several factors. Primarily, this trend could stem from the cumulative wear and tear that vehicles experience over time. As cars age, their mechanical components may degrade, leading to diminished efficiency.

Furthermore, technological advancements play a significant role. Older vehicles, typically equipped with dated technology, may not match the fuel efficiency standards of their modern counterparts. This aspect highlights the rapid evolution of automotive technology and its impact on vehicle performance.

4.2 Car Mileage versus Fuel Type

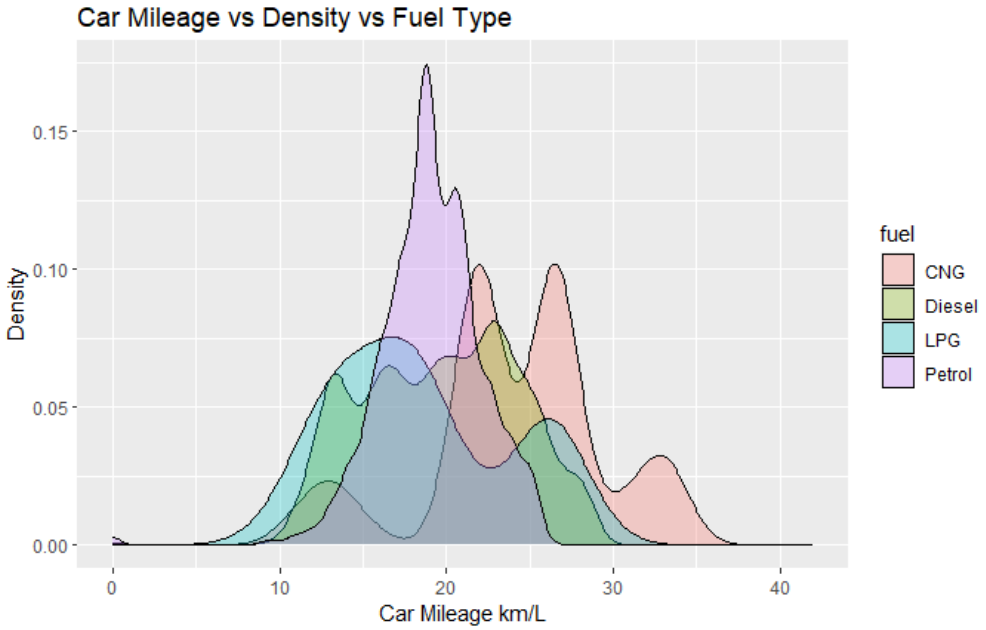

Figure 2: Fuel Type vs Car Mileage

4.2.1 Comparative Analysis of Mileage Efficiency Across Different Fuel Types

The graph under scrutiny delineates the probability distribution of mileage efficiencies for cars segmented by fuel type. The curves, each representing a unique fuel category, serve as a visual tool to compare the concentration and distribution of mileage efficiencies.

4.2.2 CNG Cars: A Case for High Efficiency

The CNG curve, highlighted in blue, peaks within the 25-30 km/L range, which suggests a strong propensity for CNG vehicles to achieve high mileage efficiency. This peak denotes the most probable mileage efficiency for CNG cars and is indicative of the fuel type's overall superior performance in terms of fuel economy within the dataset.

4.2.3 Diesel Vehicles: Diversity in Efficiency

In contrast, the green curve, symbolizing Diesel-fueled cars, exhibits a broad peak. This wide-ranging peak suggests a considerable diversity in mileage efficiencies among Diesel vehicles, albeit with a concentration around 20-25 km/L. Despite the breadth of the curve, it reflects a generally high level of fuel efficiency, underscoring the versatility of Diesel cars in the dataset.

4.2.4 LPG Vehicles: Lower Efficiency with Less Variability

The LPG-associated curve, depicted in red, sharply peaks around the 10 km/L mark. This pronounced peak suggests a narrower distribution of mileage efficiencies, with most LPG vehicles in the dataset clustered around a lower fuel efficiency. The sharpness of the peak underscores a lesser variability in the performance of LPG cars compared to their CNG and Diesel counterparts.

4.2.5 Petrol Cars: Broad Distribution Signifying Variability

The pink curve for Petrol cars shows a much broader and less distinct peak, implying a wide range of mileage efficiencies with a central tendency around 15 km/L. This spread suggests a high degree of

variability in the fuel efficiency of Petrol vehicles, with the peak pointing to a generally lower efficiency when compared to CNG and Diesel.

4.2.6 Overlapping Efficiencies Among Fuel Types

The intersection of the Diesel and Petrol curves around the 15 km/L mileage efficiency signifies an area of commonality, where vehicles powered by both fuel types exhibit similar fuel efficiencies. This overlap presents an interesting facet of the comparative analysis, indicating that while fuel type is a determinant of efficiency, there is a non-negligible range where efficiencies converge.

In synthesis, the graphical analysis reveals that CNG cars are the frontrunners in terms of mileage efficiency within this dataset, closely followed by Diesel. LPG vehicles, while less varied, tend to occupy the lower end of the efficiency spectrum, with Petrol cars displaying a substantial diversity in efficiencies. The depicted trends and overlaps in efficiencies provide valuable insights into the mileage performance of cars based on their fuel type, which is essential for consumers' decision-making and policymakers focusing on energy sustainability in the automotive sector.

4.3 Number of Cars versus Owners versus Fuel Type

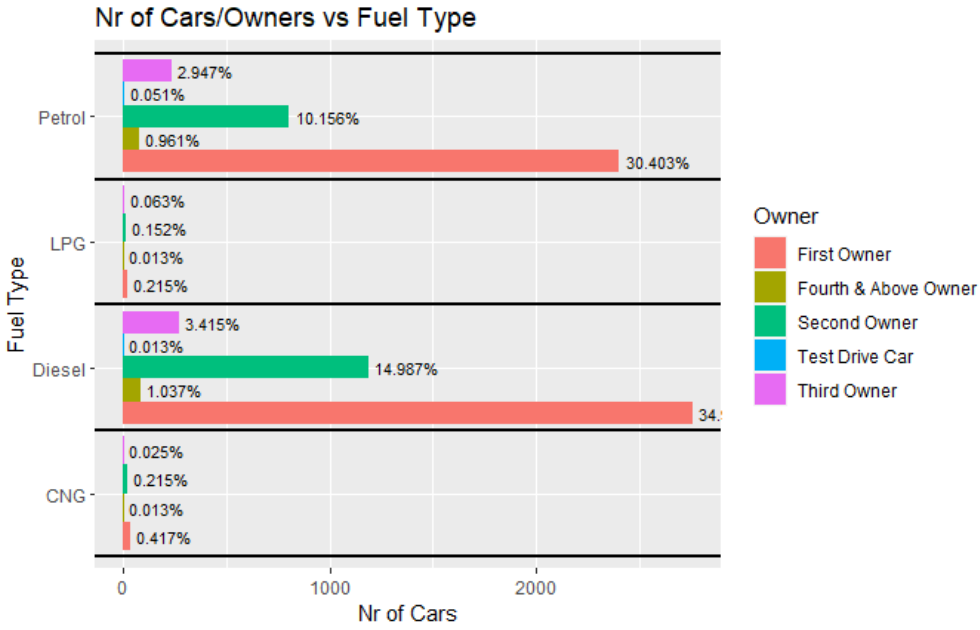

Figure 3: Cars with the appropriated fuel types in correlation with the ownership history

4.3.1 Distribution of Vehicle Ownership by Fuel Type

The analysis focuses on the stratification of vehicle ownership within the used car market, as categorized by fuel type. The stacked horizontal bar chart offers a comprehensive view of this distribution, where the length of each bar corresponds to the total count of vehicles for a given fuel type, and the colors within the bars delineate different ownership categories.

4.3.2 Dominance of Petrol Cars

The dataset prominently features Petrol vehicles, which constitute the majority. Notably, a significant share of these Petrol cars is retained by the original owner. This suggests a trend towards longer-term ownership among Petrol car owners, which could reflect consumer satisfaction or a lower propensity to switch vehicles within this segment.

4.3.3 Diesel Cars and Ownership

Diesel vehicles emerge as the second most prevalent fuel type. Similar to Petrol, a large fraction of Diesel cars are held by first owners, albeit marginally lower in comparison. This might indicate a slightly higher turnover rate for Diesel vehicles or differing market dynamics that influence the ownership lifecycle of these cars.

4.3.4 Lesser Variety in CNG and LPG Cars

CNG and LPG vehicles display a smaller footprint in the dataset. The range of ownership for these fuel types is noticeably narrower, which could be attributed to a smaller market share or a less mature secondary market for these vehicles. This observation warrants further investigation into consumer behavior regarding alternative fuel vehicles.

4.3.5 Insights into Ownership History

First-owner vehicles dominate across all fuel types, particularly among Petrol and Diesel cars. This finding underscores a broader trend of purchasing new or relatively new vehicles within these categories.

Second-owner vehicles, while representing a smaller portion, constitute a significant segment, especially within the Petrol and Diesel fuel types. This reflects a healthy secondary market and could also be indicative of the lifecycle of vehicle ownership where a change typically occurs.

The categories of third owner, fourth and above owner, and test drive cars collectively form a minor component of the dataset. The particularly low prevalence of test drive cars could be due to their limited availability in the used car market or their transitional nature as vehicles that are not often retailed.

4.3.6 Quantitative Analysis of Ownership Proportions

The detailed percentages accompanying each segment provide a granular understanding of ownership distribution. For instance, the precise figure of 30.403% for first-owner Petrol cars and 10.156% for second-owner Petrol cars allows for an accurate assessment of market composition. Such quantification is invaluable for stakeholders seeking insights into consumer trends and ownership patterns in the automotive sector.

4.4 Engine Capacity versus Mileage

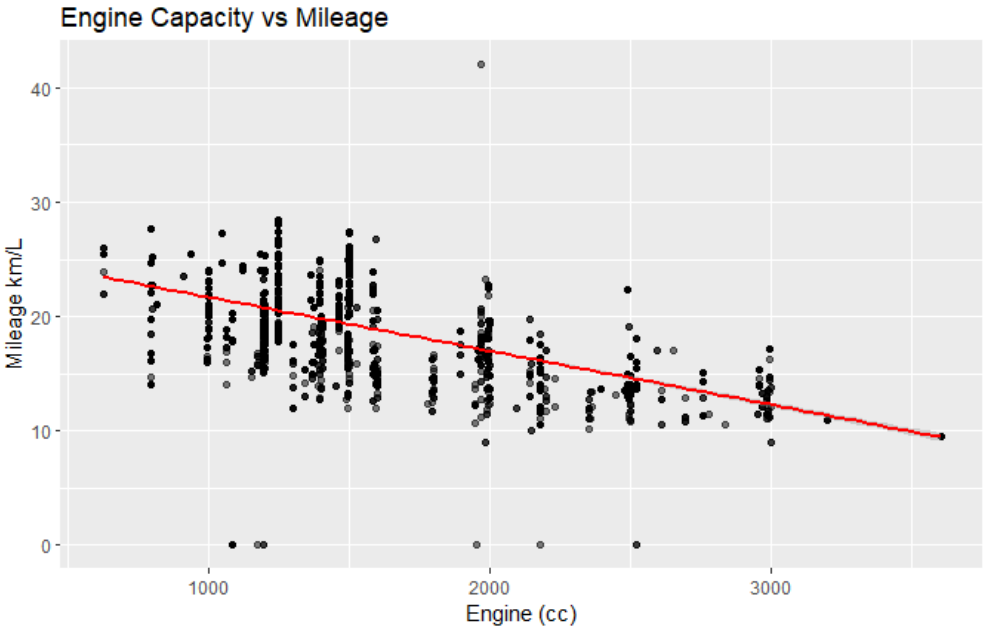

Figure 4: Engine capacity in comparison to its mileage

4.4.1 The Interplay Between Engine Capacity and Mileage Efficiency

This study presents an analysis of the relationship between engine capacity and mileage among a diverse set of vehicles. Each data point on the graph represents an individual car's engine capacity juxtaposed with its corresponding mileage, offering a snapshot of the vehicle's fuel efficiency.

4.4.2 Trend Analysis Through Regression

Central to the analysis is the red line that traverses the plot, symbolizing the regression line of best fit. The declination of this line denotes a negative correlation, providing evidence that larger engine capacities are generally associated with lower mileage. This inverse relationship suggests that, within the confines of this dataset, an increase in engine capacity tends to result in decreased fuel efficiency.

4.4.3 Variability in Mileage Across Engine Capacities

An interesting aspect of the data is the observed variance in mileage, particularly at the lower end of engine capacities. Here, the spread of data points is broad, indicating a heterogeneity in fuel efficiency that diminishes as engine capacity escalates. This pattern implies that smaller engines exhibit a more diverse range of mileage outcomes, whereas larger engines tend to converge towards a narrower band of fuel efficiency.

4.4.4 Efficiency Dynamics of Small Versus Large Engines

The overarching narrative of the graph points to a general trend where vehicles with smaller engines are more fuel-efficient, consuming less fuel per kilometer. Larger engines, as indicated by their positioning on the graph, are less fuel-efficient, nevertheless, the presence of outliers within the dataset challenges this generalization. These exceptions are represented by vehicles that defy the predominant trend—some with larger engines demonstrate high mileage, and similarly, some with smaller engines show unexpectedly low mileage.

4.4.5 Multifaceted Nature of Fuel Efficiency

The anomalies observed, where vehicles deviate from the expected trend, underscore the multifactorial nature of fuel efficiency. Engine capacity, while a significant factor, is not the sole determinant of a vehicle's fuel economy. Other factors such as aerodynamics, weight, transmission type, and even driving habits can influence a car's mileage, contributing to the variations and outliers noted in the data.

4.5 Fuel Types in the Market

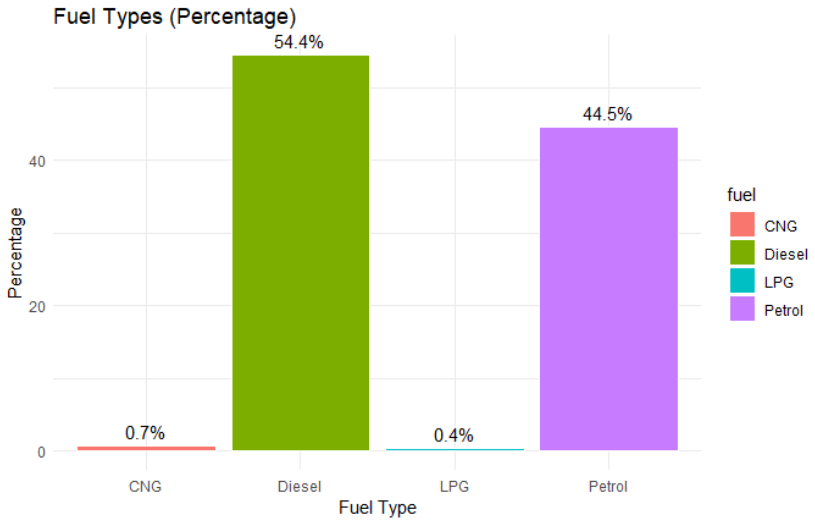

Figure 5: Number of Cars with its appropriate fuel type

4.5.1 Fuel Type Distribution in the Vehicle Dataset

This dataset presents a compelling visual representation of the current landscape of fuel type distribution within a specified vehicle population. The data is succinctly illustrated by a bar chart, where each bar's length and color correspond to the proportion of vehicles running on a particular type of fuel.

4.5.2 Minority Share of Alternative Fuels

A striking feature of the dataset is the nominal representation of Compressed Natural Gas (CNG) vehicles, as denoted by the red bar, accounting for a mere 0.7% of the total. Similarly, the cyan bar representing Liquefied Petroleum Gas (LPG) indicates a fraction even smaller than that of CNG, comprising only 0.4% of the vehicles.

4.5.3 Dominance of Diesel and Petrol

Conversely, the green bar highlights that more than half of the dataset's vehicles, precisely 54.4%, operate on diesel. This is closely followed by petrol, with the purple bar indicating that 44.5% of the cars utilize this fuel type. The prominence of diesel and petrol underscores their status as the predominant fuel choices in the dataset.

4.5.4 Interpreting the Fuel Type Preferences

The predominance of diesel and petrol could be interpreted as a reflection of several underlying factors. Consumer preferences may lean towards these fuel types due to their widespread availability and the perceived reliability of vehicles powered by them. Fuel efficiency and cost also play a crucial role, as diesel and petrol engines traditionally offer a balance of performance and economy that aligns with consumer expectations.

4.5.5 Implications of the Presence of CNG/LPG Vehicles

The marginal representation of CNG and LPG vehicles might suggest a niche adoption, potentially due to the specialized nature of these fuels. It may indicate the presence of heavy-duty commercial vehicles within the dataset, such as trucks and buses, which often utilize CNG or LPG for economic reasons or to meet environmental regulations.

4.6 Number of Seats versus Number of Cars

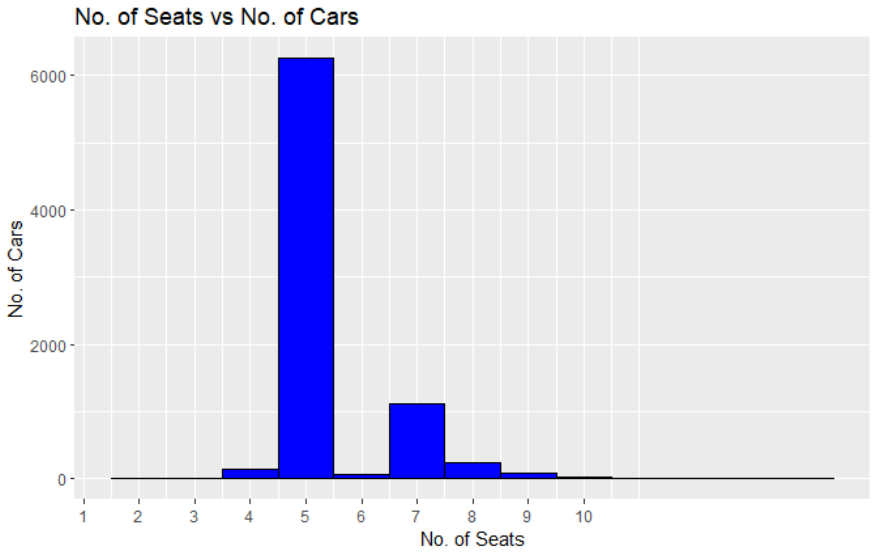

Figure 6: Number of cars with the number of seats

4.6.1 Seating Configuration Preferences in the Automotive Market

The dataset provides a graphical representation of vehicle seating configurations through a bar chart, revealing the prevalence of various seat counts in the current automotive market. Each bar's height corresponds to the number of vehicles with a specific seating capacity, providing a clear visual of consumer preferences and market trends.

4.6.2 Predominance of 5-Seat Vehicles

It is immediately evident that the 5-seat configuration is the most popular, as demonstrated by the tallest blue bar on the graph. This overwhelming majority indicates a market inclination towards this seating arrangement, which is characteristic of a broad spectrum of vehicle types, such as sedans, hatchbacks, and compact SUVs. The preference for 5-seat cars likely reflects their versatility, meeting the needs of the average family size and being conducive to daily commuting and occasional long-distance travel.

4.6.3 Market Position of 7-Seat Vehicles

The graph also indicates a significant but notably smaller proportion of 7-seat vehicles. While they do not match the ubiquity of 5-seaters, the presence of these vehicles highlights a substantial segment of the market that caters to larger families or consumers seeking additional space for passengers and cargo, commonly found in larger SUVs and minivans.

4.6.4 Rarity of Other Seating Configurations

Conversely, the bar chart shows very few cars with 2, 4, 6, 8, 9, or 10 seats, with some configurations not even present. This scarcity points to niche markets, such as two-seater sports cars, luxury coupes, or specialized vehicles designed for specific uses like transport vans or limousines. The minimal presence or absence of these seating configurations suggests that they cater to specific consumer needs and preferences, which do not represent the majority of the market.

4.6.5 Implications for the Automotive Industry

The dominance of 5-seat cars within this dataset is reflective of broader consumer demand and can influence manufacturing and marketing strategies within the automotive industry. It underscores the importance of understanding consumer demographics and their practical needs. The industry must also recognize the smaller but significant demand for vehicles with larger seating capacities, which fulfill a different set of consumer requirements.

The presence of 7-seat vehicles suggests a smaller but notable market for larger family cars or SUVs. The very low numbers or absence of bars for other seat counts suggest that vehicles with these configurations are relatively rare in this dataset(ranging from sports cars to large family carrier).

4.7 Number of Seats versus Selling Price

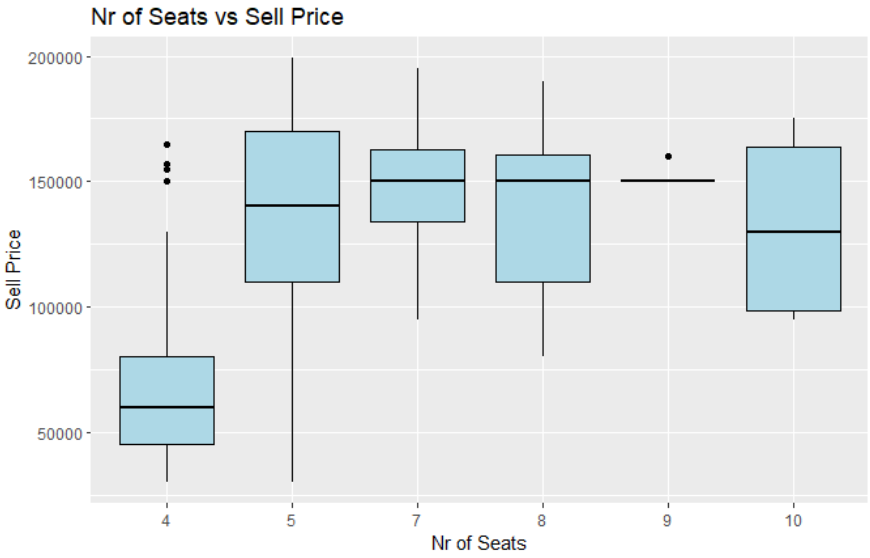

Figure 7: Number of seats in comparison to the selling price of the car

4.7.1 Insights into Selling Price Variability by Seating Capacity

The dataset's box plots provide a vivid depiction of the distribution of car selling prices segmented by seating capacity. Each box plot encapsulates the central tendency and dispersion of selling prices within each category, offering a succinct statistical summary from which we draw several observations.

4.7.2 Price Trends Across Seating Configurations

The analysis reveals that 4-seat cars have the lowest median selling price, complemented by a narrow interquartile range (IQR). This compact IQR signifies less variability in the selling prices, suggesting a market consensus on the value of these vehicles. The positioning of the median, or second quartile (Q2), within the box indicates where the central 50% of data points lie, providing a reliable indicator of the typical selling price for this seating category.

In contrast, 5-seat cars exhibit a higher median selling price than their 4-seat counterparts, coupled with a broader IQR. The extended range implies a greater diversity in the selling prices of 5-seat cars, which could be attributed to a wider variety of makes and models within this segment, catering to different market tiers.

The box plot for 7-seat cars shows a median selling price that mirrors that of 5-seat cars but with a tighter IQR. This reduced variability points to a more consistent market valuation, potentially reflecting a narrower scope of vehicle types within this seating capacity that command similar prices.

The 8-seat category presents an intriguing case, with a median selling price that is lower than that of 7-seat cars and skewed towards the higher end of the price range, as suggested by the median's proximity to the third quartile (Q3). This indicates a clustering of values towards the more expensive end of the spectrum, which may reflect a premium placed on larger capacity vehicles within this particular market niche.

Finally, vehicles with 9 and 10 seats boast the highest median selling prices. Notably, the IQR for 10-seat cars is expansive, denoting a substantial variance in selling prices. This could reflect a market segment that is less uniform, encompassing a range from utilitarian models to luxurious, feature-rich variants.

4.7.3 Considerations Beyond Seating Capacity

While seating capacity provides a structural framework for this analysis, it is imperative to acknowledge that selling prices are influenced by an array of factors. The make and model of the cars, their condition, age, mileage, and the inclusion of additional features all play pivotal roles in shaping the selling price. Furthermore, the presence of outliers—data points that lie beyond the whiskers of the box plot—indicates exceptional cases where selling prices significantly deviate from the norm for a given seat number.

4.7.4 Implications for Stakeholders

Understanding the distribution of selling prices in relation to seating capacity is invaluable for various market stakeholders. For consumers, it can guide purchasing decisions, while for manufacturers and dealers, it informs pricing strategies and inventory selection. The variability and outliers highlighted in the box plots underscore the complexity of the automotive market and the multifaceted nature of vehicle valuation.

4.8 Number of Ownership Types

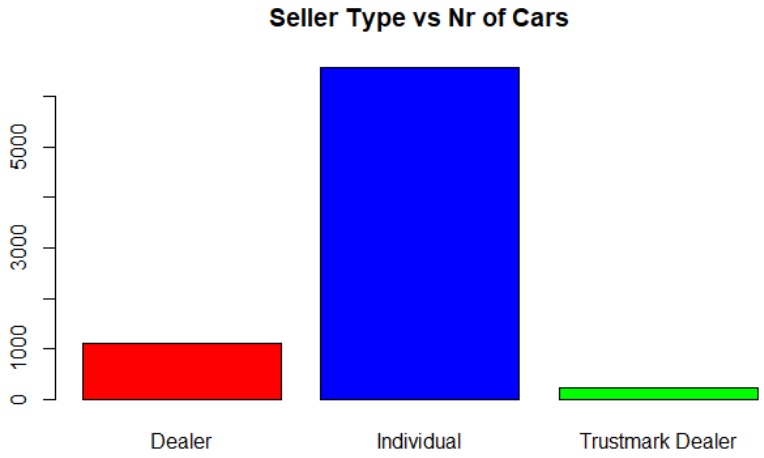

Figure 8: Number and Types of Sellers for each car

4.8.1 Seller Type Dynamics in the Car Sales Market

The analysis employs a bar chart to elucidate the proportions of car sales attributable to different types of sellers within a particular dataset. This visual tool distinctively employs colored bars to signify each seller category, allowing for an immediate comprehension of market trends in terms of sales channels.

4.8.2 Dealer Versus Individual Sales

The data points to a relatively lower number of cars sold by dealers, as indicated by the red bar's modest stature. This suggests that, in the context of the dataset, dealer transactions constitute a smaller segment of the market compared to sales by individuals. The reasons for this could be multifaceted, potentially including the perception of higher prices due to dealer overheads, a preference for direct negotiation with sellers, or a perceived lack of flexibility in the dealer sales process.

In stark contrast, individual sellers, represented by the blue bar, dominate the sales figures. The prominence of this bar underscores a marketplace where private car sales are prevalent. This could reflect a consumer preference for the perceived simplicity and transparency of dealing with individual sellers, as well as the potential for lower prices due to the absence of dealer margins.

4.8.3 The Role of Trustmark Dealers

Trustmark dealers, denoted by the green bar, account for the smallest proportion of sales among the three groups. Trustmark dealers typically offer added assurances such as certified pre-owned cars, which may come at a premium. The modest contribution of Trustmark dealer sales to the overall market could be indicative of a consumer base that prioritizes cost savings over the benefits of certification and warranty that come with purchasing from a Trustmark dealer.

4.8.4 Market Implications

The chart's insights suggest that private car sales may be a more accessible or attractive option for a majority of consumers within this dataset's market. The relatively limited sales by dealers, including Trustmark dealers, could prompt these businesses to investigate the drivers behind consumer behavior. It raises questions about the potential need for competitive pricing strategies, improved customer trust and satisfaction initiatives, or enhanced marketing efforts to better highlight the value proposition offered by dealers and Trustmark dealers alike.

4.9 Odometer Selling Price

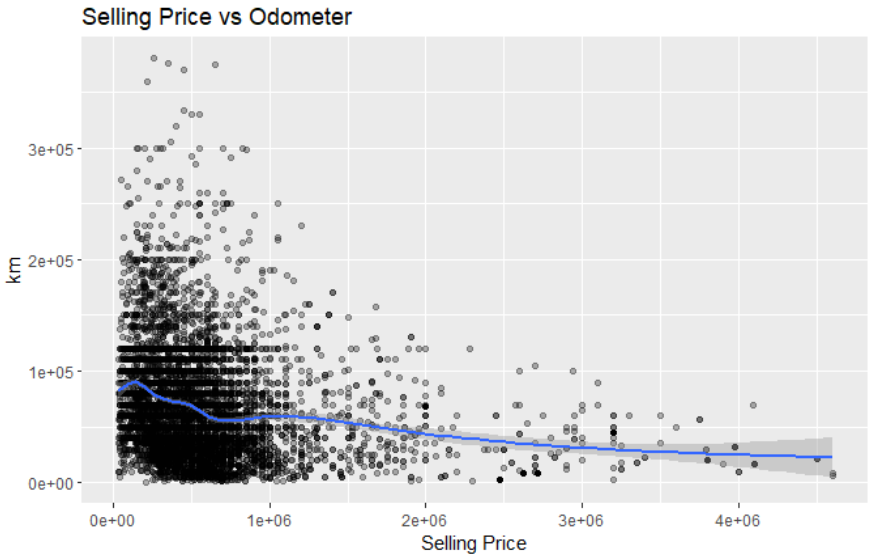

Figure 9: Percentage of Cars with their appropriate odometer

4.9.1 Evaluating the Price-Mileage Correlation in Car Sales

The provided scatter plot serves as a potent analytical tool, enabling a visualization of the correlation between car selling prices and their accumulated mileage as indicated by odometer readings. By representing individual vehicles as discrete data points, the graph facilitates an understanding of how these two crucial variables interact within the marketplace.

4.9.2 Price Depreciation Relative to Mileage

A salient observation from the scatter plot is the inverse relationship between selling price and odometer reading. Generally, as the selling price ascends, the odometer reading descends, aligning with conventional wisdom that higher-priced vehicles tend to be newer or less utilized. This trend is graphically encapsulated by the descending line through the data points, which presumably reflects a statistical mean or median trend in the relationship between the two variables.

4.9.3 Diverse Mileage in the Lower Price Bracket

The graph also indicates a substantial variance in odometer readings among the more affordably priced cars. This diversity could be reflective of a range of factors including, but not limited to, vehicle age, maintenance history, and usage patterns. Such variation suggests that within the lower price echelons, buyers may encounter a broad spectrum of vehicle conditions and histories.

4.9.4 Sparse Data at Premium Prices

The scarcity of data points at the upper echelons of the selling price spectrum suggests fewer transactions occur in this segment, which is often characteristic of luxury or newer model cars. These vehicles tend to exhibit lower odometer readings, potentially due to a combination of less frequent use and a higher retention of value over time.

4.9.5 Outliers and Their Implications

Outliers within the dataset—those points divergent from the general trend—merit particular attention. These anomalies could represent exceptional cases, such as vintage cars that command high selling

prices regardless of mileage due to their rarity or collector's value, or alternatively, high-mileage recent models that have depreciated rapidly due to intensive use.

4.9.6 Logarithmic Scale Considerations

The application of logarithmic scales to both axes is a deliberate choice to accommodate the expansive range of values and to enhance the interpretability of the relationship, which may not be linear across the entire spectrum of data. Logarithmic scaling is particularly adept at elucidating patterns that adhere to exponential growth or decay, making it a suitable method for this analysis.

4.10 Selling Price versus Mileage

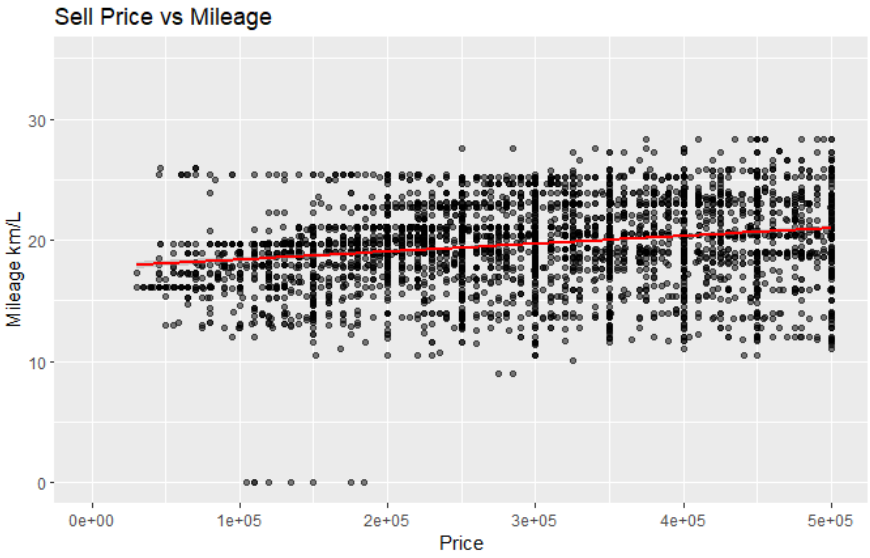

Figure 10: Selling Price comparison to the car mileage

4.10.1 Analyzing the Correlation Between Selling Price and Mileage in Cars

The scatter plot that was analyzed provides a revealing snapshot of the relationship between the selling prices of cars and their fuel efficiency, quantified in terms of mileage. By plotting individual cars as dots, with their selling price and mileage as coordinates, we can discern patterns and trends in how these two critical factors interrelate in the automotive market.

4.10.2 The Subtle Link Between Price and Fuel Efficiency

A particularly striking aspect of the plot is the relatively flat slope of the red line that cuts through the data points. This line, which represents the trend or average relationship between the two variables, indicates only a weak correlation between selling price and mileage. Contrary to what might be a more intuitive expectation, the fuel efficiency of a car does not appear to significantly fluctuate with changes in its selling price.

4.10.3 Observations and Implications

From the scatter plot, several key observations can be made:

- Variability Across Price Ranges: The graph displays a wide range of mileage values across all selling price categories. This suggests that a car's selling price is not a reliable predictor of its fuel efficiency. Cars at similar price points can exhibit vastly different mileage figures, implying that factors other than price play a significant role in determining a car's fuel efficiency.

- Absence of a Distinct Pattern: There isn't a conspicuous trend where higher-priced cars consistently offer better or worse mileage. The flatness of the trend line reinforces the idea that the relationship between these two variables is not strongly pronounced.

- Presence of Outliers: Some data points stand apart from the main cluster, indicating the existence of cars with exceptionally high or low mileage for their respective selling prices. These outliers highlight the diversity within the car market and suggest that unique characteristics of certain vehicles, such as make, model, or specific features, might significantly influence their fuel efficiency independently of price.

4.10.4 Broader Market Implications

The scatter plot's overall narrative suggests that while there might be a slight inclination for higher-priced cars to have marginally better mileage, this is not a robust or consistent trend. The market presents numerous exceptions, underscoring the multifaceted nature of car valuation and fuel efficiency determinants. For consumers, this means that relying solely on price as an indicator of fuel efficiency may be misguided. For manufacturers and dealers, it emphasizes the need to consider a broader range of attributes when pricing vehicles or highlighting their efficiency.

4.11 Transmission Car Types

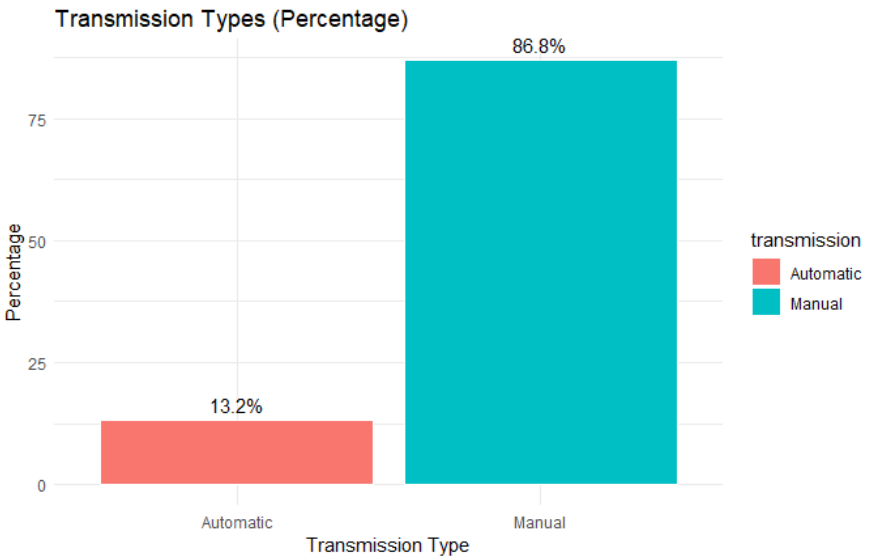

Figure 11: Percentage of the cars with the manual and automatic transmission

4.11.1 Transmission Type Preferences in the Automotive Market

The bar chart at hand offers a clear depiction of the distribution of cars by transmission type within a specific dataset. By categorizing the vehicles into two distinct groups, Automatic and Manual, and representing these groups through colored bars, we gain valuable insights into consumer preferences and market trends in transmission types.

4.11.2 Dominance of Manual Transmission

A striking aspect of the chart is the overwhelming majority of manual transmission cars, as evidenced by the blue bar covering 86.8% of the dataset. This prevalence suggests a strong market inclination towards manual transmissions. Several reasons could underpin this trend:

- Cost Factors: Manual transmission vehicles are often more affordable than their automatic counterparts, both in terms of initial purchase price and maintenance costs. This cost-effectiveness could be a driving factor behind their popularity.
- Fuel Efficiency: Historically, manual transmissions have been perceived to offer better fuel efficiency compared to automatic transmissions. Although technological advancements have narrowed this gap, the perception might still influence consumer choices.
- Driver Preference and Control: In many regions and among certain driver demographics, manual transmissions are preferred for the greater control and engagement they offer while driving.
- Market Specific Dynamics: The dataset may be representative of a market where manual cars are more accessible, or there's a cultural inclination towards manual driving.

4.11.3 Lesser Representation of Automatic Transmission

On the other hand, the red bar representing automatic transmission cars constitutes only 13.2% of the dataset. This smaller proportion could be attributed to various factors:

- Higher Costs: Automatic cars generally come with a higher price tag and potentially greater maintenance expenses, which might deter certain segments of the market.

- Perceived Complexity: In regions where manual transmissions are the norm, automatic transmissions might be perceived as more complex or less reliable, affecting their popularity.
- Driving Culture and Experience: The preference for manual or automatic transmissions can also be influenced by the prevailing driving culture and the typical driving experiences in a region.

4.11.4 Market Implications and Trends

The predominance of manual transmissions in the dataset indicates a market segment where cost-effectiveness, fuel efficiency, and driving engagement are likely highly valued. For automotive manufacturers and dealers, these insights are crucial for tailoring their product offerings and marketing strategies to align with consumer preferences. It also hints at potential shifts in market dynamics, as technological advancements in automatic transmissions continue to evolve.

4.12 Transmission Types impact on the price

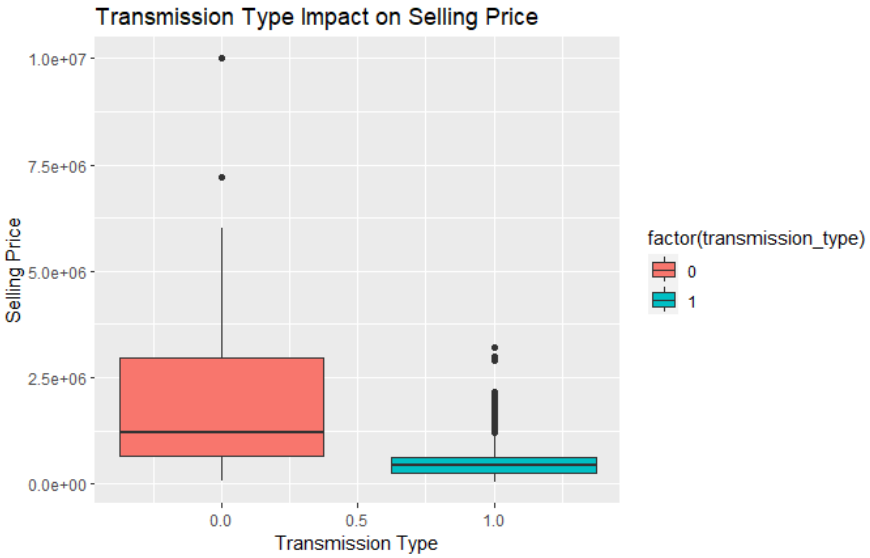

Figure 12: Transmission types impacting the price

4.12.1 Influence of Transmission Type on Car Selling Prices

The box plot analysis offers a deep dive into how transmission types impact the selling prices of cars. By categorizing the data into two distinct groups, represented by red and blue boxes, the plot provides a clear visual representation of the price distribution associated with each transmission type.

4.12.2 Variability and Median Prices of Automatic Transmission Cars

The red box, likely representing automatic transmission cars given standard coding conventions, exhibits a wider range of selling prices. This is evidenced by the taller stature of the box, suggesting a greater variability in the prices at which automatic cars are sold. Additionally, the median price, indicated by the line within the red box, is notably higher than its counterpart. This could reflect the premium often associated with automatic transmission vehicles, which are generally perceived as offering more convenience and advanced technology.

4.12.3 Compact Price Range of Manual Transmission Cars

Conversely, the blue box, presumably depicting manual transmission cars, shows a more compact range of selling prices. The lower median price within this box indicates that manual transmission cars generally sell at lower prices compared to automatic ones. This could be attributed to various factors, such as lower manufacturing costs, a market preference for more affordable vehicles, or the perception of manual cars as being more basic in terms of features.

4.12.4 The Significance of Outliers

Outliers in the graph, particularly in the automatic transmission category, are indicative of exceptional cases where cars have sold for prices significantly higher than the general trend. These could represent luxury or high-performance automatic vehicles, which command a premium due to their brand, features, or performance capabilities.

4.12.5 Price Consistency Among Manual Transmission Cars

The tighter distribution of selling prices for manual transmission cars, as shown by the more compact blue box, suggests less variability. This could be indicative of a more uniform market segment, where

manual cars are clustered around a specific price range, likely due to similarities in vehicle specifications and market positioning.

4.12.6 Implications for the Automotive Market

This box plot underscores the significant impact of transmission type on the selling price of cars. For stakeholders in the automotive industry, these insights are vital for understanding market dynamics and consumer preferences. The data highlights the need for targeted strategies when pricing and marketing vehicles with different transmission types, catering to the distinct segments within the market.

5 Models

Model Type	Mean Square Error	R-Squared
Multiple Linear Regression	553,200	%56.42
Lasso Regression	317,173,992,855	%56.61
Random Forest	46,163,273,873	%93.42

5.1 Multiple Linear Regression (MLR)

- The MLR model shows a Residual Standard Error (RSE) of 553,200 and an R-squared (R^2) of 56.47%.
- The moderate R^2 indicates that MLR captures a reasonable proportion of the variance in the selling price, but there's still a significant unexplained part. The RSE provides an estimate of the standard deviation of the residuals and is relatively high, suggesting that prediction errors can be significant.
- MLR's performance is decent but not exceptional. It suggests that while the linear relationship it assumes holds to some extent, there are other factors and possibly non-linear relationships that MLR cannot capture. This model could be more informative if combined with rigorous feature selection and possibly polynomial terms to capture non-linear effects.

5.2 Random Forest

- Random Forest has a high Mean of Squared Residuals (MSE) at 46,163,273,873, but an impressive R^2 of 93.42%.
- The high R^2 value indicates a strong fit to the data, meaning the model explains a large portion of the variance in selling prices. However, the high MSE suggests that when the model does ERR, those errors can be quite large.
- The Random Forest model's high R^2 value is promising, showing its strength in capturing complex, non-linear relationships in the data. However, the high MSE raises questions about the model's prediction accuracy for individual predictions. It could be a result of overfitting or the scale of your target variable. Fine-tuning the model parameters and implementing more robust validation techniques like cross-validation might help improve its performance.

5.3 Lasso Regression

- Lasso Regression's MSE is extremely high at 317,173,992,855, with an R^2 of 56.62%.
- The R^2 is similar to MLR, indicating a moderate fit to the data. However, the extremely high MSE is a major concern, suggesting that the model's predictions are often far off the mark.
- The Lasso model's poor performance, as indicated by the high MSE, could be due to several factors, such as inappropriate lambda value leading to over-regularization and underfitting. Lasso is designed to perform feature selection, but if critical features are penalized too heavily, the model's predictive power diminishes. A careful review of the regularization parameter and feature selection process is necessary. It's also worth exploring if the data and the relationships it holds are suitable for a linear model like Lasso.

6 Discussion

The findings from the data analysis offer insightful perspectives on the dynamics of the used car market. This section discusses the implications of these findings, their alignment with existing literature, and potential applications.

6.1 Interpretation of Findings

The significant negative correlation between a car's age and its selling price confirms the widely accepted notion of depreciation in the automotive industry. This aligns with the findings of Borthakur (2023)[Bor23], who noted a similar trend in their study. The preference for automatic transmission vehicles, reflected in their higher resale values, corroborates with Lin (2022)[Hua23] study, which highlighted the growing demand for convenience in driving. The less significant impact of fuel type on the selling price, contrary to Knittel (2009)[BKZ09] findings, may indicate a shift in consumer preferences, possibly due to environmental concerns or advancements in fuel efficiency across different engine types.

6.2 Practical Implications

- For buyers: The results suggest that consumers looking for value purchases should consider older and manual transmission vehicles. Understanding these trends can help buyers make more informed decisions.
- For sellers: Sellers can leverage this information to price their vehicles competitively, focusing on attributes like transmission type and engine capacity to attract potential buyers.
- For market analysts: The insights can guide market analysts in predicting future trends, particularly concerning the shifting preferences towards automatic transmission and environmental considerations.

The study's findings largely corroborate the existing literature on used car valuation, especially regarding depreciation with age and mileage. However, the diminished importance of fuel type presents a deviation from some earlier studies, reflecting the dynamic nature of consumer preferences and market conditions.

6.3 Limitations and Future Research

The primary limitation of this study is its reliance on data from a single source, which may not capture the full spectrum of the used car market. Future research could expand the analysis to include multiple datasets from different regions or time periods. Another limitation is the exclusion of certain variables, such as car brand and model, which could have a significant impact on the selling price. Incorporating these factors could provide a more detailed understanding of pricing dynamics. The evolving nature of the automotive industry, with emerging trends like electric vehicles and increased environmental consciousness, calls for ongoing research to stay abreast of how these developments influence the used car market.

7 Conclusion

This research has presented a detailed analysis of the used car market, focusing on the factors that influence the selling price of used cars. Through the application of various statistical methods and data visualization techniques, the study has highlighted several key insights into the dynamics of car valuation.

The analysis revealed that the age and mileage of a car significantly impact its selling price, confirming the traditional view of vehicle depreciation over time. Additionally, the preference for automatic transmission and larger engine capacities were found to have a notable influence on a car's resale value. Contrary to some earlier studies, the type of fuel used by the car was not as significant a factor in determining its selling price, suggesting a shift in consumer preferences and market trends.

These findings have important implications for various stakeholders in the used car market. Buyers can use this information to make more informed decisions, focusing on factors that will ensure the best value for their investment. Sellers can better understand how to price their vehicles and highlight certain features to attract buyers. Market analysts can leverage these insights to predict future trends and advise on strategic market positioning.

However, the study is not without its limitations. The reliance on a single data set and the exclusion of certain variables such as brand and model may limit the comprehensiveness of the findings. As the automotive industry continues to evolve, further research incorporating these factors and emerging trends like electric vehicles will be essential to provide a more complete understanding of the used car market.

In conclusion, this research offers valuable insights into the used car market, shedding light on the factors that influence car valuation. The findings contribute to a deeper understanding of consumer behavior and market dynamics, providing a foundation for future research in this area.

References

[BKZ09] Meghan Busse, Christopher Knittel, and Florian Zettelmeyer. Pain at the pump: how gasoline prices affect automobile purchasing in new and used markets. 03 2009.

[Bor23] Pragyan Borthakur. Evolution of car purchasing behaviour and the reasons behind it among indian consumers: A comprehensive analysis from 2010 to present. 07 2023.

[Esz23] Maklári Eszter. Economic comparison between conventionally powered and electric cars. *Acta Academiae Beregsasiensis. Economics*, pages 75–85, 09 2023.

[GH00] Guillaume Gaulier and Séverine Haller. The Convergence of Automobile Prices in the European Union: an Empirical Analysis for the Period 1993-1999. (2000-14), November 2000.

[Hua23] Zhiqiu Huang. The transaction price prediction of second-hand cars based on model fusion. *Applied and Computational Engineering*, 6:820–830, 06 2023.

GitHub link or https://github.com/Tesla0Maximum/data-analysis/