**Exploring Car Price Through Analytical Insight**

Aina Syazzween Suraya Binti Abdul Latif (B032010027), Thiveya A/P Mahendran (B032010290), Nursyaza Nisa Binti Arfarizal (B032010244), Nur Syuhada Binti Azhar (B032010378)

**1.0 Executive Summary**

Exploring car price through analytical insight project aimed to investigate the factors influencing car prices in the market. Through a comprehensive analysis of car attributes such as make, model, year, mileage, engine horsepower, and transmission type, several key findings emerged. Factors such as make, model, year, mileage, and engine horsepower are considered as significant roles in determining car prices. The study highlighted the impact of specific car features, including advanced safety systems, and luxury enhancements, on pricing dynamics. Furthermore, the analysis provided valuable insights into the preferences and priorities of car buyers, shedding light on the factors that drive perceived value. The findings contribute to a better understanding of the complex interactions between car attributes, market dynamics, and consumer behaviour. This knowledge empowers stakeholders in the automotive industry to make data-driven decisions, optimize pricing strategies, and enhance their competitive edge. By

leveraging these insights, buyers and sellers can make informed choices, while industry professionals can refine their marketing efforts and product offerings. Overall, this car price analysis project delivers valuable insights that foster improved pricing accuracy, informed decision-making, and enhanced competitiveness in the automotive market.

**2.0 Introduction**

Exploring car price through analytical insight is a process of studying and understanding the factors that influence the prices of cars in the market. It involves examining various features and attributes of cars to determine their impact on pricing. The goal of car price analysis is to develop models or insights that can accurately predict or explain the prices of cars based on these factors.

Analysing car prices can be valuable for various stakeholders, including car buyers, sellers, manufacturers, and insurers. For car buyers, understanding the factors that contribute to car prices can help make informed purchasing decisions and negotiate better deals. Car sellers can benefit from pricing their vehicles competitively based on market trends and customer preferences. Car manufacturers can gain insights into the pricing dynamics of different models and adjust their production and marketing strategies accordingly. Insurers can assess the risk associated with insuring different car models based on their prices and features.

Analysing car retail price typically involves examining a range of features and variables that influence car prices. Some common factors considered in car price analysis include the brand, model, year of manufacture, mileage, engine size, fuel efficiency, transmission type, number of seats, vehicle condition, and additional features such as safety equipment, infotainment systems, and luxury enhancements.

Analytical techniques such as regression analysis, machine learning algorithms, and statistical modelling can be employed to analyse historical car sales data and uncover relationships between the features and car prices. These models then for instance, can then be used to predict car prices for new or used vehicles, estimate the value of a car in the market, or identify overpriced or under-priced cars.

Exploring car price through analytical insight can provide valuable insights into market trends, price fluctuations, and the value associated with different car features. It enables stakeholders to make data-driven decisions and optimize their strategies related to car pricing, sales, and purchasing.

Overall, exploring car price through analytical insight plays a significant role in the automotive industry, helping stakeholders understand and navigate the complex factors that drive car prices and enabling them to make informed decisions in a dynamic market. This analytical also helps buyers decide on purchases wisely and bargain better car deals.

Objectives:

1. To analyse the link between automotive features and MSRP to assist in identifying the primary elements that influence price decisions.
2. To assess consumer preferences and understand which features are most valued by car buyers.
3. To evaluate the impact of market trends, such as changes in fuel prices, economic conditions, or technological advancements, on car retail prices.
4. To provide insights for forecasting and pricing optimization purposes.

Problem Statements:

1. The retail pricing of cars is influenced by numerous factors, including make, model, age, condition, mileage, location, and market dynamics.
2. Car manufacturers require insights into the key drivers of car prices to set competitive pricing strategies.
3. There is a need to collect and analyse a comprehensive dataset of car listings, including car attributes, historical sales data, geographic data, and market trends.

Critical literature reviews:

Shiqi Ou, W. Li, Jie Li, Zhenhong Lin, Xin He, Jessey Bouchard, and S. Przesmitzki (2020) attempted this research, which aims to identify common car price trends in the Chinese market by measuring statistical connections between important vehicle elements ranging from intrinsic powertrain systems to extrinsic market positioning. Almost all passenger car models sold from 2013 to 2019 are represented in the data samples. After evaluating several statistical approaches, the best one was determined to be a log-transformation variation of the multinomial linear regression model, and the goodness of fit demonstrates that this model can provide consistent predictions, which were tested using 2019 market data. The pricing and primary performance aspects of SUVs/crossovers are comparable to those of sedans, according to the findings.

A picture containing text, diagram, screenshot, line

Description automatically generated

Figure 2.1: Distributions of vehicle model Miles per Gallon (MPG) by year in 2013–2018; (b) distributions of vehicle model MPG by vehicle brand in 2013–2018.

Damilola Felix Arawomo and Augustine C. Osigwe (2016) was approach this study where it investigated the relationship between gasoline use, automobile attributes, and pricing in Ibadan. It also determined if automobile customers underestimate the importance of fuel savings. A sample of 600 automobile owners from Ibadan's six major universities was used. The hedonic automobile pricing function was calculated using OLS and IV estimation approaches in addition to the descriptive analysis. The IV estimate results were favoured and analysed. According to empirical research, respondents do not underestimate the importance of fuel savings. More specifically, 98.8% of respondents said they would spend more naira to acquire a fuel-efficient automobile if the extra naira is less than the current value of the greater future cost of gasoline for an inefficient car. Furthermore, when asked to list the most essential criteria to consider when purchasing a car, fuel economy was in third place. The estimated hedonic price model revealed that the influence of fuel consumption on automobile price was strongly negative, meaning that respondents were prepared to pay less for a car that was inefficient in terms of fuel consumption. Furthermore, security and luxury automobile features were shown to raise car prices. The report advised that vehicle market players conduct advocacy and orientations to improve users' and potential purchasers' awareness of car pricing drivers. Aside from that, if all other stated characteristics stay constant, the price of a Japanese midsize sedan is 62% greater than that of a Chinese midsize sedan, with European midsize automobiles having the highest total pricing. (3) The extra cost of gasoline use varies depending on vehicle type and fuel economy. For example, going from 30 to 50 MPG raises the car price by $119 for a Chinese brand sedan and $69 for a Chinese brand SUV.

Yixiang Zhang, Juan Li, and Wenwen Tao (2021) were approached to employ the hedonic pricing model (HPM) in this study to investigate consumers' WTP for various air conditioner (AC) features. Sales statistics for 1459 ACs were gathered from a prominent Chinese e-commerce site. Platform factors have no significant influence on pricing, according to the findings. While the premium for energy-efficiency labelling was 12.4%, the premium for brands was up to 59.4%. Meanwhile, propensity score matching (PSM) was used to investigate energy-efficiency labelling premiums; PSM can overcome the problem of selection bias and has advantages over typical HPM. The findings revealed a 19.4% premium between energy-efficiency levels 1 and 2, and a 21.5% premium between levels 1 and 3. Aside from that, if all other stated characteristics stay constant, the price of a Japanese midsize sedan is 62% greater than that of a Chinese midsize sedan, with European midsize automobiles having the highest total pricing. (3) The extra cost of gasoline use varies depending on vehicle type and fuel economy. For example, going from 30 to 50 MPG raises the car price by $119 for a Chinese brand sedan and $69 for a Chinese brand SUV.

P. Bansal, K. Kockelman (2017) were approached to a study which examine the existing vehicle ownership models for India and analyses the findings of nine expert interviews conducted to get insights into Indians' travel patterns and vehicle preferences. According to experts, vehicle price, fuel efficiency, and brand (in decreasing order of significance) are the most important variables in Indians' automobile purchasing decisions. Using Census 2011 data, this study also approximated household automobile ownership levels throughout India's 35 states. The findings imply that states with a greater number of computer-owning families and a higher proportion of households residing in rural regions with bigger household size are more likely to possess a car.

Jang, S., & Choi, J. (2021) were approaching a study about consumer attributes that act crucial roles for the fast market adoption of electric vehicles. In this study, we looked at customer preferences for EV and traditional ICEV features. They collected data by conducting 350 face-to-face discrete choice surveys of people of urban regions in Korea, which were administered by Gallup Korea using conjoint surveys. They used a mixed logit model to predict consumer preferences for EV features, including reference points, in order to empirically validate the prospect theory. They also ran market share simulations by distinguishing the attribute levels while accounting for RP reliance to forecast future EV market shares. Fuel efficiency, purchase price, distance to charging station, driving range, charging time, and autonomous driving were all shown to be statistically significant by the model.

**3.0 Problem Formulation**

**3.1 Data Wrangling**

Data wrangling, also known as data munging or data preprocessing, involves the process of cleaning, transforming, and preparing raw data for analysis. Its purpose is to convert data from its initial format into a more suitable format that can be easily analysed, interpreted, and utilized for various purposes. The data wrangling process typically includes several steps which are data collection, data assessment, data cleaning, data transformation, data integration, data validation, data documentation and data storage.

**3.2 Data Gathering**A screenshot of a computer

Description automatically generated with medium confidence

Figure 3.1: Car\_data.csv

A picture containing text, screenshot, font, number

Description automatically generated

Figure 3.2: Variable of the dataset and its descriptions

Firstly, we performed data gathering by collecting data from Kaggle. This car\_data.csv in Figure 3.1 is a US dataset containing 16 columns with 11914 rows as in Figure 3.2. This dataset combines multiple data types such as object, integer, and float. The car data is extracted from year 1990 to year 2017 including the cars from Chevrolet, Ford, Volkswagen, Chrysler, Mazda, Nissan, Volvo, Mercedes, Mitsubishi, Ferrari, BMW, Toyota, FIAT, Maybach, Saab, Porsche, Hyundai, GMC, Honda, Plymouth, Suzuki, Ford, Oldsmobile, Cadillac, Kia, Bentley, Dodge, Lamborghini, Lincoln, Subaru, Land Rover, Lotus, Buick, Lexus, Infiniti, Pontiac, Maserati, HUMMER, Acura, and Audi. The dataset is utilized for analysing car prices across various manufacturers and models, considering different features. Additionally, it enables users to predict car prices by leveraging existing data on fuel usage and incorporating web-based features. This facilitates the determination and prediction of car prices based on the available information.

**3.3 Data Inspection**

Once the process of data collection is completed, the subsequent crucial step is data inspection. This entails a comprehensive evaluation of the data to gain an understanding of its structure, format, and overall quality. During this phase, the primary objective is to identify and assess various aspects that could potentially impact the subsequent analysis.

Furthermore, data inspection entails a meticulous review of the data quality, aiming to uncover any data anomalies or irregularities that may have arisen during the data collection process. This includes identifying missing values, which can significantly impact the validity and reliability of subsequent analyses. Additionally, the identification of outliers is crucial as they can skew statistical measures and potentially indicate errors or unique observations requiring further investigation.

By undertaking a comprehensive data inspection, analysts and researchers can ensure the reliability and robustness of the collected data. This critical step lays the foundation for subsequent data cleaning, transformation, and analysis processes, ultimately contributing to the generation of accurate and meaningful insights.

A picture containing text, screenshot, font, number

Description automatically generated

Figure 3.3: Display Statistical Summary

Figure 3.3 shows .describe() function which is a commonly used method in data analysis that provides descriptive statistics for numerical columns in a DataFrame. This summary provides a quick overview of the central tendency, spread, and distribution of the numerical data in the DataFrame, allowing for initial insights into the dataset. It helps identify potential outliers, skewed distributions, and the overall range of values for each column.

A screenshot of a computer

Description automatically generated

Figure 3.4: Checking of Missing Values

As we can see in Figure 3.4 coding, there are 5 columns that have missing values which are Market Category, Engine HP, Engine Cylinder, Number of Doors, and Engine Fuel Type. To handle these values, data preprocessing will be performed to clear out the missing values.

A picture containing text, screenshot, font

Description automatically generated

A picture containing diagram, line, plot, screenshot

Description automatically generated

Figure 3.5: Detecting of outliers.

In detecting outliers, boxplot is used to provide valuable insight into data distribution. The outliers are represented by individual data points that are located beyond the whiskers, which are the lines extending from the boxes. It can be seen there are outliers present in numerical columns of the dataset.

**3.4 Data Cleaning**

After inspecting the data, data cleaning is performed to resolve the quality issues through tasks such as handling missing values. Before handling missing values, it is important to take note of the data types of the column of missing values as the way to handle these values differs. For categorical columns such as Market Category and Engine Fuel Type, the missing values are replaced with the mode of the column while the missing values for the remaining numerical columns are replaced with the mean of the column. This ensures the maintenance of the overall distribution and characteristics of the data.

A screen shot of a computer

Description automatically generated with low confidence

Figure 3.6: Replacing missing values with mode and mean.

A screenshot of a computer program

Description automatically generated with medium confidence

Figure 3.7: Standardize format.

In Figure 3.7, Standard Scalar is used to standardize and normalize numerical input variables. Standardizing the format refers to ensuring consistency and uniformity in the representation of data across different variables or fields within a dataset. This process involves transforming data into a common format or structure, typically following predefined guidelines or standards. Standardizing the format has several benefits, including facilitating data integration, improving data quality, and enhancing data analysis and processing efficiency.

**3.5 Data Transformation**

In the data transformation, we restructured and reformatted the data to make it suitable for analysis. This can involve tasks like merging datasets, reshaping data structures, creating new variables, or deriving features. In this analysis we combined the columns of “highway MPG” and “city MPG” to form a combined MPG.

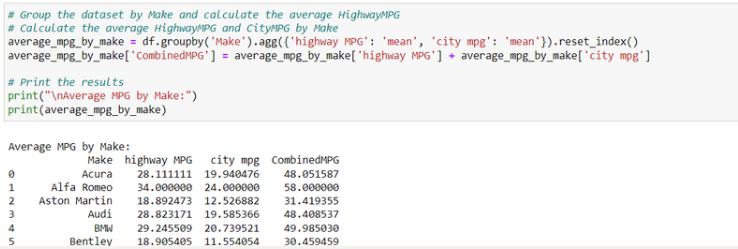


Figure 3.8: Reshaping columns

Data wrangling requires a combination of technical skills, such as programming, data manipulation, and data cleaning techniques, as well as domain knowledge to understand the context and meaning of the data. Effective data wrangling is crucial for ensuring data quality, reliability, and consistency, which are essential for accurate analysis, modelling, and decision-making.

**3.6 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a data analysis technique that examines and comprehends the properties, patterns, and relationships inherent in a dataset. It entails a set of techniques and visualizations that summarize, visualize, and analyze the data, revealing its underlying structure and features. Before undertaking more advanced analysis or modelling, the fundamental purpose of EDA is to obtain a better understanding of the data. It aids in the identification of patterns, trends, anomalies, and probable links between variables, allowing data scientists and analysts to develop hypotheses, confirm assumptions, and make informed decisions.

A screen shot of a computer code

Description automatically generated with low confidence

Figure 3.9: The code of visualizing a horizontal chart for number of cars for each company.

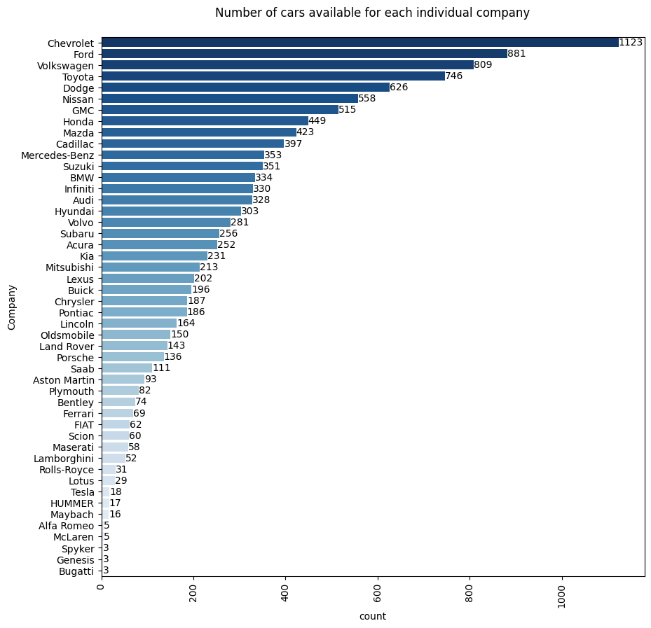


Figure 3.10: Number of cars available for each individual company

Figure 3.9 utilizes the seaborn and matplotlib libraries to create a horizontal bar chart showing the number of cars available for each individual company in the dataset. Figure 3.10 is a visualization that provides a quick and intuitive overview of the distribution of cars among different companies, allowing viewers to identify which companies have a larger or smaller presence in the dataset. The bar labels on the side of each bar provide the exact count for each company.

Figure 3.11: Percentage of vehicle sizes available in the dataset.A blue pie chart with white text

Description automatically generated with low confidence

Figure 3.11 shows a pie chart visualization representing the count and percentage distribution of different vehicle sizes in a dataset. The chart helps to visually communicate the relative proportions of each vehicle size category and allows for easy comparison. By utilizing colors and text labels, the chart provides an intuitive representation of the data, enabling viewers to quickly grasp the distribution of vehicle sizes and potentially gain insights into trends or patterns within the dataset.

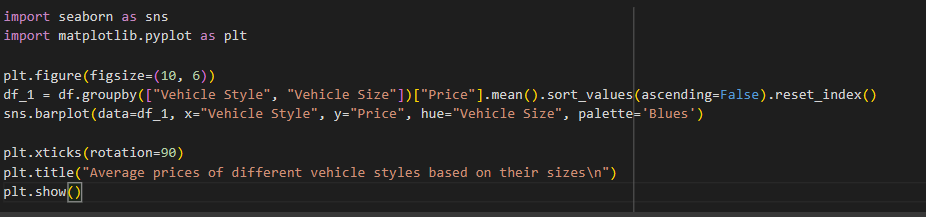


Figure 3.12: The code to create bar plot for average prices of different vehicle styles based on sizes.

A graph of different types of vehicles

Description automatically generated with low confidence

Figure 3.13: Average prices of different vehicle styles based on their sizes.

The provided code on the figure 3.12 generates a bar plot that displays the average prices of different vehicle styles based on their sizes. The plot in figure 3.13 is useful for manufacturers as it provides valuable insights into pricing trends across various vehicle styles and sizes. By examining the chart, manufacturers can identify which combinations of vehicle styles and sizes tend to command higher average prices. This information can guide their product development and pricing strategies, helping them understand customer preferences and market demand. Additionally, manufacturers can use this insight to position their products effectively within the market, target specific customer segments, and make informed decisions about production volumes and pricing tiers.

A picture containing text, screenshot, font, diagram

Description automatically generated

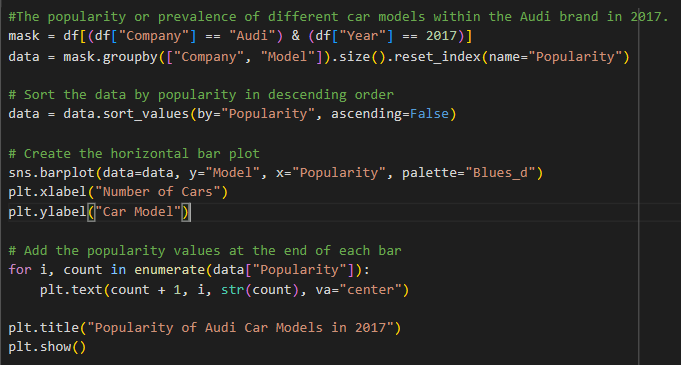
Figure 3.14: Average price of Volkswagen based on vehicle size in 2017.

A picture containing text, screenshot, font, diagram

Description automatically generated

Figure 3.15: Average price of BMW based on vehicle size in 2017.

The visualization on figures 3.14 and 3.15 present the average price distribution of Volkswagen and BMW vehicles based on their size in the year 2017. It offers valuable insights for manufacturers by showcasing the proportionate contribution of each vehicle size category to the overall average price. The pie chart with labeled slices allows easy comprehension of the market distribution. Furthermore, including price values inside each slice offers exact information for understanding the typical prices connected with each category. According to the findings in figure 3.14 and 3.15, large vehicles have the greatest average price, followed by medium and compact vehicles. Manufacturers may use this visualization to rapidly comprehend price patterns across vehicle sizes and make educated decisions about their product offers and pricing strategies.



A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 3.16: Popularity of Audi Car Models in 2017

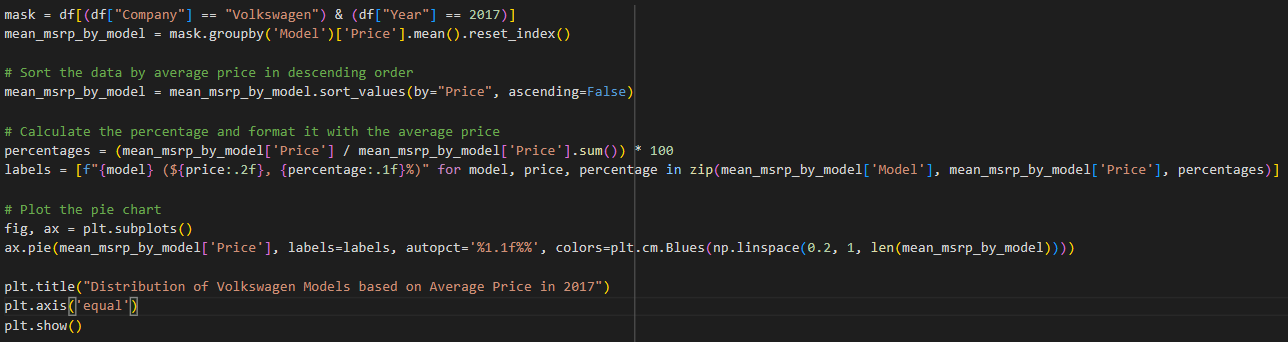
The provided code on figure above generates a horizontal bar plot representing the popularity or prevalence of different car models within the Audi brand in the year 2017. The plot displays the number of cars for each Audi model, sorted in descending order based on their popularity. Visualization allows manufacturers and Audi brand stakeholders to quickly understand and compare the popularity of different car models, identifying which models were more prevalent in the market during 2017. The values displayed at the end of each bar provide precise information about the number of cars for each model, aiding in making data-driven decisions related to production, marketing, and resource allocation.

Figure 3.17: The code of visualizing the distribution of Volkswagen and Audi models based on average price in 2017.

A picture containing text, screenshot, font, circle

Description automatically generated

Figure 3.18: The distribution of Volkswagen Models based on average price in 2017.

A picture containing text, screenshot, font, circle

Description automatically generated

Figure 3.19: The distribution of Audi Models based on average price in 2017.

Based on the code in figure 3.17, the goal is to visualize the distribution of Volkswagen and Audi models based on their average price in the year 2017. By filtering the dataset to include only a particular brand of vehicles in 2017, calculating the mean price for each model, and sorting the data in descending order of average price, we obtain a ranked list of car models for a particular brand. The pie chart is then created (figure 3.18 & 3.19) to display the proportion of each model's average price relative to the total sum of average prices. The chart provides a visual representation that allows manufacturers to quickly identify which models have higher or lower average prices compared to others. Figure 3.18, for example, indicates that the Touareg (a Volkswagen model) had the largest average price distribution in 2017, at 15.9%. This implies that the Touareg is the most expensive automobile model in the Volkswagen brand in 2017. While The information provided in Figure 3.19 indicates that the Audi R8 has the highest percentage of distribution of average price among all the car models in 2017, with a value of 13.2%. This means that the Audi R8 accounted for a significant portion of the average price distribution in that year.

A screen shot of a computer program

Description automatically generated with low confidence

Figure 3.20: The code of visualizing the correlation heatmap between numerial features.

A picture containing text, screenshot, parallel, number

Description automatically generated

Figure 3.21: The correlations between different numerical features

The provided code in figure 3.21 generates a correlation heatmap to visualize the relationships between different numerical features in a dataset. In this case, the heatmap in figure 3.21 highlights the correlation between city MPG (miles per gallon) and highway MPG. This means that there is a strong relationship between these two variables, suggesting that vehicles with higher city MPG tend to have higher highway MPG as well. The heatmap provdes a visual representation of this correlation, allowing manufacturers to easily identify this pattern. This insight can be valuable for manufacturers in the automotive industry as it indicates that improvements in fuel efficiency for city driving are likely to translate into improved fuel efficiency on the highway as well. Besides that, since our data related to retail price, we can observe that engine HP and engine cylinders exhibit a relatively strong positive correlation with the retail price. This indicates that as the values of these variables increase, there is a tendency for the retail price of the product to increase as well.

Figure 3.22: The distribution of transmission typesA picture containing text, circle, screenshot, data storage device

Description automatically generated

The pie chart in figure 3.22 visually presents the proportions of different transmission types, allowing manufacturers to gain insights into the popularity or prevalence of each type. Based on the figure, we can infer that Americans, or the USA market have a strong preference for automatic transmissions which is almost 70%. Automatic transmissions offer convenience and ease of use, as they do not require the driver to manually shift gears. This preference for automatic transmissions may be influenced by various factors, including driving habits, traffic conditions, and personal preferences. This insight can be valuable for manufacturers targeting the American market, as it suggests that prioritizing the production or availability of vehicles with automatic transmissions may align with consumer preferences and potentially lead to higher sales or market acceptance in the USA.

A picture containing text, screenshot, diagram, font

Description automatically generated

Figure 3.23: Engine fuel types for automatic category

The pie chart in figure 3.23 displays the proportion of different fuel types used in vehicles with automatic transmissions. Based on the visualization, it appears that Regular unleaded fuel has the highest percentage (36.4%) among the engine fuel types. There could be a reason, such as the fact that automatic transmission vehicles often include a wide range of models, including those designed for everyday commuting and general use. These vehicles are often optimized for regular unleaded fuel, making it the recommended or default fuel type. As a result, a larger number of automatic transmission vehicles are likely to be compatible with regular unleaded fuel, leading to a higher percentage in the dataset.

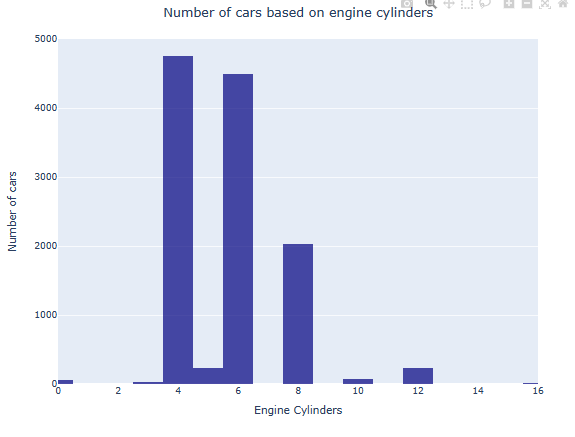


Figure 3.24: The number of cars based on engine cylinders.

A screen shot of a computer program

Description automatically generated with low confidence

Figure 3.25: The car’s specifications for 16 engine cylinders

The histogram in figure 3.24 represents the number of cars based on the count of engine cylinders. The x-axis represents the number of engine cylinders, while the y-axis represents the number of cars. The visualization shows that four engine cylinders have the highest number of cars, which is 4752. This means that in the USA market, cars typically have four engine cylinders. Additionally, we could analyze car’s specifications based on number of engine cylinders. For example, figure 3.25 shows the Bugatti Veyron 16.4 model specifications. We extract specific information and analyze the data associated with the car that has the highest number of engine cylinders, which is 16 engine cylinders.

A picture containing text, screenshot, font, plot

Description automatically generated

Figure 3.26: The number of cars sold for less than $5000 across different years.

By analyzing the bar chart in figure 3.26 showing the count of cars sold per year, manufacturers can identify the years with higher demand for lower-priced vehicles. The decreasing trend in the mean price of cars suggests that, on average, cars were priced lower in 1999 compared to the preceding years. This can be interpreted as a potential market shift towards more affordable or budget-friendly cars during that time. The decrease in mean price could be attributed to various factors such as changes in market demand, economic conditions, or the availability of lower-priced car models. Simultaneously, the decline in the number of cars sold for less than $5000 further supports the observation of a changing market trend. This decrease in count indicates a potential decrease in the availability of cars priced below $5000 in the market. It could be indicative of a shift towards higher-priced vehicles or a decrease in the supply of inexpensive used cars during that specific year. This information can guide production planning and inventory management strategies. Additionally, the line plot showcasing the mean price per year helps manufacturers understand the average pricing dynamics for lower-priced cars over time. This insight can aid in pricing strategies, competitive analysis, and product positioning.

**4.0 Model or Methods and Results (Modeling and Evaluation)**

Two different types of machine learning algorithm models were used, decision tree and random forest to train the car price dataset and obtain the results for further model evaluation.

* **Random Forest Regression**

Random Forest is a popular machine learning algorithm which is an ensemble learning method that combines multiple decision trees to make predictions. Firstly, data pre-processing must be performed as explained in Problem Formulation (Data Wrangling) part above in the report then dataset will be split into training and testing sets. Then the model training is performed by importing the necessary libraries, including the Random Forest regressor from scikit-learn.

Then an instance of the Random Forest regressor with desired parameters, such as the number of trees and maximum depth were created. Using the fit(**)** method, providing the features (X\_train) and target variable (y\_train), the model was fit to the training data. Model prediction was calculated to further calculate R2 score, hence the trained Random Forest model was used to predict car prices for the testing dataset by calling the predict(**)** method, providing the features (X\_test). The predicted prices (y\_pred) for the test set were obtained to use it for R2 calculation.

As for the evaluation, the performance of the Random Forest model using appropriate evaluation metrics like R-squared (R2) which represents the proportion of variance in the target variable (car price) explained by the model were assessed. R-squared value were obtained to understand how well the model captures the variability in the car prices. Higher R-squared values indicate a better fit and it ranges between 0 and 1, with higher values indicating a better fit of the model to the data. Below is the prediction, R2 results and scatter plot of R2 score of random forest regression.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 4.1: Prediction of car prices and R2 results of random forest regression.

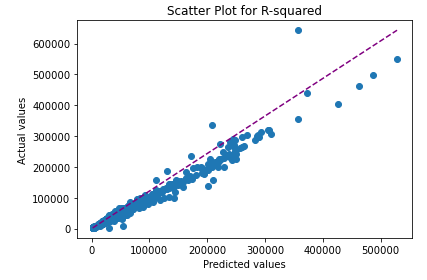


Figure 4.2: Scatter plot of R2 score for random forest regression.

* **Decision Tree Regression**

The decision tree algorithm has been used for the modelling part. It is a tree-like model which is constructed to make predictions based on the features of the cars. Firstly, the data was pre-processed as explained in Data Wrangling part. The dataset is then split into training and testing sets.

Next, the decision tree model was trained on the training data. The decision tree algorithm builds a tree structure by recursively partitioning the data based on the feature values that best separate the car prices. The splitting process continues until a certain stopping criterion is met, such as reaching a maximum depth or minimum number of samples in each leaf node.

Once the decision tree is trained, it was used to predict car prices for the testing dataset. The features of the cars in the test set were passed through the decision tree, and the corresponding predicted prices are obtained. These predictions are then compared to the actual prices to evaluate the performance of the model.

Evaluation of the decision tree model involves using appropriate metrics for regression tasks. Common evaluation metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2). For this car analysis, R2 score has been used as an evaluation metric. R2 score indicates the proportion of variance explained by the model and in order to obtain the R2 score, the prediction value has to be calculated first.

Interpreting the evaluation metrics helps in assessing the performance of the decision tree model. A higher R2 value suggests that the model captures a larger portion of the variability in the car prices. Below are the prediction, R2 results and scatter plot of R2 score of decision tree regression.

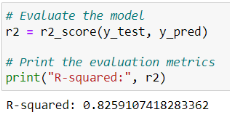


Figure 4.3: R2 results of decisions tree regression.

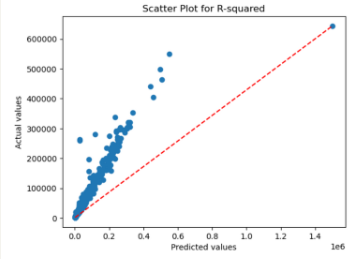


Figure 4.4: Scatter plot of R2 score for decision tree regression.

**5.0 Limitations and Future Recommendations**

During the study of our analysis on car retail prices based on features and MRSP, we encountered certain limitations that affected the scope and applicability of the study. Recognizing these limitations and reflecting on our analysis process, we can identify areas for improvement and offer recommendations for future research.

A major limitation of our analysis is that the dataset spans from 1990 to 2017. While this dataset provides valuable insights into historical trends, it does not capture the most recent market dynamics and pricing patterns. Future research should aim to acquire more up-to-date data to ensure the relevance and accuracy of the analysis, especially considering the potential impact of economic, technological, and consumer behavior changes in recent years.

To overcome the limitations related to dataset timeframe, future studies should consider utilizing real-time or more recent data to capture the current market dynamics accurately. This could involve partnering with automotive industry stakeholders, such as manufacturers, dealerships, or data providers, to access updated and comprehensive datasets that reflect the latest pricing trends.

One limitation of our analysis is that it may not provide manufacturers with actionable insights into the relationship between car features and retail prices. The analysis primarily focuses on understanding the correlation between features and prices without delving into the underlying factors that drive pricing decisions. Future research should explore additional methodologies, such as market research surveys or interviews with industry experts, to gain a more nuanced understanding of the factors influencing pricing decisions from a manufacturer's perspective.

To provide manufacturers with more valuable insights, future research should incorporate qualitative approaches, such as interviews or focus groups with industry experts or professionals in the automotive sector. This would enable a deeper understanding of the decision-making process behind pricing strategies, allowing manufacturers to optimize their pricing models based on a comprehensive understanding of consumer preferences, market competition, and other relevant factors.

Furthermore, to enhance the analysis and provide a more comprehensive understanding of car retail prices, future studies should consider including external factors that influence pricing decisions. These could include economic indicators (inflation rates and GDP growth), market trends (consumer preferences and technological advancements), and regulatory factors (emission standards and government incentives). Incorporating these external factors would contribute to a more accurate and holistic analysis of car retail prices.

The exclusion of Malaysia, our home country, from the analysis due to the unavailability of relevant data is indeed a limitation. Analyzing the car retail prices in our own country would have provided valuable insights into the local market dynamics, consumer preferences, and factors specific to the Malaysian automotive industry.

To address this limitation and gain a better understanding of car retail prices in Malaysia, future research should focus on collecting a comprehensive dataset specific to the Malaysian market. This dataset should include relevant features such as car brands, models, year of manufacture, mileage, and other factors that influence pricing decisions.

Furthermore, incorporating Malaysia in the analysis would enable comparisons with other regions or countries, providing a broader perspective on global trends, pricing differentials, and market competitiveness. This comparative analysis would be particularly useful for stakeholders seeking to expand their operations or make strategic decisions in the international automotive market.

Consider exploring other ensemble learning techniques in addition to random forests. Ensemble methods, such as AdaBoost, Gradient Boosting, or XGBoost, can potentially improve the predictive power of the models by combining multiple weaker models into a stronger ensemble. These methods often result in higher accuracy and better generalization.

Decision trees and random forest models offer good interpretability, as they allow for feature importance analysis. Explore the feature importance rankings provided by the models to gain insights into which features have the greatest impact on predicting car retail prices. This information can be valuable for understanding the driving factors behind pricing decisions and informing decision-making processes for manufacturers, buyers, and other stakeholders.

Lastly, the performance of machine learning models heavily depends on the quality and relevance of the input features. Conduct further feature engineering to enhance the predictive power of the models. This could involve creating new features based on domain knowledge or extracting additional information from the existing features. Feature selection techniques, such as recursive feature elimination or feature importance analysis, can help identify the most relevant features for predicting car retail prices.

**6.0 Conclusion**

In conclusion, the analysis conducted on car prices through analytical insights provides a comprehensive understanding of the multifaceted factors that influence the automotive market. By examining a wide range of attributes such as brand, model, year of manufacture, mileage, engine size, fuel efficiency, transmission type, vehicle condition, and additional features, this analysis equips stakeholders in the automotive industry with valuable knowledge to inform their decision-making processes.

Car sellers can leverage the analysis to optimize their pricing strategies and enhance their competitiveness. By staying attuned to market trends and customer preferences, sellers can adjust their pricing strategies to remain competitive and maximize their sales potential. The insights obtained from this analysis enable sellers to effectively position their vehicles, highlight desirable features, and respond to fluctuations in demand. This knowledge empowers sellers to optimize their pricing decisions, ultimately leading to increased sales and profitability.

For car manufacturers, understanding the key drivers of car prices is crucial for setting competitive pricing strategies. The insights obtained from this analysis allow manufacturers to make data-driven decisions regarding production volumes, model configurations, and marketing initiatives. By considering factors such as brand value, market demand, and the impact of specific features on pricing, manufacturers can effectively price their vehicles to align with consumer expectations and market conditions. This analysis equips manufacturers with the knowledge needed to enhance their competitiveness and profitability in the dynamic automotive industry.

**7.0 References**

Arawomo, D.F., & Osigwe, A.C. (2016). Nexus of fuel consumption, car features and car prices: Evidence from major institutions in Ibadan. Renewable & Sustainable Energy Reviews, 59, 1220-1228.

Zhang, Y., Li, J., & Tao, W. (2021). Does energy efficiency affect appliance prices? Empirical analysis of air conditioners in China based on propensity score matching. Energy Economics, 101, 105435.

Ou, S., Li, W.Z., Li, J., Lin, Z., He, X., Bouchard, J., & Przesmitzki, S.V. (2020). Relationships between Vehicle Pricing and Features: Data Driven Analysis of the Chinese Vehicle Market. Energies.

Bansal, P., & Kockelman, K.M. (2017). Indian Vehicle Ownership: Insights from Literature Review, Expert Interviews, and State-Level Model. Journal of the Transportation Research Forum, 56.

Jang, S., & Choi, J. (2021). Which consumer attributes will act crucial roles for the fast market adoption of electric vehicles?: Estimation on the asymmetrical & heterogeneous consumer preferences on the EVs. Energy Policy, 156, 112469.

Jacyna, M., & Semenov, I. (2020). Models of vehicle service system supply under information uncertainty. Eksploatacja I Niezawodnosc-maintenance and Reliability, 22, 694-704.

Nguyen, P. (2021). Investigating car purchasing decision-making process using Multi- Objective Optimization Ratio Analysis based Analytical Hierarchy Process Model: An empirical case from Vietnam.

Mustapa, S.I., Ayodele, B.V., Mohamad Ishak, W.W., & Ayodele, F.O. (2020). Evaluation of Cost Competitiveness of Electric Vehicles in Malaysia Using Life Cycle Cost Analysis Approach. Sustainability.

**8.0 Acknowledgement**

We would like to extend our sincere appreciation to our esteemed lecturer Intrduction od Data Science (BITI2513), Dr. Noor Fazillah Abd Yusof, for her invaluable guidance and instruction throughout the process of conducting this data analysis. Her expertise, knowledge, and dedication to teaching have been instrumental in our understanding of data analysis techniques and methodologies.

First and foremost, we extend our appreciation to the dataset provider, whose efforts in collecting and curating the car features and MRSP data were instrumental in our analysis. Their dedication in ensuring data accuracy and accessibility deserves acknowledgment.

We would also like to thank the authors and researchers whose studies and publications on car pricing analysis provided valuable insights and served as a foundation for our research. Their contributions to the field of automotive economics and pricing dynamics have been immensely helpful in shaping our understanding and approach.

Finally, we would like to acknowledge our team members for their collaborative efforts and dedication in conducting this analysis. Each team member's unique skills and perspectives have greatly contributed to the success of this project.