**MACHINE LEARNING ALGORITHMS-1**

**CIA 1**

**DOMAIN-SPECIFIC MODEL BUILDING**

**(AUTOMOTIVE DOMAIN)**

**BY**

**SHREYA RUSTAGI**

**2327958**

**UNDER THE GUIDANCE OF**

**DR. HELEN JOSEPHINE**



**MBA PROGRAMME**

**SCHOOL OF BUSINESS AND MANAGEMENT**

**CHRIST (DEEMED TO BE UNIVERSITY), BANGALORE**

**JULY**

**2024**

#### Business understanding

#### The automotive industry is characterized by intense competition, where pricing strategies play a crucial role in determining the market success of a vehicle. Manufacturers and dealerships must set prices that are attractive to consumers while ensuring profitability. The challenge lies in accurately predicting the market price of a car based on its various attributes and features. Accurate price prediction models can help automotive companies set competitive prices, optimize their production costs, and forecast potential profits, thus maintaining a strategic edge in the market. Determining the factors affecting the price of the car is very crucial to understand.

#### Problem Identification

To understand the key factors that influence the pricing of cars. This can help car manufacturers, dealers, and consumers make informed decisions. Specifically, we aim to identify the relationships between car features (e.g., engine size, horsepower, dimensions) and their prices.

1. **Variables**

Here are the key variables in the dataset:

1. car: The make of the car.
2. wheel-base: The distance between the centers of the front and rear wheels.
3. length: The length of the car.
4. width: The width of the car.
5. height: The height of the car.
6. curb-weight: The weight of the car without passengers or cargo.
7. engine-size: The size of the car's engine.
8. bore: The diameter of the cylinders in the engine.
9. stroke: The distance the piston travels in the cylinder.
10. compression-ratio: The ratio of the cylinder's volume at the bottom of the piston stroke to the volume at the top of the stroke.
11. horsepower: The power output of the engine.
12. peak-rpm: The maximum revolutions per minute of the engine.
13. city-mpg: The car's fuel efficiency in miles per gallon in city driving.
14. highway-mpg: The car's fuel efficiency in miles per gallon on the highway.
15. price: The price of the car.

### **Objectives**

1. **Data Exploration**: Understand the distribution and relationships between different variables.
2. **Feature Selection**: Identify the most influential features that affect car pricing.
3. **Model Development**: Develop a regression model to predict car prices based on the identified features.
4. **Model Evaluation**: Assess the model's performance using appropriate metrics to ensure accuracy and reliability.

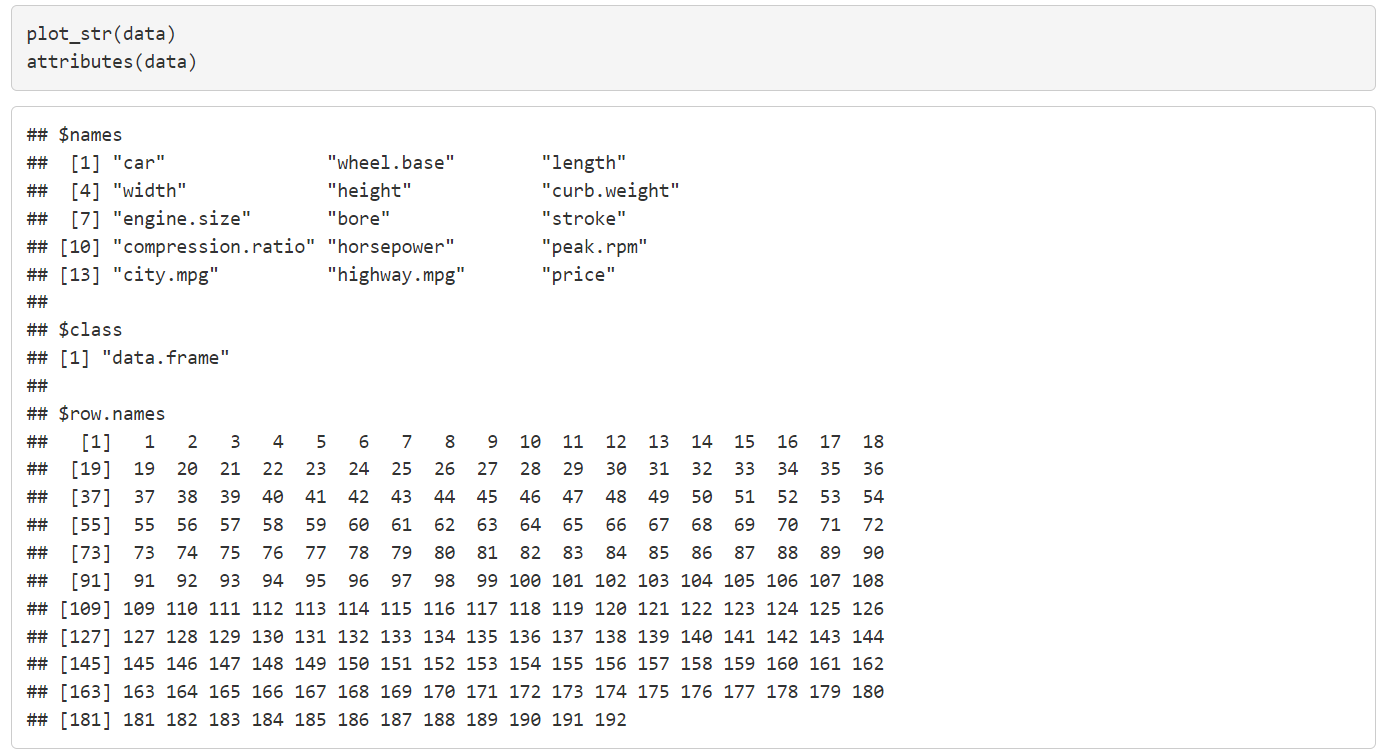
### **Importance of the Problem**

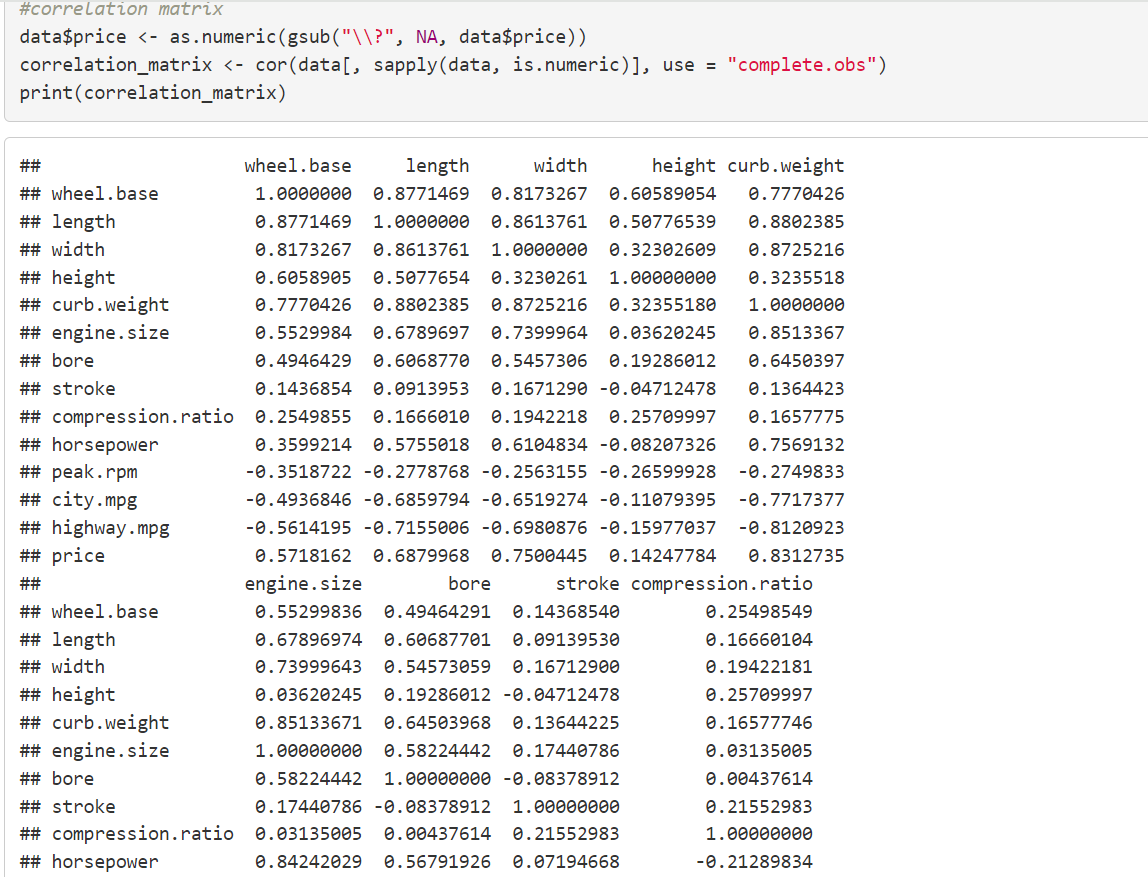
Accurate pricing models are vital for automotive companies for several reasons:

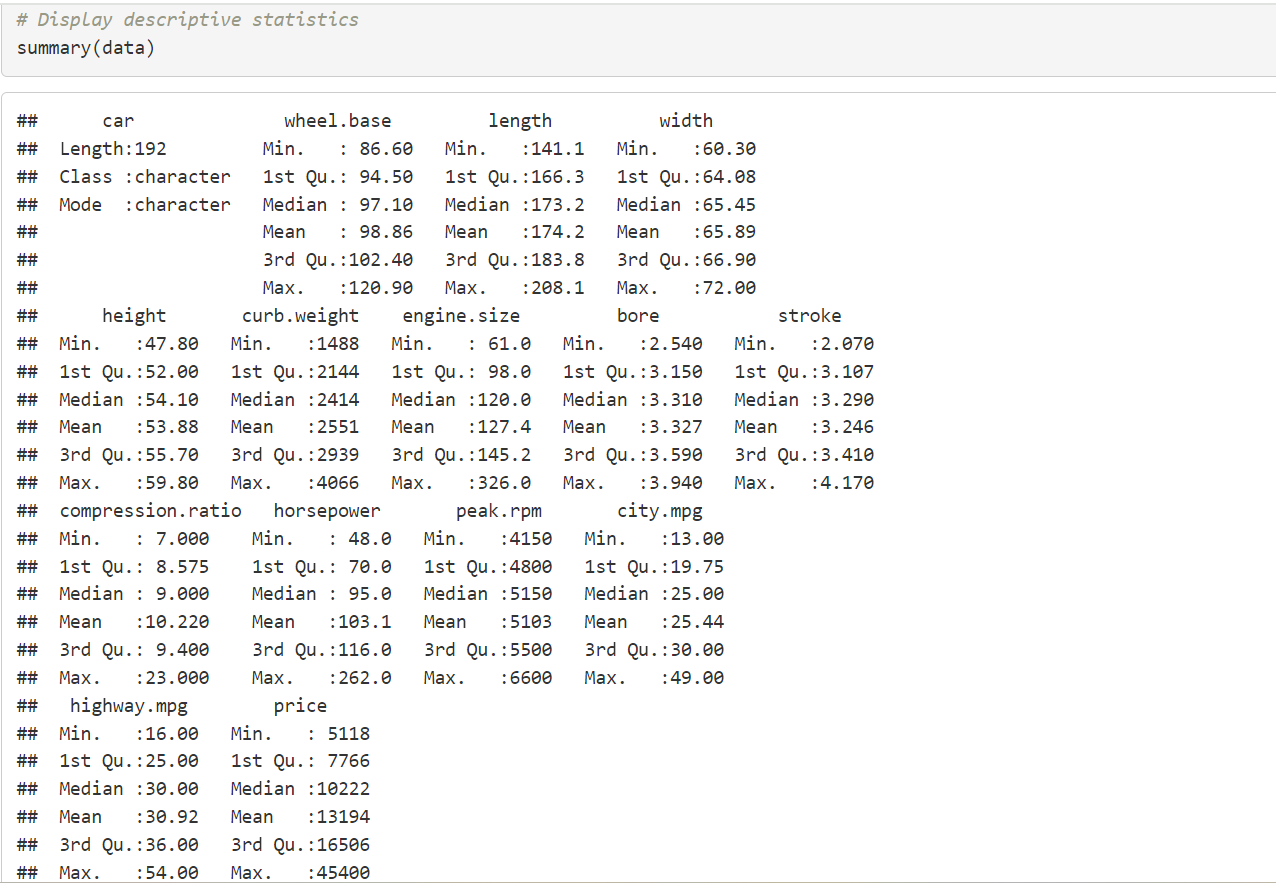
* **Competitive Pricing**: Ensuring prices are competitive within the market to attract potential buyers.
* **Profit Optimization**: Balancing between an attractive price point for consumers and the profitability for the manufacturer.
* **Cost Management**: Optimizing production and operational costs based on predicted sales prices.
* **Market Strategy**: Informing strategic decisions related to marketing, product development, and inventory management.

**2.Data understanding**

* 1. **Data collection:** The data has been taken from UCI machine learning repository. (https://archive.ics.uci.edu/datasets)
  2. **Data exploration:** The dataset was explored to understand its structure, distribution of values, and relationships between variables. Summary statistics and data visualization techniques were used.

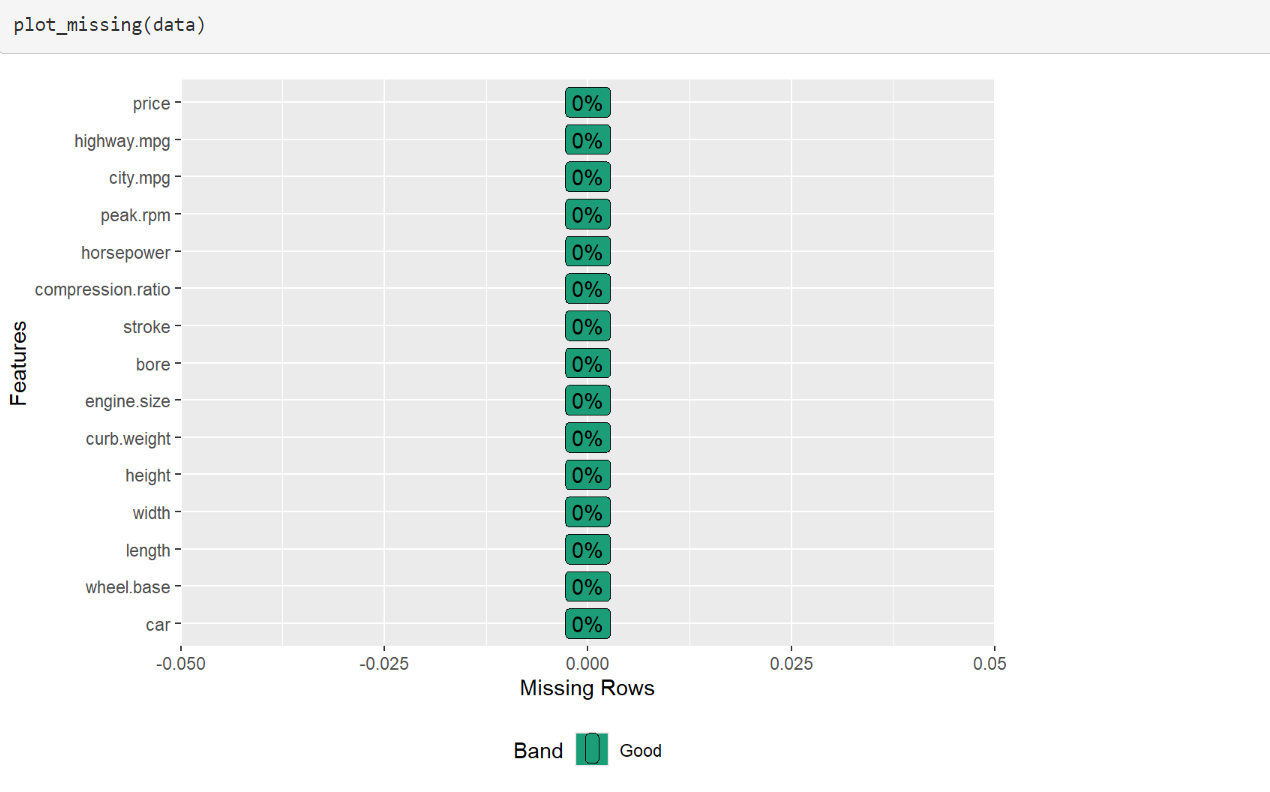


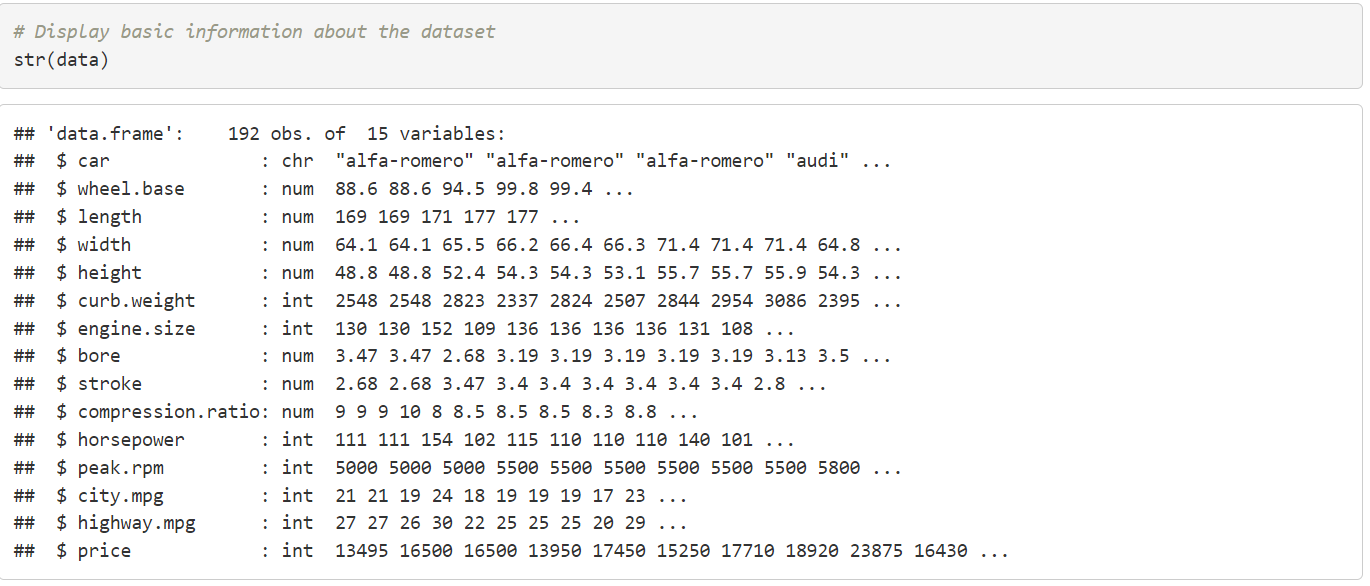




**Explanation of the code**:

* str(data), plot\_str(data), attributes(data), head(data, 10), and tail(data, 10) provide a structural overview and sample records of the dataset.
* summary(data) provides summary statistics for each variable.
* plot\_missing(data) visualizes the missing data.
* The dataset is cleaned by removing duplicate rows, converting the price column to numeric, and removing rows with missing price values.
  1. **Assessing data quality:**
     + **Missing Values**: Identified missing values in the dataset.
     + **Outliers**: Detected and treated outliers to ensure data quality.
     + **Data Types**: Ensured that variables had appropriate data types





## **Data Preparation**

### **Data Integration**: The data was integrated from a single source; hence, no integration was required.

### **Data Cleaning**:

### **Missing value analysis**: Identified missing values in the dataset but there were no missing values in the dataset.

#### Data Imputation: There was no missing values so imputation wasn’t required for the dataset.

#### Variable Standardization: Standardized numerical variables to ensure they have a mean of 0 and a standard deviation of 1.

#### Feature Selection/Engineering: Selected features based on their correlation with the target variable, price.

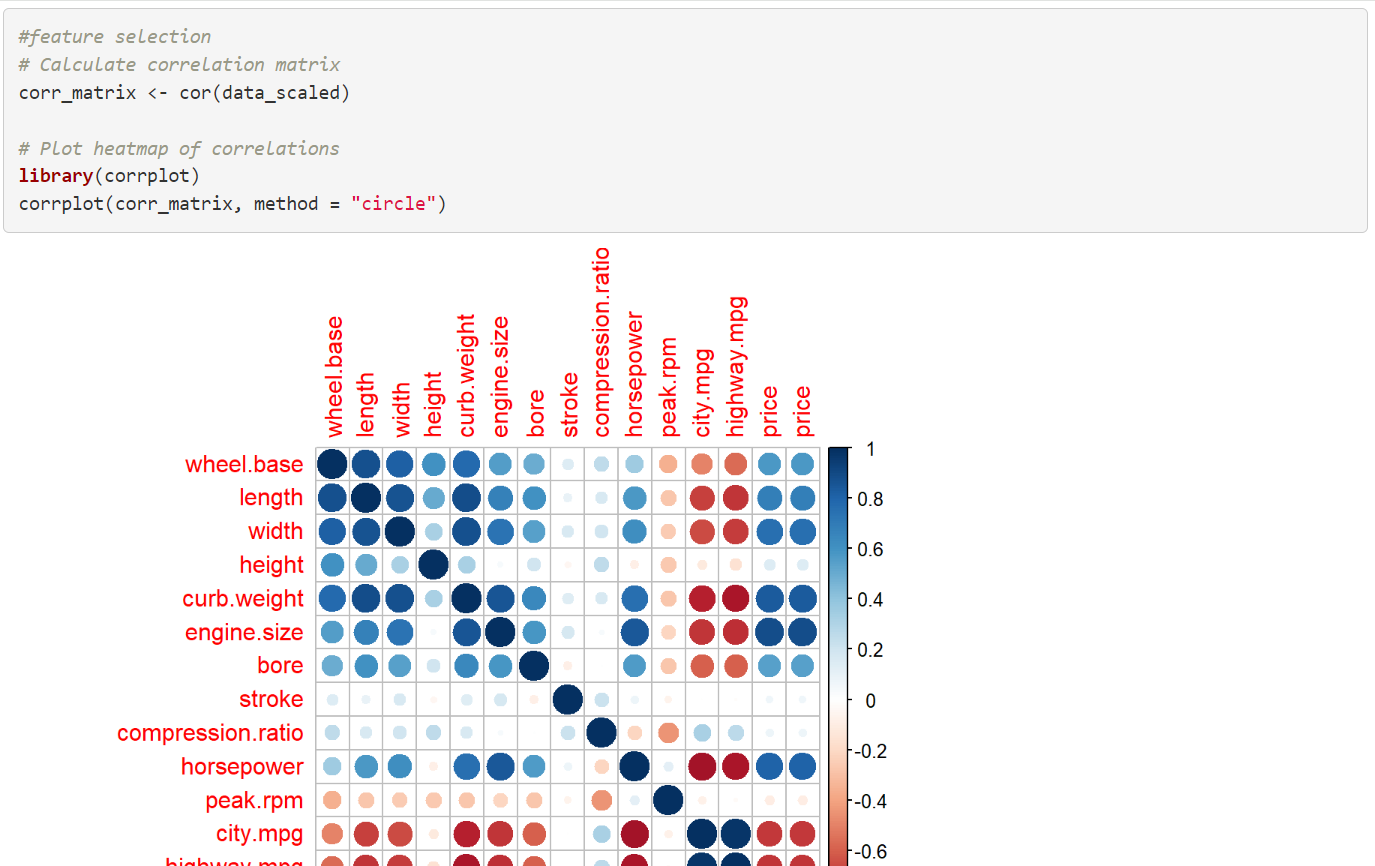
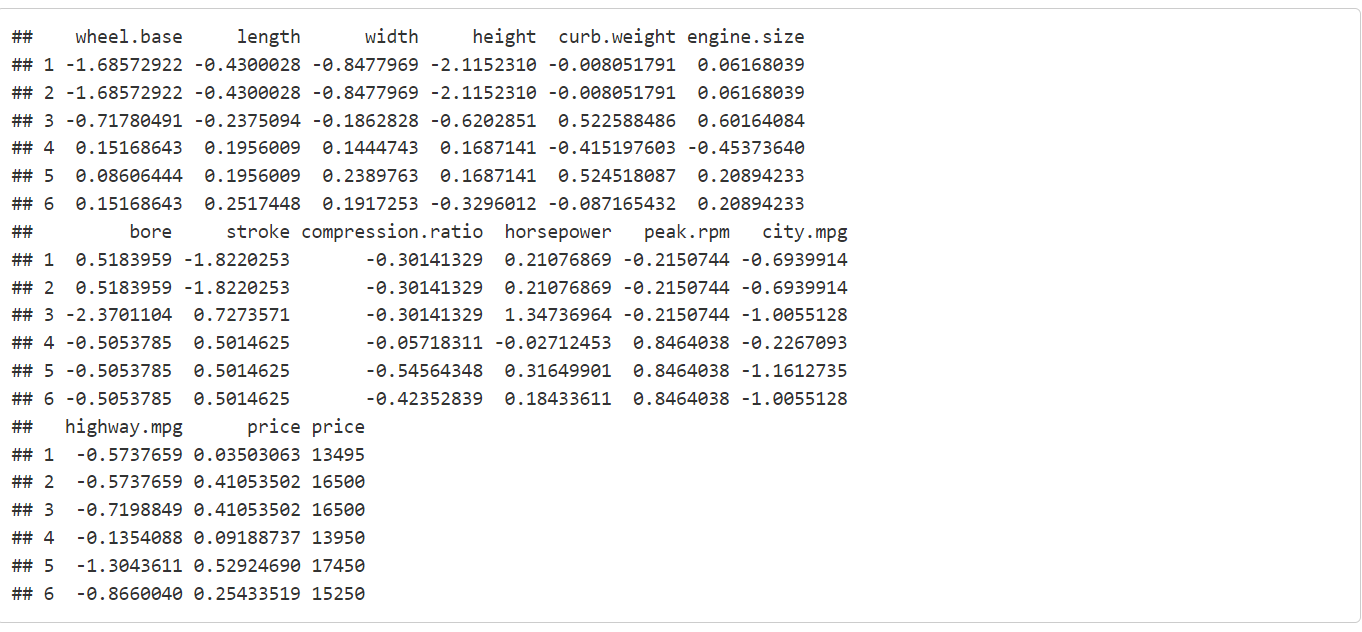
#### Detection and Treatment: Detected and treated outliers to avoid skewed results and ensure model accuracy. There were no outliers in the dataset.

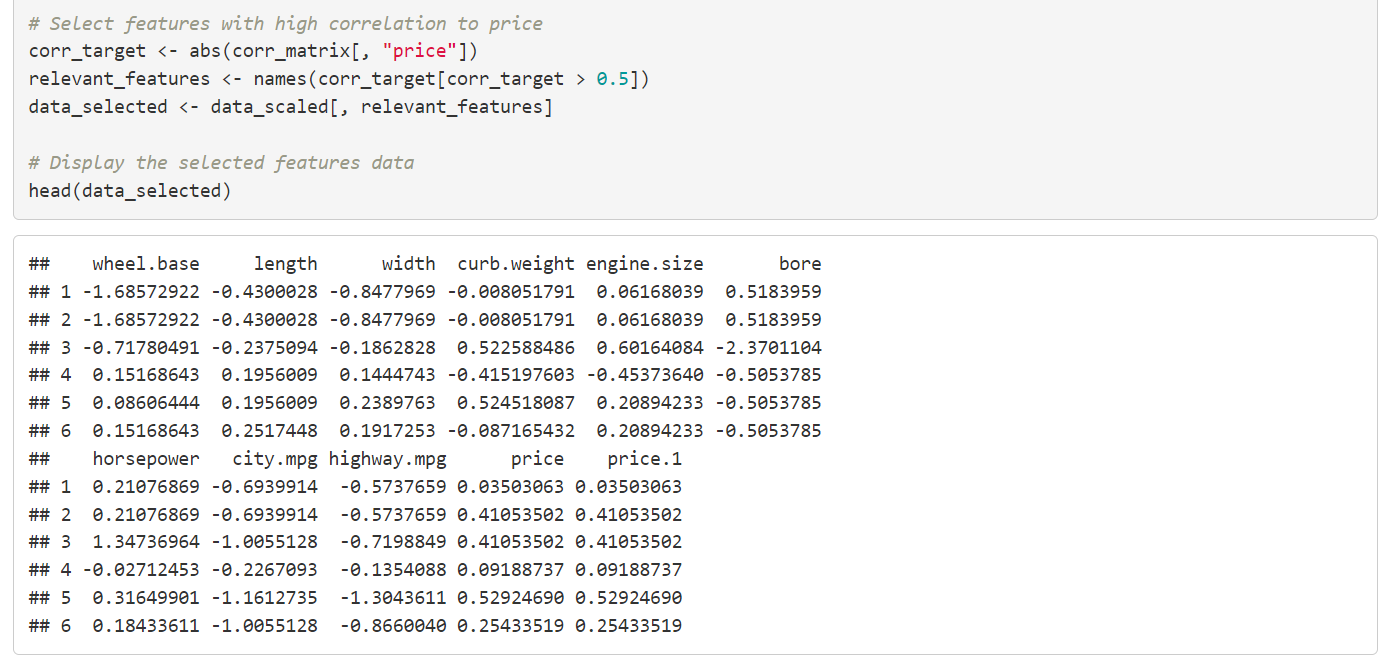




**Explanation of the code**:

* Numerical columns are selected and standardized to have a mean of 0 and a standard deviation of 1.
* The standardized data is combined back with the price column.

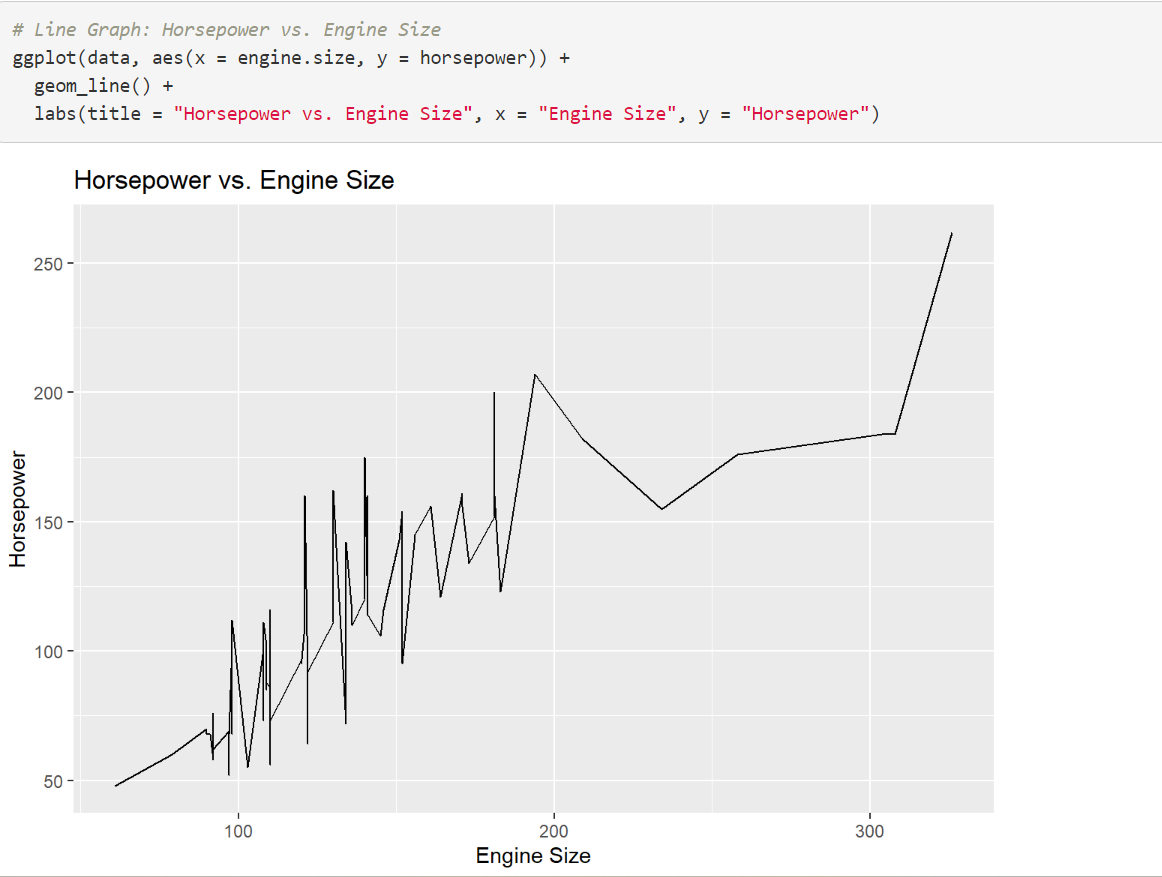




**Explanation of the code**:

* A correlation matrix is calculated and visualized using corrplot.
* Features with a high correlation to price (correlation > 0.5) are selected for further analysis.
  + 1. **Visualisation :**

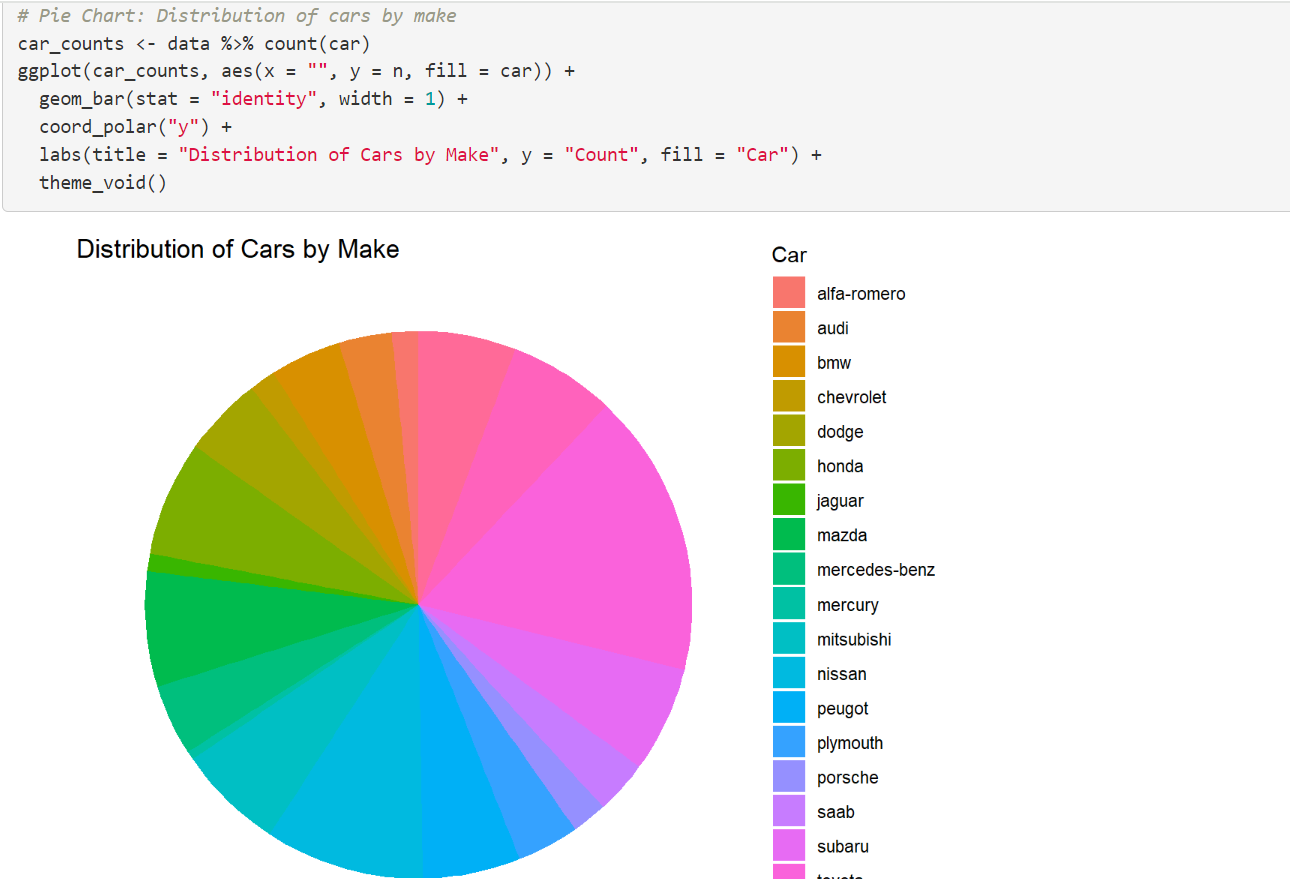
1. **Line graph**



**Interpretation**

The line graph shows a generally positive relationship between engine size and horsepower, indicating that as engine size increases, horsepower typically rises. However, significant variability is observed, particularly in the 150-200 engine size range, suggesting that other factors besides engine size, such as engine design, turbocharging, and fuel type, influence horsepower. Sharp fluctuations and a steep increase towards larger engine sizes imply that certain high-performance engines are designed to deliver significantly more power. This relationship can help automotive manufacturers differentiate products, target performance-oriented customers, and develop pricing strategies.

1. **Pie chart**



**Interpretation**

The pie chart shows a diverse and balanced representation of various car manufacturers in the dataset. Each segment, representing a different car make, is of roughly equal size, indicating that no single manufacturer overwhelmingly dominates the dataset. This balanced distribution ensures that analyses and predictive models derived from the data are unbiased and generalizable across different brands. The diverse representation of car makes facilitates comprehensive market analysis and helps in understanding competitive dynamics.

Additionally, it enables the development of reliable pricing models that can be applied broadly, aiding manufacturers and dealerships in devising targeted marketing strategies and making informed business decisions.

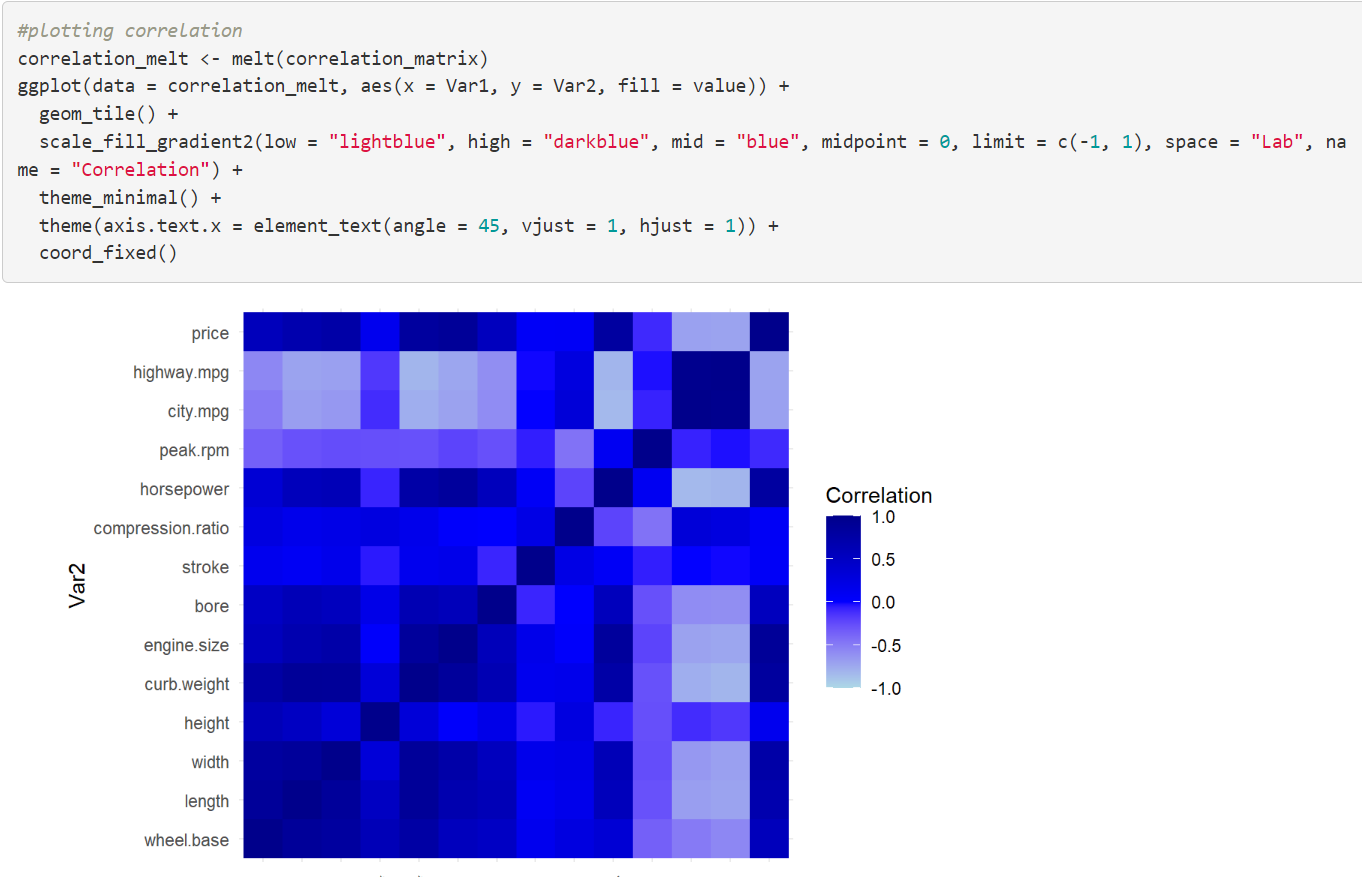
**Explanation of the code:**

 This code creates a pie chart showing the distribution of cars by make.

 The count function calculates the number of cars for each make, and geom\_bar with coord\_polar("y") creates a pie chart.

1. **Correlation matrixTop of Form**

Bottom of Form



**Interpretation**

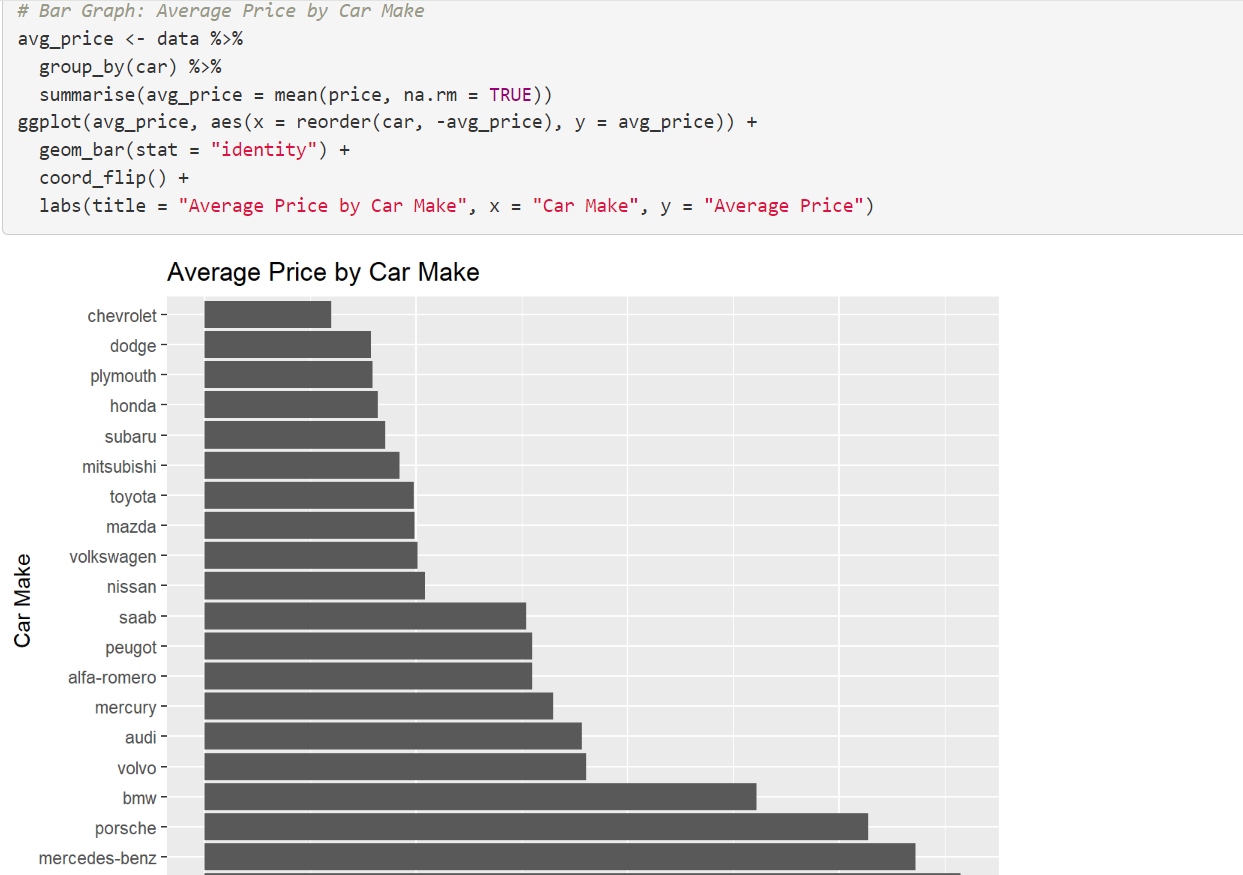
The correlation matrix heatmap shows the relationships between numerical variables in the dataset. Darker blue indicates stronger positive correlations, while lighter shades indicate weaker or negative correlations. Notably:

* **Price** is strongly positively correlated with **engine size**, **curb weight**, and **horsepower**, indicating higher prices for cars with larger engines, more weight, and more power.
* **Price** is negatively correlated with **city-mpg** and **highway-mpg**, suggesting higher prices for less fuel-efficient cars.
* **Engine size** and **horsepower** are strongly correlated, as are **wheel base**, **length**, and **width**.

**Explanation of the code:**

* The correlation matrix is melted into a long format suitable for ggplot2.
* A matrix of correlations is plotted using geom\_tile to visualize the relationships between variables.

1. **Bar graph**



**Interpretation**

The bar graph shows the average car prices across different manufacturers. From the graph, we can see that luxury brands like Mercedes-Benz, Porsche, BMW, and Volvo have the highest average prices, reflecting their premium market positioning and the higher value associated with their vehicles. In contrast, more mainstream and budget-friendly brands like Chevrolet, Dodge, and Plymouth show significantly lower average prices, indicating their focus on affordability and mass-market appeal. This distribution highlights the varied market segments catered to by different manufacturers, from high-end luxury to economy. Such insights are crucial for understanding market dynamics, setting competitive pricing strategies, and targeting appropriate consumer segments.

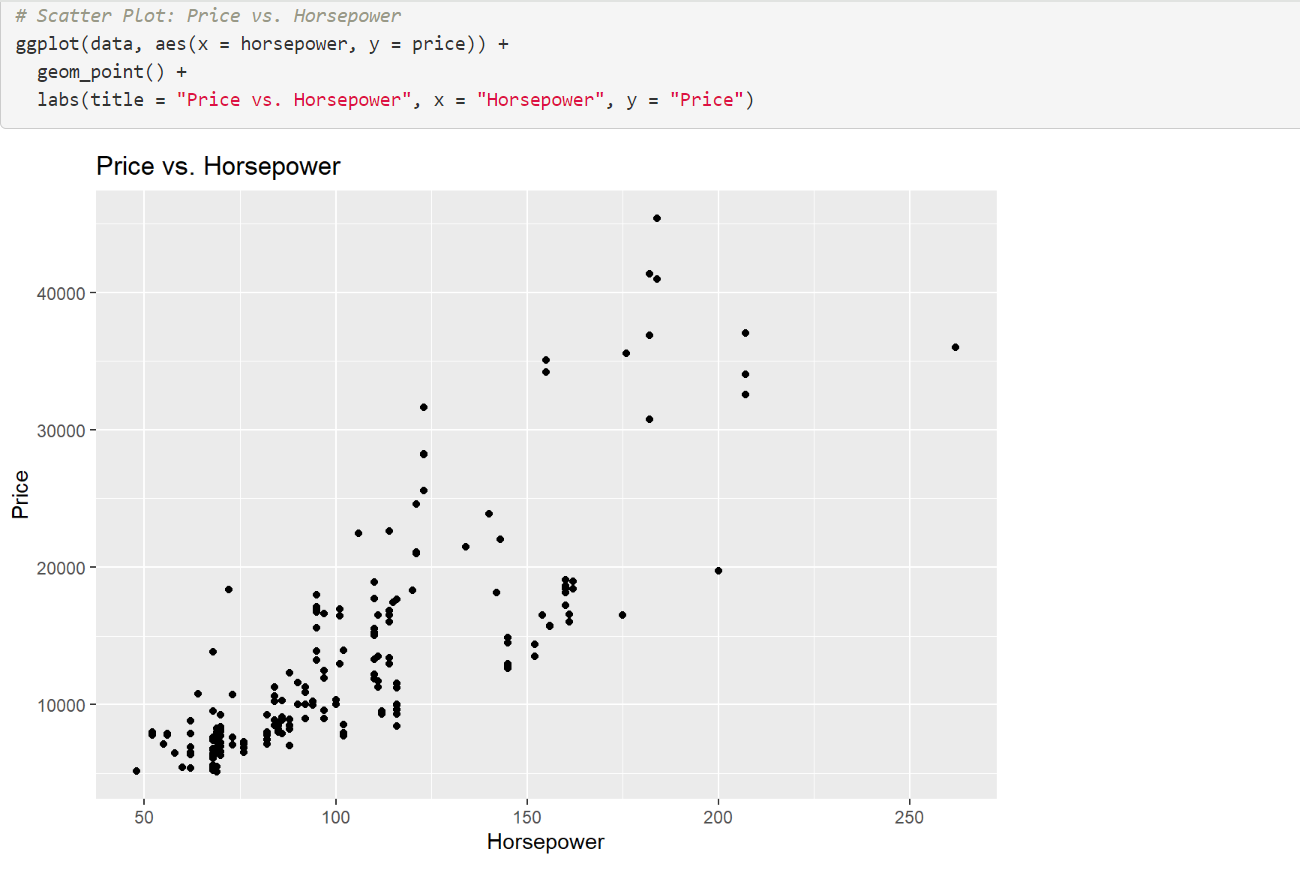
**Explanation of the code**:

* This code calculates the average price of cars by make and plots it using a bar graph.
* coord\_flip() is used to flip the coordinates for better readability.

**Top of Form**

**Bottom of Form**

1. **Scatterplot**



**Interpretation**

The scatter plot depicts the relationship between a car's horsepower and its price. The plot shows a clear positive correlation, indicating that cars with higher horsepower tend to be priced higher. This trend is expected as more powerful engines often come with higher production costs and are typically found in higher-end or performance vehicles. However, the plot also reveals significant variability in prices for a given horsepower level, suggesting that other factors such as brand, luxury features, and market positioning also play crucial roles in determining a car's price. The presence of outliers, particularly in the higher horsepower range, highlights specific models that are priced substantially higher than others, likely due to their premium or specialized nature. Overall, the scatter plot underscores the importance of horsepower as a key factor influencing car prices, while also pointing to the multifaceted nature of vehicle pricing strategies.

**5. MODEL EVALUATION AND MODEL DIAGNOSTICS**

## **1. Multiple Linear Regression (MLR)**

### **Model Explanation**

* **Multiple Linear Regression** models the relationship between a dependent variable (price) and multiple independent variables (car features).
* The model assumes a linear relationship between the predictors and the target variable.

### **Performance Metrics**

* **R-squared**: 0.9125956
* **MAD**: 1839.089
* **MSE**: 6895630
* **RMSE**: 2625.953

### **Interpretation**

* **R-squared** indicates that 91.26% of the variability in car prices is explained by the model, which is quite high.
* **MAD** is relatively low, indicating that the average deviation of predicted prices from actual prices is about $1839.
* **MSE** and **RMSE** values are the lowest among the models, suggesting that the model has the smallest average squared errors and root mean squared errors.

## **2. Lasso Regression**

### **Model Explanation**

* **Lasso Regression** (Least Absolute Shrinkage and Selection Operator) is a type of linear regression that includes a penalty for large coefficients, which can lead to simpler models with some coefficients reduced to zero (feature selection).
* It helps in reducing overfitting and improving model generalization.

### **Performance Metrics**

* **R-squared**: 0.8510303
* **MAD**: 1988.028
* **MSE**: 10257651
* **RMSE**: 3202.757

### **Interpretation**

* **R-squared** is 85.10%, indicating that 85.10% of the variability in car prices is explained by the model.
* **MAD** is slightly higher than that of MLR, showing greater average deviation.
* **MSE** and **RMSE** values are higher than MLR, indicating larger average squared errors and root mean squared errors.

## **3. Ridge Regression**

### **Model Explanation**

* **Ridge Regression** is similar to linear regression but includes a penalty on the size of coefficients (L2 regularization), which helps in reducing multicollinearity and preventing overfitting.
* Unlike Lasso, Ridge Regression does not shrink coefficients to zero.

### **Performance Metrics**

* **R-squared**: 0.788812
* **MAD**: 2476.874
* **MSE**: 11940289
* **RMSE**: 3455.472

### **Interpretation**

* **R-squared** is 78.88%, indicating that 78.88% of the variability in car prices is explained by the model.
* **MAD** is the highest among the models, showing the greatest average deviation.
* **MSE** and **RMSE** values are also the highest, indicating the largest average squared errors and root mean squared errors.

## **4. Decision Tree**

### **Model Explanation**

* **Decision Tree** regression splits the data into subsets based on the values of input features, creating a tree-like structure to predict the target variable.
* It can capture non-linear relationships between features and the target variable.

### **Performance Metrics**

* **R-squared**: 0.8692891
* **MAD**: 2294.663
* **MSE**: 9294170
* **RMSE**: 3048.634

### **Interpretation**

* **R-squared** is 86.93%, indicating that 86.93% of the variability in car prices is explained by the model.
* **MAD** is higher than MLR and Lasso but lower than Ridge, showing moderate average deviation.
* **MSE** and **RMSE** values are lower than Ridge but higher than MLR and Lasso, indicating moderate average squared errors and root mean squared errors.

**6.Comparison leaderboard**

**LEADER BOARD**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODELS | R SQUARE | MAD | MSE | RMSE |
| MULTIPLE LINEAR REGRESSION | 0.9125956 | 1839.089 | 6895630 | 2625.953 |
| LASSO REGRESSION | 0.8510303 | 1988.028 | 10257651 | 3202.757 |
| RIDGE REGRESSION | 0.788812 | 2476.874 | 11940289 | 3455.472 |
| DECISION TREE | 0.8692891 | 2294.663 | 9294170 | 3048.634 |

## **Leaderboard**

1. **Multiple Linear Regression**
   * Best performance with highest R-squared (0.9126)
   * Lowest error metrics (MAD: 1839.089, MSE: 6895630, RMSE: 2625.953)
   * Most accurate predictions
2. **Lasso Regression**
   * Good performance with high R-squared (0.8510)
   * Effective feature selection but higher errors than MLR (MAD: 1988.028, MSE: 10257651, RMSE: 3202.757)
   * Balances complexity and accuracy
3. **Decision Tree**
   * Moderate performance with R-squared (0.8693)
   * Captures non-linear relationships but higher errors than MLR and Lasso (MAD: 2294.663, MSE: 9294170, RMSE: 3048.634)
   * Useful for understanding feature interactions
4. **Ridge Regression**
   * Lowest performance with R-squared (0.7888)
   * Highest error metrics (MAD: 2476.874, MSE: 11940289, RMSE: 3455.472)
   * Handles multicollinearity but less accurate for this dataset

In conclusion, Multiple Linear Regression is the most effective model for predicting car prices in this dataset, followed by Lasso Regression, Decision Tree, and Ridge Regression.

Top of Form

Bottom of Form

### **Recommendation**

* **Multiple Linear Regression (MLR)** is the best-performing model based on the given metrics. It has the highest R-squared value, indicating the best fit, and the lowest MAD, MSE, and RMSE values, indicating the most accurate predictions.
* **Lasso Regression** can be considered if feature selection is important, as it reduces the impact of less important features by shrinking their coefficients to zero. However, it performs slightly worse than MLR.
* **Ridge Regression** is less preferable due to its lower R-squared and higher error metrics, suggesting less accuracy and higher variability in predictions.
* **Decision Tree** provides a good alternative, especially for capturing non-linear relationships, but it also shows higher error metrics compared to MLR.

**7.DEMOCRATIZING SOLUTION**

To democratize the car price prediction solution, we will develop a user-friendly interface, such as a web or mobile application, allowing users to easily input car features and obtain price predictions. The application will feature interactive visualizations to help users understand the data and predictions better. Comprehensive documentation, tutorials, and training sessions will ensure that users of all technical backgrounds can effectively use the tool. We will provide API integration for seamless incorporation into other systems and gather user feedback to continually improve the tool. Transparency in the model's decision-making process, accessibility for users with disabilities, and strong data privacy and security measures will further support widespread and responsible use of the solution. This approach ensures that advanced predictive analytics are accessible, understandable, and beneficial to a broad audience within the automotive industry