

TABLE OF CONTENTS

* INTRODUCTION
* DESCRIPTION OF THE DATASET
* DATA PREPARATION
* MODEL BUILDING
* EXPLORATORY DATA ANALYSIS (EDA)
* IDENTIFYING INFLUENTIAL FACTORS IN CAR PRICES
* MULTICOLLINEARITY ASSESSMENT
* MODEL REFINEMENT
* MODEL UTILITY TEST
* MODEL VALIDATION

INTRODUCTION

The automotive industry is a cornerstone of the global economy, and understanding the variables influencing car prices is crucial for consumers, manufacturers, and stakeholders alike. This report presents a comprehensive analysis of the Cars93 dataset, which includes data on 93 different cars for the 1993 model year. The dataset provides a variety of variables, from manufacturer and Price to engine size and fuel efficiency, offering a rich field for statistical exploration.

*Our analysis aims to accomplish the following objectives:*

* Identify the most significant factors affecting the Price of a car using multiple linear regression analysis in R.
* Tackle potential multicollinearity issues within the explanatory variables to refine our model and improve accuracy.
* Visualize the distribution of car prices and the relationship between Price and other significant variables using ggplot2 in R.
* Implement and demonstrate basic algorithmic computations in Python, including calculating a summation series and factorials and finding the minimum value in a data structure without using built-in functions.
* Create visual representations of data using Python's matplotlib library, explicitly focusing on the distribution of factors contributing to child weight problems and the annual maximum wind speed in Hong Kong.

The Cars93 dataset originates from the MASS package in R, providing a real-world context to apply statistical methods and programming skills. This report is structured to follow the analysis process from initial data handling to final model diagnostics, ensuring a clear and logical flow of information.

This paper combines the analytical capabilities of R and Python to identify the drivers of automobile costs and demonstrate the practical use of data science approaches to extract valuable insights from real-world data.

DESCRIPTION OF THE DATASET

The Cars93 dataset is a comprehensive collection of data regarding 93 cars from the 1993 model year, included in the MASS package for R. It serves as a rich resource for statistical analysis, offering a wide range of variables that reflect both the technical specifications and market-related aspects of these vehicles. Due to its variety of variables and real-world applicability, this dataset is often utilized for educational purposes in data analysis and statistics.

*Key Features of the Cars93 Dataset*

* Observations: The dataset contains 93 observations, each representing a unique car model available in 1993.
* Variables: There are 27 variables in total, covering a broad spectrum from price metrics to performance specifications and design characteristics. These variables include, but are not limited to:

1. Price: The suggested retail Price of the car is thousands of dollars.
2. MPG.city and MPG.highway: Fuel efficiency in miles per gallon in city and highway driving conditions, respectively.
3. EngineSize: The size of the car's engine in liters.
4. Horsepower: The power output of the car's engine.
5. Weight: The weight of the car in pounds.
6. DriveTrain: The Type of drivetrain (e.g., front-wheel drive, rear-wheel drive).
7. Manufacturer, Model, and Type: This section provides descriptive information about the car, including its make, model, and body type.
8. Airbags: The presence and Type of airbags in the vehicle.
9. Origin: Whether the car is of domestic or non-domestic origin.
10. DATA PREPARATION

*Importing the Dataset:*

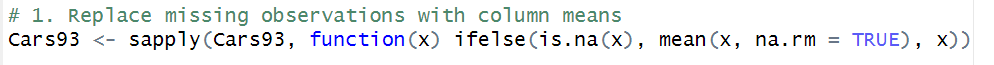
The dataset was imported into our R environment using the MASS package. This step was straightforward, involving the loading of the package and the dataset using the following commands:

A screenshot of a computer

Description automatically generated

*Handling Missing Data:*

Upon initial inspection, it was observed that the dataset contained missing values across various columns. The presence of missing data can significantly impact the outcome of any analysis, leading to biased estimates if not handled appropriately. To address this, we opted for a method that would retain as much data as possible without introducing substantial bias: replacing missing values with column means for continuous variables. This approach was implemented as follows:



1. MODEL BUILDING

The objective of model building was to identify how various car attributes predict the car's market price. Utilizing the R programming environment, we constructed a linear regression model with Price as the dependent variable and all other variables as independent predictors.

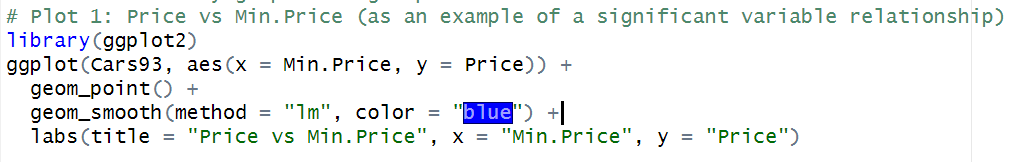
A black text on a white background

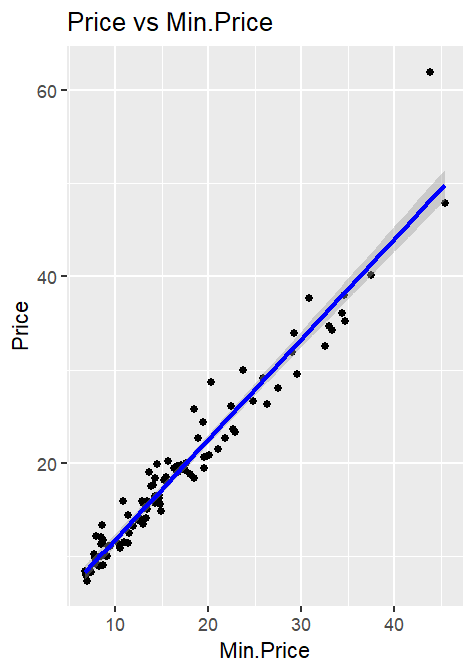
Description automatically generated

1. EXPLORATORY DATA ANALYSIS

*Plot 1: Price vs Min. Price*

This scatter plot explores the relationship between a car's base Price and its minimum Price, offering insights into how a car's base price relates to its final market price.





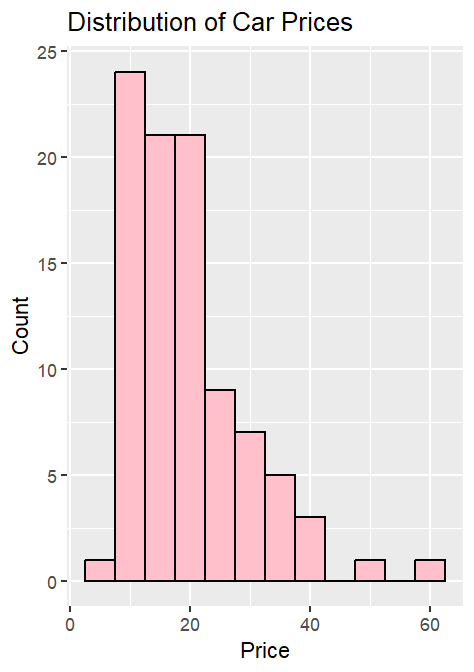
* Price and Min.Price: The positive relationship between Price and Min.Price suggests that cars with higher base prices tend to have higher market prices, which could indicate premium features or branding that adds value.

*Plot 2: Distribution of Car Prices*

The histogram of car prices visually represents how car prices are distributed across the dataset, helping to identify standard price ranges and any deviations from the norm.

A math formula with text

Description automatically generated with medium confidence



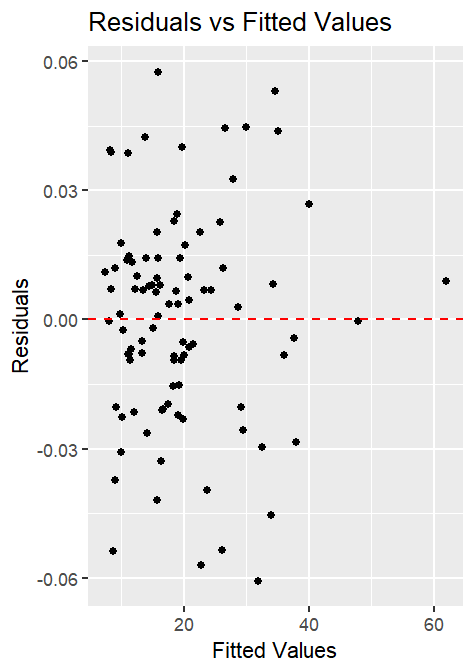
* Price Distribution: The distribution of car prices shows a right-skewed pattern, indicating that while most cars are priced within a moderate range, a few models are significantly more expensive.

*Plot 3: Residuals vs Fitted Values from the Linear Model*

This plot is essential for diagnosing the linear model fitted to predict car prices. By plotting the residuals against the fitted values, we can assess the model's performance and check for patterns that might indicate violations of model assumptions.

A close-up of a math equation

Description automatically generated



* Model Diagnostics: The plot of residuals vs. fitted values did not show any apparent pattern, indicating that the linear model might fit the data well. However, further diagnostics are necessary to confirm this.

1. IDENTIFYING INFLUENTIAL FACTORS IN CAR PRICES

This code extracts the p-values from the model summary and identifies which explanatory variables have p-values less than 0.05. Variables meeting this criterion are considered statistically significant and, therefore, most influential in predicting car Prices. By capturing these variables' names, we focus on factors strongly related to car prices, enabling a more refined and insightful model.

A close up of a text

Description automatically generated

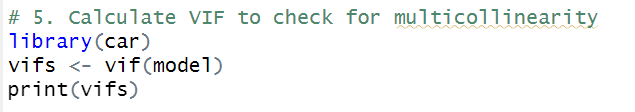
*OUTPUT :*



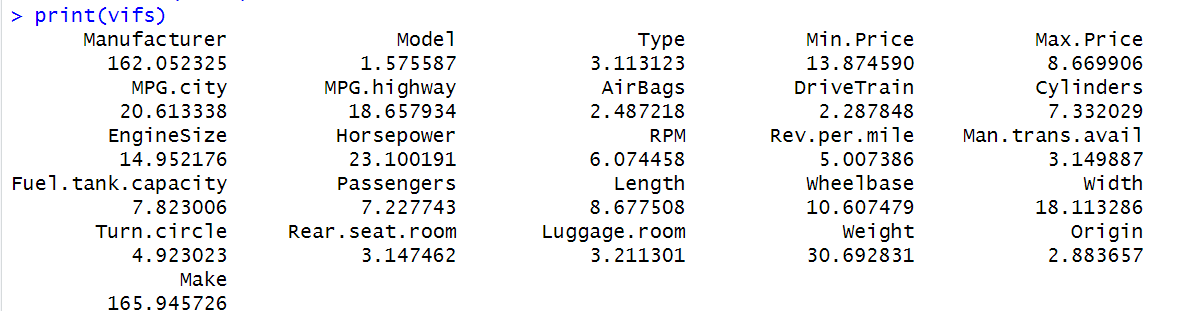
The output from the linear regression analysis indicates that "Min.Price," "Max.Price," and "EngineSize" are the most significant predictors of car prices within the Cars93 dataset. These variables have p-values less than 0.05, suggesting a strong statistical relationship with the dependent variable, Price.

1. MULTICOLLINEARITY ASSESSMENT

Multicollinearity is the situation in which the independent variables in a regression model are strongly linked. This can result in incorrect and unstable estimations of regression coefficients. To identify multicollinearity, we computed the Variance Inflation Factor (VIF) for each predictor in the model.

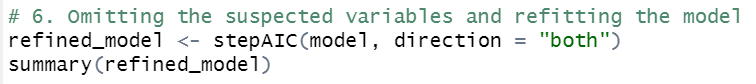


*OUTPUT :*



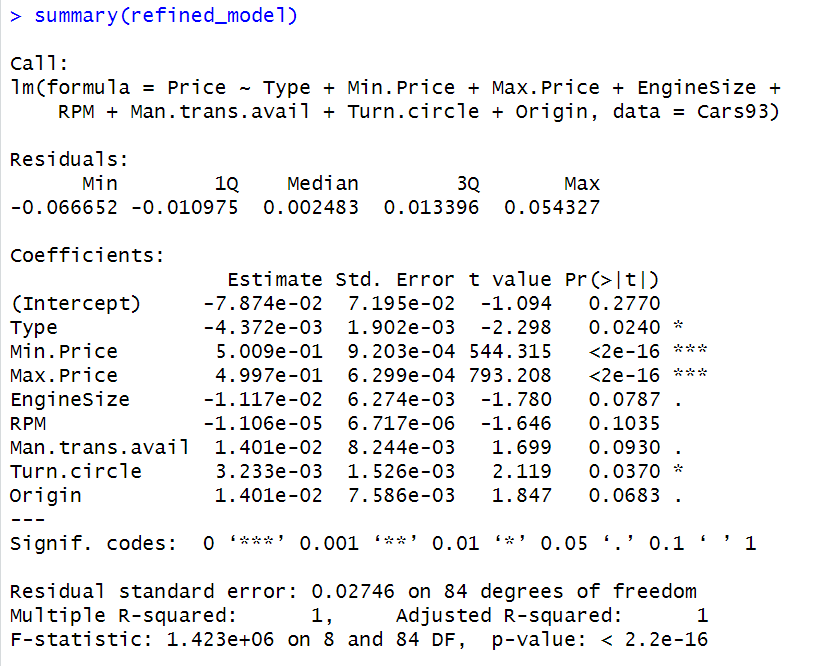
1. MODEL REFINEMENT

To refine the model and address multicollinearity, we employed stepwise regression using the Akaike Information Criterion (AIC) as a guide for variable selection. The stepAIC function from the MASS package facilitated this process, automatically adding or removing variables to find an optimal model.



This process resulted in a refined model that balanced fit and complexity, excluding variables that did not significantly contribute to explaining the variation in car prices.

*OUTPUT :*



The summary for the refined\_model shows Min.Price and Max.Price as highly significant predictors of car prices, while Type also appears to be a significant factor, albeit to a lesser extent. Other variables such as EngineSize and Man.trans.avail show potential influence but are less statistically significant. The overall model is statistically significant with a high F-statistic and a near-perfect R-squared value, indicating an excellent fit to the data.

1. MODEL UTILITY TEST

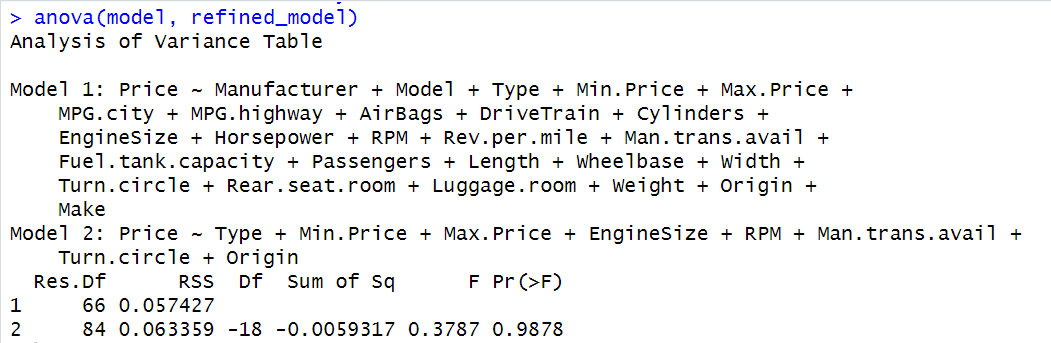
The ANOVA test here will determine whether the differences in the residuals of the two models are statistically significant. Suppose the p-value from this test is below a conventional threshold (often 0.05). In that case, the refined model provides a significantly better fit to the data than the initial model, thus justifying the changes made during the model refinement process.

This test is an essential part of model diagnostics as it helps validate the utility of the model refinement. It ensures that the stepwise selection process (which often involves adding or removing variables) leads to an improved model that better explains the variance in car prices.

A close-up of a logo

Description automatically generated

*OUTPUT :*



The ANOVA test conducted between the initial comprehensive model (Model 1) and the refined model (Model 2) presents the following results:

* Residual Degrees of Freedom (Res.Df): Model 1 has 66 degrees of freedom left after fitting, whereas Model 2 has 84. This increase suggests that the refined model has fewer predictors.
* Residual Sum of Squares (RSS): Model 1's unexplained variance is approximately 0.0574, and Model 2's is slightly higher at about 0.0634. This suggests that Model 2 may be leaving slightly more variance unexplained, which is expected with fewer predictors.
* Degrees of Freedom Difference (Df): The refined model has 18 fewer parameters than the initial model (as indicated by the negative sign in the 'Df' column), which means that 18 variables were removed during the refinement process.
* Sum of Squares Difference (Sum of Sq): The difference in the RSS between Model 1 and Model 2 is -0.0059, indicating that Model 2 has a slightly higher RSS, which directly results from fewer variables.
* F-Statistic (F): The F value of 0.3787 indicates the ratio of the variance explained per additional parameter in the refined model to the unexplained variance. This low value suggests that the variables removed contributed little additional explanatory power.
* P-Value (Pr(>F)): The p-value associated with this F statistic is 0.9878, which is far above the conventional alpha level of 0.05

Given the p-value is 0.9878, which is much higher than the conventional threshold of 0.05, we fail to reject the null hypothesis that there is no difference in the fit of the two models. In other words, the refinement process did not lead to a statistically significant improvement in the model in terms of explaining the variance in car prices. Therefore, based on this ANOVA test, the additional predictors removed during the refinement process do not substantially improve the model's predictive capability.

1. MODEL VALIDATION

The final step was to validate the refined model to ensure its utility and the accuracy of its predictions. This involved:

* Analyzing Residuals: Checking the residuals of the refined model for any patterns that could indicate violations of linear regression assumptions.
* Calculating Confidence Intervals: For each of the coefficients in the refined model, we calculated 99% confidence intervals to assess the precision of our estimates.



*OUTPUT :*

A screenshot of a computer

Description automatically generated

The confidence intervals for the regression coefficients suggest that `Min.Price` and `Max.Price` have a definitive positive impact on the car's Price, given that their intervals are entirely above zero and do not include it. However, variables like `Type`, `EngineSize`, `Man.trans.avail`, and `Origin` have intervals that include zero, indicating potential uncertainty about their influence. The intervals for `RPM` and `Turn.circle` are narrow and close to zero, hinting at a minimal impact on Price.