Statistical Analysis of Used Cars Data

Welbeck Achiampong, Muhyadin Yusuf, Minh Pham, and Jordan Addo

oday

Table of Contents

# set your global options here and load your packages  
knitr::opts\_chunk$set(fig.width = 10, fig.height = 5, echo = TRUE, eval = TRUE)  
library(knitr)  
library(tidyverse)  
used\_cars <- read.csv("used\_cars\_data.csv") #redirect

# Introduction

This project aims to explore the used car market by analyzing the factors that influence the pricing of used cars. With the increasing demand for used vehicles, understanding how various features like the manufacturer, model, year, engine type, and kilometers driven impact the price can help consumers and dealers alike make informed decisions. The dataset, used\_cars\_data.csv, contains information on these attributes, offering insights into how different features contribute to the sale price of a car. By developing a predictive model, we can provide a more accurate estimate of a used car’s value based on its characteristics, making this research valuable for both buyers and sellers in the used car market.

To address this, we will build a linear regression model to estimate the price based on these features. The model will follow the general equation where is the response variable(pricted price) and the represents the coefficients and the are the input variables. Through pre-processing the data and ensuring it is clean and normalized yo be able to accurately predict the price of used cars.

## Reserach questions

Our research question is to determine whether a model can effectively predict the price of a used car using input variables such as the fuel type, mileage, power etc.

## Data set desription

The dataset, used\_cars\_data.csv, contains information about used cars and their sale prices. There are several variables present in the dataset, including:

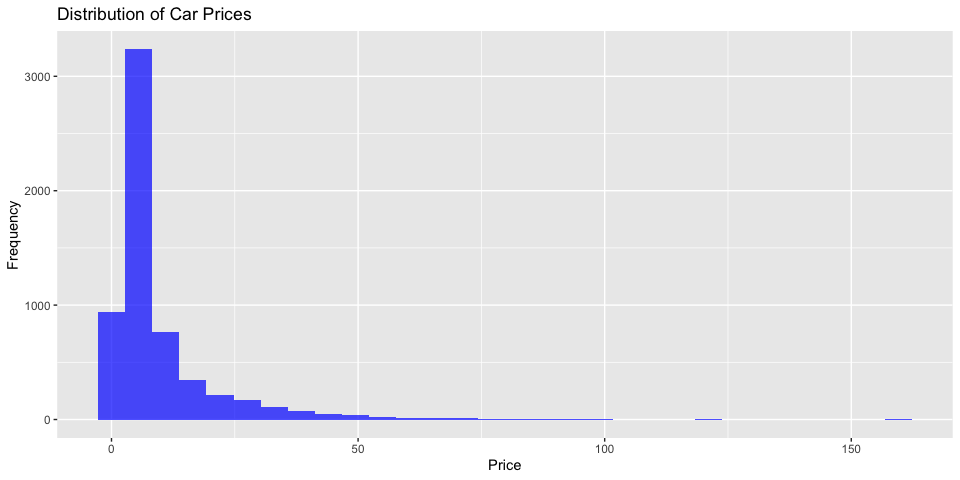
* Manufacturer: the car manufacturer (e.g. Toyota, Ford, Honda, etc.)
* Model: the specific car model (e.g. Corolla, Mustang, Accord, etc.)
* Location: The place of the car (e.g. Mumbai, Pune, Jaipur, etc.)
* Year: the year the car was released (e.g. 2010, 2014, 2015, etc.)
* Engine: the number of cylinders in the engine
* Fuel\_type : the type of fuel the car uses (e.g. gas, diesel, electric)
* Kilometer\_Driven: the total distance the car has traveled
* Transmission: the type of transmission (e.g. automatic, manual)

Using this data, we can develop a model to predict the sale price of a used car based on these variables.

# Exploratory Data Analysis

ggplot(used\_cars, aes(x = Price)) +  
 geom\_histogram(bins = 30, fill = "blue", alpha = 0.7) +  
 labs(title = "Distribution of Car Prices", x = "Price", y = "Frequency")

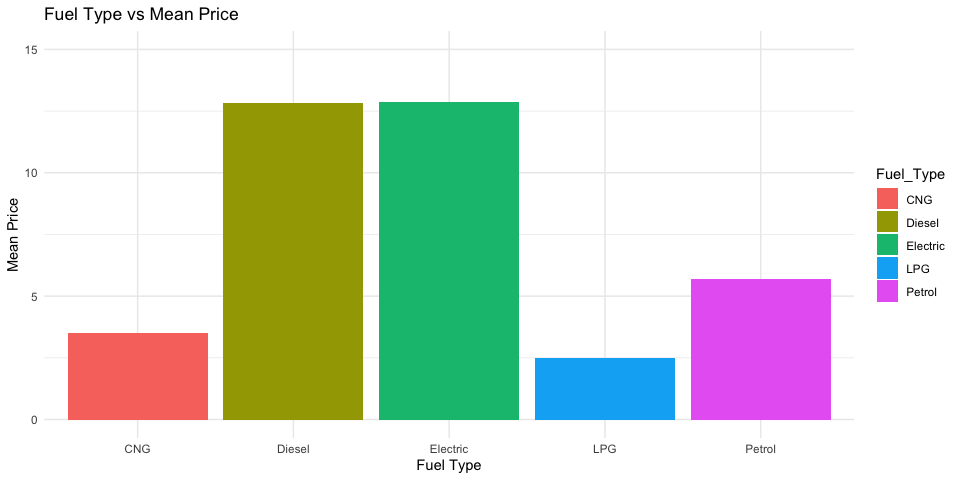
## Warning: Removed 1234 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

 In our exploratory of the data, we observe that the price range of most used cars is around $10,000 USD.

levels(used\_cars$Fuel\_Type)

## NULL

summarized\_data <- used\_cars %>%  
 group\_by(Fuel\_Type) %>%  
 summarise(mean\_price = mean(Price, na.rm = TRUE))  
  
ggplot(summarized\_data, aes(x = Fuel\_Type, y = mean\_price, fill = Fuel\_Type)) +  
 geom\_bar(stat = "identity") + # Use 'identity' since we're passing the pre-calculated values  
 ylim(0,15) +  
 labs(title = "Fuel Type vs Mean Price", x = "Fuel Type", y = "Mean Price") +  
 theme\_minimal()

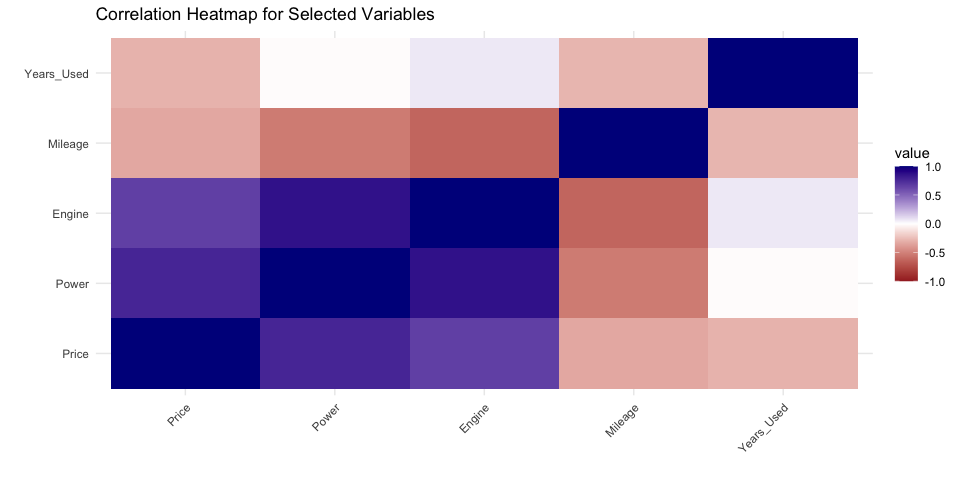
 Need explanation:

# Load necessary libraries  
library(ggplot2)  
library(reshape2)

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

library(dplyr)  
  
# Assuming the 'Mileage' column needs cleaning  
used\_cars$Mileage <- as.numeric(gsub("[^0-9.]", "", used\_cars$Mileage))  
used\_cars$Engine <- as.numeric(gsub("[^0-9.]", "", used\_cars$Engine))  
used\_cars$Power <- as.numeric(gsub("[^0-9.]", "", used\_cars$Power))  
  
# Calculate 'Years\_Used'  
used\_cars$Years\_Used <- 2024 - used\_cars$Year  
  
selected\_data <- used\_cars[, c("Price", "Power", "Engine", "Mileage", "Years\_Used")]  
  
  
selected\_data <- na.omit(selected\_data)  
  
cor\_matrix <- cor(selected\_data, use = "complete.obs")  
  
cor\_melted <- melt(cor\_matrix)  
  
ggplot(cor\_melted, aes(Var1, Var2, fill = value)) +  
 geom\_tile() +  
 scale\_fill\_gradient2(low = "brown", mid = "white", high = "darkblue", midpoint = 0, limits = c(-1, 1)) +  
 theme\_minimal() +  
 labs(title = "Correlation Heatmap for Selected Variables", x = "", y = "") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



Jordan :Need explain the correlation graph The correlation heatmap highlights the relationships between different factors and vehicle price, providing insights into what affects pricing the most. Engine size and power show strong positive correlations with price, meaning cars with bigger engines and more horsepower tend to have higher prices. This is expected, as vehicles with better performance often come at a premium price. On the other hand, kilometers driven has a clear negative correlation with price—indicating that cars with more kilometers on the odometer tend to sell for less. This shows the wear and tear that comes with higher mileage. Years used has a weak correlation with price. This suggests that a car’s age alone doesn’t have a significant effect on its value. It seems that other factors, like mileage and condition, are more influential when determining a car’s price. Overall, the graph shows that engine performance and usage are the main drivers of vehicle price, while age plays a much smaller role.

# Statistical Methods

Regression

Regression is a statistical method used to understand the relationship between one dependent variable and one or more independent variables. This model helps us to quantify how a change in the independent variables are linked with changes in the dependent variable, making this model a tool to help us make predictions and analysis. For example in our project, the dependent variable is the price while the independent variables are (Kilometers\_Driven, Fuel\_Type, Mileage, Transmission, Power, etc.). So we are using the multi linear regression model to see how each independent variable effects the price of used cars.

The statistical technique we plan to use is a multi linear regression model to answer our regression question. The model could be expressed as:

# Results

## Regression

library(dplyr)  
library(ggplot2)  
  
#used\_cars <- read.csv("used\_cars\_data.csv")  
  
used\_cars$Mileage <- as.numeric(gsub(" km/kg| kmpl", "", used\_cars$Mileage))  
used\_cars$Engine <- as.numeric(gsub(" CC", "", used\_cars$Engine))  
used\_cars$Power <- as.numeric(gsub(" bhp", "", used\_cars$Power))  
  
  
used\_cars <- used\_cars %>%  
 mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm = TRUE), .)))  
  
used\_cars$Fuel\_Type <- as.factor(used\_cars$Fuel\_Type)  
used\_cars$Transmission <- as.factor(used\_cars$Transmission)  
used\_cars$Location <- as.factor(used\_cars$Location)  
  
used\_cars$Year <- factor(used\_cars$Year)  
  
unique\_years <- sort(unique(used\_cars$Year))  
used\_cars$Year <- factor(used\_cars$Year, levels = unique\_years)  
  
model <- lm(Price ~ Fuel\_Type + Mileage + Transmission + Power, data = used\_cars)  
summary(model)

##   
## Call:  
## lm(formula = Price ~ Fuel\_Type + Mileage + Transmission + Power,   
## data = used\_cars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -63.629 -2.491 -0.126 2.030 131.532   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.007745 1.164819 -3.441 0.000584 \*\*\*  
## Fuel\_TypeDiesel 0.286751 0.924755 0.310 0.756507   
## Fuel\_TypeElectric 7.647819 5.144294 1.487 0.137148   
## Fuel\_TypeLPG 0.466173 2.259530 0.206 0.836551   
## Fuel\_TypePetrol -1.525186 0.929188 -1.641 0.100754   
## Mileage 0.134369 0.022471 5.980 2.34e-09 \*\*\*  
## TransmissionManual -3.411968 0.242876 -14.048 < 2e-16 \*\*\*  
## Power 0.119207 0.002399 49.689 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.148 on 7245 degrees of freedom  
## Multiple R-squared: 0.5183, Adjusted R-squared: 0.5178   
## F-statistic: 1113 on 7 and 7245 DF, p-value: < 2.2e-16

predicted\_prices <- predict(model, used\_cars)  
used\_cars$Predicted\_Price <- predicted\_prices  
head(used\_cars)

## S.No. Name Location Year Kilometers\_Driven  
## 1 0 Maruti Wagon R LXI CNG Mumbai 2010 72000  
## 2 1 Hyundai Creta 1.6 CRDi SX Option Pune 2015 41000  
## 3 2 Honda Jazz V Chennai 2011 46000  
## 4 3 Maruti Ertiga VDI Chennai 2012 87000  
## 5 4 Audi A4 New 2.0 TDI Multitronic Coimbatore 2013 40670  
## 6 5 Hyundai EON LPG Era Plus Option Hyderabad 2012 75000  
## Fuel\_Type Transmission Owner\_Type Mileage Engine Power Seats New\_Price Price  
## 1 CNG Manual First 26.60 998 58.16 5 1.75  
## 2 Diesel Manual First 19.67 1582 126.20 5 12.50  
## 3 Petrol Manual First 18.20 1199 88.70 5 8.61 Lakh 4.50  
## 4 Diesel Manual First 20.77 1248 88.76 7 6.00  
## 5 Diesel Automatic Second 15.20 1968 140.80 5 17.74  
## 6 LPG Manual First 21.10 814 55.20 5 2.35  
## Years\_Used Predicted\_Price  
## 1 14 3.087564  
## 2 9 10.553976  
## 3 13 4.074259  
## 4 12 6.238676  
## 5 11 15.105737  
## 6 12 2.461858

Write estimated final models. Give details. Interpret models and/or coefficients. What do your models say? Do the two models send the same message? What are the important inputs?

## Final Estimated Model

The estimated model for the price of a used car considering the different variables such as Fuel type(Diesel,Electric,LPG,Petrol), Mileage, Transmission and power. The co-efficient in out model tells us the change in our price by one unit. In other words, as the price of a used car increases, the input variables changes based on their respective co-efficient. The model consist of few on our numerical, our residual standaed error calculated was on average 7.148 with 7245 degrees of freedom. the multiple R - squared calculated was 0.5178 which indicates the proportion of variance explained by our predictors. Our Adjusted R-squared which penalize the model for adding imput variables that do not improve our model. Our p- value(< 2.2e-16) indicates that the model is statistically significant.

## Model Interpertation

anova(model)

## Analysis of Variance Table  
##   
## Response: Price  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Fuel\_Type 4 65054 16264 318.32 < 2.2e-16 \*\*\*  
## Mileage 1 76151 76151 1490.48 < 2.2e-16 \*\*\*  
## Transmission 1 130857 130857 2561.23 < 2.2e-16 \*\*\*  
## Power 1 126144 126144 2468.97 < 2.2e-16 \*\*\*  
## Residuals 7245 370159 51   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

levels(used\_cars$Fuel\_Type)

## [1] "CNG" "Diesel" "Electric" "LPG" "Petrol"

From our analysis table, the Sum of squares explains the variation in the sale price of used car for each variable, Transmission has a large amount of variation with a sum of squares of 130,857 indicating that the variable transmission explains more variation of the price of a used car. Similarly, Power also has slightly large amount of variation with sum of squares of 126,144. Mileage and fuel type also as some significance of variation in explaining of our response variable(Price).

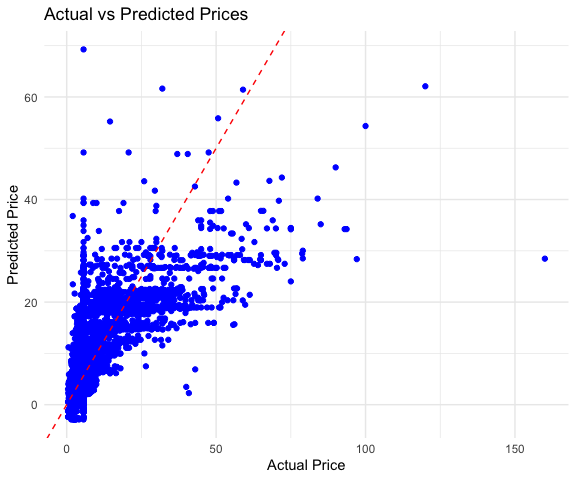
Each of the mean square of our variable was calculated by dividing our sum of squares by the degree of freedom which helps us understand how much variation each factor explains per degree of freedom. Overall each of our variables tells us how much variation each factors explains per degree of freedom.

The F-value is the ratio of the mean square of each variable to the mean square of the residuals. A higher F-value indicates a more significant effect on the price of a used car. As an illustration, the F-value for the variable transmission is 2561.23, suggesting that transmission has a highly significant effect on the price of a used car, similar, Power, Mileage and Fuel type also have some significance effect on the price of used cars.

Similar to F-value, the p-value also tells us whether the variable has a statistically significant effect on the price of used cars. A very small p-value which is less than 0.05 indicates that our predicted variable has statistically significant impact on the price. We observe our p-value for all of our variables(Fuel type, Mileage, Transmission and Power), we have highly statistical significant impact on price because their p - value are less than 0.05, suggesting that they each tell us some significance affect on the price of a used car.

In conclusion, the key drivers of used car prices are Fuel Type, Mileage, Transmission, and Power, all of which significantly affect pricing.

ggplot(used\_cars, aes(x = Price, y = predicted\_prices)) +  
 geom\_point(color = "blue") +  
 geom\_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +  
 ggtitle("Actual vs Predicted Prices") +  
 xlab("Actual Price") +  
 ylab("Predicted Price") +  
 theme\_minimal()

 This scatter plot shows the relationship between the actual price and the predicted price of used cars, with the actual price on the x-axis and the predicted price on the y-axis. Each blue dot represents a data point, which corresponds to a specific car’s actual and predicted price. The red dashed line represents the line of perfect prediction, where the predicted price would exactly match the actual price (i.e., the line where ). Notice most of the data points are clustered near the lower values suggesting that the predictive model makes relatively accurate prediction for lower priced cars, but is less accurate for higher priced cars. For the points over or under the the red line, this could be factors of not including specific variables in our data, or the model has over predicted oor under predicted the price compared to the actual value. Overall, while the model performs well for lower-priced cars, it seems to struggle with accurately predicting the prices of higher-end cars.

# Conclusions

Overall, Our research on the used car market helped us to discover some of the features that is considered when determining the price of a used car. By building a multiple linear regression model, we were able to discover that fuel type, mileage, transmission and enginepower are the influence in predicting the price of used cars. Notably, engine power and transmission type were found to be critical, with automatic transmission and higher engine power associated with higher vehicle prices. dditionally, while mileage has a negative correlation with price, indicating that cars with higher kilometers tend to be priced lower, the car’s age (years used) showed only a modest effect on pricing, suggesting that other factors like engine condition and mileage play a more significant role.

The model achieved reasonable predictive accuracy with an R-squared value of 0.518, indicating that approximately 51.8% of the variance in used car prices can be explained by the variables included in the model. This makes it a useful tool for buyers, sellers, and analysts in estimating fair market prices. While the model performs well, future enhancements could include additional variables such as the car’s condition, accident history, or even regional pricing differences to further improve its predictive power. This research offers valuable contributions to the understanding of used

# Appendix A

[Kaggle Used Car Dataset](https://www.kaggle.com/datasets/ayushparwal2026/cars-dataset?resource=download)

[Introduction to Data Science:rafalab.dfci.harvard.edu/dsbook/regression.html](https://rafalab.dfci.harvard.edu/dsbook/regression.html)

[Introduction to Statistical Learning](https://www.statlearning.com)

[Car Resale Value Prediction](https://www.enjoyalgorithms.com/blog/car-resale-value-predictor-using-random-forest-regressor)

[Resale Value Car Price](https://carchase.com.au/resources/car-valuation-guide/9-factors-that-impact-the-value-of-your-car/)