Statistical Analysis of Used Cars Data

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

October 23, 2024

Contents

List of Figures

List of Tables

```r
# 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
```

1 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 $Y = B_0 + B_1 x_1 + B_2 x_2$ where Y is the response variable(pricted price) and the $B's$ represents the coefficients and the $x's$ 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.

1.1 Reserach questions

Our research question is to determine whether a model can effectively predict the price of a used car using variables such as the year of the car, kilometers driven, engine capacity, and fuel type.

1.2 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.)

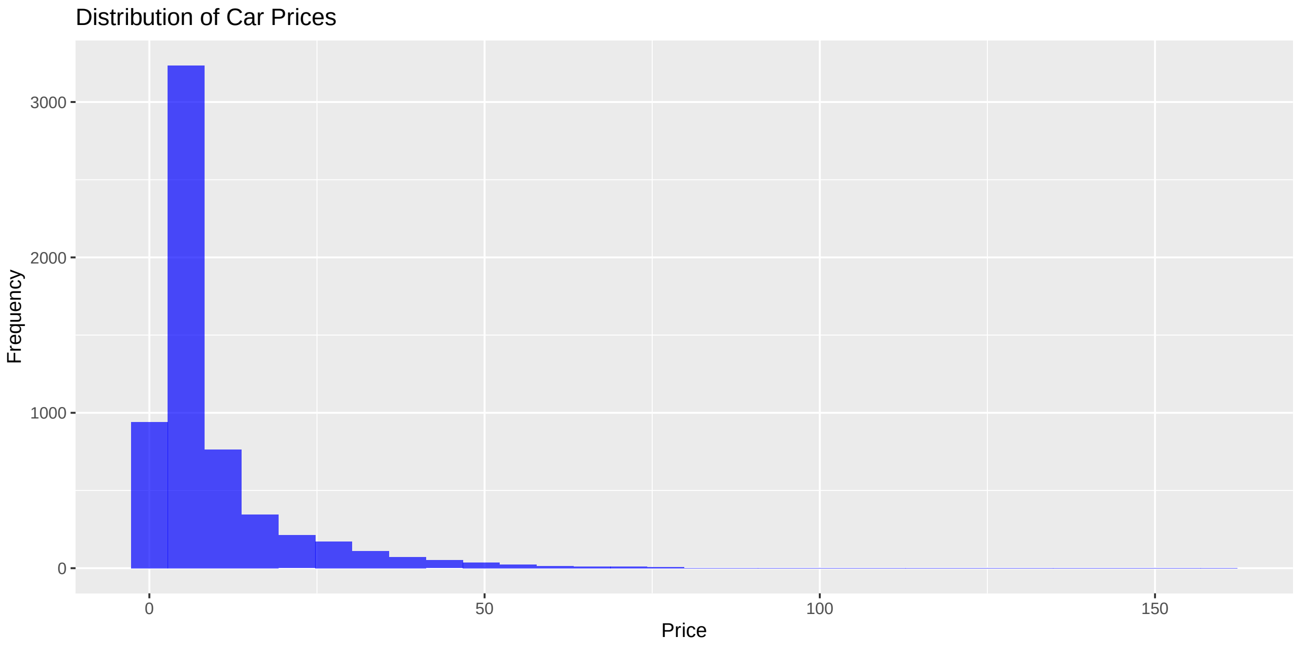
- Location: The place of the car (e.g. Mumbai, Pune, Jaipur, etc.)
- Year: the year the car was released (e.g. 2010, 2014, 2015, etc.)
- Engine: the number of cylinders in the engine
- Fuel_type : the type of fuel the car uses (e.g. gas, diesel, electric)
- Kilometer_Driven: the total distance the car 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.

2 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()‘).
```



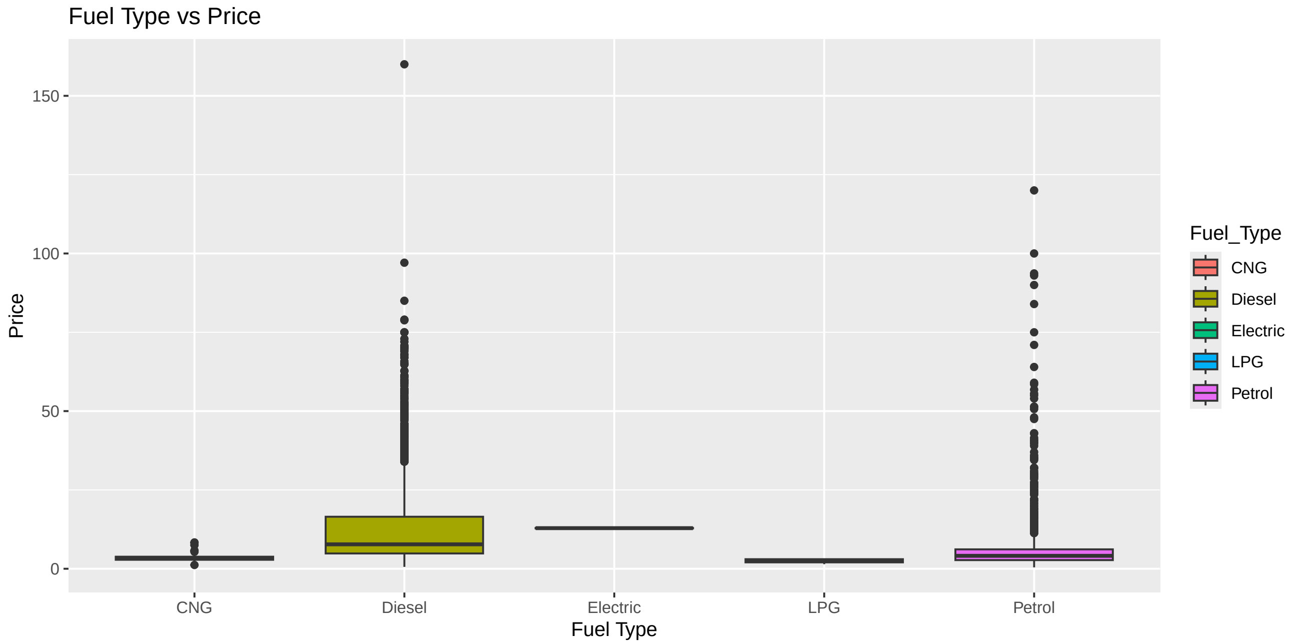
In our exploratory of the data, we observe that the price range of most used cars is around $10,000 USD.

```r
levels(used_cars$Fuel_Type)
```

```
## NULL
```

```r
ggplot(used_cars, aes(x = Fuel_Type, y = Price, fill = Fuel_Type)) +
  geom_boxplot() +
  labs(title = "Fuel Type vs Price", x = "Fuel Type", y = "Price")
```

```
## Warning: Removed 1234 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```

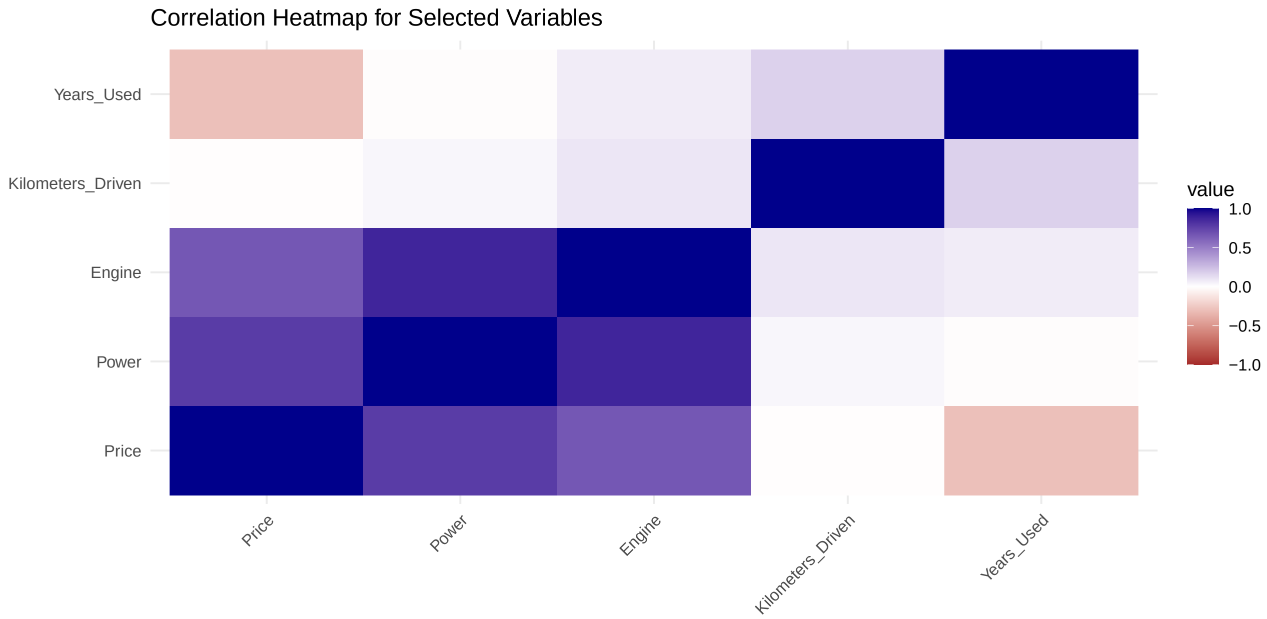


Need explanation:

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

```
##
## Attaching package: 'reshape2'

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

```r
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", "Kilometers_Driven", "Years_Use


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, lim
  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

3 Statistical Methods

3.1 Regression

What are the two statistical techniques you plan to use to answer your regression question? Give details here. Name variables, write out models (using β_i's), let me know if you're using CV, backward selection, etc. Write formulas. Explain why you chose to use these two techniques.

Explain Regression(Jordan) Regression is a statistical method used to understand the relationship between different variables.The goal of regression is to understand how the dependent variable changes when the independent variables are varied. This helps in predicting future values of the dependent variable. Our statistical approach will be to create a linear regression model.

4 Results

4.1 Regression

```
library(dplyr)
library(ggplot2)
```

```r
#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 ~  Kilometers_Driven  + Fuel_Type +  Mileage + Transmission + Power, dat
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ Kilometers_Driven + Fuel_Type + Mileage +
##     Transmission + Power, data = used_cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.635  -2.479  -0.102   2.007 131.384
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -3.636e+00  1.168e+00  -3.112 0.001864 **
## Kilometers_Driven     -3.666e-06  1.007e-06  -3.642 0.000273 ***
## Fuel_TypeDiesel        3.079e-01  9.240e-01   0.333 0.738975
## Fuel_TypeElectric   7.613e+00  5.140e+00   1.481 0.138613
## Fuel_TypeLPG        4.350e-01  2.258e+00   0.193 0.847217
## Fuel_TypePetrol    -1.595e+00  9.286e-01  -1.718 0.085927 .
## Mileage             1.258e-01  2.258e-02   5.572 2.61e-08 ***
## TransmissionManual -3.360e+00  2.431e-01 -13.823  < 2e-16 ***
## Power               1.191e-01  2.397e-03  49.660  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.142 on 7244 degrees of freedom
## Multiple R-squared:  0.5191, Adjusted R-squared:  0.5186
## F-statistic: 977.6 on 8 and 7244 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        3.009842
## 2          9       10.660176
## 3         13        4.089619
## 4         12        6.172488
## 5         11       15.197553
## 6         12        2.389628
```

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?

4.2 Final Estimated Model

$Price = -3.636e+00 - 3.666e-06 * Kilometers_Driven + 3.079e-01 * Fuel_TypeDiesel + 7.613e+00 * Fuel_TypeElectric + 4.350e-01 * Fuel_TypeLPG - 1.595e+00 * Fuel_TypePetrol + 1.258e-01 * Mileage - 3.360e+00 * TransmissionManual + 1.191e-01 * Power + \epsilon$

The estimated model for the price of a used car considering the different variables such as Kilometer Driven, 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 imput variables changes based on their respective co-efficient.

4.3 Model Interpertation

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Price
##                     Df Sum Sq Mean Sq   F value Pr(>F)
## Kilometers_Driven    1     94      94    1.8434 0.1746
## Fuel_Type            4  66417   16604  325.5408 <2e-16 ***
## Mileage              1  78520   78520 1539.4475 <2e-16 ***
## Transmission         1 128069  128069 2510.9031 <2e-16 ***
## Power                1 125783  125783 2466.0773 <2e-16 ***
## Residuals         7244 369482      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, our Sum of squares explains the variation by each variable, Transmission has a large samount of variation with a sum of squares of 128,069 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 125,783. Mileage and fuel type also as some significance of variation in explaining of our response variable(Price). Lastly Kilometer driven xplains very little of the variation of the price of used cars.

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.

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 2510.90, 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. Kilometer driven on the other hand, has a small F-value of 1.84, suggesting there's a small significant effect on price.

Similar to F-value, the p-value also tells us whether the variable has a statistically sihnificant 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 kilometer driven to be 0.1746 which is greater than 0.05, suggesting that kilometer driven does not significantly impact the price of used cars. All of the other variables(Fuel type, Mileage, Transmission and Power) have highly statistical significant impact on price because their p - value are less than 0.05.

In conclusion, Fuel_Type, Mileage, Transmission, and Power are key drivers of Price. Kilometers_Driven does not significantly affect Price, as indicated by its high p-value (0.1746).

Make plots of your models if possible.

```r
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()
```

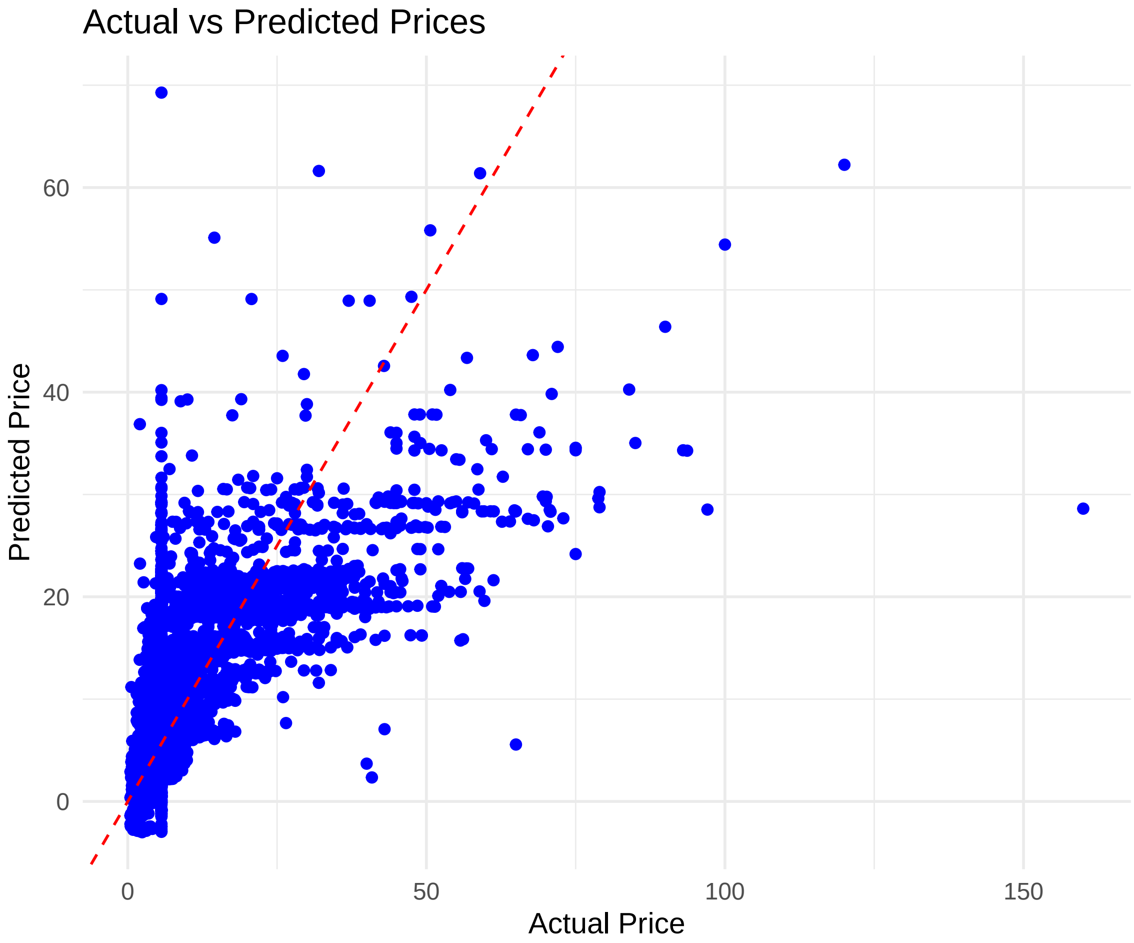


Figure 1: The relationship between Actual and Predicted Prices

5 Conclusions

What did you learn? What else would you have wanted to know but couldn't? Is something looking weird / surprising / unexpected?

Don't use formulas here or be too statistical. This section is for the wider audience.

Everybody should write three to four sentences.

6 Appendix A

Introduction to Data Science:rafalab.dfci.harvard.edu/dsbook/regression.html

Kaggle Used Car Dataset

Introduction to Statistical Learning