Statistical Analysis of Used Cars Data

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

oday

Contents

List of Figures

```r
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")
```

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 $Y = B_0 + B_1x_1 + B_2x_2$ 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 input variables such as the fuel type, mileage, power etc.

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.)

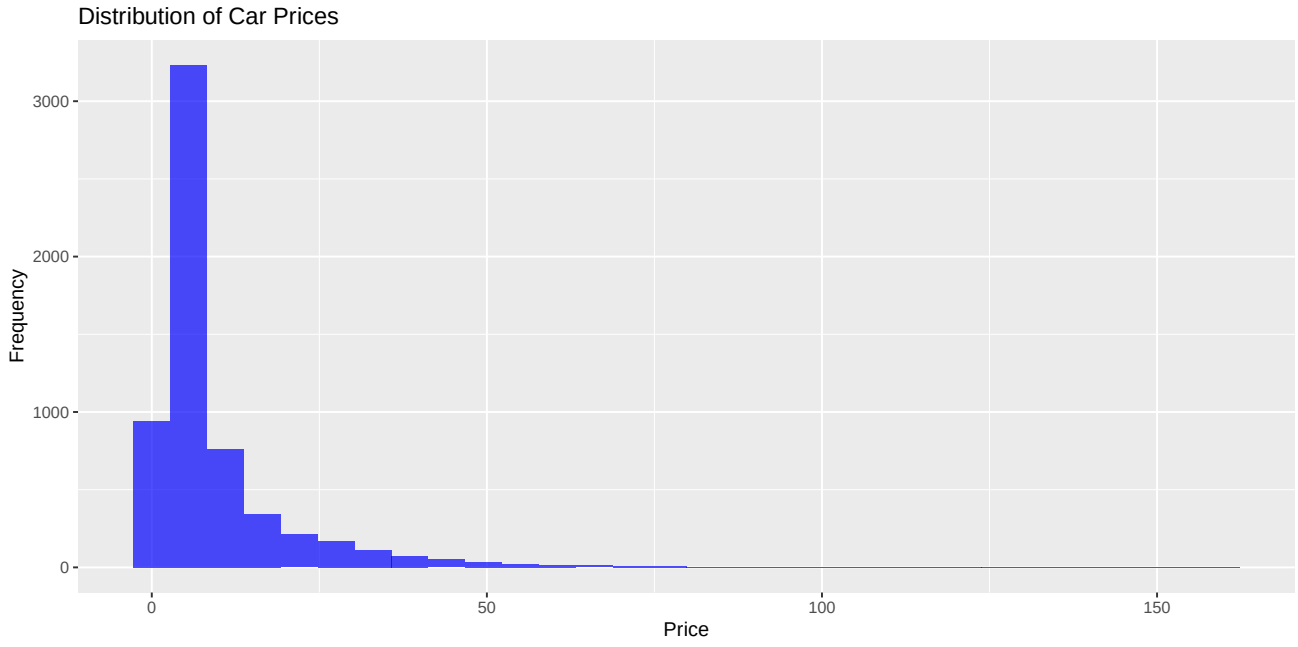
- Year: the year the car was released (e.g. 2010, 2014, 2015, etc.)
- Engine: the number of cylinders in the engine
- Fuel_type : the type of fuel the car uses (e.g. gas, diesel, electric)
- Kilometer_Driven: the total distance the car has traveled
- Transmission: the type of transmission (e.g. automatic, 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 the price on the x-axis and the number of cars (frequency) on the y-axis. We observe that the price range of most used cars is around $10,000 USD.

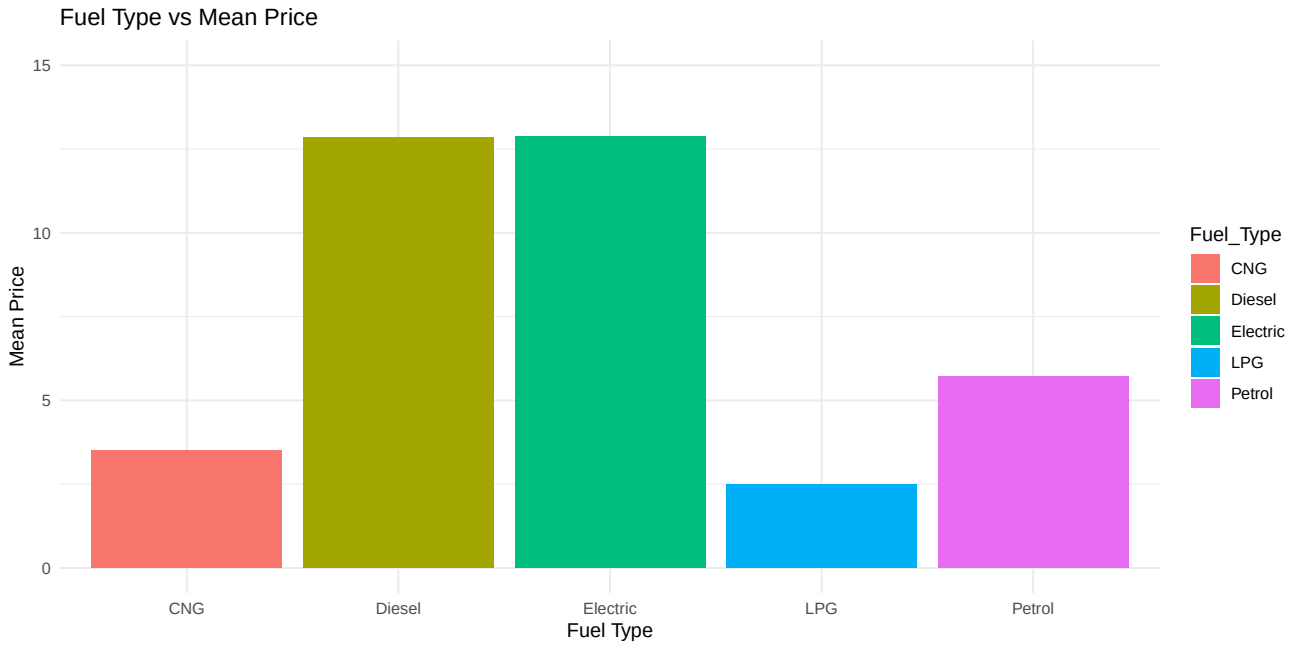
```r
levels(used_cars$Fuel_Type)
```

```
## NULL
```

```r
summarized_data <- used_cars %>%
  group_by(Fuel_Type) %>%
  summarise(mean_price = mean(Price, na.rm = TRUE))
summarized_data
```

```
## # A tibble: 5 x 2
##   Fuel_Type mean_price
##   <chr>          <dbl>
## 1 CNG             3.52
## 2 Diesel         12.8
## 3 Electric       12.9
## 4 LPG             2.49
## 5 Petrol          5.70
```

```r
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
  ylim(0,15) +
  labs(title = "Fuel Type vs Mean Price", x = "Fuel Type", y = "Mean Price") +
  theme_minimal()
```



The bar chart shows the mean price of vehicles for different fuel types (CNG, Diesel, Electric,

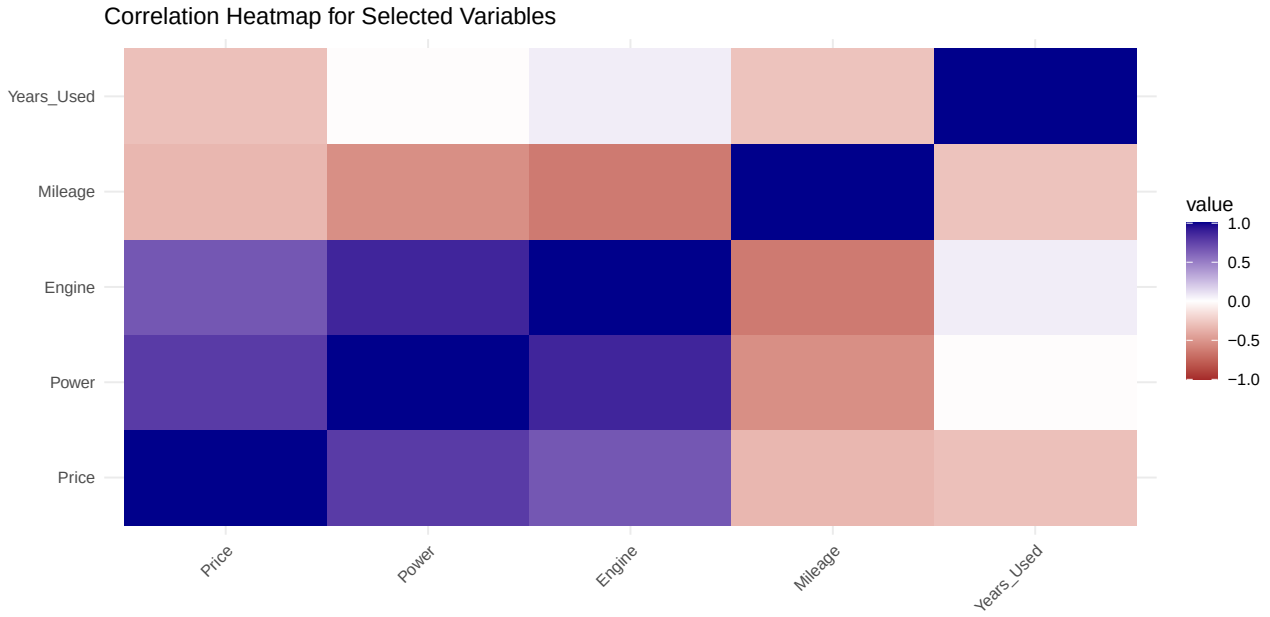
LPG, and Petrol). Electric vehicles lead with the highest average price, likely due to their advanced battery technology and the growing market demand for environmentally friendly transportation. This reflects the higher initial costs associated with producing and purchasing electric vehicles compared to conventional fuel types. Diesel vehicles follow closely behind, with a high mean price attributed to their fuel efficiency and durability, making them more valuable, particularly for long-distance driving. Diesel cars are often favored in markets where longevity and fuel economy are prioritized. Petrol vehicles sit in the middle, with a lower average price than both electric and diesel vehicles. This reflects their widespread availability and affordability, offering a balance between performance and cost that appeals to a broad consumer base. LPG and CNG vehicles have the lowest mean prices, with CNG being the most affordable. These fuel types are typically more economical and are favored for their cost-effectiveness in certain markets, although they are less common. The lower demand and reduced market value for these vehicles are reflected in their lower average prices. In summary, electric and diesel vehicles command higher prices due to technological advancements and performance benefits, while petrol, CNG, and LPG vehicles offer more budget-friendly options.

```
library(ggplot2)
library(reshape2)
```

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

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

```
library(dplyr)


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))


used_cars$Years_Used <- 2024 - used_cars$Year

selected_data <- used_cars[, c("Price", "Power", "Engine", "Mileage", "Years_Used")]
```

```r
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))
```



The correlation heatmap illustrates the relationships between several variables in the dataset. There is a strong positive correlation between years used and mileage, indicating that older cars tend to have more kilometers driven, which is expected as vehicles accumulate usage over time. Both years used and mileage have negative correlations with price, showing that as a car gets older and its mileage increases, its value usally decreases, with mileage having a slightly stronger impact. On the other hand, engine size and power have a positive correlation, as cars with larger engines tend to produce more power. Both of these variables also shows strong positive correlations with price, suggesting that vehicles with larger engines and higher horsepower are generally priced higher due to their performance capabilities. There is a weak correlation between

mileage and engine size, suggesting that engine size doesn't significantly impact how much a vehicle is driven. Overall, the heatmap highlights how engine performance factors positively influence price, while usage-related factors like mileage and years used negatively affect a vehicle's value.

3 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:

```
$Y= B_0 + B_1 X_1 + B_2X_2... B_nX_n + \epsilon$
```

Where Y is the dependent variable, B_i are the coefficients and $X's$ are the input or independent variables.

4 Results

4.1 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))
```

```r
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?

4.2 Final Estimated Model

$Price = -4.007745 + 0.286751 * Fuel_TypeDiesel + 7.647819 * Fuel_TypeElectric + 0.466173 * Fuel_TypeLPG - 1.525186 * Fuel_TypePetrol + 0.134369 * Mileage - 3.411968 * TransmissionManual + 0.119207 * Power + \epsilon$

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.

4.3 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 $y = x$). Notice most of the data points are clustered

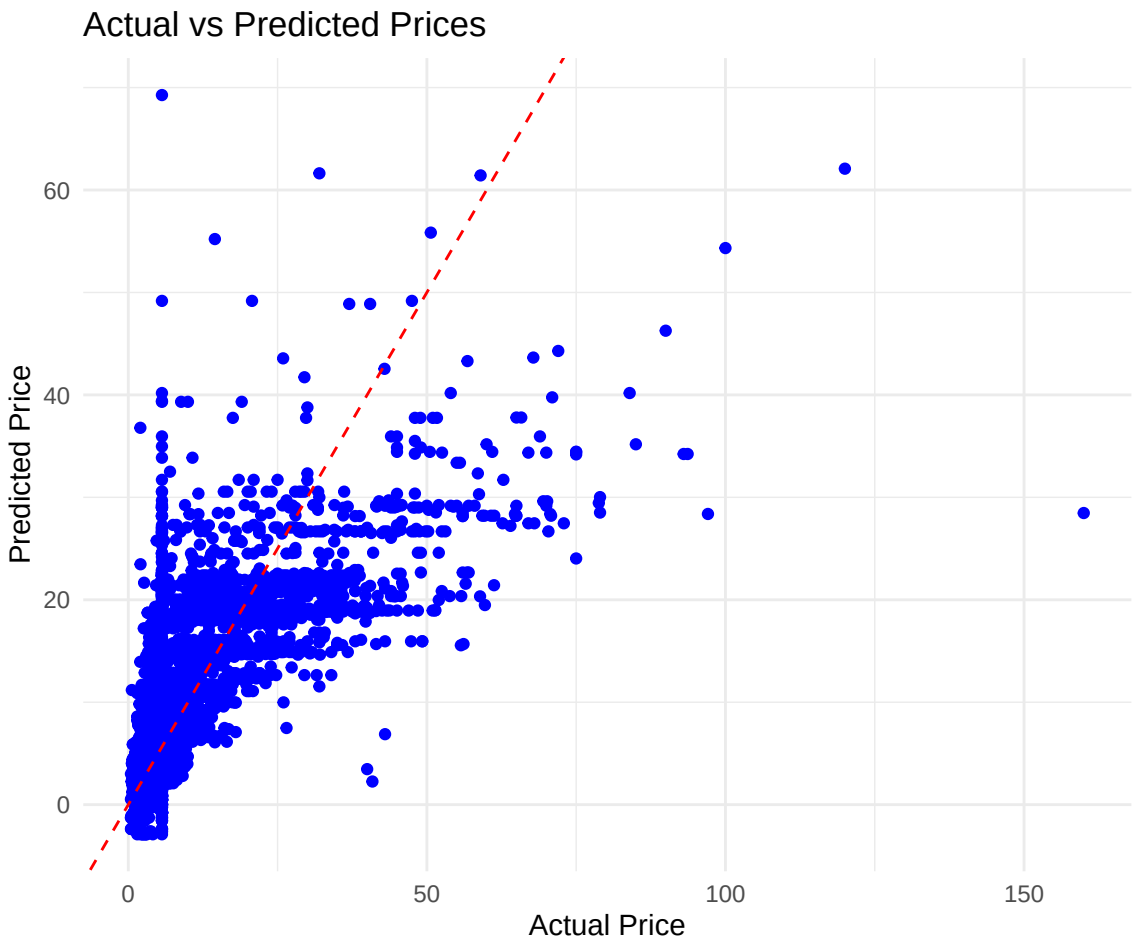


Figure 1: The relationship between Actual and Predicted Prices

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 a higher priced cars.

5 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

6 Appendix A

Kaggle Used Car Dataset

Introduction to Data Science

Introduction to Statistical Learning

Car Resale Value Prediction

Resale Value Car Price

Electric Cars vs Petrol Cars