Business Analytics of Craigslist car sales dataset

Adettoun

2023-05-18

```
library(ggplot2)#To explore the dataset
library(gplots)
library(ggpubr)#To explore and visualize dataset
library(corrplot)#for visualization of correlation
library(dplyr)#for data manipulation
library(tidyverse)#Data manipulation
library(tidyr)
library(plotly)
library(treemapify)#for treemaps
library(treemap)
library(gridExtra) #to combine ggplots
library(factoextra)#to view clusters
library(NbClust) #for clustering
library(reshape2)
library(psych)#to describe the dataset
library(rmarkdown)#to knit to word or pdf
library(tmaptools)#to create tmaps using shapefiles
library(tmap)
library(RColorBrewer)
library(e1071)

#Importing Dataset

project_new1 <- read.csv('Used_Vehicles.csv')

#To view first 20 rows
head(project_new1,20)
```

```
##            id url                region price year manufacturer model
## 1  7222695916  NA              prescott  6000   NA
## 2  7218891961  NA          fayetteville 11900   NA
## 3  7221797935  NA          florida keys 21000   NA
## 4  7222270760  NA worcester / central MA  1500   NA
## 5  7210384030  NA            greensboro  4900   NA
## 6  7222379453  NA         hudson valley  1600   NA
## 7  7221952215  NA         hudson valley  1000   NA
## 8  7220195662  NA         hudson valley 15995   NA
## 9  7209064557  NA       medford-ashland  5000   NA
## 10 7219485069  NA                  erie  3000   NA
## 11 7218893038  NA               el paso     0   NA
## 12 7218325704  NA               el paso     0   NA
## 13 7217788283  NA               el paso     0   NA
```

```
## 14 7217147606  NA                el paso     0   NA
## 15 7209027818  NA                el paso     0   NA
## 16 7223509794  NA             bellingham 13995   NA
## 17 7222753076  NA             bellingham 24999   NA
## 18 7222206015  NA             bellingham 21850   NA
## 19 7220030122  NA             bellingham 26850   NA
## 20 7218423006  NA             bellingham 11999   NA
##    condition cylinders fuel odometer title_status transmission drive size
type
## 1                                 NA
## 2                                 NA
## 3                                 NA
## 4                                 NA
## 5                                 NA
## 6                                 NA
## 7                                 NA
## 8                                 NA
## 9                                 NA
## 10                                NA
## 11                                NA
## 12                                NA
## 13                                NA
## 14                                NA
## 15                                NA
## 16                                NA
## 17                                NA
## 18                                NA
## 19                                NA
## 20                                NA
##    state VIN description posting_date
## 1     AZ  NA          NA           NA
## 2     AR  NA          NA           NA
## 3     FL  NA          NA           NA
## 4     MA  NA          NA           NA
## 5     NC  NA          NA           NA
## 6     NY  NA          NA           NA
## 7     NY  NA          NA           NA
## 8     NY  NA          NA           NA
## 9     OR  NA          NA           NA
## 10    PA  NA          NA           NA
## 11    TX  NA          NA           NA
## 12    TX  NA          NA           NA
## 13    TX  NA          NA           NA
## 14    TX  NA          NA           NA
## 15    TX  NA          NA           NA
## 16    WA  NA          NA           NA
## 17    WA  NA          NA           NA
## 18    WA  NA          NA           NA
## 19    WA  NA          NA           NA
## 20    WA  NA          NA           NA
```

##DATA CLEANING STEP

```
library(dplyr)


# Drop unnecessary variables from the dataset
project_new1 <- select(project_new1, -region, -url, -VIN, -description, -
size, -posting_date)

# Remove rows where price = 0
project_new1 <- filter(project_new1, price != 0)

# Remove rows where all major variables have missing values
project_new1 <- filter(project_new1, !is.na(year) & !is.na(odometer) &
manufacturer != "" & model != ""
                       & condition != "" & cylinders != "" & fuel != "" &
transmission != "")

#Check the missing values
missing_values <- colSums(is.na(project_new1))

# Print the count of missing values for each variable
print(missing_values)

##           id        price         year manufacturer        model
condition
##            0            0            0            0            0
0
##    cylinders         fuel     odometer title_status transmission
drive
##            0            0            0            0            0
0
##         type        state
##            0            0

# Impute missing values in numeric variables using the median
project_new1$year <- ifelse(is.na(project_new1$year),
median(project_new1$year, na.rm = TRUE), project_new1$year)

# Check if there are any missing values in numeric variables after imputation
#summary_data <- summarise_all(project_new1, list(n, n_miss = sum(is.na(.))))


# Sort the dataset by the year variable
project_new1 <- arrange(project_new1, year)

# Impute missing values in the odometer variable using the mean grouped by
year
project_new1 <- project_new1 %>%
  group_by(year) %>%
```

```
  mutate(odometer = ifelse(is.na(odometer), mean(odometer, na.rm = TRUE),
odometer))

# Check if all the numeric values are imputed
#summary_data <- summarise_all(project_new1, list(n, n_miss = sum(is.na(.))))

# Drop the id column
project_new1 <- select(project_new1, -id)

# Impute categorical variables
project_new1$manufacturer[project_new1$manufacturer == ""] <- "missing"
project_new1$condition[project_new1$condition == ""] <- "good"
project_new1$cylinders[project_new1$cylinders == ""] <- "6 cylinders"
project_new1$title_status[project_new1$title_status == ""] <- "clean"
project_new1$transmission[project_new1$transmission == ""] <- "automatic"
project_new1$drive[project_new1$drive == ""] <- "4wd"
project_new1$type[project_new1$type == ""] <- "SUV"

# Remove price outliers
percentiles <- quantile(project_new1$price, probs = seq(0.95, 1, by = 0.01),
na.rm = TRUE)
project_new1 <- filter(project_new1, price >= 3000 & price <= 70000)


#create logprice variable
project_new1$logprice <- log(project_new1$price)

# Print the first 20 rows of the final dataset
head(project_new1, 20)

## # A tibble: 20 × 14
## # Groups:   year [7]
##    price  year manufacturer model             condition cylinders fuel
odometer
##    <dbl> <int> <chr>        <chr>             <chr>     <chr>     <chr>
<int>
##  1 38250  1900 acura        "rdx"             new       4 cylind… gas
4500
##  2  3990  1905 chevrolet    "astro cargo"     excellent 8 cylind… gas
202570
##  3 27000  1913 ford         "\"t\""           like new  4 cylind… gas
9999999
##  4  5000  1915 ford         "model t"         good      4 cylind… gas
12345
##  5 16000  1918 ford         "model t"         good      4 cylind… gas
56000
##  6 15000  1922 ford         "t-bucket roadst… good      4 cylind… gas
80000
##  7 19500  1923 buick        "touring model"   good      4 cylind… gas
```

```
58
##  8 15000  1923 ford         "t-bucket roadst… good      8 cylind… gas
1000
##  9 18500  1923 ford         "t bucket"        like new  8 cylind… gas
2500
## 10 18990  1923 ford         "model t"         like new  8 cylind… gas
6652
## 11 18500  1923 ford         "t bucket"        new       8 cylind… gas
2500
## 12 18500  1923 ford         "model t"         excellent 8 cylind… gas
4963
## 13 18500  1923 ford         "t bucket"        like new  8 cylind… gas
2500
## 14 18500  1923 ford         "t-bucket"        like new  8 cylind… gas
2500
## 15 19000  1923 ford         "t - bucket"      excellent 8 cylind… gas
1600
## 16 29995  1923 ford         "t bucket"        excellent 8 cylind… gas
3700
## 17 30000  1923 ford         "t-bucket"        like new  8 cylind… gas
4400
## 18 15000  1923 ford         "model t"         excellent 8 cylind… gas
1000
## 19 18500  1923 ford         "t-bucket"        excellent 8 cylind… gas
2500
## 20 17500  1923 ford         "23 t"            excellent 8 cylind… gas
1000
## # ℹ 6 more variables: title_status <chr>, transmission <chr>, drive
<chr>,
## #   type <chr>, state <chr>, logprice <dbl>

#Recheck for missing values
mising_values <- colSums(is.na(project_new1))

# Print the count of missing values for each variable
mising_values

##        price         year manufacturer        model    condition
cylinders
##            0            0            0            0            0
0
##         fuel     odometer title_status transmission        drive
type
##            0            0            0            0            0
0
##        state     logprice
##            0            0
```

##Exploratory Data Analysis

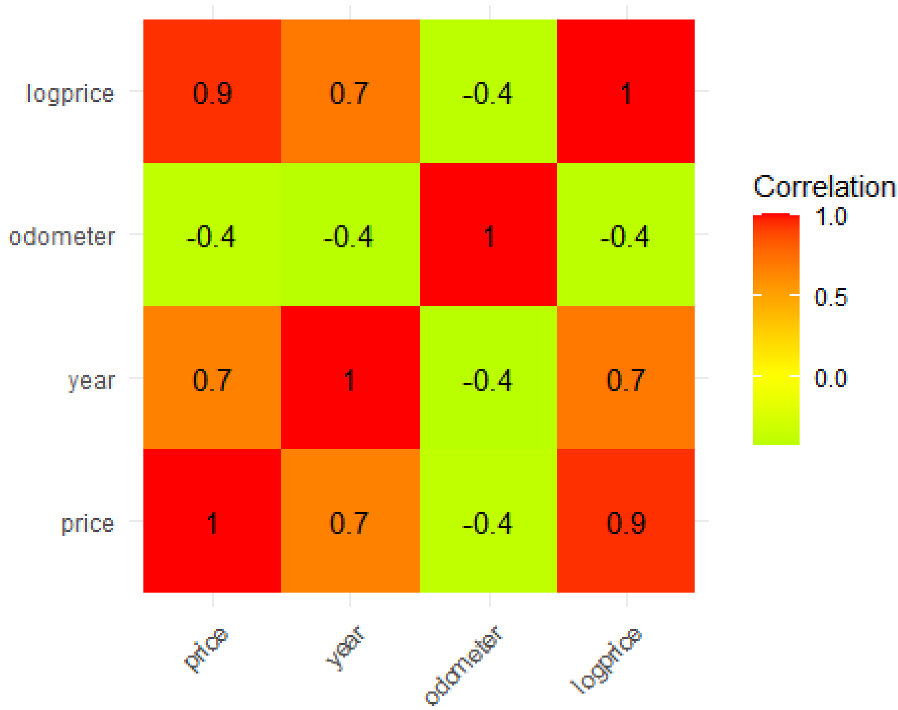
```r
library(ggplot2)

# Subset the data for the desired years
project_new1 <- project_new1[project_new1$year >= 2003 & project_new1$year <=
2021, ]

# Print the first 20 observations
head(project_new1, 20)

## # A tibble: 20 × 14
## # Groups:   year [1]
##    price  year manufacturer model             condition cylinders fuel
odometer
##    <dbl> <int> <chr>        <chr>             <chr>     <chr>     <chr>
<int>
##  1  9500  2003 chrysler     town & country    excellent 6 cylind… gas
30376
##  2  3500  2003 toyota       camry             good      4 cylind… gas
237000
##  3  8900  2003 ford         f250 lariat       good      8 cylind… dies…
247000
##  4  3950  2003 honda        civic ex          excellent 4 cylind… gas
236890
##  5  3500  2003 toyota       camry             good      4 cylind… gas
237000
##  6  6500  2003 chevrolet    tahoe z71         like new  8 cylind… gas
245700
##  7  6800  2003 ford         f-150             excellent 8 cylind… gas
190000
##  8  3500  2003 honda        accord            good      4 cylind… gas
201580
##  9 29990  2003 ford         super duty f-550… good      8 cylind… dies…
54703
## 10  3950  2003 honda        civic ex          excellent 4 cylind… gas
236890
## 11  3800  2003 chrysler     pt cruiser gt     good      4 cylind… gas
150000
## 12  3500  2003 toyota       camry             good      4 cylind… gas
237000
## 13  3950  2003 honda        civic ex          excellent 4 cylind… gas
236890
## 14  3900  2003 gmc          envoy             excellent 6 cylind… gas
165000
## 15 28750  2003 chevrolet    silverado 2500    like new  8 cylind… dies…
20000
## 16  5250  2003 chevrolet    trailblazer       good      6 cylind… gas
221000
## 17  5500  2003 ford         explorer sport t… excellent 6 cylind… gas
191000
## 18  6500  2003 chevrolet    tahoe             excellent 8 cylind… gas
```

```
257000
## 19  3995  2003 chrysler     pt cruiser        excellent 4 cylind… gas
142000
## 20  6900  2003 ford         ranger edge 4x4   good      6 cylind… gas
230000
## # ℹ 6 more variables: title_status <chr>, transmission <chr>, drive
<chr>,
## #   type <chr>, state <chr>, logprice <dbl>
```

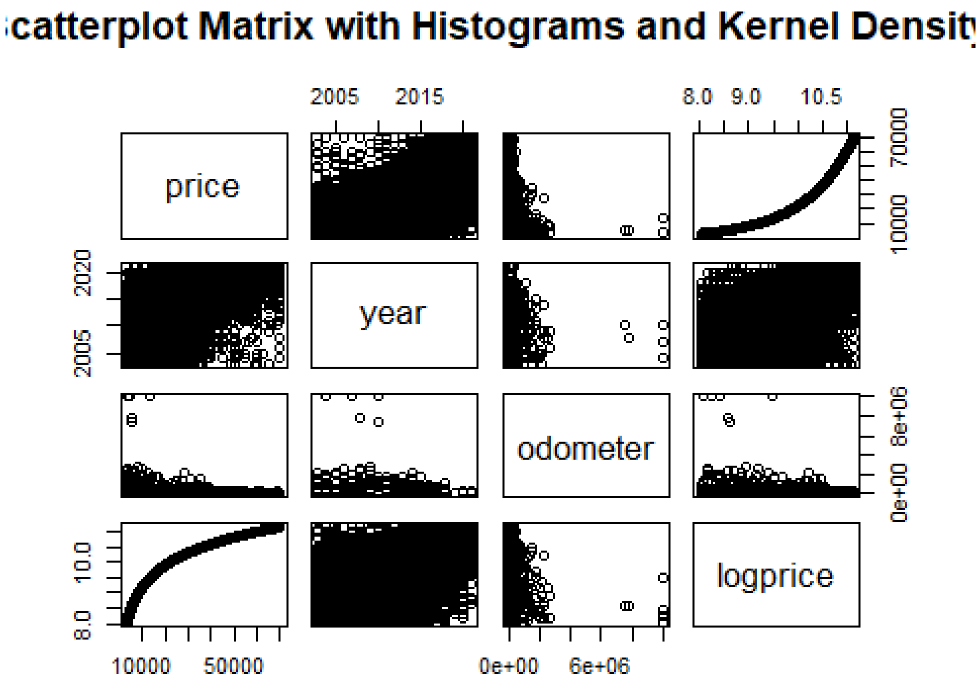
```
# Calculate correlation matrix
correlation <- cor(project_new1[, c("price", "year", "odometer",
"logprice")], use = "pairwise.complete.obs")
print(correlation)
```

```
##                price       year   odometer   logprice
## price     1.0000000  0.6553425 -0.4055012  0.9424302
## year      0.6553425  1.0000000 -0.4365961  0.6988973
## odometer -0.4055012 -0.4365961  1.0000000 -0.4267439
## logprice  0.9424302  0.6988973 -0.4267439  1.0000000
```

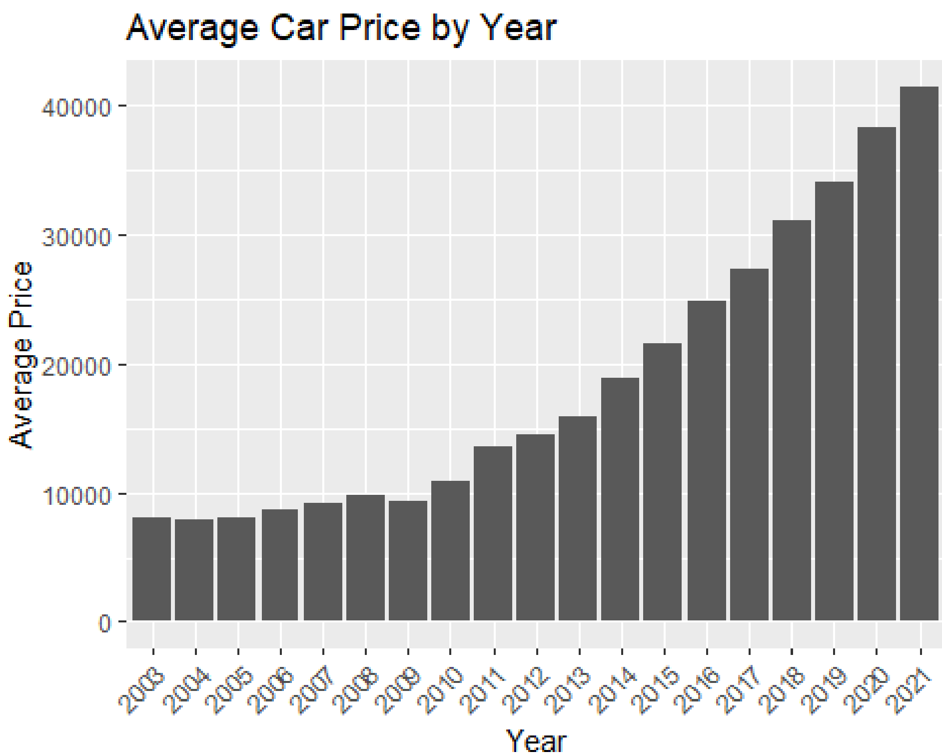
```
# Create heatmap of correlation matrix
library(ggplot2)
library(reshape2)

# Melt correlation matrix
correlation_melted <- melt(correlation)

# Create heatmap
ggplot(correlation_melted, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low="green", mid="yellow", high="red", midpoint=0) +
  geom_text(aes(label=round(value, 1)), color="black") +  # Add correlation
value to each box
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x="", y="", fill="Correlation")
```



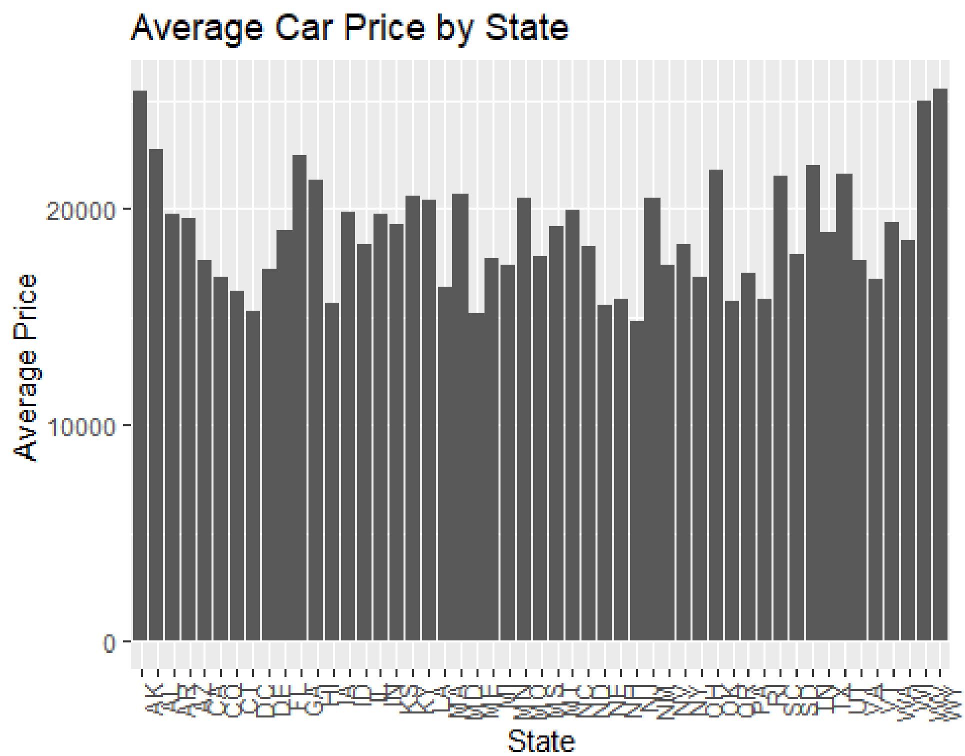

```r
# Select the variables for the scatterplot matrix
variables <- c("price", "year", "odometer", "logprice")

# Subset the data to include only the selected variables
subset_data <- project_new1[, variables]

# Create the scatterplot matrix with histograms and kernel density plots
pairs(subset_data, main = "Scatterplot Matrix with Histograms and Kernel
Density")
```



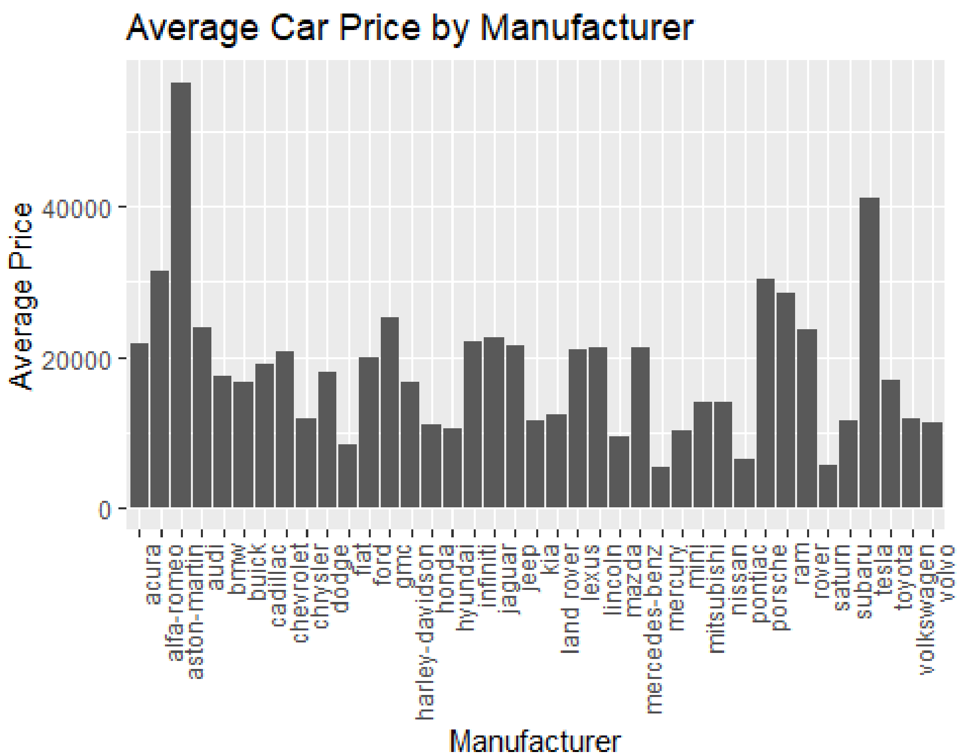

```r
# Subset the data for the desired years
project_new1 <- project_new1[project_new1$year >= 2003 & project_new1$year <=
2021, ]

# Calculate mean price by year
mean_price_by_year <- aggregate(price ~ year, project_new1, mean)

# Bar plot of average price by year
ggplot(mean_price_by_year, aes(x = as.factor(year), y = price)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Car Price by Year", x = "Year", y = "Average Price")+
   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



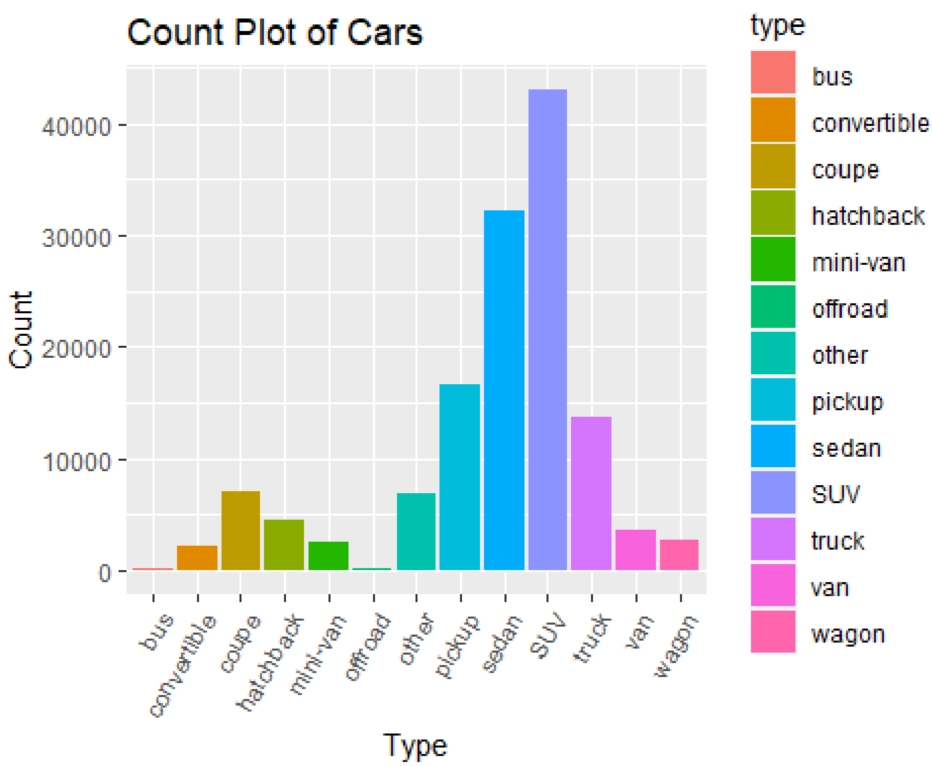

```
# Calculate mean price by state
mean_price_by_state <- aggregate(price ~ state, project_new1, mean)

# Bar plot of average price by state
ggplot(mean_price_by_state, aes(x = state, y = price)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Car Price by State", x = "State", y = "Average
Price") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



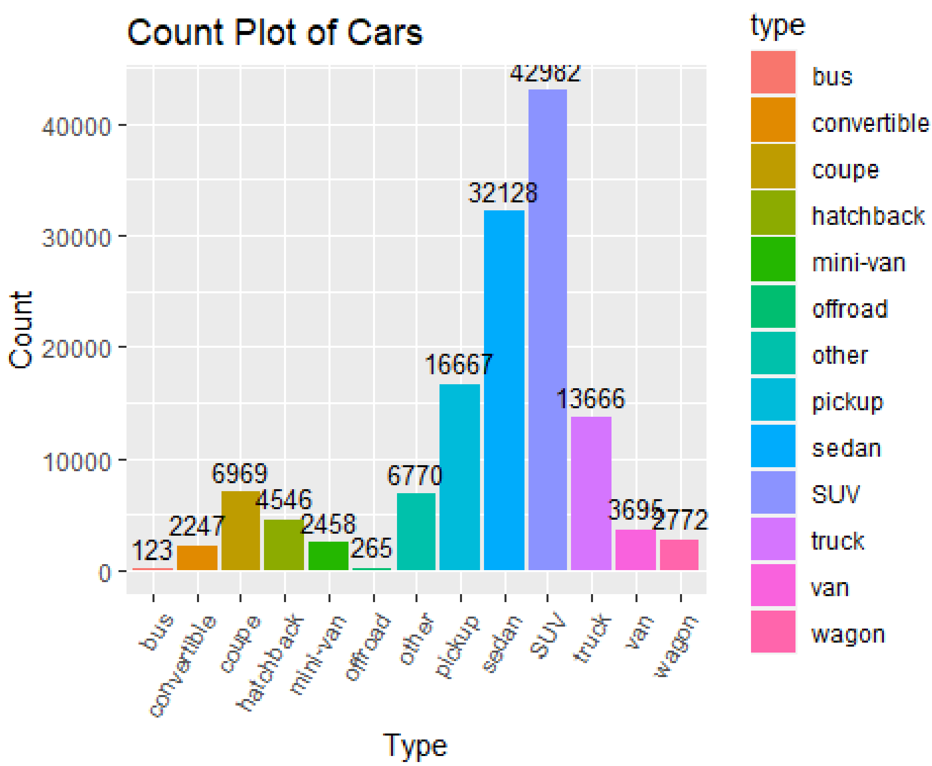

```r
# Calculate mean price by manufacturer
mean_price_by_manufacturer <- aggregate(price ~ manufacturer, project_new1,
mean)

# Bar plot of average price by manufacturer
ggplot(mean_price_by_manufacturer, aes(x = manufacturer, y = price)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Car Price by Manufacturer", x = "Manufacturer", y =
"Average Price") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



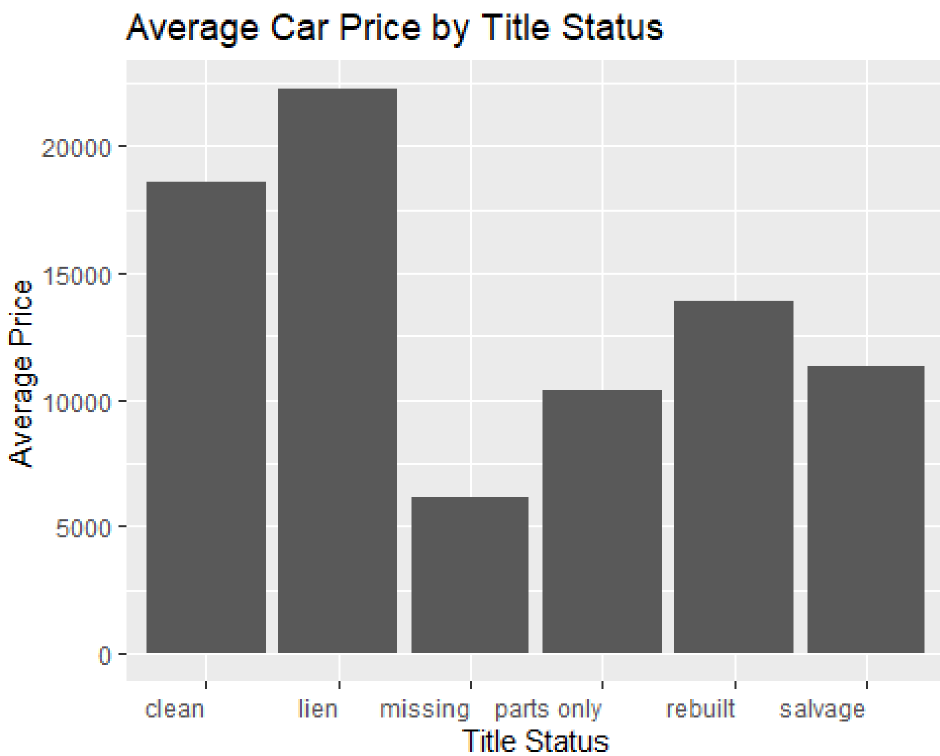

```r
# Count plot of cars by type
ggplot(project_new1, aes(x = type, fill = type)) +
  geom_bar() +
  labs(title = "Count Plot of Cars", x = "Type", y = "Count") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```




```r
library(ggplot2)

# Count the number of cars by type
car_counts <- project_new1 %>%
  group_by(type) %>%
  summarise(count = n())

# Create a count plot with labels
count_plot <- ggplot(car_counts, aes(x = type, y = count, fill = type)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = count), vjust = -0.5, color = "black", size = 3.5) +
# Add count labels
  labs(title = "Count Plot of Cars", x = "Type", y = "Count") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))

# Display the count plot
print(count_plot)
```




```
# Calculate mean price by title status
mean_price_by_title_status <- aggregate(price ~ title_status, project_new1,
mean)

# Bar plot of average price by title status
ggplot(mean_price_by_title_status, aes(x = title_status, y = price)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Car Price by Title Status", x = "Title Status", y =
"Average Price") +
  theme(axis.text.x = element_text(angle = 0, hjust = 1))
```



- It is clear that cars from recently produced cars are more expensive than older ones. The car market has experienced a significant increase in prices due to the introduction of new technologies. Cars made in recent years are equipped with more features and are generally in better condition than those from the early 2000s. Furthermore, new cars hold their value better in the resale market compared to older cars, which are expected to fetch lower prices.

- As seen in the above plot, Montana, Washington, and West Virginia, have the most expensive used cars car by average price. It's not surprising that car prices are higher in wealthier states, given the varying geographical locations across the country.

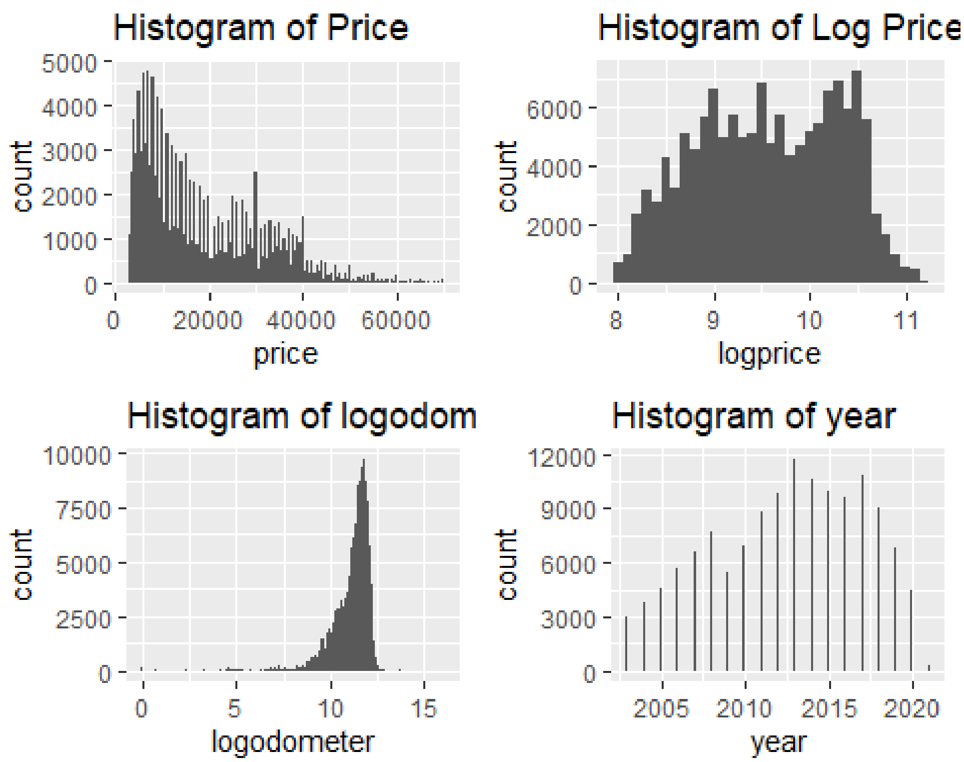
- Luxury car manufacturers like Aston Martin, Tesla, and Porsche have the highest prices on average, which is justified by their branding as luxury. On the other hand, Saturn, Kia, and Mercury, which are typically known for their budget-friendly smaller cars, are on the lower end of the spectrum. These brands are more accessible to the common people.

- The majority of cars listed are SUVs and sedans. This is because many people in the United States prefer larger vehicles such as SUVs, which provide ample space for families and offer generous amounts of storage in the boot. Sedans are also a popular choice among small families, office workers, doctors, and young people due to their stylish design and comfortable ride. There are also a significant number of trucks listed. However, coupes and hatchbacks are less common as they are considered luxury cars and come with a higher price tag.

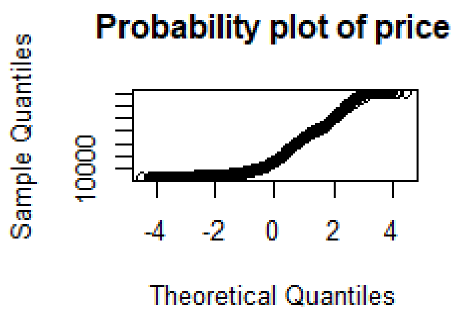
- When looking to purchase a car, many people prefer those with clean and lien titles. A lien title means that the car is still under a mortgage or EMI that must be transferred to the new owner. This is why cars with clean and lien titles tend to have higher prices. However, some cars have “missing” title statuses, which don’t provide much information about the car’s status. Customers may need to consider other features to make an informed decision.

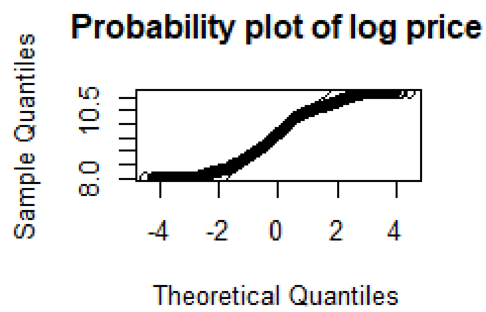
Distribution Analysis

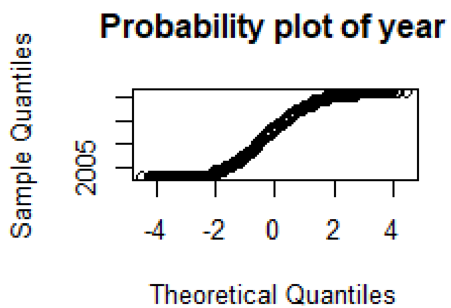
```
# Univariate analysis of price and logprice
univariate_price <- summary(project_new1$price)
univariate_logprice <- summary(project_new1$logprice)
univariate_price

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3000    7900   14500   18414   27444   70000

univariate_logprice

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.006   8.975   9.582   9.565  10.220  11.156

# Histogram of price and logprice
hist_price <- ggplot(project_new1, aes(x = price)) +
  geom_histogram(binwidth = 500) +
  labs(title = "Histogram of Price")


# Histogram of logprice
hist_logprice <- ggplot(project_new1, aes(x = logprice)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "Histogram of Log Price")


#create logodometer variable
project_new1$logodometer <- log(project_new1$odometer)

#Histogram of odometer
hist_odo <- ggplot(project_new1, aes(x = logodometer)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "Histogram of logodometer")

#Histogram of year
hist_year <- ggplot(project_new1, aes(x = year)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "Histogram of year")

grid.arrange(hist_price, hist_logprice, hist_odo, hist_year,nrow = 2, ncol =
2)
```




```r
par(mfrow = c(2, 2))
# Probplot of price
qqnorm(project_new1$price, main="Probability plot of price")
qqline(project_new1$price)

# Probplot of logprice
qqnorm(project_new1$logprice, main="Probability plot of log price")
qqline(project_new1$logprice)

# Probplot of year
qqnorm(project_new1$year, main="Probability plot of year")
qqline(project_new1$year)
```







- The odometer variable has been logged and transformed to reduce the skewness. The accompanying p-value from the Goodness-of-Fit tests confirms the variable follows a normal distribution.

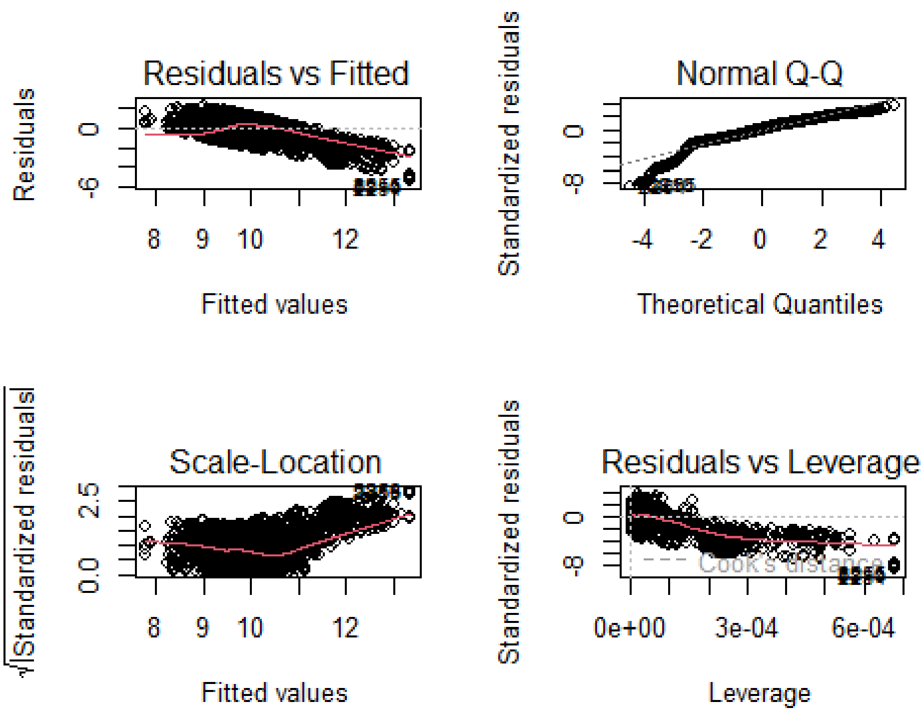
- The distribution of the price variable is skewed to the right. Although the probability plot has data points that are along the line. It is necessary to adjust the variable's distribution. This is done by doing a log transformation to normalize the data.

Simple Linear Regression

```
#H0: No significant price difference between cars with higher odometer
readings and those with lower ones.
#H1: There is significant price difference between cars with higher odometer
readings and those with lower


project_new2 <- read.csv('project_new2.csv')
# Fit a simple linear regression model
model <- lm(logprice ~ logodometer, data = project_new2)

# Print the summary of the model
summary(model)

##
## Call:
## lm(formula = logprice ~ logodometer, data = project_new2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -5.2613 -0.4439  0.0396  0.4609  2.1470
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.580378   0.017387   781.1   <2e-16 ***
## logodometer -0.360242   0.001553  -232.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6261 on 134842 degrees of freedom
## Multiple R-squared:  0.2853, Adjusted R-squared:  0.2853
## F-statistic: 5.383e+04 on 1 and 134842 DF,  p-value: < 2.2e-16
```

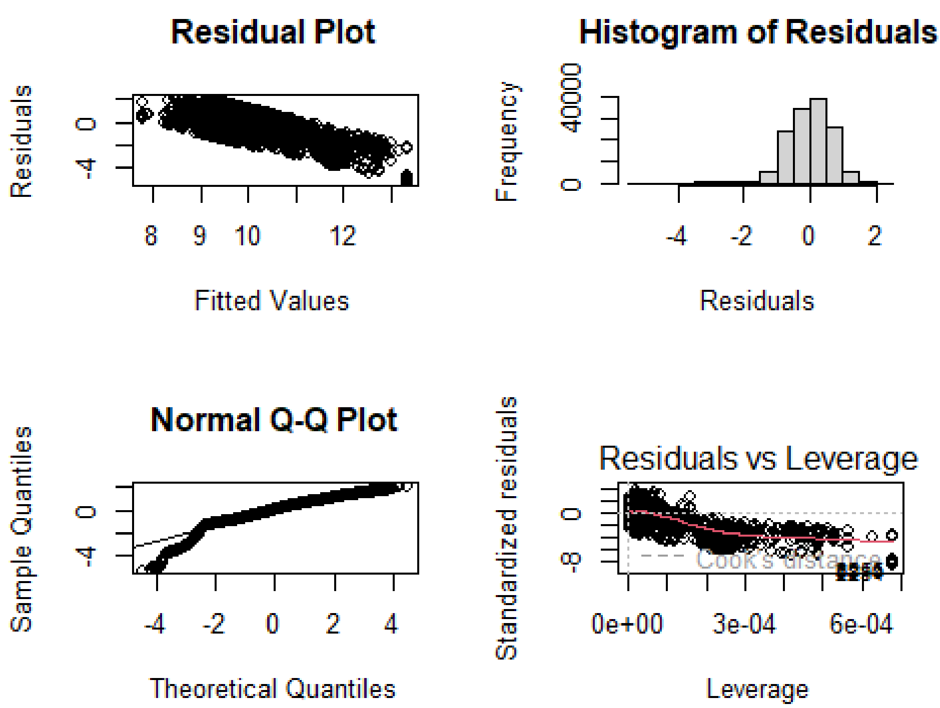
```
#plot(model)

# Generate diagnostic plots
par(mfrow = c(2, 2))  # Set the layout to display four plots in a 2x2 grid
plot(model, which = c(1, 2, 3, 5))  # Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage
```



- From the result, F test p-value is less than 0.0001, which indicates that the model is significant. Estimated logdometer is -0.23. The t-value is -267.87. The p-value is less than 0.0001. We can reject the null hypothesis. The price and odometer have a negative relationship. For every percent increase in odometer, the price of used cars will be decreased 0.23%.

```
# Get residuals
residuals <- residuals(model)
```

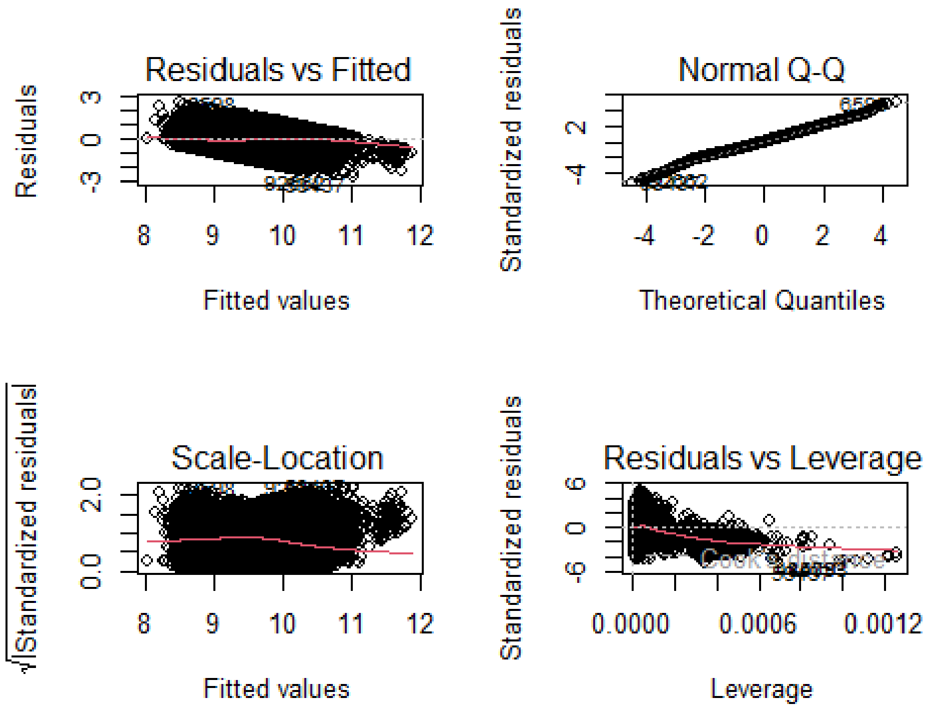
```r
par(mfrow = c(2, 2))
# Residual plot
plot(model$fitted.values, residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residual Plot")

# Histogram of residuals
hist(residuals,
     xlab = "Residuals",
     main = "Histogram of Residuals")

# Normal Q-Q plot
qqnorm(residuals,
       main = "Normal Q-Q Plot")
qqline(residuals)



# Residuals vs Leverage plot
plot(model, which = 5)
```



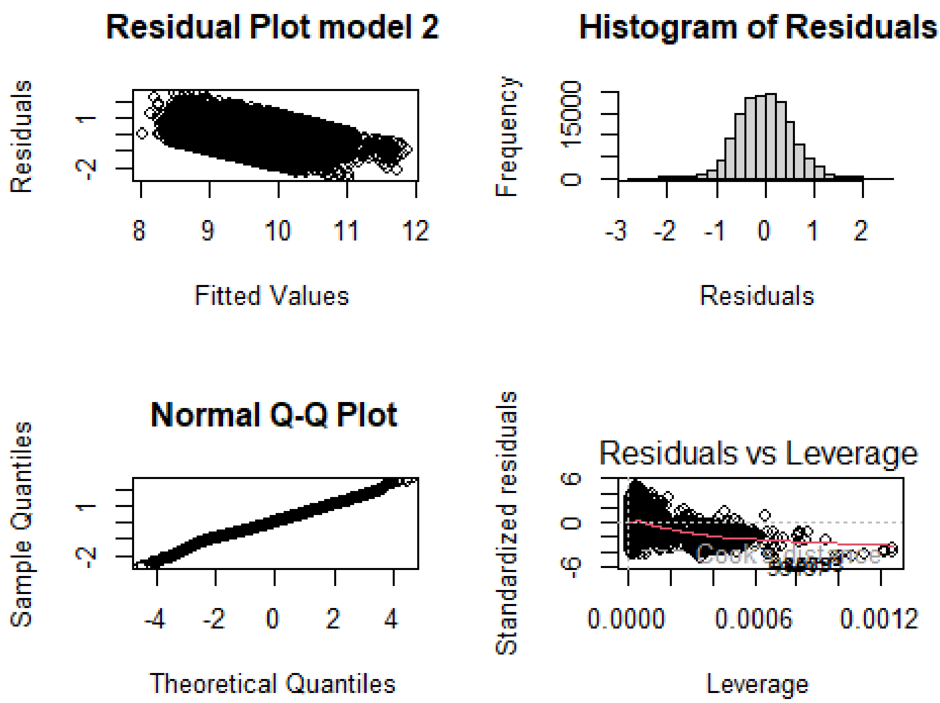
From the diagnostic outputs. The density and quantile of the residuals look visually ok. This might be attributed to the luxury cars in the dataset which are still expensive despite being used cars

##Multiple linear Regression

```r
# Fit a simple linear regression model

#H0: No significant Average car price difference based on year of manufacture
#H1: There is significant difference in Average car price based on year of
manufacture

model2 <- lm(logprice ~ logodometer + year, data = project_new2)

# Print the summary of the model
summary(model2)

##
## Call:
## lm(formula = logprice ~ logodometer + year, data = project_new2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60922 -0.36261 -0.00655  0.34513  2.48429
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.800e+02  7.519e-01 -239.35   <2e-16 ***
## logodometer -1.440e-01  1.524e-03  -94.47   <2e-16 ***
## year         9.497e-02  3.689e-04  257.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5127 on 134841 degrees of freedom
## Multiple R-squared:  0.5209, Adjusted R-squared:  0.5209
## F-statistic: 7.329e+04 on 2 and 134841 DF,  p-value: < 2.2e-16

#plot(model)

# Generate diagnostic plots
par(mfrow = c(2, 2))  # Set the layout to display four plots in a 2x2 grid
plot(model2, which = c(1, 2, 3, 5))  # Residuals vs Fitted, Normal Q-Q,
Scale-Location, Residuals vs Leverage
```




```r
residuals2 <- residuals(model2)

par(mfrow = c(2, 2))
# Residual plot
plot(model2$fitted.values, residuals2,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residual Plot model 2")

# Histogram of residuals
hist(residuals2,
     xlab = "Residuals",
     main = "Histogram of Residuals")

# Normal Q-Q plot
qqnorm(residuals2,
       main = "Normal Q-Q Plot")
qqline(residuals2)



# Residuals vs Leverage plot
plot(model2, which = 5)
```

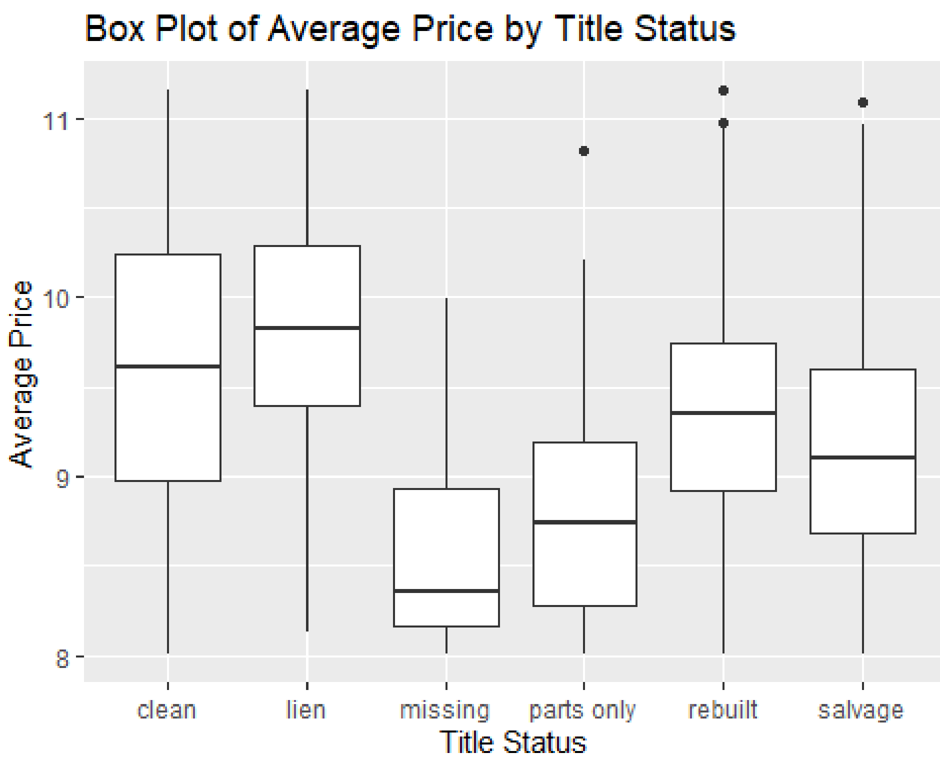


- The year variable has estimated value of 0.10 and t-value of 438.18. The p-value is also below 0.0001. The null hypothesis will be rejected. There exists a positive relationship between price and year. The closer the year of manufacturer is to the present date, the more the price increases by about 10%. Based on the results presented above, we can conclude that the null hypothesis stating that higher odometer values result in higher costs for cars is rejected. Additionally, it is evident that used cars produced in recent years tend to cost more.

##ANOVA by Title_Status

```
##ANOVA
#H0: The title status of the used car has no significant impact on the price
#H1: The title status of the used car has a significant impact on the price

ANova_1 <- aov(project_new2$logprice ~ project_new2$title_status)
summary(ANova_1)

##                               Df Sum Sq Mean Sq F value Pr(>F)
## project_new2$title_status      5    589  117.72   216.3 <2e-16 ***
## Residuals                 134838  73376    0.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Calculate average price by Title_status
avg_price <- project_new2 %>%
  group_by(title_status) %>%
  summarize(avg_price = mean(logprice))
```

```
# Create a box plot
boxplot <- ggplot(project_new1, aes(x = title_status, y = logprice)) +
  geom_boxplot() +
  labs(x = "Title Status", y = "Average Price") +
  ggtitle("Box Plot of Average Price by Title Status")

# Display the box plot
print(boxplot)
```

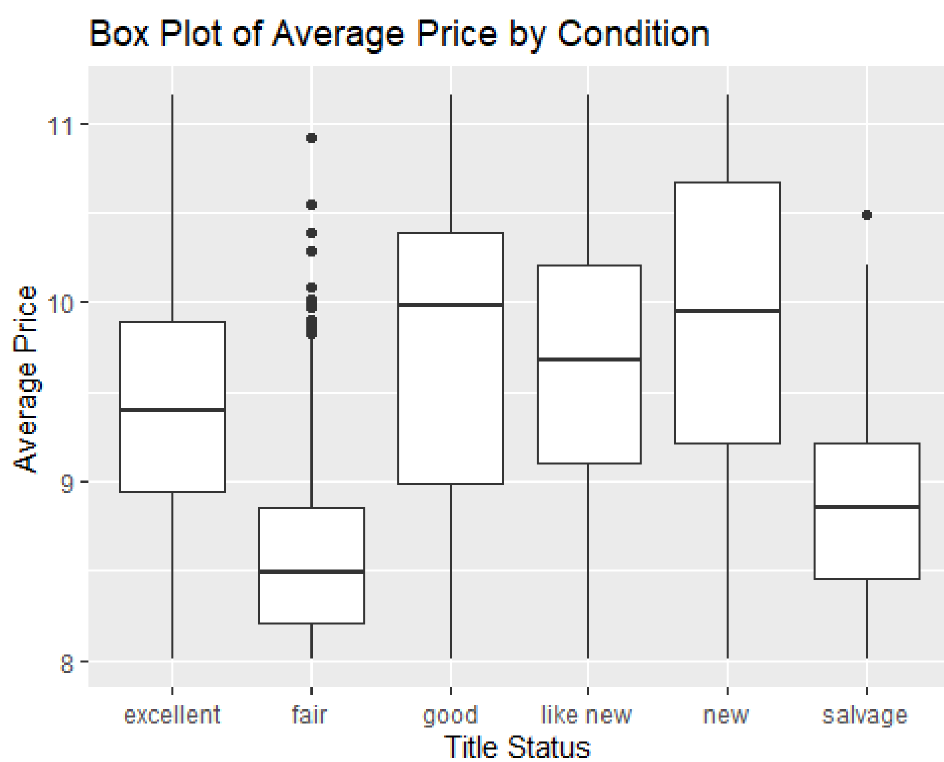


- As seen there 6 categories of used cars based on title_status. The null hypothesis is that the title of status has no significant impact on price. The F-value 705.78 with p-value is lesser than 0.0001, indicating significance of the model. Based on our analysis, the null hypothesis can be rejected, indicating that the title status of used cars does indeed impact their price. The box plot reveals that a clean or lien title is associated with a higher price, whereas a missing or parts title corresponds to a comparatively lower price.

##Anova by Condition

```
##ANOVA

#H0: The condition of the used car has no significant impact on price
#H1: The condition of the used car has a significant impact on the price

ANova_2 <- aov(project_new2$logprice ~ project_new2$condition)
summary(ANova_2)
```

```
##                                Df Sum Sq Mean Sq F value Pr(>F)
## project_new2$condition      5   3700   740.1    1420 <2e-16 ***
## Residuals               134838  70264     0.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Calculate average price by Condition
avg_price <- project_new2 %>%
  group_by(title_status) %>%
  summarize(avg_price = mean(logprice))


# Create a box plot
boxplot2 <- ggplot(project_new1, aes(x = condition, y = logprice)) +
  geom_boxplot() +
  labs(x = "Title Status", y = "Average Price") +
  ggtitle("Box Plot of Average Price by Condition")

# Display the box plot
print(boxplot2)
```



As seen, there are 6 categories of used cars based on condition. The null hypothesis is that the condition has no significant impact on price. The F-value is 3352.23 and p-value lesser than 0.0001, indicating a significant model. Based on our analysis, we can conclude that we reject the null hypothesis and confirm that the condition of used cars does indeed have an impact on their price. By examining the box plot, we can infer those cars in new, excellent, like new, or good condition tend to have higher prices, while those in fair or salvage condition are comparatively lower. It is worth noting that there are outliers in the fair group, which may affect the overall trend.

In this project, we analyzed sales data of used vehicles from Craigslist to understand the impact of odometer, year, title status, and condition on the price of the car in the US. Our exploratory data analysis resulted in some interesting findings. We observed that the price of a used car is positively correlated with its manufacturing year, and cars with a clean title status tend to fetch a higher price. Also, as anticipated, the correlation matrix revealed a negative association between price and odometer. We utilized linear regression models, diagnostic tests, and ANOVA to achieve our goal of determining the significance of variables on price and rejecting the null hypothesis. Our findings confirmed that newer cars are priced higher, while those with higher odometer values are less expensive. In particular, the price decreases by 0.23% for every percent increase in odometer, and the price increases by approximately 10% for every year newer. Furthermore, we conducted ANOVA tests to assess the title status and condition of cars and found that those with cleaner titles and better conditions generally have higher prices. This evaluation holds significant worth for both buyers and sellers in the pre-owned auto industry, as it offers valuable insights into market trends and aids in informed decision-making. Manufacturers can leverage the findings to gain an understanding of consumer preferences, enhance their production processes, and boost their brand's market share. Meanwhile, purchasers can make more informed decisions by keeping up with the current market trends.