ANOVA project

Michael Yun, Priyanshu Dey

Collaboration rules:

Students are encouraged to work with a partner on this project. Be sure to register your team: "Canvas » People » ANOVA Project Teams" and write the full name of each teammate in the "author" line at the top of this Rmd document.

### Instructions

Write a report that includes an introduction to the data, appropriate EDA, model specification, the checking of conditions, and in context conclusions. To include sections in your report use the # as illustrated by the # Instructions for this section. Larger section headings have one #, smaller subsection headings have ## or ### or even ####. There should be a coherent and well-organized narrative in addition to appropriate code and figures. You may also reference your MLR project as a framework.

Introduction

In this study, we examine the pricing of three popular automobile models—the Honda Civic, Ford F-150, and Jeep Cherokee—in two distinctly different geographic locations within the United States: State College, Pennsylvania (ZIP code 16801) and Charlotte, North Carolina (ZIP code 28207). This analysis aims to uncover potential pricing disparities and patterns based on vehicle model and geographical location. By analyzing data collected from Autotrader listings provided by St. Lawrence University's dataset portal, we intend to identify how external factors such as location and internal factors like car model influence the pricing of used vehicles,as well as potential covariates include the mileage and year/age.

The selection of car models provides a broad spectrum of vehicle types and market segments:

Honda Civic: A staple in the compact car segment, known for its reliability and efficiency. Ford F-150: A leading model in the full-size pickup truck category, renowned for its capability and versatility. Jeep Cherokee: A popular SUV that balances off-road capability with on-road comfort.

The two chosen locations offer contrasting demographics and economic landscapes:

State College, PA (16801): Known primarily as a college town, home to Penn State University, which may influence vehicle demand and pricing,coded as 1. Charlotte, NC (28207): A major metropolitan area with a diverse economy and a larger market for various types of vehicles,coded as 0.

Research Question

How does car model,location and mileage affect car prices, specifically examining how much of the variance in car prices can be attributed to differences in model and location after controlling for mileage?

Two-Way Factorial ANCOVA Model with Interaction

Hypotheses

- **Null Hypotheses (H0):**
  - **H0:** Interaction between location and model does not significantly influence affect the price of cars when adjusting for mileage, meaning all models would have the same pricing adjusted for mileage.
- **Alternative Hypotheses (Ha):**
  - **Ha:** Interaction between location and model significantly affects the price of cars when adjusting for mileage.

EDA

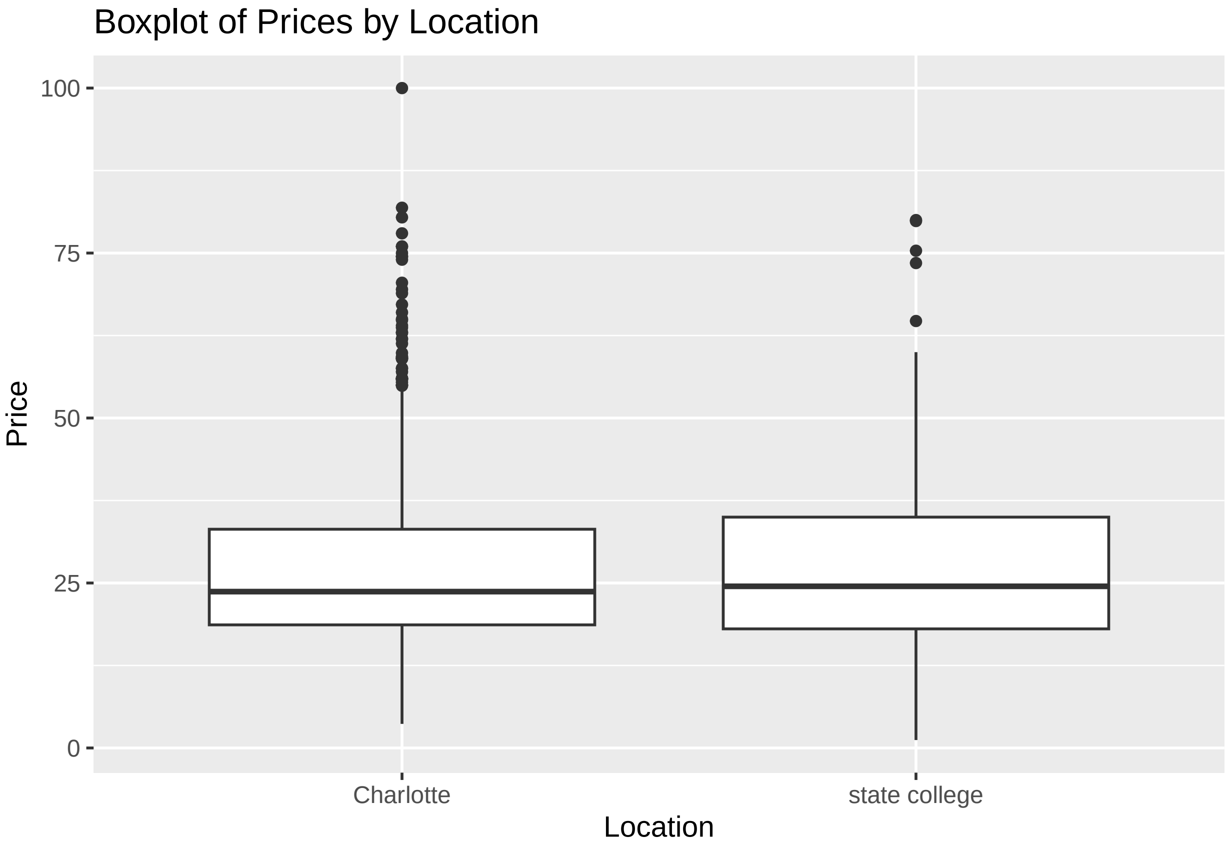
Data collation

```
#Load the datasets
sc_civic<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/16801_civic.csv")
nc_civic<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/28207_civic.csv")
sc_Cherokee<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/16801_Cherokee.csv")
nc_Cherokee<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/28207_Cherokee.csv")
sc_F150<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/16801_F150.csv")
nc_F150<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/28207_F150.csv")
# Add a new column 'location' to each datasets,SC coded as 1,NC coded as 0
sc_civic$location<-"state college"
nc_civic$location<-"Charlotte"
sc_Cherokee$location<-"state college"
nc_Cherokee$location<-"Charlotte"
sc_F150$location<-"state college"
nc_F150$location<-"Charlotte"
# Assigning the model names to each dataframe
sc_civic$model <- "Civic"
nc_civic$model <- "Civic"
sc_Cherokee$model <- "Cherokee"
nc_Cherokee$model <- "Cherokee"
sc_F150$model <- "F150"
nc_F150$model <- "F150"
#perform full join to merge the data sets
combined_data <- sc_civic %>%full_join(nc_civic)%>%full_join(sc_Cherokee)%>%full_join(nc_Cherokee)%>%ful
```

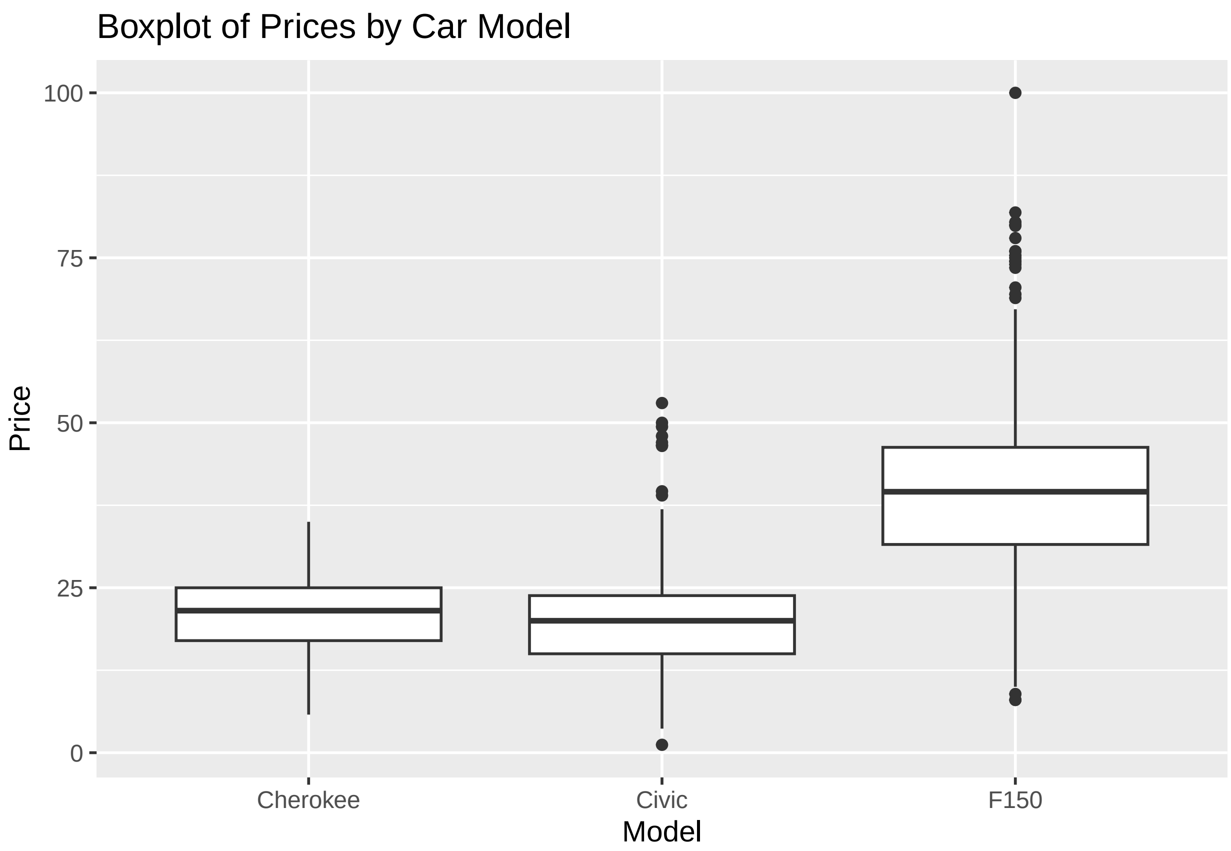
```
## Joining with ‘by = join_by(year, price, mileage, location, model)‘
## Joining with ‘by = join_by(year, price, mileage, location, model)‘
## Joining with ‘by = join_by(year, price, mileage, location, model)‘
## Joining with ‘by = join_by(year, price, mileage, location, model)‘
## Joining with ‘by = join_by(year, price, mileage, location, model)‘
```

```r
# Add a new column 'age'
combined_data $age <- 2024 - combined_data $year
# Remove NA values specifically in 'mileage' and 'price'
combined_data <- na.omit(combined_data, cols = c("mileage", "price"))
# Remove rows where 'mileage' or 'price' equals zero
combined_data <- combined_data[combined_data$mileage != 0 & combined_data$price != 0, ]
```

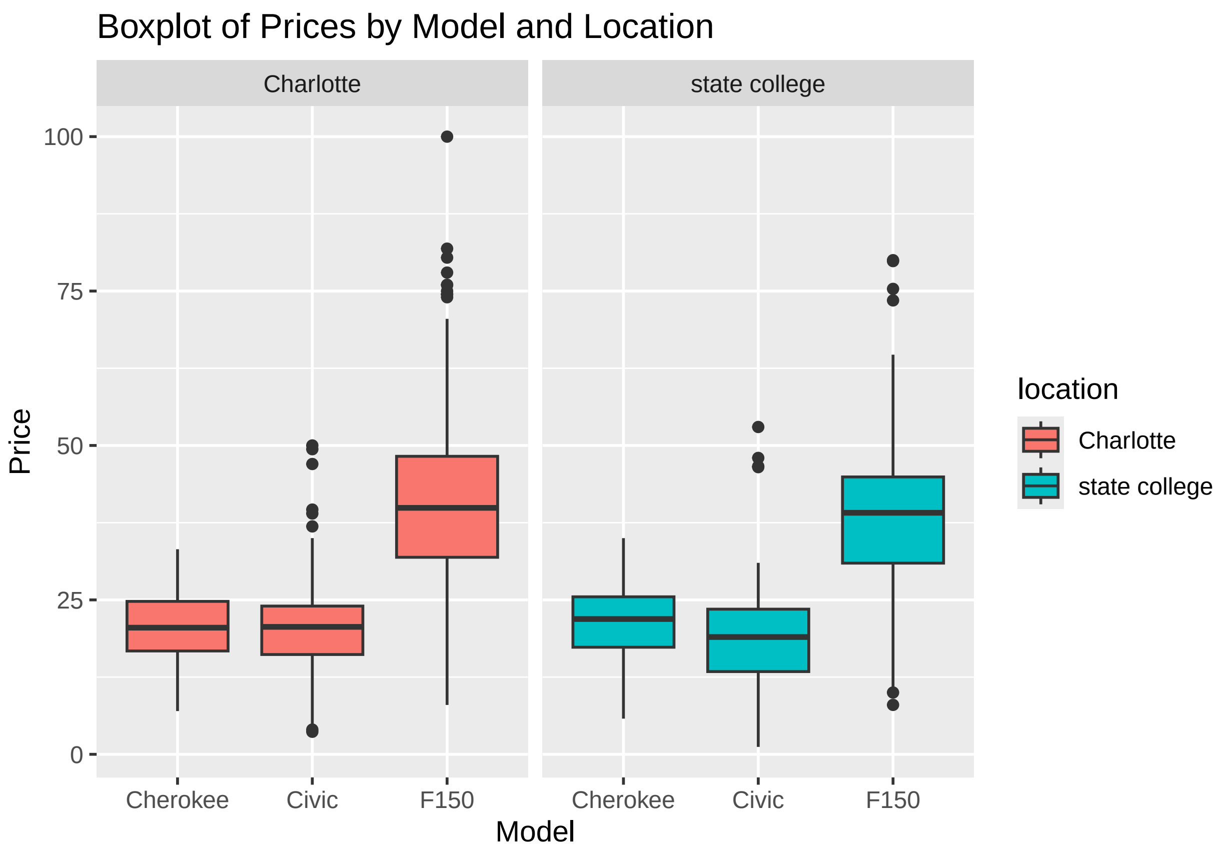
```r
# Boxplot for price by location
ggplot(combined_data, aes(x = location, y = price)) +
  geom_boxplot() +
  labs(title = "Boxplot of Prices by Location", x = "Location", y = "Price")
```



```r
# Boxplot for price by model
ggplot(combined_data, aes(x = model, y = price)) +
  geom_boxplot() +
  labs(title = "Boxplot of Prices by Car Model", x = "Model", y = "Price")
```

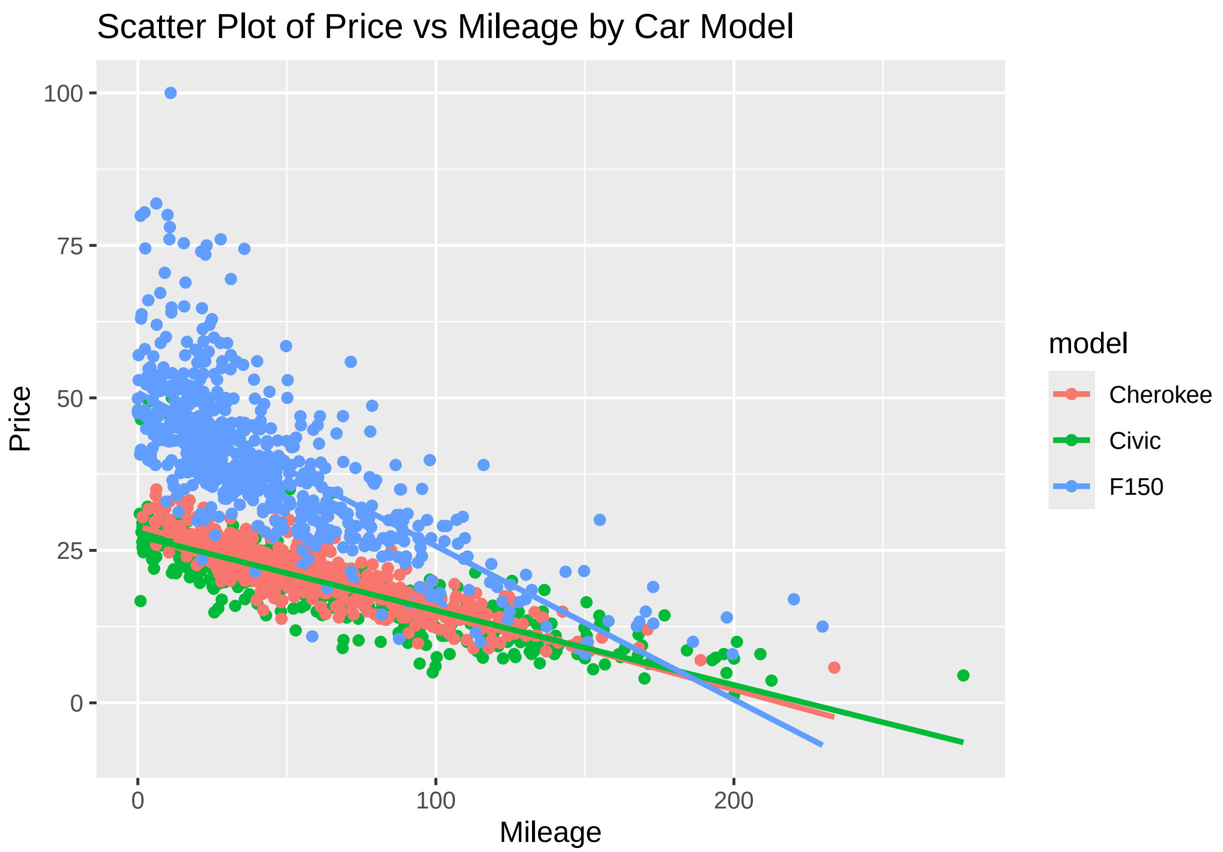


```r
# Boxplot for price by model and location interaction
ggplot(combined_data, aes(x = model, y = price, fill = location)) +
  geom_boxplot() +
  labs(title = "Boxplot of Prices by Model and Location", x = "Model", y = "Price") +
  facet_wrap(~ location)
```



```
# Scatter plot for price vs mileage colored by model
ggplot(combined_data, aes(x = mileage, y = price, color = model)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatter Plot of Price vs Mileage by Car Model", x = "Mileage", y = "Price")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
# Interaction plot for price by location and model
interaction.plot(combined_data$location,combined_data$model, combined_data$price,
                 fun = mean, type = "b", legend = TRUE,
                 xlab = "Location", ylab = "Price", trace.label = "Model")
```

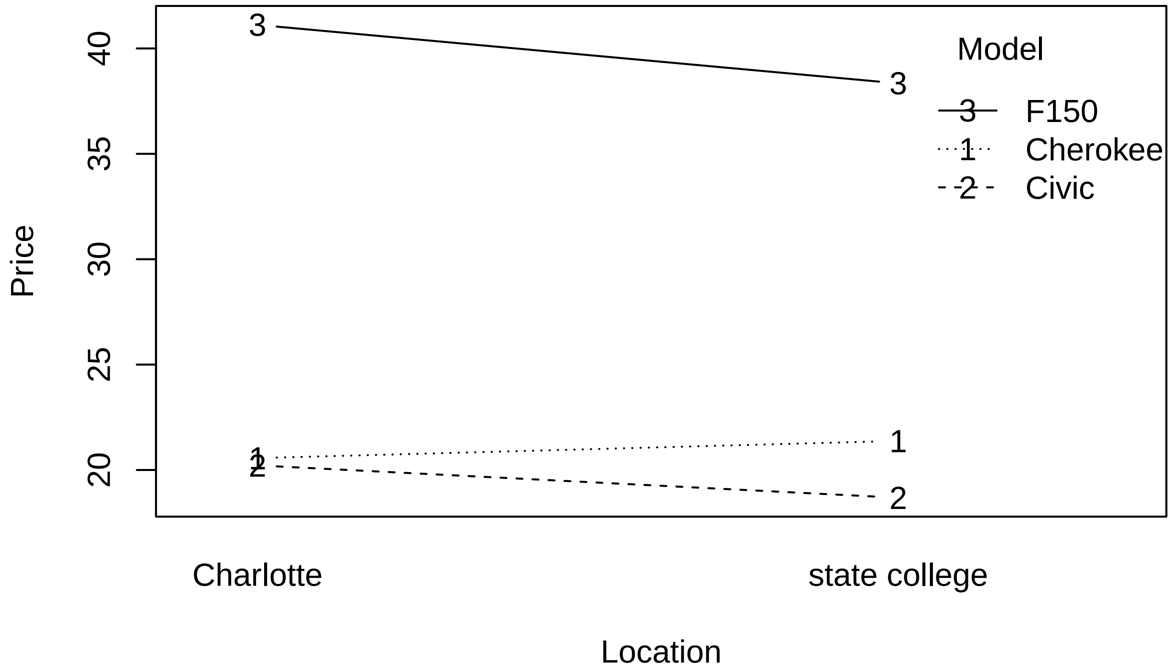


Image 1: Boxplot of Prices by Location

This boxplot displays the distribution of prices for each location, without considering the car model. The median price in Charlotte appears to be lower than in State College. There are some outliers (dots above the whiskers) in both locations, indicating the presence of extremely high-priced vehicles.

Image 2: Boxplot of Prices by Car Model

This boxplot shows the distribution of prices for each car model, irrespective of location. The F150 model has the highest median price, followed by the Civic and then the Cherokee. The F150 also exhibits a larger spread in prices compared to the other models.

Image 3: Boxplot of Prices by Model and Location

This faceted boxplot combines the information from the previous two plots, illustrating the price distributions for each combination of location and car model. In both locations, the F150 consistently has the highest median price, followed by the Civic and then the Cherokee. The price distributions for the Civic and Cherokee appear to be relatively similar across locations, while the F150 shows a more noticeable difference, with higher prices in State College.

Image 4: Scatter Plot of Price vs. Mileage by Car Model

This scatter plot displays the relationship between price and mileage for each car model, with different colors representing different models. There is a clear negative correlation between price and mileage, indicating that vehicles with higher mileage tend to have lower prices. The F150 model generally has higher prices compared to the Civic and Cherokee for similar mileage levels. The Civic and Cherokee models exhibit a more overlapping range of prices and mileages.

Image 5: Line Plot of Prices by Location and Model

This line plot presents the mean prices for each combination of location and car model. The F150 model has the highest mean price in both locations, followed by the Civic and then the Cherokee. The mean prices for the Civic and Cherokee are relatively similar across locations, while the F150 shows a more substantial difference, with a higher mean price in State College.

Model Fitting

```
#ANCOA
anova_model <- aov(lm(price ~ model+mileage+location+location*model, data = combined_data))
summary(anova_model)
```

```
##                   Df Sum Sq Mean Sq  F value   Pr(>F)    
## model              2 147173   73586 1878.772  < 2e-16 ***
## mileage            1  72619   72619 1854.071  < 2e-16 ***
## location           1   1250    1250   31.923 1.88e-08 ***
## model:location     2    539     270    6.887  0.00105 ** 
## Residuals       1695  66389      39                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. **Model*location :**

    - **Degrees of Freedom (Df):** The model has 2 degrees of freedom, which suggests there are three different car models being compared.
    - **Sum of Squares (Sum Sq):** The Sum Sq for the model is 539, which is a measure of the total variation attributed to the differences in the mean prices across the car models.
    - **Mean Square (Mean Sq):** The Mean Sq, which is the Sum Sq divided by the Df, is 270. This represents the average variation per model category.
    - **F-value:** The F-value is 6.887, which is substantially large, indicating a strong effect of the model on price.
    - **p-value (Pr(>F)):** The p-value is less than 0.00105, which is highly significant. This means interaction between location and model significantly affects the price of cars when adjusting for mileage.

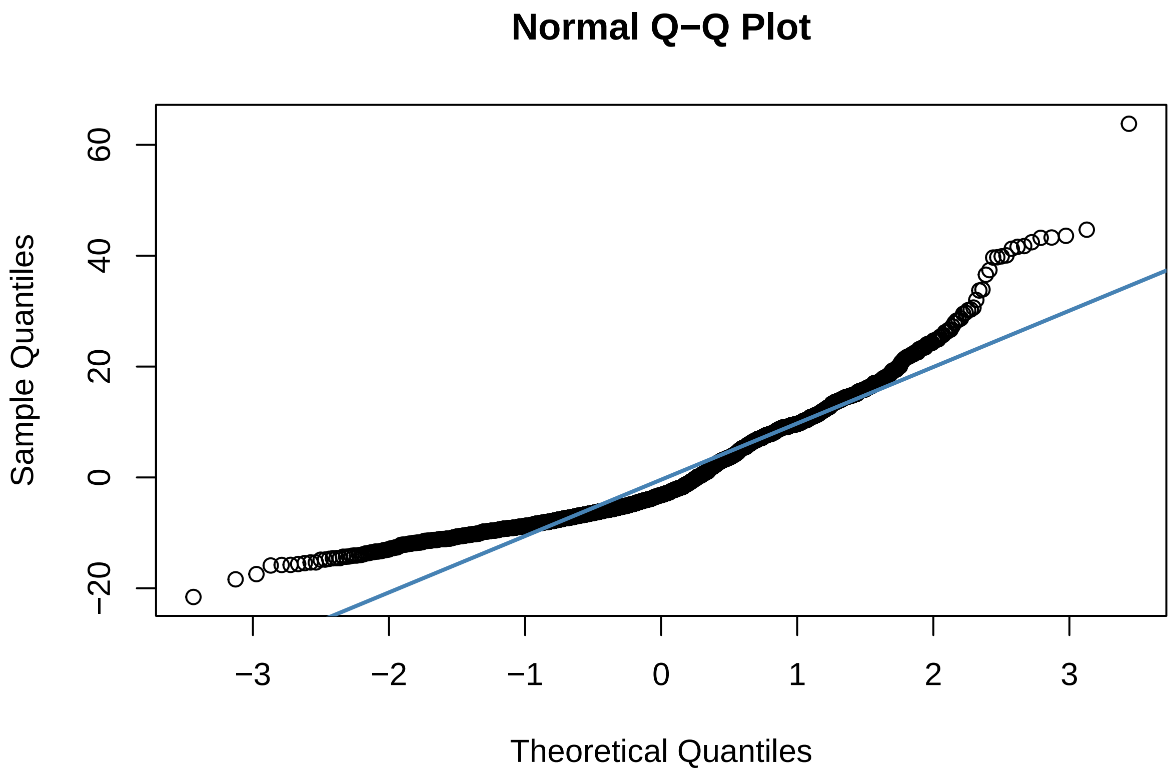
    The ANCOVA results demonstrated that both the car model and location significantly affected car prices ($p < 0.05$ for the interaction term). Thus, we reject the null hypotheses (H0) and accept the alternative hypotheses (Ha), concluding that different car models and locations significantly affect car prices, after adjusting mileage(p=0.00105).

Assess Model Conditions

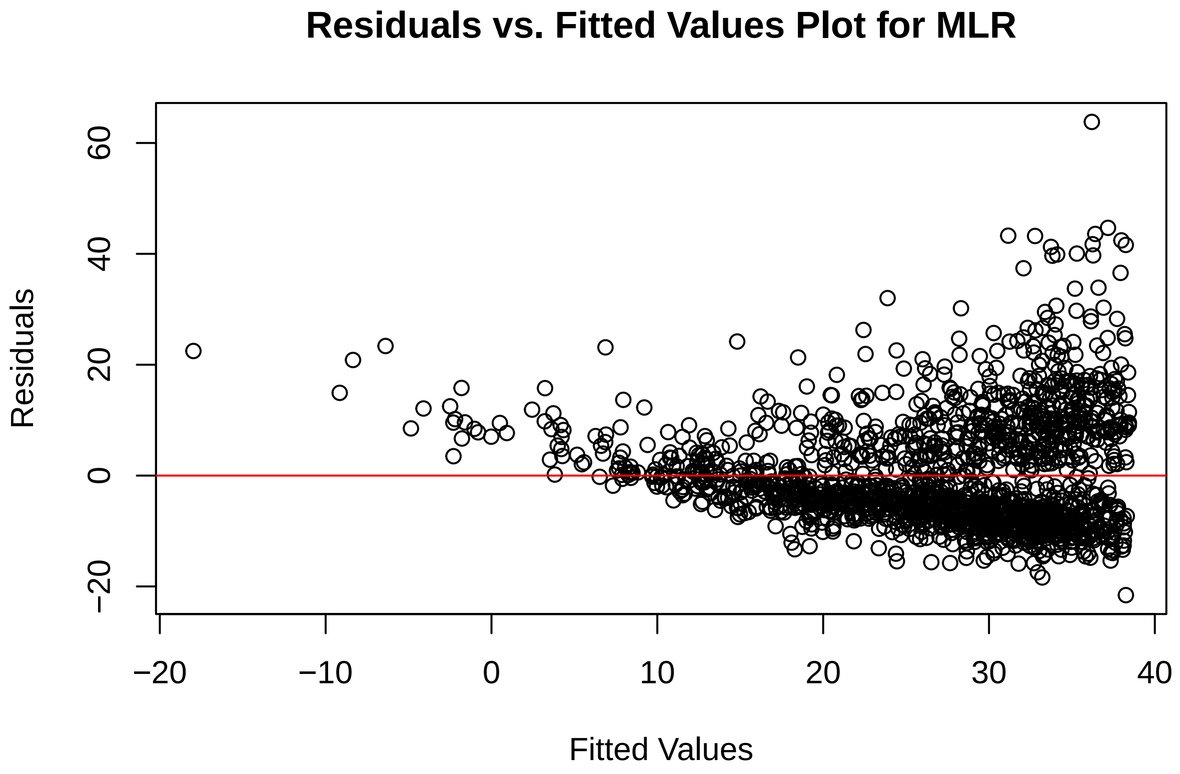
MLR

```
# Fitting the MLR model
mlr_model <- lm(price ~ mileage, data = combined_data)
```

```r
# Check for normality of residuals
qqnorm(residuals(mlr_model))
qqline(residuals(mlr_model), col = "steelblue", lwd = 2)
```



```r
# Check for homoscedasticity
plot(fitted(mlr_model), residuals(mlr_model),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values Plot for MLR")
abline(h = 0, col = "red")
```



### Comments

Linearity: The data seems to be random and the linearity condition seems to be met with the data points being scattered randomly across the Residuals vs Fitted Values.
#### Independence The data is assumed to be independent due to the design of the data.

Normality The QQ-plot shows that the tail ends are not symmetric with the data which could lead to some issues with normality. #### Independence

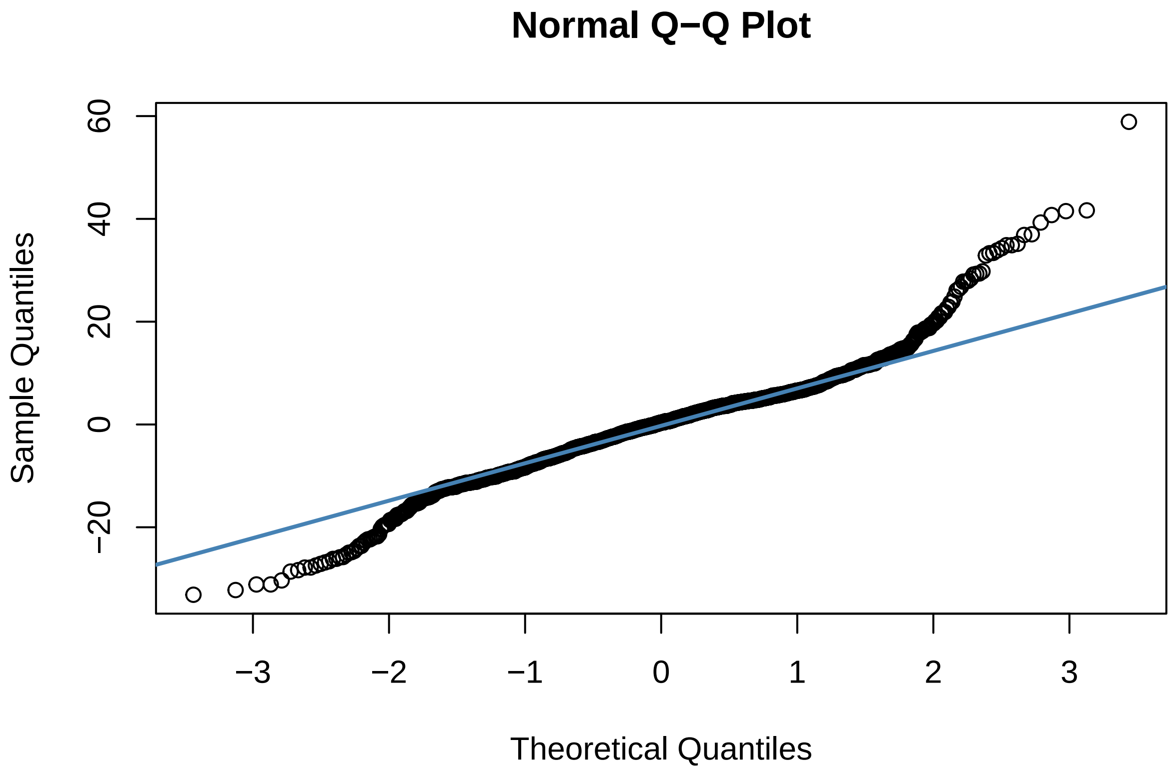
Equal Variance The Residuals seem to have non constant variance in which the points seem to be skewed to the right

Randomness The data collected from the source seems to be a random sampling of data points.

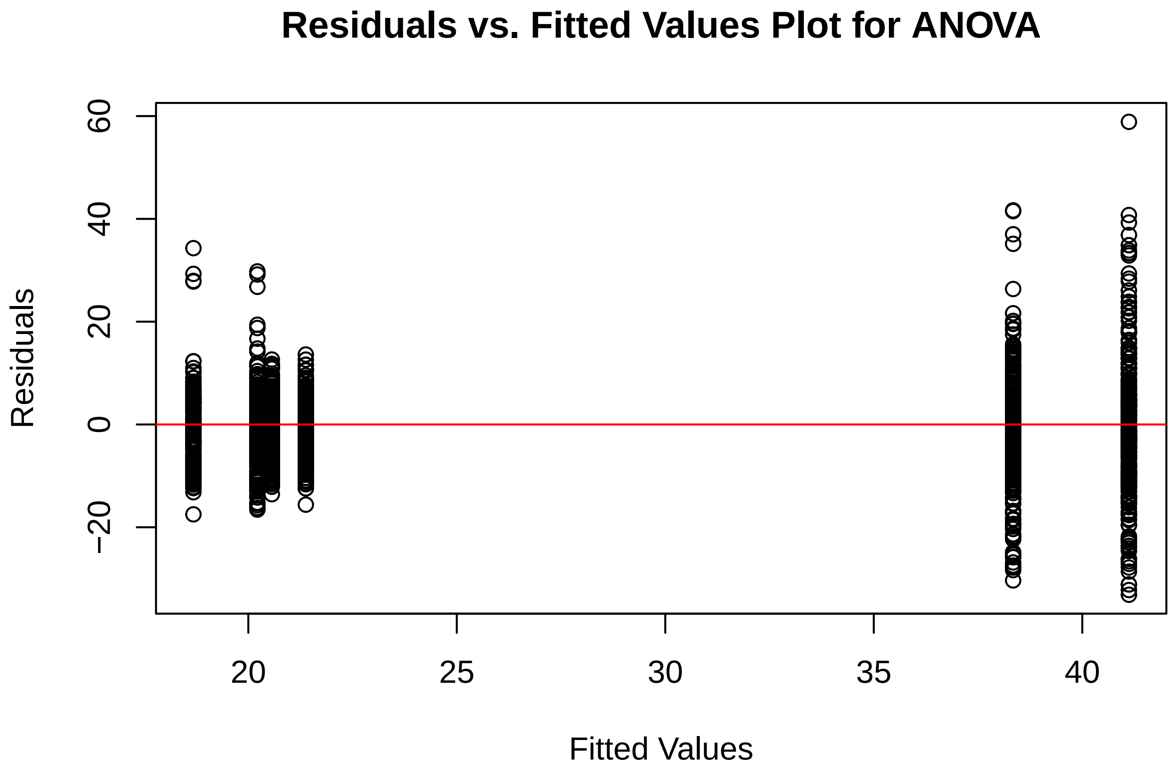
Anova

```
# Fitting the ANOVA model
anova_model1 <- aov(price ~ model + location + model:location, data = combined_data)

# Check for normality of residuals
qqnorm(residuals(anova_model1))
qqline(residuals(anova_model1), col = "steelblue", lwd = 2)
```



```r
# Check for homoscedasticity
plot(fitted(anova_model1), residuals(anova_model1),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values Plot for ANOVA")
abline(h = 0, col = "red")
```



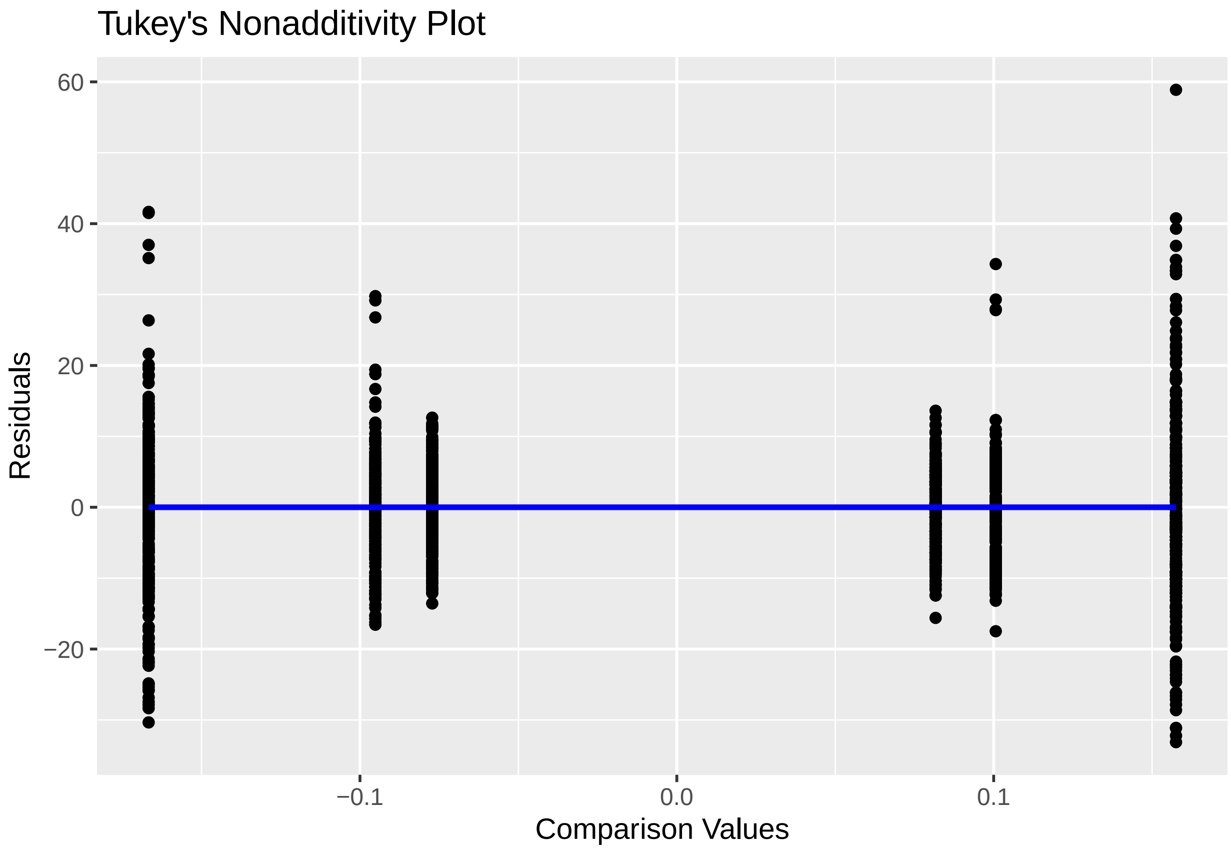
```
library(dplyr)
library(ggplot2)

# Fit the model
model_fit <- lm(price ~ model + location + model:location, data = combined_data)

# Summarize overlays with needed effects
Data_Overlays <- combined_data %>%
  mutate(
    grand_mean = mean(price), # Calculate the grand mean of price
    model_effect = ave(price, model, FUN = mean) - grand_mean, # Effect of each model
    location_effect = ave(price, location, FUN = mean) - grand_mean # Effect of each location
  ) %>%
  mutate(
    # Compute residuals from the model
    Residuals = residuals(model_fit),
    # Compute Comparison Values for Tukey Nonadditivity Plot
    Comparison_Values = model_effect * location_effect / grand_mean
  )

# Plot residuals vs comparison values & fit simple linear regression
ggplot(Data_Overlays, aes(x = Comparison_Values, y = Residuals)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(
    title = "Tukey's Nonadditivity Plot",
    x = "Comparison Values",
    y = "Residuals"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
# Fit a linear model to residuals vs comparison values
comparison_fit <- lm(Residuals ~ Comparison_Values, data = Data_Overlays)
summary(comparison_fit)
```

```
##
## Call:
## lm(formula = Residuals ~ Comparison_Values, data = Data_Overlays)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.128  -5.171   0.371   4.654  58.880
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4.536e-16  2.194e-01       0        1
## Comparison_Values  2.080e-14  1.831e+00       0        1
##
## Residual standard error: 9.05 on 1700 degrees of freedom
## Multiple R-squared:  1.092e-31,  Adjusted R-squared:  -0.0005882
## F-statistic: 1.856e-28 on 1 and 1700 DF,  p-value: 1
```

```
# Calculate the optimal transformation power
slope <- coef(comparison_fit)["Comparison_Values"]
power <- 1 - slope
power <- ifelse(power == 0, "logarithmic transformation", power)

# Output the optimal transformation
print(paste("The transformation power is", power))
```

```
## [1] "The transformation power is 0.999999999999979"
```

Comments

Constant Additivity The data seems to have to have some outriders and a lot deviation from the residual line. However, there seems to be no pattern therefore this condition acceptable.

Independence This is assumed based on how the data was collected from the provided data set. #### Normality According to the QQ-plot the data seems to be Normal.

Equal Variance The points look to be normally distributed through the data. #### Randomness The data is assumed to be random from the way the data was collected.

After calculating the optimal transformation, the data was close to 1 therefore there is no transformation needed. ##MLR Transformation

```
# Check if any prices are zero or negative which would prevent log transformation
sum(combined_data$price <= 0)
```

```
## [1] 0
```

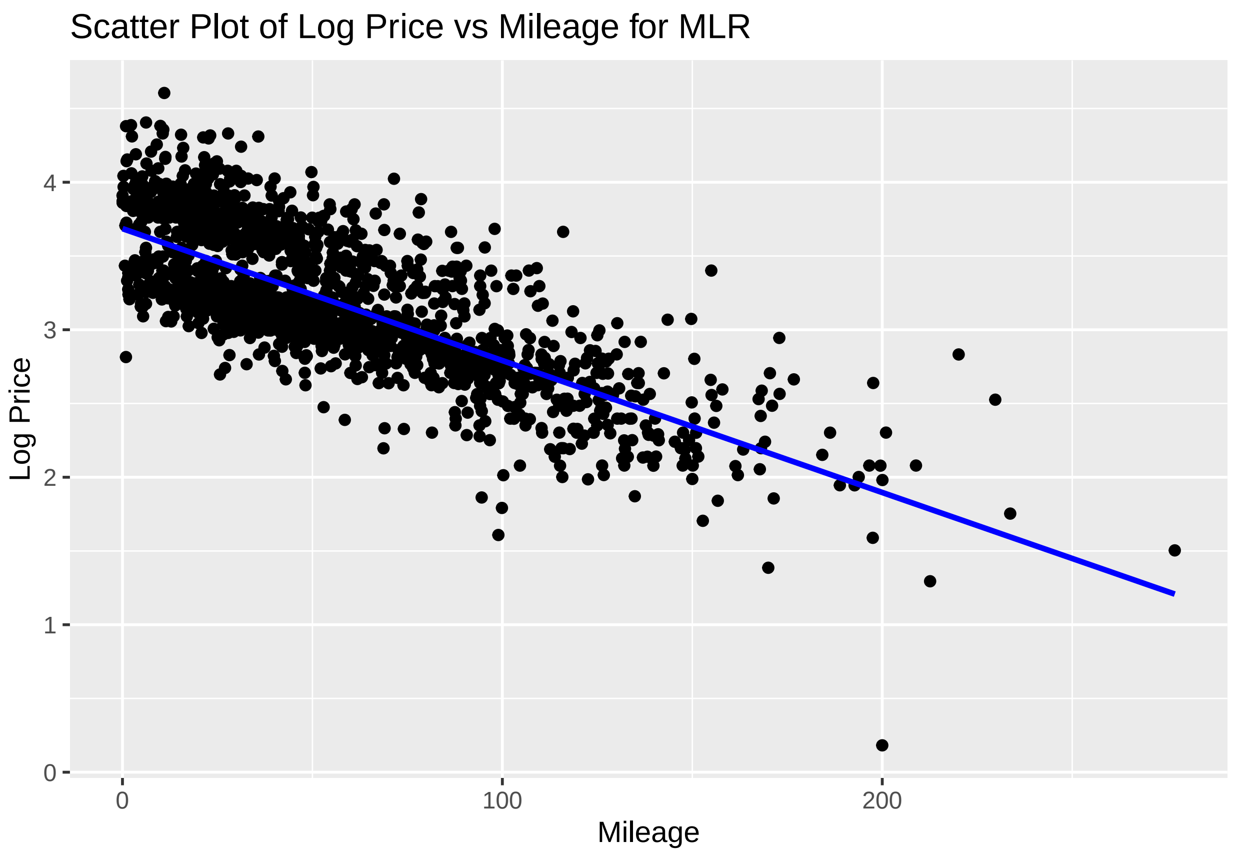
```
# Assuming all prices are positive
combined_data$log_price <- log(combined_data$price)

# Fit the MLR model with the transformed price
log_mlr_model <- lm(log_price ~ mileage, data = combined_data)
summary(log_mlr_model)
```

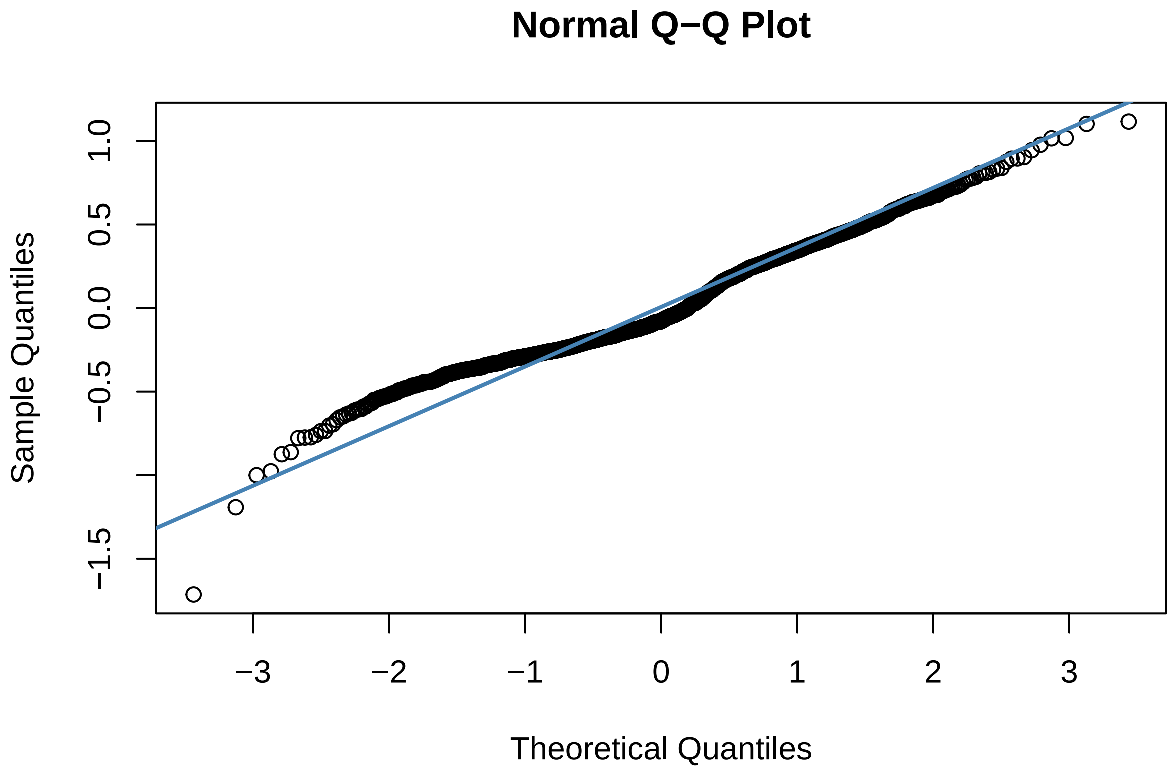
```
##
## Call:
## lm(formula = log_price ~ mileage, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71417 -0.23392 -0.07919  0.24696  1.11581
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.6855637  0.0130538  282.34   <2e-16 ***
## mileage     -0.0089453  0.0001898  -47.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3189 on 1700 degrees of freedom
## Multiple R-squared:  0.5666, Adjusted R-squared:  0.5663
## F-statistic:  2222 on 1 and 1700 DF,  p-value: < 2.2e-16
```

```
# Checking linearity and independence for transformed MLR
ggplot(combined_data, aes(x = mileage, y = log_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Scatter Plot of Log Price vs Mileage for MLR",
       x = "Mileage", y = "Log Price")
```

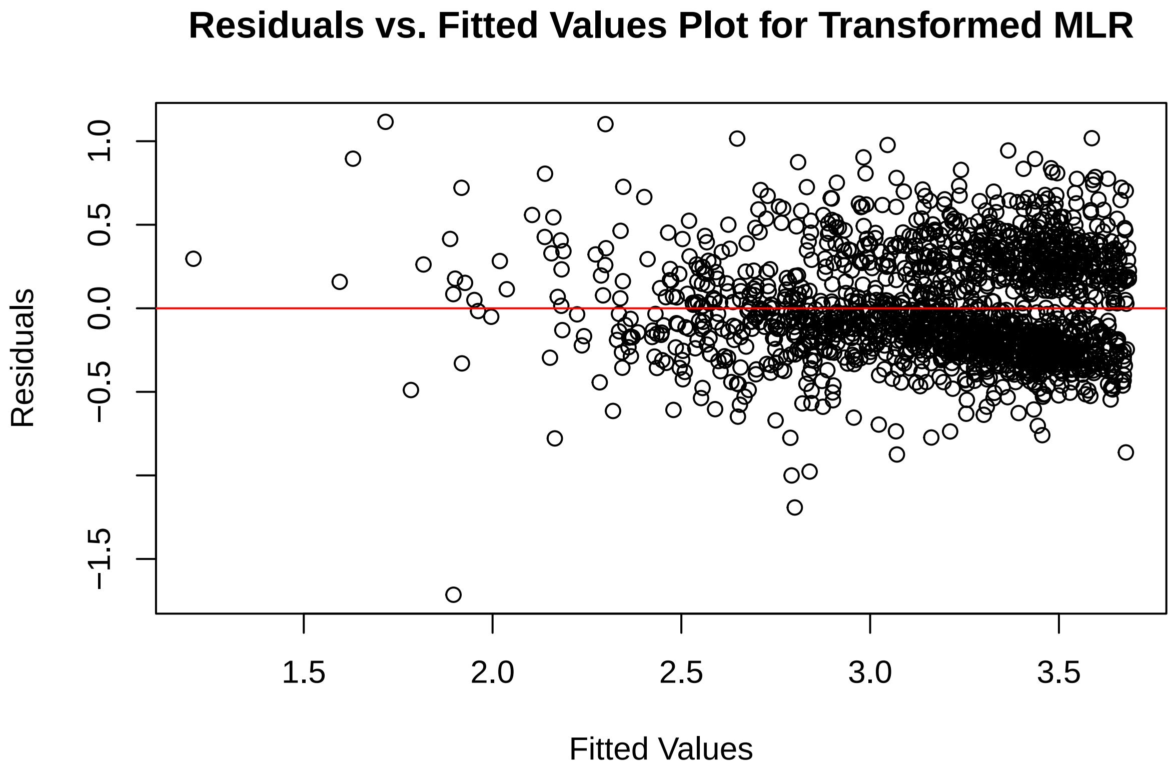
```
## `geom_smooth()` using formula = 'y ~ x'
```



```
# Checking normality of residuals for transformed MLR
qqnorm(residuals(log_mlr_model))
qqline(residuals(log_mlr_model), col = "steelblue", lwd = 2)
```



```r
# Checking homoscedasticity for transformed MLR
plot(fitted(log_mlr_model), residuals(log_mlr_model),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values Plot for Transformed MLR")
abline(h = 0, col = "red")
```



Comments

After the transformation, the data seems to fit the LINER conditions more suggesting a log transformation was needed for the Normality condition to be met.

Conclusions

The primary objective of this study was to investigate the influence of used vehicle pricing based on the geographical location, mileage, and car model. The analysis revealed that there is a significant difference in pricing based on the anova model. The Ford F-150 had an average price that was higher than both the Jeep Cherokee and Honda Civic, about $18,000 and $20,000 respectively. While the location did not show a significant impact on the pricing the interaction suggests there is a price difference between the Ford F-150 and other models was more pronounced in State College than in Charlotte. It is better to sell the F-150 at the State College and avoid buying them at State College. In the model as mileage increases the price decreases which is around $8.95 per 1,000 miles added to the vehicle. This research provided a foundational understanding of how car pricing varies by model, location, and model.