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

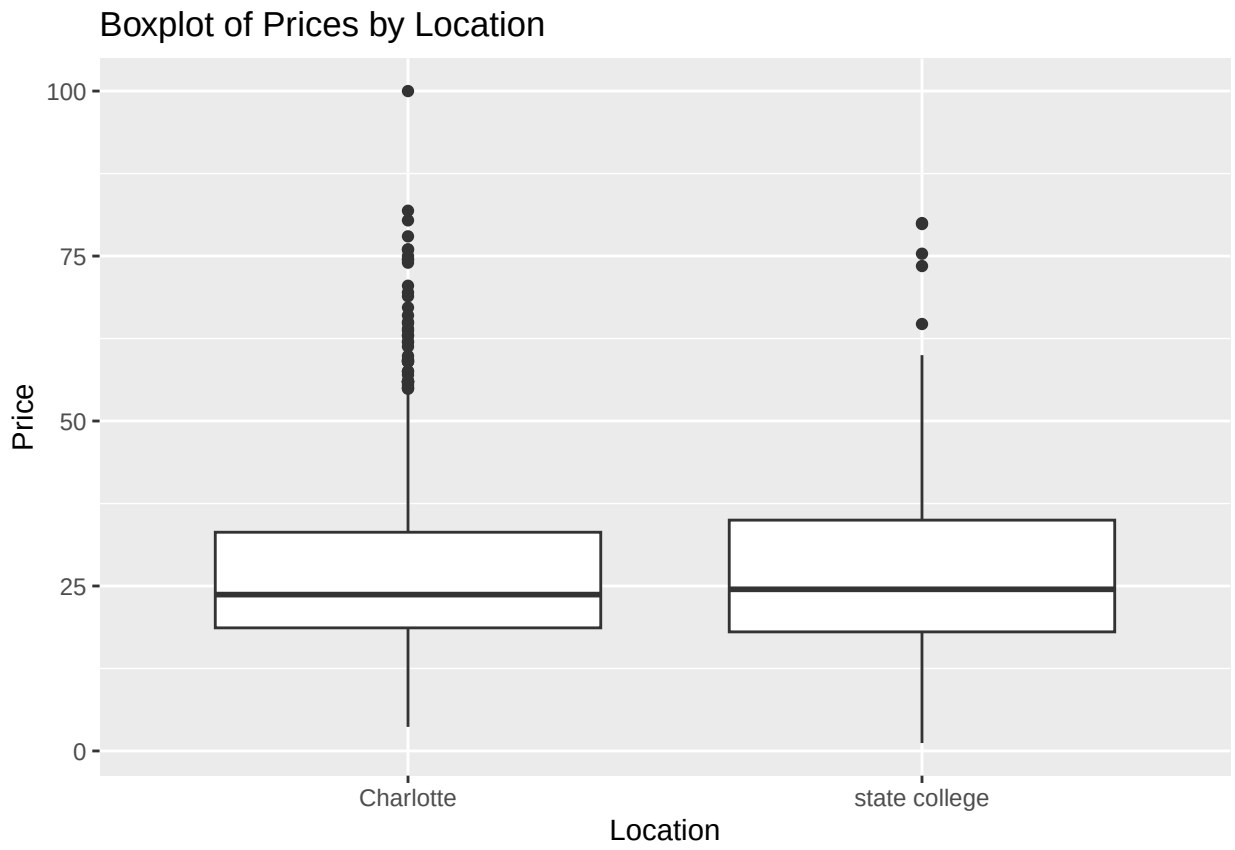
Data collation

```
#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
```

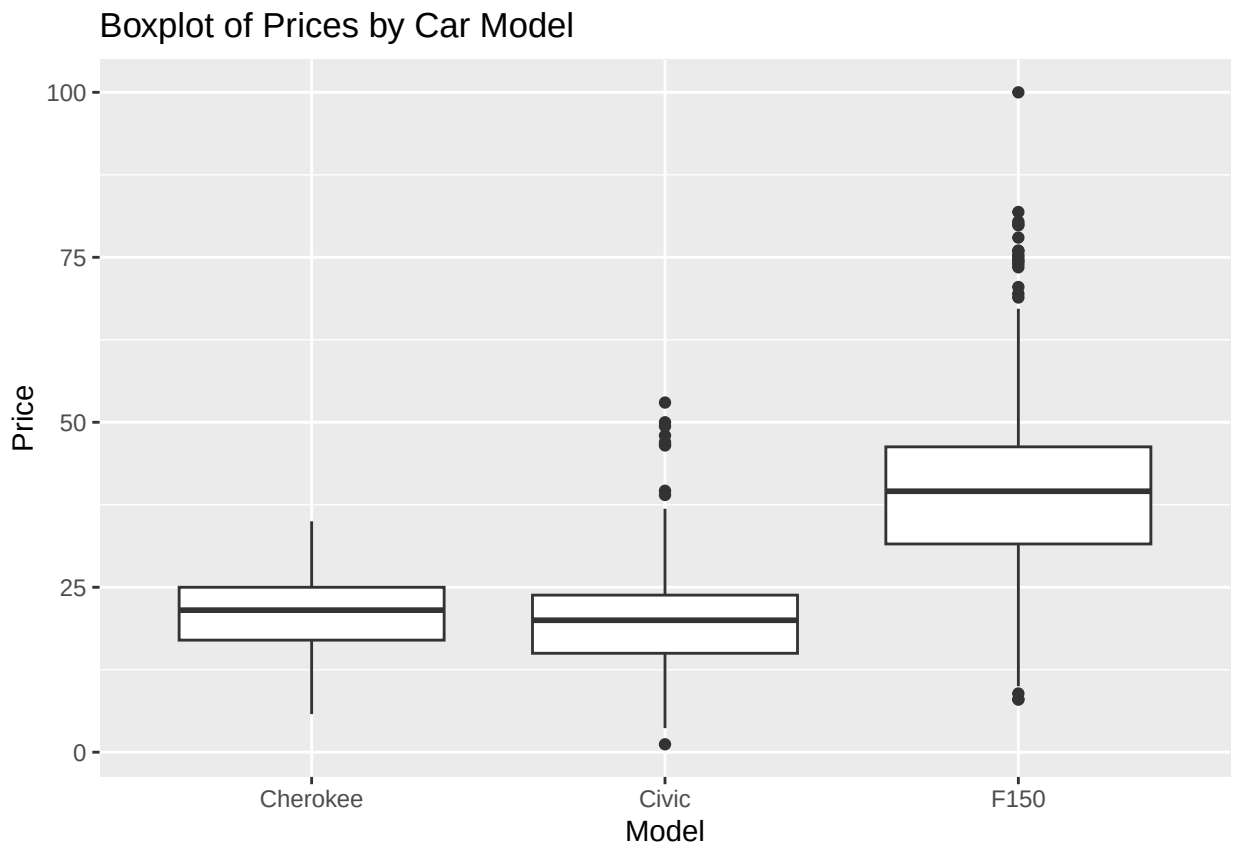
```
## Joining with `by = join_by(year, price, mileage, location, model)`
## Joining with `by = join_by(year, price, mileage, location, model)`
## Joining with `by = join_by(year, price, mileage, location, model)`
## Joining with `by = join_by(year, price, mileage, location, model)`
## Joining with `by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

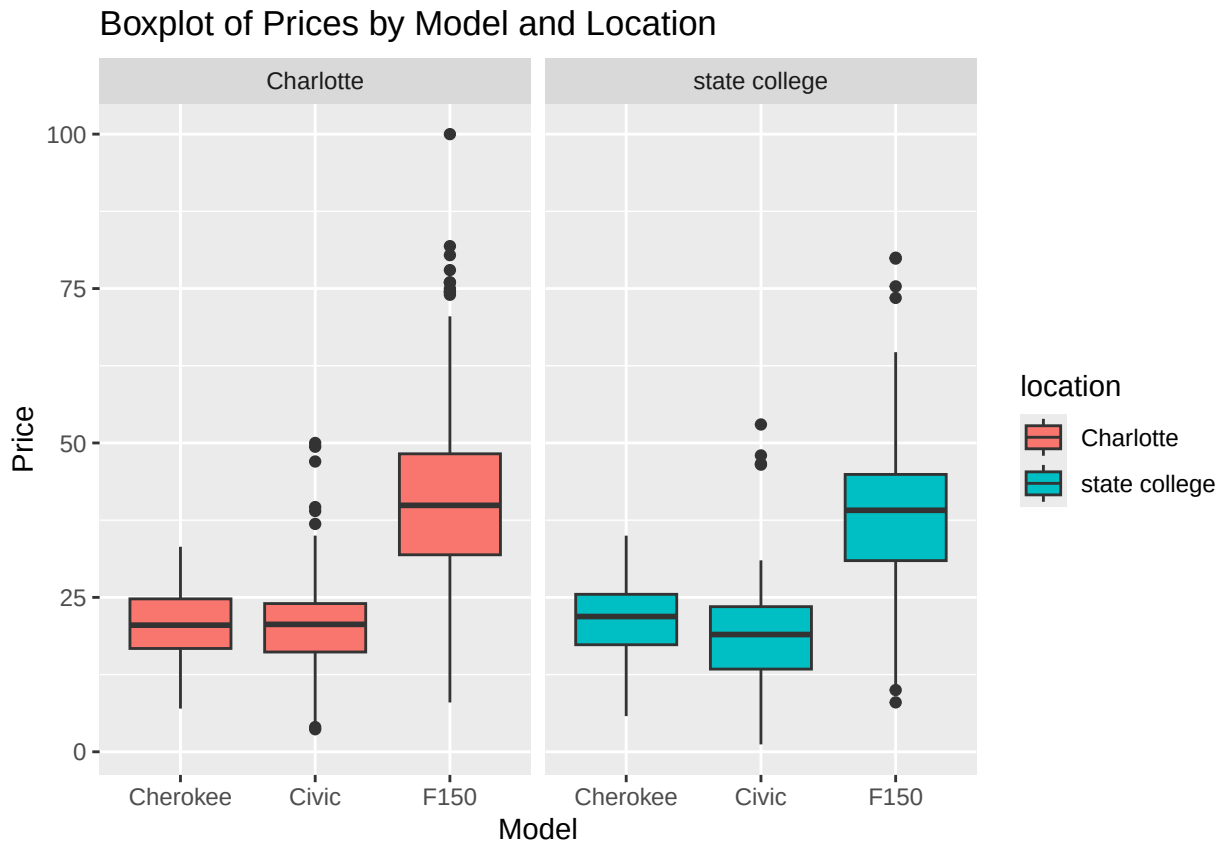
```r
# Boxplot for price by location
ggplot(combined_data, aes(x = location, y = price)) +
  geom_boxplot() +
  labs(title = "Boxplot of Prices by Location", x = "Location", y = "Price")
```




```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")
```

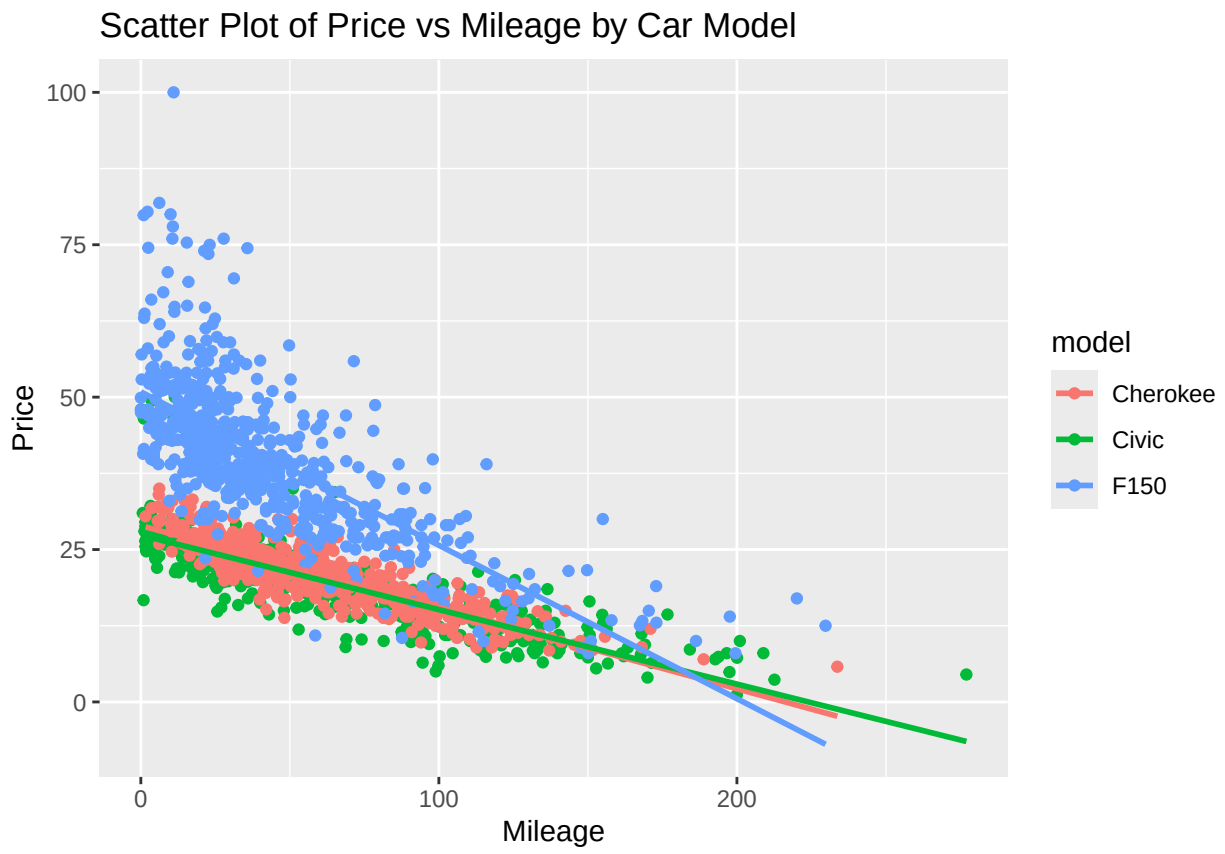



```r
# Boxplot for price by model and location interaction
ggplot(combined_data, aes(x = model, y = price, fill = location)) +
  geom_boxplot() +
  labs(title = "Boxplot of Prices by Model and Location", x = "Model", y = "Price") +
  facet_wrap(~ location)
```




```
# 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'
```




```r
# 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")
```

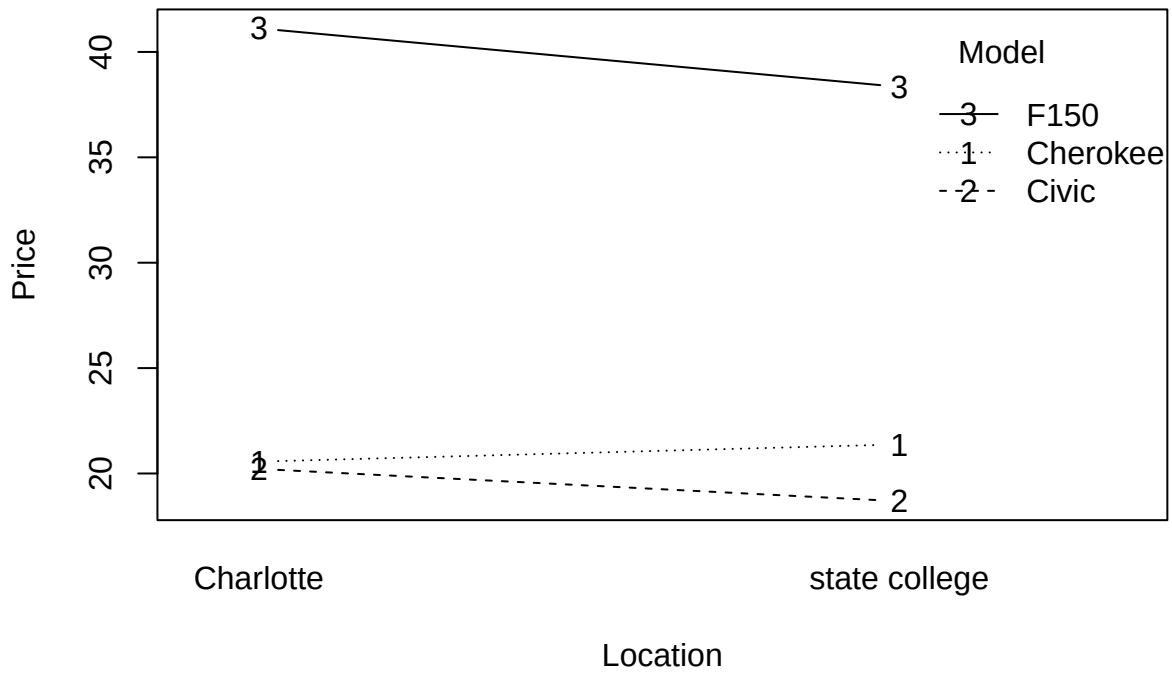


Image 1: Boxplot of Prices by Location

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

```
#ANCOVA
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 :**

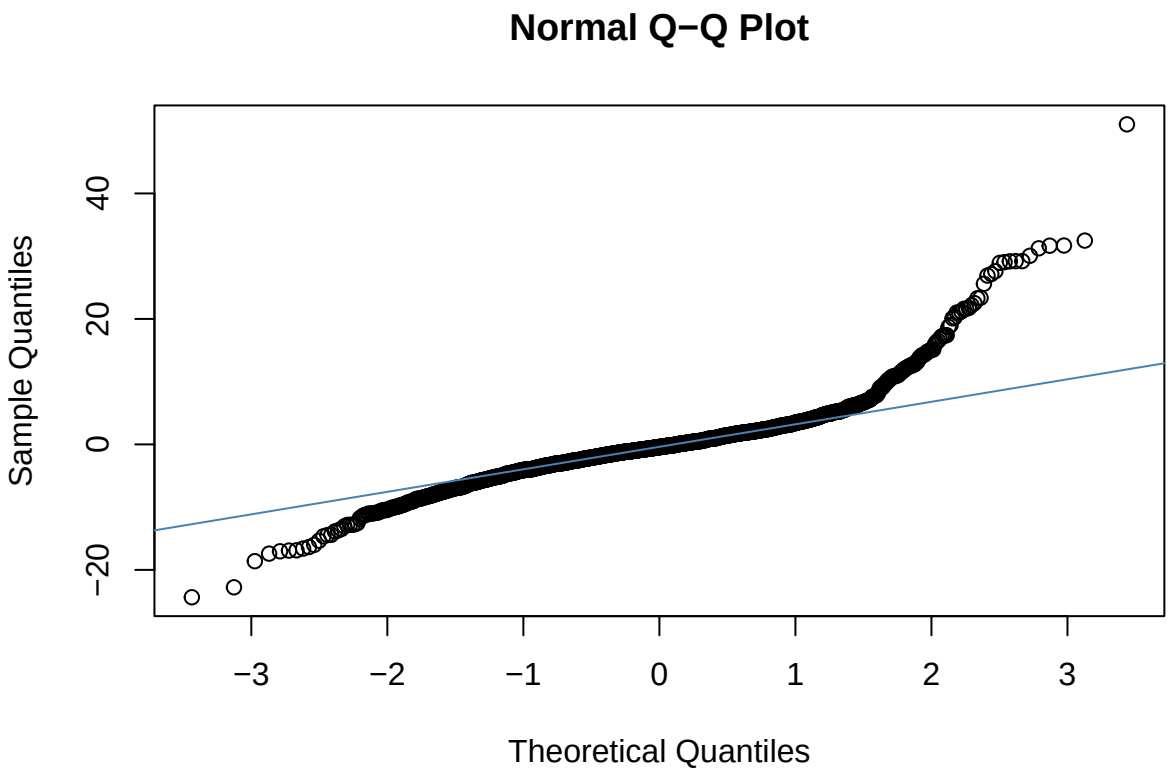
   - **Degrees of Freedom (Df):** The model has 2 degrees of freedom, which suggests there are three different car models being compared.
   - **Sum of Squares (Sum Sq):** The Sum Sq for the model is 539, which is a measure of the total variation attributed to the differences in the mean prices across the 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

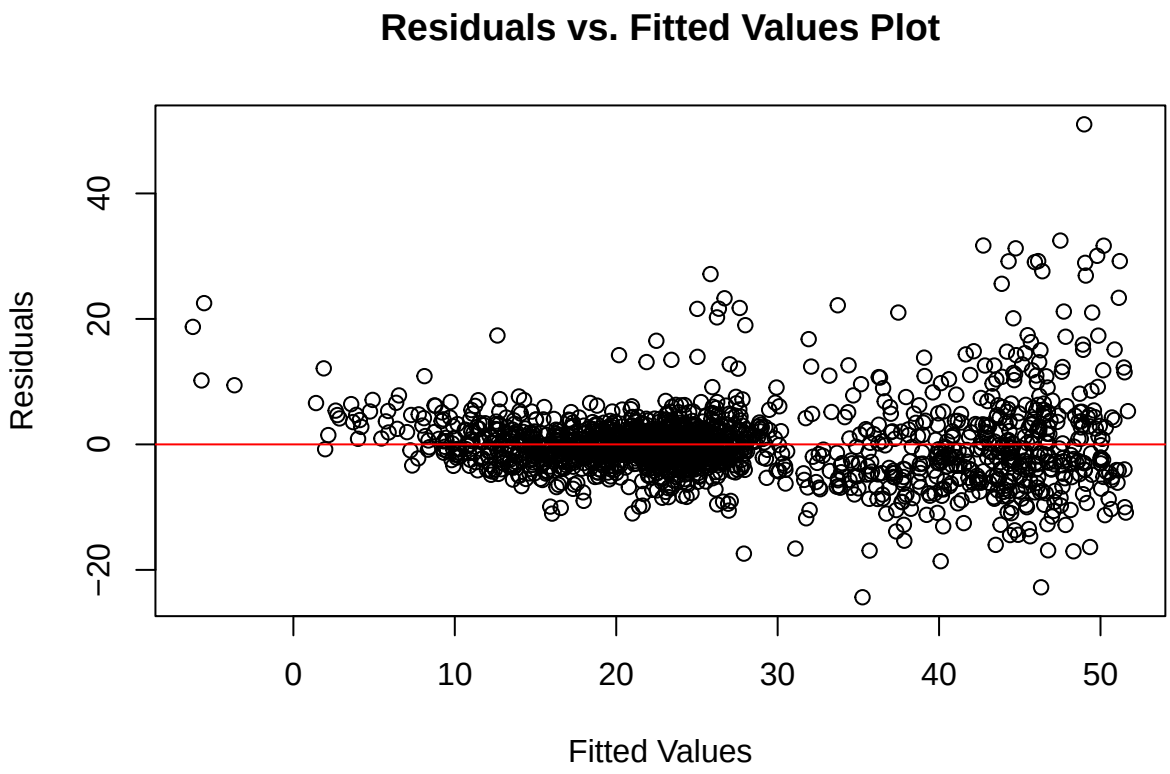
Checking Normality

```
# Fit the ANCOVA model
ancova_model <- lm(price ~ model * mileage + location, data = combined_data)

# Checking normality of residuals using a Q-Q plot
qqnorm(residuals(ancova_model))
qqline(residuals(ancova_model), col = "steelblue")
```



Checking Homoscedasticity

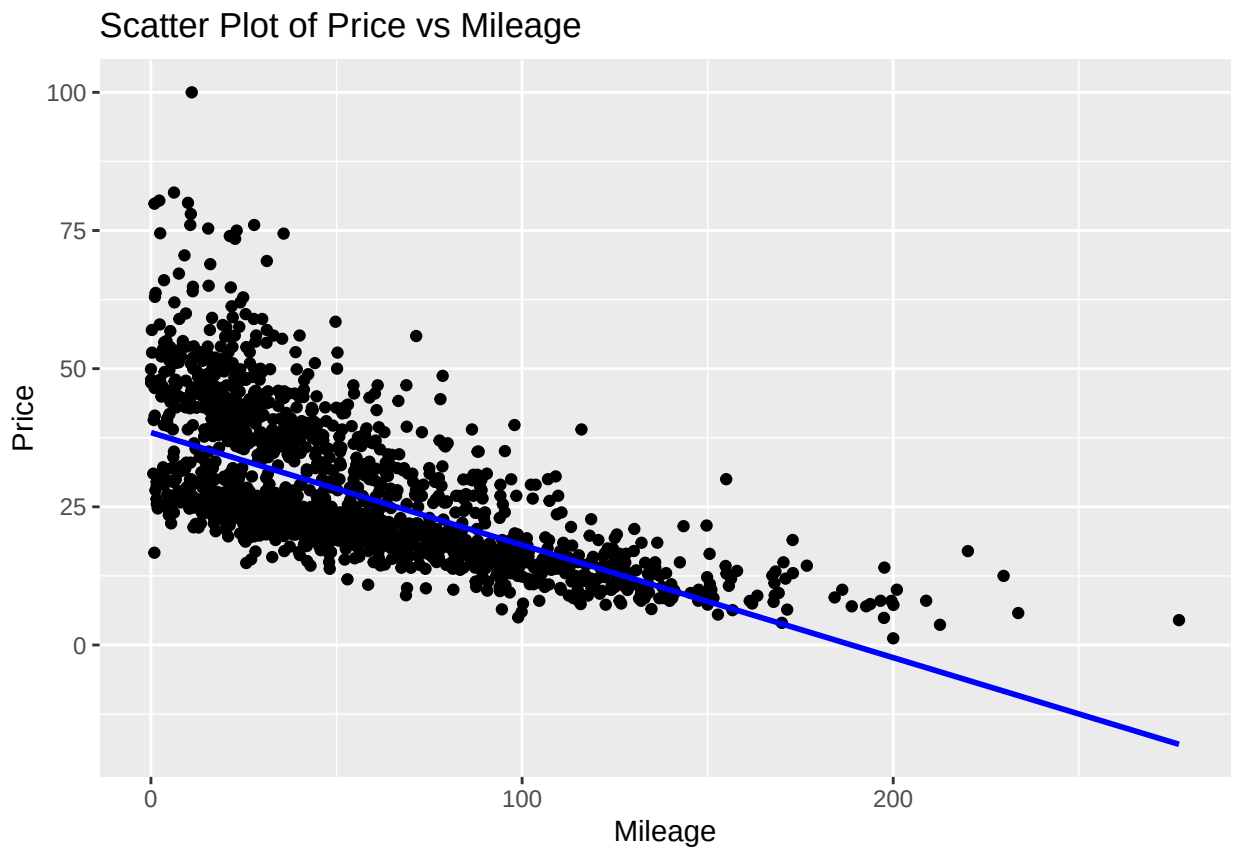
```
# Scatter plot of residuals against fitted values to assess homoscedasticity
plot(fitted(ancova_model), residuals(ancova_model),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values Plot")
abline(h = 0, col = "red")
```



Checking Independence and Linearity

```
# Scatter plot to check the linearity assumption between mileage and price
ggplot(combined_data, aes(x = mileage, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Scatter Plot of Price vs Mileage",
       x = "Mileage", y = "Price")
```

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



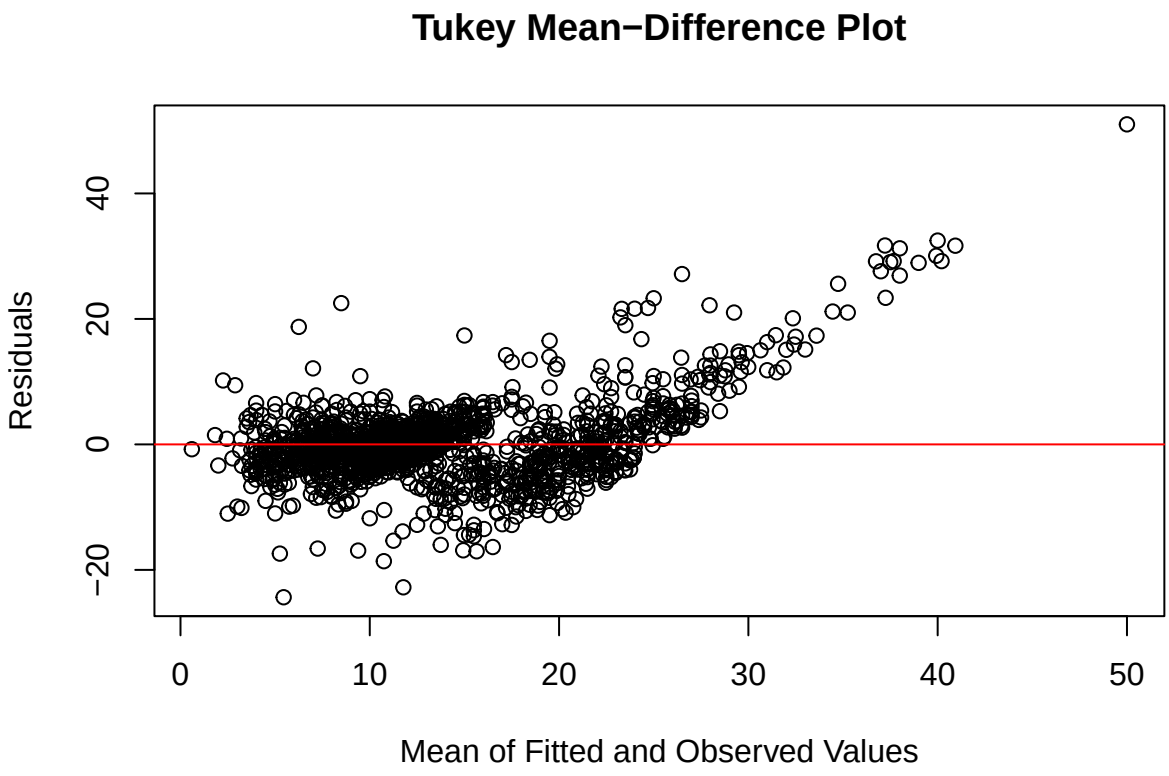
Comments

Normality The qq-plot of shows devations from normality esspecially at the tails of the distribution #### Homoscedasticity - The residuals vs. fitted values plot shows a random dispersion of residuals around the horizontal line, suggesting that homoscedasticity is reasonable.

Independence and Linearity

- The scatter plot of price vs. mileage shows a clear negative trend, supporting the assumption of linearity. The random dispersion of residuals suggests that the assumption of independence is met.

```
# Fit the ANCOVA model
ancova_model <- lm(price ~ model * mileage + location, data = combined_data)

# Calculate the mean of each fitted value and its corresponding residual
residuals <- residuals(ancova_model)
fitted_values <- fitted(ancova_model)
mean_residuals <- apply(cbind(residuals, fitted_values), 1, mean)

# Create the Tukey mean-difference plot
plot(mean_residuals, residuals,
     xlab = "Mean of Fitted and Observed Values",
     ylab = "Residuals",
     main = "Tukey Mean-Difference Plot")
abline(h = 0, col = "red")
```



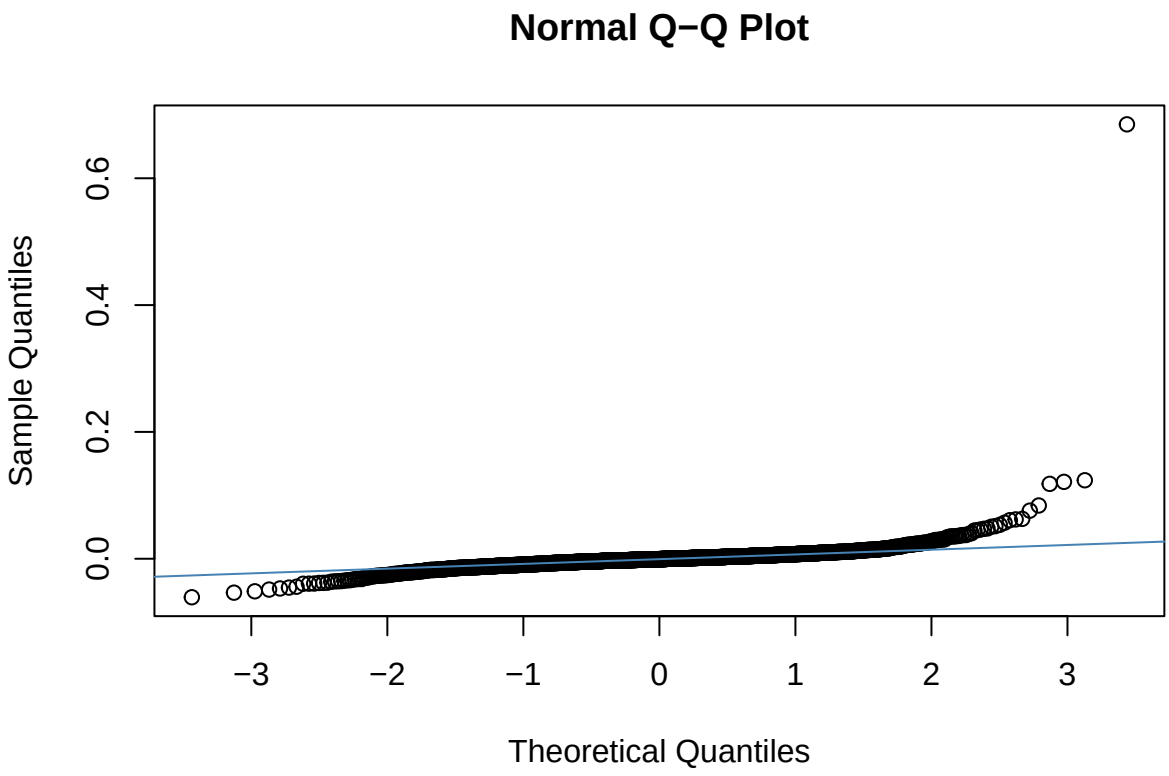

```r
anova_model <- aov(price ~ model + location, data = combined_data)

# Perform Tukey HSD test
TukeyHSD(anova_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = price ~ model + location, data = combined_data)
##
## $model
##                       diff       lwr        upr     p adj
## Civic-Cherokee -1.432657 -2.718115 -0.1471993 0.0244604
## F150-Cherokee  18.741300 17.497313 19.9852867 0.0000000
## F150-Civic     20.173957 18.901371 21.4465439 0.0000000
##
## $location
##                                diff       lwr        upr     p adj
## state college-Charlotte -1.170832 -2.035209 -0.3064547 0.0079639
```

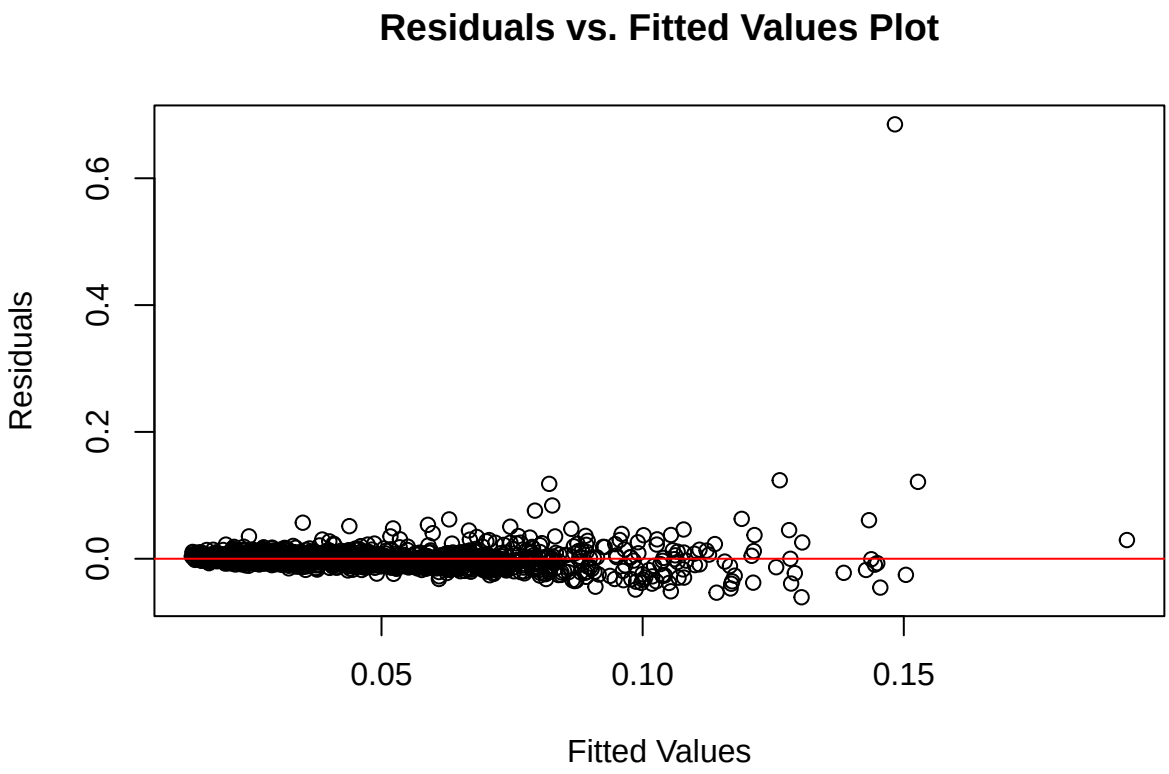
```r
# Apply log transformation to the 'price' variable to stabilize variance
combined_data$inv_price <- 1/combined_data$price

inv_ancova_model <- lm(inv_price ~ model * mileage + location, data = combined_data)
summary(inv_ancova_model)
```

```
##
## Call:
## lm(formula = inv_price ~ model * mileage + location, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06071 -0.00571 -0.00061  0.00437  0.68503
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.459e-02  1.804e-03  13.628  < 2e-16 ***
## modelCivic             -3.959e-03  2.299e-03  -1.722 0.085191 .
## modelF150              -1.099e-02  2.162e-03  -5.083 4.13e-07 ***
## mileage                 4.281e-04  2.445e-05  17.510  < 2e-16 ***
## locationstate college   3.418e-03  1.023e-03   3.342 0.000848 ***
## modelCivic:mileage      1.932e-04  3.127e-05   6.179 8.04e-10 ***
## modelF150:mileage      -1.221e-04  3.381e-05  -3.611 0.000314 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02097 on 1695 degrees of freedom
## Multiple R-squared:  0.5693, Adjusted R-squared:  0.5678
## F-statistic: 373.4 on 6 and 1695 DF,  p-value: < 2.2e-16
```

```r
# And then check the model assumptions again, starting with normality
qqnorm(residuals(inv_ancova_model))
qqline(residuals(inv_ancova_model), col = "steelblue")
```




```r
# Homoscedasticity check using residuals vs. fitted values plot
plot(fitted(inv_ancova_model), residuals(inv_ancova_model),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values Plot")
abline(h = 0, col = "red")
```



Conclusions

The primary objective of this study was to investigate the influence of used vechiles pricing based on the geographical location, milage, and car model.The car model has a significant impact on the pricing of the used vechiles. The ANOVa results indicate that differnt car modesl have differnt average prices with the F-150 having the higher car price between the two cars. While locatin was not found to be significant the ANOVA model, there was a significant interaction effect between location and car model on car prices. This implies that the influnce of car model on prive varies by geographic location. The dignostic plots revealed some devations from nomaloty and the presence of outliers but suporrted the CAINER assumotions well enogh to go on with the experiment. The findings highlight the complextiy of used cehile priceing and the importance of cinsidering multifacted set of factors. Consumers and auto dealers can make a estimated guess on what the car vehicle price should be based on the milage, model and location of where it is being sold or bought. While the study provided beneficial insights, further research could explore additional factors such as the vehicle's condition, features, and market demand to enhance the predictive accuracy of car pricing models. In conclusion tje rei;ts offer a foundation for a more informed understanding of vehicle valuation.