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(s)

Two-Way Factorial Model with Interaction (ANOVA Model 1)

Research Question How do location and car model both independently and interactively influence car prices? Specifically, does the effect of car model on pricing differ depending on the location?

Hypotheses

- **Null Hypotheses (H0):**
  - **H0a:** There is no significant main effect of location on car prices, meaning that car prices do not differ between locations.
  - **H0b:** There is no significant main effect of model on car prices, meaning that car prices do not differ among models.
  - **H0c:** There is no significant interaction effect between location and model on car prices, meaning that the effect of the car model on price does not vary by location.
- **Alternative Hypotheses (Ha):**
  - **Ha1:** There is a significant main effect of location on car prices.
  - **Ha2:** There is a significant main effect of model on car prices.
  - **Ha3:** There is a significant interaction effect between location and model on car prices.

ANCOVA Model (ANOVA Model 2)

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

Hypotheses

- **Null Hypotheses (H0):**
  - **H0d:** Car model does not significantly affect the price of cars when adjusting for mileage, meaning all models would have the same pricing adjusted for mileage.
  - **H0e:** Mileage does not significantly influence the price of cars when adjusting for the model, suggesting that changes in mileage do not affect the pricing once the model is accounted for.
- **Alternative Hypotheses (Ha):**
  - **Ha4:** Car model significantly affects the price of cars when adjusting for mileage.
  - **Ha5:** Mileage significantly influences the price of cars when adjusting for the model.

EDA

Data collation

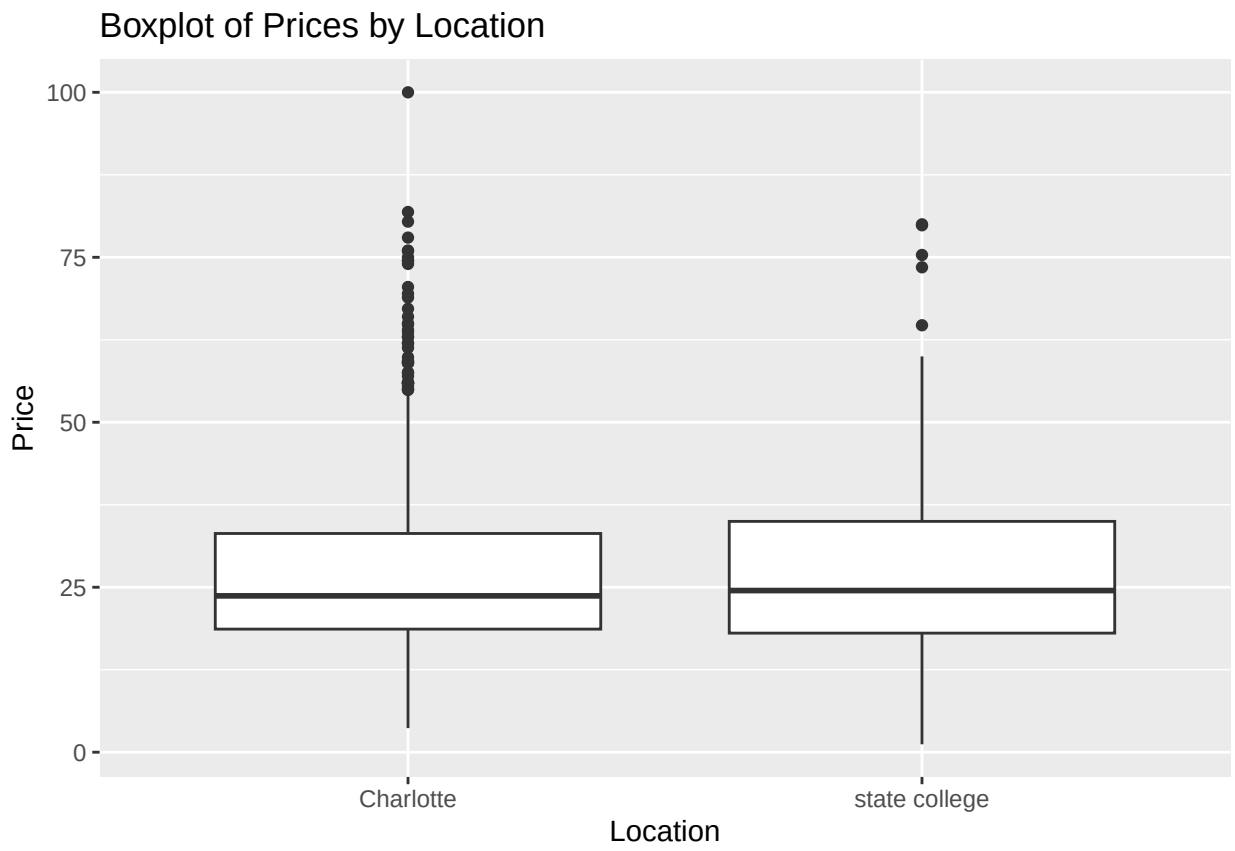
```
#Load the datasets
sc_civic<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/16801_civic.csv")
nc_civic<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/28207_civic.csv")
sc_Cherokee<-read.csv("https://raw.githubusercontent.com/Migueldesanta/Anova/main/16801_Cherokee.csv")
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"
```

```r
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
```

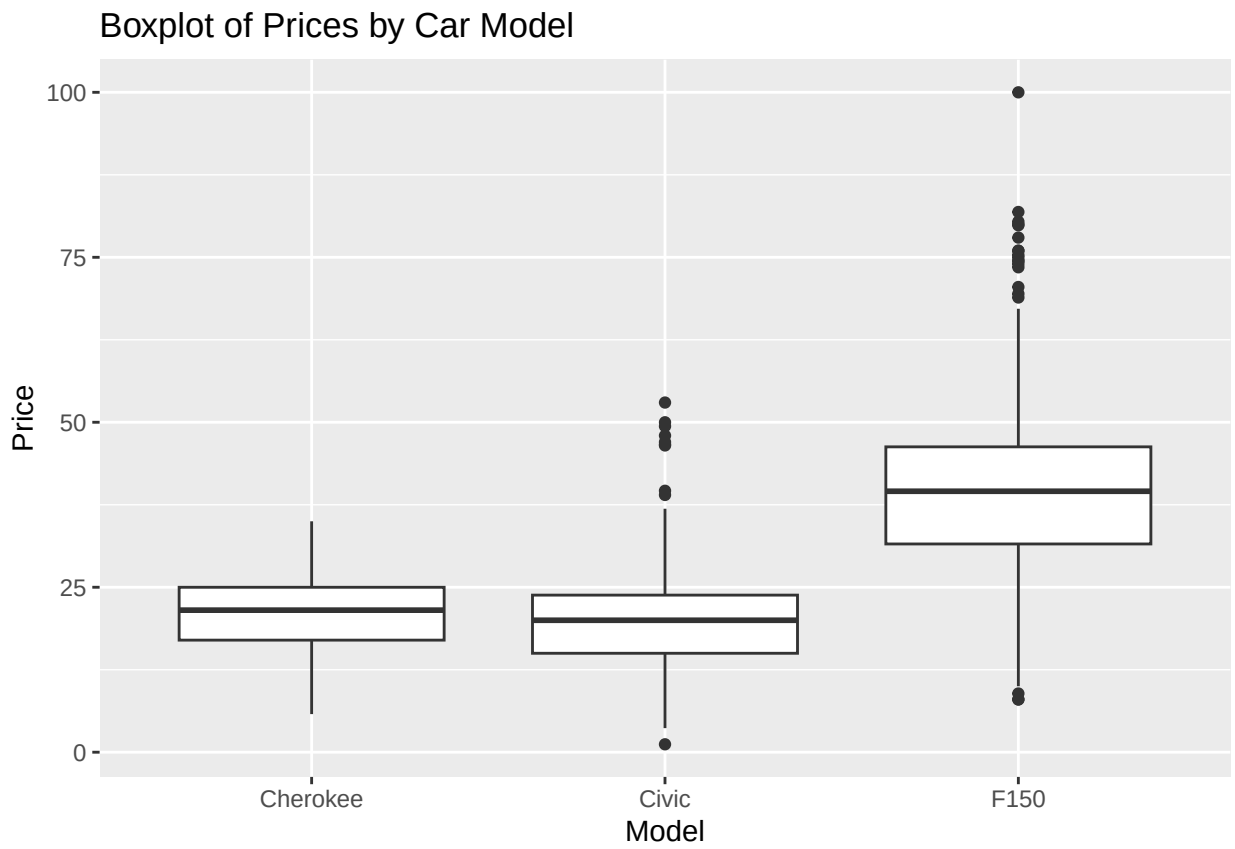
```
## Joining with `by = join_by(year, price, mileage, location, model)`
## Joining with `by = join_by(year, price, mileage, location, model)`
## Joining with `by = join_by(year, price, mileage, location, model)`
## Joining with `by = join_by(year, price, mileage, location, model)`
## Joining with `by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

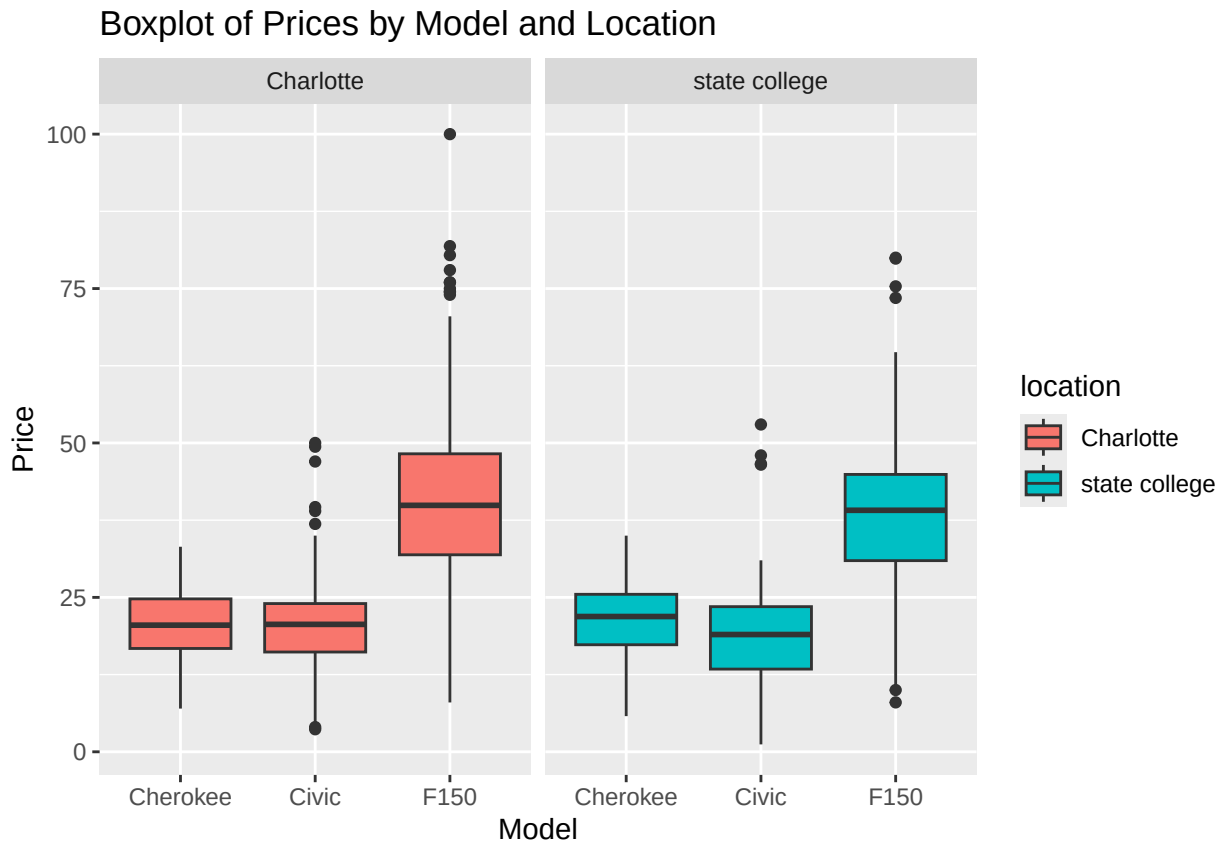
```r
# Boxplot for price by location
ggplot(combined_data, aes(x = location, y = price)) +
  geom_boxplot() +
  labs(title = "Boxplot of Prices by Location", x = "Location", y = "Price")
```




```
# 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")
```

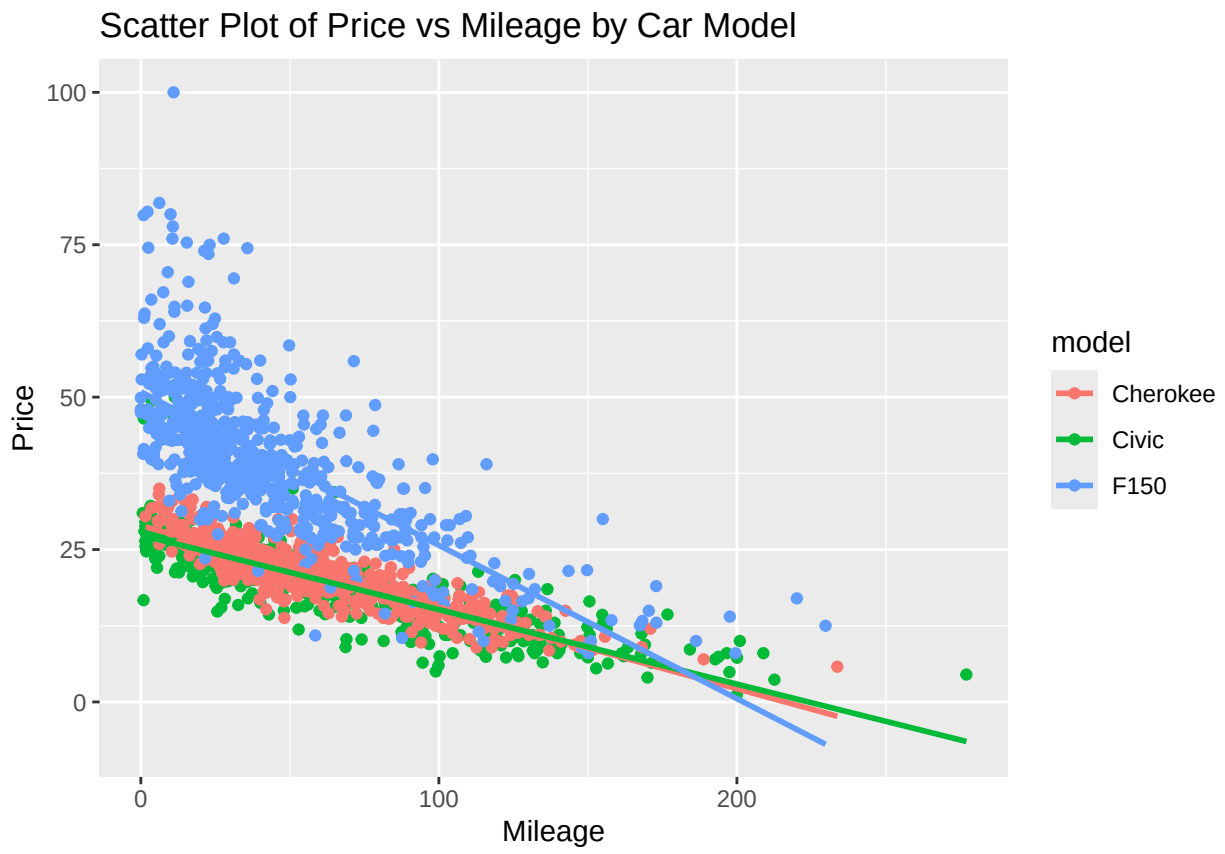



```r
# Boxplot for price by model and location interaction
ggplot(combined_data, aes(x = model, y = price, fill = location)) +
  geom_boxplot() +
  labs(title = "Boxplot of Prices by Model and Location", x = "Model", y = "Price") +
  facet_wrap(~ location)
```




```
# 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")
```

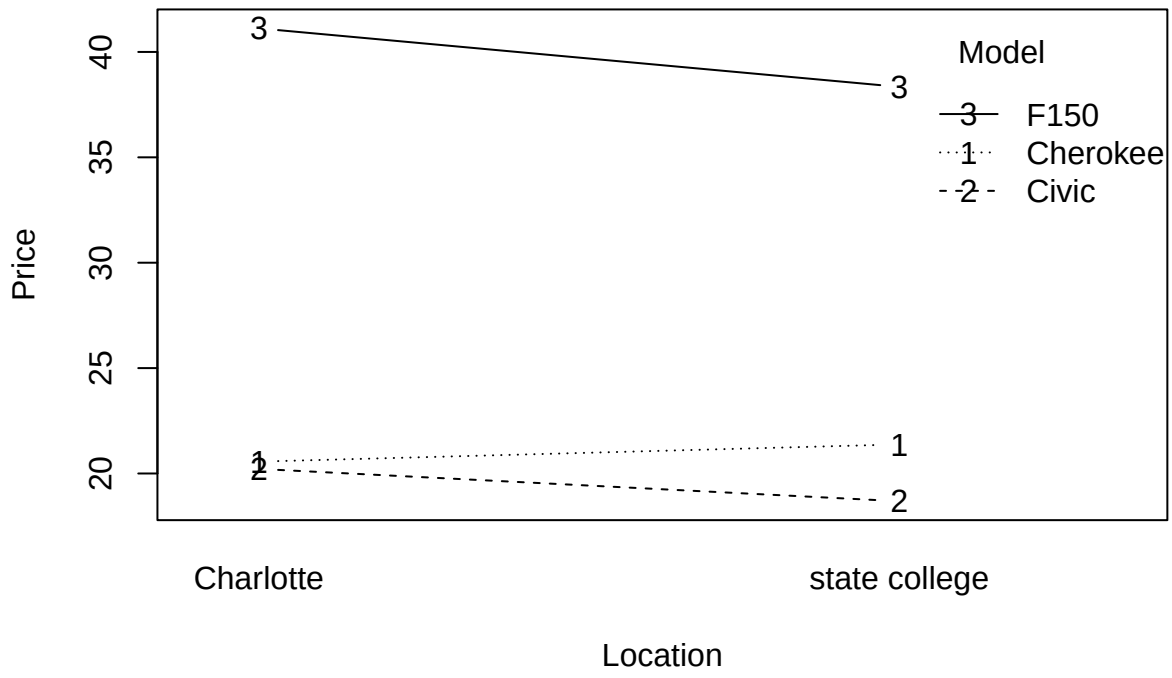


Image 1: Boxplot of Prices by Location

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.

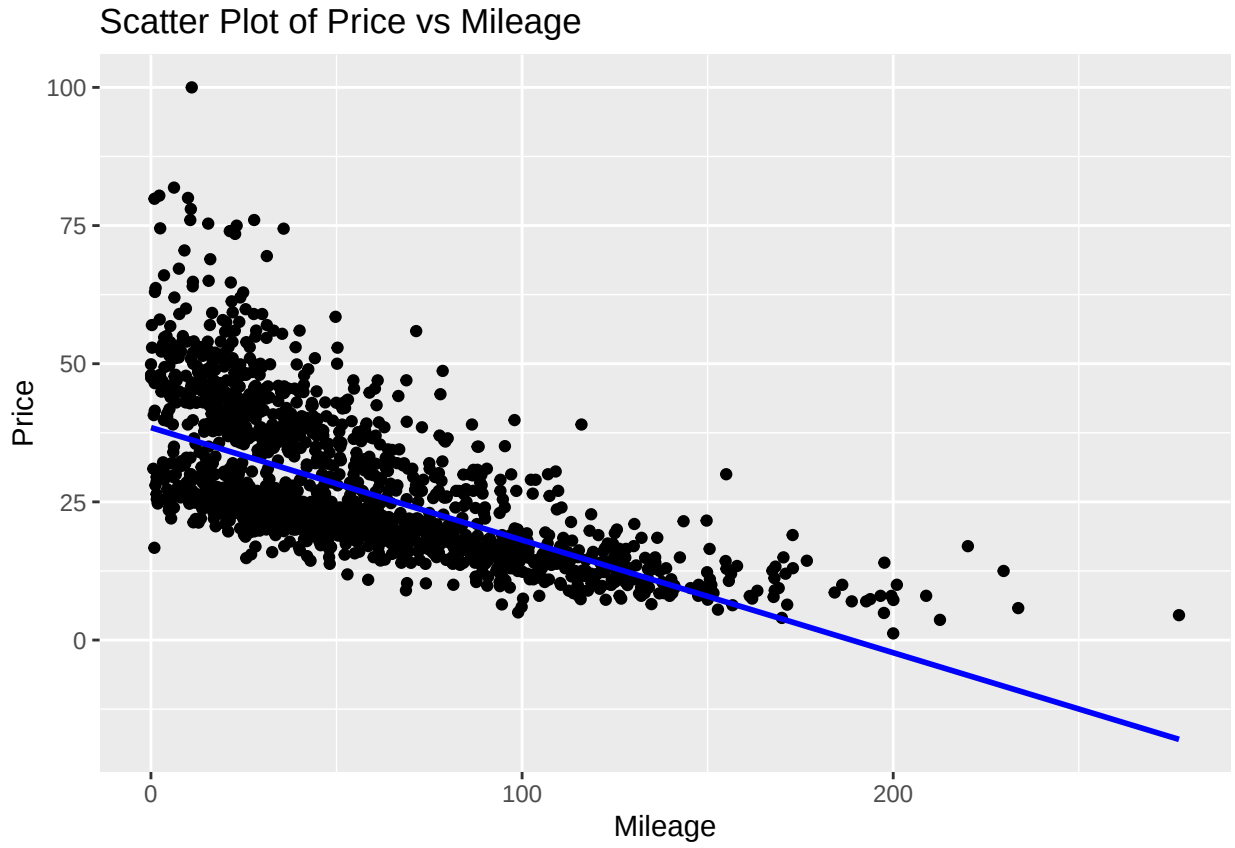
Image 5: Line Plot of Prices by Location and Model

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

```
# Plot mileage vs. price with a regression line to check for linearity
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'
```



The plot shows random dispersion of residuals around the hornizontal line, indicating that the assumption of indepencdence met.

```
#Two-Way Factorial Model with Interaction
anova_model_1 <- aov(lm(price ~ location+model+location * model, data = combined_data))
summary(anova_model_1)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## location           1    208     208   2.537 0.11139
## model              2 147551   73775 898.604 < 2e-16 ***
## location:model     2    970     485   5.905 0.00278 **
## Residuals       1696 139242      82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. **Location:**
    - The location has 1 degree of freedom (Df) indicating it's a comparison between two groups (e.g., two different locations).
    - The sum of squares (Sum Sq) for location is 208, and the mean square (Mean Sq), which is the Sum Sq divided by the Df, is also 208.
    - The F value is 2.537, which is the ratio of the Mean Sq of location to the Mean Sq of the residuals.
    - The p-value for location is 0.11139, which is greater than 0.05, suggesting that there is no statistically significant difference in car prices between the two locations at the conventional alpha level of 0.05.
2. **Model:**
    - The model has 2 degrees of freedom, indicating three different car models are being compared.
    - The Sum Sq for model is 147,551, with a Mean Sq of 73,775, which is quite large relative to the Sum Sq for location.
    - The F value for the model is 898.604, which is highly significant with a p-value less than 2e-16 (practically zero). This indicates a very strong effect of the car model on price; there are statistically significant differences in car prices among the three models.
3. **Location:Model (Interaction):**
    - The interaction term has 2 degrees of freedom, which aligns with the number of interaction comparisons available in a 2 x 3 setup (two locations times three models).
    - The Sum Sq for the interaction is 970, with a Mean Sq of 485.
    - The F value for the interaction is 5.905, which has a p-value of 0.00278. This is less than 0.01, indicating that there is a statistically significant interaction effect between location and model on car prices. This means that the effect of the car model on price is different across locations.

      The ANOVA results indicated that the main effect of location on car prices was not statistically significant ($p > 0.05$), leading us to retain the null hypothesis (H0a) that there is no significant difference in car prices between the locations tested. However, the car model had a highly significant effect on price ($p < 0.001$), so we reject the null hypothesis (H0b) and accept the alternative hypothesis (H1b) that at least one model's mean price is different from the others. Furthermore, there was a statistically significant interaction between location and model ($p < 0.01$), leading us to reject the null hypothesis (H0c) and accept the alternative hypothesis (H1c) that the effect of car model on price varies by location.

```r
#ANCOVA
anova_model_2 <- aov(lm(price ~ model+mileage, data = combined_data))
summary(anova_model_2)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## model           2 147173   73586    1833 <2e-16 ***
## mileage         1  72619   72619    1809 <2e-16 ***
```

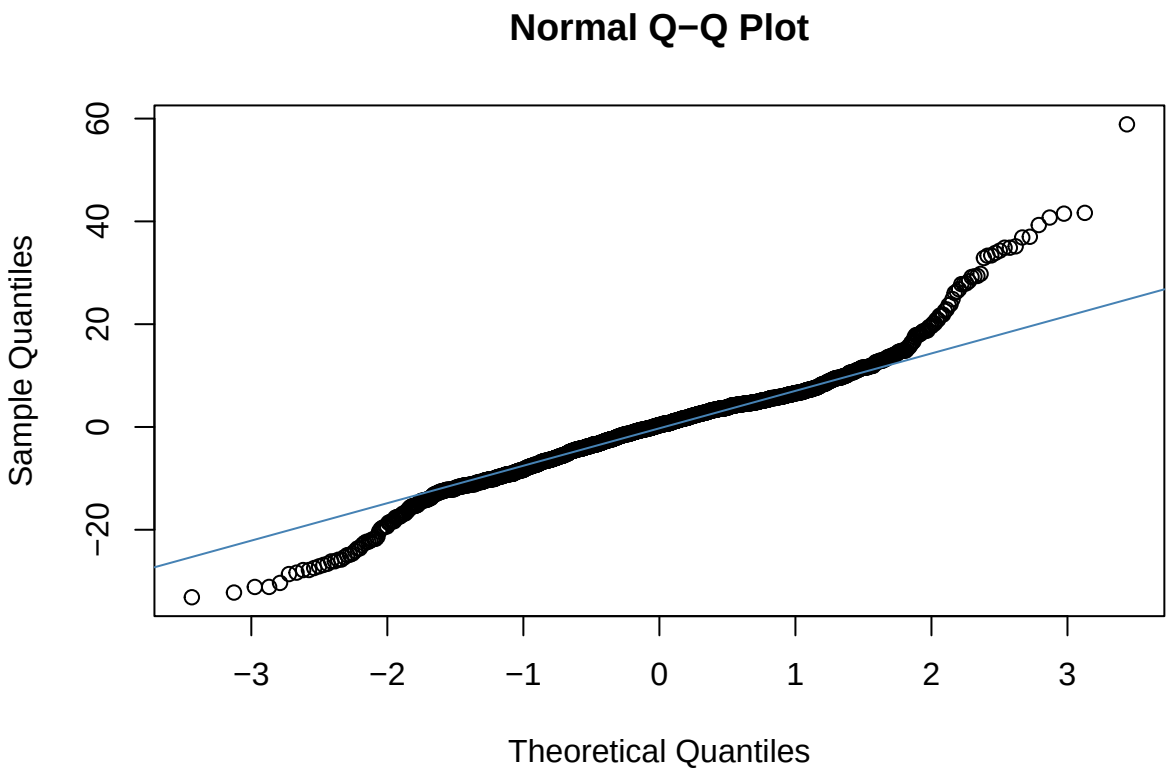
```
## Residuals   1698  68178      40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. **Model:**
    - **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 147,173, 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 73,586. This represents the average variation per model category.
    - **F-value:** The F-value is 1833, which is substantially large, indicating a strong effect of the model on price.
    - **p-value (Pr(>F)):** The p-value is less than 2e-16 (indicating a value very close to zero), which is highly significant. This means that there is a statistically significant difference in car prices among the different models.
2. **Mileage:**
    - **Degrees of Freedom (Df):** Mileage has 1 degree of freedom, as it is a continuous predictor.
    - **Sum of Squares (Sum Sq):** The Sum Sq for mileage is 72,619, indicating the variation in car prices due to mileage.
    - **Mean Square (Mean Sq):** The Mean Sq is also 72,619 since there is only one mileage variable.
    - **F-value:** The F-value for mileage is 1809, which is also quite large and indicative of a strong effect.
    - **p-value (Pr(>F)):** Like the model, the p-value for mileage is less than 2e-16, suggesting a very strong statistical significance. This implies that mileage has a significant effect on car prices.

    The ANCOVA results demonstrated that both the car model and mileage significantly affected car prices ($p < 0.001$ for both factors). Thus, we reject both null hypotheses (H0d and H0e) and accept the alternative hypotheses (H1d and H1e), concluding that different car models significantly affect car prices, and mileage has a significant influence on car prices when adjusting for the model.

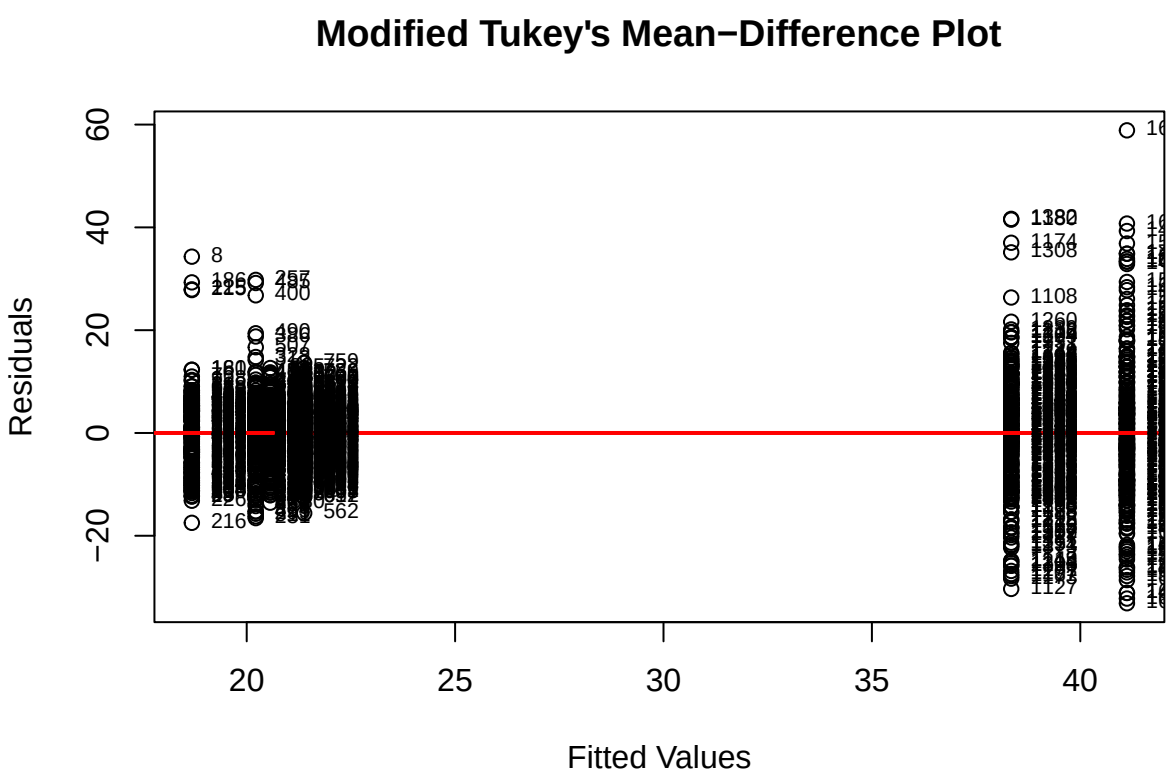
Assess Model Conditions

```r
# Fit the ANOVA model
anova_model_1 <- aov(price ~ location + model + location:model, data = combined_data)

# Plotting Q-Q plot of residuals to check for normality
qqnorm(residuals(anova_model_1))
qqline(residuals(anova_model_1), col = "steelblue")
```



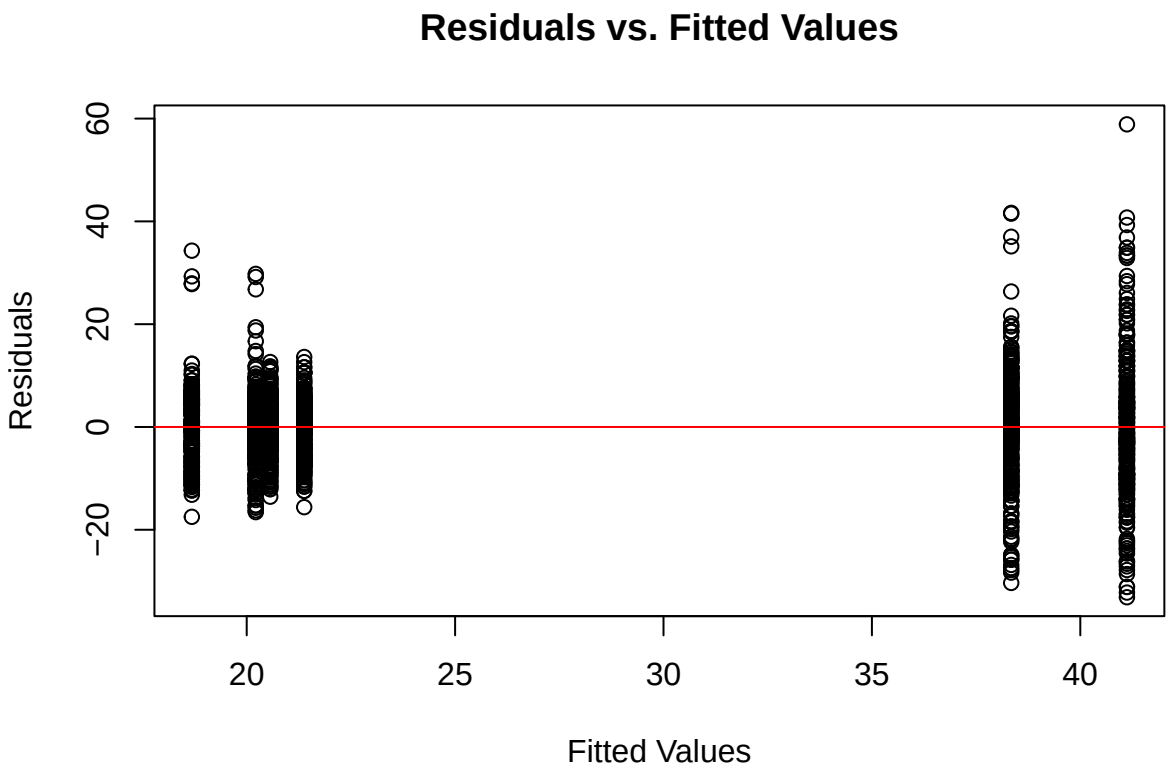
The qq plot shows some deviation from normailty particularly in the tails, but the data is approximately normal.

```r
# Fit the ANOVA model

# Calculate residuals and fitted values
residuals <- residuals(anova_model_1)
fitted_values <- fitted(anova_model_1)

# Create a Tukey's Mean-Difference Plot (residuals vs. fitted values)
plot(fitted_values, residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Modified Tukey's Mean-Difference Plot")

# Adding a horizontal line at zero to help assess even spread
abline(h = 0, col = "red", lwd = 2)


# Adding labels if needed
text(fitted_values, residuals, labels = row.names(combined_data), cex = 0.7, pos = 4)
```



The spread of residuals seems consistent across different fitted values, although some outliers are evident. There doesn't appear to be a funnel-shaped pattern, which suggests that homoscedasticity is reasonable.

```
# Plot residuals against fitted values to look for patterns
plot(fitted(anova_model_1), residuals(anova_model_1),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values")
abline(h = 0, col = "red")
```



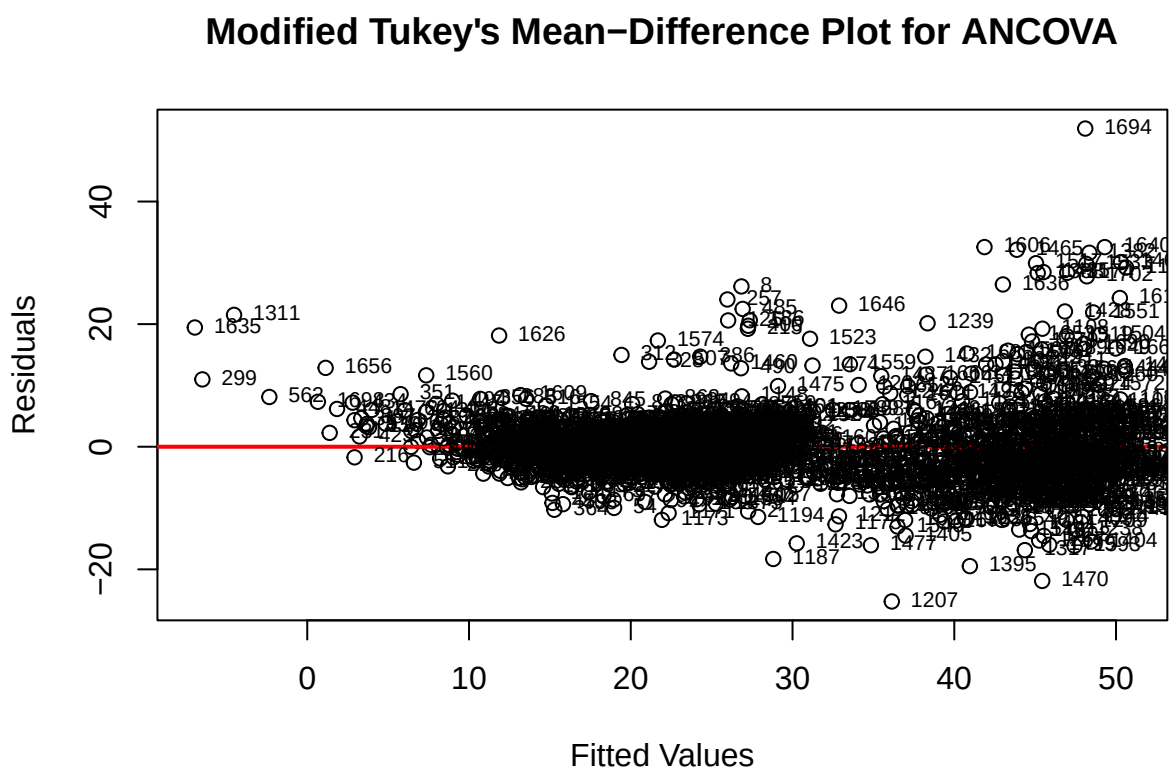
The plot for residuals against shows a random dispersion of residuals around the horizontal line, suggesting that the assumption of independence is met.

```
# Fit model with interaction
ancova_model_check <- lm(price ~ model * mileage, data = combined_data)

# Use ANOVA to test if interaction terms are significant
anova_interaction_check <- anova(ancova_model_check)
print(anova_interaction_check)
```

```
## Analysis of Variance Table
##
## Response: price
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## model             2 147173   73586 2104.56 < 2.2e-16 ***
## mileage           1  72619   72619 2076.89 < 2.2e-16 ***
## model:mileage     2   8877    4439  126.95 < 2.2e-16 ***
## Residuals      1696  59301      35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows that both the car model and mileage, as well as their interaction, are highly significant predictors of price ($p < 2.2e\text{-}16$), indicating that,the model of the car (model) has a strong effect on the car's price., mileage (mileage) also significantly affects the car's price, the interaction between car model and mileage (model:mileage) is significant, suggesting that the effect of mileage on price is not consistent across different car models.

```r
# Fit the ANCOVA model
ancova_model_2 <- lm(price ~ model + mileage + model:mileage, data = combined_data)

# Calculate residuals and fitted values
residuals_ancova <- residuals(ancova_model_2)
fitted_values_ancova <- fitted(ancova_model_2)

# Create a Tukey's Mean-Difference Plot (residuals vs. fitted values)
plot(fitted_values_ancova, residuals_ancova,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Modified Tukey's Mean-Difference Plot for ANCOVA")

# Adding a horizontal line at zero to help assess even spread
abline(h = 0, col = "red", lwd = 2)


text(fitted_values_ancova, residuals_ancova, labels = row.names(combined_data), cex = 0.7, pos = 4)
```



The plot also does not indicate obvious signs of changing variance across the range of fitted values. Outliers are present however, which may be worth investigating further.

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.