Bill Le 47590 & Thanonon Choundom 47573

Introduction and Literature Review

Automobiles, approximately 1.47 billion of them worldwide, are shaped by the waves of innovation and traditional craftsmanship, presenting a dynamic interaction between car manufacturing and cutting-edge automotive technology (stumpf) . When it comes to purchasing one, price is just one of the many factors to consider. It plays a crucial role in our lives, from the moment we complete our drivers exam, to the moment when we use it everyday for convenience. However, consumer preferences are rapidly evolving and the rising economic inflation, investment decisions become more difficult as car parts fluctuate in price and quality. Taking into consideration, the significance of research into car quality and pricing cannot be overstated. With the agreement between my partner and I, we decided to research what affects car prices and why some cars are considered better or more valuable to others. Specifically, we are curious about the power of the car, what kind of body style it possesses, and how these factors change the game when it comes to price tags. It has been a hot topic for a while now, and we have decided to research the trend behind it. Researching this niche has the possibility of assisting car enthusiasts to understand the best combination of car parts that will be worth the investment.

Therefore, as analysts, our study adopts an Analysis of Covariance (ANCOVA) - moderated regression analysis approach, to examine the moderating effect of car body types on the relationship between horsepower and price. While there was previous research regarding how car parts influence price, there remains a gap in comprehensively understanding how these components interact. Our aim is to close the gap, providing a closer analysis into how moderating a covariate might change the relationship between an independent and dependent variable.

We are confident that with our investigation into the complex dynamics shaping car valuations will unveil insights for relevant stakeholders within and beyond the automotive industry. This includes manufacturers and dealers who work with craftsmanship and sales strategies, as well as buyers who want to make more informed decisions that suit their needs.

Methodology

Objective:

The primary objective of our research is to employ our knowledge of R and analyze the relationship between horsepower and car prices while moderating the body types of cars. We will analyze our dataset sourced from Kaggle.com in R studio, which provides a rich base for analysis. We will employ analytical strategies that fit best to our project. As mentioned above, using a moderate regression from ANCOVA, we still specifically assess the relationship between a car's horsepower and its market price, taking into account how different body types might influence this relationship. For example, the relationship between the horsepower and the car's price in a sedan body type, might not be as significant as if it has a hatchback body type. Using R studio's visualization capabilities to create scatter plots and heat graphs are also crucial, to graphically depict a clear representation of our findings. By conducting this analysis, we aim to solve how various features and designs contribute to the overall value of a vehicle, which provides a deeper understanding of the automotive market dynamics for both buyers and sellers.

Research Question:

"How does horsepower influence the price of a car, and how does this relationship vary among different car body types?"

Data Source:

The data source we found is on Kaggle.com

Dependent Variable:

- Car Price

Independent Variable:

- Horsepower
- Car body types (covariate - moderated)

Hypothesis:

Higher horsepower is associated with higher car prices, and this association is moderated by the car's body type. We expect that the body types such as sedans and hatchbacks will show a stronger relationship between horsepower and price compared to others.

Exploratory Data Analysis

Data Acquisition

The dataset that we found on Kaggle through extensive research consisted of 26 columns, and 205 rows. We imported our dataset into R studio, setting working directory and importing all necessary libraries (more imported libraries are shown later). Finally, we created a car_data variable to store and read our CSV file from the pathname.

```
setwd("/Users/le.bill/Desktop/uni /statistics")
library(readr)
library(dplyr)
library(tidyr)
library(MASS)
car_data <- read.csv("/Users/le.bill/Desktop/uni /statistics/scrap price.csv")

```

We made a checklist of what was necessary to be achieved in the EDA, shown below:

- [x] ~~Formulation of The Question~~
- [x] ~~Check the Packaging~~
- [x] ~~Assess general statistics~~
- [x] ~~EDA techniques~~
- [x] ~~Validate with external sources~~
- [x] ~~Trying different solutions~~
- [x] ~~Questions and Analyses~~

Package checking

With the library we added on R studio (dplyr), we first checked the string and the top and bottom of the dataset.

str(car_data)

```
'data.frame':	205 obs. of  27 variables:
 $ ID              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
 $ name            : chr  "alfa-romero giulia" "alfa-romero stelvio" "alfa-romero Quadrifoglio" "audi 100 ls" ...
 $ fueltypes       : chr  "gas" "gas" "gas" "gas" ...
 $ aspiration      : chr  "std" "std" "std" "std" ...
 $ doornumbers     : chr  "two" "two" "two" "four" ...
 $ carbody         : chr  "convertible" "convertible" "hatchback" "sedan" ...
 $ drivewheels     : chr  "rwd" "rwd" "rwd" "fwd" ...
 $ enginelocation  : chr  "front" "front" "front" "front" ...
 $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ carlength       : num  169 169 171 177 177 ...
 $ carwidth        : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
 $ carheight       : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
 $ curbweight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
 $ enginetype      : chr  "dohc" "dohc" "ohcv" "ohc" ...
 $ cylindernumber  : chr  "four" "four" "six" "four" ...
 $ enginesize      : int  130 130 152 109 136 136 136 136 131 131 ...
 $ fuelsystem      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
 $ boreratio       : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.13 ...
 $ stroke          : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
 $ compressionratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
 $ horsepower      : int  111 111 154 102 115 110 110 110 140 160 ...
 $ peakrpm         : int  5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
 $ citympg         : int  21 21 19 24 18 19 19 19 17 16 ...
 $ highwaympg      : int  27 27 26 30 22 25 25 25 20 22 ...
 $ price           : num  13495 16500 16500 13950 17450 ...
 $ price_transformed: num  3181 3758 3758 3270 3936 ...
```

head(car_data)

```
> head(car_data)
  ID symboling                     name fueltypes aspiration doornumbers     carbody drivewheels enginelocation wheelbase carlength carwidth
1  1         3       alfa-romero giulia       gas        std         two convertible         rwd          front      88.6     168.8     64.1
2  2         3      alfa-romero stelvio       gas        std         two convertible         rwd          front      88.6     168.8     64.1
3  3         1 alfa-romero Quadrifoglio       gas        std         two   hatchback         rwd          front      94.5     171.2     65.5
4  4         2              audi 100 ls       gas        std        four       sedan         fwd          front      99.8     176.6     66.2
5  5         2               audi 100ls       gas        std        four       sedan         4wd          front      99.4     176.6     66.4
6  6         2                 audi fox       gas        std         two       sedan         fwd          front      99.8     177.3     66.3
  carheight curbweight enginetype cylindernumber enginesize fuelsystem boreratio stroke compressionratio horsepower peakrpm citympg highwaympg
1      48.8       2548       dohc           four        130       mpfi      3.47   2.68              9.0        111    5000      21         27
2      48.8       2548       dohc           four        130       mpfi      3.47   2.68              9.0        111    5000      21         27
3      52.4       2823       ohcv            six        152       mpfi      2.68   3.47              9.0        154    5000      19         26
4      54.3       2337        ohc           four        109       mpfi      3.19   3.40             10.0        102    5500      24         30
5      54.3       2824        ohc           five        136       mpfi      3.19   3.40              8.0        115    5500      18         22
6      53.1       2507        ohc           five        136       mpfi      3.19   3.40              8.5        110    5500      19         25
  price price_transformed
1 13495          3181.324
2 16500          3757.953
3 16500          3757.953
4 13950          3269.947
5 17450          3936.357
6 15250          3520.489
```

tail(car_data)

```
> tail(car_data)
     ID symboling            name fueltypes aspiration doornumbers carbody drivewheels enginelocation wheelbase carlength carwidth carheight
200 200        -1    volvo diesel       gas      turbo        four   wagon         rwd          front     104.3     188.8     67.2      57.5
201 201        -1 volvo 145e (sw)       gas        std        four   sedan         rwd          front     109.1     188.8     68.9      55.5
202 202        -1     volvo 144ea       gas      turbo        four   sedan         rwd          front     109.1     188.8     68.8      55.5
203 203        -1     volvo 244dl       gas        std        four   sedan         rwd          front     109.1     188.8     68.9      55.5
204 204        -1       volvo 246    diesel      turbo        four   sedan         rwd          front     109.1     188.8     68.9      55.5
205 205        -1     volvo 264gl       gas      turbo        four   sedan         rwd          front     109.1     188.8     68.9      55.5
    curbweight enginetype cylindernumber enginesize fuelsystem boreratio stroke compressionratio horsepower peakrpm citympg highwaympg price
200       3157        ohc           four        130       mpfi      3.62   3.15              7.5        162    5100      17         22 18950
201       2952        ohc           four        141       mpfi      3.78   3.15              9.5        114    5400      23         28 16845
202       3049        ohc           four        141       mpfi      3.78   3.15              8.7        160    5300      19         25 19045
203       3012       ohcv            six        173       mpfi      3.58   2.87              8.8        134    5500      18         23 21485
204       3217        ohc            six        145        idi      3.01   3.40             23.0        106    4800      26         27 22470
205       3062        ohc           four        141       mpfi      3.78   3.15              9.5        114    5400      19         25 22625
    price_transformed
200          4214.705
201          3822.941
202          4232.204
203          4676.733
204          4853.683
205          4881.406
```

General Statistical Analysis

We checked for missing values as well as to check the number of unique values there are. The dataset showed no missing values, which is highly convenient since we do not need to manually remove these values off of the dataset. There was a range of different unique values across all variables, shown below.

```
> sum(is.na(car_data))
[1] 0
> #check for unique values
> sapply(car_data, function(x) length(unique(x)))
               ID         symboling              name         fueltypes        aspiration       doornumbers           carbody       drivewheels
              205                 6               147                 2                 2                 2                 5                 3
   enginelocation         wheelbase         carlength          carwidth         carheight        curbweight        enginetype    cylindernumber
                2                53                75                44                49               171                 7                 7
       enginesize        fuelsystem         boreratio            stroke  compressionratio        horsepower           peakrpm           citympg
               44                 8                38                37                32                59                23                29
       highwaympg             price price_transformed
               30               189               189
```

Descriptive Statistics

We summarized the overall dataset first using *summary(car_data),* and found the results to show a general right skew among most of the numerical variables, a few being symmetrical, and one having a slight left skew. We believe that it is common in automotive data, as consumers typically purchase lower-end cars with smaller engines, less features, and lower prices, which can skew the distribution. Specifically, looking at the numerical variable *price* and *horsepower,* both visibly present a right skew.

```
> summary(car_data)
       ID        symboling           name            fueltypes          aspiration        doornumbers          carbody
 Min.   :  1   Min.   :-2.0000   Length:205         Length:205         Length:205         Length:205         Length:205
 1st Qu.: 52   1st Qu.: 0.0000   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :103   Median : 1.0000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :103   Mean   : 0.8341
 3rd Qu.:154   3rd Qu.: 2.0000
 Max.   :205   Max.   : 3.0000
 drivewheels        enginelocation        wheelbase        carlength        carwidth       carheight       curbweight     enginetype
 Length:205         Length:205         Min.   : 86.60   Min.   :141.1   Min.   :60.30   Min.   :47.80   Min.   :1488   Length:205
 Class :character   Class :character   1st Qu.: 94.50   1st Qu.:166.3   1st Qu.:64.10   1st Qu.:52.00   1st Qu.:2145   Class :character
 Mode  :character   Mode  :character   Median : 97.00   Median :173.2   Median :65.50   Median :54.10   Median :2414   Mode  :character
                                       Mean   : 98.76   Mean   :174.0   Mean   :65.91   Mean   :53.72   Mean   :2556
                                       3rd Qu.:102.40   3rd Qu.:183.1   3rd Qu.:66.90   3rd Qu.:55.50   3rd Qu.:2935
                                       Max.   :120.90   Max.   :208.1   Max.   :72.30   Max.   :59.80   Max.   :4066
 cylindernumber       enginesize      fuelsystem          boreratio        stroke      compressionratio   horsepower        peakrpm
 Length:205         Min.   : 61.0   Length:205         Min.   :2.54   Min.   :2.070   Min.   : 7.00    Min.   : 48.0   Min.   :4150
 Class :character   1st Qu.: 97.0   Class :character   1st Qu.:3.15   1st Qu.:3.110   1st Qu.: 8.60    1st Qu.: 70.0   1st Qu.:4800
 Mode  :character   Median :120.0   Mode  :character   Median :3.31   Median :3.290   Median : 9.00    Median : 95.0   Median :5200
                    Mean   :126.9                      Mean   :3.33   Mean   :3.255   Mean   :10.14    Mean   :104.1   Mean   :5125
                    3rd Qu.:141.0                      3rd Qu.:3.58   3rd Qu.:3.410   3rd Qu.: 9.40    3rd Qu.:116.0   3rd Qu.:5500
                    Max.   :326.0                      Max.   :3.94   Max.   :4.170   Max.   :23.00    Max.   :288.0   Max.   :6600
    citympg        highwaympg        price
 Min.   :13.00   Min.   :16.00   Min.   : 5118
 1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7788
 Median :24.00   Median :30.00   Median :10295
 Mean   :25.22   Mean   :30.75   Mean   :13277
 3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16503
 Max.   :49.00   Max.   :54.00   Max.   :45400
```

```
> desc_stats_select <- summary(car_data[c("price", "horsepower")])
> desc_stats_select
     price         horsepower
 Min.   : 5118   Min.   : 48.0
 1st Qu.: 7788   1st Qu.: 70.0
 Median :10295   Median : 95.0
 Mean   :13277   Mean   :104.1
 3rd Qu.:16503   3rd Qu.:116.0
 Max.   :45400   Max.   :288.0
```

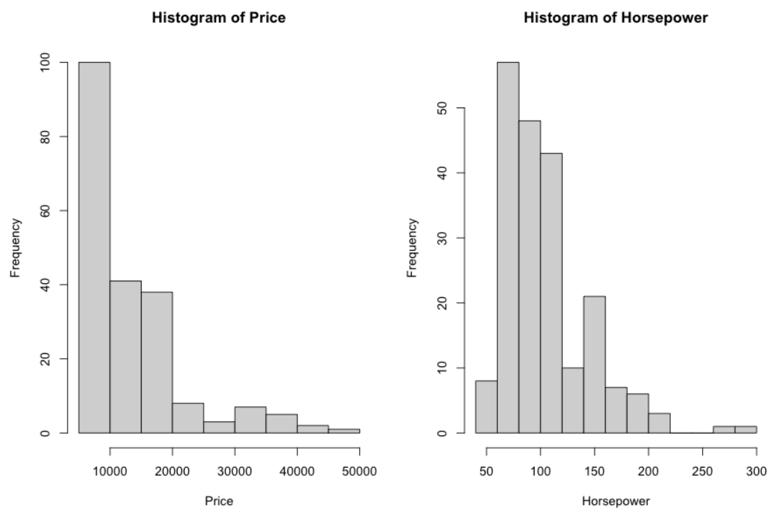
As you can see, the minimum is 5118, the median of the price is 10,295, with a maximum of 45400. Similar to horsepower, the median of 95 with a maximum of 288, while only having a minimum of 48.

Visualization of Graphs and Boxplots

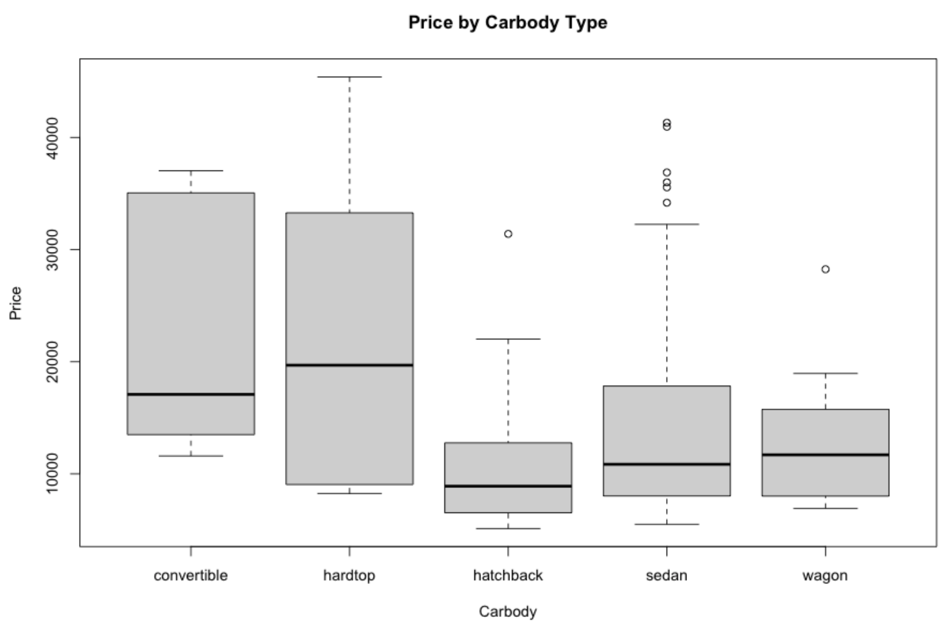
Next, we visualized the data points to really grasp the distribution across the selected variables for the analysis. Using the library ggplot2 and tidyr, we graphed plots, box plots, histograms and barplots to examine their distribution and the relationship between horsepower and price, as well as their relationship for each car body category

```
par(mfrow = c(1, 2))
# Visualizing distribution of price variable
hist(car_data$price, main="Histogram of Price", xlab="Price")
# Visualizing distribution of horsepower variable
hist(car_data$horsepower, main="Histogram of Horsepower", xlab="Horsepower")
```

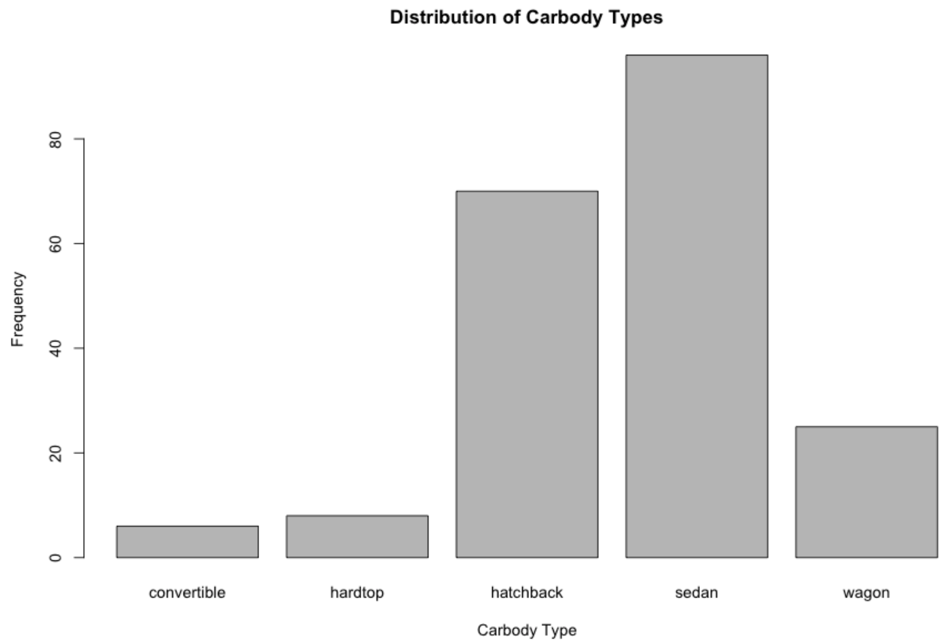
Histograms of price and horsepower (g-1)



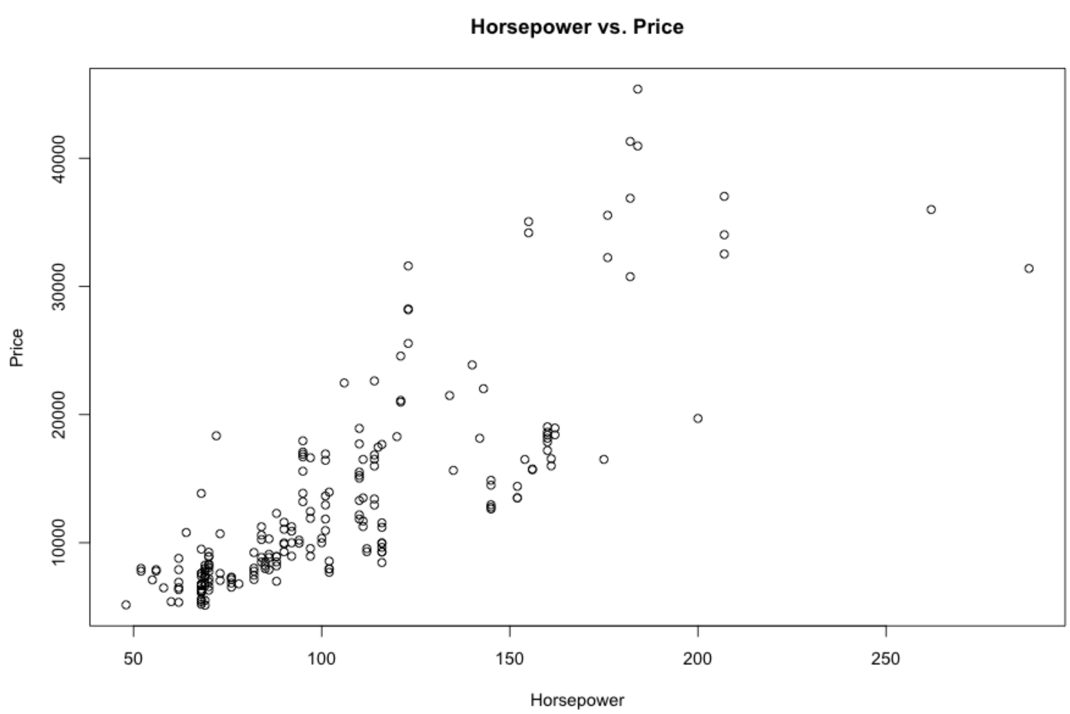
Box & Whisker plot; Price by Car Body Type (g-2)



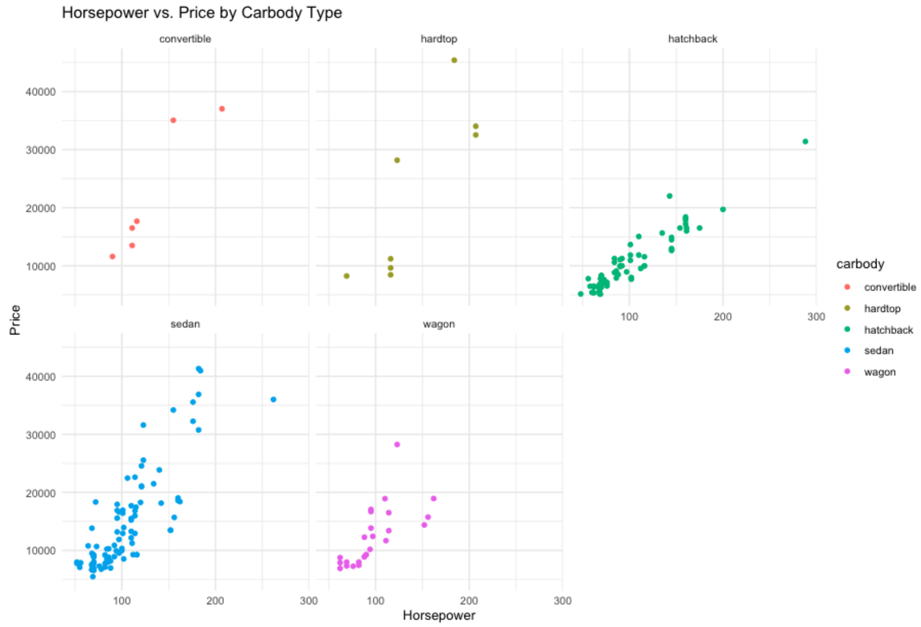
Barplot of distribution of Car Body Types (g-3)



Scatter Plot of Horsepower Vs. Price (g-4)



Scatterplot for Horsepower Vs. Price by Car body Type (g-5)



G-1: Graph 1 depicts a right skewness on both histograms, suggesting that there are a number of higher-than-average (or median) values pulling the mean to the right.

G-2: Graph 2 displays a box & whisker plot for price by car body types, suggesting that car prices vary significantly across different car body types, with convertibles and hardtops generally priced higher than hatchbacks, sedans, and wagons.

G-3: Graph 3 displays the frequencies of each car body type, indicating that hatchbacks and sedans are most frequent.

G-4: Graph 4 displays a scatterplot for Horsepower vs Price, indicating a linear relationship (more HP = higher price)

G-5: Graph 5 displays HP vs price by car body type, which indicates positive linear relationship on all, as well as showing the frequencies of each type (more dots = more cars with that body type)

Creating a Correlation matrix

The next most important analysis is identifying whether there is a strong correlation between the two numerical variables; horsepower and price.

```
library(corrplot)
par(mfrow=c(1, 1))

numerical_data <- data[,sapply(data, is.numeric)]  # Select only numerical columns
cor_matrix <- cor(numerical_data, use="complete.obs")# Compute correlation matrix
corrplot(cor_matrix, method="circle")

price_correlations <- cor_matrix['price',]
threshold <- 0.5
significant_variables <- names(price_correlations[abs(price_correlations) > threshold & names(price_correlations) != "price"])
significant_correlations <- price_correlations[significant_variables]
print(significant_correlations)
```

The code above first selects only numeric columns from *car_data*, then creates the correlation matrix, and visualizes the correlation using method=circle. It then identifies significant correlations with *price* by setting a threshold of 0.5. I chose a threshold of 0.5 because according to Cohen's conventional guidelines on interpreting the effect size of statistical findings and correlation coefficients, he argued there were 3 effect sizes (small ~0.1-0.29, medium ~ 0.3-49, large ~ 0.5-1.0), and with these guidelines it helped countless of researchers and practitioners contextualize the strength of an observation given data. Because of its reliability, we used his guideline to depict our significant variables. The code then prints the significant correlations along with their corresponding variables. It helps to identify the strongest correlated variables.

```
> print(significant_correlations)
 wheelbase  carlength   carwidth curbweight enginesize  boreratio horsepower    citympg highwaympg
 0.5778156  0.6829200  0.7593253  0.8353049  0.8741448  0.5531732  0.8081388 -0.6857513 -0.6975991
```

As shown, the correlation between *horsepower* and *price* is approx 0.808, which indicates strong correlation, ensuring reliability and accuracy of future models.

Model Selection and Creation

Fitting and creating the model

Before creating our moderated regression model, we first asked ourselves, *"are all of the elements within the car body types necessary"*, because as shown on page 8 graph 5, we can see that most of the automobiles were hatchbacks, sedan's, and wagon's, and the others did not seem to have enough data for an accurate model. Therefore we created another model subset, one subset labeling all the body types, and the second subset only selecting the 3 types.

```
model_subset <- car_data[car_data$carbody %in% c("hatchback", "sedan", "wagon"), ]
```

```
model_data <- lm(price ~ horsepower * carbody, data = model_subset)
```

In order to perform a moderated regression analysis, we first created a linear regression model to predict the *price* of cars based on *horsepower* and *carbody* type. The ~ symbol separates the DV from the predictors, and the * symbol indicates that both predictions and their interaction are included in the model. We used car_data for the data of this regression model.

Meeting Assumptions for Moderated Regression

Meeting the assumptions of a statistical analysis such as moderated regression ANCOVA analysis is essential to check before creating the final model. By ensuring that the assumptions are met, the statistical methods will accurately reflect the data relationships without unbiased estimates and incorrect conclusions about the relationships. It also results in easier interpretation of these tests.

We have created a checklist for assumptions that need to be met for a successful analysis:

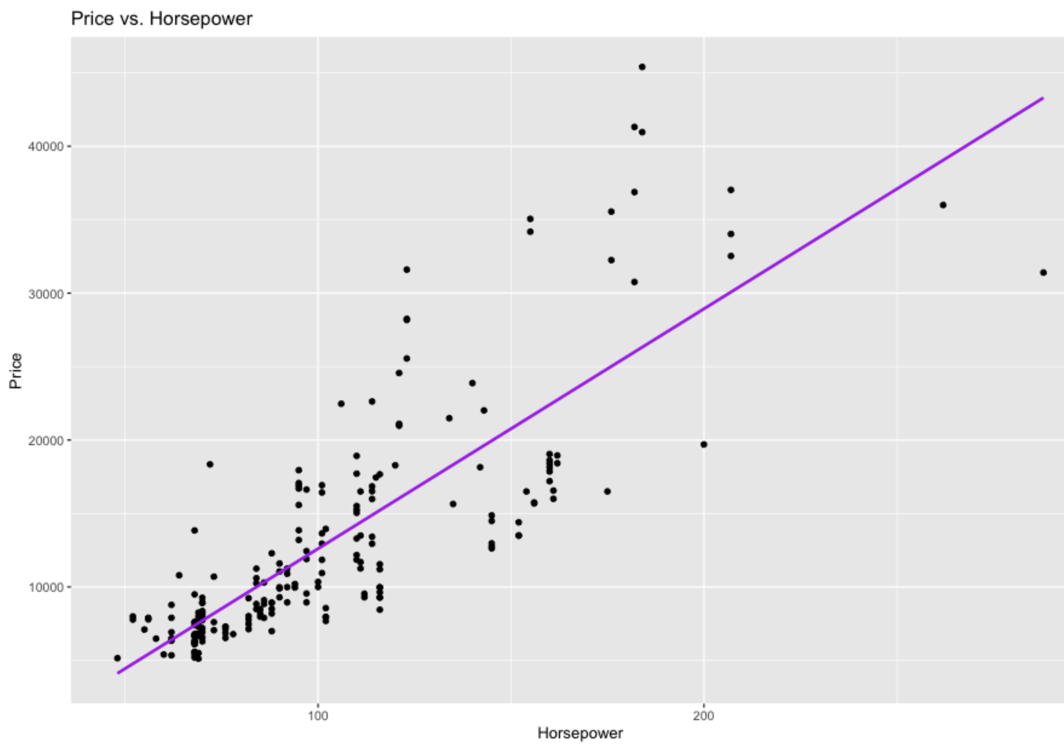
- ☑ ~~Linearity~~
- ☑ ~~Independence of Residuals~~
- ☑ ~~Normality of Residuals~~
- ☑ ~~Homoscedasticity~~
- ☐ No perfect Multicollinearity
- ☑ *~~Checking Homogeneity for regression slopes~~*

Checking Linearity

Checking the relationship between the DV, IV, and covariates, to make sure they are linear. We created a plot with ggplot that shows a clear upward trend, indicating that cars with more horsepower tend to be more expensive, confirming the linear relationship between these variables.

```
ggplot(car_data, aes(x = horsepower, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Price vs. Horsepower", x = "Horsepower", y = "Price")
```

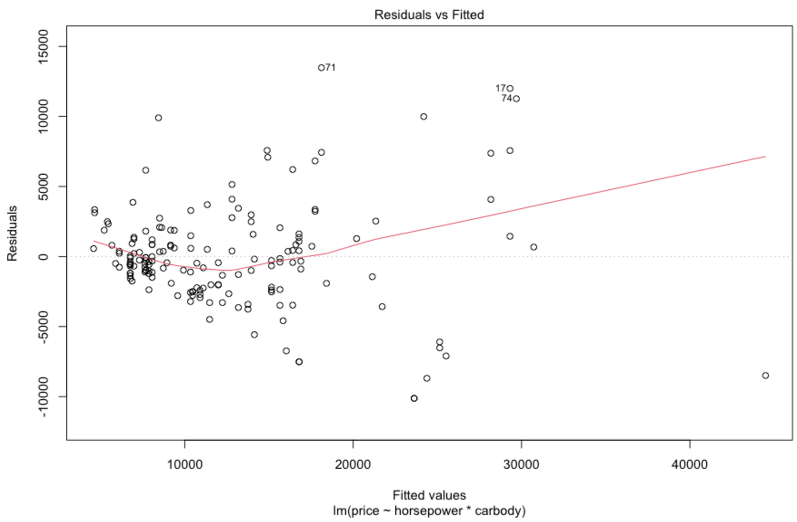
Scatterplot showing linearity



Checking Independence of Residuals

The assumption of independence of residuals states that errors in a statistical model are unrelated to each other. We check this assumption to ensure the validity of statistical tests. Violations can lead to biased results and incorrect conclusions.

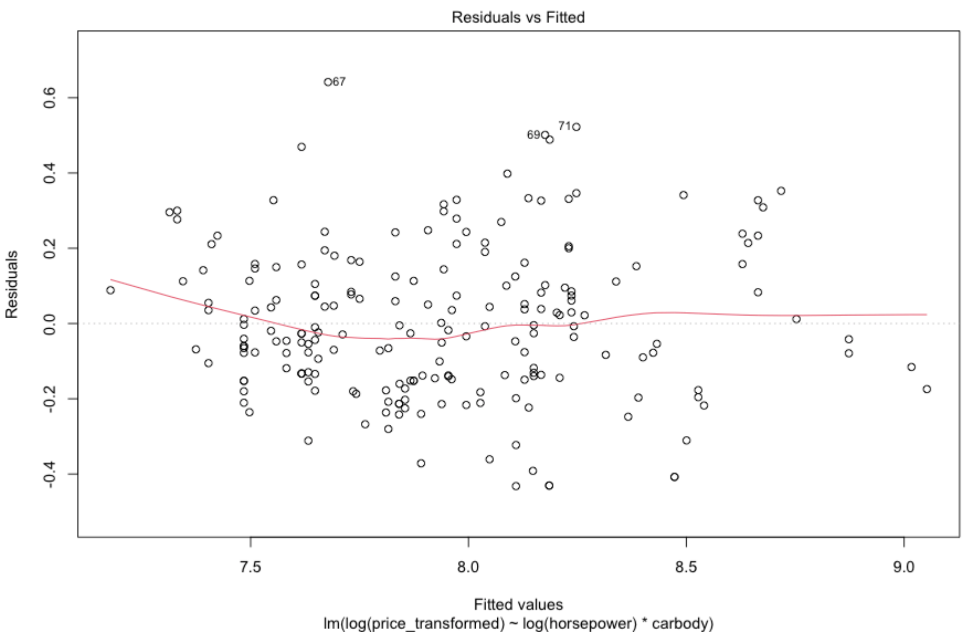
```
plot(model_data, which = 1)
```



The plot above shows the residuals vs fitted values from the model, basically the difference between observed and predicted prices vs predicted prices. Independence of Residuals are met if the scattered points should form a random pattern around the horizontal line at 0. The red line indicates the overall trend of the residuals. While there is no visible pattern, the scattered dots are more towards one side, resulting in the red line to be slightly angled, which could indicate possible issues like non-linearity.

We counteracted this problem by performing multiple steps: We applied a logarithmic transformation to our IV and DV variables, which helps to linearize the relationship. Furthermore, we also conducted a box cox transformation to stabilize variance and further normalized the data.

```
c <- boxcox(model)
lambda <- bc$x[which.max(bc$y)]
car_data$price_transformed <- (car_data$price^lambda - 1) / lambda
model_bc <- lm(log(price_transformed) ~ log(horsepower) * carbody, data = model_subset)
plot(model_bc_clean, which = 1)
```

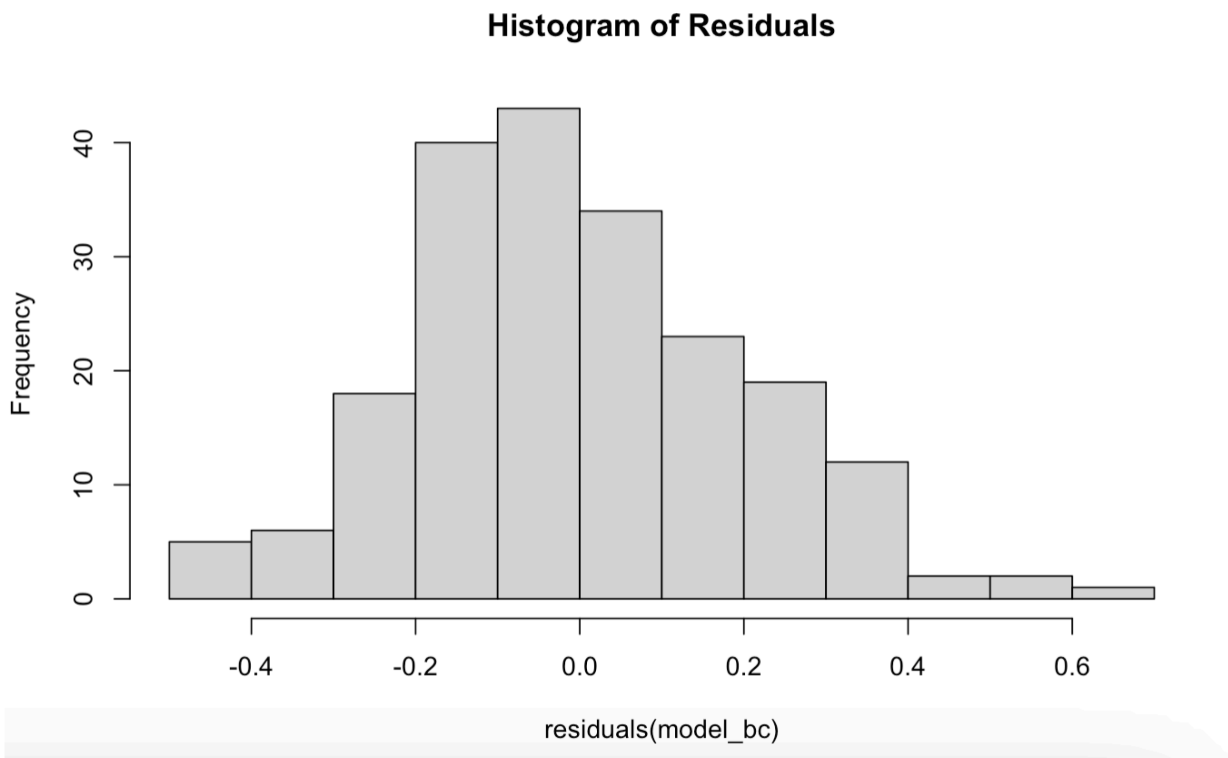


With the transformed logarithmic model, we can see that the points are more scattered and the line is more horizontal, showing that it suggests homoscedasticity, meeting the assumption.

Checking Normality of Residuals

The assumption of normality of residuals typically assumes that the residuals from a regression model are normally distributed. Checking the assumption ensures the validity of the model's p-values and confidence intervals. I first created a histogram to visualize the normality of residuals

```
#checking for normality of residuals
hist(residuals(model_bc), breaks = "FD", main = "Histogram of Residuals")
```



As shown above, we see that the data follows a slight right skew, which shows a degree of normality, but we're not exactly sure how much. In order to verify that, we'll have to follow up with a shapiro-wilk test that validates whether it is normally distributed or not. The test states that if the p-value is less than (0.5), the null hypothesis is rejected and there is evidence of non-normality. The code below tests its normality:

shapiro.test(residuals(model_bc))

```
        Shapiro-Wilk normality test

data:  residuals(model_bc)
W = 0.98599, p-value = 0.04031
```

The p-value of 0.040 shows that the null hypothesis has been rejected, and there is evidence of the data not being normally distributed. There were a couple of ways to fix the p-value, and we chose to remove outliers using the cook's distance.

Cook's distance is primarily used in regression analysis to identify the outliers or any significant element in the dataset. Using Cook's distance, it can tell us how much the regression coefficients would change if that specific data point were to be removed.

```
#identifying outliers through cooks distance
cooks_d <- cooks.distance(model_bc)
#print(cooksd)

# creating threshold
threshold <- 4 / (nrow(car_data) - length(coef(model_bc)))

# Identify outliers based on the threshold
outliers <- which(cooks_d > threshold)

# Remove outliers from the dataset
car_data_clean <- car_data[-outliers, ]
model_subset <- car_data_clean[car_data_clean$carbody %in% c("hatchback", "sedan", "wagon"), ]
model_bc_clean <- lm(log(price_transformed) ~ log(horsepower) * carbody, data = model_subset)
```

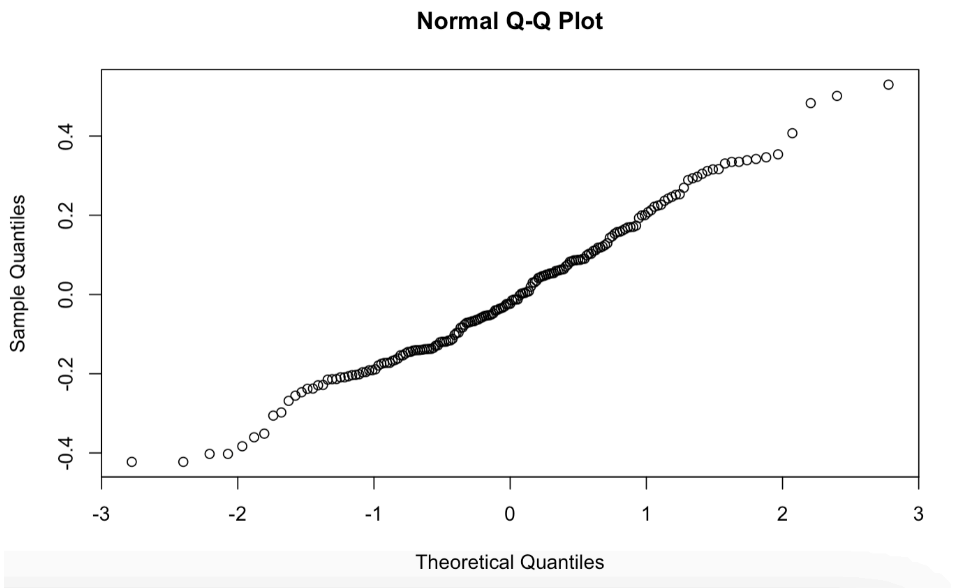
We first identified the outliers using cooks_distance() formula, then created a general threshold of *4/(n - p - 1),* this is the general rule of thumb for identifying outliers, which was provided in a documentation by scikit-yb.

We then moved the outliers that exceeded the threshold into a variable called *outliers*, and created a new variable *car_data_clean*, to remove outliers from *car_data*. The variable *model_subset* would then need to be updated, implementing *car_data_clean*, and selecting the body types again. The linear model is updated again from the variable *model_bc_clean.*

```
	Shapiro-Wilk normality test

data:  residuals(model_bc_clean)
W = 0.98663, p-value = 0.08001
```

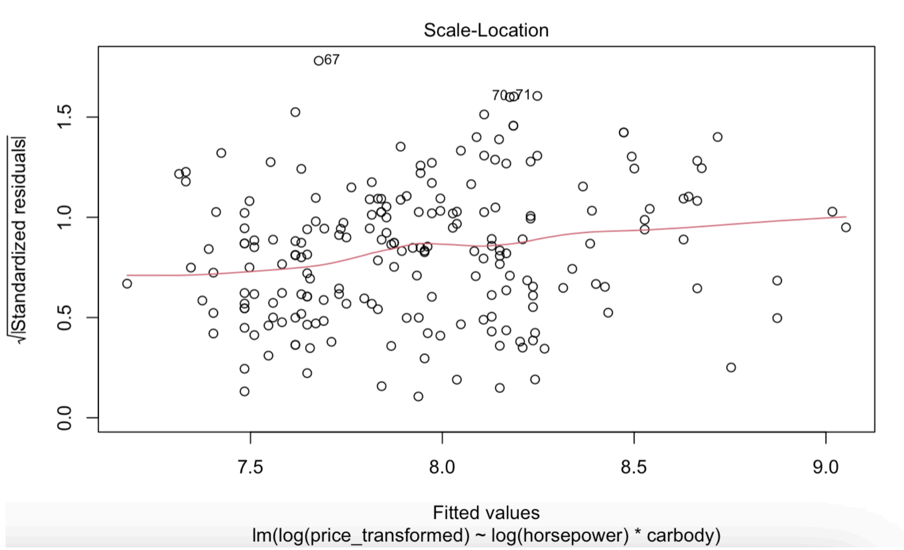
We can see now that the p-value is greater than 0.05, verifying that the data has evidence of normality.



Our Q-Q plot also verifies normality, given that the plot follows a relatively straight line, with some slight deviations in the upper tail, which could signify some skewness, but not significant.

Checking for Homoscedasticity

Homoscedasticity is similar to the independence of residuals, which is an indicator that the variance of the residuals is consistent across all levels of IV. It should remain constant no matter what the value of the predictor is. We used a Scale-Location plot that shows the spread of residuals at each level of fitted values



The plot shows the spread of residuals that has a slight upward trend with the fitted values, as shown in the upward trend of the red line. There might be potential violation of the assumption, but in general there is a greater indication of homoscedasticity, meeting the assumption.

Checking for no perfect multicollinearity

No perfect collinearity ensures stability of the estimates of regression coefficients. A perfect collinearity occurs when one IV can be exactly predicted from another IV, which results in an infinite number of solutions. It is similar to trying to estimate two predictors when the two predictors are essentially the same thing. The model cannot provide unique estimates for the coefficients when there is perfect multicollinearity.

We tested this assumption with the car library, using VIF (variance inflation factor).

```
#checking for no perfect multicollinearity
library(car)
vif(model_bc_clean)
```

```
                                GVIF Df GVIF^(1/(2*Df))
log(horsepower)             2.330762  1        1.526683
carbody                 47688.361185  2       14.777573
log(horsepower):carbody 48523.475603  2       14.841848
```

As shown above, we see that the GVIF value is 14.77 and 14.84 for the two variables, which suggests there is significant multicollinearity, most likely due to interaction term. VIF values suggest that any value above 10 is considered a significant collinearity, which does not meet the assumption. However, we chose to proceed with the model because of the primary interest in understanding the interaction effect itself, which is central to our research question. Moreover, significant collinearity does not bias our predictions of the model itself. Therefore, even with the unmet assumption, we will choose to continue with our model.

Checking Homogeneity for regression slopes

We tested the assumption which presents the effect of the covariate on the dependent variable and that it should be consistent across all levels of the independent variable. We compared two models; one that included an interaction term between *horsepower* and *car body,* and one that did not.

```
#checking for homogenity of regression slopes
reduced_model <- lm(log(price_transformed) ~ log(horsepower) + carbody, data = model_subset)
model_comparison <- anova(reduced_model, model_bc_clean)

model_comparison
```

```
> model_comparison
Analysis of Variance Table

Model 1: log(price_transformed) ~ log(horsepower) + carbody
Model 2: log(price_transformed) ~ log(horsepower) * carbody
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    179 6.8709
2    177 6.6903  2   0.18059 2.3889 0.09468 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA comparison yielded a p-value of 0.09468, which is greater than the significance level of 0.05, suggesting that the interaction terms do not improve the model. Therefore, these results show evidence of homogeneity of regression slopes, and a signal to proceed with the model.

Justification of Model Choice

In short, our model has addressed and met the required assumptions. The adjustments made to address the model’s assumptions further justified its selection, which ensures the model’s conclusions are reliable, given the availability of the data. Our chosen model is justified for its appropriateness in finding the relationships within our data and its alignment with our research objectives.

Predictor Selection

***Rationale*:**
To finalize the predictor selection for our regression model on car prices, we leveraged both EDA and inferential statistics. As shown in pages 3-9 in our report, we visualized histograms, scatterplots, and boxplots, which revealed important relationships between car prices and variables, like horsepower and car body types. The correlation analysis was also critical for the model as it claims that *horsepower* is a critical predictor because of its strong linear association with price, which indicates that cars with higher HP tend to be priced higher. This is a factor that is likely reflective of consumer valuation of performance. The inclusion of the variable *car body*, as well as the interaction between HP and car body types were justified by the variability in price distribution across the different body types, which shows the impact of HP on price that varies by car type. We understood that different body types differ in distinct market segments, thus affecting their valuation differently. Even though there was some analysis that exhibits only moderate relationships with price, the inclusion was important for a comprehensive model. Even some variables with moderate correlation with price such as engine size and curb weight, contributed to the overall explanatory power of the model. Our thoughtful selection of predictors is backed up by a combination of strong evidence and theoretical considerations, which aims to build a model that reflects the dynamic nature of car pricing.

Model Fitting and Interpretation

Fitting the chosen model

```
#fitting the chosen model into data
model_bc_clean <- lm(log(price_transformed) ~ log(horsepower) * carbody, data = model_subset)
summary(model_bc_clean)
```

```
Call:
lm(formula = log(price_transformed) ~ log(horsepower) * carbody,
    data = model_subset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42254 -0.14047 -0.02357  0.11980  0.52979

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.736216   0.296656  12.594   <2e-16 ***
log(horsepower)              0.886521   0.065126  13.612   <2e-16 ***
carbodysedan                -0.668677   0.408815  -1.636   0.1037
carbodywagon                 0.126122   0.743482   0.170   0.8655
log(horsepower):carbodysedan  0.188386   0.089373   2.108   0.0365 *
log(horsepower):carbodywagon  0.009826   0.162981   0.060   0.9520
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1944 on 177 degrees of freedom
Multiple R-squared:  0.7682,    Adjusted R-squared:  0.7617
F-statistic: 117.3 on 5 and 177 DF,  p-value: < 2.2e-16
```

Interpretation of results from summary

First off, let us restate our initial objective for this analysis: to understand the dynamic relationship between a car's horsepower, its body type, and how these factors influence the car's price. After a comprehensive exploration of our dataset and statistical modeling, we came to the conclusion that horsepower significantly impacts the price, confirming that more high-end cars tend to command higher prices in the market. The relationship between them however, is differed by the car's body type, with sedans showing a distinct price premium associated with an increase of horsepower.

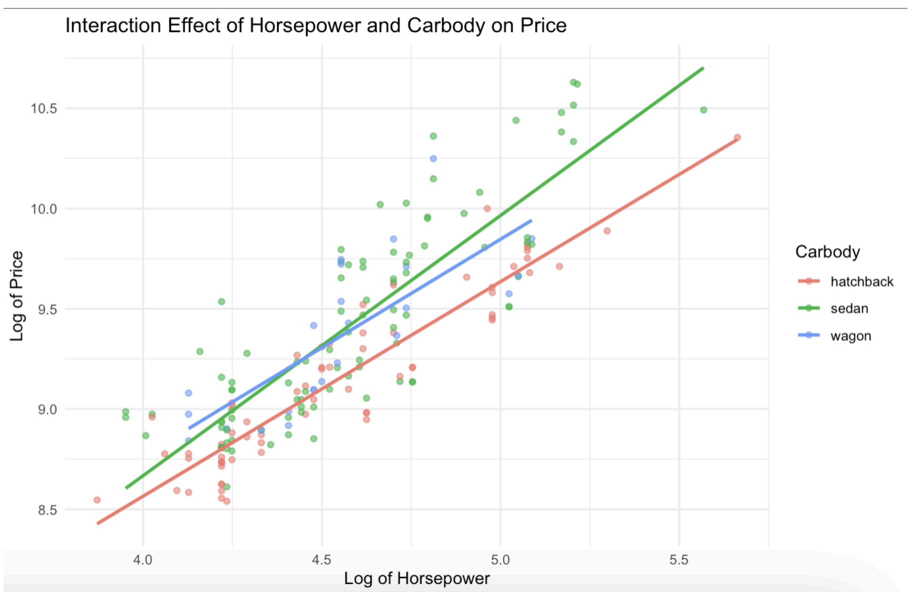
We move on to the statistical evidence that backs up our claims and conclusions. As shown above, we summarized our *model_bc_clean* model that was fitted with the transformed price as the DV and the logarithm of HP and car body types (and their interaction) as predictors. Our summary indicated a positive effect of horsepower on price, backed by the positive coefficient for *log(horsepower)* and its high p-value ($p < 2e-16$). The coefficients for car body types (sedan and wagons), when considered without the interaction term, are not statistically different at the

0.05 significance level. It suggests that the body types do not have a large effect on the price that is distinguishable from the baseline category (such as the body type hatchback).

On the other hand, the interaction term *log(horsepower):carbodysedan* is significant (p = 0.0365). This indicates the effect of HP on price is different for sedans compared to the baseline car body type. The coefficient for sedans is positive, which shows a higher price premium that is associated with horsepower. The interaction term for wagons is not significant (p = 0.9520), suggesting that the impact of horsepower on wagon prices is not statistically different from the baseline.

The adjusted R-squared value (0.7617) indicates that 76.17% of the variability in the transformed price is explained by the model, which suggests a good fit.

Interpretation of results from graph



We also visualized a graph from our regression model. The graph, presenting regression lines for each car body type (hatchbacks, sedans, and wagons) against horsepower (log), reveals that an increase in horsepower also increases car prices (log) across all body types. The sedan category presents the steepest increase, showing that sedans experience a higher price elevation with rising horsepower compared to hatchbacks and wagons. We found it fascinating, as it explains the price the consumers are willing to pay for, for performance in specific car types.

Suggestion of improvements in the model for further research

While the idea behind the model seems relevant and promising, the execution could be improved to develop a more accurate model to help with investment and analysis ventures. Such improvements could include:

- Broader Dataset: given our dataset and the amount of brands and car models out there in the world, future studies could benefit from a more extensive dataset. As shown in the graph results, the lines weren't fully completed due to the lack of data. Inclusion of more diverse car models, years, sales, or car specifics to increase accuracy.
- Model Complexity: The model could be expanded for more complex relationships, such as higher-order interactions and non-linear effects, which could be captured through polynomial regression.
- Validation with External Data: Cross-validating the findings with external data sources, such as car sales data from dealerships or price listings from car sales websites.

Sources

https://www.kaggle.com/datasets/erolmasimov/price-prediction-multiple-linear-regression?resource=download - kaggle dataset

https://www.thedrive.com/guides-and-gear/how-many-cars-are-there-in-the-world

https://www.statology.org/how-to-identify-influential-data-points-using-cooks-distance/