Cars93

Member 1, Member 2, Member 3

Introduction

For this project, we will making use of `Cars93` dataset from the `MASS` package. `Cars93` is a data set about 93 cars on sale in the USA in 1993 with 93 rows and 27 columns. It contains information such as `Manufracturer`, `Model`, `Type` of car, average `Price`, `MPG`, miles per US gallon, in city and highway, `EngineSize`, `Horsepower`, and so on. The data set contains a total of 13 missing values. 2 missing values from `Rear.seat.room` and 11 missing values from `Luggage.room` columns. The rest of the columns have 0 missing values. The goal of this project is to understand the relationship between Price and various other variables within the data set through hypothesis tests and predictive modeling.

Research Question and Hypotheses

Research Question

Based on the introduction, we would like to study the `Price` variable. We would like to ask the following questions:

- Does the presence of manual transmission affect the price of the car?
- Can the price of the car be explained/predicted by `MPG.city`, `MPG.highway`, `EngineSize`, `Horsepower`, `RPM`, `Fuel.tank.capacity` and `Weight`?

Hypotheses

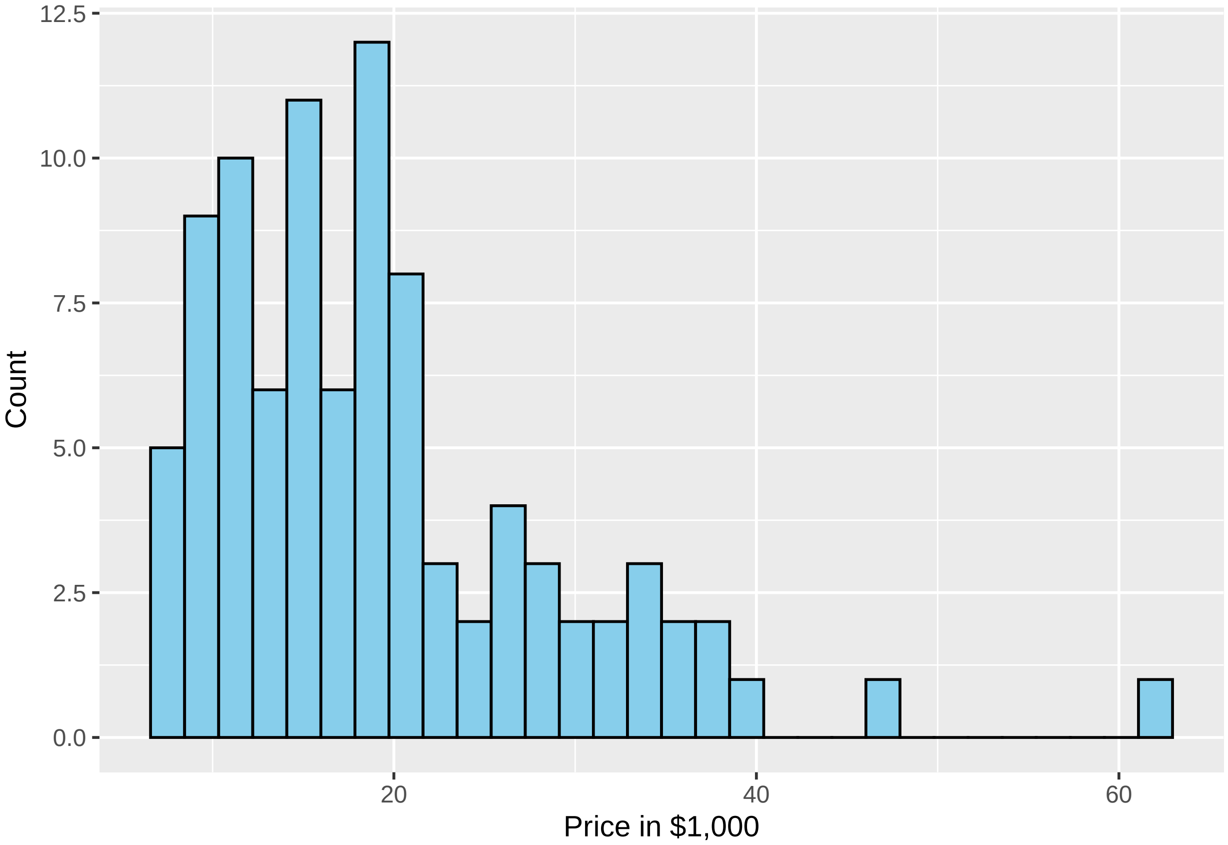
We will test out the first research question using hypothesis testing with a 95% confidence level. Our hypothesis is that there is no difference in car prices for automatic and manual transmission cars from 1993. In mathematical form it would be

$H_0 : Price_{Manual} - Price_{Automatic} = 0, \quad H_A : Price_{Manual} - Price_{Automatic} \neq 0.$

Exploratory Data Analysis

First, we need to look at the distribution of our main variable before we begin any sort of analysis. We will make use of a histogram plot.

```r
Cars93%>%
  ggplot(aes(x = Price))+
  geom_histogram(color = "black", fill = "skyblue")+
  labs(x = "Price in $1,000",
       y = "Count")
```



It looks like the car Prices are skewed to the right. This is evident because of the presence of the outliers and tail towards the right. Let's look at some summary statistics of our main variable.

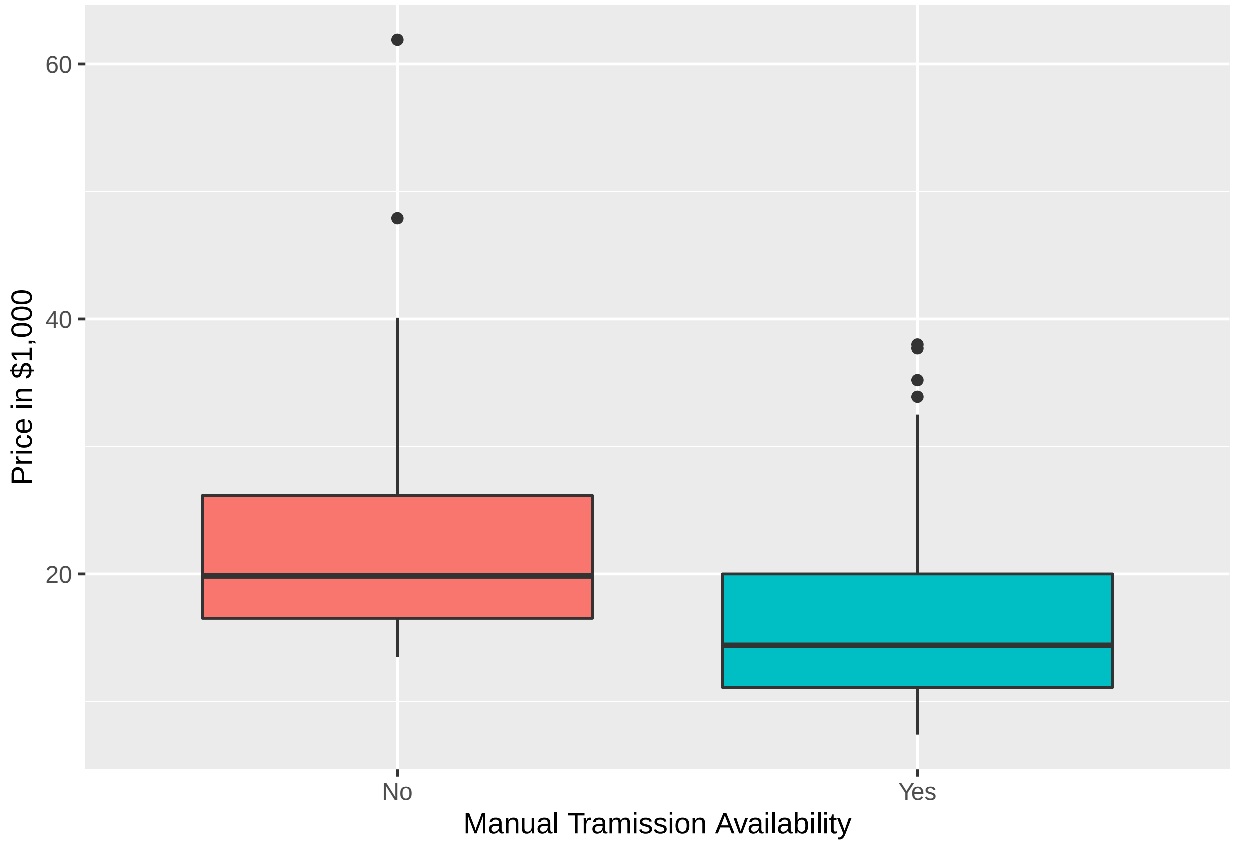
```
Cars93%>%
  summarise(Mean = mean(Price),
            SD = sd(Price),
            Min = min(Price),
            Q1 = quantile(Price, 0.25),
            Median = median(Price),
            Q3 = quantile(Price, 0.75),
            Max = max(Price))
```

```
##       Mean      SD Min   Q1 Median   Q3  Max
## 1 19.50968 9.65943 7.4 12.2   17.7 23.3 61.9
```

The right skewness is supported by the fact that the mean price of a car is quite higher than the median. Price has a large variability with a standard deviation of 9.66. The minimum price of a car was $7,400 and the maximum price was $61,900.

Let's look at `Man.trans.avail`, the variable of interest for our first question and price.

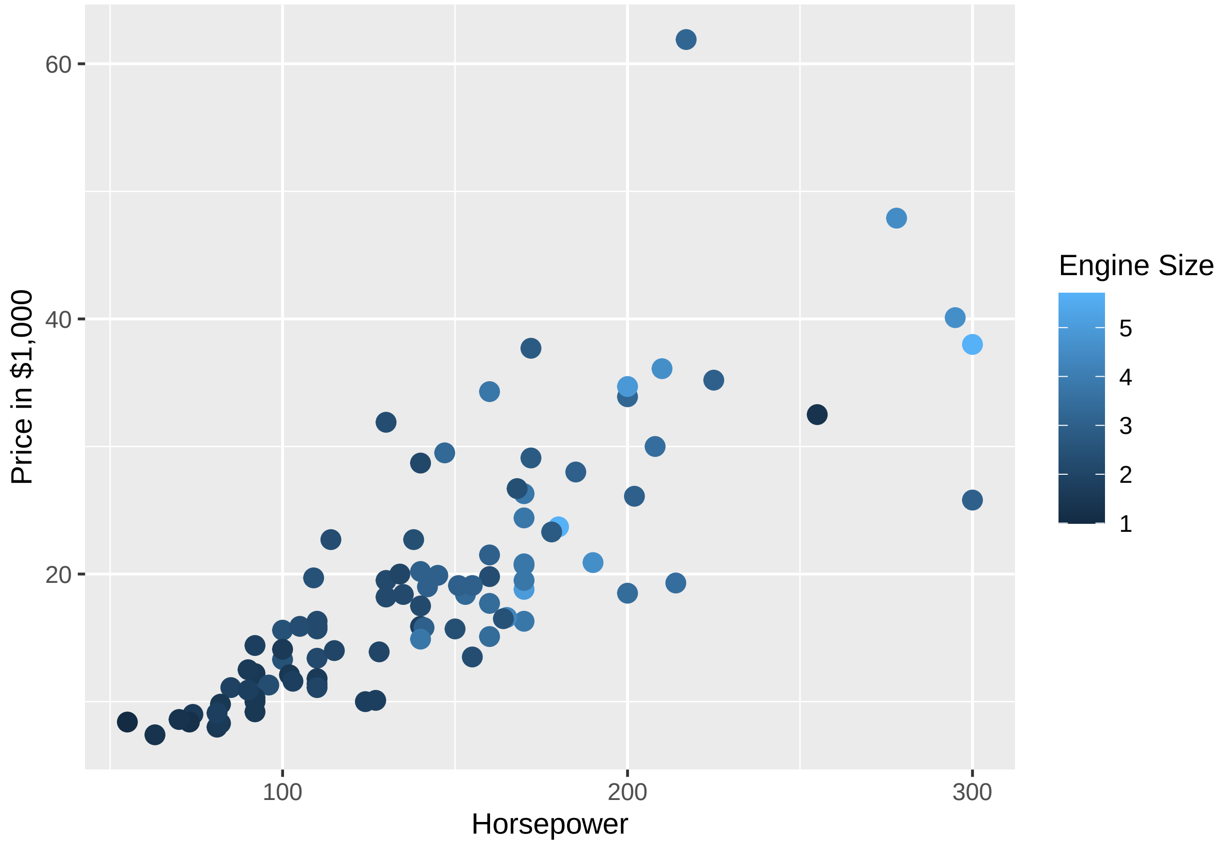
```
Cars93%>%
  ggplot(aes(x = Man.trans.avail, y = Price, fill = Man.trans.avail))+
  geom_boxplot(show.legend = F)+
  labs(x = "Manual Tramission Availability",
       y = "Price in $1,000")
```



We can see that there is a clear difference between the two box plot. Automatic cars seems to be priced a lot higher than manual cars. But this plot isn't enough evidence for us to deem that the difference is significant enough.

Before we can answer the second research question using predictive modeling. We need to understand the relationship between the main response variable (`Price`) and the explanatory variables (`MPG.city`, `MPG.highway`, `EngineSize`, `Horsepower`, `RPM`, `Fuel.tank.capacity` and `Weight`). We can understand their relationship making use of `ggplot()` function to make plots and find their correlation.

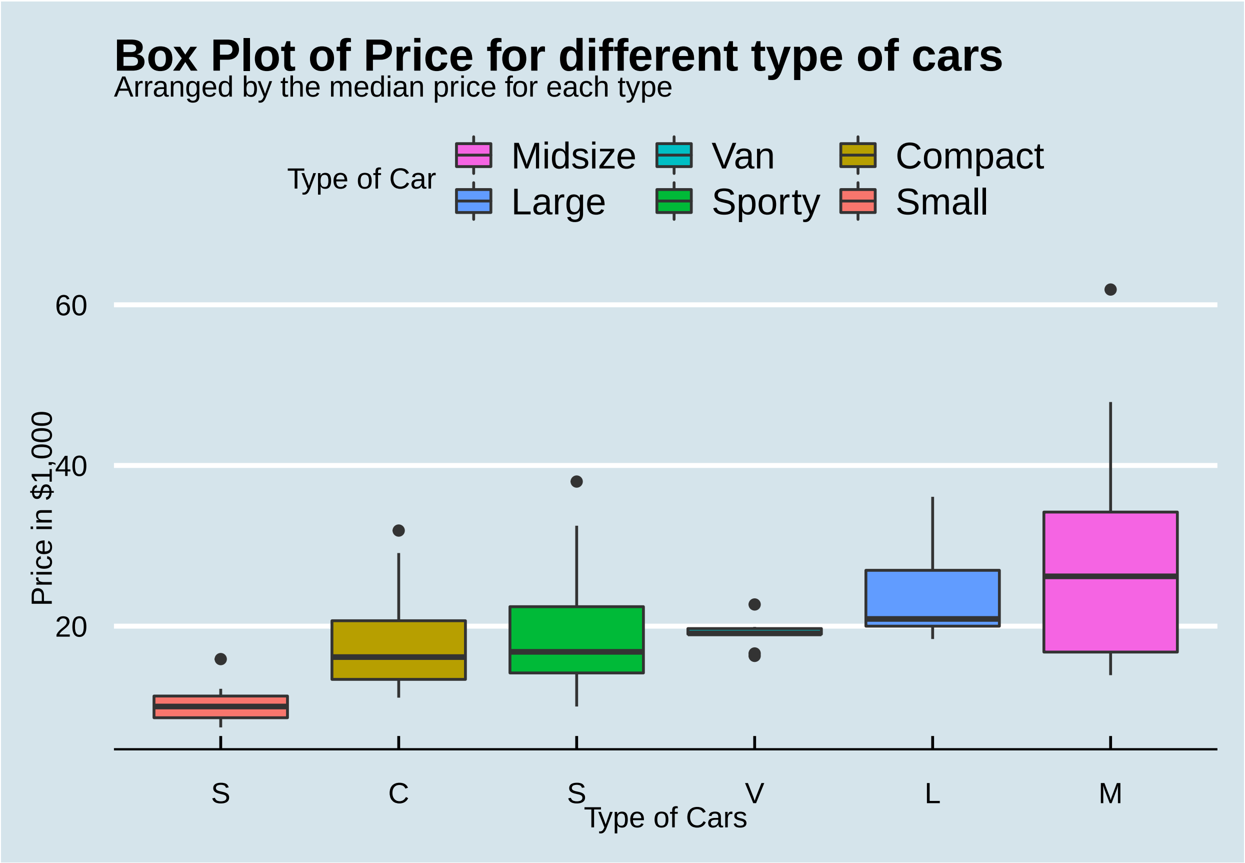
```
Cars93%>%
  ggplot(aes(x = Horsepower, y = Price, col = EngineSize))+
  geom_point(size = 3)+
  labs(x = "Horsepower",
       y = "Price in $1,000",
       col = "Engine Size")
```



In the plot above, we plotted horsepower against price. We can see a positive relationship between them. So as Horsepower in a car increases, we tend to see a increase in the price of the car. We also included the Engine Size variable as the color for the points. We can notice that as Horsepower increases, the color of the points tend to move towards the lighter shade of blue indicating that the engine size is also increasing. So from this plot we can see that Price and Horsepower have a positive relationship. Also, Price and Engine size have a positive relationship.

Next we want to look at how the price is affect by the type of cars.

```
Cars93%>%
  mutate(Type = reorder(Type, Price, median))%>%
  ggplot(aes(x = Type, y = Price, fill = Type))+
  geom_boxplot()+
  labs(x = "Type of Cars",
       y = "Price in $1,000",
       title = "Box Plot of Price for different type of cars",
       subtitle = "Arranged by the median price for each type")+
  theme(axis.text.x = element_text(angle = 45))+
  scale_x_discrete(labels = c("S","C","S","V","L","M"))+
  theme(legend.title = element_text(color = "red", size = 20),
          legend.text = element_text(color = "blue"))+
  scale_fill_discrete(name = "Type of Car")+
  guides(fill = guide_legend(reverse=TRUE))+
  theme_economist()
```

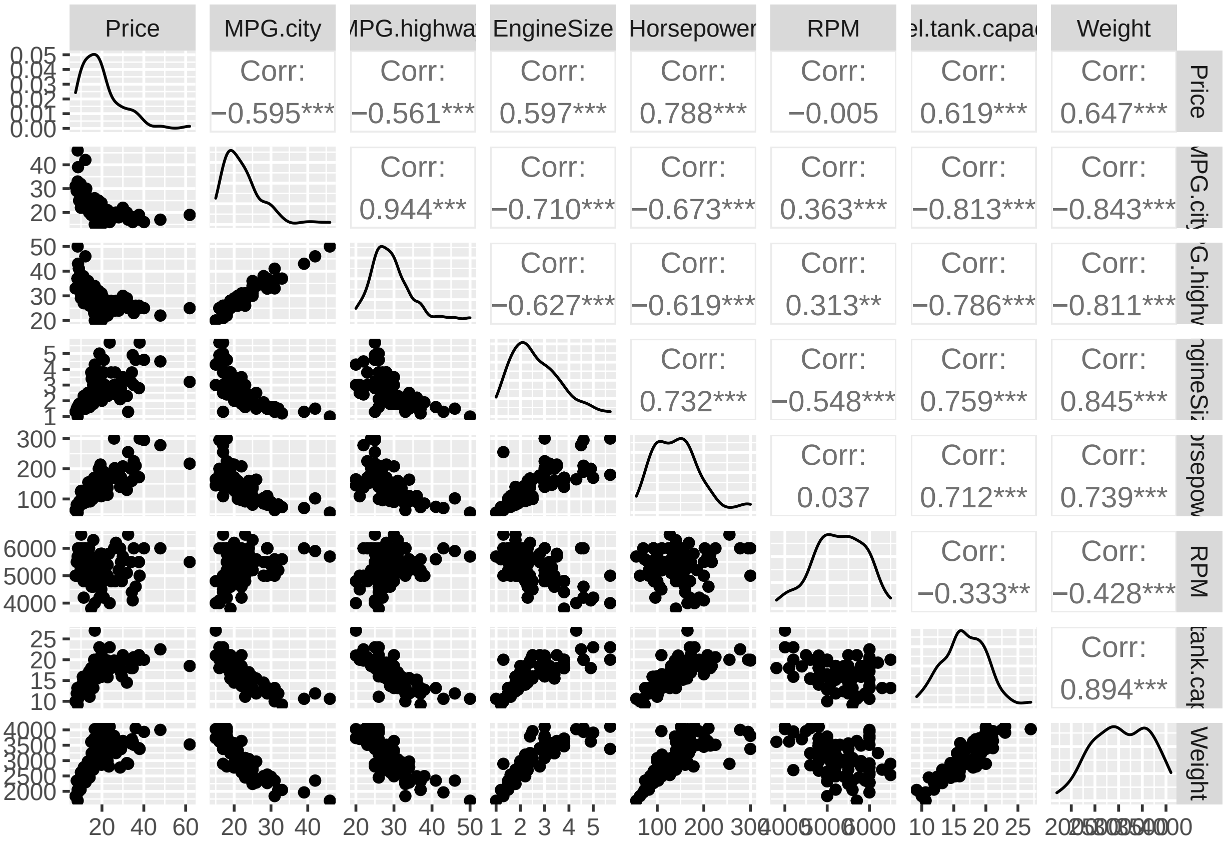


For the above side-by-side box plot, we arranged the types of cars using their median price. From the plot, we can see that Midsize cars has the highest median price while Small cars have the lowest median price. Even though Midsize has the highest median prices, it also has the highest variability while Vans have the least variability. This could indicate that there are various other factors that would have affect on the price of a car than just the type of the car alone.

Correlation

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. Correlation gives us a quantitative sense of how strong the relationship between two variables is. Correlation ranges from -1 to 1 inclusive. A positive correlation means there is a positive relationship (slope) between the two variables, i.e., as one variable increases, the second variable tends to increase as well. A negative correlation means there is a negative slope between the two variables. The close the correlation is to -1 or 1, the closer it is for the relationship between the variables to be in a straight line with no variation. From the price vs horsepower plot above, we can see that the points aren't really close to each other but there is a positive slope, then we can guess that the correlation is positive but isn't very close to 1. We can find the correlation between the response and the explanatory variables using the `ggpairs()` function from the `GGally` package. `ggpairs()`, not only calculates the correlation but it also plots the relationship for us to have a visual check for the corresponding correlation. The code is shown below:

```
Cars93%>%
  dplyr::select(Price, MPG.city, MPG.highway,
                EngineSize, Horsepower, RPM,
                Fuel.tank.capacity, Weight)%>%
  ggpairs()
```



We can see that Price and Horsepower has the highest positive correlation of 0.788. This is pretty close to 1 as we suspected. Price and MPG city has a negative correlation of -0.595 and we can clearly see a negative relation in the corresponding plot. The asterisks beside the correlation number dictates the significance of the correlation. If there are no asterisks then the correlation isn't strong at all. As we can, Price and RPM has a correlation of -0.005 which is very close to 0.

Statistical Inference

For the hypothesis test, we will make us of the `t_test()` function from the `infer` package. We will show the code down below to test our hypothesis.

```
Cars93%>%
  filter(!is.na(Price), !is.na(Man.trans.avail))%>%
  group_by(Man.trans.avail)%>%
  summarise(mean = mean(Price))
```

```
## # A tibble: 2 x 2
##   Man.trans.avail  mean
##   <fct>           <dbl>
## 1 No               23.8
## 2 Yes              17.2
```

As we can see, there is a difference between the prices but we aren't sure if this difference is significant enough for us to be able to reject the null hypothesis based on the table above. So to get a robust result, we will make use of statistical methods we have used in labs before.

```
Cars93%>%
  filter(!is.na(Price), !is.na(Man.trans.avail))%>%
  t_test(response = Price,
         explanatory = Man.trans.avail,
         order = c("No","Yes"),
         mu = 0,
         conf_int = T,
         conf_level = 0.95)
```

```
## # A tibble: 1 x 7
##   statistic  t_df p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      3.05  51.1 0.00365 two.sided       6.60     2.25     11.0
```

From the results table above, we can see that the p-value is less than 0.05 meaning that we can reject the null hypothesis. We can see that the estimate is 6.60292 which means that the difference between the mean price of automatic cars and the mean price of manual cars is \$ 6,602.92. The difference is positive. Thus we know that automatic cars on average are higher in price compared to manual cars. Based on the result, we can construct the confidence interval as CI: (2.25, 10.95). The interpretation of this confidence interval is that 95% of the time the mean price difference between automatic and manual cars is between \$2,254.3 and \$10,952.5.

Predictive Modeling

Methodology

We will be making use of multiple linear regression model. From the `ggpairs()` plot above, we see that the relationship between Price and all the explanatory variables is very linear. So this method is fitting for us to use. First, we will include all the variables in the model. Then we will remove one variable at a time till all the variables in the model are significant. We decide the significance of a variable using p-value. If the p-value is less than 0.05 then the variable is significant. The code is down below for the full model.

```
full_model = lm(Price ~ MPG.city + MPG.highway +
                  EngineSize + Horsepower + RPM +
                  Fuel.tank.capacity + Weight, data = Cars93)
summary(full_model)
```

```
##
## Call:
## lm(formula = Price ~ MPG.city + MPG.highway + EngineSize + Horsepower +
##     RPM + Fuel.tank.capacity + Weight, data = Cars93)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.406  -2.916  -0.535   1.430  31.977
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5.5822199 16.4516871  -0.339 0.735213
## MPG.city           -0.0261593  0.3931094  -0.067 0.947101
## MPG.highway        -0.0782824  0.3822441  -0.205 0.838220
## EngineSize         -0.3476376  1.7889194  -0.194 0.846383
## Horsepower          0.1212092  0.0316744   3.827 0.000247 ***
## RPM                 0.0006197  0.0021207   0.292 0.770819
## Fuel.tank.capacity  0.0276658  0.4515211   0.061 0.951286
## Weight              0.0025104  0.0032621   0.770 0.443677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.081 on 85 degrees of freedom
## Multiple R-squared:  0.6338, Adjusted R-squared:  0.6037
## F-statistic: 21.02 on 7 and 85 DF,  p-value: 3.701e-16
```

The estimated regression equation for the model above is:

$$\widehat{Price} = -5.582 - 0.026 \times MPG.city - 0.078 \times MPG.highway - 0.348 \times EngineSize + 0.121 \times Horsepower + 0.0006 \times RPM + 0.028 \times Fuel.tank.capacity + 0.003 \times Weight$$

From the summary table above, MPG.city, MPG.highway, and EngineSize has a negative estimate of -0.026, -0.078 and -0.348 respectively. This means that a unit increase in MPG.city correlates to a decrease of 0.026 units ($26) in Price on average. We can make similar interpretations for the negative estimate variables. Horsepower, RPM, Fuel tank capacity and Weight has a positive estimate. We interpret the result for Horsepower as a unit increase in Horsepower correlates to a increase of 0.121 ($121) units in Price on average. From the results, we can conclude that only 60.37% of total variation in the outcome (response) variable is explained by the explanatory variables collectively. The whole regression model isn't statistically significant as majority of the variables have a p-value much higher than 0.05.

Here we can see that fuel tank capacity has the highest p-value. So we will proceed to remove the variable from the full model and keep repeating this procedure till we have a model with all significant variables like Horsepower from the full model.

Results

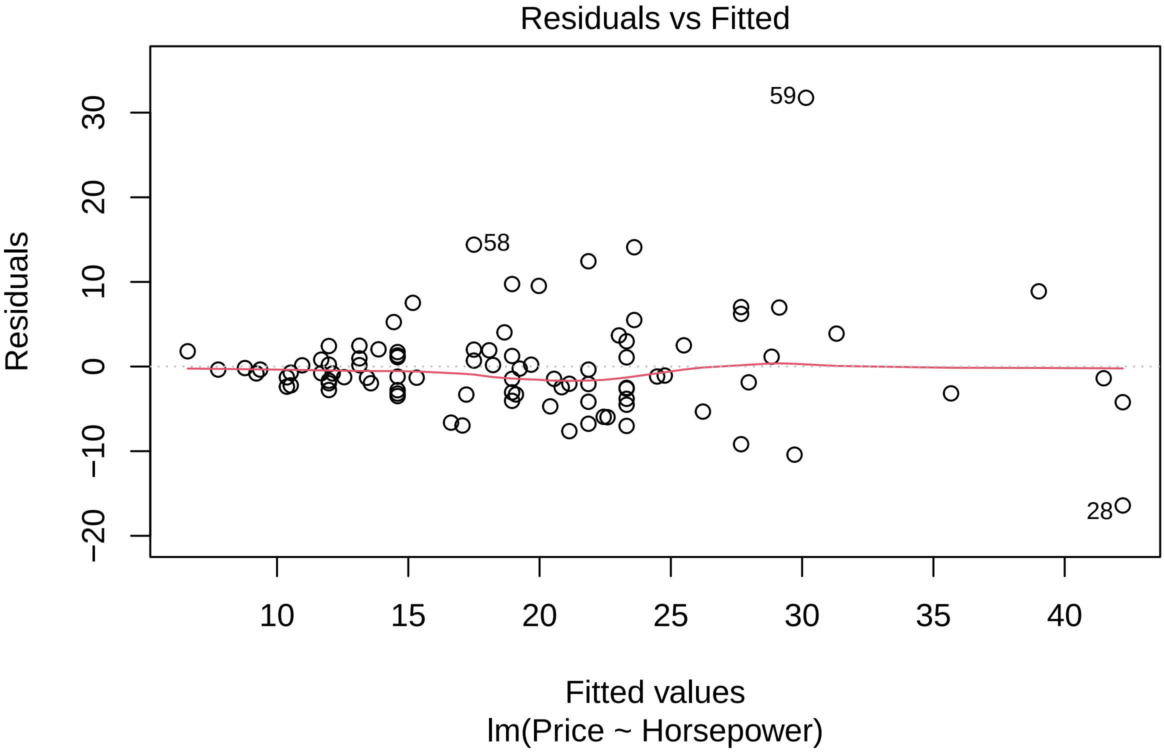
Below we have the final variable model with the selection method applied as discussed above. After removing all the highly insignificant variables, we are only left with Horsepower as the only variable with a high significance as the p-value much less than 0.05.

```r
final_model = lm(Price ~ Horsepower, data = Cars93)
summary(final_model)
```

```
##
## Call:
## lm(formula = Price ~ Horsepower, data = Cars93)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.413  -2.792  -0.821   1.803  31.753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.3988     1.8200  -0.769    0.444
## Horsepower    0.1454     0.0119  12.218   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.977 on 91 degrees of freedom
## Multiple R-squared:  0.6213, Adjusted R-squared:  0.6171
## F-statistic: 149.3 on 1 and 91 DF,  p-value: < 2.2e-16
```

Before we interpret the results from the model, we want to check if the model satisfies the assumptions of linear regression. To test this, we will plot the residual from the model.

```r
plot(final_model, which = 1)
```



Here we can see the residual plot for the final model. We can observe that the residual plot starts off being close to zero and clustered and starts to spread out as the fitted values start to increase. This violates the constant variance assumption for the residuals.

Interpretation

We can see that the estimate for Horsepower has changed compared to the full model. So based on the new model, we can say that a unit increase in Horsepower correlated to a increase of 0.1454 units ($145.4) in Price on average. The Adjusted R^2 for the final model is 61.71% which is a 1.34% higher than the full model that we built earlier. Thus, this one variable model has a higher percentage of total variation in the outcome variable is explained by the explanatory variable.

Conclusion

We were able to test out our hypothesis regarding the presence of a manual transmission affects the price. We found out that automatic transmissions cars are more expensive on average compared to a manual transmission car. We can conclude that the most of the numerical variables in the data set isn't useful enough to be used to explain the Price of a given car. Only Horsepower was a good variable that was able to attain a good p-value in the linear regression model.

Discussion

One of the ways we could improve the results from this work would be to make use of the categorical variables available in the data set. We can also improve by using a data set with more observations. A data set with close to 1000 observations will make the results of the linear regression model better and reliable.

Reference

Venables, W. N. and Ripley, B. D. (1999) Modern Applied Statistics with S-PLUS. Third Edition. Springer. https://www.rdocumentation.org/packages/MASS/versions/7.3-56/topics/Cars93

Packages and Data

```r
library(MASS)
library(tidyverse)
library(GGally)
library(infer)
library(ggthemes)
data("Cars93")
```