Regression Models Project

Executive Summary

In this project, we're asked to use the *mtcars* dataset and write up an analysis to answer two main questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

To answer these questions we'll build a number of linear regression models in increased complexity and compare them in terms of their performance. As a result of this study we find that manual transmission cars in this dataset have better *mpg* by about **2** miles/gallon compared to automatic transmission cars. However this difference in performance is found not to be statistically significant based on the final model used here.

Analysis

Let us begin by investigating the dataset first:

```
dim(mtcars)
```

```
## [1] 32 11
```

```
head(mtcars,2)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4       21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag   21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

As can be seen, there are a total of 32 observables (cars) that have 11 measurements. One can simply do `?mtcars` in the console to see what each measurement corresponds to, and what their units are. For the sake of bervity, we'll not do this here.

In order to tackle the afore mentioned questions, let us build a few linear models, in increasing complexity, following the course contet:

```
model1 <- lm(mpg ~ factor(am) - 1, data = mtcars)
model2 <- update(model1, . ~ . + wt)
model3 <- update(model2, . ~ . + factor(cyl))
model4 <- update(model3, . ~ . + hp)
model5 <- update(model4, . ~ . + disp)
```

Now, let's look at the coefficients of the first model:

```
## factor(am)0 factor(am)1
##    17.14737    24.39231
```

Here the outcome (*mpg*) is explained by a single regressor (*am* - transmission type) that is a factor variable. **-1** in the formula means we don't want to have the constant term. As a result, we end up with two coefficients, one corresponding to the automatic transmission (0), and the other to the manual transmission (1). The numerical values of the coefficients show the median *mpg* for each transmission type, namely **17** miles/gallon for the automatic transmission and **24** miles/gallon for the manual transmission. This can be verified looking at the first box plot in the Appendix.

Now, the question is “Is this the end of the story? Do manual transmission cars significantly outperform automatic counterparts based on this dataset?”. If one looks at the correlation matrix (second figure in the Appendix), which shows how much the variables are correlated to each other in the dataset, s/he immediately sees there are other variables that are more correlated to the *mpg* than the transmission type. This means that if we add them to our model we have a better chance to describe the variability in *mpg*. Let’s use ANOVA to see how the other models, where we add other variables, compare to the first one:

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ factor(am) - 1
## Model 2: mpg ~ factor(am) + wt - 1
## Model 3: mpg ~ factor(am) + wt + factor(cyl) - 1
## Model 4: mpg ~ factor(am) + wt + factor(cyl) + hp - 1
## Model 5: mpg ~ factor(am) + wt + factor(cyl) + hp + disp - 1
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     29 278.32  1    442.58 73.5623 6.452e-09 ***
## 3     27 182.97  2     95.35  7.9244   0.00216 ** 
## 4     26 151.03  1     31.94  5.3093   0.02980 *  
## 5     25 150.41  1      0.62  0.1025   0.75149    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

As one can see, *Model 4*, which includes the weight, cylinder type, and horsepower in addition to the transmission type performs the best. Adding displacement doesn’t gain us much since it is highly correlated to the other regressors. Let’s look at the summary of this model:

```
## 
## Call:
## lm(formula = mpg ~ factor(am) + wt + factor(cyl) + hp - 1, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9387 -1.2560 -0.4013  1.1253  5.0513 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## factor(am)0  33.70832    2.60489  12.940 7.73e-13 ***
## factor(am)1  35.51754    2.03171  17.482 6.81e-16 ***
## wt           -2.49683    0.88559  -2.819  0.00908 ** 
## factor(cyl)6 -3.03134    1.40728  -2.154  0.04068 *  
## factor(cyl)8 -2.16368    2.28425  -0.947  0.35225    
## hp           -0.03211    0.01369  -2.345  0.02693 *  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.9892, Adjusted R-squared:  0.9868 
## F-statistic: 398.6 on 6 and 26 DF,  p-value: < 2.2e-16
```

In this model, one can see that the cylinder type, horsepower, and the weight are all negatively correlated to the milage (i.e. the coefficients are negative). The coefficient of the weight is **-2.5**, which means that if all other varibles are kept fixed, the *mpg* will go down by **2.5** gallons for each additional **1000** lbs in weight. The R-squared value of this model tells us that it explains **99%** of the variation in *mpg*, which is very good. A few diagnostic plots for this model, which don’t show any striking issues, are shown in the third figure in the Appendix. Now let’s see the confidence intervals for the predicted milage for a hypothetical car that has the mean weight and horsepower as of the original dataset, and 6 cylinders:

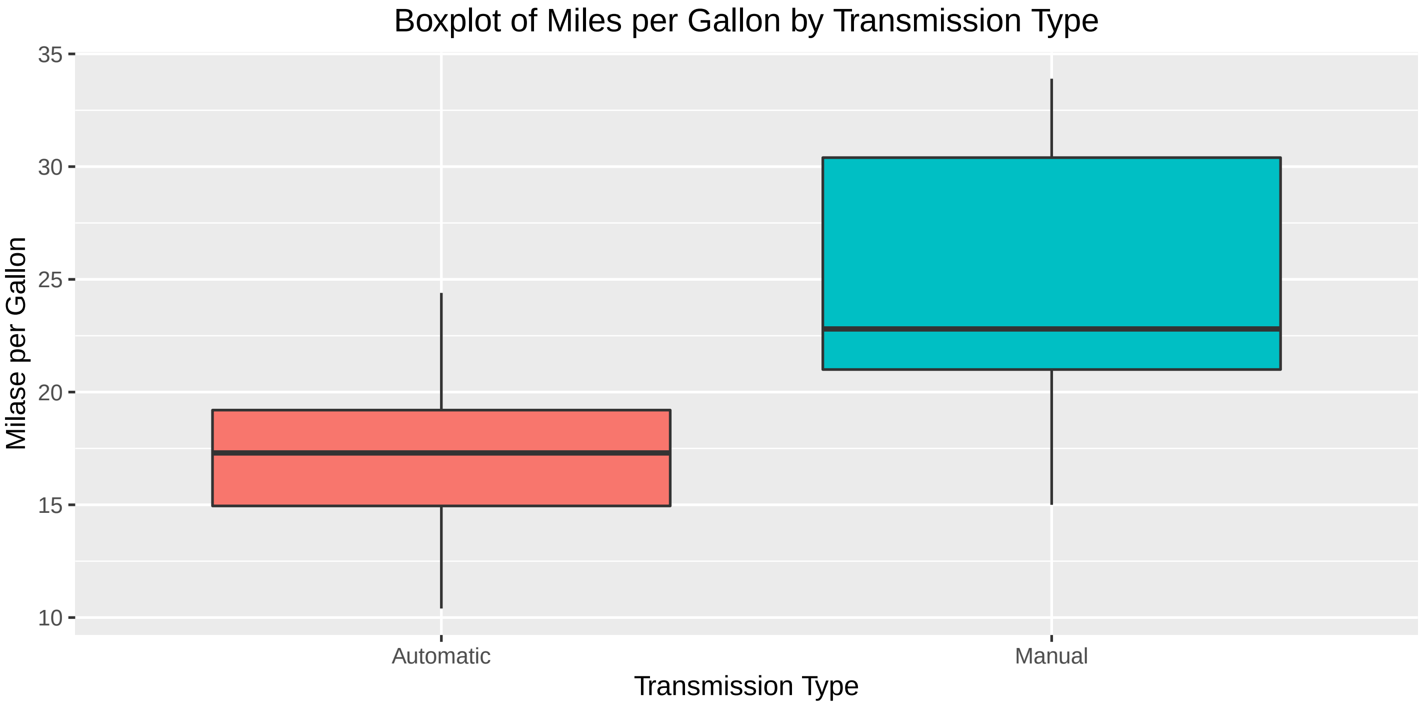
```
# Predict
model4.prediction <- predict(model4,
                             data.frame(am = c(0,1), wt = mean(mtcars$wt), cyl = c(6), hp = mean(mtcars$hp)),
                             interval="confidence")
print(model4.prediction)
```

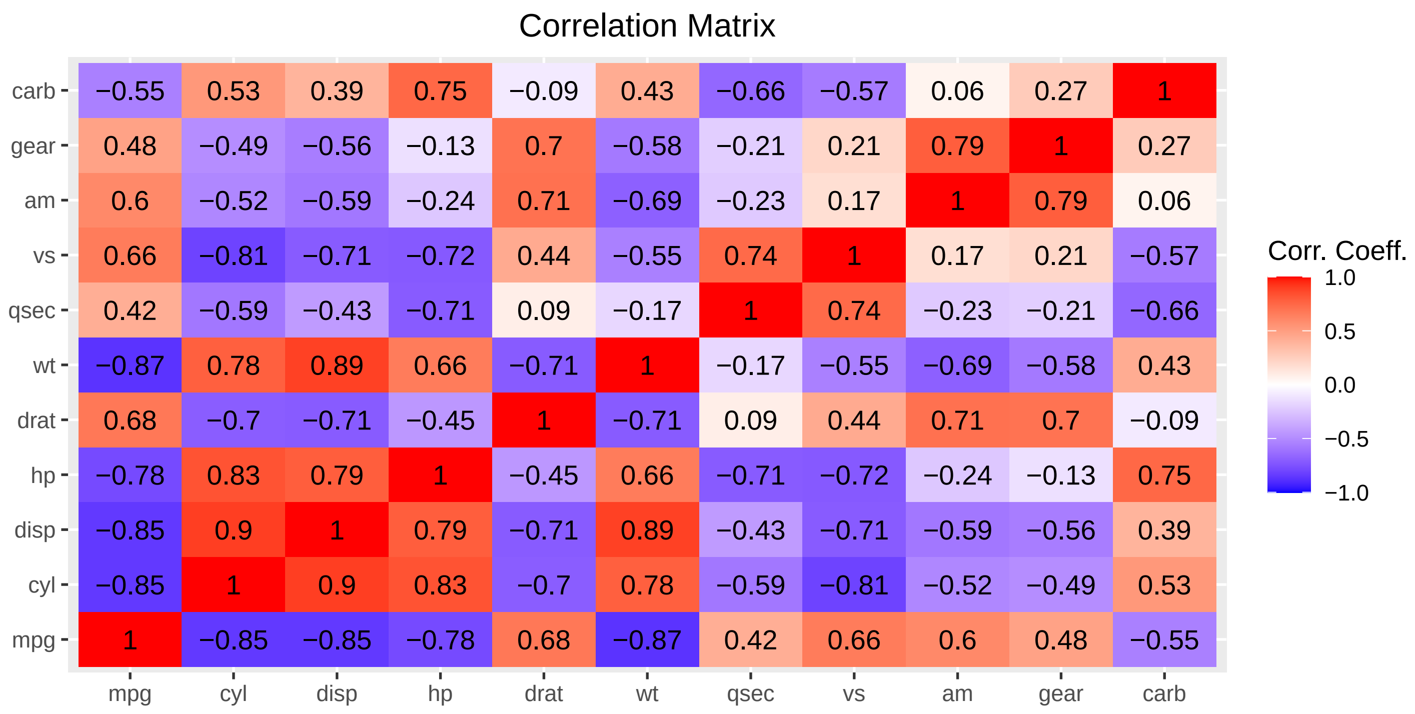
```
##        fit      lwr      upr
## 1 17.93400 15.47757 20.39043
## 2 19.74321 17.32716 22.15927
```

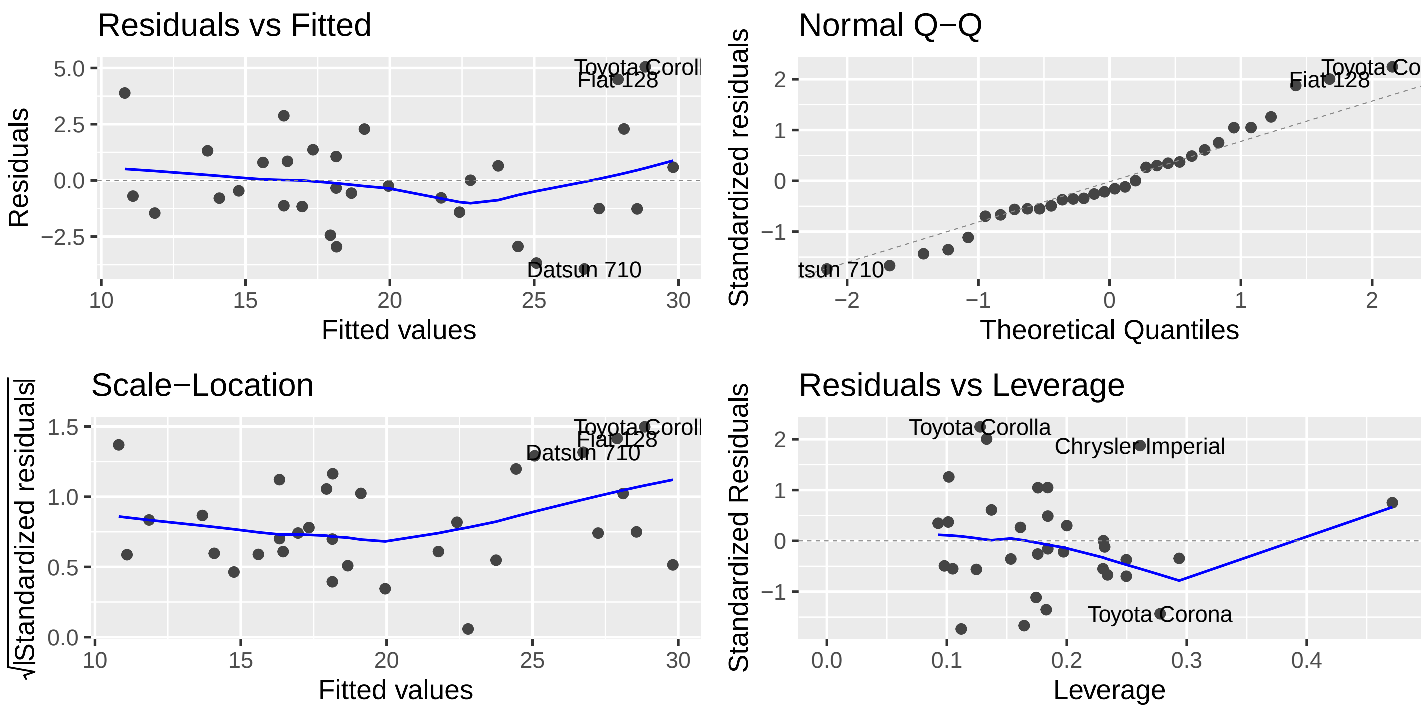
Conclusions

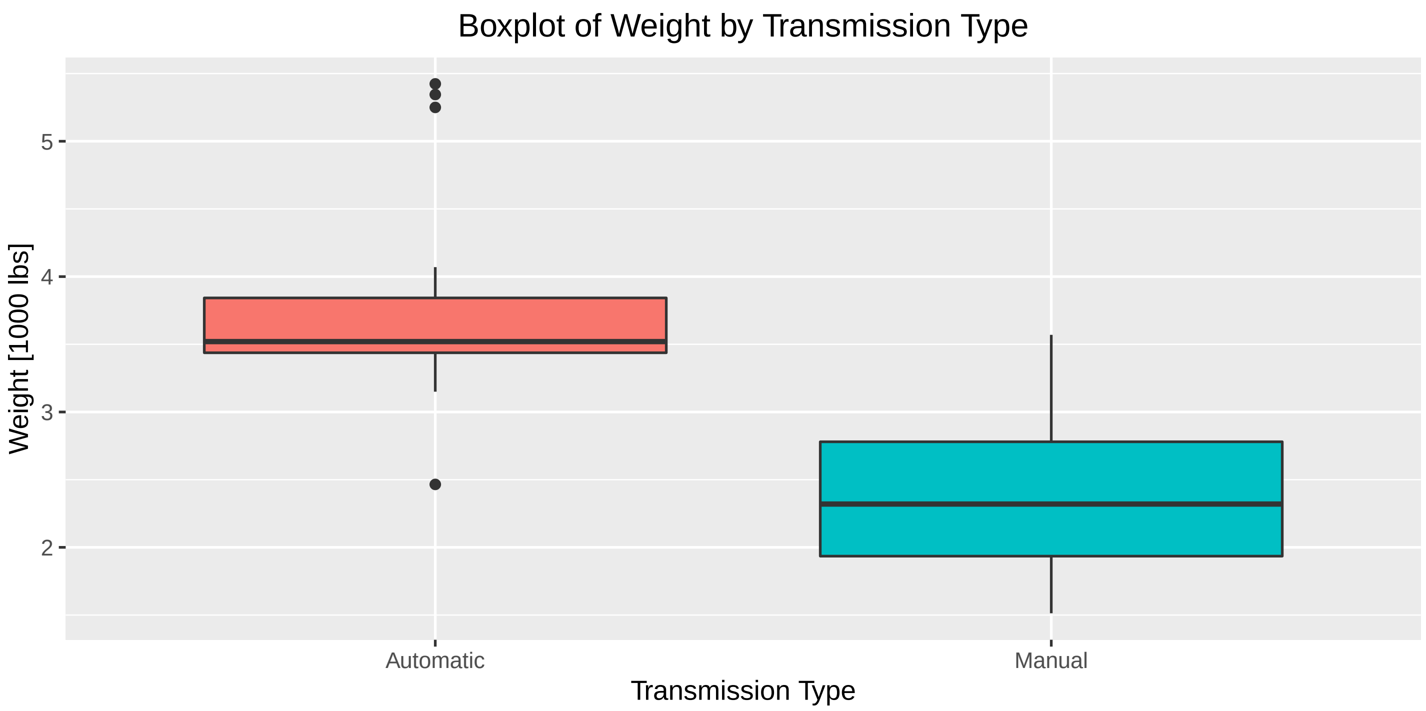
What we see is that if our hypothetical car has automatic transmission it has an expected *mpg* of about **18** miles/gallon, and if it has manual transmission about **20** miles/gallon. However, the **95%** confidence level intervals are not exclusive, i.e. although the manual transmission looks to have better milage (about **2** miles per gallon - which can be verified by the fitted model coefficients for the transmission type), it's not statistically significant.

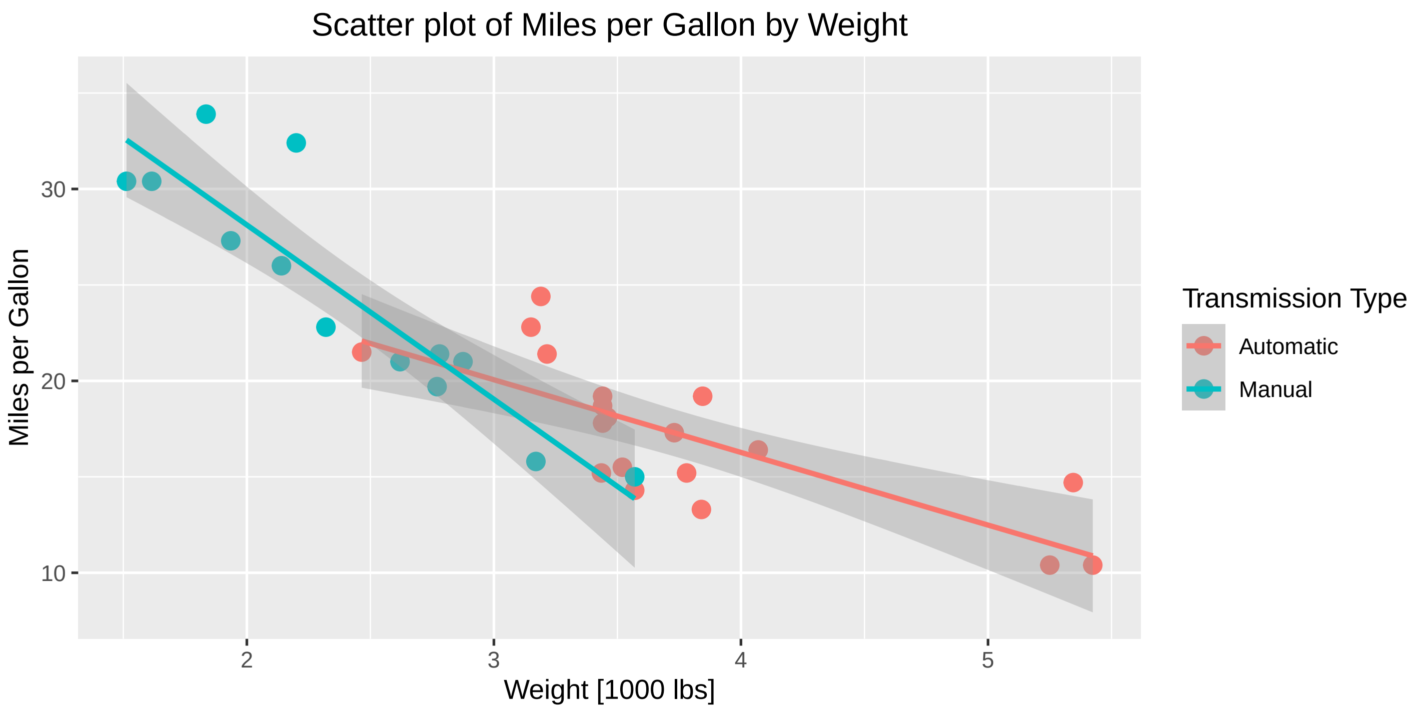
Appendix

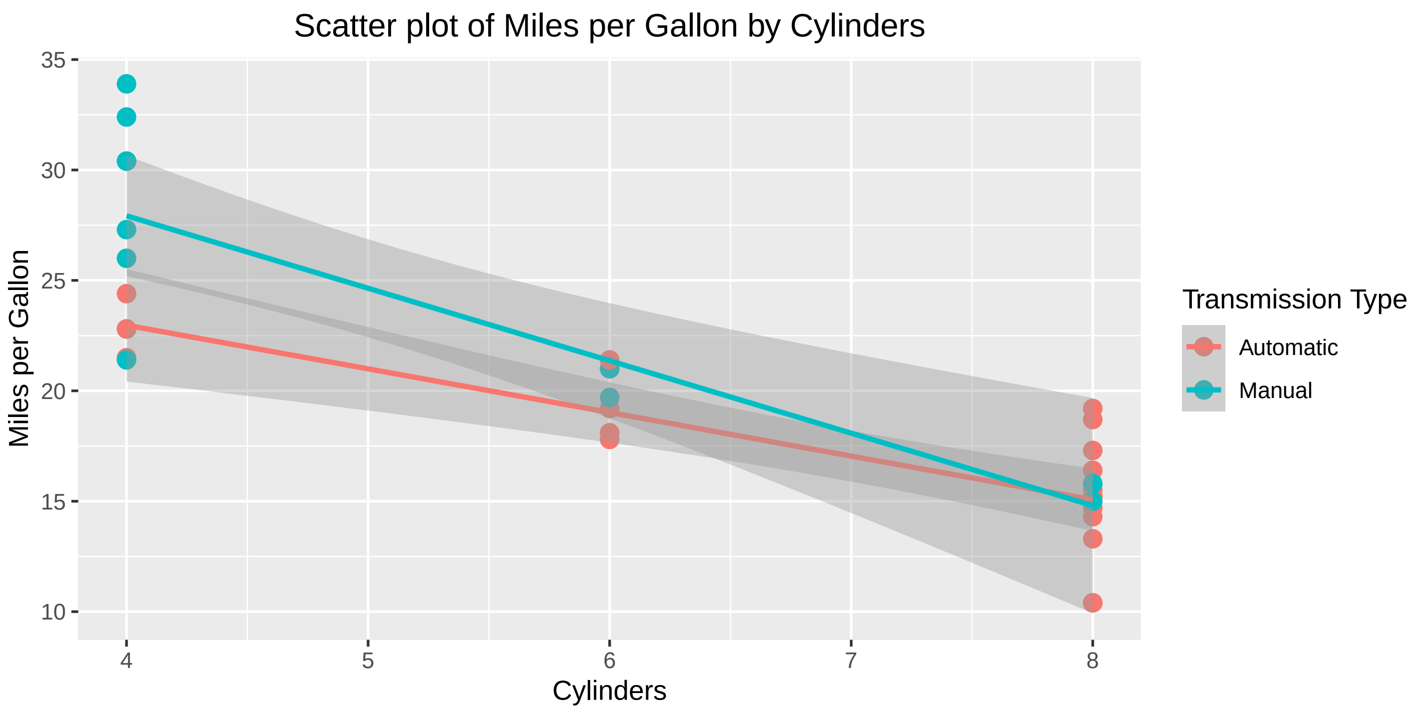



Correlation Matrix
carb −0.55 0.53 0.39 0.75 −0.09 0.43 −0.66 −0.57 0.06 0.27 1
gear 0.48 −0.49 −0.56 −0.13 0.7 −0.58 −0.21 0.21 0.79 1 0.27
am 0.6 −0.52 −0.59 −0.24 0.71 −0.69 −0.23 0.17 1 0.79 0.06
vs 0.66 −0.81 −0.71 −0.72 0.44 −0.55 0.74 1 0.17 0.21 −0.57
qsec 0.42 −0.59 −0.43 −0.71 0.09 −0.17 1 0.74 −0.23 −0.21 −0.66
wt −0.87 0.78 0.89 0.66 −0.71 1 −0.17 −0.55 −0.69 −0.58 0.43
drat 0.68 −0.7 −0.71 −0.45 1 −0.71 0.09 0.44 0.71 0.7 −0.09
hp −0.78 0.83 0.79 1 −0.45 0.66 −0.71 −0.72 −0.24 −0.13 0.75
disp −0.85 0.9 1 0.79 −0.71 0.89 −0.43 −0.71 −0.59 −0.56 0.39
cyl −0.85 1 0.9 0.83 −0.7 0.78 −0.59 −0.81 −0.52 −0.49 0.53
mpg 1 −0.85 −0.85 −0.78 0.68 −0.87 0.42 0.66 0.6 0.48 −0.55
mpg cyl disp hp drat wt qsec vs am gear carb
Corr. Coeff.
1.0
0.5
0.0
−0.5
−1.0


Residuals vs Fitted
Residuals
5.0
2.5
0.0
−2.5
Toyota Coroll
Fiat 128
Datsun 710
10
15
20
25
30
Fitted values
Normal Q−Q
Standardized residuals
2
1
0
−1
Toyota Co
Fiat 128
tsun 710
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
Toyota Coroll
Fiat 128
Datsun 710
10
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized Residuals
2
1
0
−1
Toyota Corolla
Chrysler Imperial
Toyota Corona
0.0
0.1
0.2
0.3
0.4
Leverage


Boxplot of Weight by Transmission Type
Weight [1000 lbs]
5
4
3
2
Automatic
Manual
Transmission Type


Scatter plot of Miles per Gallon by Weight
Miles per Gallon
30
20
10
2
3
4
5
Weight [1000 lbs]
Transmission Type
Automatic
Manual


Scatter plot of Miles per Gallon by Cylinders
Miles per Gallon
35
30
25
20
15
10
Cylinders
4
5
6
7
8
Transmission Type
Automatic
Manual