Bridget Erb
MAT3333
Final Project: Modeling

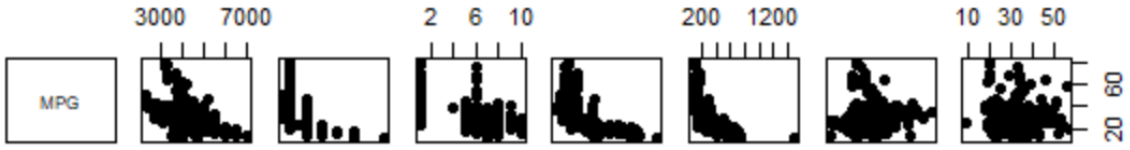
Data Exploration:
The first thing to do is explore the relationship between the potential predicting variables and the outcome, so we can assess which variables will need transformations as well as which ones we might consider adding to our model.

First, we compute the correlation coefficient between MPG and the continuous potential predicting variables:

```
                    MPG
MPG           1.0000000
Weight       -0.5796093
Cylinders    -0.6151101
Gears        -0.4829639
Displacement -0.6399918
Horsepower   -0.6649640
AxleRatio     0.1131194
NVratio       0.1735610
```

From this output, we can see that Weight, Displacement, and Horsepower seem to have the strongest relationship with MPG. However, this does not include the relationship between MPG and the categorical predictors in the dataset (FuelType, Transmission, and DriveSystem) and does not take into account the issue of multicollinearity, so this is not enough to go off of for our model.

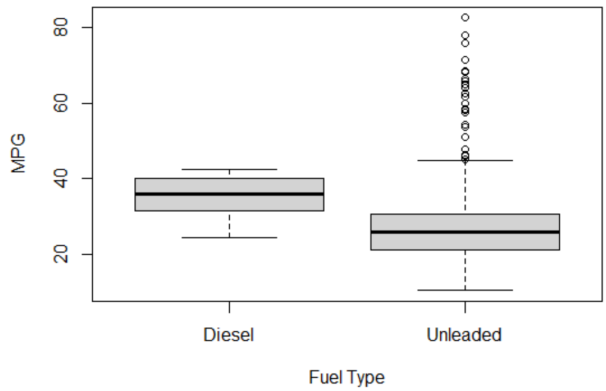
Next, we will look at a scatter matrix between all of the continuous variables and MPG:

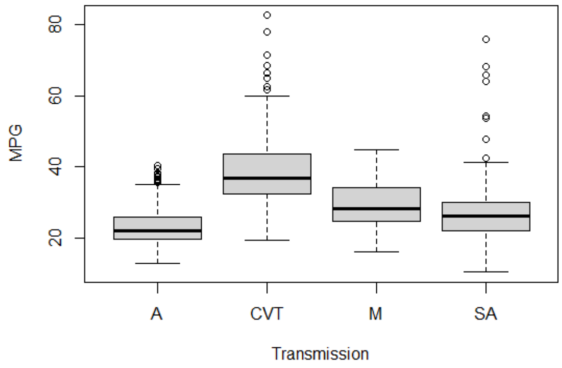


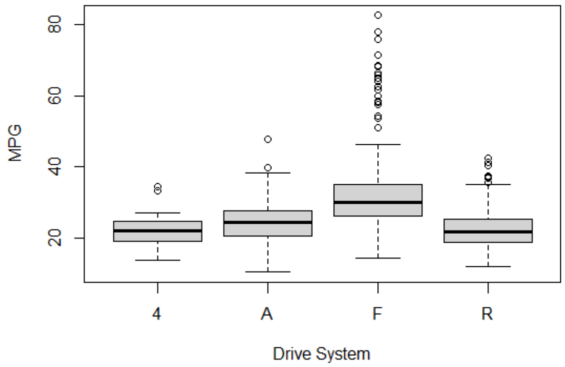
(From left to right: Weight, Cylinders, Gears, Displacement, Horsepower, AxleRatio, NVratio)

These plots seem to align with the outputs of the correlation coefficients--The first five variables appear to have some sort of linear relationship while AxleRatio and NVratio do not, so those two can be omitted from the model at the very least. However, we can also see that most of these plots violate at least one of the conditions for a linear model. This means that some transformations will be required before we consider them as predictors.

Next, we look at boxplots for each of the categorical variables:



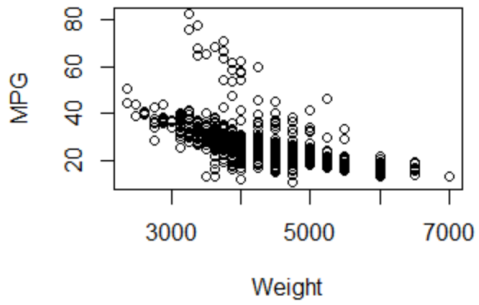
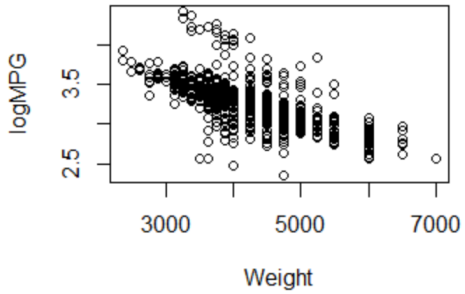
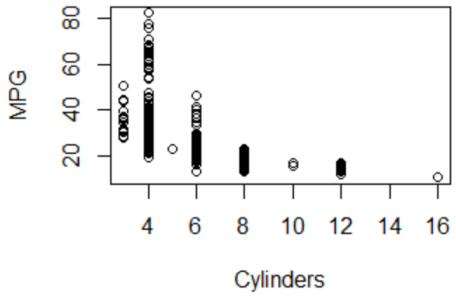
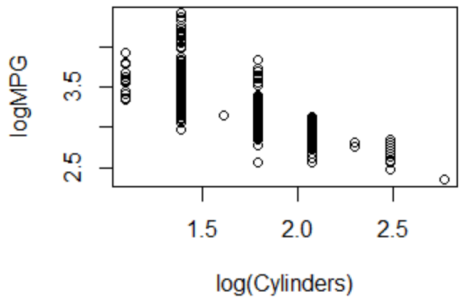
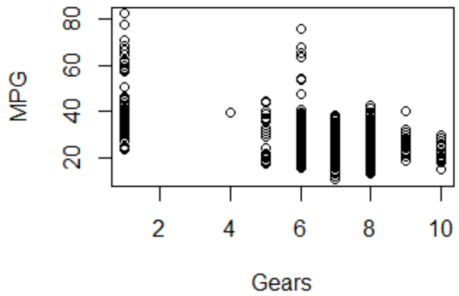
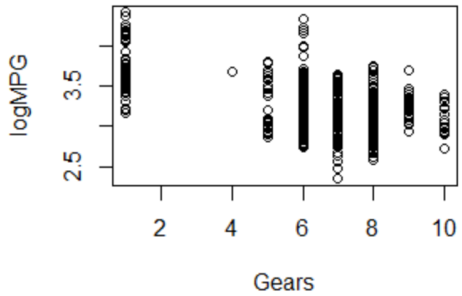
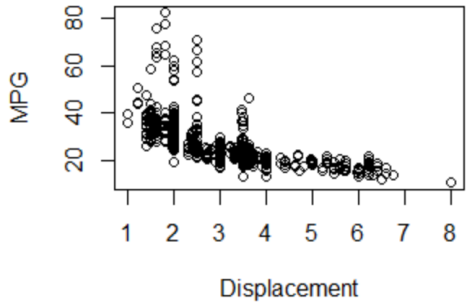
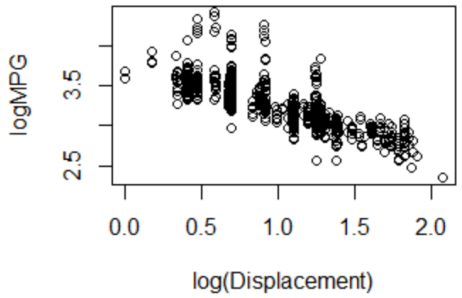




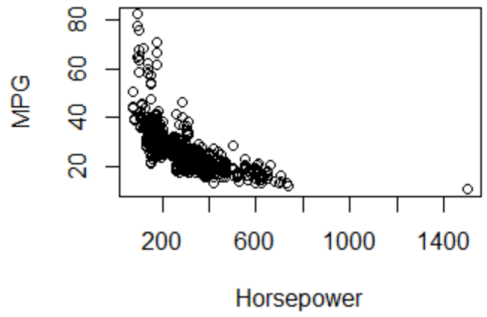
All of these plots seem to suggest that MPG seems to vary between each category. Therefore, we will need to make the three variables compatible with our linear model to assess whether or not there is a significant relationship between them and MPG.

Transformations:

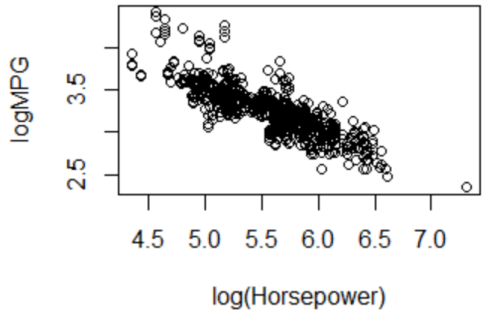
I chose to do a logarithmic transformation of MPG. Since this is the outcome, I will be predicting logMPG instead of the variable MPG. Luckily, each of the potential predicting variables seemed to have benefited from this transformation, as the correlations drastically improved between the variables. I then opted to do logarithmic transformations of Cylinders, Displacement, and Horsepower. I did not do any transformations for Weight or Gears, since it did not improve linearity or the correlations that much.

Before:	After:
Correlation Coefficient: -0.5796093 	Correlation Coefficient: -0.6664316 
Correlation Coefficient: -0.6151101 	Correlation Coefficient: -0.764058 
Correlation Coefficient: -0.4620066 	Correlation Coefficient: -0.4829639 
Correlation Coefficient: -0.6399918 	Correlation Coefficient: -0.7822278 

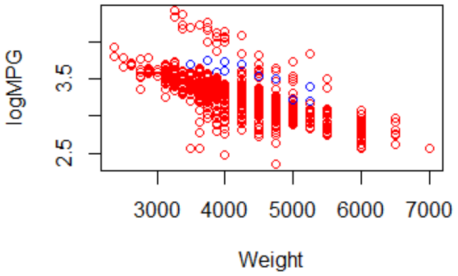
Correlation Coefficient: -0.664964

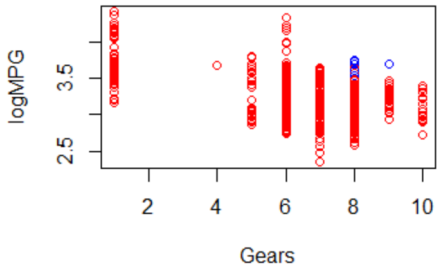


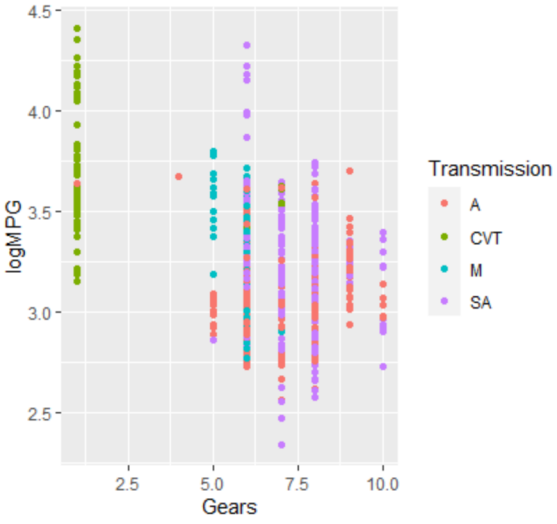
Correlation Coefficient: -0.823113

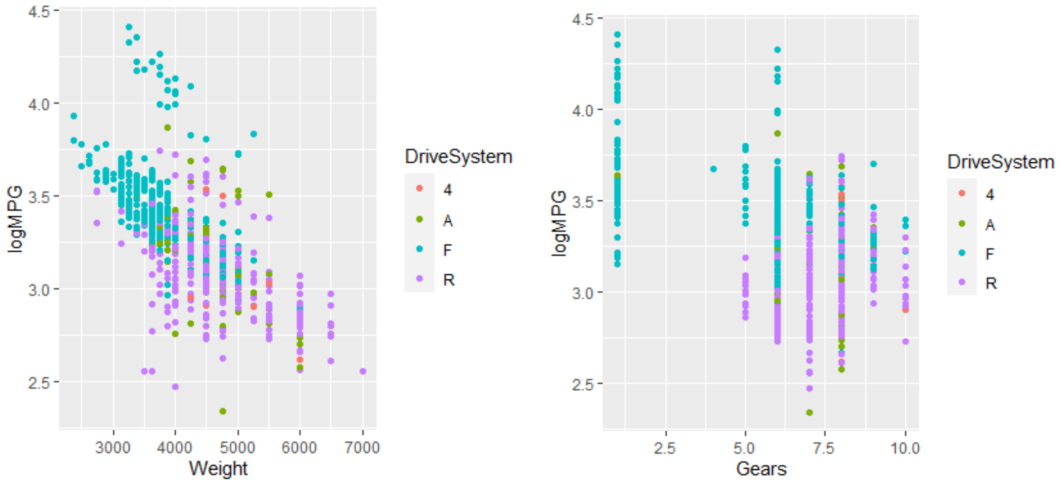


Although this solved the majority of problems with linearity in some of these correlations, there still appears to be some issues with the correlation with Weight and Gears. This may be Due to an interaction with a categorical variable, so I explored this further with the following plots:





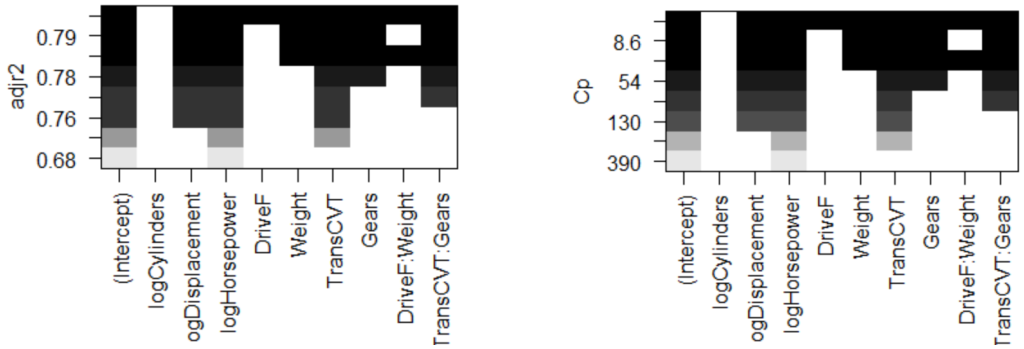




These plots show that although there does not appear to be an interaction between FuelType for Weight or Gears, Transmission and Drive System potentially could. Therefore, we will create two new interaction variables-- WeightDriveF (Weight*DriveSystemF) and GearTransCVT (Gears*TransmissionCVT). We will see in the next section if these provide any significant contributions to the model.

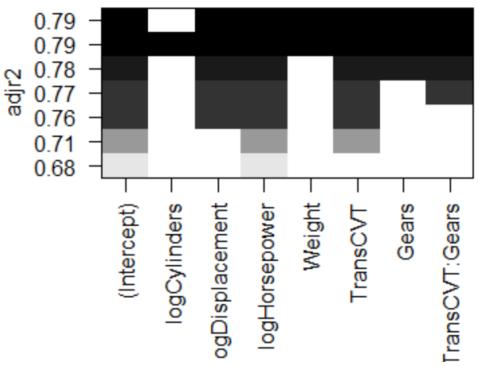
Variable Selection:

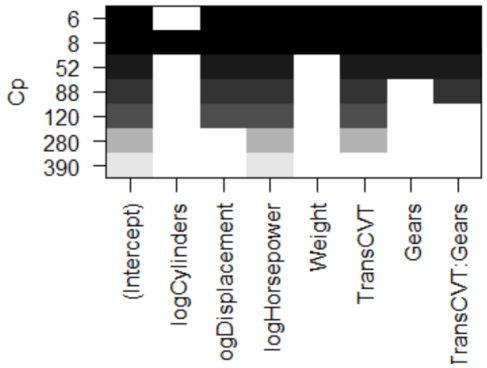
In order to select the variables for our model, I used a "Best Subsets" method to figure out which predictors I should include within my model. Since there are only 7 variables we will be looking at (Weight, Gears, logCylinders, logDisplacement, logHorsepower, GearTrans, and WeightDrive and only 700 rows of data, this was not particularly computationally intensive for my computer to handle. The criteria I used to determine which variables to use as predictors were Mallow's Cp, and $R^2_{adj.}$



These plots show that all except for logCylinders, each of these variables was useful in predicting the outcome of logMPG. This is quite a few variables to include and I was worried

about potentially overfitting the model, so I also created a summary of all those variables together. This did result in a slightly lower $R^2_{adj.}$ value, but it also resulted in a lower Mallow's Cp.





Although DriveF and its interaction with Weight are less than a p-value of 0.05, I decided to omit them since they were the largest p-values in the table. This left us with this as the summary:

```
> summaryHH(all)
                    model p   rsq  rss adjr2    cp   bic stderr
1                      lH 2 0.678 19.9 0.677 387.4  -779  0.169
2               lH-TrCVT 3 0.711 17.8 0.710 276.5  -850  0.160
3            lD-lH-TrCVT 4 0.757 15.0 0.756 124.1  -964  0.147
4     lD-lH-TrCVT-TCVT: 5 0.768 14.3 0.767  89.5  -990  0.143
5   lD-lH-TrCVT-G-TCVT: 6 0.780 13.6 0.778  52.7 -1019  0.140
6 lD-lH-W-TrCVT-G-TCVT: 7 0.794 12.7 0.792   7.0 -1059  0.135
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.241e+00  1.030e-01  50.883  < 2e-16 ***
logDisplacement -1.905e-01  2.544e-02  -7.489 2.12e-13 ***
logHorsepower   -3.225e-01  2.326e-02 -13.865  < 2e-16 ***
Weight          -6.586e-05  9.535e-06  -6.907 1.12e-11 ***
TransCVT         5.568e-01  3.953e-02  14.084  < 2e-16 ***
Gears            3.702e-02  4.974e-03   7.444 2.91e-13 ***
TransCVT:Gears  -8.530e-02  9.296e-03  -9.176  < 2e-16 ***
```

Final Model:

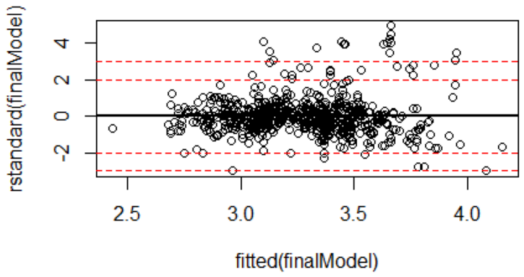
$$\widehat{logMPG} = 5.24 - 0.191(logDisplacement) - 0.323(logHorsepower) - 0.0000656(Weight) + 0.0557(TransCVT) + 0.037(Gears) - 0.0853(TransCVT * Gears)$$

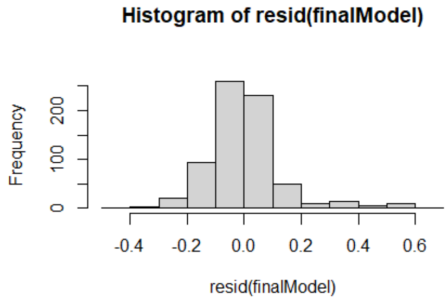
To compare the effectiveness of this model to a single variable model, I created an ANOVA table:

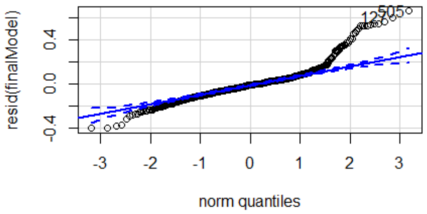
```
Analysis of Variance Table

Response: logMPG
                Df Sum Sq Mean Sq   F value    Pr(>F)
logDisplacement   1 37.664  37.664 2055.5431 < 2.2e-16 ***
logHorsepower     1  6.057   6.057  330.5879 < 2.2e-16 ***
Weight            1  0.672   0.672   36.6964 2.264e-09 ***
TransCVT          1  2.751   2.751  150.1180 < 2.2e-16 ***
Gears             1  0.169   0.169    9.2409  0.002456 **
TransCVT:Gears    1  1.543   1.543   84.2007 < 2.2e-16 ***
Residuals       693 12.698   0.018
---
```

Diagnostic Plots:







The heteroscedasticity in the residual plot looks relatively good. However, there does appear to be some issues with normality, as the histogram is slightly skewed right and the normal quantile plot deviates drastically along the tails. This can be seen also in the residual plot, where there are multiple variables above 3 standard deviations away from the mean. This could be an issue

with potential outliers, or perhaps I should have opted for different transformations for some of my variables.

Application and Prediction/Confidence Intervals:
In addition, use your final model to predict the gas mileage for a Toyota Rav4 which weighs 3875 pounds, has 4 cylinders, 6 gears (a 6-speed transmission), displacement of 2.494 liters, 176 horsepower, an axle ratio of 4.07, an N/V ratio of 30.6, an automatic transmission, front-wheel drive, and uses unleaded gas.

$$\widehat{logMPG} = 5.24 - 0.191(0.396) - 0.323(2.245) - 0.0000656(3875) + 0.0557(0) + 0.037(6) - 0.0853(0)$$

$$\widehat{logMPG} = 4.408748$$

$$MPG = 82.166$$

95% confidence interval: $\widehat{logMPG} : (4.279871, 4.537625)$ $\widehat{MPG} : (72.231, 93.468)$

95% prediction interval: $\widehat{logMPG} : (4.113378, 4.704117)$ $\widehat{MPG} : (61.152, 110.401)$