Project #1

Bailey Perry *ID: 4895940 perry520umn.edu*

Due: February 12, 2014

1. Data Pre-Processing

This data set provides data for variables related to fuel use and income throughout the United States. The goal is to identify how taxes affect fuel consumption and how fuel consumption varies over through the different states. The data set includes the following variables:

- `Drivers`: State Number of licensed drivers in the state
- `FuelC`: Gasoline sold for road use, thousands of gallons
- `Income`: Per person personal income for the year 2000, in thousands of dollars
- `Miles`: Miles of Federal laid highway miles in the state
- `MPC`: Estimated miles driven per capita
- `Pop`: 2001 population age 16 and over
- `Tax`: Gasoline state tax rate, cents per gallon

It is important to note that it is known that MPC has a masking effect on the effect due to taxes on Fuel Consumption, this means that MPC will not be included in the models.

The following command shows the data types of the variables in the dataframe in R:

```
require(alr3)
```

```
## Loading required package: alr3
```

```
## Loading required package: car
```

```
lapply(fuel2001, class)
```

```
## $Drivers
## [1] "integer"
##
## $FuelC
## [1] "integer"
##
## $Income
## [1] "integer"
##
## $Miles
## [1] "integer"
##
## $MPC
## [1] "numeric"
```

```
##
## $Pop
## [1] "integer"
##
## $Tax
## [1] "numeric"
```

```
fuel2001$Miles <- log(fuel2001$Miles)
fuel2001$Drivers <- fuel2001$Drivers*1000/fuel2001$Pop
fuel2001$FuelC <- fuel2001$FuelC*1000/fuel2001$Pop
```

As recommended, the miles variable was replaced by its logarithm before completing any further data analysis. Drivers and FuelC were also manipulated to help make comparisons. Since state size would have an effect on these values, we divided them by the population to create comparable ratios. The variables were also multipled by 1000 to remove decimals, also making the unit for FuelC into gallons.

2. Graphical and Numerical Exploration of the Data

a) Summary, Boxplots, and Histograms:

Let us do a quick numerical summary of the variables:

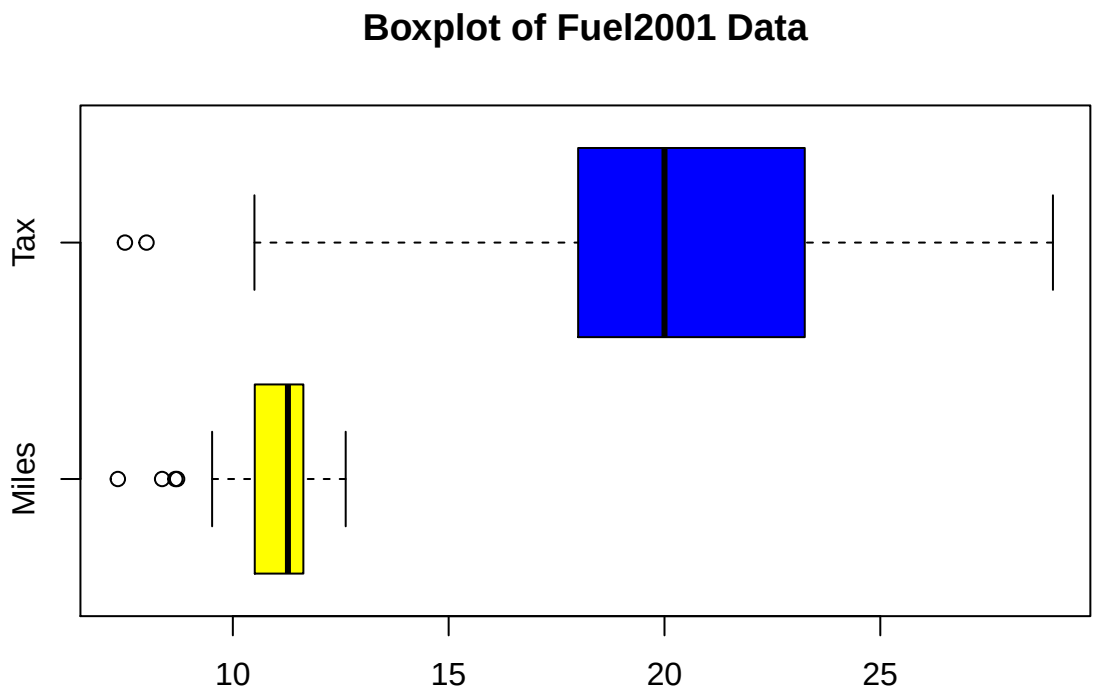
```
summary(fuel2001)
```

```
##     Drivers           FuelC           Income          Miles
##  Min.   : 700.2   Min.   :317.5   Min.   :20993   Min.   : 7.336
##  1st Qu.: 864.1   1st Qu.:575.0   1st Qu.:25323   1st Qu.:10.507
##  Median : 909.1   Median :626.0   Median :27871   Median :11.276
##  Mean   : 903.7   Mean   :613.1   Mean   :28404   Mean   :10.914
##  3rd Qu.: 943.0   3rd Qu.:666.6   3rd Qu.:31209   3rd Qu.:11.634
##  Max.   :1075.3   Max.   :842.8   Max.   :40640   Max.   :12.614
##       MPC             Pop               Tax
##  Min.   : 6556   Min.   :  381882   Min.   : 7.50
##  1st Qu.: 9391   1st Qu.: 1162624   1st Qu.:18.00
##  Median :10458   Median : 3115130   Median :20.00
##  Mean   :10448   Mean   : 4257046   Mean   :20.15
##  3rd Qu.:11311   3rd Qu.: 4845200   3rd Qu.:23.25
##  Max.   :17495   Max.   :25599275   Max.   :29.00
```

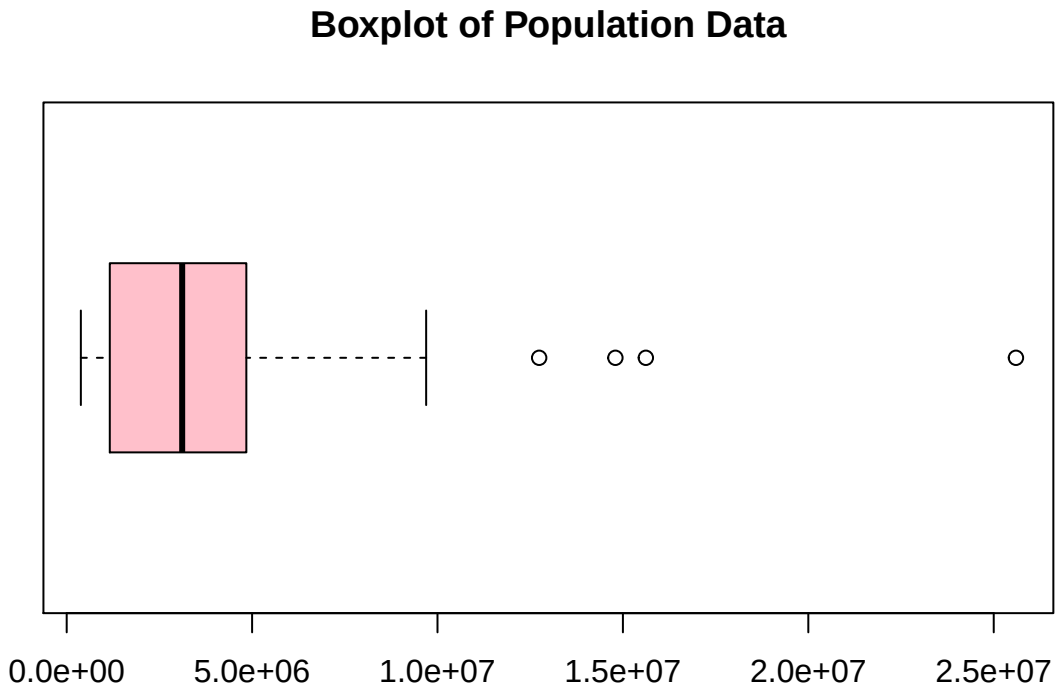
Looking at the summaries and comparing means and medians we can determine which variables are skewed. Based on the output, Pop seems to have the largest skew of all the variables since the mean and median are quite different.

Next let us do box-plots of the quantitative variables:

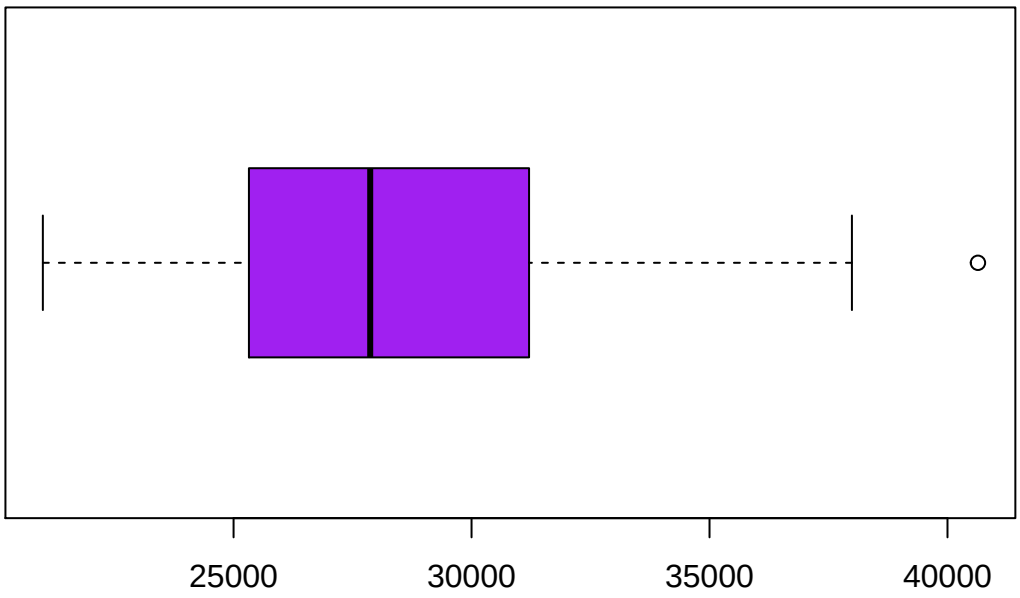
```
with( fuel2001, boxplot(Miles, Tax, horizontal = TRUE, names=c("Miles", "Tax"), col=c("yellow", "blue")
```



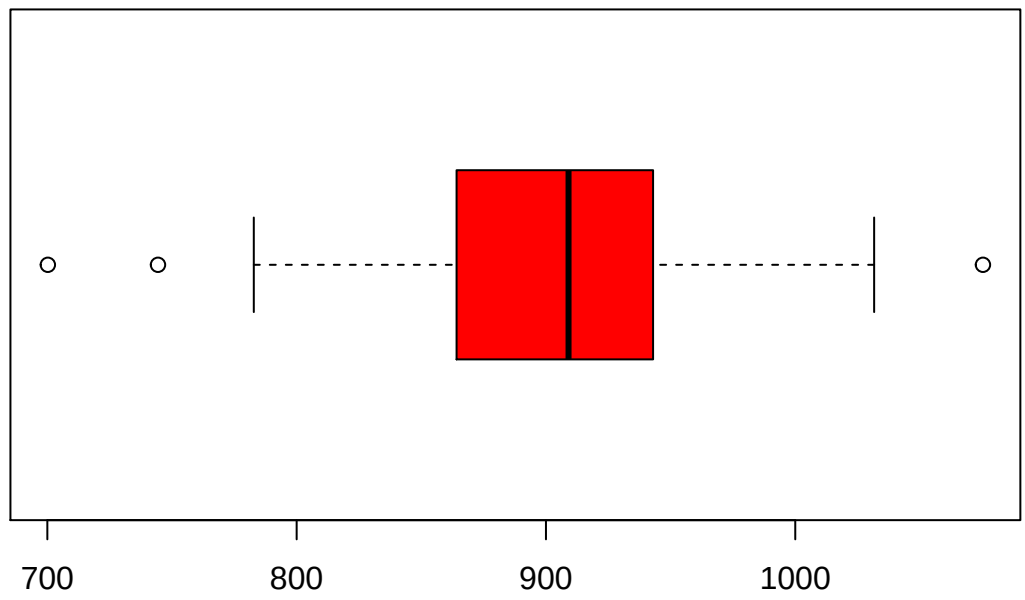

```
boxplot(fuel2001$Pop, names=c("Pop"), col="pink", horizontal= TRUE, main =
          "Boxplot of Population Data")
```



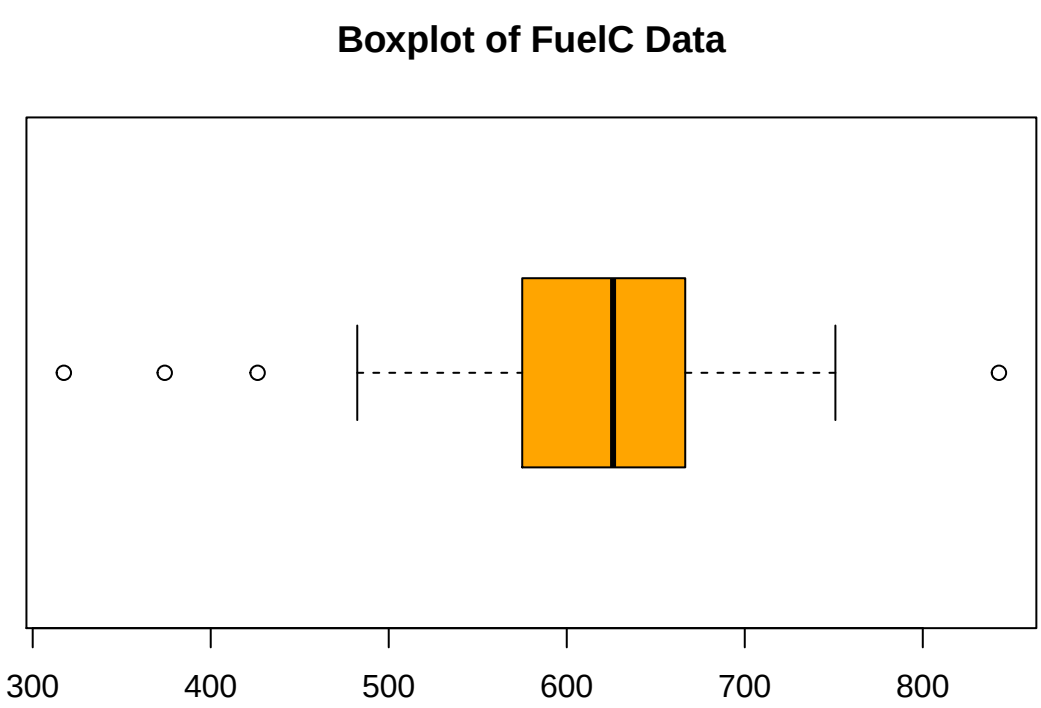

```
boxplot(fuel2001$Income, names=c("Income"), col="purple", horizontal = TRUE, main
        = "Boxplot of Income Data")
```



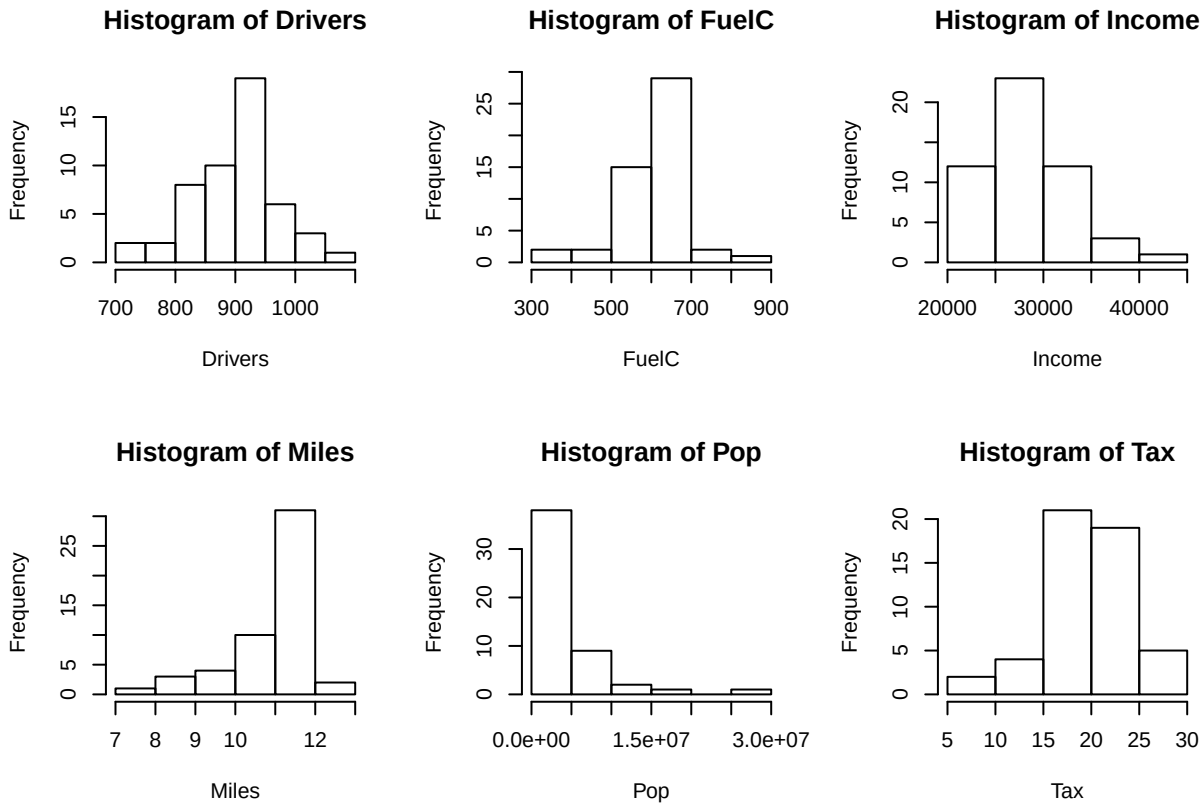

```
boxplot(fuel2001$Drivers, names=c("Drivers"), col="red", horizontal = TRUE, main=
          "Boxplot of Drivers Data")
```




```
boxplot(fuel2001$FuelC, names=c("FuelC"), col="orange", horizontal = TRUE, main =
          "Boxplot of FuelC Data")
```




```
par(mfrow = c(2,3))
with(fuel2001, hist(Drivers))
with(fuel2001, hist(FuelC))
with(fuel2001, hist(Income))
with(fuel2001, hist(Miles))
with(fuel2001, hist(Pop))
with(fuel2001, hist(Tax))
```



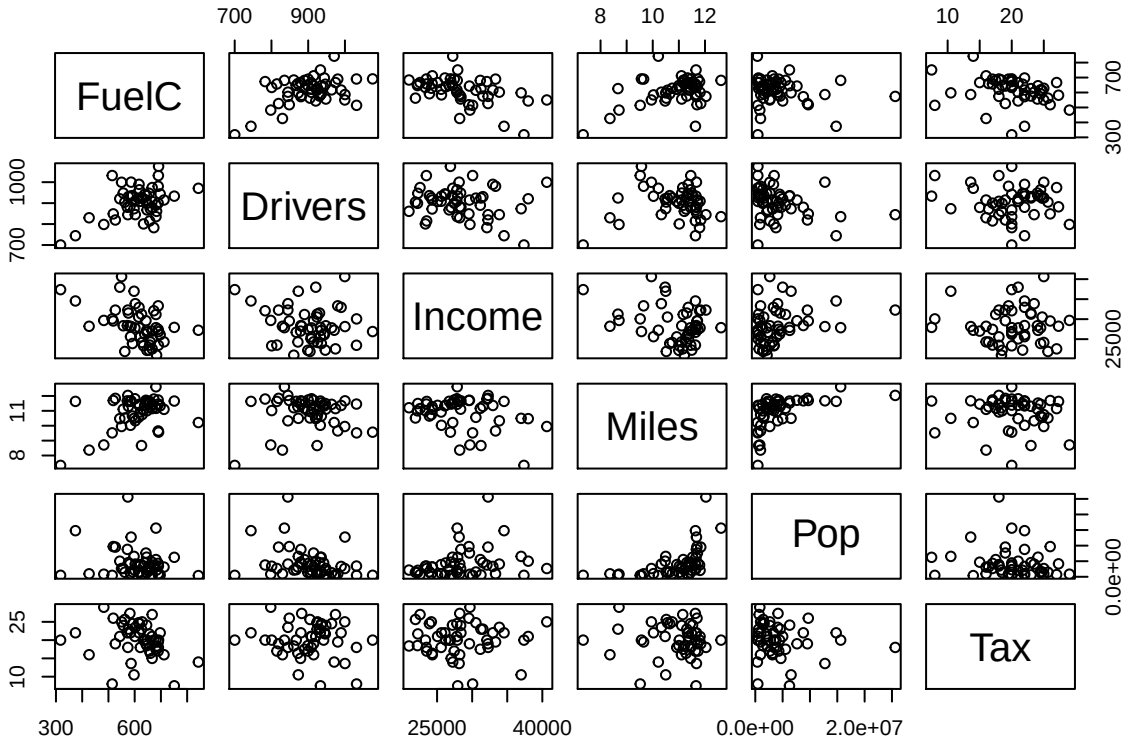
Since the variables all had different magnitudes it was necessary to separate many of the boxplots so that an accurate view of the data could be obtained. These boxplots allow us to see the skew and any possible outliers.

The histograms give us an even better view of the skew in each of the particular variables. We can see that Drivers is mostly symmetric and FuelC is somewhat symmetric. The five remaining variables all suggest some sort of skew: Population is heavily skewed to the right, and Miles is heavily skewed to the left. Income is slightly skewed to the right, and Tax is slightly skewed to the left.

b) Scatterplots and Correlation:

Just like we did the scatterplot for the linear regression, we will do a similar plot here as well:

```
with(fuel2001, pairs(FuelC ~ Drivers + Income + Miles + Pop + Tax))
```




```
with(fuel2001, cor(cbind(Drivers, FuelC, Income, Miles, Pop, Tax)))
```

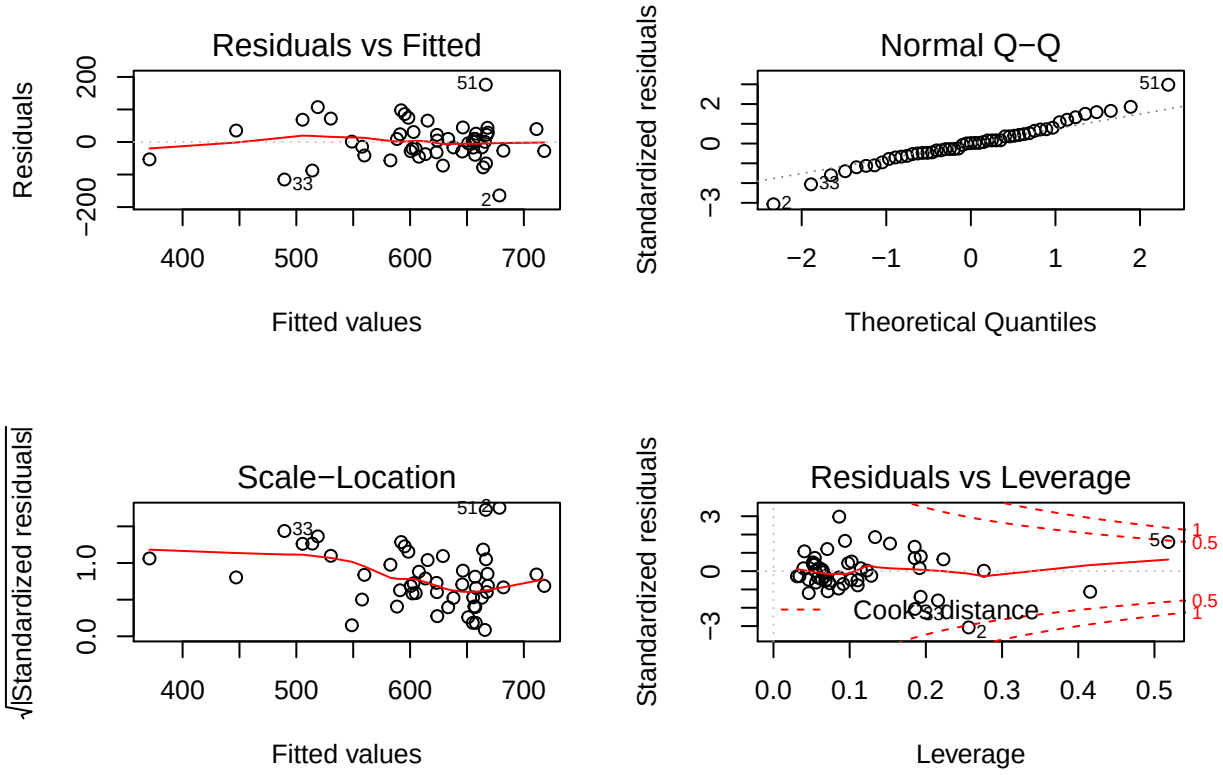
```
##              Drivers      FuelC      Income       Miles        Pop
## Drivers  1.00000000  0.4685063 -0.17596063  0.03059068 -0.2868008
## FuelC    0.46850627  1.0000000 -0.46440498  0.42203233 -0.1639005
## Income  -0.17596063 -0.4644050  1.00000000 -0.29585136  0.2650850
## Miles    0.03059068  0.4220323 -0.29585136  1.00000000  0.5066833
## Pop     -0.28680081 -0.1639005  0.26508498  0.50668329  1.0000000
## Tax     -0.08584424 -0.2594471 -0.01068494 -0.04373696 -0.1458658
##                 Tax
## Drivers -0.08584424
## FuelC   -0.25944711
## Income  -0.01068494
## Miles   -0.04373696
## Pop     -0.14586581
## Tax      1.00000000
```

From the scatterplots, we can give a basic analysis of the correlation between variables. Miles and Pop have a moderate positive correlation. It is not strong enough to suggest a masking effect though. Miles and Drivers both have moderate positive correlations with the FuelC variable.

3. Run a Model:

The preliminary model is done without transforming any of the variables because it is first necessary to see whether the assumptions for linear regression are already met.

```
m1 <- lm(fuel2001$FuelC ~ fuel2001$Drivers + fuel2001$Income + fuel2001$Miles +
           fuel2001$Pop + fuel2001$Tax)
par(mfrow = c(2,2))
plot(m1)
```




```
summary(m1)
```

```
##
## Call:
## lm(formula = fuel2001$FuelC ~ fuel2001$Drivers + fuel2001$Income +
##     fuel2001$Miles + fuel2001$Pop + fuel2001$Tax)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -164.242  -30.506    1.196   29.708  176.263
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.588e+01  1.952e+02   0.081 0.935522
```

```
## fuel2001$Drivers  3.723e-01  1.299e-01   2.867 0.006286 **
## fuel2001$Income  -3.445e-03  2.389e-03  -1.442 0.156205
## fuel2001$Miles    4.473e+01  1.177e+01   3.801 0.000431 ***
## fuel2001$Pop     -6.258e-06  2.675e-06  -2.340 0.023799 *
## fuel2001$Tax     -5.105e+00  1.974e+00  -2.586 0.013017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.95 on 45 degrees of freedom
## Multiple R-squared:  0.5636, Adjusted R-squared:  0.5151
## F-statistic: 11.62 on 5 and 45 DF,  p-value: 3.059e-07
```

Interpretation: The graphs show that there may be some points that are skewing our results so it is necessary to perform outlier tests to see if their are values we need to transform to create a better fitting model. We also note that the Residuals v Fitted plot is only slightly skewed, but fits the assumption pretty well. The QQ Plot showed variation in its tails from the normal QQ line. These convey that the data may not fit the assumptions well enough, and it may be necessary to transform the data. Based on the summary output, the R-squared value = 0.5636, which isn't particularly high so we will compare that to the other models.

```
require(car)
outlierTest(m1)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferonni p
## 2 -3.41993          0.0013619     0.069457
```

{Interpretation:} Since the outlier test showed the point number 2, with a p-value of 0.069, we can conclude that the value does not have a low enough p-value to decide that it is actually an outlier. The output also tells us directly "No studentized residuals".

4. Transforming the Variables:

Based on the plots above, we will perform a transformation to see if transforming the variables creates a better match with the assumptions.

```
require(alr3)
trans <- (powerTransform(cbind(FuelC, Drivers, Income, Miles, Pop, Tax) ~ 1,
                         fuel2001))
summary(trans)
```

```
## bcPower Transformations to Multinormality
##
##         Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
## FuelC      1.8626   0.5728           0.7399           2.9853
## Drivers    1.6435   1.3573          -1.0168           4.3039
## Income    -0.1808   0.7734          -1.6967           1.3351
## Miles      4.6712   1.2228           2.2745           7.0679
## Pop        0.1426   0.0996          -0.0527           0.3379
```

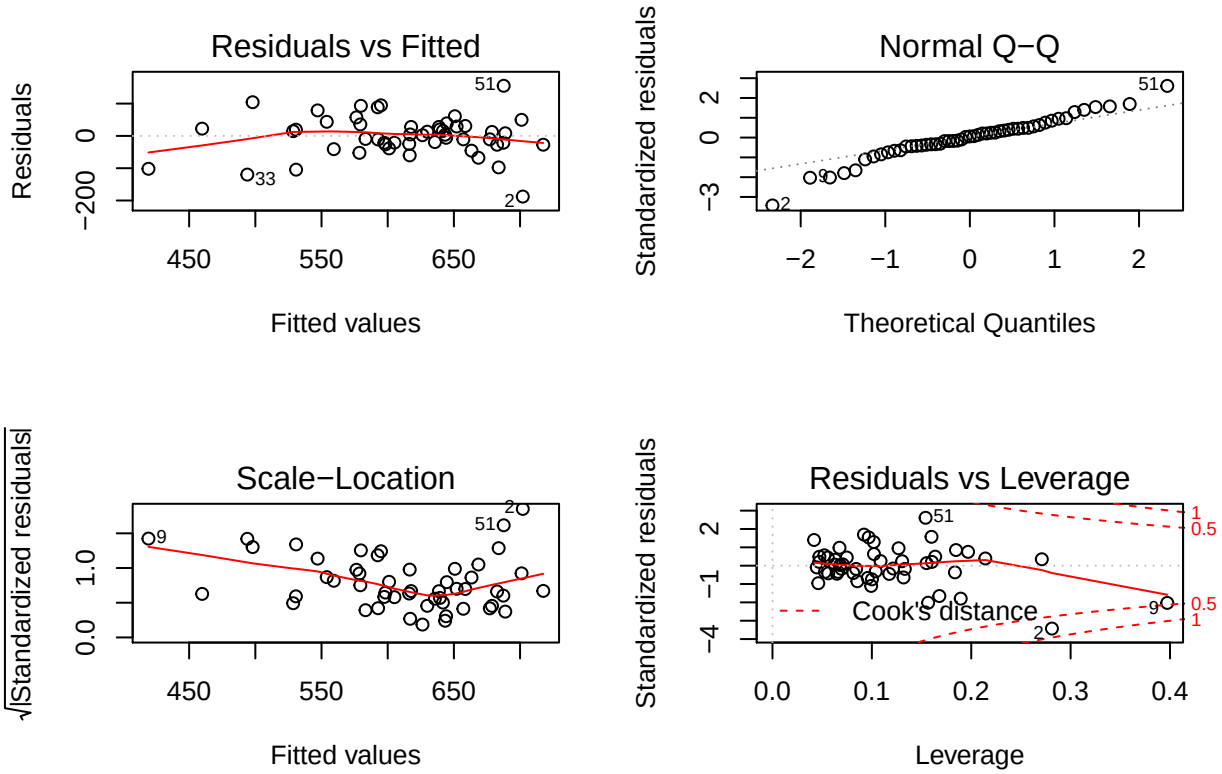
```
## Tax        1.8866   0.4549           0.9951           2.7782
##
## Likelihood ratio tests about transformation parameters
##                                          LRT df         pval
## LR test, lambda = (0 0 0 0 0 0)     49.78719  6 5.186226e-09
## LR test, lambda = (1 1 1 1 1 1)    101.31001  6 0.000000e+00
## LR test, lambda = (1 1 1 4.67 0 1)  11.05855  6 8.658374e-02
```

Interpretation: Based on the p-values derived from the transformation, it shows that all of the transformations have p-values that are less than 0.05, except for the power transformation. We know that if the transformation p-value is less than 0.05, we should NOT choose to use that transformation. In the next section we will implement the power transformation.

5. Fitting the Model with Transformed Variables:

a) Power Transformation Model:

```
fuel2001.T <- with(fuel2001, data.frame(FuelC=FuelC, Drivers=Drivers,
                                        Income=Income, NewMiles=(Miles^5), NewPop=log(Pop), Tax=Tax))
m2 <- lm(FuelC ~ Drivers + Income + NewMiles + NewPop + Tax, data=fuel2001.T)
par(mfrow= c(2,2))
plot(m2)
```



Now that we have transformed the variables with the fitting transformation, we want to look at the summary of

the resulting linear regression model and the new plots of the data. It is necessary to see if the transformation has helped make the assumptions be met.

```
summary(m2)
```

```
##
## Call:
## lm(formula = FuelC ~ Drivers + Income + NewMiles + NewPop + Tax,
##     data = fuel2001.T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -187.780  -26.752    3.616   29.879  155.156
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.862e+02  2.493e+02   2.753  0.00849 **
## Drivers      4.592e-01  1.333e-01   3.446  0.00124 **
## Income      -4.087e-03  2.567e-03  -1.592  0.11833
## NewMiles     7.648e-04  2.583e-04   2.961  0.00488 **
## NewPop      -2.740e+01  1.548e+01  -1.771  0.08340 .
## Tax         -4.708e+00  2.049e+00  -2.297  0.02633 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.62 on 45 degrees of freedom
## Multiple R-squared:  0.5251, Adjusted R-squared:  0.4723
## F-statistic: 9.951 on 5 and 45 DF,  p-value: 1.856e-06
```

After running this model, we see that the R squared value decreased compared to model 1, which suggests that the model did not improve our variables. The Normal QQ Plot did not show much of a stronger linear trend, with no less variation in the tails. The Residuals v Fitted plot still does not meet the assumption perfectly, but the data is still spread evenly around the line. Overall, model 2 does not show much improvement from the first model.

We do not need to run an outlier test for model 2, because it is already clear that the model does not fit as well as the first model did.

b) Removing Variables:

When considering implementing a third model, it was noted that there were multiple explanatory variables and one useful thing would be to decrease that number.

But, since none of the variable had a strong enough correlation, it is better to keep them in the model than remove them. They may have some impact on explaining variation in the data. Thus, it is solely a decision between using model 1 or model 2 (power transformed).

6. Conclusions:

Based on the two previous implemented models and corresponding analysis, Model 1 is chosen as the best model. This model did not implement any transformation for the variables. The result of this model was the highest R-squared value as compared to model 2, and the closest fitting graphs for meeting the assumptions, although the assumptions were not met perfectly.

```
summary(m1)
```

```
##
## Call:
## lm(formula = fuel2001$FuelC ~ fuel2001$Drivers + fuel2001$Income +
##     fuel2001$Miles + fuel2001$Pop + fuel2001$Tax)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -164.242  -30.506    1.196   29.708  176.263
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.588e+01  1.952e+02   0.081 0.935522
## fuel2001$Drivers  3.723e-01  1.299e-01   2.867 0.006286 **
## fuel2001$Income  -3.445e-03  2.389e-03  -1.442 0.156205
## fuel2001$Miles    4.473e+01  1.177e+01   3.801 0.000431 ***
## fuel2001$Pop     -6.258e-06  2.675e-06  -2.340 0.023799 *
## fuel2001$Tax     -5.105e+00  1.974e+00  -2.586 0.013017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.95 on 45 degrees of freedom
## Multiple R-squared:  0.5636, Adjusted R-squared:  0.5151
## F-statistic: 11.62 on 5 and 45 DF,  p-value: 3.059e-07
```

When reviewing the summary, the significant variables are any of the variables with p-values less than 0.05. Based on the model 1 summary, it is known that the significant variables are Drivers, Miles, Pop, and Tax.

Based on the Coefficient Estimate values, it is good to analyze what the coefficients represent. First we want to analyze the effect tax has on FuelC, since that was the main objective of this project. In terms of tax, for every one cent increase in tax, the fuel consumption decreases by approximately 5 gallons per person. It is known that the variable decrease occurs per person, because the FuelC variable was modified to be in units of gallon/person before any models were implemented.

For every one unit increase in Drivers, there is an increase in FuelC by 0.37 gallons. In terms of Miles of federal Highway, for every unit increase in Miles there is a 45 gallon increase in FuelC. Finally, for every unit increase in Population, there is an incrementally small (0.000006) decrease in FuelC.

It is noted that it is not generally practical, or necessary, to include the Population variable in the model because Population was implemented in some of the other variables. But, it was noted as acceptable by Professor Nandy for this first project.