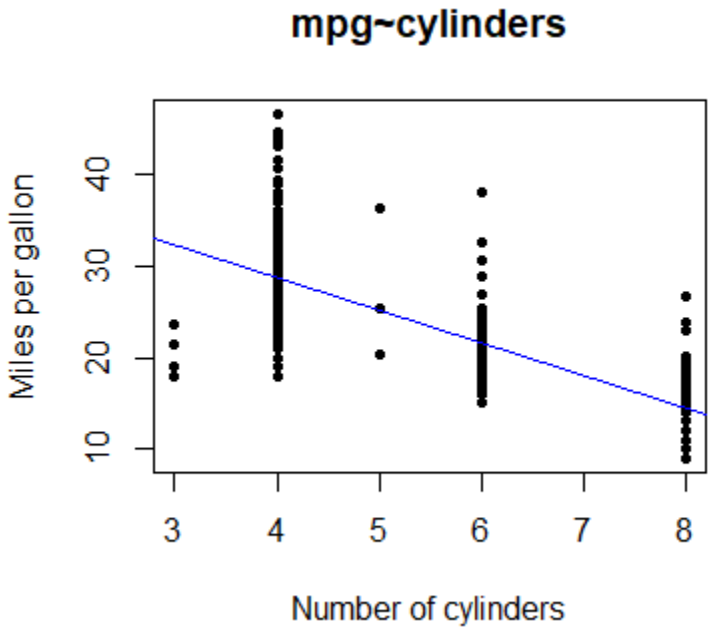
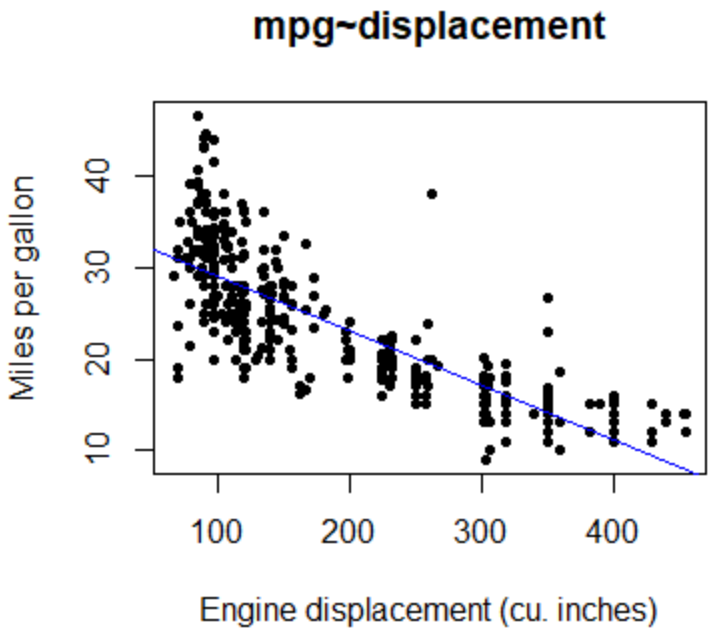
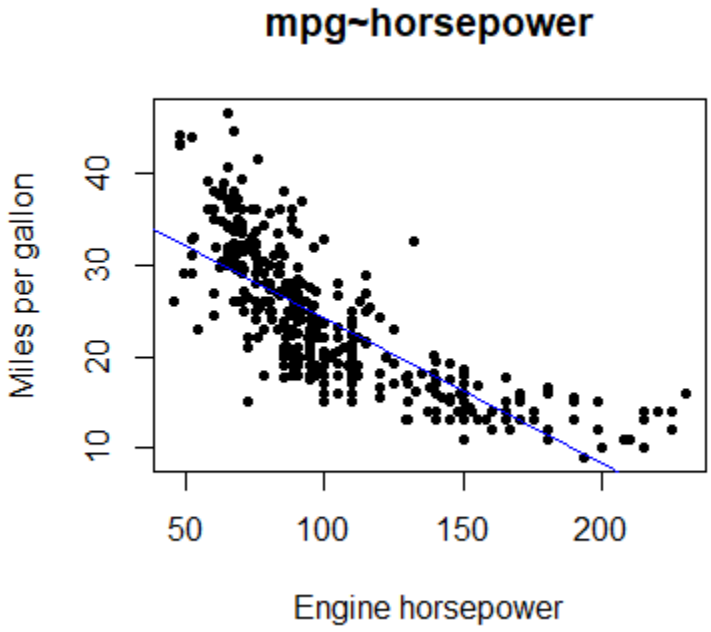
Question1)

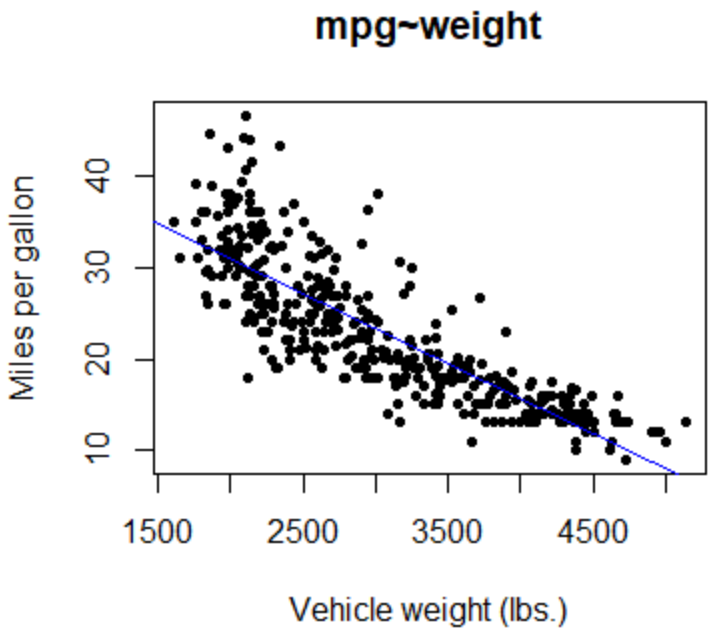
The 'weight' variable will provide the best prediction since its value for R-squared value (0.6926) is the highest among all other variables. The R-squared value for the given data is:

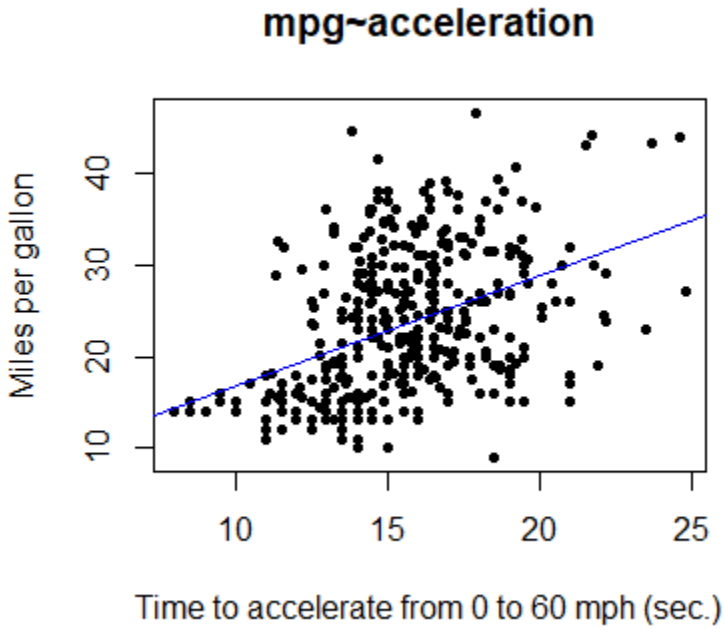
Cylinders	0.6047
Displacement	0.6482
Horsepower	0.6059
Weight	**0.6926**
Acceleration	0.1792
Year	0.3370
Origin	0.3195

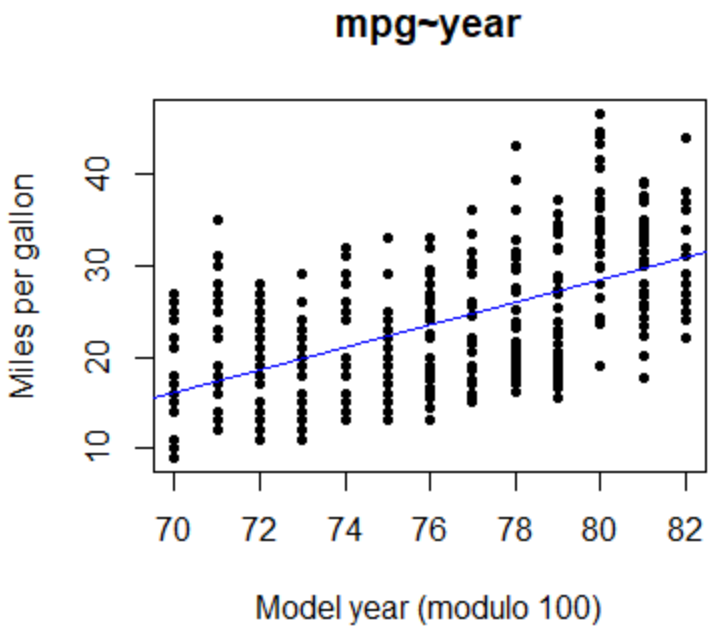


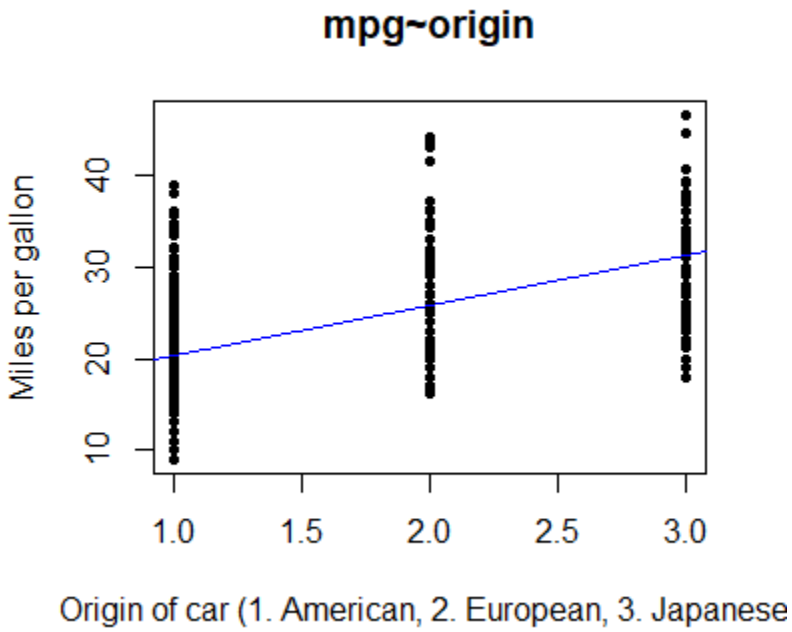

mpg~displacement
Miles per gallon
10
20
30
40
100
200
300
400
Engine displacement (cu. inches)


mpg~horsepower
Miles per gallon
10
20
30
40
50
100
150
200
Engine horsepower


mpg~weight
Miles per gallon
10
20
30
40
1500
2500
3500
4500
Vehicle weight (lbs.)


mpg~acceleration
Miles per gallon
10
20
30
40
10
15
20
25
Time to accelerate from 0 to 60 mph (sec.)





b)

The resulting multiple regression equation for the model is:

mpg= -17.218435 + (-0.493376) cylinders + (0.019896) displacement + (-0.016951) horsepower + (-0.006474) weight + (0.080576) acceleration + (0.750773) year + (1.426141) origin

We can compare the values of coefficients by running ztests on the values of slopes and coefficients of individual regression models with the multiple regression model.

Question 2)

The training accuracy can be found by using Confusion Matrix.

Confusion Matrix gives us a 2*2 matrix of the Positive/Negative values of the Predicted model and the Positive/Negative values of the Reference model.

In our case, the test.data is used as the reference model which consists of actual values of the "Direction" variables whereas the predicted_val is used as the predicted model.

Running confusion matrix on these two values for each data point gives us the number of True Positives/ True Negatives/ False Positives / False Negatives.

Accuracy can be found by taking the ratio of True Positives + True Negatives / Total number of outcomes, where True Positives are the number of predicted outcomes that are actually positives (Up) and True Negatives are the number of outcomes that are actually negative (Down)

Output of the confusionMatrix function is as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160   1
         1   0 152

               Accuracy : 0.9968
                 95% CI : (0.9823, 0.9999)
    No Information Rate : 0.5112
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9936
 Mcnemar's Test P-Value : 1

            Sensitivity : 1.0000
            Specificity : 0.9935
         Pos Pred Value : 0.9938
         Neg Pred Value : 1.0000
             Prevalence : 0.5112
         Detection Rate : 0.5112
   Detection Prevalence : 0.5144
      Balanced Accuracy : 0.9967

       'Positive' Class : 0
```

The Accuracy for the computed model is **0.9968**

Question 3)

Error Analysis: We can analyze the deviation of the values in the model using anova function. For the variables whose deviation is significantly less can be eliminated to improve the accuracy. The deviation analysis for our model is given as follows:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Direction

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev
NULL                     936     1295.7
Lag1    1     2.41       935     1293.3
Lag2    1     0.83       934     1292.5
Lag3    1     0.07       933     1292.4
Lag4    1     0.02       932     1292.4
Lag5    1     0.03       931     1292.4
Volume  1     0.13       930     1292.2
Today   1  1292.24       929        0.0
```

Since, the deviance of Lag2, Lag3, Lag4, Lag5, Volume is significantly lower and close to zero, we can eliminate those variables and compute the model using Lag1 and Today. After Computation, the confusion matrix output is as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 161   0
         1   0 152

               Accuracy : 1
                 95% CI : (0.9883, 1)
    No Information Rate : 0.5144
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.5144
         Detection Rate : 0.5144
   Detection Prevalence : 0.5144
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

In this way, we can improve the accuracy to **1**. This is better than the previous accuracy which was **0.9968 .**

We can use some of the following method to improve the accuracy in general models where there is more scope for improvement:

Normalization: We can normalize all the variables to avoid the predicted values to be dominated by a single variable. In our question, the summary of the dataset is as follows:

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
1	2001	0.381	-0.192	-2.624	-1.055	5.010	1.1913	0.959	Up
2	2001	0.959	0.381	-0.192	-2.624	-1.055	1.2965	1.032	Up
3	2001	1.032	0.959	0.381	-0.192	-2.624	1.4112	-0.623	Down
4	2001	-0.623	1.032	0.959	0.381	-0.192	1.2760	0.614	Up
5	2001	0.614	-0.623	1.032	0.959	0.381	1.2057	0.213	Up
6	2001	0.213	0.614	-0.623	1.032	0.959	1.3491	1.392	Up

Since the Year is of a much different magnitude than others, we can normalize the independent variables either by eliminating the “Year” column or standardize all the variables.

Question 4)

The resulting regression equation of the big data set is:

y= 1.001 + 0.6 x

When we divide the data into the given 5 samples, the resulting equations respectively are:

$$Y = 1.0000474 + 0.5995133\ x$$

$$Y = 1.0000564 + 0.5996065\ x$$

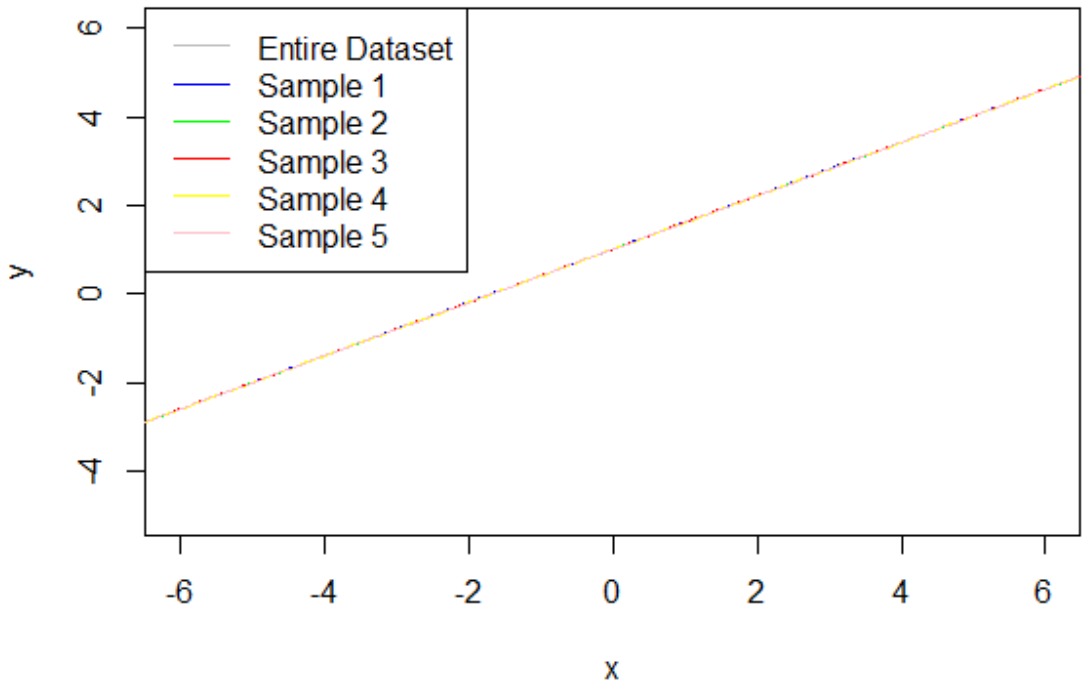
$$Y = 1.0002622 + 0.6003393\ x$$

$$Y = 0.9996778 + 0.5996503\ x$$

$$Y = 1.0001878 + 0.5999795\ x$$

From the above equation generated from the five different samples, we can observe that the value of intercept and x- coefficient remains almost the same. If we plot the lines on a graph, we notice that

these lines are linear in nature. The value of y linearly increases with that of x.



Question 5)

The Z-test function is implemented using the 5-step Hypothesis procedure:

Step 1: State the hypotheses and identify the claim.

Step 2: Find the critical value(s) from the appropriate table.

Step 3: Compute the test value.

Step 4: Make the decision to reject or not reject the null hypothesis.

Step 5 Summarize the results

In the function, we calculate the critical values using the qnorm function available in R. We then compute the test value using the z-test formula:

Z= sample mean- population mean / (std.deviation / squareroot(sample size))

Then, we compare the z-value with critical value and reject the hypothesis if the z-value is present in the critical region.

Question 6)

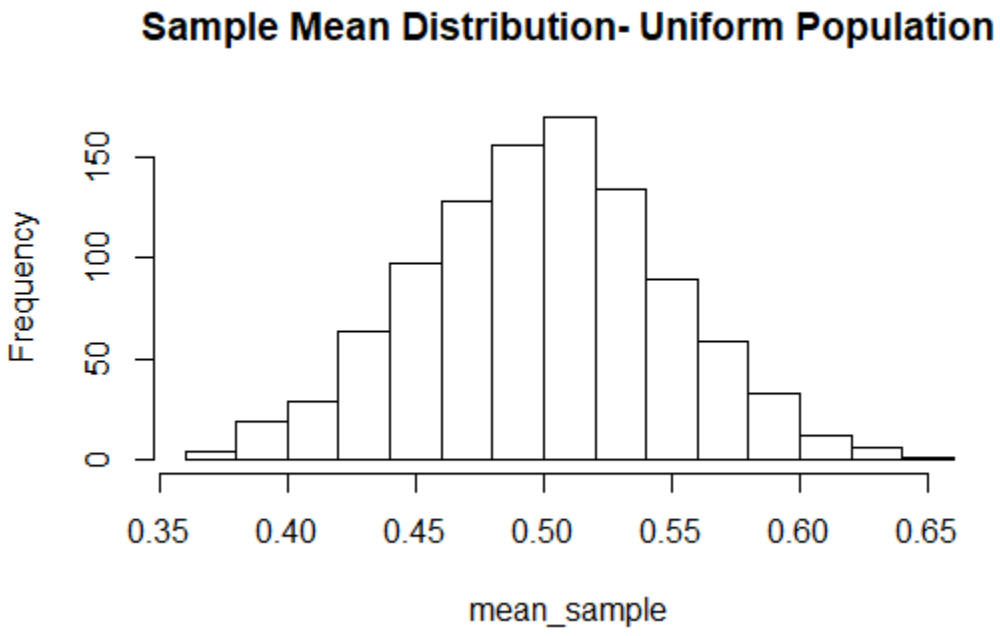
Central Limit Theorem states that:

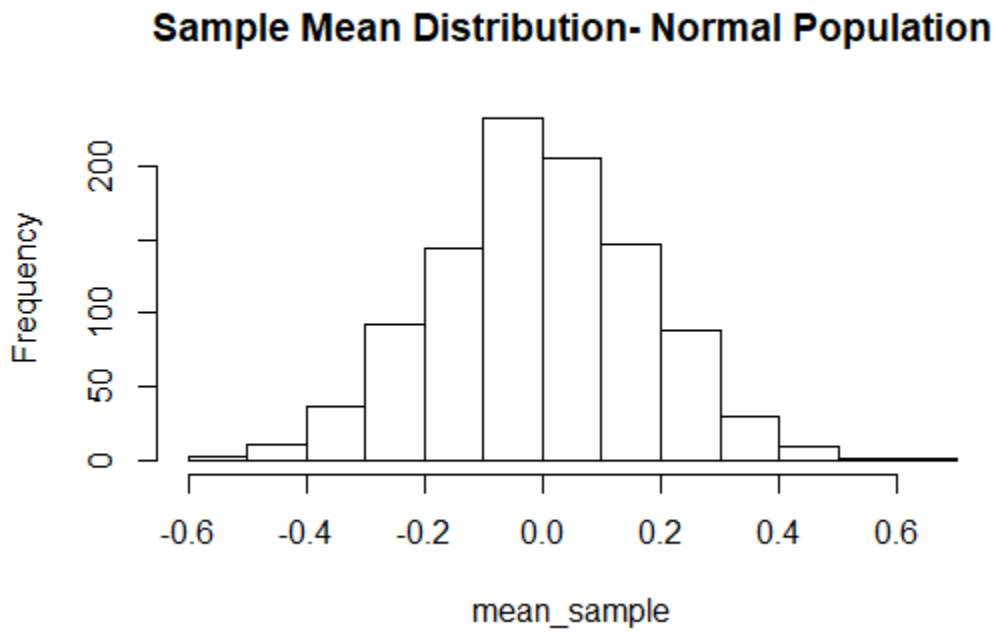
For a given population with a finite mean and a non zero finite variance, the sampling distribution of the mean approaches a normal distribution with a mean as same as the population mean and has a standard error as the sample size increases.

This is true regardless of the nature of the parent population. The sampling distribution will always approach normal distribution.

In the R code, the nature population distribution could be taken as a uniform distribution or normal distribution. Based on the distribution, random variables are generated. Then, samples of a given sample size are generated from the population. The mean of each of these samples are stored in a variable. We then calculate the population mean and std deviation. Next, the mean and standard deviation of the samples is calculated.

We observe that, the population mean and the average of the mean of samples is equal. This supports the Central Limit Theorem. We then, plot the distribution of the means of samples. The plots are as follows:





We observe that irrespective of the nature of the population (Uniform/ Normal), the distributions of the mean of samples approach normal distribution. Hence, we can use this to demonstrate the Central Limit Theorem.