Part 1

Find the mean, standard deviation, and range for all the features.

Table 1 Mean, Standard Deviation, and Range of each feature

	cyl	dis	hor	wei	acc
Mean	5.47	194.41	104.47	2977.58	15.54
Standard Deviation	1.71	104.64	38.49	849.4	2.76
Range	3 - 8	68 - 455	46 - 230	1613 - 5140	8 - 24.8

Part 2

Display the histograms of all the features together with their normal density functions.

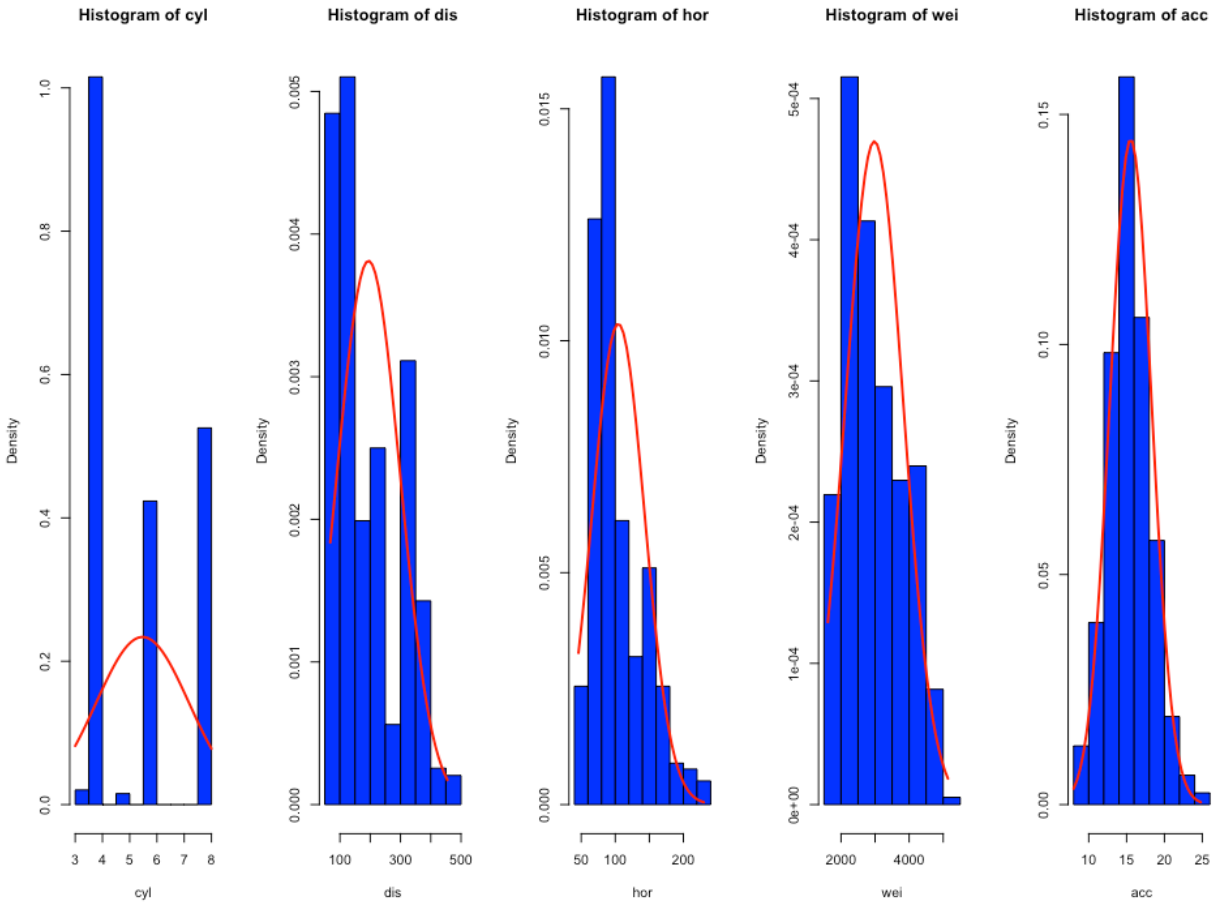

Figure 1 Histogram with density of all the features and normal density function with mean and standard deviation respectively

From the histograms, we can see that for Number of Cylinders, which is a categorical feature meaning the value is not continuous (3, 4, 5, 6, and 8), it will not follow normal distribution.

For the other features, they mostly follow normal distribution which is reasonable. In the car market, premium-class vehicle will have high displacement, high horsepower, high weight, and high acceleration. However, those also mean high cost which are not preferred by most customers. Most of the customers will purchase the middle-class cars which are cheaper and more durable.

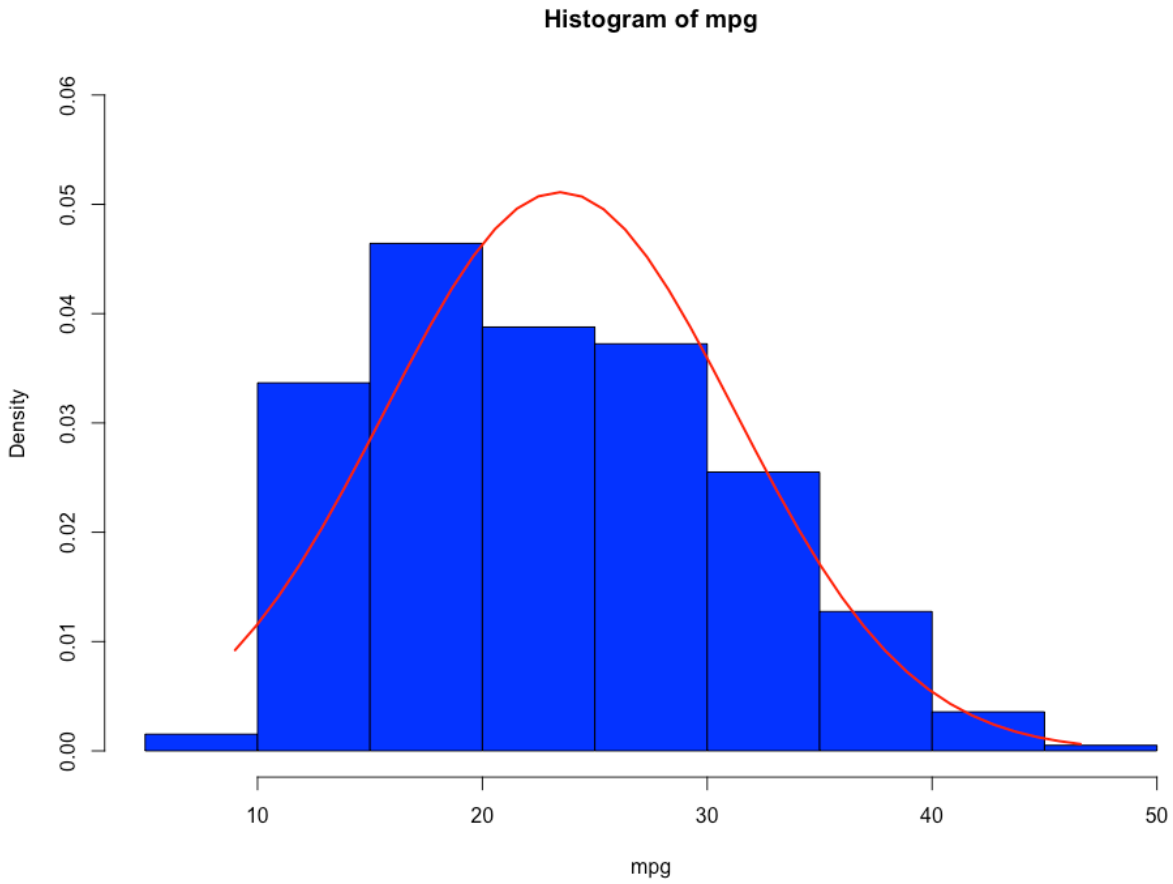

Figure 2 Histogram of mpg and normal density function with its mean and standard deviation

Histogram of mpg in Figure 3 also follows normal distribution. This is also reasonable since that most of the cars in the market has mpg around 15 to 30. Some vehicles will have lower mpg such as old vehicles which are less fuel-efficient or pick-ups which have heavy weight, so they consume lots of fuel. Due to the constraint of technology, it is very rare that engine can have mpg higher than 35. Even the vehicles meet mpg of 40, they will become costly.

Part 3

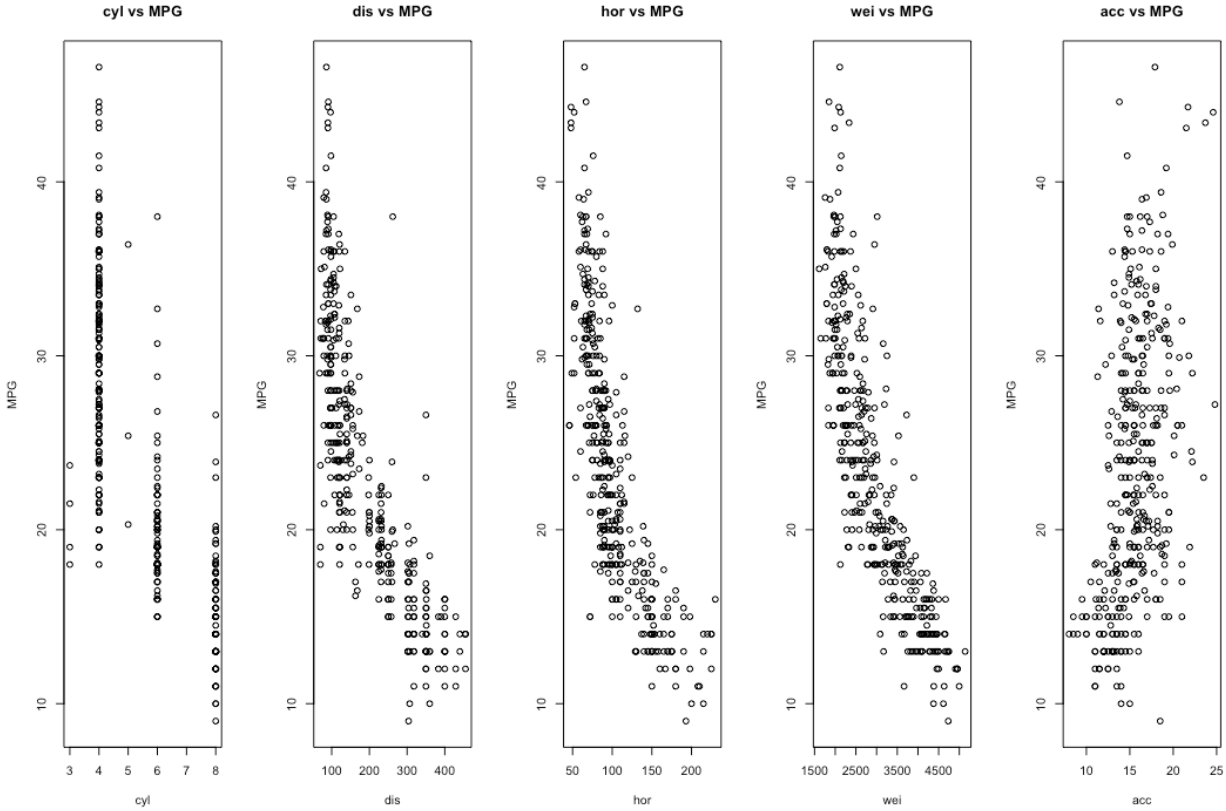

Figure 3 Scatter Plot of each feature vs MPG

From these scatter plots, we observe that displacement, horsepower, weight, and acceleration seem to have a liner relation with mpg. We can confirm these by the correlation table below.

Table 2 Correlation Table between each feature vs MPG

	(cyl , mpg)	(dis , mpg)	(hor , mpg)	(wei , mpg)	(acc , mpg)
Correlation	-0.78	-0.81	-0.78	-0.83	0.42

From the table, it is surprising that cylinder also has a strong correlation with mpg, since it is a categorical variable. However, it makes sense that cars with high number of cylinders do require more gas which means lower mpg i.e., negative correlation between number of cylinders and mpg. In addition, we can see that acceleration has less linear relation than all the other features. We can conclude that acceleration is not a good feature to do the linear regression model to estimate mpg since the correlation value is below 0.7.

Contrarily, number of cylinders, displacement, horsepower, and weight have a strong correlation with mpg i.e., absolute value of correlation bigger than 0.7. We believe that number of cylinders, displacement, horsepower, and weight can have strong capacity to predict mpg.

Part 4

Table 3 Correlation Matrix for all the features

	cyl	dis	hor	wei	acc
cyl	1.00	0.95	0.84	0.90	-0.50
dis	0.95	1.00	0.90	0.93	-0.54
hor	0.84	0.90	1.00	0.86	-0.69
wei	0.90	0.93	0.86	1.00	-0.42
acc	-0.50	-0.54	-0.69	-0.42	1.00

From the correlation matrix in Table 3, we can observe that except for the acceleration, all the other four features have strong correlation between each other. The highest correlation of all is between the number of cylinders and the displacement. This could be the result of this relation:

$$Displacement = stroke\ length\ \times \pi \times (\frac{1}{2} \times bore)^2 \times number\ of\ cylinders$$

Part 5

Quantile plot of MPG from Q1% to Q100%

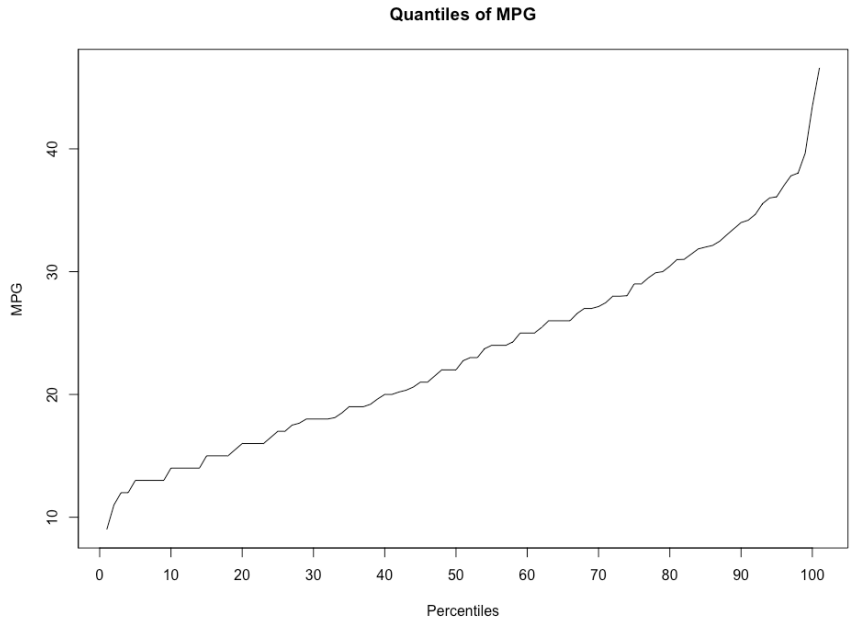

Figure 4 Percentile Plot for MPG

Part 6

Linear regressions for the 5 features versus MPG:

Table 4 Coefficient, Root Mean Squared Error, Relative Accuracy of Liner Model for each feature vs MPG

	cyl	dis	hor	wei	acc
A (Intercept)	42.92	35.12	39.94	46.22	4.83
B (Slope)	-3.56	-0.06	-0.16	-0.01	1.20
Root Mean Squared Error	4.91	4.64	4.91	4.33	7.08
Relative Accuracy	0.21	0.20	0.21	0.18	0.30

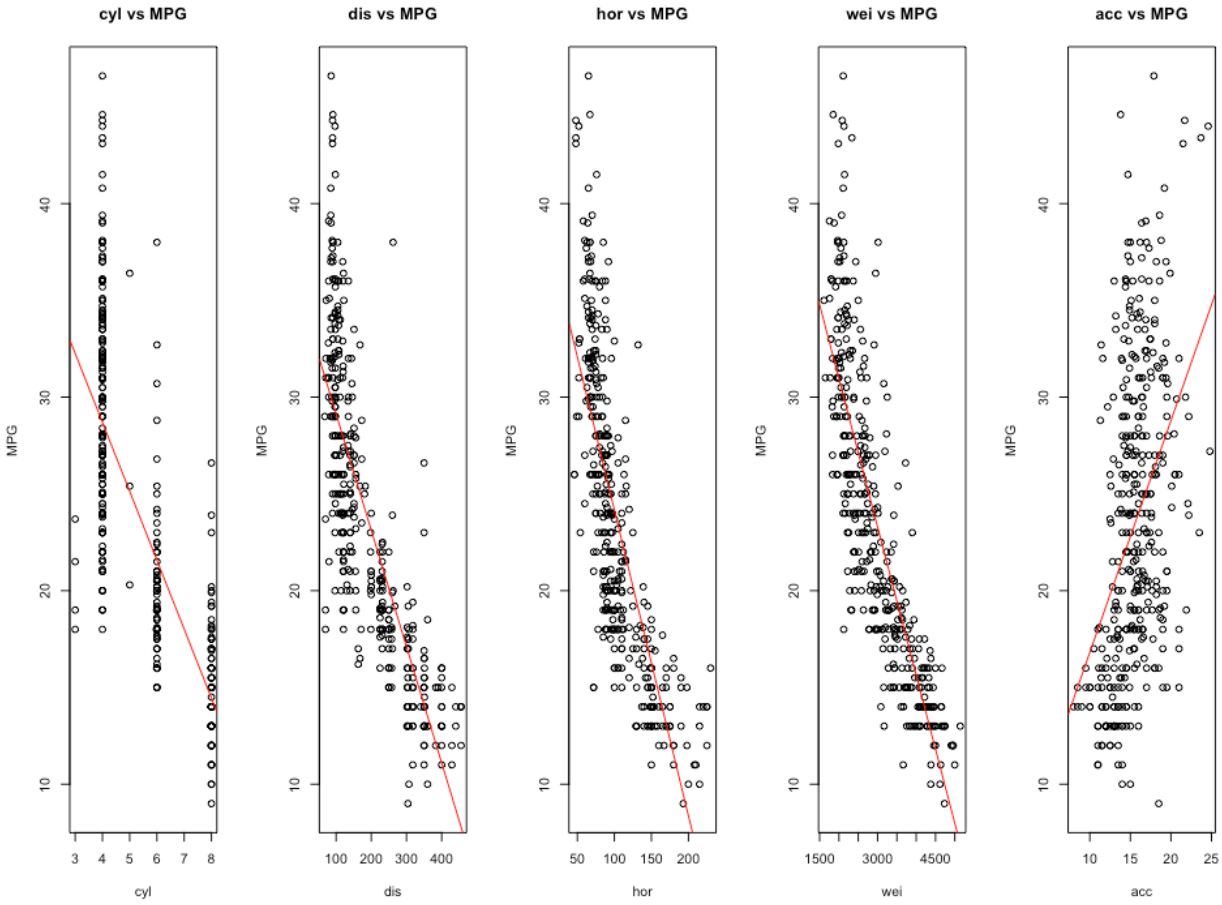

Figure 5 Scatter Plot with Liner Regression Line for each feature vs MPG

From Figure 6, we can observe that all the linear regression lines are pretty align with the scatter pattern. However, we need numerical value to prove that.

From Table 2 we know that weight has the highest correlation with mpg. The relative accuracy values in table 4 also support the idea that weight has significant linear relationship with mpg. Similarly, acceleration has a low correlation value with mpg and a poor relative accuracy, leading us to the conclusion that it would be the worst explanatory variable to use in a linear regression model.

Part 8

Please refer to the script at the end of this report.

Method: The Q33% equals to 18.503 and the Q66% equals to 26.6. We use these two quantiles to distinguish all the cases into three classes. Thus, we got 130 cases under LOWmpg, 130 cases under MEDmpg, and 132 cases under HIGHmpg.

Part 9

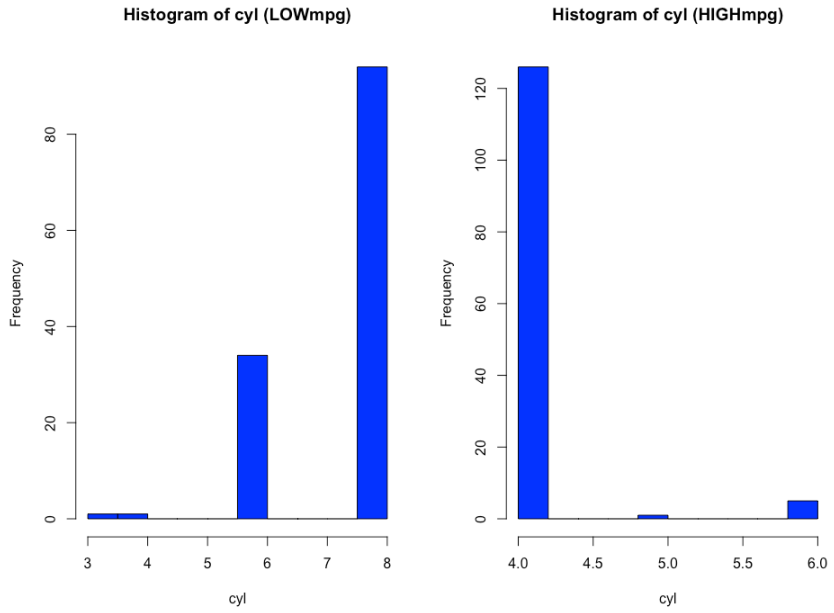

Figure 6 Left: Histogram of Cylinders in LOWmpg class. Right: Histogram of Cylinders in HIGHmpg class

The histograms in Figure 7 displaying Number of Cylinders of LOWmpg and HIGHmpg indicate significant levels of discrimination between the different mpgs. Number of Cylinders in cars with LOWmpg are typically at 6 or 8, while those for HIGHmpg cars are at 4.

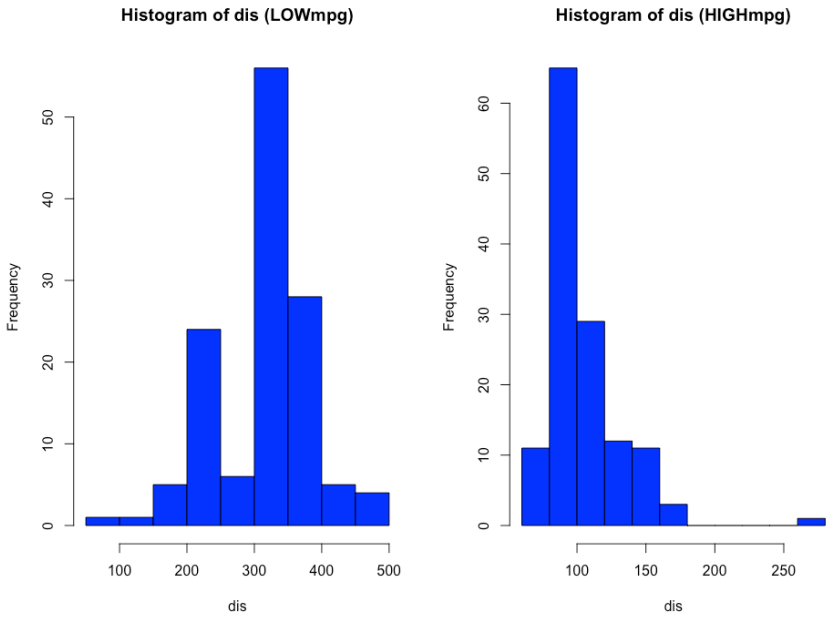

Figure 7 Left: Histogram of Displacement in LOWmpg class. Right: Histogram of Displacement in HIGHmpg class

The histograms in Figure 8 displaying displacement of LOWmpg and HIGHmpg indicate significant levels of discrimination between the two groups of mpgs. Displacement levels in cars with LOWmpg are typically in the 200-400 range, while the levels for HIGHmpg cars is around

100. The LOWmpg cars appear to have a left-skewed curve while the HIGHmpg is skewed to the right.

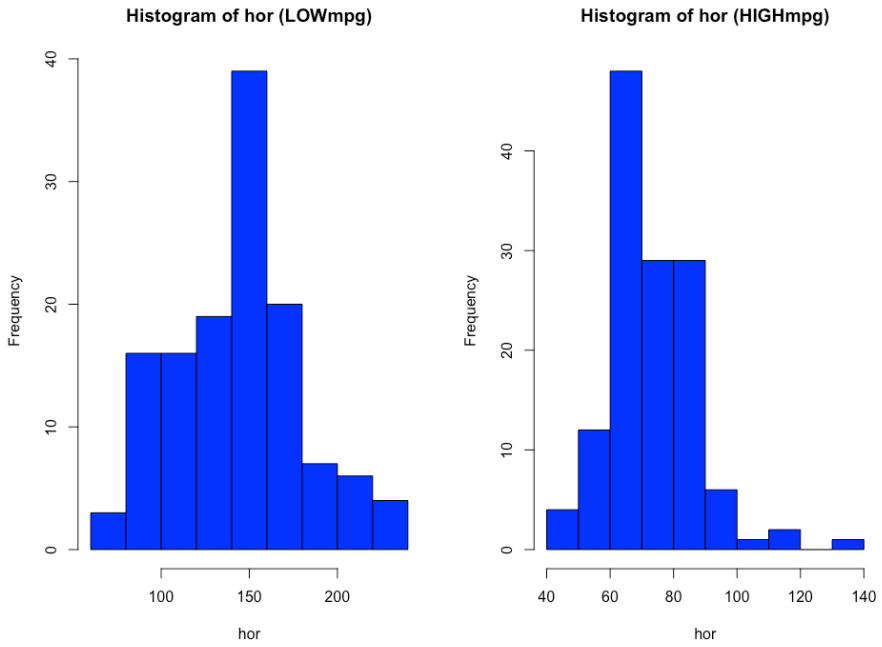

Figure 8 Left: Histogram of Horsepower in LOWmpg class. Right: Histogram of Horsepower in HIGHmpg class

The histograms in Figure 9 displaying Horsepower of LOWmpg and HIGHmpg indicate significant levels of discrimination between the different mpgs. Horsepower in cars with LOWmpg are typically above 100, while the levels for HIGHmpg cars are below 100. The LOWmpg cars appear to have a bell-shape curve while the HIGHmpg cars have a right-skewed distribution.

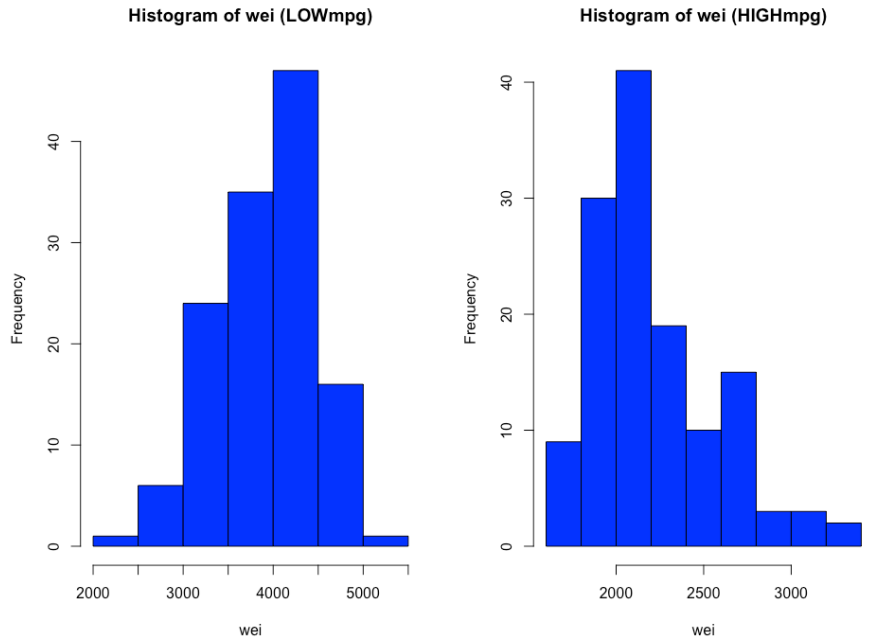

Figure 9 Left: Histogram of Weight in LOWmpg class. Right: Histogram of Weight in HIGHmpg class

The histograms of the weights of LOWmpg cars compared to HIGHmpg cars indicates that on average LOWmpg cars tend to weigh more than HIGHmpg cars. The shape of the curve for the LOWmpg cars is approximately normal with the typical weight being around 4,000 lbs. The HIGHmpg curve is skewed to the right with a typical weight of only 2,000 lbs. These differences lead us to believe that there will be discrimination between the classes.

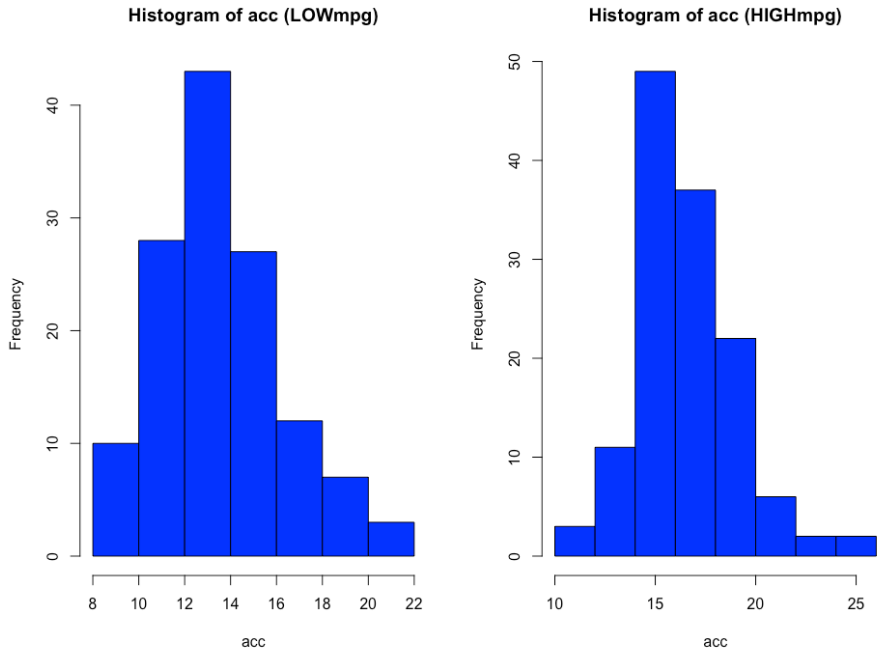

Figure 10 Left: Histogram of Acceleration in LOWmpg class. Right: Histogram of Acceleration in HIGHmpg class

The histograms in Figure 11 for the acceleration of the different mpg classes indicate that acceleration will have poor discrimination power. Both curves are approximately normal with similar means and standard deviations.

Part 10

Calculate the Mean, Standard Deviation, and Confidence Interval for each feature with two classes (LOWmpg and HIGHmpg).

Table 5 Mean, Standard Deviation, and Confidence Interval of each feature that classified as LOWmpg

	cyl	dis	hor	wei	acc
Mean	7.41	315.31	145.62	3937.33	13.78
Standard Deviation	1.01	71.11	35.84	557.19	2.65
Confidence Interval	[7.26 , 7.56]	[305.05 , 325.57]	[140.45 , 150.79]	[3856.95 , 4017.71]	[13.4 , 14.16]

Table 6 Mean, Standard Deviation, and Confidence Interval of each feature that classified as HIGHmpg

	cyl	dis	hor	wei	acc
Mean	4.08	106.4	74.39	2226.09	16.56
Standard Deviation	0.39	26.42	13.88	345.88	2.52
Confidence Interval	[4.02 , 4.14]	[102.62 , 110.18]	[72.4 , 76.38]	[2176.57 , 2275.61]	[16.2 , 16.92]

Comparing the confidence interval for both LOWmpg and HIGHmpg, we can find that there is no overlapping for all the features. However, the 90% confidence interval for acceleration are close between both classes even though they are not overlapping. We can conclude that acceleration has poor discrimination power. This can prove that our interpretation from the histograms is correct.

Part 11

Apply kNN classifier to the training and test set which are randomly selected with 80%/20% respectively. Fix the parameter k equals to 5.

Method: First, we use the training set for train and the training set to get the accuracy prediction. Then, we use the test set for test to get the accuracy prediction. Finally, compute the confusion matrix to better observe the prediction accuracy by the kNN in each class.

Table 7 Accuracy of Prediction for Training Set and Test Set when k = 5

AccTrain	AccTest
0.83	0.73

Table 8 Confusion Matrix for Training Set when k = 5. Global training performance = 83.8%

	Pred High	Pred Low	Pred Med
True High	0.833	0.000	0.167
True Low	0.000	0.939	0.061
Ture Med	0.148	0.111	0.741

Table 9 Confusion Matrix for Test Set when k = 5. Global test performance = 76.9%

	Pred High	Pred Low	Pred Med
True High	0.731	0.000	0.269
True Low	0.000	1.000	0.000
Ture Med	0.212	0.212	0.576

From the Accuracy Prediction table in Table 7, we can see that Accuracy Prediction of Training Set is above 0.8, a good prediction to the training set. Meanwhile, Accuracy Prediction of the Test Set is above 0.7 which is a good prediction as well.

From the Confusion Matrices in Table 8 and Table 9, we can observe that kNN classifier did a better prediction to classify between LOWmpg and HIGHmpg. There is no true LOWmpg case being classified as HIGHmpg case and vice versa. This is matching with the results we predicted from histograms in Problem 9 since they have very different distribution is case of all the features except acceleration.

However, the accuracy prediction of MEDmpg is lower. In some of the cases of true LOWmpg and HIGHmpg are predicted as MEDmpg and vice versa. This could happen due to that

MEDmpg have features fall between LOWmpg and HIGHmpg cases and cause kNN classifier will have difficulty to distinguish that.

Part 12

Perform the kNN classifier with same data set but different value of k.

Method: For each value k, we must implement kNN automatic classifier and calculate the prediction accuracies of the training and test sets. The optimal k will have the highest accuracy values for both the training and the test sets. Typically, an acceptable accuracy value for the test set would be around 70%.

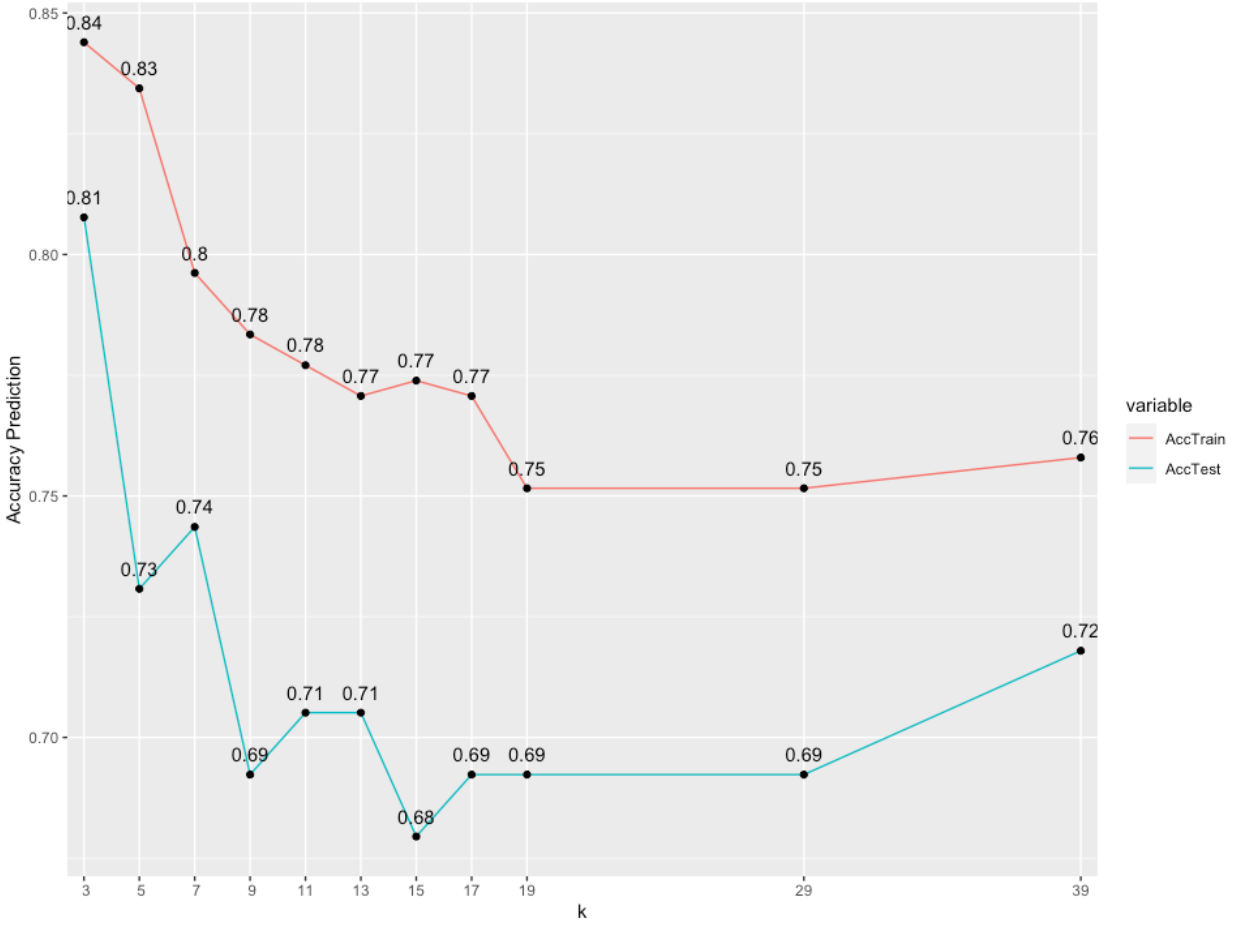

Figure 11 Accuracy Curve for both Training Set and Test Set when k = 3, 5, 7, 9, 11, 13, 15, 17, 19, 29, 39

In Figure 12, we observe that the accuracy predictions of the training and test sets are both at their highest when k is equal to three. As the value of k increases the accuracies begin to sharply decline. Increasing the value of k will lead to a model that is not very flexible and with low training accuracy, this will lead to bias in our model.

Table 10 Performance Table and Margin on Performance for Test Set when k = 3, 5, 7, 9, 11, 13, 15, 17, 19, 29, 39

k	Performance	Margin on Performance
3	0.831	0.042
5	0.769	0.048
7	0.771	0.048
9	0.719	0.051
11	0.738	0.050
13	0.744	0.049
15	0.720	0.051
17	0.738	0.050
19	0.745	0.049
29	0.732	0.050
39	0.752	0.049

From Table 10, we observe that k = 3 has the highest performance as well as the lowest margin on performance. However, it must be noted that the size of our data set is quite small which will lead to very unstable results when doing data analysis.