7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0. The equation to calculate Euclidean distance between two vectors u = (u1; u2; : : : ; un) and v = (v1; v2; : : : ; vn) is

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_n - v_n)^2}.$$

The Euclidean distance between Observation 1 and the test point is:

$\sqrt{[(0 - 0)^2 + (3 - 0)^2 + (0 - 0)^2]} = \sqrt{3^2} = 3.00$

The Euclidean distance between Observation 2 and the test point is:

$\sqrt{[(2 - 0)^2 + (0 - 0)^2 + (0 - 0)^2]} = \sqrt{2^2} = 2.00$

The Euclidean distance between Observation 3 and the test point is:

$\sqrt{[(0 - 0)^2 + (1 - 0)^2 + (3 - 0)^2]} = \sqrt{(1 + 3^2)} = \sqrt{10} \approx 3.16$

The Euclidean distance between Observation 4 and the test point is:

$\sqrt{[(0 - 0)^2 + (1 - 0)^2 + (2 - 0)^2]} = \sqrt{(1 + 2^2)} = \sqrt{5} \approx 2.24$

The Euclidean distance between Observation 5 and the test point is:

$\sqrt{[(-1 - 0)^2 + (0 - 0)^2 + (1 - 0)^2]} = \sqrt{(1 + 1)} = \sqrt{2} \approx 1.41$

The Euclidean distance between Observation 6 and the test point is:

$\sqrt{[(1 - 0)^2 + (1 - 0)^2 + (1 - 0)^2]} = \sqrt{(1 + 1 + 1)} = \sqrt{3} \approx 1.73$

(b) What is our prediction with K = 1? Why?

We use the training point that is closest (training observation 5 (or, -1,0,1)) to the new test point (0,0,0)—which has the smallest Euclidean distance from the test observation of ≈1.41—to predict which binary outcome the new test point is most likely to be according to majority. In this case, the majority outcome of the point nearest the test point is Green, so we predict the value to be Green.

$P(Y = Red \mid X = x_0) = 1/1\ (\Sigma i \varepsilon N_0)\ I(y_i = Red) = I(y_5 = Red) = 0/1$
$P(Y = Green\ n \mid X = x_0) = 1/1\ (\Sigma i \varepsilon N_0)\ I(y_i = Green) = I(y_5 = Green) = 1/1$

(c) What is our prediction with K = 3? Why?

We use the three training points that are closest (training observation 2 (or, (2,0,0), training observation 5 (or -1,0,1) and training observation 6 (or, (1,1,1)) to the new test point (0,0,0)—which have the has smallest Euclidean distances from the test observation of ≈1.41, ≈1.73 and 2.00, respectively—to predict which binary outcome the new test point is most likely to be according to majority. In this case, the majority outcome of the three points nearest the test point is Red so, we predict the value to be Red.

$P(Y = Red \mid X = x_0) = 1/3\ (\Sigma i \varepsilon N_0)\ I(y_i = Red) = I(y_{2,5,6} = Red) = (1/3)2$ (or, 1 + 0 + 1) = 2/3
$P(Y = Green \mid X = x_0) = 1/3\ (\Sigma i \varepsilon N_0)\ I(y_i = Green) = I(y_{2,5,6} = Green) = 1/3(1)$ (or, 0 + 1 + 0) = 1/3

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

Because (0,0,0) is centered at the meeting point of the three axes, K should not be too small and should not be too large. Since there are only 6 training points, K should be around 3 or 4 (in the middle). If K is smaller than that, it could be too flexible and will have a weak decision boundary. If K is larger than that, it could be too inflexible—like a linear decision boundary—and will have a high bias. Especially since this is a small training data set, we should check all possible values for K in order to choose the K value that has the optimal or smallest error rate to use for KNN.

9. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data. *First, download the Auto data set from the course Elearning. Use the read.csv() function to import the data to R. Make sure that the missing values have been removed from the data before you start the problem. Missing values are recorded with "?" so you would need na.strings = "?" option when you import the data to R using read.csv() function. You can also check 2.3.4 Loading Data in our textbook. After loading the data use the na.omit() function to remove rows containing missing values.*

(a) Which of the predictors are quantitative, and which are qualitative?

Origin and Name are qualitative predictors, while the rest of the predictors—MPG, Cylinders, Displacement, Horsepower, Weight, Acceleration and Year—are quantitative.

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

Ranges:
MPG: 9.0 - 46.6
Cylinders: 3 - 8
Displacement: 68 - 455

Horsepower: 46 - 230
Weight: 1613 - 5140
Acceleration: 8.0 - 24.8
Year: 70.0 – 82.0

(c) What is the mean and standard deviation of each quantitative predictor?

Means and Standard Deviations:
MPG: 23.45 (mean); 7.805007 (standard deviation)
Cylinders: 5.472 (mean); 1.705783 (standard deviation)
Displacement: 194.4 (mean); 104.644 (standard deviation)
Horsepower: 104.5 (mean); 38.49116 (standard deviation)
Weight: 2978 (mean); 849.4026 (standard deviation)
Acceleration: 15.54 (mean); 2.758864 (standard deviation)
Year: 75.98 (mean);3.683737 (standard deviation)

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

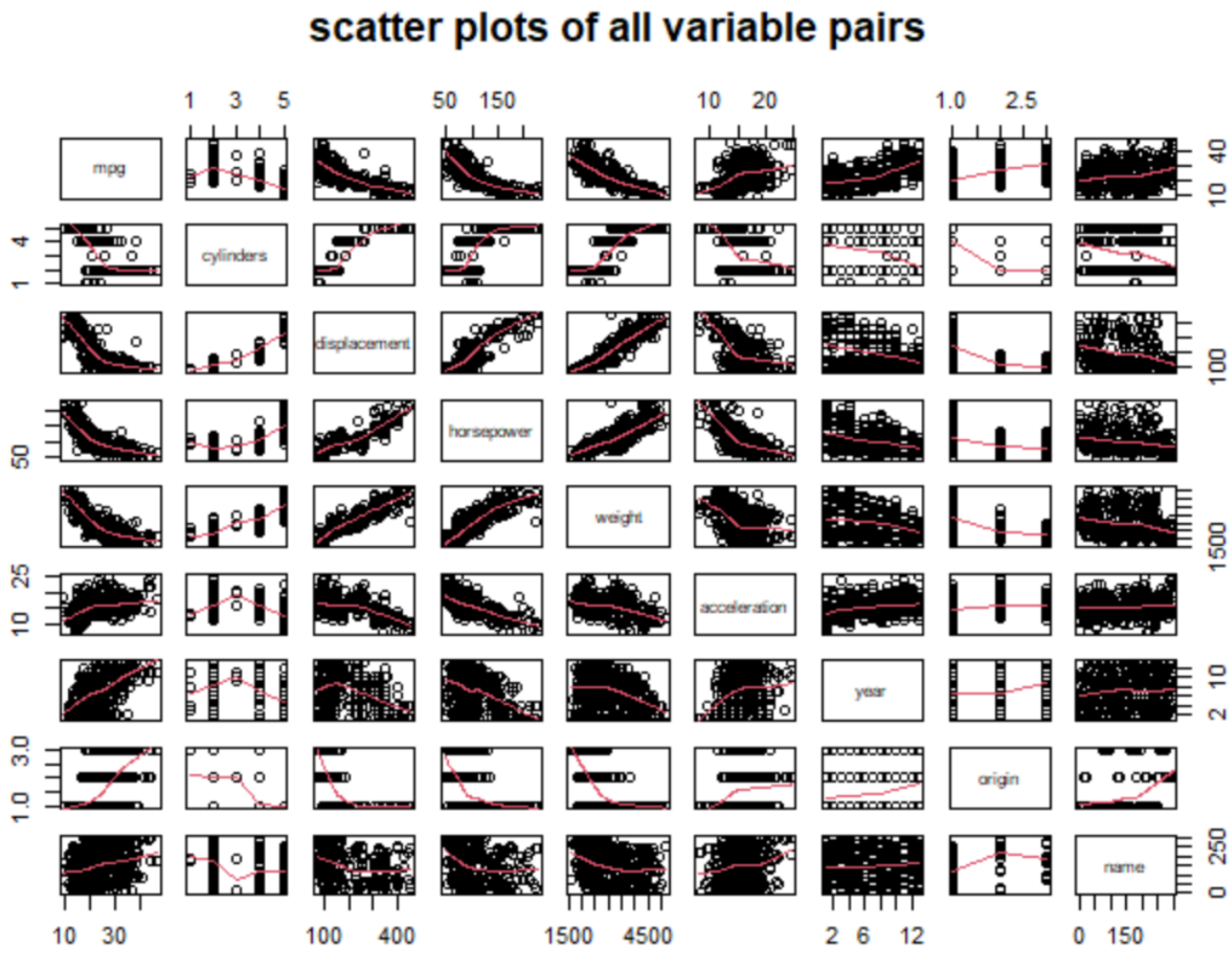
	Range	**Mean**	**SD**
Mpg	11.0 – 46.6	24.404430	7.867283
Cylinders	3 – 8	5.373418	1.654179
Displacement	68 – 455	187.240506	99.678367
Horsepower	46 – 230	100.721519	35.708853
Weight	1649 – 4997	2935.971519	811.300208
Acceleration	8.5 – 24.8	15.726899	2.693721
Year	70 – 82	77.145570	3.106217
Origin	1 – 3	1.601266	0.819910

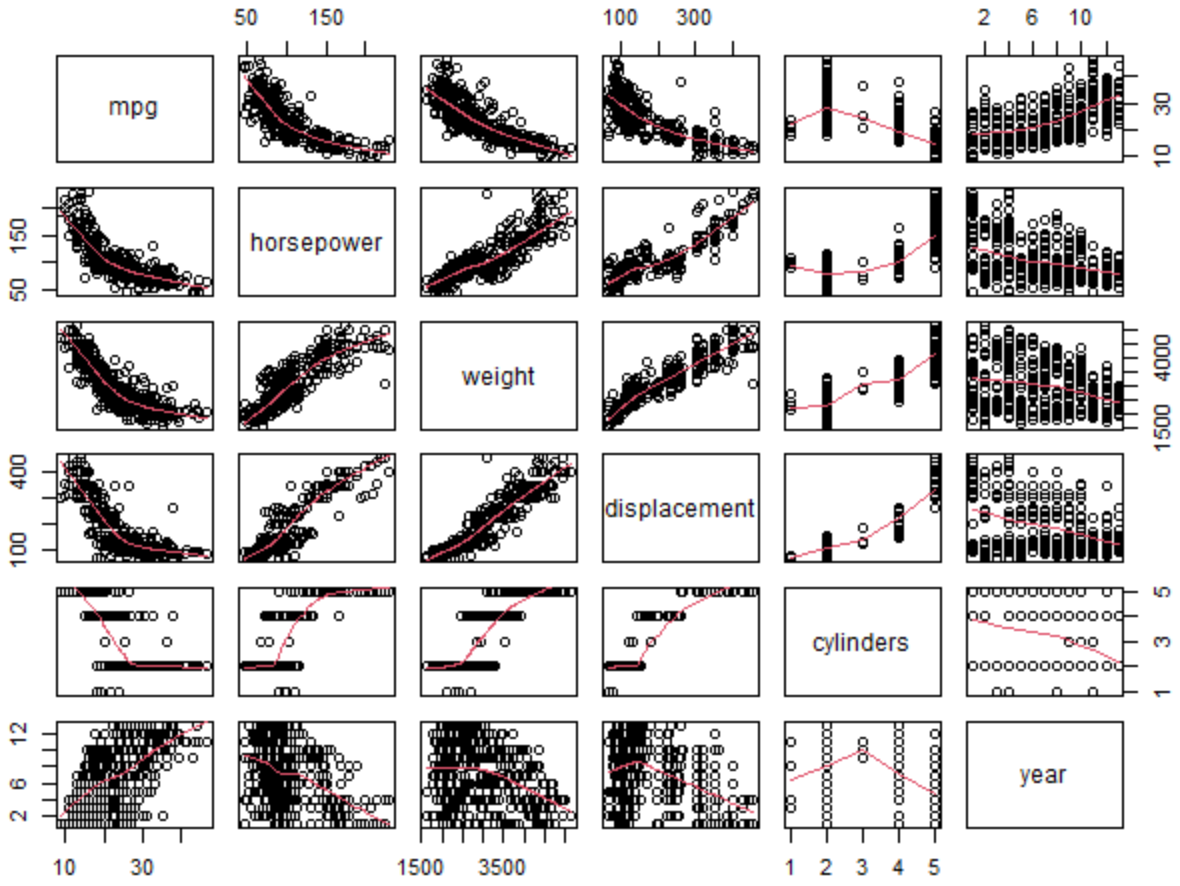
(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

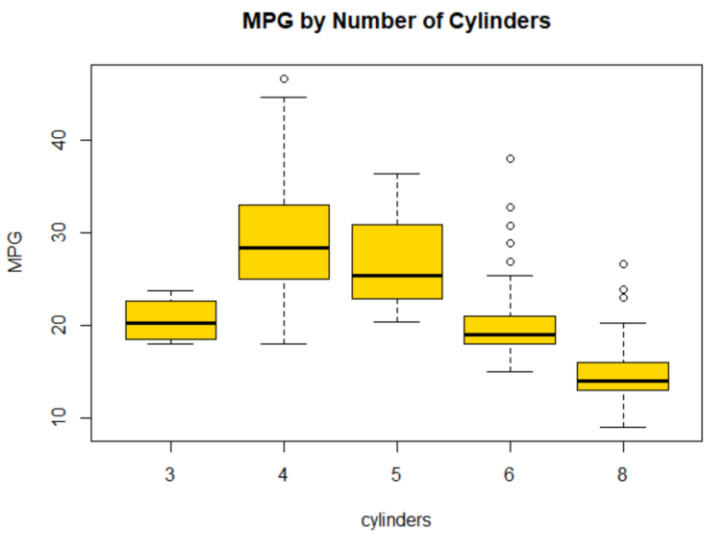
At first, I chose to take a look at all pairs that comprised of MPG and another variable with which MPG had a distinct “shape” in its pair plot: cylinders, displacement, horsepower, weight, year and country of origin.

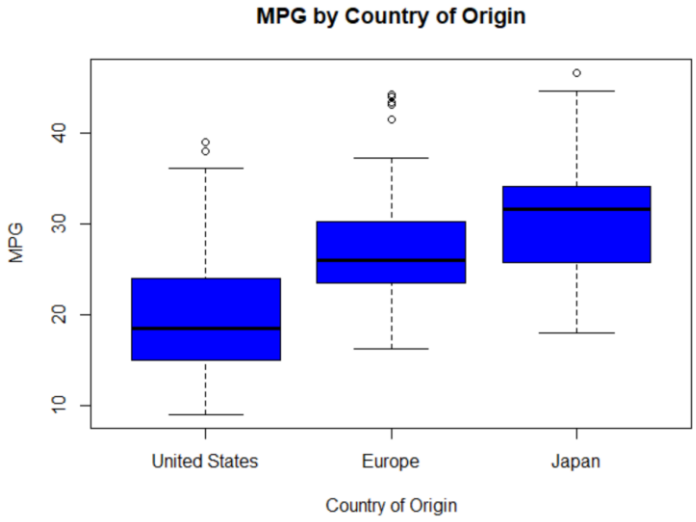
We can see by the definition of shape in certain pairs, such as MPG and displacement, MPG and horsepower and MPG and weight that that there is a distinct relationship (an inverse relationship between MPG and the other three—or, the higher MPG goes, the lower the others go). The three variables of displacement, horsepower and weight can be seen in the “scatter plots of all variable pairs” as having a direct relationship with each other. Horsepower is directly related to weight (the higher one goes, the higher the other goes). Displacement and weight are directly related and displacement and horsepower are directly related.

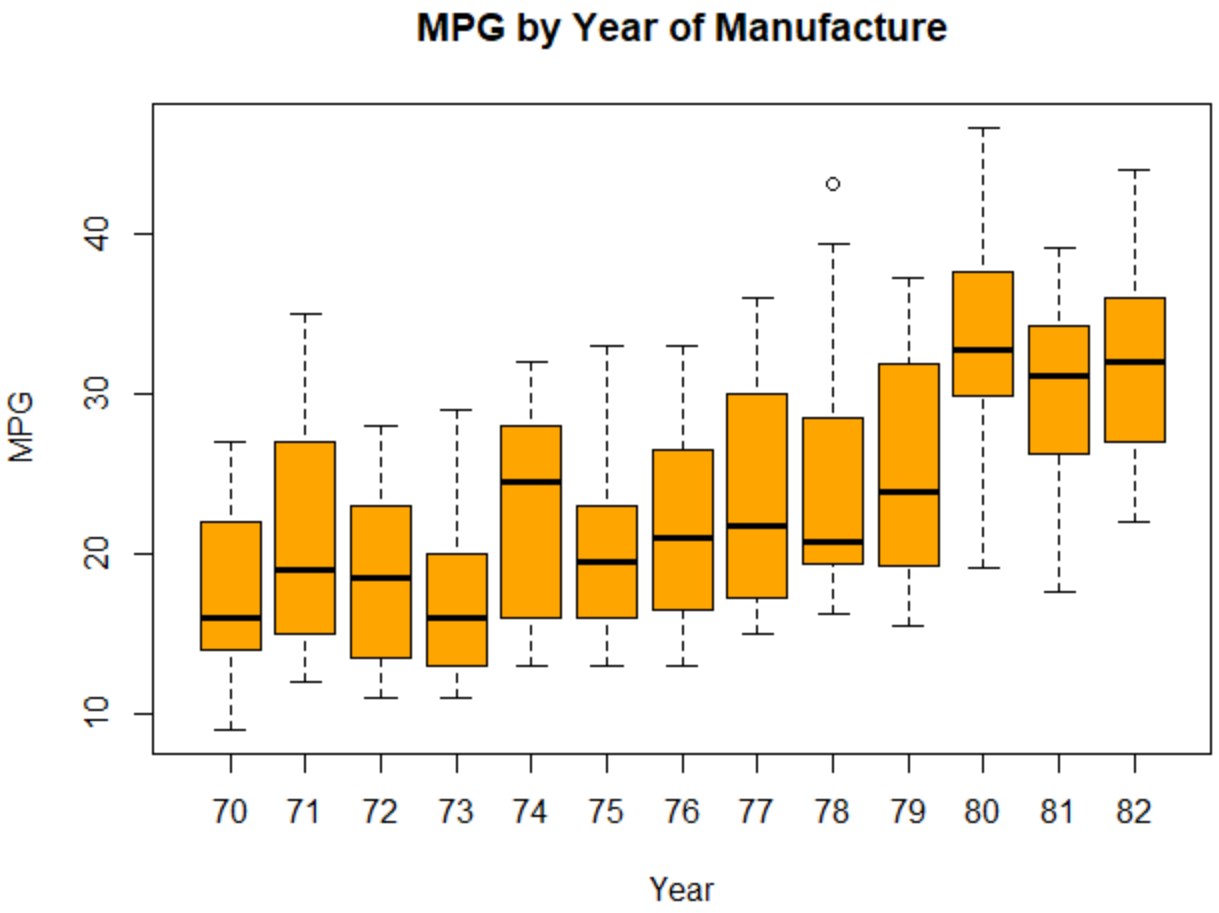
Were I to investigate these direct relationships I could learn more about the automobile mechanics that contribute to or detract from better MPG. That said, I include the relationships between MPG and weight, displacement, horsepower in my below graphics. However, I do not fully understand how they are related to each other. I also do not understand the variable “acceleration”.

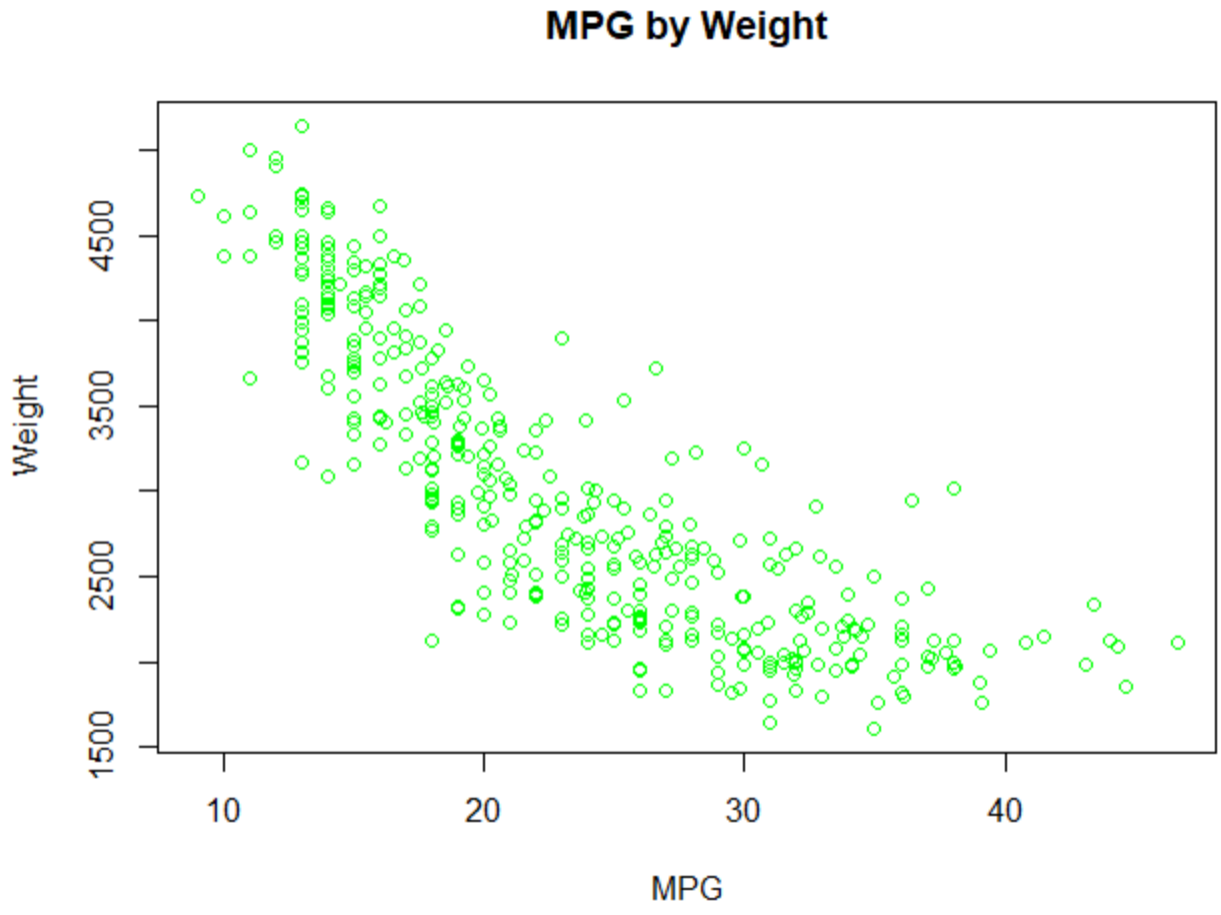


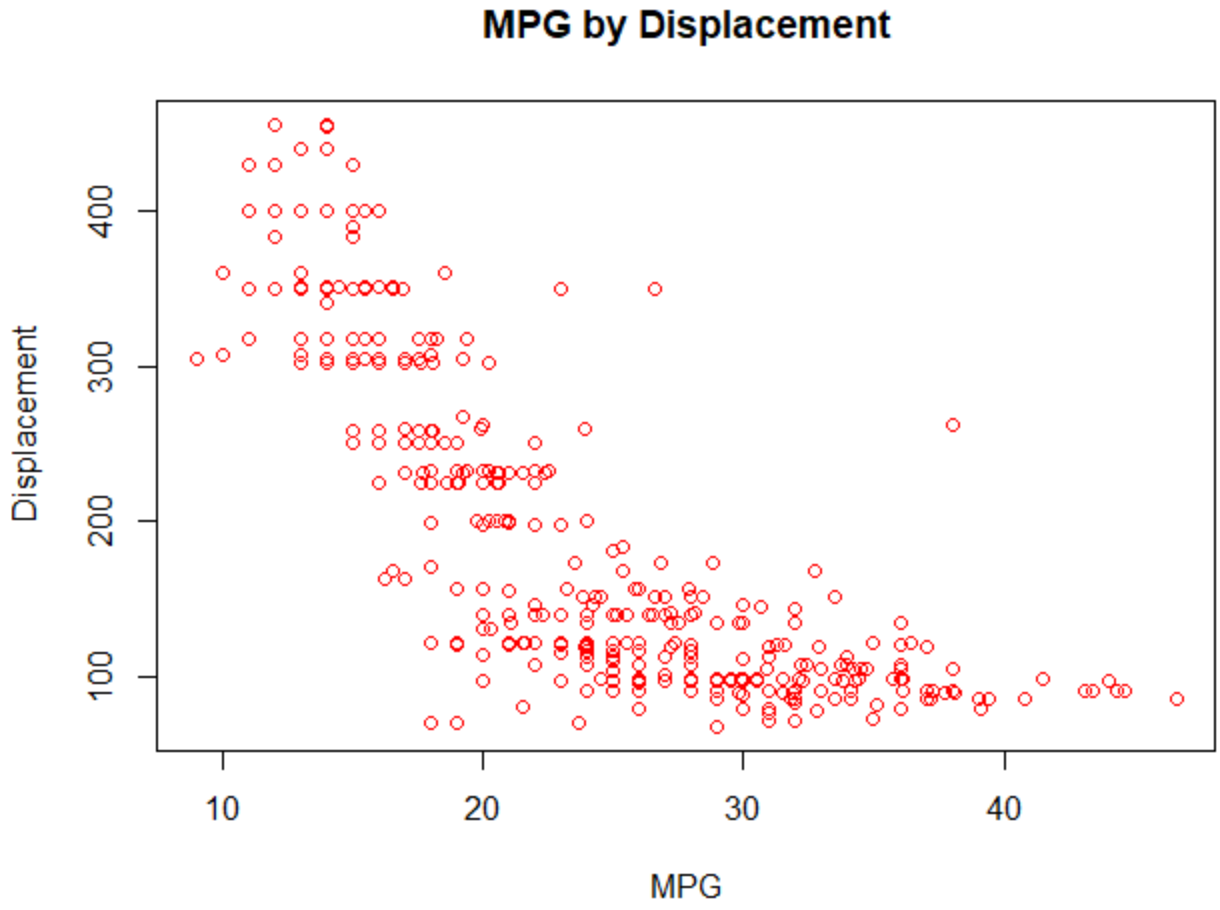

mpg
horsepower
weight
displacement
cylinders
year

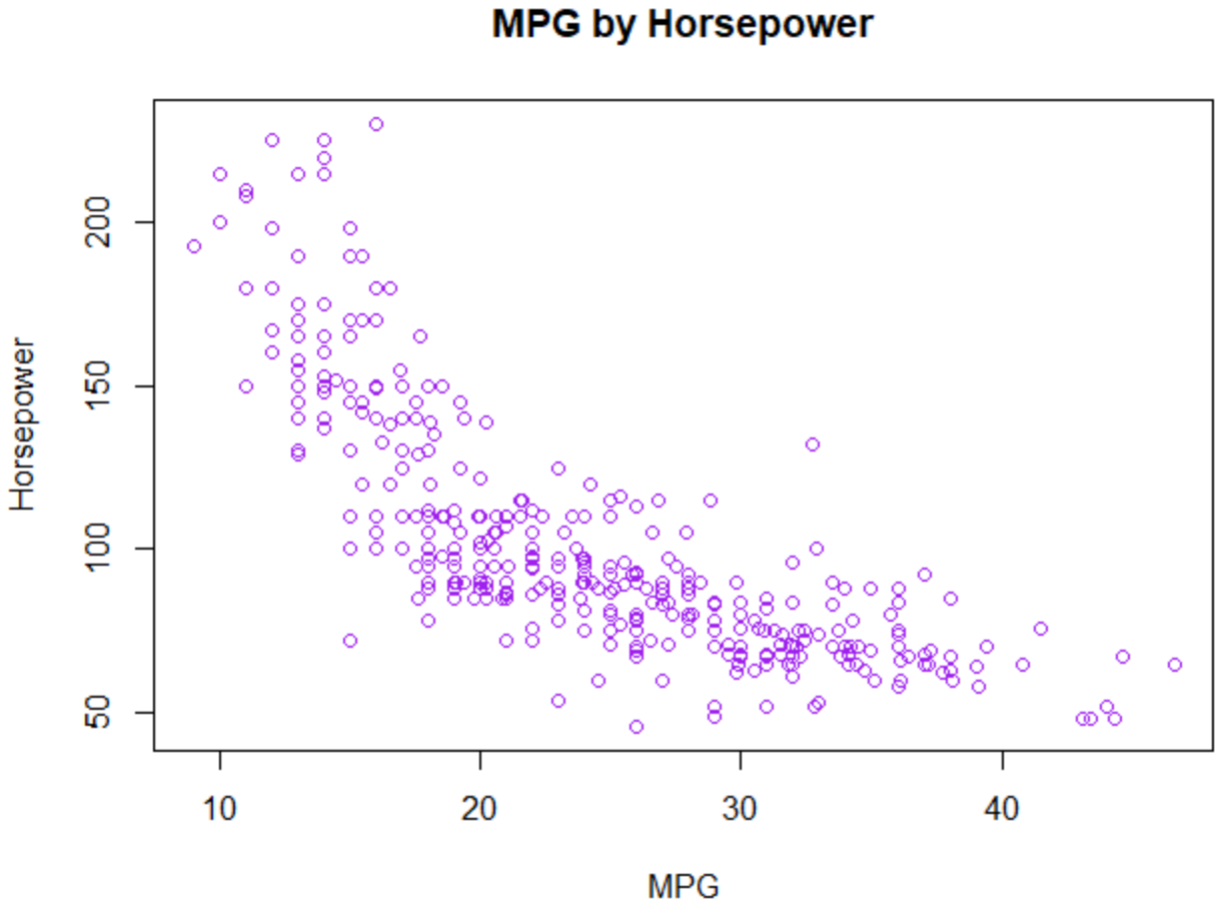

MPG by Number of Cylinders
MPG
cylinders

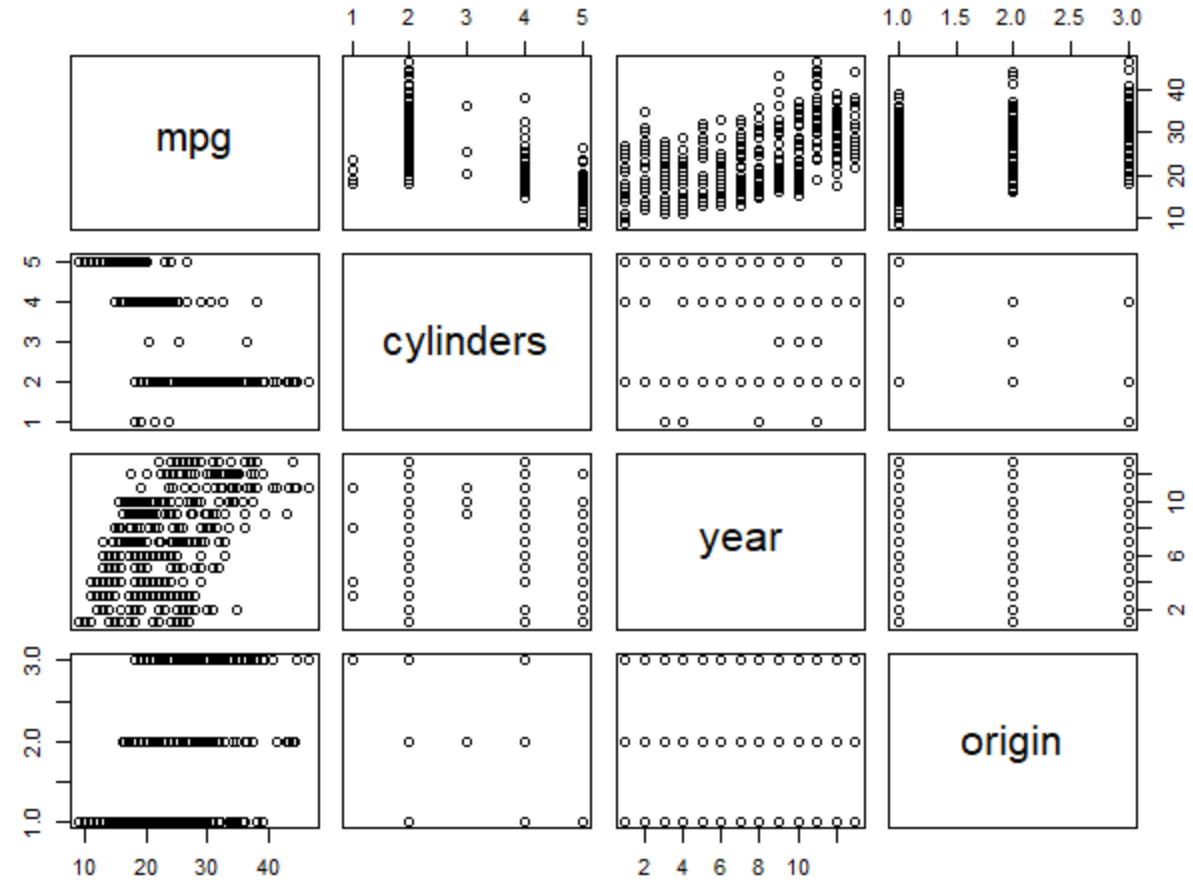

MPG by Country of Origin
MPG
United States
Europe
Japan
Country of Origin

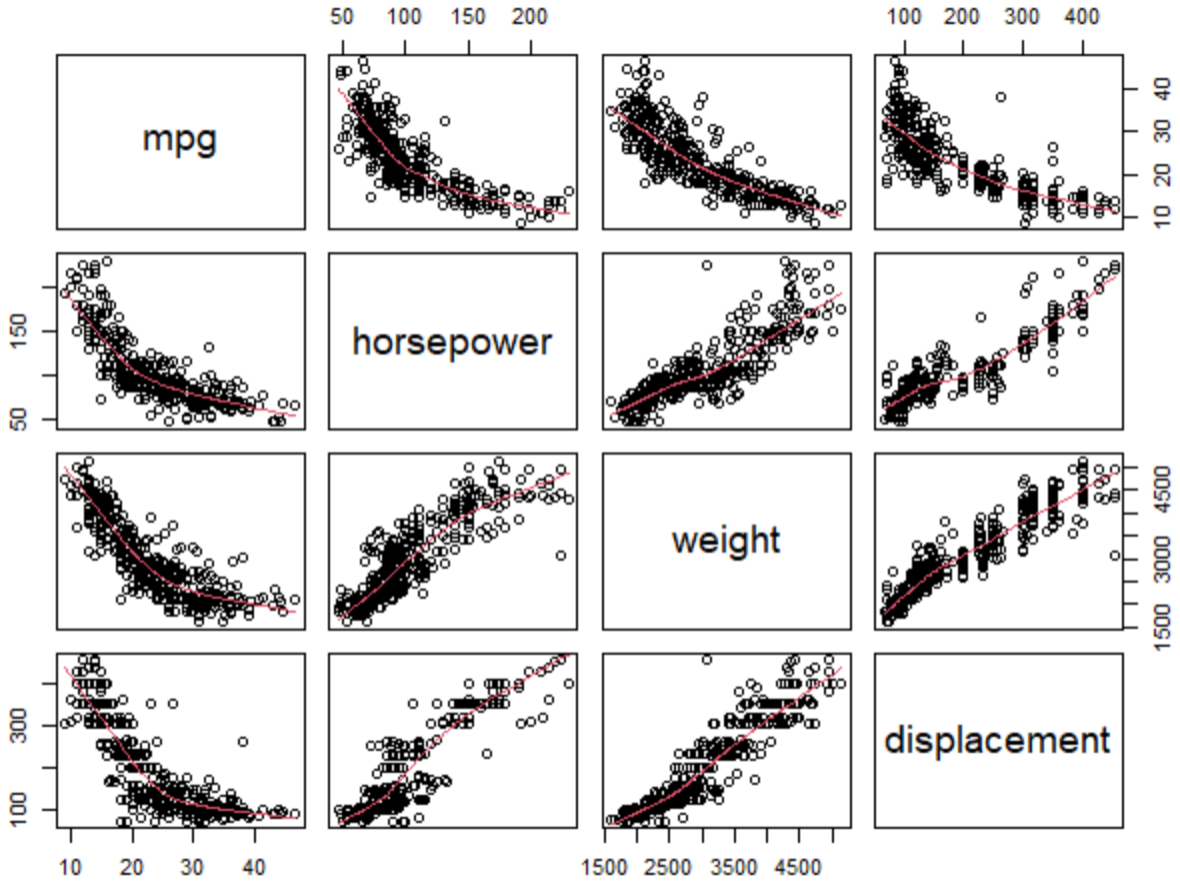

MPG by Year of Manufacture
MPG
10
20
30
40
70
71
72
73
74
75
76
77
78
79
80
81
82
Year

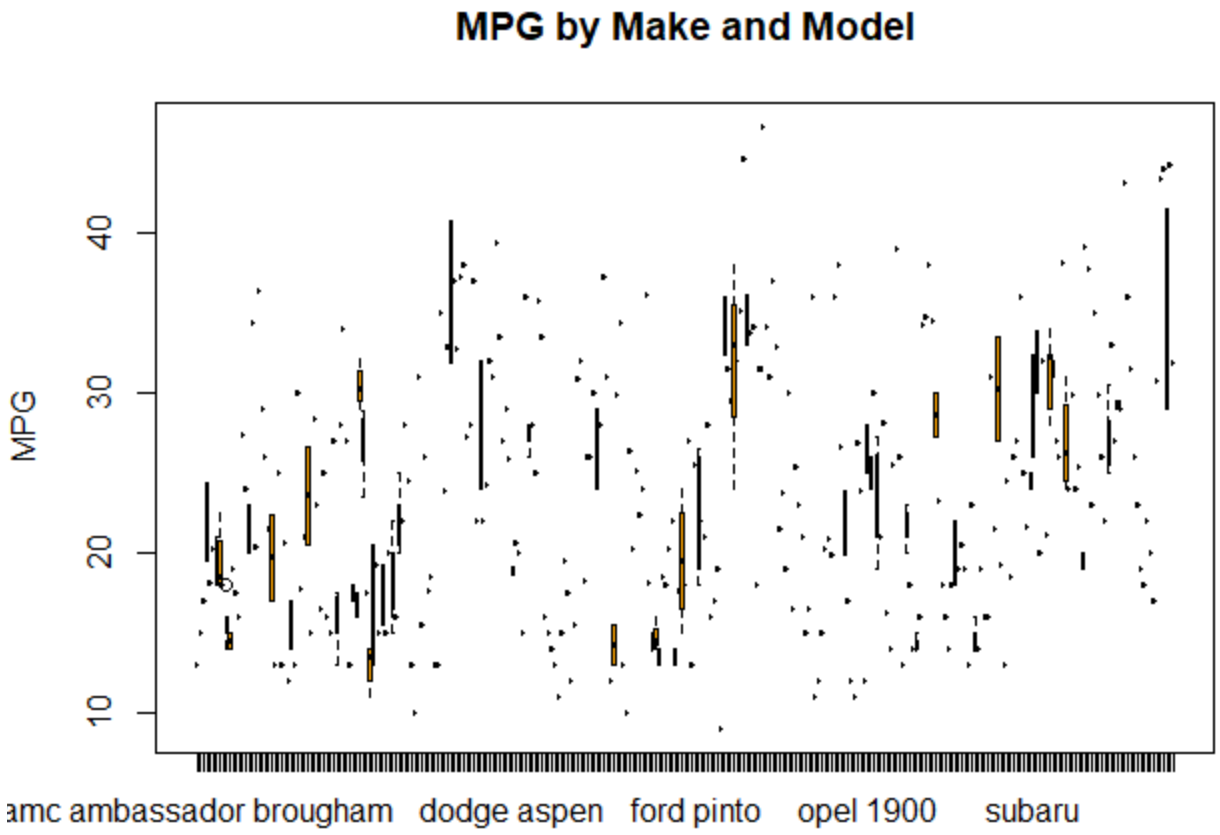

MPG by Weight
Weight
1500
2500
3500
4500
10
20
30
40
MPG


MPG by Displacement
Displacement
100
200
300
400
10
20
30
40
MPG


MPG by Horsepower
Horsepower
50
100
150
200
10
20
30
40
MPG


mpg
cylinders
year
origin


mpg
horsepower
weight
displacement



(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

We can reasonably predict that cars with 4 cylinders and/or that are made in Japan will have better gas mileage, and that the newer the car (at least in the year range of 1970 – 1982), the better the gas mileage.

The greater the weight, the more displacement and/or the higher the horsepower, the less MPG the car will have.

In sum, horsepower, weight, displacement (the former three of which are highly correlated), cylinders, year, and country of origin are predictors of MPG. Some makes/models have better MPG than others, but it is hard to tell which ones from the chart because the names are offset (possibly Subaru, Dodge Aspen, and Ambassador?).