Car Talk

According to a website

> "There are 107 million U.S. households, each with an average of 1.9 cars, trucks or sport utility vehicles and 1.8 drivers, the Bureau of Transportation Statistics reported. That equals 204 million vehicles and 191 million drivers."

That's a lot of cars, not all parked outside in the parking lot.

We'll look at a dataset on cars to learn how to make boxplots, scatterplots and find correlations. First we need to load the data set.

```
> library(MASS)
> data(Cars93)
> attach(Cars93)
```

That loads an external package `MASS`, then a data set in that package, and finally attaches the variables so you can call them by name.

To find out what variables are present we can do either of the following `names(Cars)` or `?Cars93`. The first command list the names of the variables, the second displays the help page.

1 boxplots

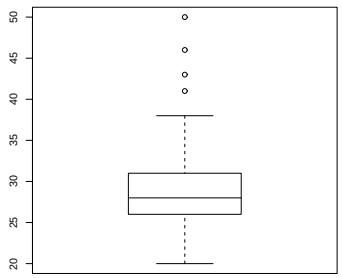
Boxplots allow one to easily visualize the five number summary of a data set. Recall, this summarizes center with the median, spread with the IQR and also gives the range. Furthermore, boxplots can show whether a data set is symmetric or skewed by comparing the two halves.

Making boxplots is easy with the `boxplot()` command. For example, a boxplot of `MPG.highway` can be found with

```
> fivenum(MPG.highway)

[1] 20 26 28 31 50

> boxplot(MPG.highway)
```



From the boxplot you can see the median is around 27 and 25% or more are less than 25 or so.

 Question 1: Make a boxplot of `MPG.city`. Does it have the same general shape as the highway boxplot? What is the cutoff for the top 75%? What is the IQR?

You can make side-by-side boxplots by listing more than one variable. For example, this command will produce the two boxplots together

```
> boxplot(MPG.highway, MPG.city, names = c("highway", "city"))
```

The argument `names=...` will label the boxplots. Otherwise, they appear with a 1 and a 2 under them.

 Question 2: Make the side-by-side boxplot of `MPG`. Do the two have the similar centers? How about spreads?

The above boxplot shows two numeric variables side by side. If one variable is a categorical variable, then a different syntax can be used which in many cases is easier.

For example, the variable `DriveTrain` contains information on the type of drive the car is (four wheel, front or rear.) To make a boxplot of miles per gallon for each level of this variable can be done as follows

```
> boxplot(MPG.highway ~ DriveTrain)
```

The tilde, `~`, notation is used for many `R` functions.

 Question 3: Make the boxplots above. Are the spreads similar? Are the centers similar?

 Question 4: Make boxplots of mileage versus the number of passengers with

```
> boxplot(MPG.highway ~ Passengers)
```

What is going on with the plots for 1 and 8 passengers? Can you describe any trend appearing in the data?

 Question 5: Make boxplots of highway mileage versus the type of car, `Type`. Are there any trends? Are the centers similar? Are the spreads? Why does `small` have no median marked? You can find the five number summary of this data with the command

```
> fivenum(MPG.highway[Type == "Small"])

[1] 29 33 33 37 50
```

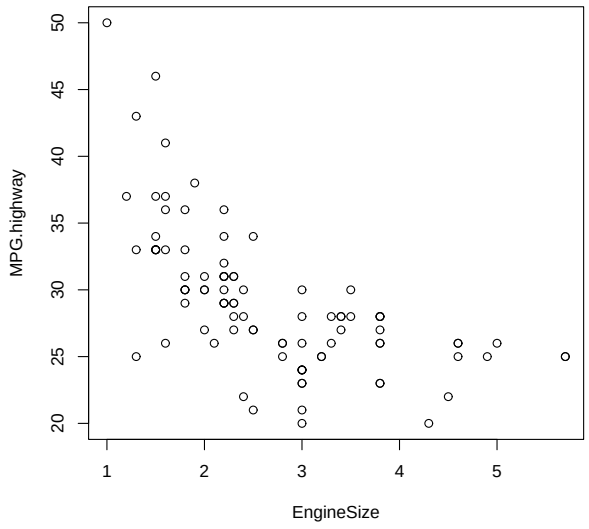
Notice the square brackets and double equals sign. This looks at only the values when `Type` is equal to `small`.

2 scatterplots

For pairs of numeric variables, we can make scatterplots to investigate relationships. A scatterplot is made with the command `plot()`. The simplest usage is of the type `plot(x,y)`.

For example, a plot of `MPG.highway` by `EngineSize` can be made with

```
> plot(EngineSize, MPG.highway)
```



The figure shows that the first variable is printed along the *x* axis and the second on the *y*.

Question 6: From the scatterplot for `EngineSize` and `MPG.highway` describe any trend.

Question 7: For the scatterplot, the command

```
> identify(EngineSize, MPG.highway, labels = Make)
```

allows you to click points to label them one at a time. Find the make of the car with the best and worst mileage and least and greatest engine size. Use the right mouse button to stop.

Question 8: You can color different points in a plot by a third variable using `col=`. For example,

```
> plot(Price, MPG.highway, col = as.numeric(Type))
```

The command `as.numeric` gives a number to each level of `Type`. You can see what the types are by

```
> identify(Price, MPG.highway, labels = Type)
```

The scatterplot shows a decreasing relationship between price and mileage. Does the coloring indicated the presence of any lurking variable? That is, an unspecified variable that can cause the relationship.

 Question 9: Make a scatterplot of `MPG.city` versus `MPG.highway`. Describe the trend.

3 correlation

Finding correlations in `R` is easy using the function `cor()`. For example, to find the correlation between `MPG.highway` and `MPG.city` is done with

```
> cor(MPG.highway, MPG.city)

[1] 0.944
```

Notice it is close to 1 as should be expected. (Why?)

 Question 10: Find the correlation of `MPG.highway` and `Price`. Is it "big"? Why is it negative?

 Question 11: Make a scatterplot of `Width` and `Height`. Guess the correlation and then check using `cor()`.

 Question 12: Most all the variables in the data set are correlated. To see uncorrelated data try these commands

```
> x = rnorm(100)
> y = rnorm(100)
> plot(x, y)
> cor(x, y)

[1] 0.07023
```

If you type this, you will get a different, but similar, answer as the `rnorm()` function returns randomly chosen numbers.

3.1 Spearman Correlation

When data is curvilinear, it may be correlated in a strong sense, yet the Pearson correlation coefficient can be 0. A forced example is the following one

```
> x = c(-2, -1, 0, 2, 1)
> y = x^2
> cor(x, y)
```

```
[1] 0
```

The correlation is 0, but the two data sets are clearly related.

When there is a *monotonic* (increasing or decreasing) relationship, the Spearman Correlation Coefficient can summarize how strong that relationship is. To find this value you first rank the data from smallest to largest and then find the correlation.

Ranking is done with `rank()`. For example

```
> x = c(1, 3, 5, 7, 2, 4, 6)
> rank(x)

[1] 1 3 5 7 2 4 6
```

When there are ties they are averaged

```
> y = c(1, 1, 7, 2, 3)
> rank(y)

[1] 1.5 1.5 5.0 3.0 4.0
```

To find the Spearman Correlation Coefficient then is done as follows

```
> cor(rank(MPG.city), rank(MPG.highway))

[1] 0.9359
```

Notice, this is about the same as `cor(MPG.city, MPG.highway)` as the data is more or less linear.

 Question 13: Compare the Pearson and Spearman correlation coefficients for `MPG.highway` and `Price`. Is there a big difference?

 Question 14: Look at the output of `names(Cars93)` and try to think which relationships should have a large correlation. Write down three such relationships. Make a scatterplot and check with `cor`. Were you correct?