Examining the number of characters in an email

Your Name Here

Put the date here

Load packages and data set.

We will be using the *email50* data set which you have already downloaded and placed in your class folder. Read this data into R by filling in your path. If you don't recall where it is at, check lab 3 and simply copy the code used there to read it in.

```
email50 <- read.delim("_____________________/email50.txt", header=TRUE, sep="\t")
```

At this point click the Knit HTML button at the top and check to see if your document is created correctly.

Summary Statistics

Let's examine the number of characters `num_char` contained in an email.

What are the `min`imum and `max`imum number of characters in emails from this dataset? Don't forget to *Knit html* after each chunk to make sure the output is being produced correctly.

```
min(email50$____);max(______$_______)
```

Use the `range()` function to find these numbers in another way.

```
_____(email50$_____)
```

What is the `mean` and `var`iance of this variable?

```
_____(email50$____)
var(_____$____)
```

Univariate visualizations (one variable)

Let's visualize the distribution of `num_char`.

Histograms

1. Create a histogram of `num_char` using the base `hist(x)` plotting function.

```
hist(_____)
```

2. Use the `breaks=` argument to create 3 breaks in the histogram. How many bins are created?

```
_____(______, breaks=3)
```

3. Does a histogram with 13 bins provide a different description of the data distribution than when there are only 4 bins? Draw one and find out.

```
_____(______, breaks=___)
```

Boxplots

1. Calculate the five number `summary`.

```
summary(______)
```

2. Draw a boxplot using `boxplot(x)`.

```
boxplot(______)
```

3. Redraw the boxplot horizontally.

```
boxplot(______, horizontal=______)
```

Bivariate visualizations (two variables)

This is when multiple variables are plotted against each other. Or one variable is plotted against levels of another variable.

Scatterplot

We would expect that the more characters in the email corresopnds with more `line_breaks`. Let's create a scatterplot with `line_breaks` on the y axis and `num_char` on the x. The syntax for a scatterplot is `plot(y~x)` OR `plot(x, y)`. The first notation is called **model notation** and we will be using it frequently in this class.

```
plot(______ ~ ______)
```

Grouped boxplot

Can you use the `num_char` or `line_breaks` to determine if an email is spam or not? Creating grouped boxplots can help us see if one group has higher measurements compared to the other. Again we use the model notation and type `boxplot(c~g)`, where in this case `c` is the continous variable and `g` is the grouping variable `spam`.

1. Create a boxplot of `num_char` against `spam`.

```
boxplot(____$num_char ~ email50$____)
```

2. Create a horizontal grouped boxplot of `line_breaks` against `spam`.

```
boxplot(email50$_____ ~ _____$spam, horizontal=_____)
```