STAT100B Lab 5 Data Visualization in R

First Name _____ Last Name _____SID (last 4 digits only)_____

The `Loblolly` data in `R` has several variables pertaining to growth records for Loblolly pines, a type of pine tree native to the Southeastern United States. Load this data in `R` and examine the help file with the following commands:

```
data("Loblolly")
?Loblolly
```

Exercise 1 What are the variables in this dataset? Are they numeric or categorical? (6 pts)

Boxplots

In order to get more comfortable examining model assumptions, we'd like to get familiar with `R`'s plotting capailibites. We will start by examining how `height` varies across different levels of `seed`. Since the `seed` variable has 14 levels, we will ask `R` for a subset of the data that includes only seeds 329, 315, and 305.

The following command creates this subset by taking `Loblolly` *such that* `Seed` is 329 *or* `Seed` is 315 *or* `Seed` is 305. The `droplevels` command cleans up the subsetted data so that it will plot nicely.

```
subset <- Loblolly[Loblolly$Seed == 329 | Loblolly$Seed == 315 | Loblolly$Seed == 305,]
subset <- droplevels(subset)
```

To examine graphically how `height` varies across different levels of `seed` (in our subset of the data), we will start with a boxplot. Remember that we can use a `~` as "by". That is, we want a boxplot of `height` *by* levels of `seed`. Recall also that we use the dollar sign `$` to tell `R` that we want a particular variable from a dataset, i.e., dataset$variable.

```
boxplot(subset$height ~ subset$Seed)
```

This plot is a nice start, but it may look somewhat incomplete. It's missing a title and could stand to have cleaner axis labels. The following command adds a `main` title and axis labels `xlab` and `ylab`:

```
boxplot(subset$height ~ subset$Seed,
        main = "Boxplot of Tree Height by Seed on Subsetted Data",
        xlab = "Seed", ylab="Height (ft)")
```

Exercise 2 Do you think that `height` differs between different values of `seed` ? (2pts)

Exercise 3 Do the three `height` groups look normally distributed? (2pts)

Histograms

While boxplots are a convenient way to make side-by-side comparisons, it can be difficult to conclusively answer questions like the one posed in Exercise 3. Histograms provide a much more straightforward way to examine the shape of a distribution. The following command creates a basic histogram for the `Loblolly height` variable:

```
hist(Loblolly$height)
```

We would again like to include a better title and new axis labels. Fortunately, `R` 's functions for plotting all use the same approach!

```
hist(Loblolly$height,
     main = "Histogram of Loblolly Pine Heights",
     xlab = "Height (ft)", ylab="Frequency")
```

Exercise 4 Do the Loblolly pine heights appear to be normally distributed? (2pts)

Previously, we wanted information on normality for a subset of the data, using only seeds 329, 315, and 305. We can build individual histograms for these data. Recall that the square brackets can be read as “such that”.

```
hist(Loblolly$height[Loblolly$Seed == 329],
     main = "Histogram of Pine Heights for Seed 329",
     xlab = "Height (ft)", ylab="Frequency")
```

Exercise 5 Create histograms of the pine heights for the other two seeds, 315 and 305. For each seed, decide whether it is reasonable to assume that the heights are normally distributed. (4pts)

Scatterplots and Regression Lines

We may run into some problems with this data because there are only 6 observations per seed! It may be more reasonable to compare `age` and `height` of trees in the complete data.

Exercise 6 How many observations are there for `height` ? (2pts) How many differet values are there for `age` ? (2pts) Show your R code to find the answser the two questions above. (4pts)

If we want to examine the relationship between `age` and `height` , it is reasonable to think that we would be interested in using `height` to predict `age` . (If we walking through a forest of Loblolly pine trees, it will be much easier to get a tree's height than its age!)

Exercise 7 In this setting, which is the explanatory variable (predictor)? Which is the response? (2pts)

Using this predictor/response variable setting, we want to look at a scatterplot of the data to get an idea of whether there might be any correlation between the two. The following command creates this scatterplot, complete with a title and reasonable axis labels.

```
plot(x = Loblolly$height, y = Loblolly$age,
     main = "Scatterplot of Age vs Height",
     xlab = "Height (feet)", ylab = "Age (Years)")
abline(a = 0.7574, b = 0.3783, col='red')
```

Exercise 8 Is there evidence of a linear relationship between tree age and tree height? (2pts) Without doing any math or using the computer, take a guess as to what the correlation might be for these two variables. (2pts)

We'll spend some time on regression next week, but for now the regression line is

$$\hat{y} = 0.7574 + 0.3783x$$

We can include this in our plot using the function `abline` . This function adds a line to an existing plot in `R` . The name "abline" refers to the way that lines are written in many an algebra class: $y = a + bx$. Include the regression line in your scatterplot by adding the following line of code right under the previous plot function:

Exercise 9 Predict the age of a tree that is 20 feet tall. (3pts)

On your own

1. Load the data using the command `data("faithful")`. Now, we will think about using eruption duration to predict how long we need to wait before seeing another eruption.

   a. What is the response variable? (2pts) What is the explanatory variable? (2pts)

   b. Create a scatterplot of this data with the predictor and response variables on the appropriate axes. Make sure to include appropriate axis labels and title. Write down your R code below. (4pts)

   c. The regression line for this data is

   $$\hat{y} = 33.47 + 10.73x$$

   Write down the R code to add this line to your scatterplot. (2pts)

2. The `ToothGrowth` dataset in `R` gives information about the effect of Vitamin C on tooth grown in Guinea Pigs. (Use the command `?ToothGrowth` for more information on the data and each individual variable.) Load this data into `R` using the command `data("ToothGrowth")`.

   a. Create boxplots of the guinea pigs' tooth length by the factor levels for supplement type *and* dosage. Be sure to include an appropriate title and axis labels to match your boxplots. Writedown your R code below. (4pts) *Hint*: you can create factor levels in a formula by using factorA * factorB.

b. Based on your boxplots, do you think there are differences among the six treatments? (2pts)

This lab was written by Lauren Cappiello for STAT 100B at the University of California, Riverside using the RMarkdown Lab style file from OpenIntro.