STAT 100B Lab 4

Wesley Chang

Summer 2020 Session B

Setup for Lab

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

Lab Exercises

Exercise 1

What are the variables in this dataset? Are they numeric or categorical?

```
names(Loblolly)
```

```
## [1] "height" "age"    "Seed"
```

Answer

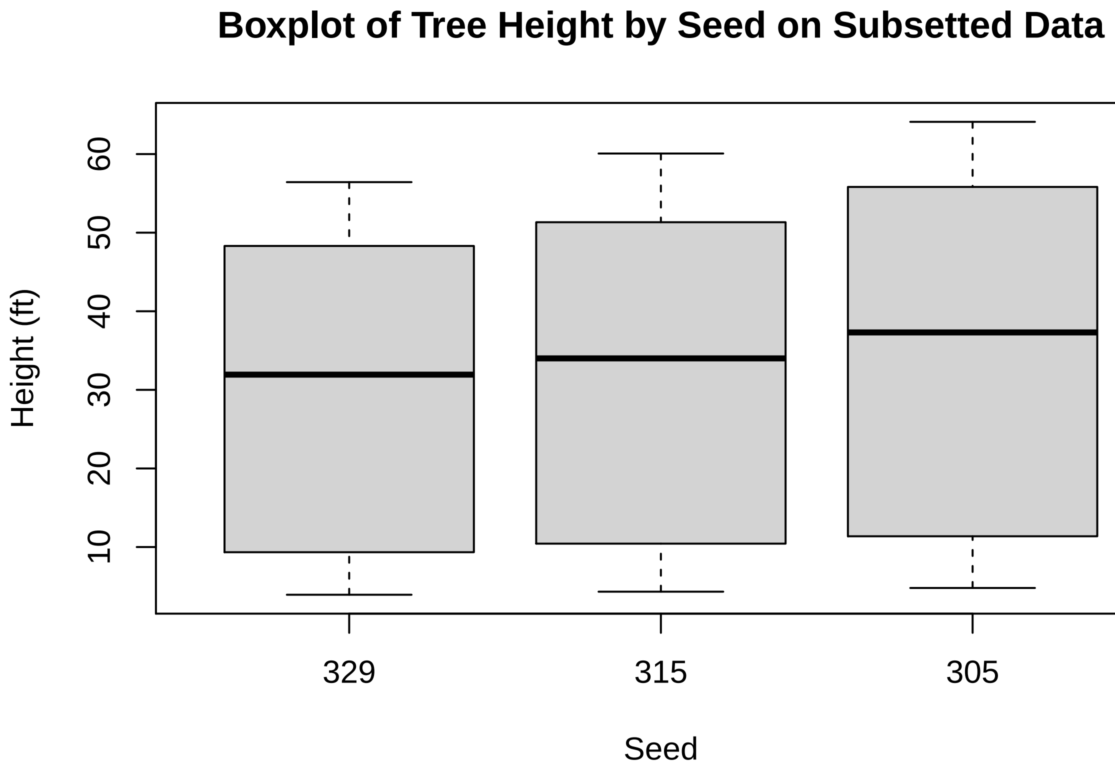
The variables in the dataset are height, age, and seed. Height and age are numeric, while seed is categorical.

Exercise 2

Do you think that `height` *differs between different values of* `seed`*?*

```
# create subset such that Loblolly is 329 or 315 or 305
# droplevel() cleans up subsetted data so that it will plot nicely
subset <- Loblolly[Loblolly$Seed == 329 | Loblolly$Seed == 315 | Loblolly$Seed == 305,]
subset <- droplevels(subset)

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



Answer

I think that height differs between different values of `seed`, but the variation is not high.

Exercise 3

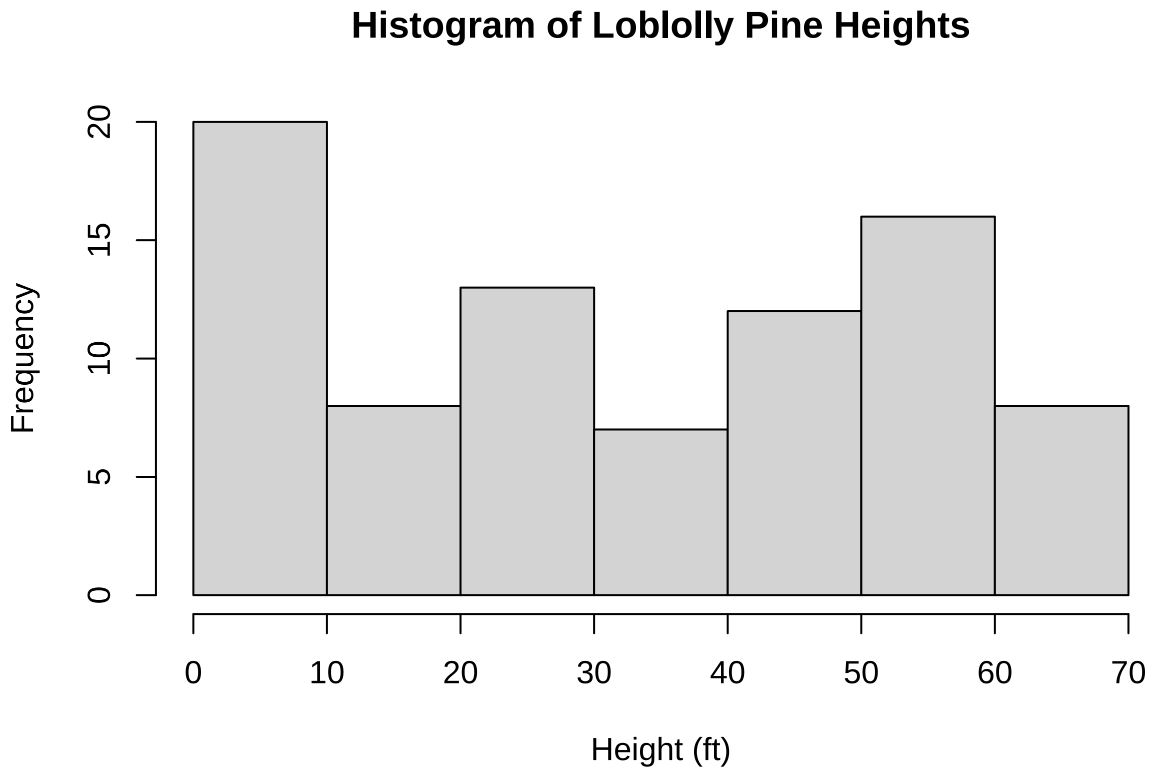
Do the three `height` *groups look normally distributed?*

Answer

The three height groups all look normally distributed, as the three respective means appear relatively centered in each distribution boxplot.

Exercise 4

```r
# construct histogram for the height variable, with label names
hist(Loblolly$height,
     main = "Histogram of Loblolly Pine Heights",
     xlab = "Height (ft)", ylab="Frequency")
```

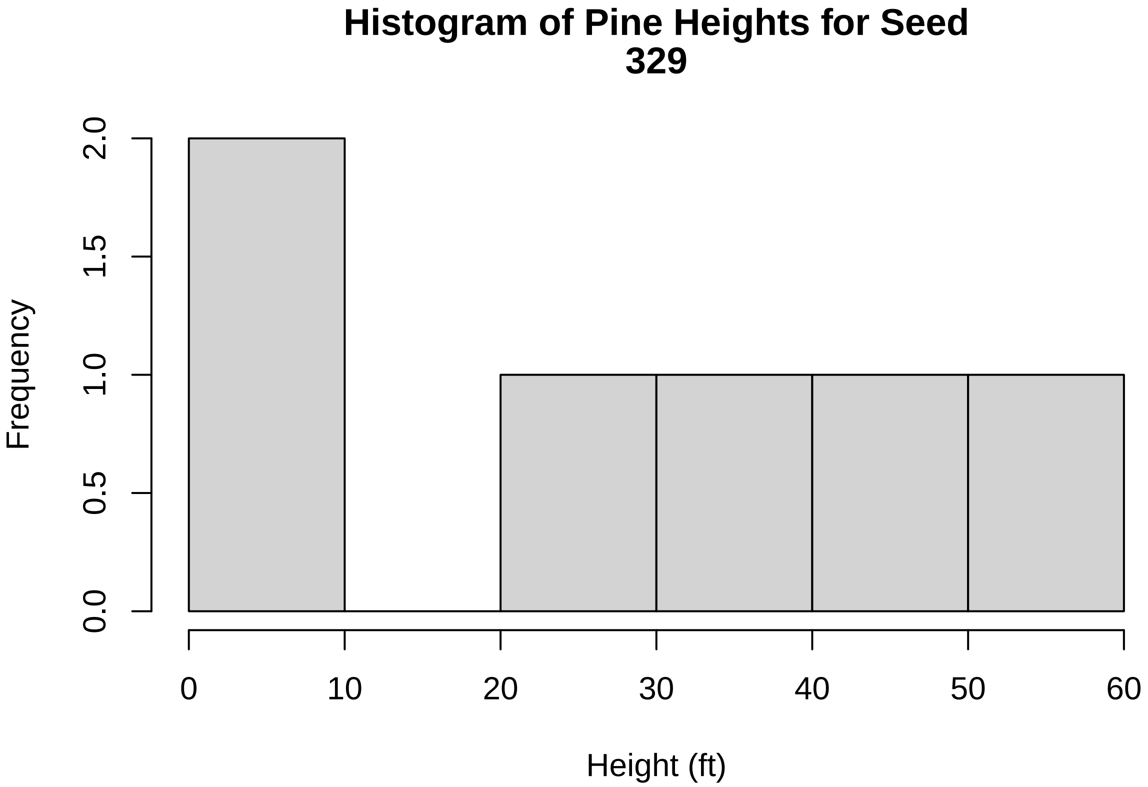


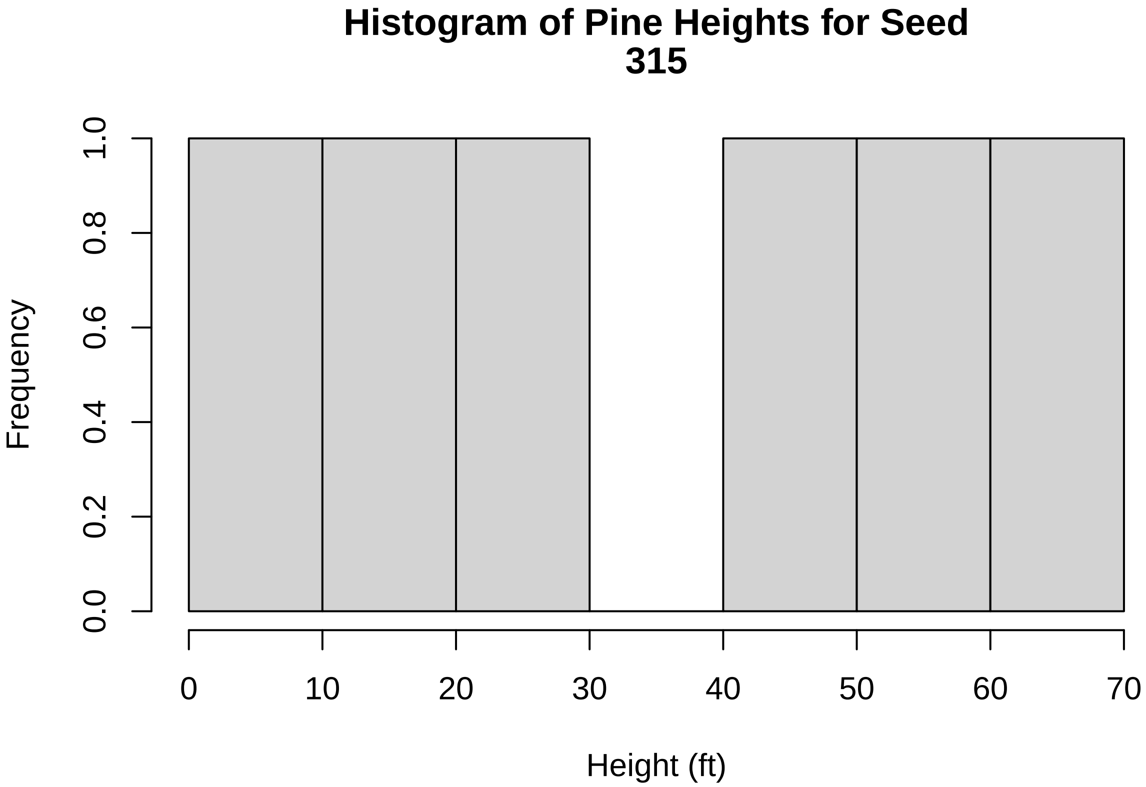
Answer

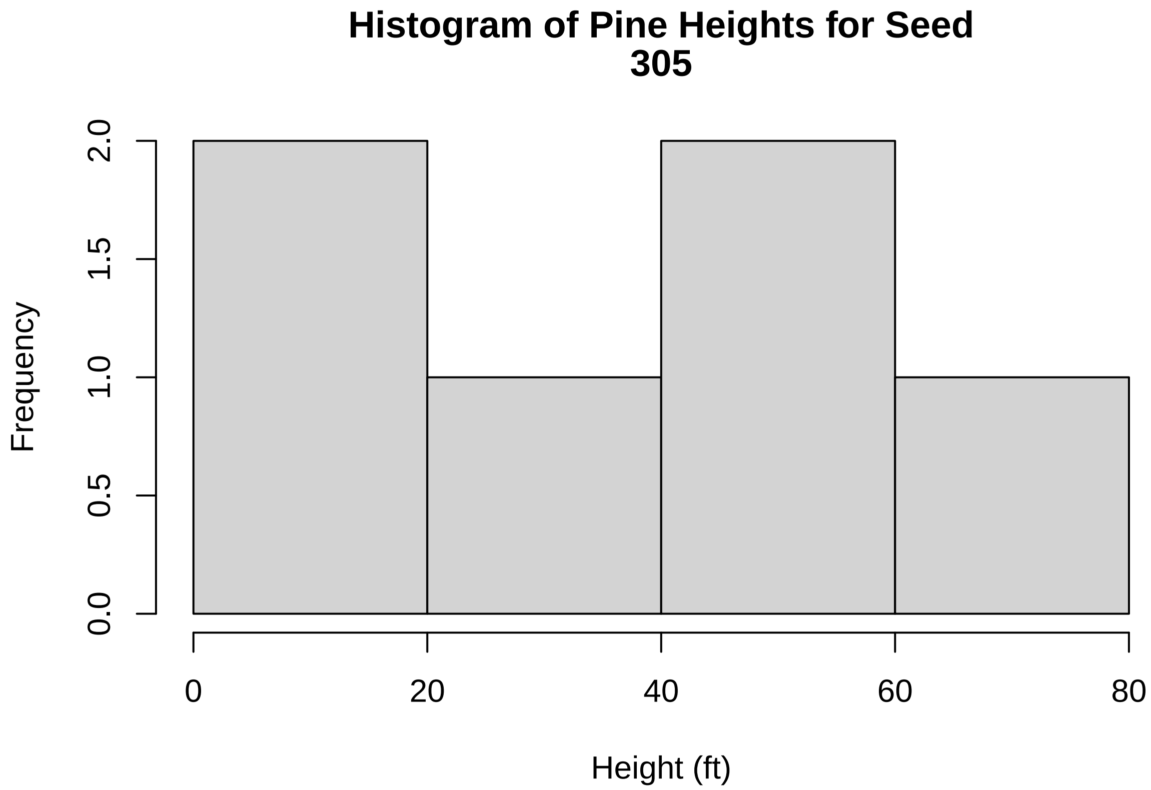
The Loblolly pine heights do not appear normally distributed, based on the shape of the histograms.

Exercise 5

```r
# build individual histograms for each subset for seeds 329, 315, and 305
for(i in levels(subset$Seed)){
    hist(Loblolly$height[Loblolly$Seed == i],
         main = c("Histogram of Pine Heights for Seed", i),
         xlab = "Height (ft)", ylab ="Frequency"
    )
}
```


Histogram of Pine Heights for Seed
329
Frequency
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50
60
Height (ft)


Histogram of Pine Heights for Seed
315
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Height (ft)



Answer

The Loblolly pine heights for each seed above do not appear to be normally distributed based on the individual histogram plots.

Exercise 6

How many observations are there for `height`*? How many different values are there for* `age`*? Show your R code to find the answer to the two questions above.*

```
length(Loblolly$height)
```

```
## [1] 84
```

```
length(unique(Loblolly$age))
```

```
## [1] 6
```

Answer

There are 84 observations for height in the Loblolly data set. There are 6 different unique values for age.

Exercise 7

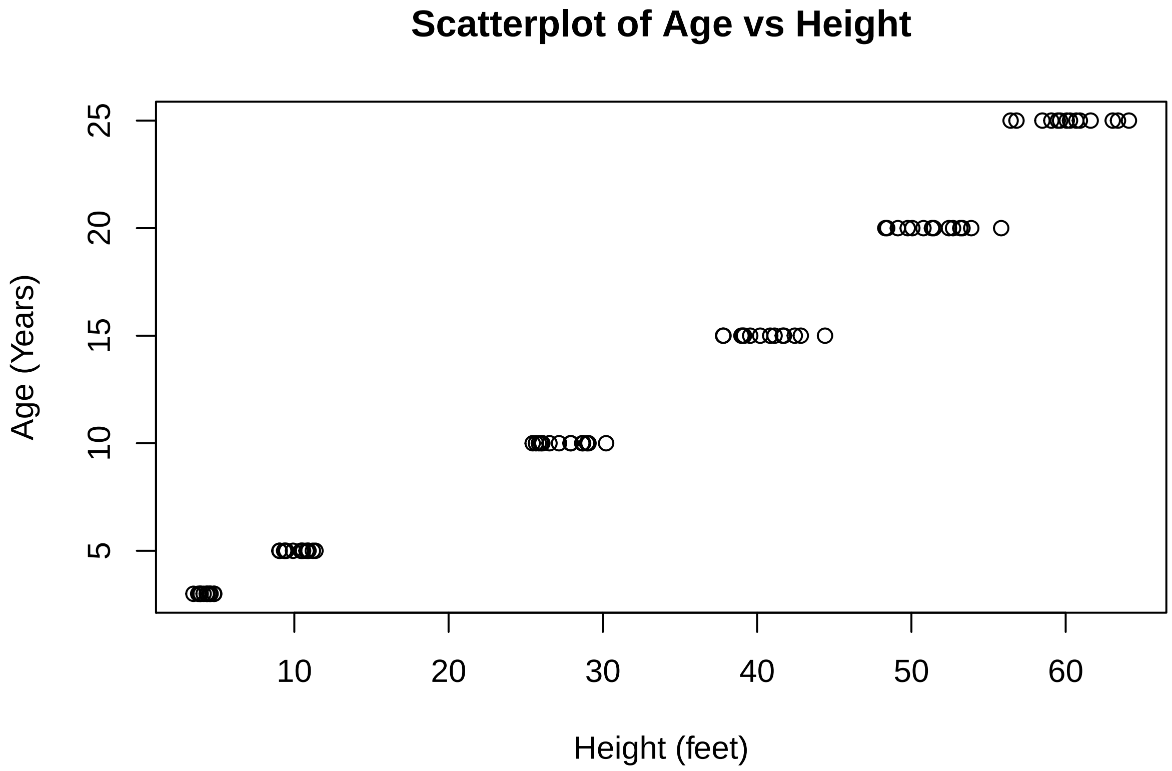
In this setting, which is the explanatory variable (predictor)? Which is the response?

Answer

In this setting, the explanatory variable or predictor is height, as we want to see how we can use this variable to predict the values of age.

Exercise 8

```
# scatterplot of age vs height from the Loblolly data
plot(x = Loblolly$height, y=Loblolly$age,
     main = "Scatterplot of Age vs Height",
     xlab = "Height (feet)", ylab = "Age (Years)")
```



Exercise 8

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

Answer

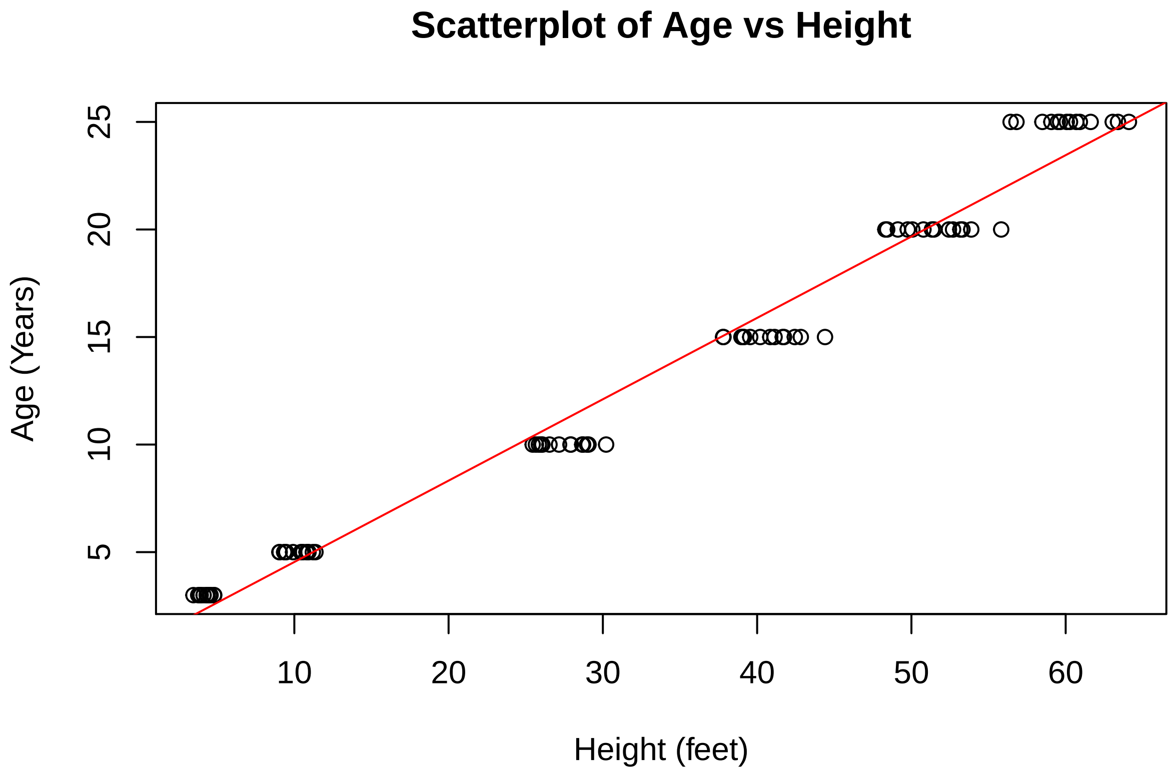
There seems to be evidence of a positive linear relationship between tree age and height. I would estimate the correlation between the variables to be about .4.

Exercise 9

Predict the age of a tree that is 20 feet tall for the regression line:

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

```
# use the function abline to add the above regression plot
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')
```



```
# predict age of a tree that is 20 feet tall for the regression line
# a = 0.7574, b = 0.3783, x = 20
0.7574 + 0.3783 * 20
```

```
## [1] 8.3234
```

Answer

For a tree that is 20 feet tall, we predict that the age will be 8.3234 years, based on the above regression equation.

On Your Own

Question 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.*

```
data("faithful")
?faithful
```

Part a:

What is the response variable? What is the explanatory variable?

Answer The response variable is waiting time between eruptions, while the explanatory variable is the duration of the eruptions.

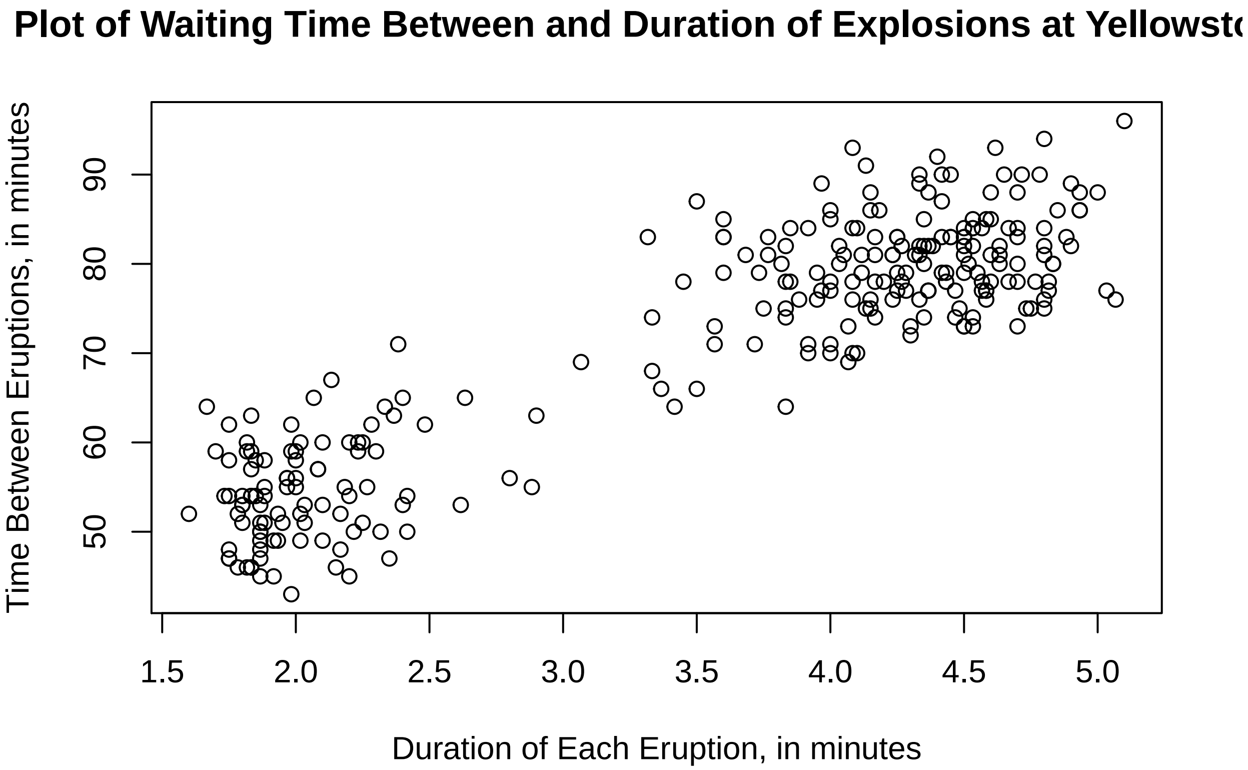
Part b:

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.

```
# find names of variables in faithful
attach(faithful)
names(faithful)
```

```
## [1] "eruptions" "waiting"
```

```
# scatterplot
plot(x = eruptions, y = waiting,
     main = "Plot of Waiting Time Between and Duration of Explosions at Yellowstone",
     xlab= "Duration of Each Eruption, in minutes",
     ylab= "Time Between Eruptions, in minutes"
     )
```



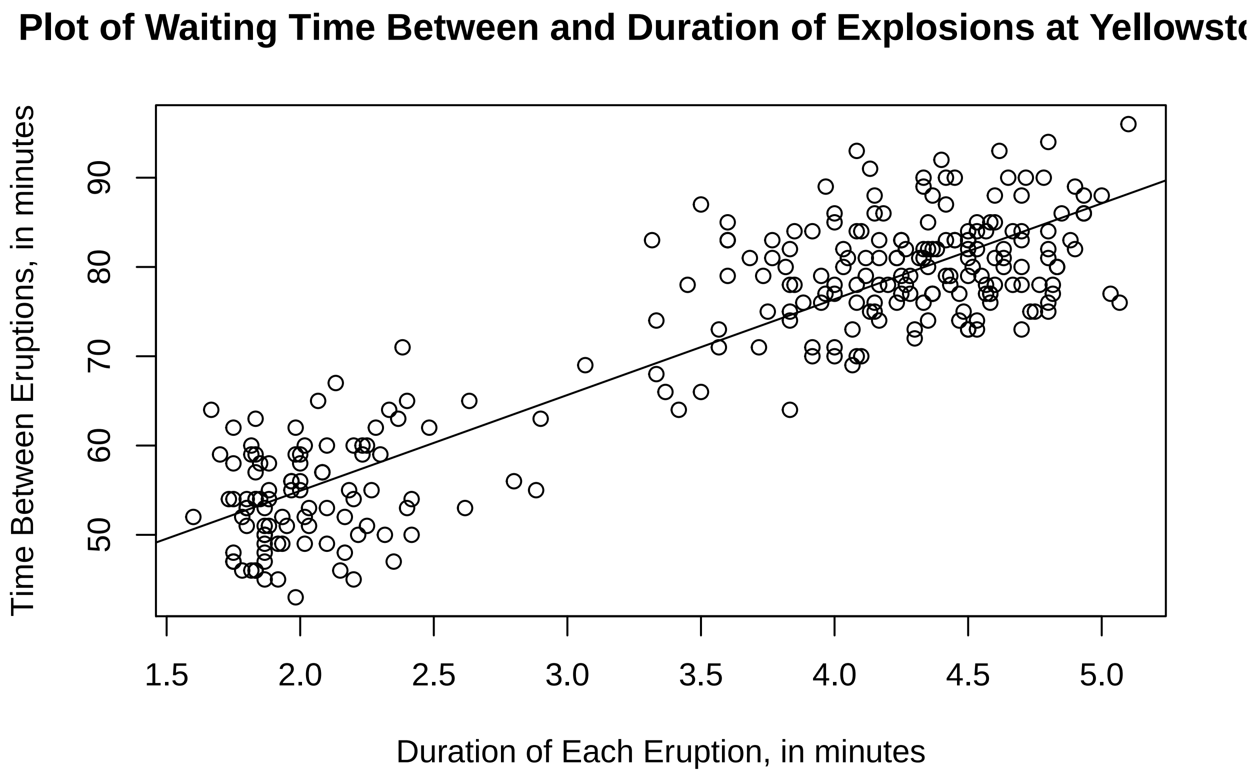
Part c:

Write down the R code to add this regression line to your scatter plot:

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

```
plot(x = eruptions, y = waiting,
     main = "Plot of Waiting Time Between and Duration of Explosions at Yellowstone",
     xlab= "Duration of Each Eruption, in minutes",
     ylab= "Time Between Eruptions, in minutes"
```

```
      )
abline(a = 33.47, b = 10.73)
```



Question 2:

The `ToothGrowth` *dataset in* `R` *gives information about the effect of Vitamin C on tooth growth in Guinea Pigs. Load this data into* `R` *using the command* `data("ToothGrowth")`*.*

```
data("ToothGrowth")
?ToothGrowth

ToothGrowth$supp <- factor(ToothGrowth$supp)
ToothGrowth$dose <- factor(ToothGrowth$dose)

names(ToothGrowth)
```

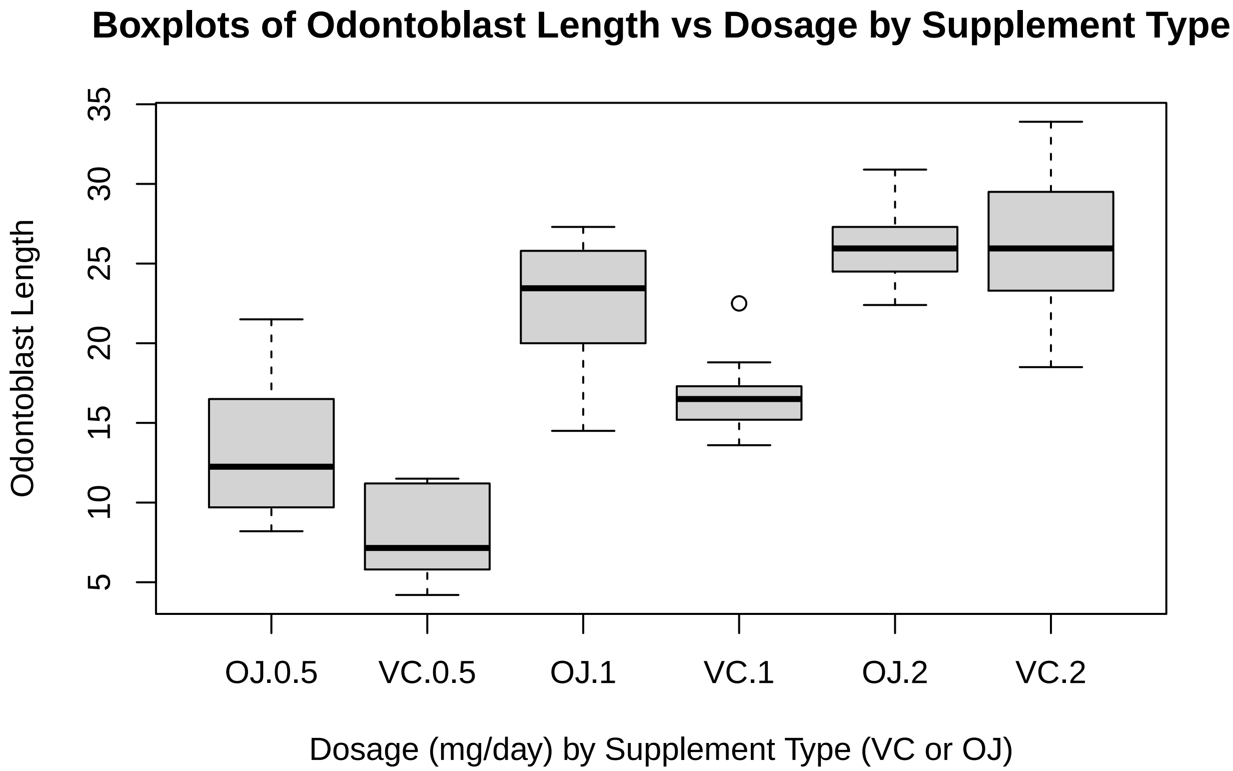
```
## [1] "len"  "supp" "dose"
```

```
attach(ToothGrowth)
```

Part a:

*Create boxplot 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. Write down your R code below. Hint: you can create factor levels in a formula by using factorA * factorB*

```r
boxplot(len ~ supp * dose,
        main = "Boxplots of Odontoblast Length vs Dosage by Supplement Type",
        xlab = "Dosage (mg/day) by Supplement Type (VC or OJ)",
        ylab = "Odontoblast Length",
        )
```



Part b:

Based on your boxplots, do you think there are differences among the six treatments?

Answer

Based on my boxplots, I do think there are differences among the six treatments. The boxplot indicate different ranges and different means for almost every one of the treatments.