Untitled

Krista Pipho

September 21, 2018

Question 1. Summary Statistics

```
# First we open the "countries2005_groomed" data file and view the variabe names
countries <- read.csv("C:\\Users\\xenon\\Desktop\\R Studio 2018\\countries2005_groomed.csv")

names(countries)
```

```
##  [1] "Country"
##  [2] "Continent"
##  [3] "Birth_rate_crude_per_1000_people"
##  [4] "Death_rate_crude_per_1000_people"
##  [5] "Life_expectancy_at_birth_female_years"
##  [6] "Life_expectancy_at_birth_male_years"
##  [7] "Life_expectancy_at_birth_total_years"
##  [8] "Literacy_rate_adult_total"
##  [9] "Personal_computers_per_100_people"
## [10] "Physicians_per_1000_people"
## [11] "Prevalence_of_HIV_total"
```

Now we will generate summary statistics on various aspects of the data, including an arument that allows for numerical analysis dispite the presence of NA values

```
#This portion calculates the summary values for all variables and stores them

mean_life = mean(countries$Life_expectancy_at_birth_total_years,na.rm=TRUE)
median_life = median(countries$Life_expectancy_at_birth_total_years,na.rm=TRUE)
sd_life = sd(countries$Life_expectancy_at_birth_total_years,na.rm=TRUE)

mean_mlife = mean(countries$Life_expectancy_at_birth_male_years,na.rm=TRUE)
median_mlife = median(countries$Life_expectancy_at_birth_male_years,na.rm=TRUE)
sd_mlife = sd(countries$Life_expectancy_at_birth_male_years,na.rm=TRUE)

mean_flife = mean(countries$Life_expectancy_at_birth_female_years,na.rm=TRUE)
median_flife = median(countries$Life_expectancy_at_birth_female_years,na.rm=TRUE)
sd_flife = sd(countries$Life_expectancy_at_birth_female_years,na.rm=TRUE)

mean_comp = mean(countries$Personal_computers_per_100_people,na.rm=TRUE)
median_comp = median(countries$Personal_computers_per_100_people,na.rm=TRUE)
sd_comp = sd(countries$Personal_computers_per_100_people,na.rm=TRUE)

mean_phys = mean(countries$Physicians_per_1000_people,na.rm=TRUE)
median_phys = median(countries$Physicians_per_1000_people,na.rm=TRUE)
sd_phys = sd(countries$Physicians_per_1000_people,na.rm=TRUE)

#This portion prints the stored values for each variable
print("Life Expectancy")
```

```
## [1] "Life Expectancy"
```

```
mean_life
```

```
## [1] 67.65946
```

```
median_life
```

```
## [1] 71
```

```
sd_life
```

```
## [1] 11.0221
```

```
print("Male Life Expectancy")
```

```
## [1] "Male Life Expectancy"
```

```
mean_mlife
```

```
## [1] 65.41935
```

```
median_mlife
```

```
## [1] 68
```

```
sd_mlife
```

```
## [1] 10.43209
```

```
print ("Female Life Expectancy")
```

```
## [1] "Female Life Expectancy"
```

```
mean_flife
```

```
## [1] 70.05882
```

```
median_flife
```

```
## [1] 74
```

```
sd_flife
```

```
## [1] 11.65553
```

```
print ("Personal computers per 100 people")
```

```
## [1] "Personal computers per 100 people"
```

```
mean_comp
```

```
## [1] 15.55866
```

```
median_comp
```

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

```
sd_comp
```

```
## [1] 22.85569
```

```
print ("Number of physicians per 1000 people")
```

```
## [1] "Number of physicians per 1000 people"
```

```
mean_phys
```

```
## [1] 2
```

```
median_phys
```

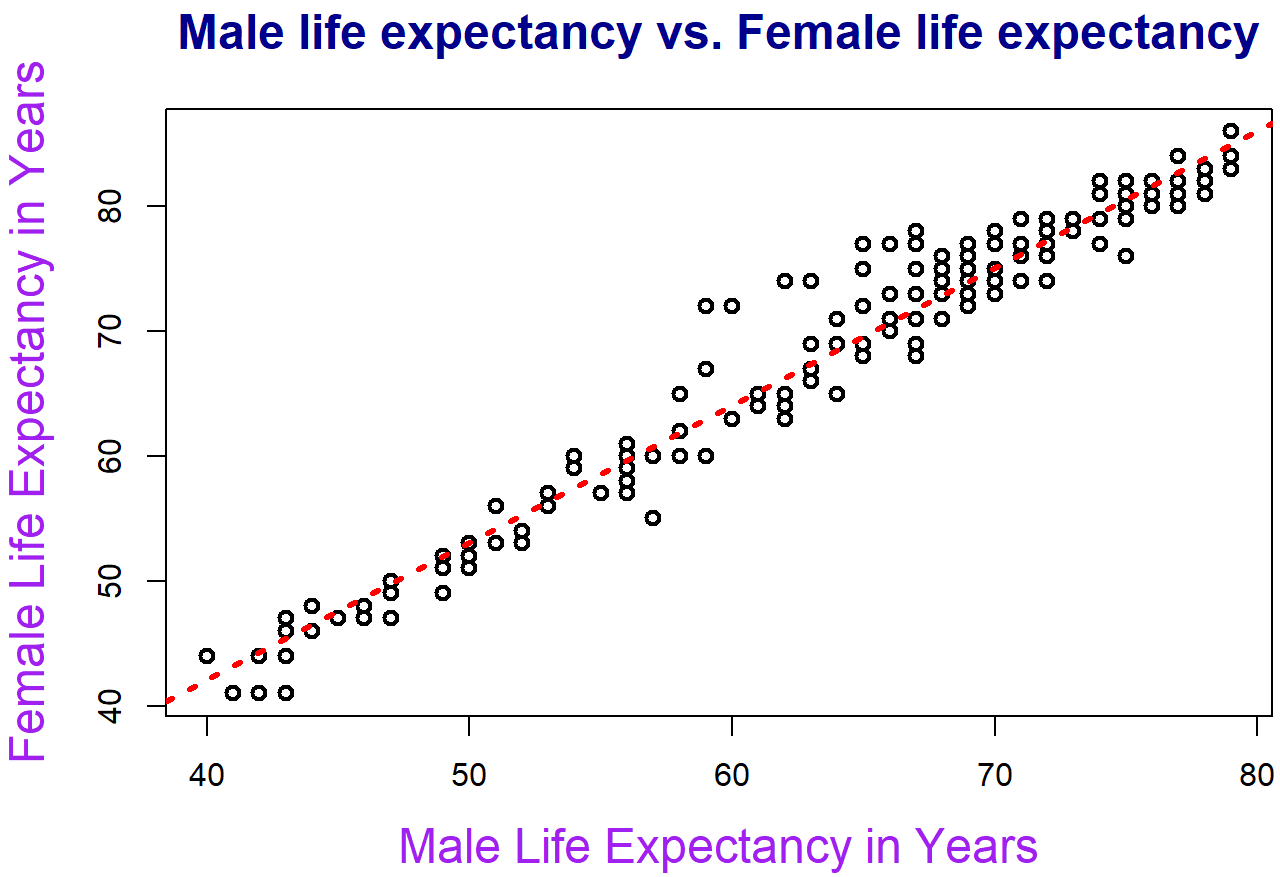
```
## [1] 2
```

```
sd_phys
```

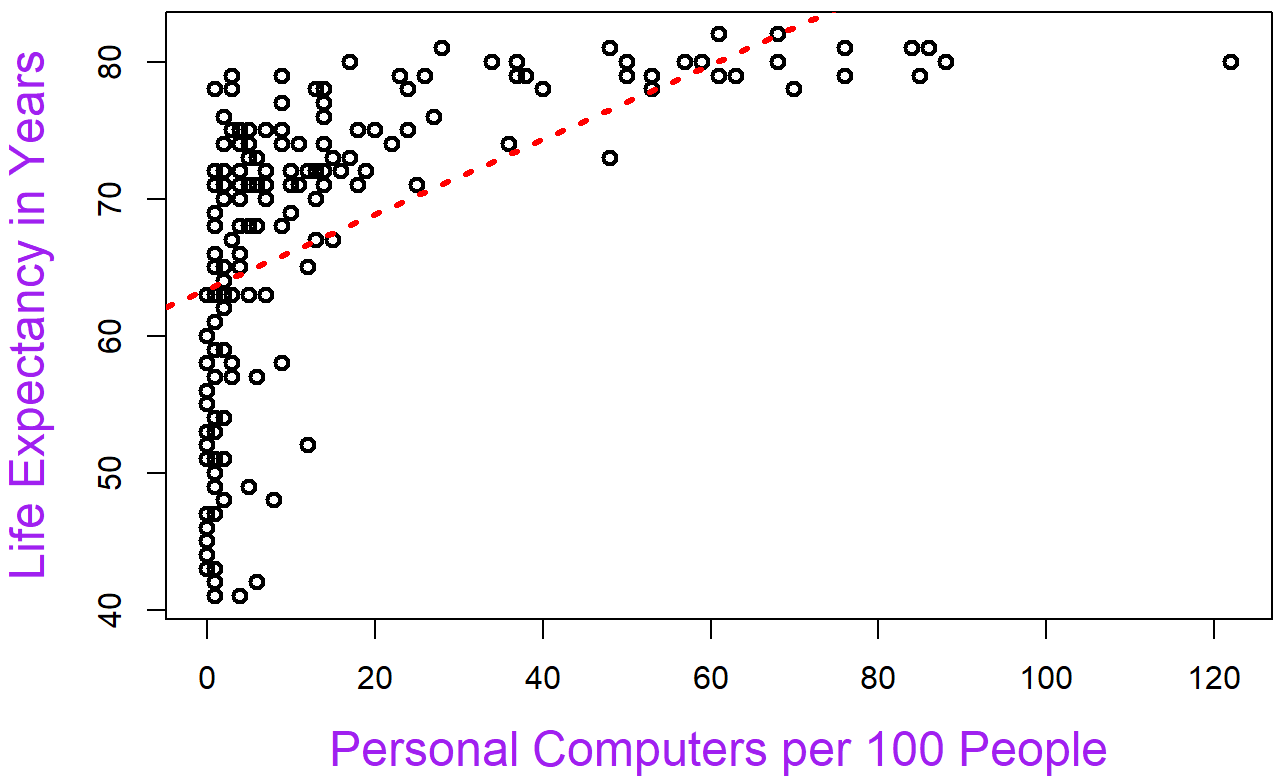
```
## [1] 1.285369
```

a. Now that we have summaries for each variable we move on to plotting relationships between variables

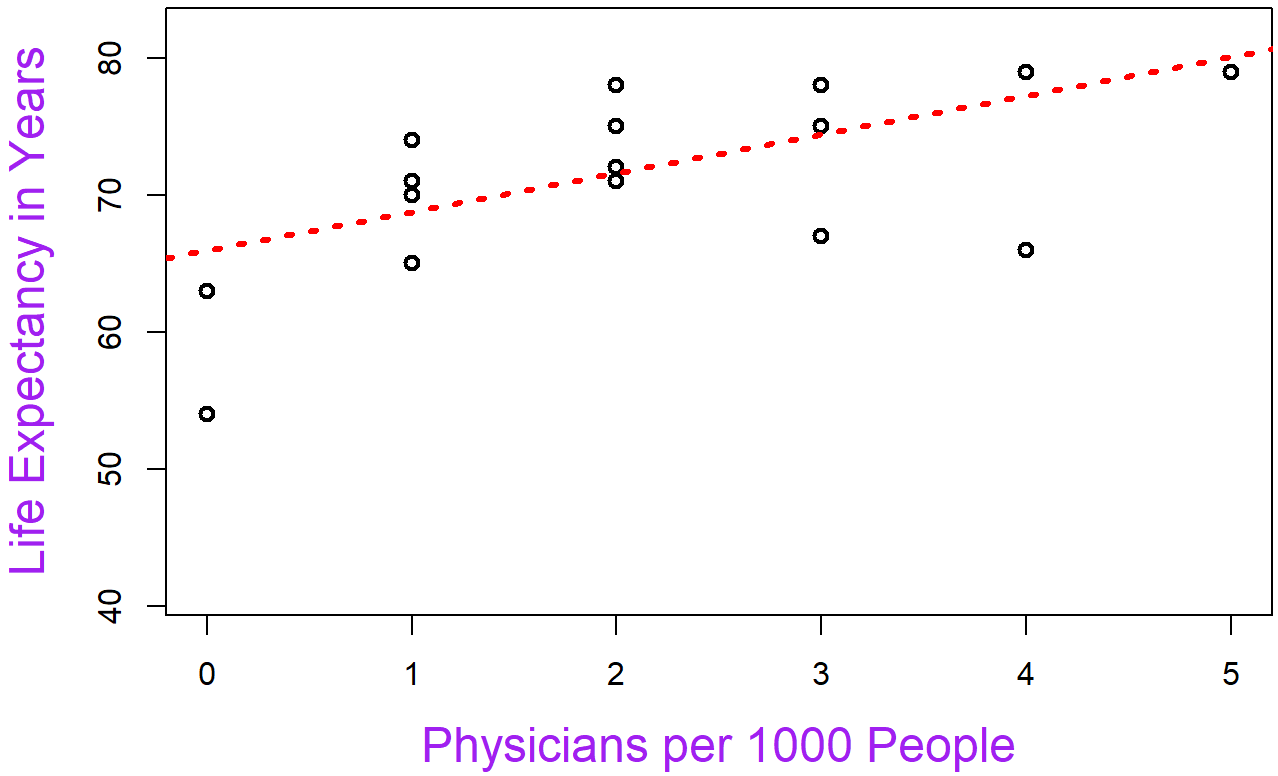
```
# Scater plot of Male life expectancy vs. Female life expectancy
plot(countries$Life_expectancy_at_birth_male_years,countries$Life_expectancy_at_birth_female_years, main="M
ale life expectancy vs. Female life expectancy",
     xlab = "Male Life Expectancy in Years", ylab = "Female Life Expectancy in Years",
     col.lab='purple',
     col.main = 'dark blue',
     cex.lab = 1.5,
     cex.main=1.5,
     lwd = 2)
abline(lm(countries$Life_expectancy_at_birth_female_years~countries$Life_expectancy_at_birth_male_years),
       col='red',
       lwd = 3,
       lty = "dotted")
```




```
# Scatter plot of Personal computers vs. Total life expectancy
plot(countries$Personal_computers_per_100_people,countries$Life_expectancy_at_birth_total_years, main="Pers
onal Computers vs. Life Expectancy",
     xlab = "Personal Computers per 100 People", ylab = "Life Expectancy in Years",
     col.lab='purple',
     col.main = 'dark blue',
     cex.lab = 1.5,
     cex.main=1.5,
     lwd = 2)
abline(lm(countries$Life_expectancy_at_birth_total_years ~ countries$Personal_computers_per_100_people),
       col='red',
       lwd = 3,
       lty = "dotted")
```



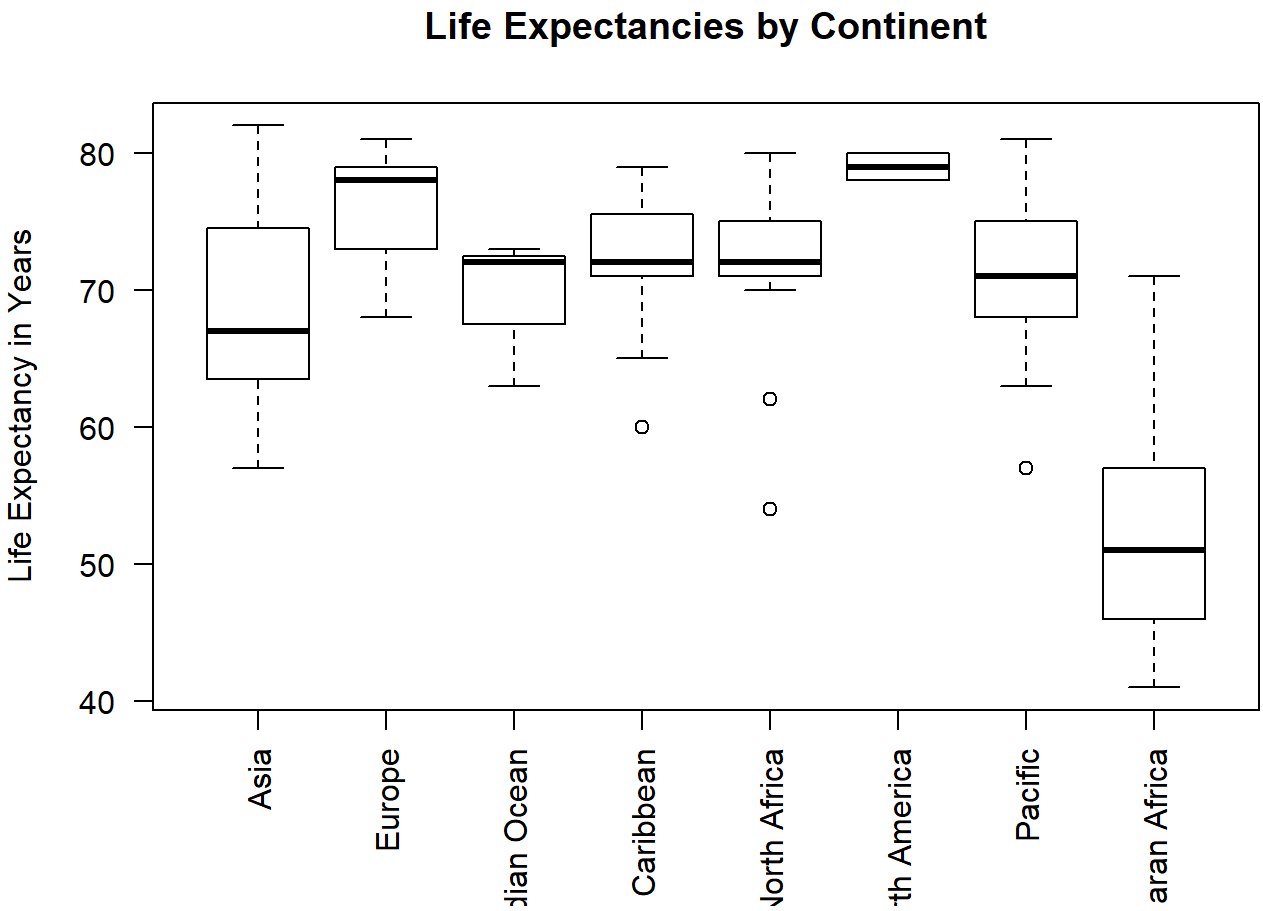

```
# Scatter plot of Physicians vs. Total life expectancy
plot(countries$Physicians_per_1000_people, countries$Life_expectancy_at_birth_total_years, main="Physicians
 vs. Life Expectancy",
     xlab = "Physicians per 1000 People", ylab = "Life Expectancy in Years",
     col.lab='purple',
     col.main = 'dark blue',
     cex.lab = 1.5,
     cex.main=1.5,
     lwd = 2)
abline(lm(countries$Life_expectancy_at_birth_total_years ~ countries$Physicians_per_1000_people),
       col='red',
       lwd = 3,
       lty = "dotted")
```



b. The relationship between personal computers and life expectancy is clearly not linear over this range. The relationship between life expectancy and number of physicians is relatively linear. Both are positive correlations. The linear approximation of the relationship for computers has a less steep slope(~10 years life expectancy / 20 computers per hundred people, or ~1 year life expectancy / 20 computers per thousand people) than the linear trendline for physicians (~5 years life expectancy / 1 physician per thousand people). Physicians and life expectancy have a stronger and more consistant correlation than computers and life expectancy. Computers only seem to have a correlation at very high concentrations. As ever, correlation does not imply causation, and it is not neccisarily true that either of these variables 'explain' life expectancy.

c. Now we will explore the relationship between continent and life expectancy.

```
# Boxplot of life expectancy by continent
boxplot(countries$Life_expectancy_at_birth_total_years~countries$Continent,main="Life Expectancies by Conti
nent", ylab="Life Expectancy in Years", las = 2)
```



Question 2. Blood Pressure Boxplot

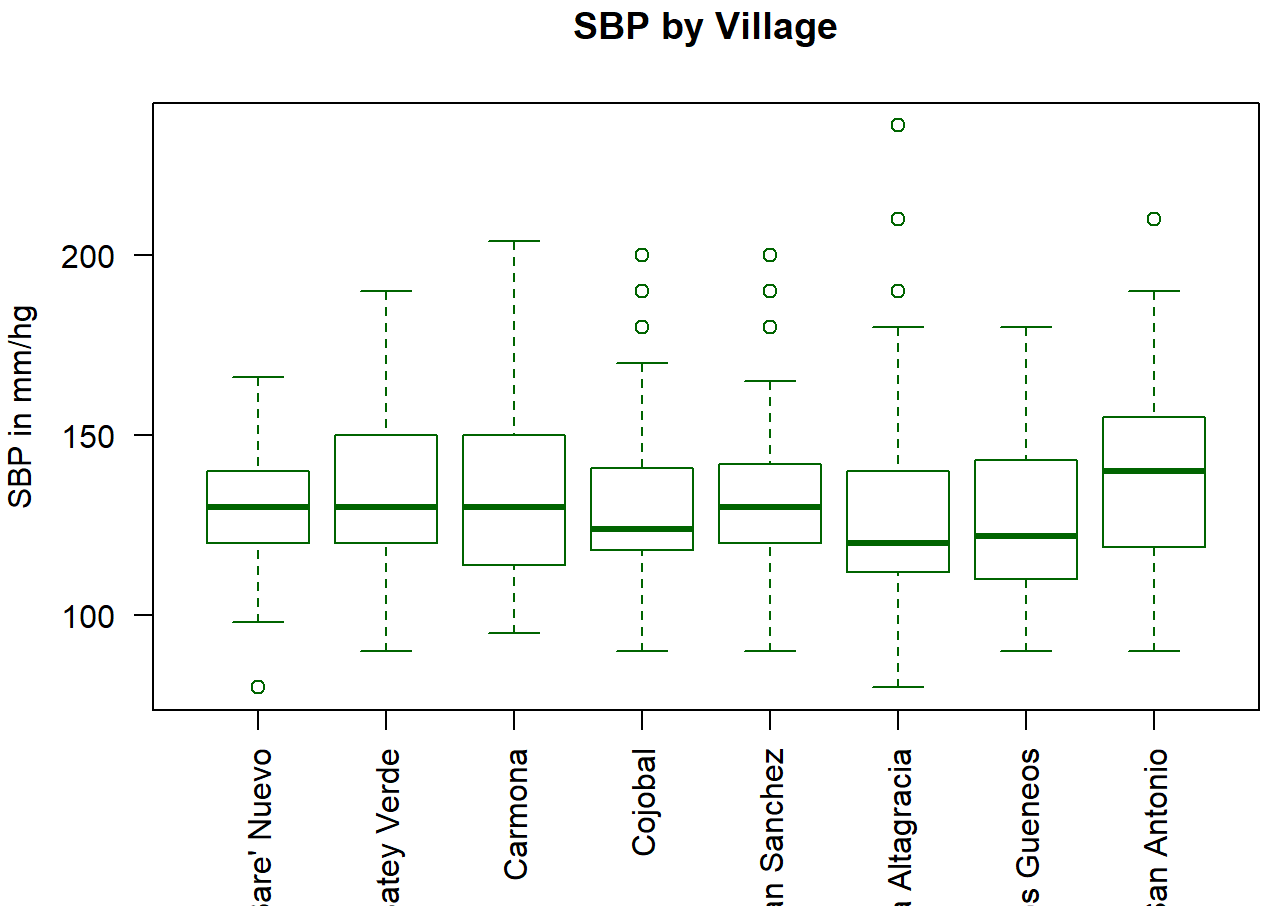
```
# Here we open the "BP" data file and view the variabe names

BP <- read.csv("C:\\Users\\xenon\\Desktop\\R Studio 2018\\BP.csv")

names(BP)
```

```
## [1] "Village" "Gender"  "Age"     "SBP"     "DBP"
```

```
# Boxplot of Systolic Blood Pressure by Village
boxplot(BP$SBP~BP$Village,main="SBP by Village", ylab="SBP in mm/hg", las = 2, border = "dark green")
```



Question 3. Bootstrap Test

a. First we will do a simple visualization of the data

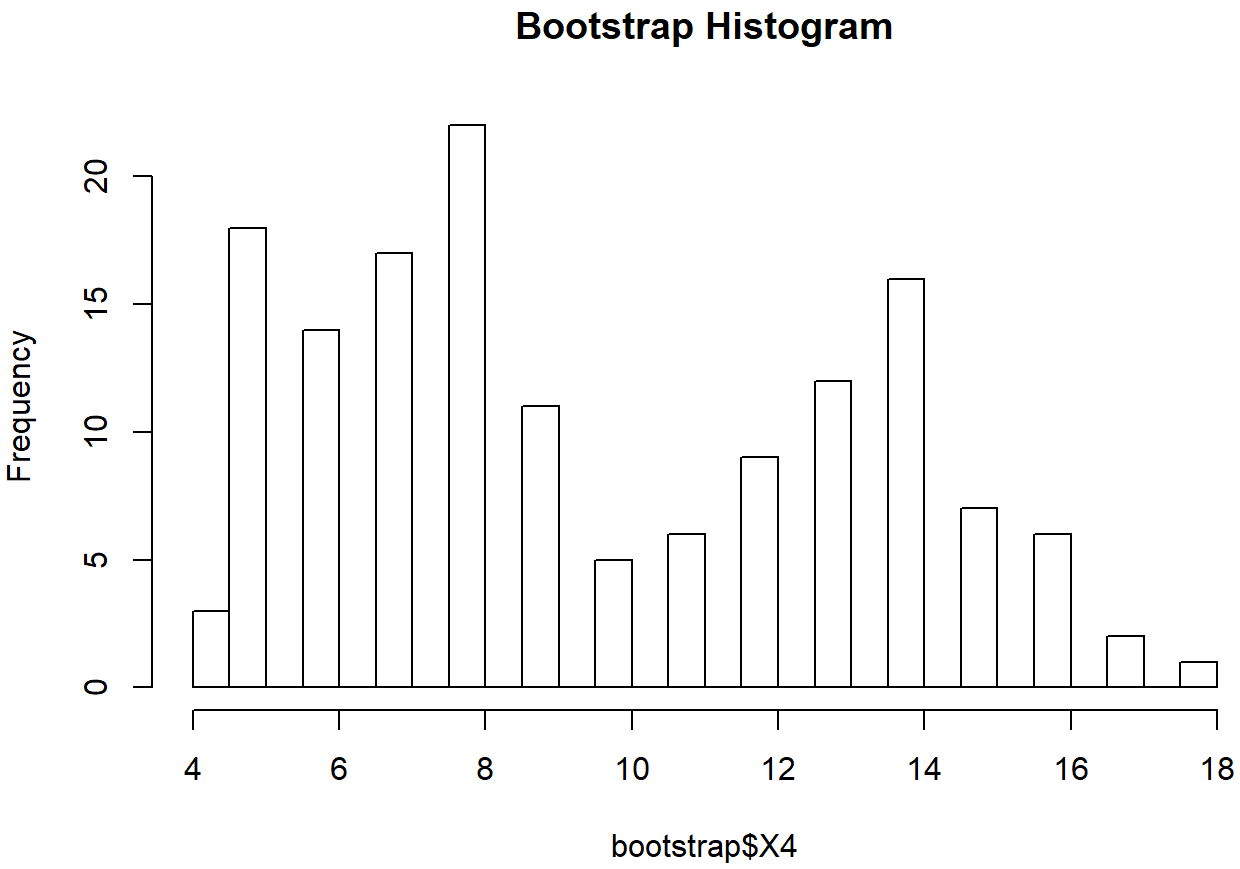
```
# Here we open the "bootstrap_test" data file and view the variabe names

bootstrap <- read.csv("C:\\Users\\xenon\\Desktop\\R Studio 2018\\bootstrap_test.csv")

names(bootstrap)
```

```
## [1] "X4"
```

```
# This creates a simple histogra of the values in bootstrap$X4
hist(bootstrap$X4, main ="Bootstrap Histogram", breaks = 20)
```



b) Next, we will explore some quantitative properties of the data

```
# Calculates the mean of bootstrap$X4
mean_boot = mean(bootstrap$X4)

# Calculates the 95% confidance interval of bootstrap$X4 using length(bootstrap$X4)-1 to express the degree
s of freedom
confidence <- qt(c(0.025,0.975),length(bootstrap$X4)-1)

# Prints the values that we saved above
print(mean_boot)
```

```
## [1] 9.644295
```

```
print(confidence)
```

```
## [1] -1.976122  1.976122
```

c. Now we will use bootstrap as an alternet means of calculating confidance intervals

```
#creates an empty list that we will fill with simulated mean values
bstrap<-c()
#ensures that it is empty by calling it before the loop has run
bstrap
```

```
## NULL
```

```
# Picks 149 values 1000 times, from the bootstrap$X4 data set- selections are made with repacement.
for(i in 1:1000){
  bsample<-sample(bootstrap$X4,149,replace=T)
  # saves generated list as a sample
  bestimate<-mean(bsample)
  # populates the originally empty list
  bstrap<-c(bstrap,bestimate)}
# calculate the quantiles of the bstrap
quantile(bstrap,0.05)
```

```
##       5%
## 9.187584
```

```
quantile(bstrap,0.95)
```

```
##      95%
## 10.15436
```

```
quantile(bstrap,0.025)
```

```
##     2.5%
## 9.066946
```

```
quantile(bstrap,0.975)
```

```
##    97.5%
## 10.26879
```