Assignment4_XinHuang

Xin Huang

Answers to Question1

1a)

```
myData <- read.csv("HW01pb1data.csv", header = FALSE)
#exam all the columns
class(myData$V1)
```

```
## [1] "integer"
```

```
class(myData$V2)
```

```
## [1] "integer"
```

```
class(myData$V3)
```

```
## [1] "integer"
```

```
class(myData$V4)
```

```
## [1] "factor"
```

```
class(myData$V5)
```

```
## [1] "factor"
```

Given the resluts, we can see that column V1, V2, V3 are quantitative V4 and V5 are qualitative.

1b)

```
levels(myData$V4)
```

```
##  [1] "0"           "10"          "100"         "110"         "120"
##  [6] "140"         "15"          "150"         "160"         "20"
## [11] "200"         "25"          "30"          "35"          "40"
## [16] "5"           "50"          "55"          "60"          "65"
## [21] "70"          "80"          "85"          "90"          "thirty five"
```

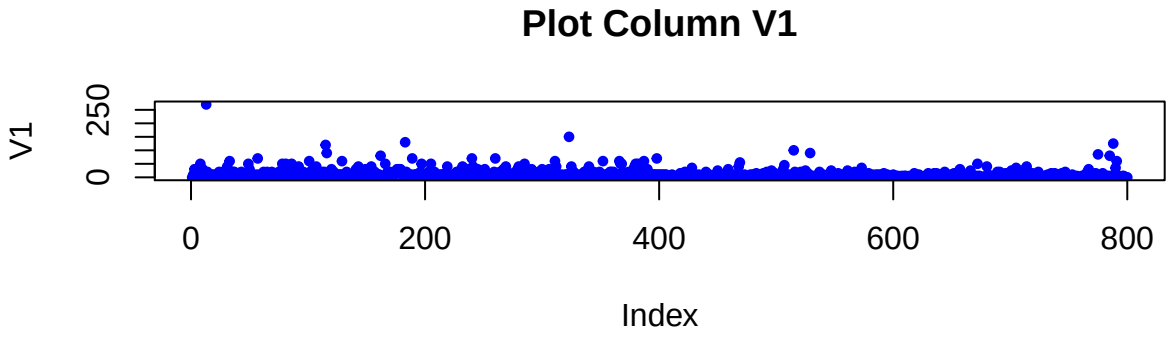
```
levels(myData$V5)
```

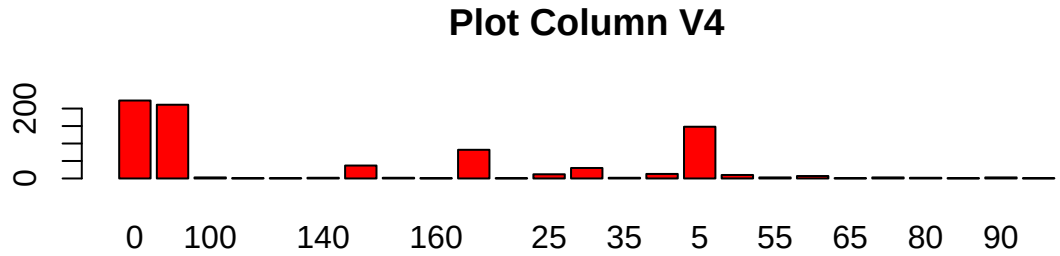
```
##  [1] "0"           "10"          "120"         "140"         "15"
##  [6] "20"          "25"          "255"         "30"          "35"
## [11] "40"          "45"          "5"           "50"          "55"
## [16] "60"          "70"          "80"          "twenty five"
```

By printing out all the levels of column V4 and V5, we can see that they both contain data different data type. So when the data was read in, they can not be treated as numeric but as factors.

1c)

```
mat <- matrix(1 : 2, nrow = 2)
layout(mat)
plot(myData[, 1], pch = 20, col = "blue",
     cex = 0.9, ylab = "V1",
     main = "Plot Column V1")
plot(myData[, 4], col = "red",
     main = "Plot Column V4")
```





In the first pic, it plots column 1 scatters data on a x-y axis. It uses index and value as a x-y values In the second pic, it plots column 4 as a histogram graph. It uses factors to count how many element are in each factors.

Answers to Question2

2a)

```
#Read original data and generate sample
myData <- read.csv("HW01pb2data.csv",header=FALSE)
sampleData <- sample(myData[, 1], 10000, replace=TRUE)
```

2b)

Compute those values using following functions:

```
paste("Mean of sample data: ", mean(sampleData))
```

```
## [1] "Mean of sample data:  9.43700644075614"
```

```r
paste("Max of sample data: ", max(sampleData))
```

```
## [1] "Max of sample data:  17.403285578256"
```

```r
paste("Var of sample data: ", var(sampleData))
```

```
## [1] "Var of sample data:  4.02784646189728"
```

```r
paste("Quantile of sample data: ", quantile(sampleData, 0.25))
```

```
## [1] "Quantile of sample data:  8.07856983918223"
```

2c)

Compute those values on original data:

```r
paste("Mean of whole data: ", mean(myData[, 1]))
```

```
## [1] "Mean of whole data:  9.4514680349268"
```

```r
paste("Max of whole data: ", max(myData[, 1]))
```

```
## [1] "Max of whole data:  18.9665681608958"
```

```r
paste("Var of whole data: ", var(myData[, 1]))
```

```
## [1] "Var of whole data:  4.00182160383524"
```

```r
paste("Quantile of whole data: ", quantile(myData[, 1], 0.25))
```

```
## [1] "Quantile of whole data:  8.10388024879266"
```

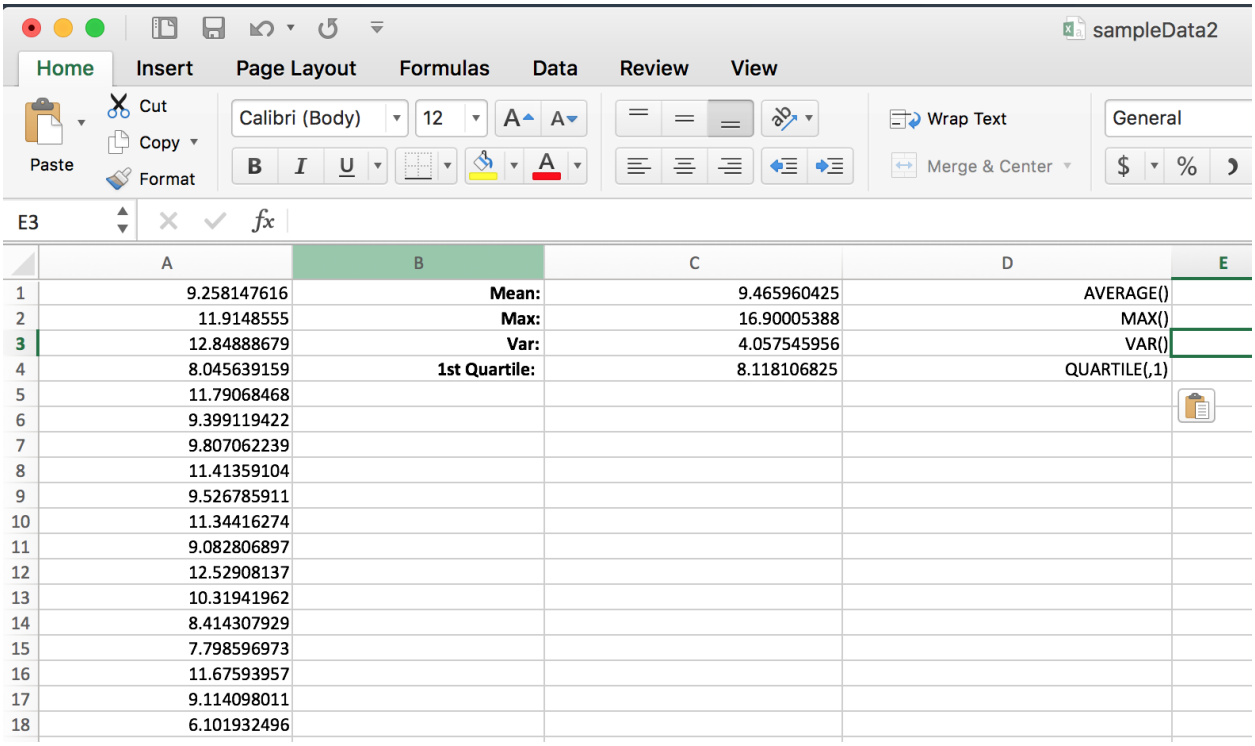
2d)

Write the data into a csv file:

```r
write.csv(sampleData, file = "sampleData.csv", row.names= FALSE, col.names = FALSE)
```

```
## Warning in write.csv(sampleData, file = "sampleData.csv", row.names =
## FALSE, : attempt to set 'col.names' ignored
```

Compute the values using Excel functions:



	A	B	C	D	E
1	9.258147616	**Mean:**	9.465960425	AVERAGE()	
2	11.9148555	**Max:**	16.90005388	MAX()	
3	12.84888679	**Var:**	4.057545956	VAR()	
4	8.045639159	**1st Quartile:**	8.118106825	QUARTILE(,1)	
5	11.79068468				
6	9.399119422				
7	9.807062239				
8	11.41359104				
9	9.526785911				
10	11.34416274				
11	9.082806897				
12	12.52908137				
13	10.31941962				
14	8.414307929				
15	7.798596973				
16	11.67593957				
17	9.114098011				
18	6.101932496				

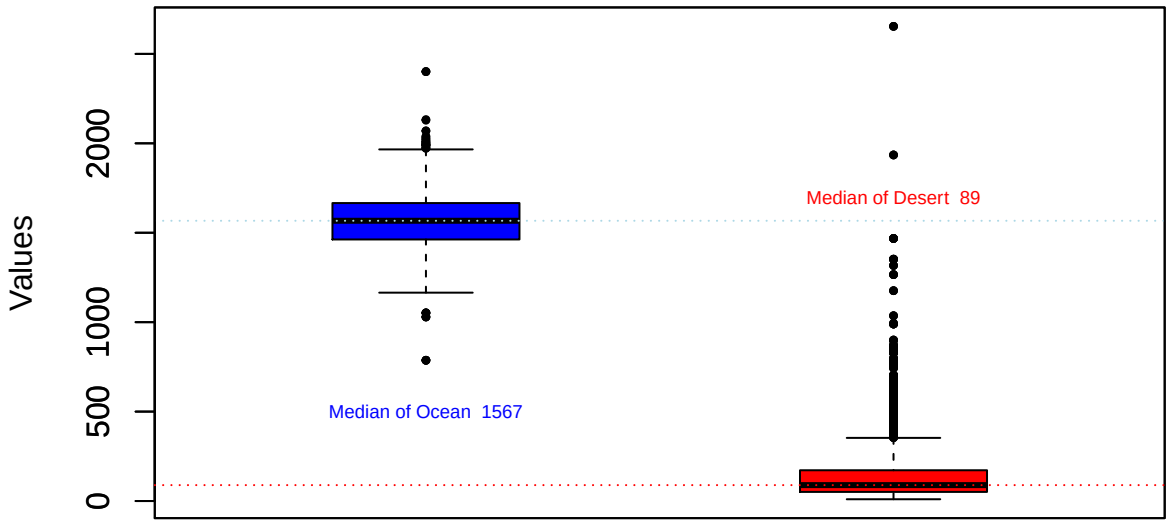
Answers to Question3

3a)

```
ocean <- read.csv("HW01pb3OceanViewdata.csv", header = FALSE)
desert <- read.csv("HW01pb3Desertdata.csv", header = FALSE)

boxplot(ocean, at = 1, xlim = c(0.5, 2.5),
        ylim = range(c(ocean, desert)),
        main = "House Box Plots",
        ylab="Values",
        pch = 20,
        cex = 0.7,
        col ="blue")
abline(h = median(ocean$V1), col = "lightblue", lty = 3)
text(1,  500,
     paste("Median of Ocean ", median(ocean$V1)),
     col = "blue", cex = 0.6)

boxplot(desert, at = 2,
        add = TRUE, pch = 20,
        col ="red", cex = 0.7)
abline(h = median(desert$V1), col="red", lty = 3)
text(2,  1700,
     paste("Median of Desert ", median(desert$V1)),
     col = "red", cex = 0.6)
```

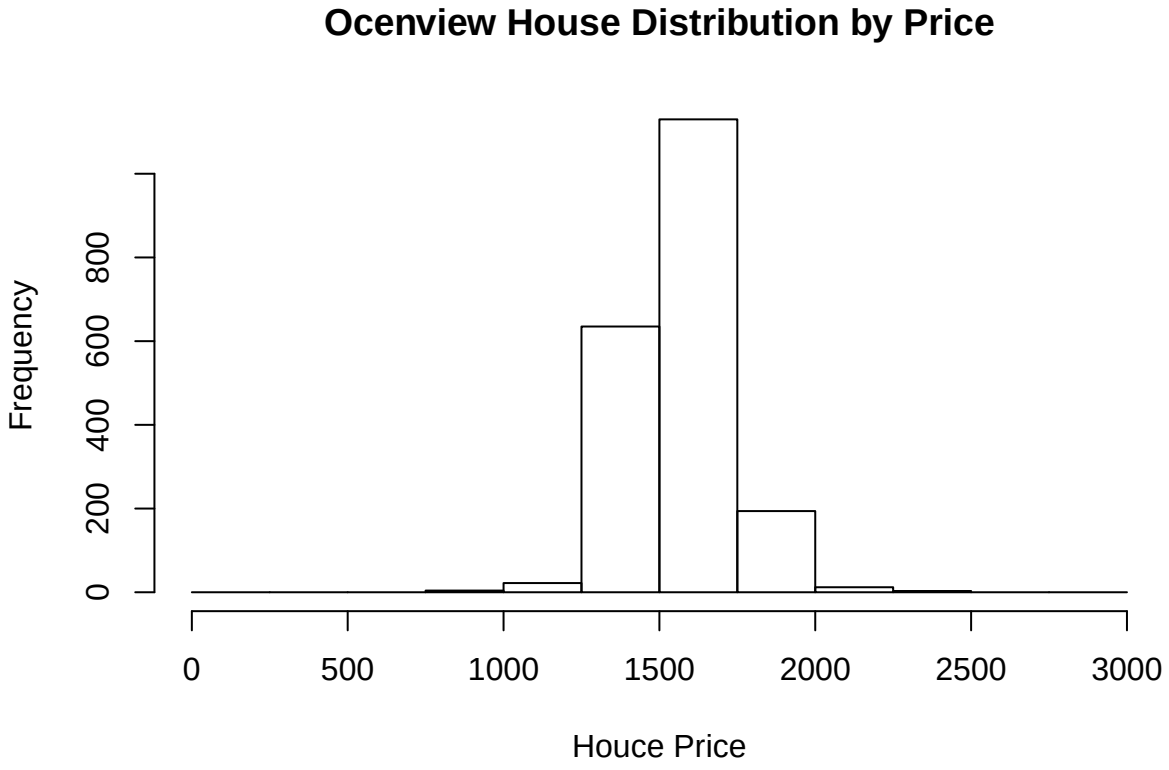


From the graphs, we can find out that: 1. The average prices of ocean view houses is much higher than that of houses in the desert. Also median prices of ocean view hourse is much higher too.

The data of ocean view whichc is houses is almost symmetrically distributed. On the other hand, the data of desert houses is more on one side

3b)

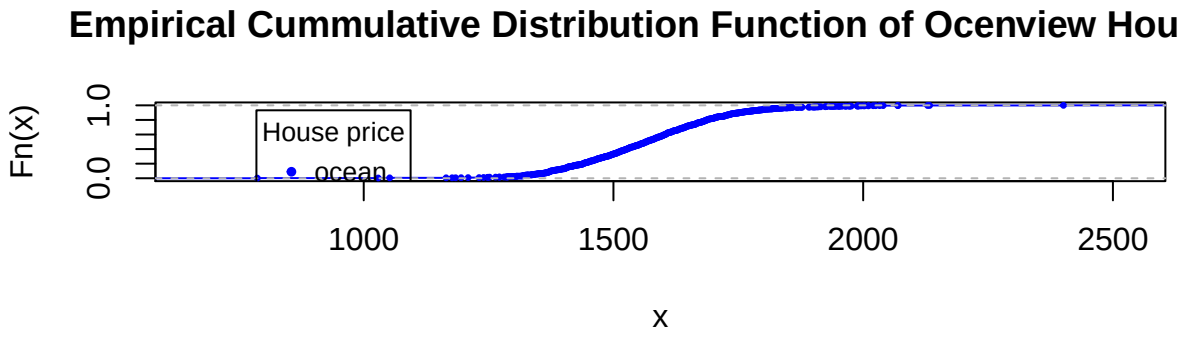
```
names(ocean)[1] <- "HousePrice"
names(desert)[1] <- "HousePrice"
breaks <- seq.int(0, 3000, by = 250)
hist(ocean$HousePrice, breaks, main = "Ocenview House Distribution by Price",
     xlab = "Houce Price")
```

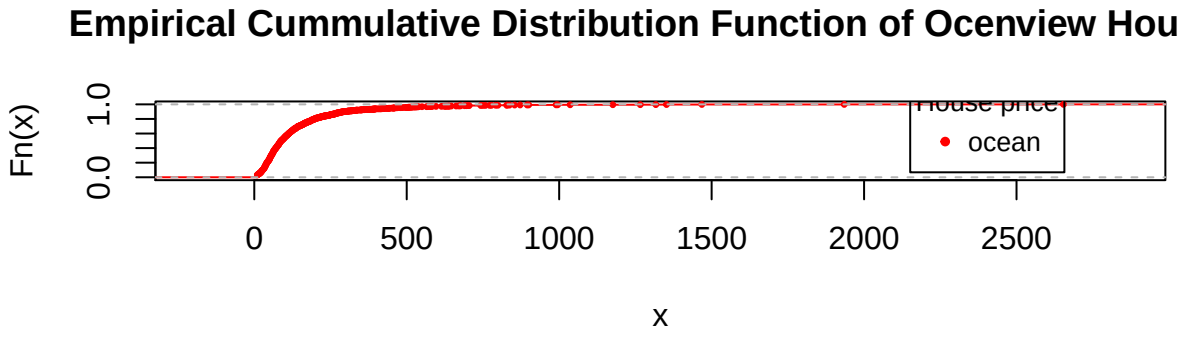


3c)

```
par(mfrow=c(2,1))
plot(ecdf(ocean$HousePrice),
     pch = 20 , cex = 0.5, col = "blue",
     main = "Empirical Cummulative Distribution Function of Ocenview Houses ")
legend("topleft", c("ocean"), col = c("blue"),
       pch = 20 ,inset =.1, title = "House price",
       cex = 0.8)

plot(ecdf(desert$HousePrice),
     pch = 20 , cex = 0.5, col = "red",
     main = "Empirical Cummulative Distribution Function of Ocenview Houses ")
legend("bottomright", c("ocean"), col = c("red"),
       pch = 20 ,inset =.1, title = "House price",
       cex = 0.8)
```

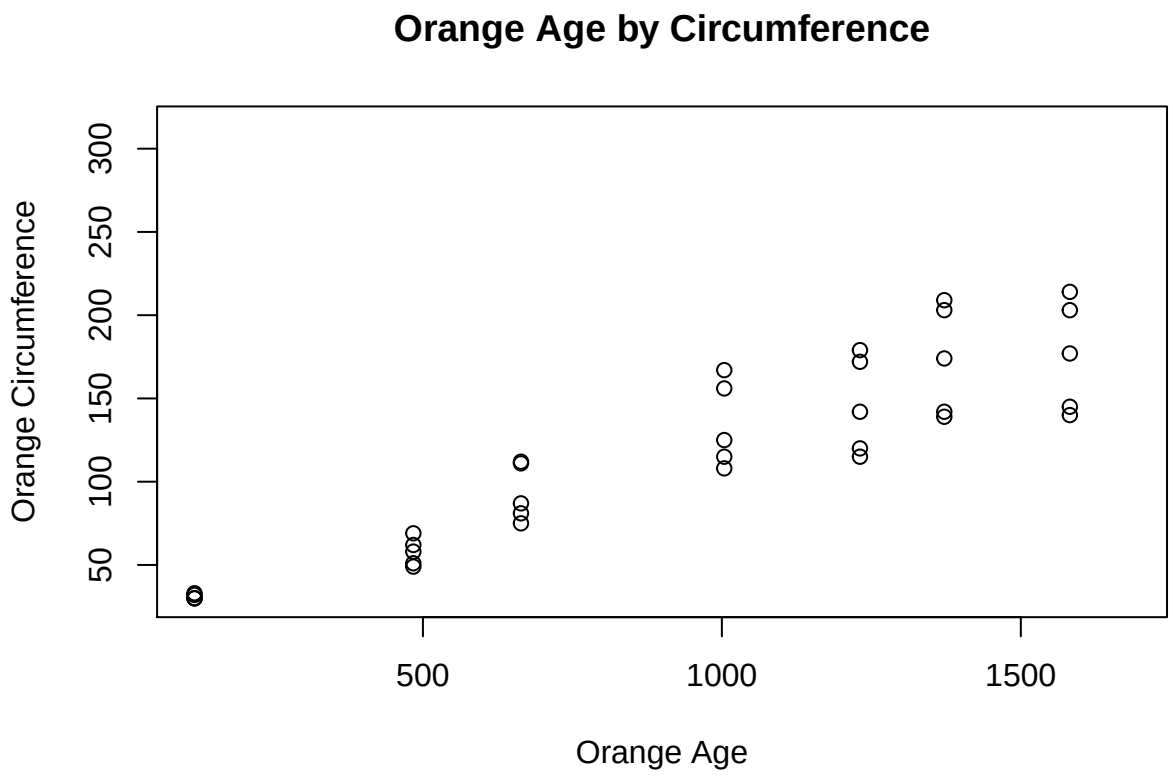




Answers to Question4

4a)

```
orange <- as.data.frame(Orange)
par(mfrow=c(1,1))
plot(orange$age, orange$circumference,
     main = "Orange Age by Circumference",
     xlim = c(min(orange$age),max(orange$age) + 100),
     ylim = c(min(orange$circumference),max(orange$circumference) + 100),
     xlab = "Orange Age",
     ylab = "Orange Circumference")
```



4b)

```
cor <- cor(orange[which(orange$Tree == 1), 2], orange[which(orange$Tree == 1), 3])
paste("Correlation between Age and Circumference is: ", cor)
```

```
## [1] "Correlation between Age and Circumference is:  0.985467542479218"
```

4c)

Use functions from ddply() to compute the resutls:

```
names(orange) <- toupper(names(orange))
result <- orange %>%
    group_by(TREE) %>%
    summarise(COVARIANCE = cov(AGE, CIRCUMFERENCE), CORRELATION = cor(AGE, CIRCUMFERENCE))

result$TREE <- as.numeric(as.character(result$TREE))
result[order(result$TREE), ]
```

```
## # A tibble: 5 x 3
##    TREE COVARIANCE CORRELATION
##   <dbl>      <dbl>       <dbl>
## 1     1   22340.07   0.9854675
## 2     2   34290.45   0.9873624
## 3     3   22239.83   0.9881766
## 4     4   37062.62   0.9844610
## 5     5   30442.81   0.9877376
```

Answer to Question5

5a)

```
median(desert$HousePrice)
```

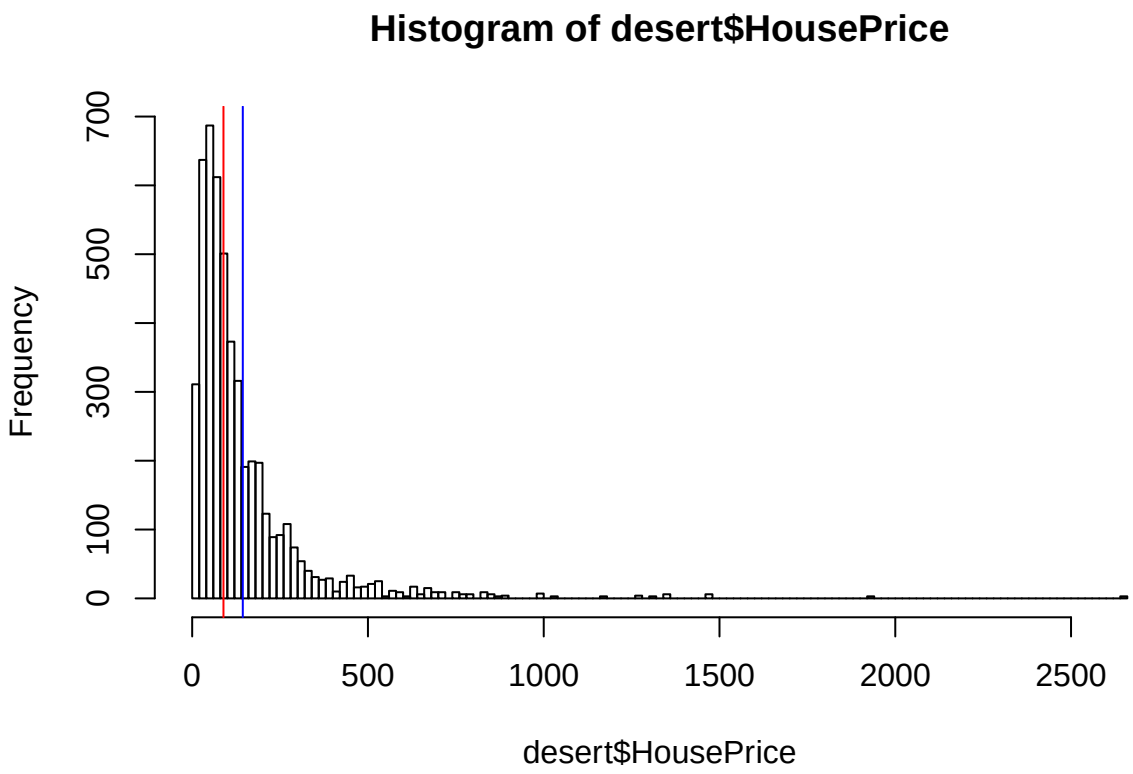
```
## [1] 89
```

```
mean(desert$HousePrice)
```

```
## [1] 144.0348
```

5b)

```
hist(desert$HousePrice, breaks = 100)
abline(v = median(desert$HousePrice), col = "red")
abline(v = mean(desert$HousePrice), col = "blue")
```



I plotted the frequecy of the hourse price. The distribution is skewed to the right if there is a long tail to the right. That is if the mean is greater than the median, the distribution is skewed to the right. A few high numbers will pull the mean above the median.

5C)

```
add10desert <- desert + 10
median(add10desert$HousePrice)
```

```
## [1] 99
```

5d)

```
mult2desert <- desert * 2
median(mult2desert$HousePrice)
```

```
## [1] 178
```