STAT 341: Assigment 2, Question 1

Jainish Mehta, 20773809
Due: October 9, 2020

Part 1a **(i)**

```r
par(oma=c(0,0,2,0),mfrow=c(1,2))
sh <- suppressPackageStartupMessages
sh(require(ggplot2))
     df1 <- read.csv("C:/Users/jaini/one500.csv",header=TRUE)
     df2<- read.csv("C:/Users/jaini/two500.csv",header=TRUE)

#Populate empty vector
vector = c()
   pixel<-rowSums(df1)
   vector <- c(vector, pixel)

 boxplot(vector,main="Digit 1 Population Boxplot", ylab="Frequency")
 #Create 25 bins
hist(vector,breaks=25,
     main="Digit 1 Population Histogram",col ="grey", xlab="Sum of Brightness Values")
```

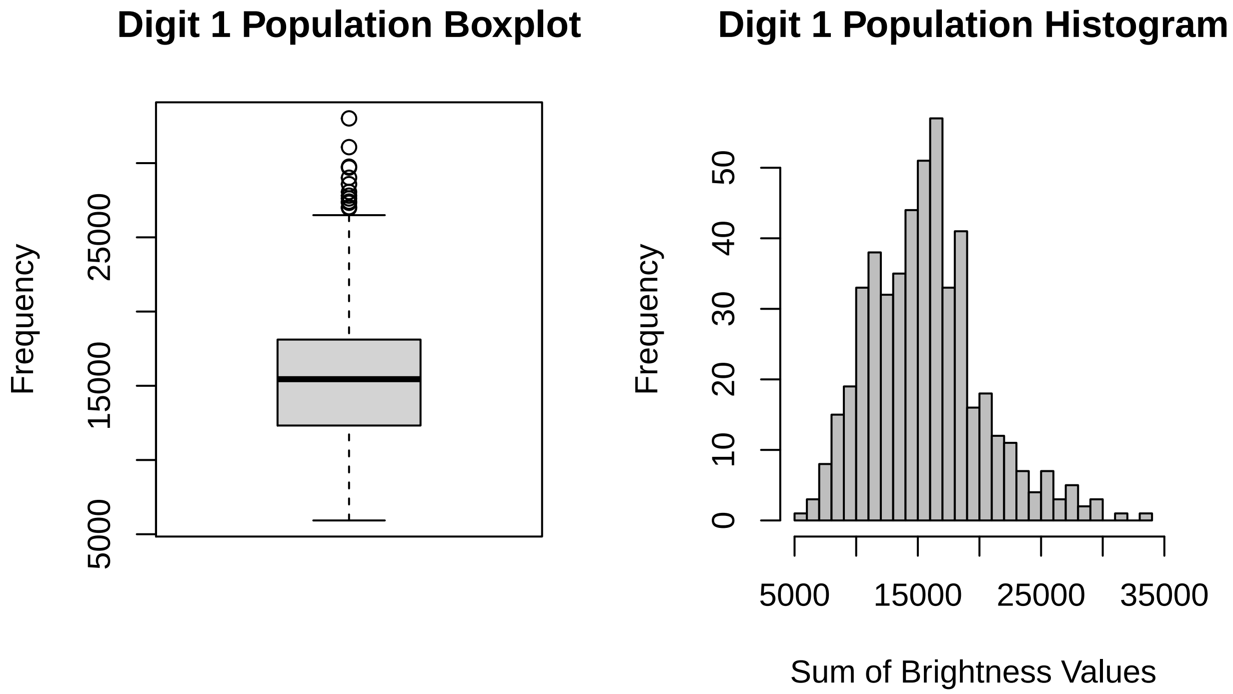

Figure 1

```r
par(oma=c(0,0,2,0),mfrow=c(1,2))
#Populate empty vector
vector2 = c()
  pixel<-rowSums(df2)
 vector2 <- c(vector2, pixel)
 boxplot(vector2, main="Digit 2 Population Boxplot",ylab="Frequency")
hist(vector2, breaks=25,
     main="Digit 2 Population Histogram", col="grey",xlab="Sum of Brightness Values")
```

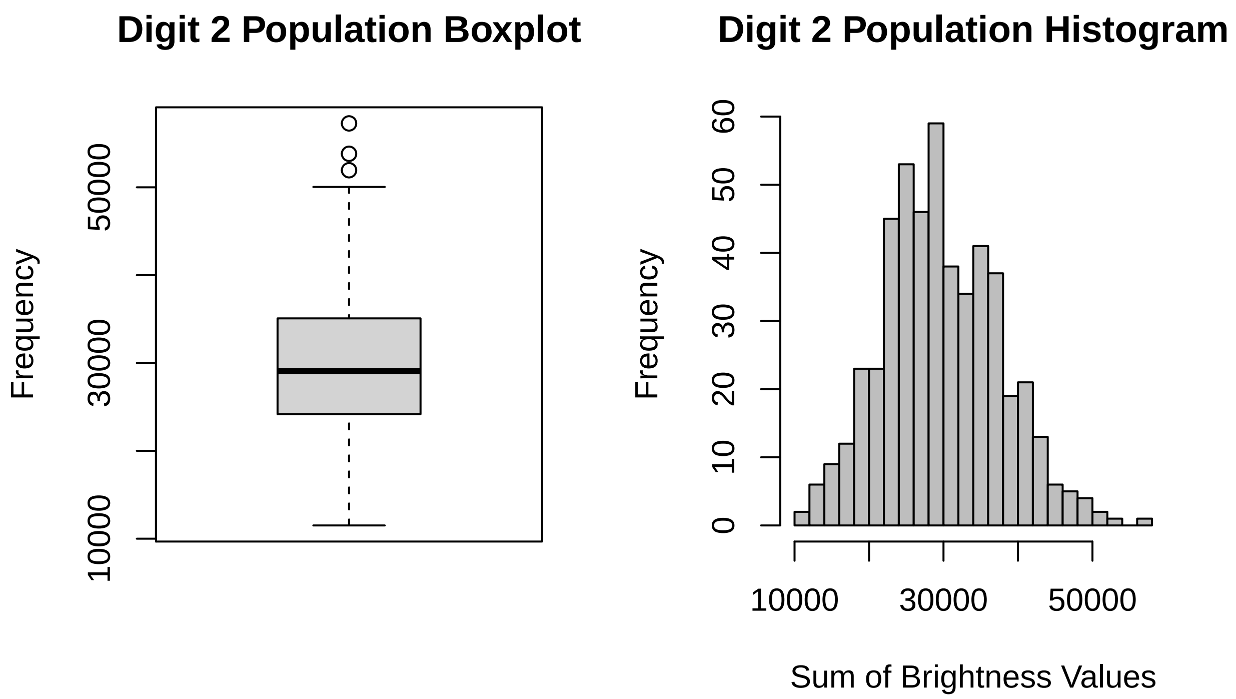

Figure 2

There are many more outliers in the one's than the two's population as shown through the boxplots. The one's boxplot also has a larger range and the median that splits the upper and lower quartile equally. The two's median is much closer to the lower quartile than the upper quartile, hence a positive skew and a larger inter-quartile range. The median for the one's is higher than that of the two's. The one's histogram is much more left-skewed than the two's histogram, which has more spread than the one's histogram. Both histograms have low frequency of values at the edge of the graph.

(ii) Yes, the boxplots are a good representation of the populations. The one's boxplot conveys how the one's population is much more left-skewed with many more outliers. The plot also depicts that there is much less spread than the two's population, hence a smaller inter-quartile range. This suggests there is lower variance within the one's population. The two's boxplots depicts the a larger spread, larger inter-quartile range and larger distribution of values.

(iii)

```r
par(oma=c(0,0,2,0))
vector<-na.omit(vector)
N<-length(vector)
vectorrank <- rank(vector, ties.method = "first")
p<-vectorrank/N
plot(p, vector, pch=19, col =adjustcolor("grey",alpha =0.5),
     main="Quantile Plot for 1s Population",
     xlab="Proportion p",ylab="Sum of Brightness Values, Q(p)")
```

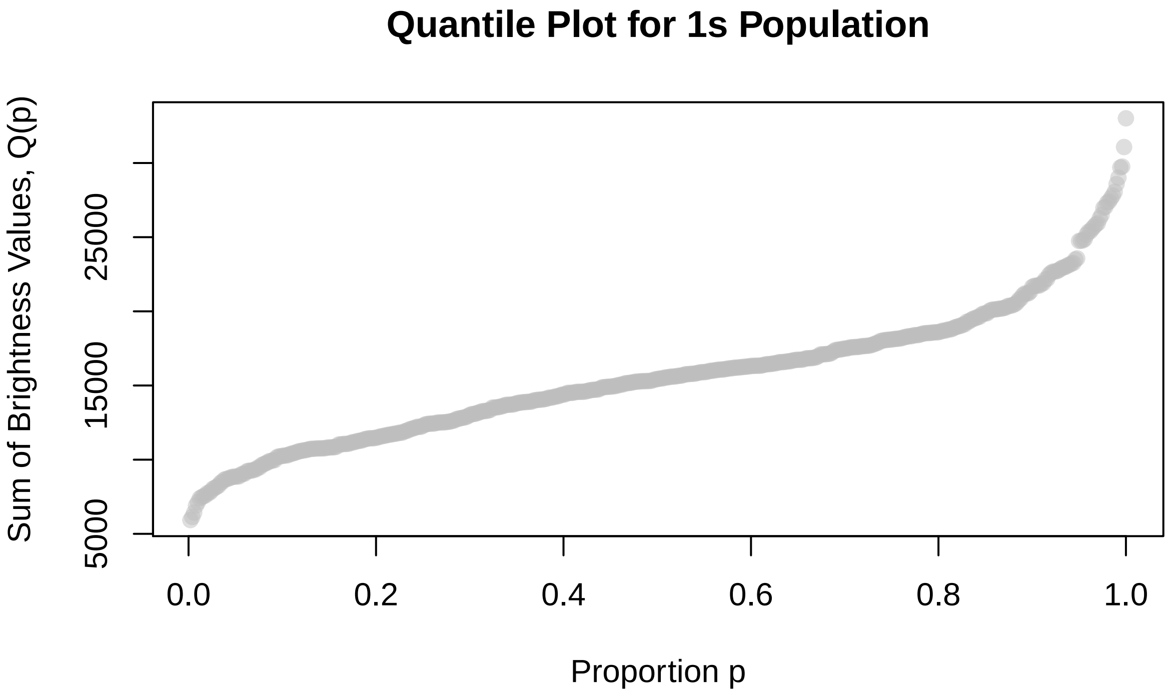

Figure 3

```
vector2<-na.omit(vector2)
vectorrank2 <- rank(vector2, ties.method = "first")
p<-vectorrank2/N
plot(p, vector2, pch=19, col =adjustcolor("grey",alpha =0.5),
     main="Quantile Plot for 2s Population",
     xlab="Proportion p",ylab="Sum of Brightness Values, Q(p)")
```

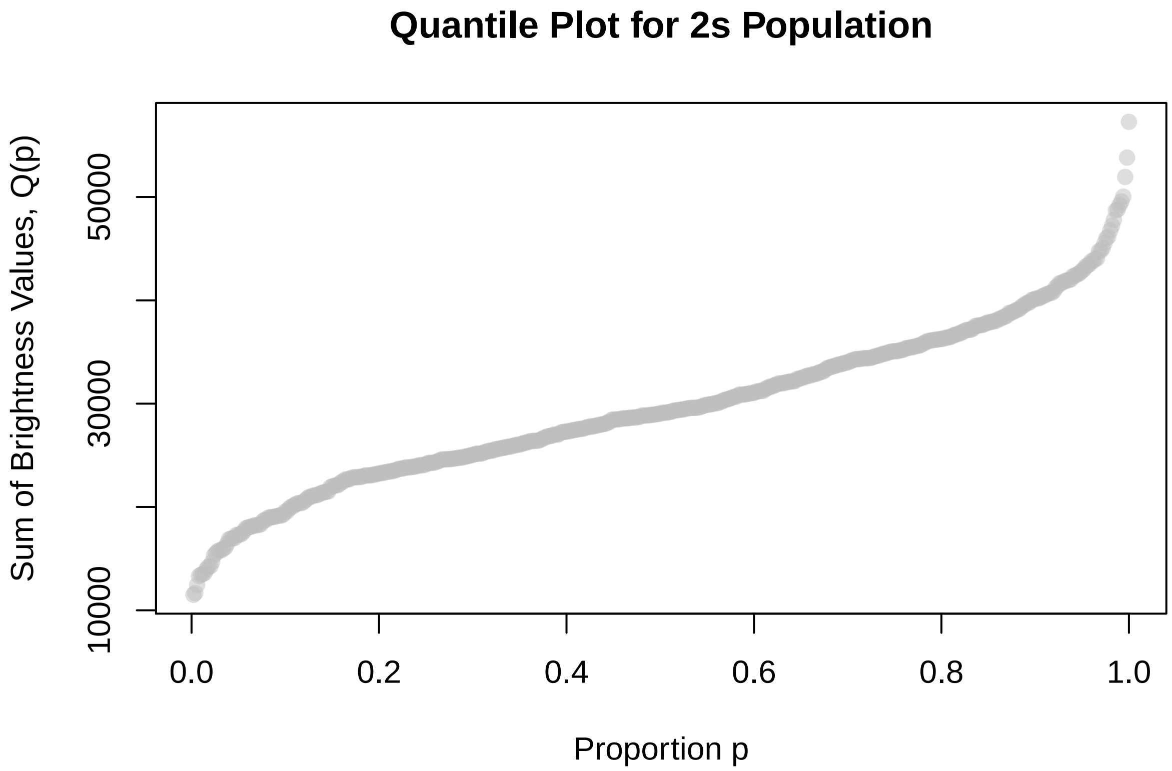

Figure 4

The shape of the quantile plots reveal that both population are almost normally distributed, with the 1's quantile plot show a slight left skewness. Both the one's and two's quantile plots show that there is a steady slope in the center of the plot. This indicates there is roughly similar frequencies and shape of the plot at the center bins. This is further shown by the darker, dense concentration of variates with similar y-values. Near the right edge of both quantile plots, more so in the 1s population, there are rapidly rising spots. In this, the y values are not similar even though the ranks are similar. The two's population depicts densely concentrated values in the center with fewer outliers, hence fewer rapidly rising spots at the edges of the quantile plot.

***Part 1b* (i)**

```
par(oma=c(0,0,2,0),mfrow=c(1,2))
#Populate empty vector
vector3 = c()
for (i in 1:500){
  pixel<-c((df1[[i,403]]))
 vector3 <- c(vector3, pixel)
}
 min=min(vector3)
 max = max(vector3)
 boxplot(vector3,main="Digit 1, Pixel 403 Population",ylab="Frequency")
 #25 Bins
hist(vector3,breaks=50,
     main="Digit 1, Pixel 403 Population", col="grey",
     xlab="Brightness Values")
```

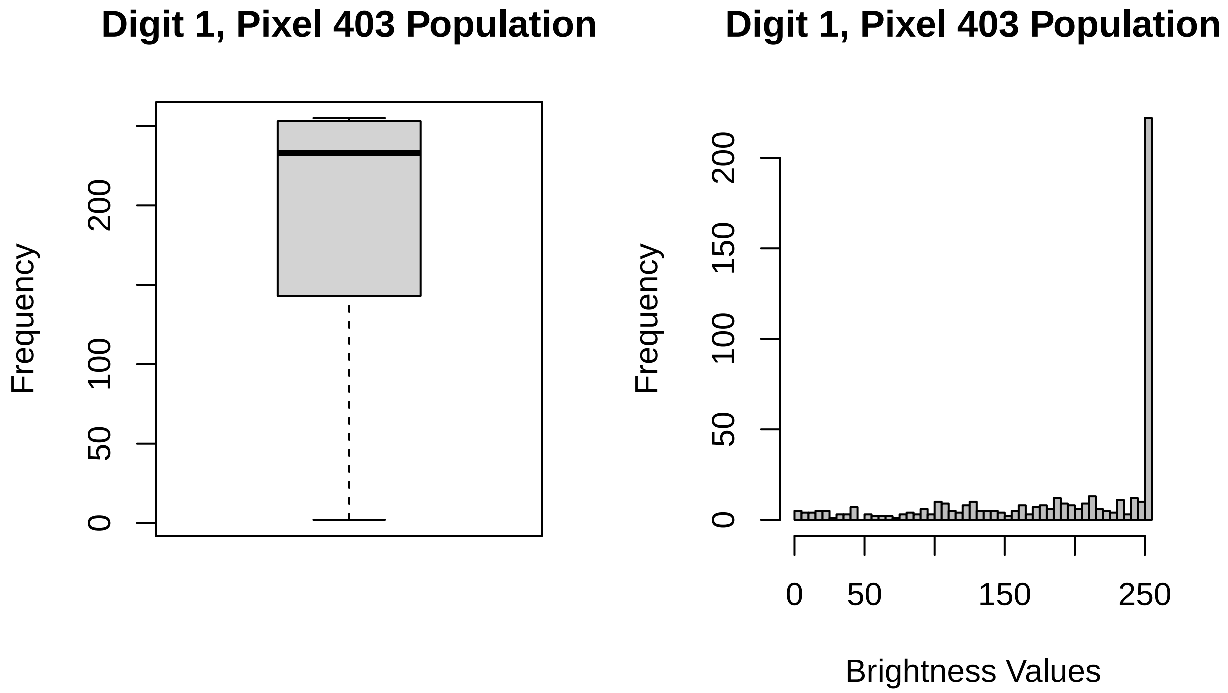

Figure 5

```
par(oma=c(0,0,2,0),mfrow=c(1,2))
vector4 = c()
for (i in 1:500){
  pixel<-c((df2[[i,403]]))
  vector4 <- c(vector4, pixel)
}
 min=min(vector4)
 max = max(vector4)
boxplot(vector4,main="Digit 2, Pixel 403 Population",ylab="Frequency")
#25 Bins
hist(vector4,breaks=seq(min,max,length.out=50),
     main="Digit 2, Pixel 403 Population", col="grey",
     xlab="Brightness Values")
```

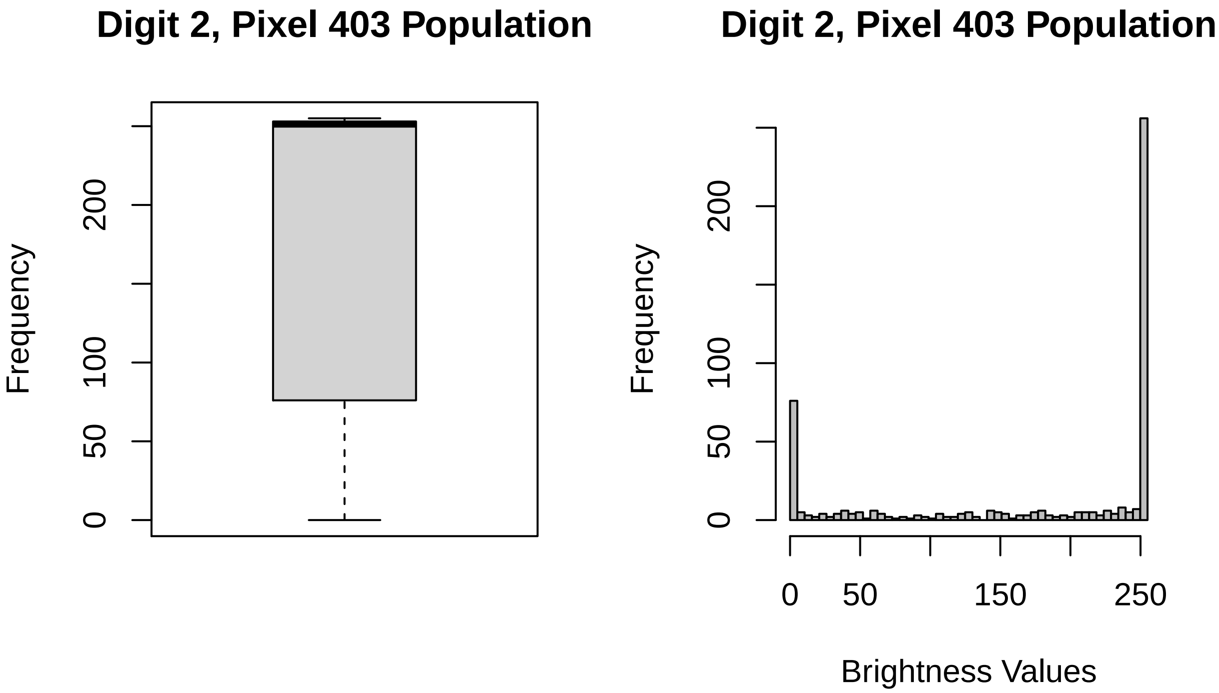

Figure 6

The 1's and 2's boxplot both convey how the population is left-skewed. The 2's population is shown to be much more left-skewed, however. The inter-quartile range of the 2's population is bigger than that of the 1's population, thus more variability in the frequency of values. The distance between the minimum and the first quartile, the lower whisker, is much larger in the 1's population than the 2's. The histograms show a similar aspect as it is negatively skewed with the 2s population having a relatively higher frequency at the lowest bin and low variance in the middle values. The 1's population is also negatively skewed with the highest frequency at the last bin. The fact that the highest frequencies are at the lowest and highest bins is because intuitively, the middle columns are usually fully darkened or not at all.

(ii) The boxplot is not as good a representation as in part (a) as it does not convey as many features of the histogram. Both boxplots look similar, with similar lower whiskers, medians, and inter-quartile ranges, even though the histograms are slightly different. The boxplot does not convey how there is a large frequency in the lowest bin in the 2's population, yet similar variance in brightness values in the middle bins.

(iii)

```r
par(oma=c(0,0,2,0))
vector3<-na.omit(vector3)
N<-length(vector3)
vectorrank3 <- rank(vector3, ties.method = "first")
p<-vectorrank3/N
plot(p, vector3, pch=19, main="Digit 1 Quantile Plot",
     col =adjustcolor("grey",alpha =0.5),xlab="Proportion p",
     ylab="Brightness Value, Q(p)")
```

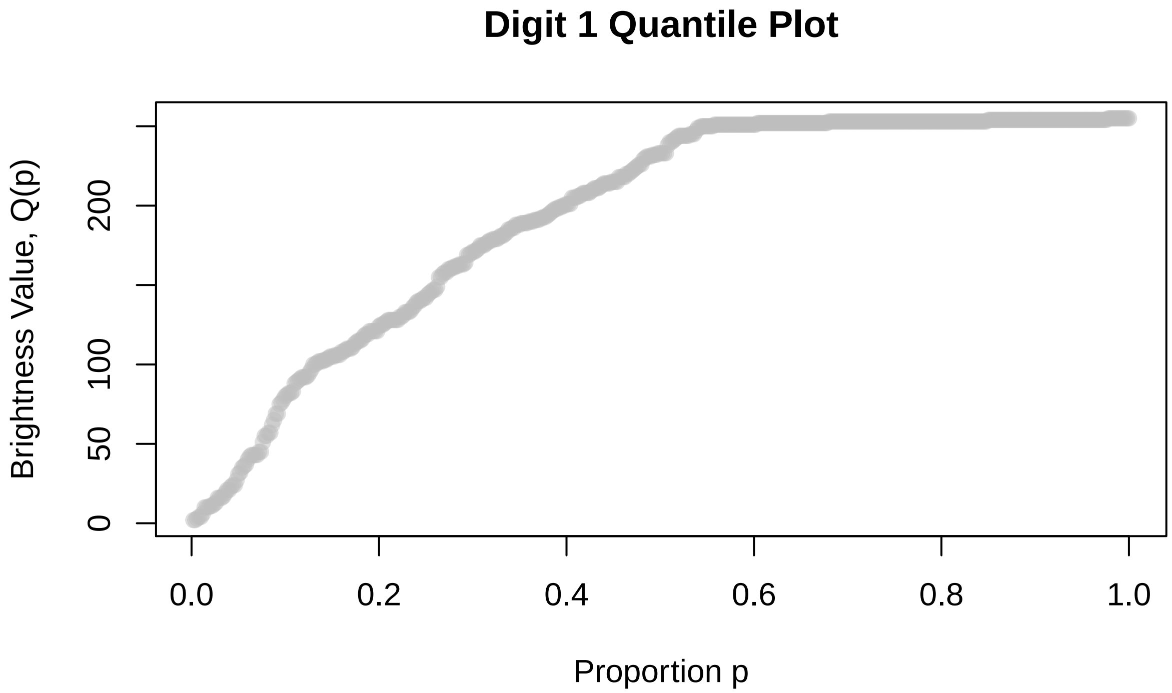

Figure 7

```
vector4<-na.omit(vector4)
N<-length(vector4)
vectorrank4 <- rank(vector4, ties.method = "first")
p<-vectorrank4/N
plot(p, vector4, pch=19, main="Digit 2 Quantile Plot",
     col =adjustcolor("grey",alpha =0.5),
     xlab="Proportion p", ylab="Brightness Value, Q(p)")
```

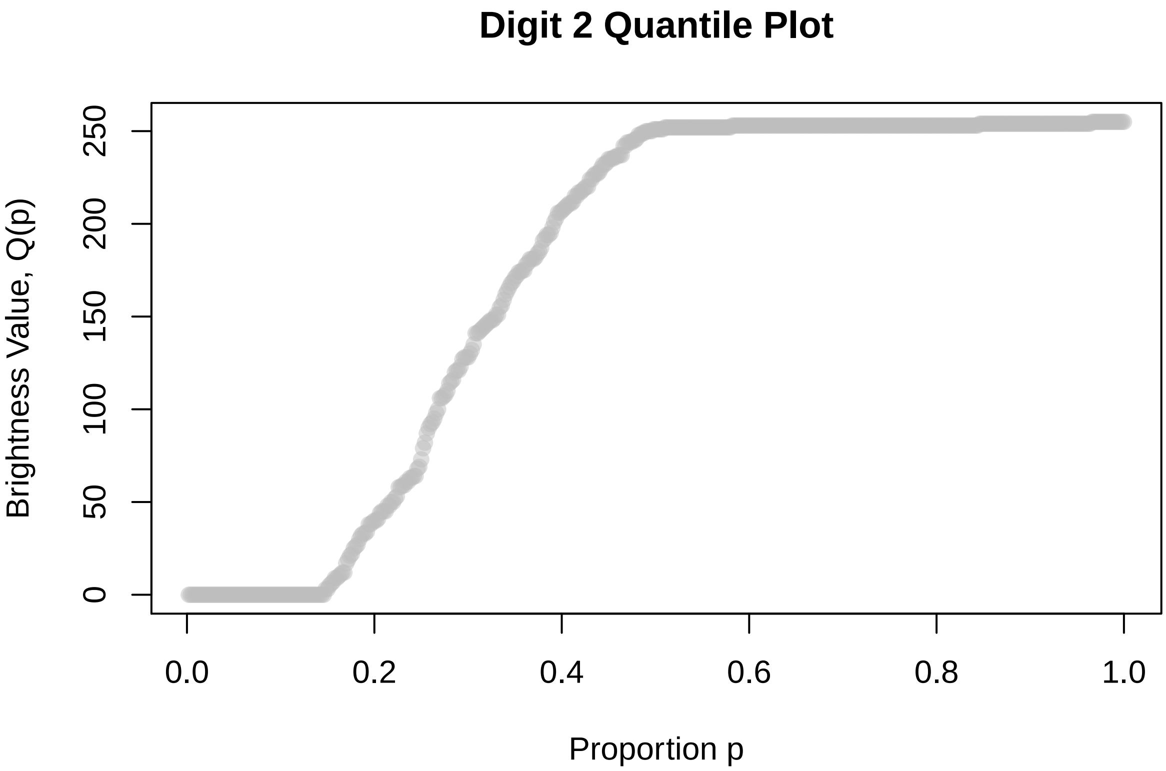

Figure 8

The 1's population quantile plot depicts a steady increase in y values as the order of the variate increases, thus conveying the low variance and low frequency of the lower and center bins. The long plateau at the end depicts the dense concentration of values at the edge of the histogram, hence beginning at around 0.55. The 2's population quantile plot illustrates the dense concentration of similar y values in the lowest and highest bins and low variance and smaller frequencies in the middle bins, hence being lighter in density. This is shown by the steeper, linear slope. There is a longer plateau with a denser concentration area in the 2's population than in the 1's population, hence likely a larger frequency for the last bin of the 2's population. Overall, both population are shown to be left-skewed.

Part 1c
The implicit assumption for a boxplot to be a "good" representation of a population is that it is normally distributed and unimodal.This is shown when comparing part *(a)* and *(b)* in which the boxplot showed much more of the characteristics of the population and the histogram, such as skewness, outliers, distribution, and more. This is not emphasized as much in part *(b)* since it very skewed and hence the other features are not as obvious or prominent through the boxplot.

Part 1d **(i)**

```
vector5 = c()
for (i in 1:500){
  pixel<-c((df1[[i,406]]))
 vector5 <- c(vector5,pixel)
}

print(IQR(vector5))
```

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

For the Freedman-Diaconis' rule for the number of bins, the equation is as follows:

$$\begin{aligned}\frac{Items}{Bins} &= 2\frac{IQR(x)}{N^{\frac{1}{3}}} \\ &= 2\frac{1}{500^{\frac{1}{3}}} \approx 0.252 \\ Bins &= \frac{Items}{\frac{Items}{Bins}} \\ &= \frac{255}{0.252} \approx 1012\end{aligned}$$

The Scott's rule for the number of bins is

$$\frac{Items}{Bins} = 3.5\frac{\sigma}{N^{\frac{1}{3}}}$$

where

$$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$$

Using code, we can determine $\Sigma(x_i-\mu)^2$.

```
total=0
for (i in 1:length(vector5)){
  total=total+((vector5[i]-mean(vector5))^2)
}
print (mean(vector5))
```

```
## [1] 247.064
```

```
print (total)
```

```
## [1] 364252
```

Hence, $\mu = 247.064$ and $\Sigma(x_i-\mu)^2 = 364252$. Therefore, $\sigma = \sqrt{\frac{364252}{N}} = \sqrt{\frac{364252}{500}} \approx 27$. To solve the equation,

$$\begin{aligned}Size = 3.5*\frac{27}{500^{\frac{1}{3}}} &\approx 3.5*3.4 \approx 11.9 \\ Bins &= \frac{Items}{\frac{Items}{Bins}} \\ Bins &= \frac{255}{11.9} \approx 21\end{aligned}$$

```
str(hist(vector5,breaks="FD",plot=FALSE))
```

```
## List of 6
##  $ breaks  : num [1:1276] 0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 ...
##  $ counts  : int [1:1275] 1 0 0 0 0 0 0 0 0 0 ...
##  $ density : num [1:1275] 0.01 0 0 0 0 0 0 0 0 0 ...
##  $ mids    : num [1:1275] 0.1 0.3 0.5 0.7 0.9 1.1 1.3 1.5 1.7 1.9 ...
##  $ xname   : chr "vector5"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"
```

Hence there are 1275 bins using Freedman-Diaconis' rule for the number of bins.

```r
y<-hist(vector5,breaks="Scott", plot=FALSE)
str(y)
```

```
## List of 6
##  $ breaks  : num [1:27] 0 10 20 30 40 50 60 70 80 90 ...
##  $ counts  : int [1:26] 1 0 1 2 0 0 0 0 0 0 ...
##  $ density : num [1:26] 2e-04 0e+00 2e-04 4e-04 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 ...
##  $ mids    : num [1:26] 5 15 25 35 45 55 65 75 85 95 ...
##  $ xname   : chr "vector5"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"
```

Hence, there are 26 bins using Scott's rule for the number of bins.

Table 1: Number of Bins Calculated and Generated using Hist()

Calculated Scott Bins	Calculated Freedman–Diaconis Bins	Programmed Scott Bins	Programmed Freedman–Diaconis Bins
21	1012	26	1275

(ii)

```r
par(oma=c(0,0,0,0),mfrow=c(2,2))
hist(vector5,breaks="FD",xlab="Brightness Value",main="Freedman-Diaconi Bins")
hist(vector5,breaks="scott",xlab="Brightness Value",main="Scott Bins")
hist(vector5, breaks=21,main=" 21 Bins", col="grey",xlab="Brightness Value")
hist(vector5, breaks=1012,main="1012 Bins", col="grey",xlab="Brightness Value")
```

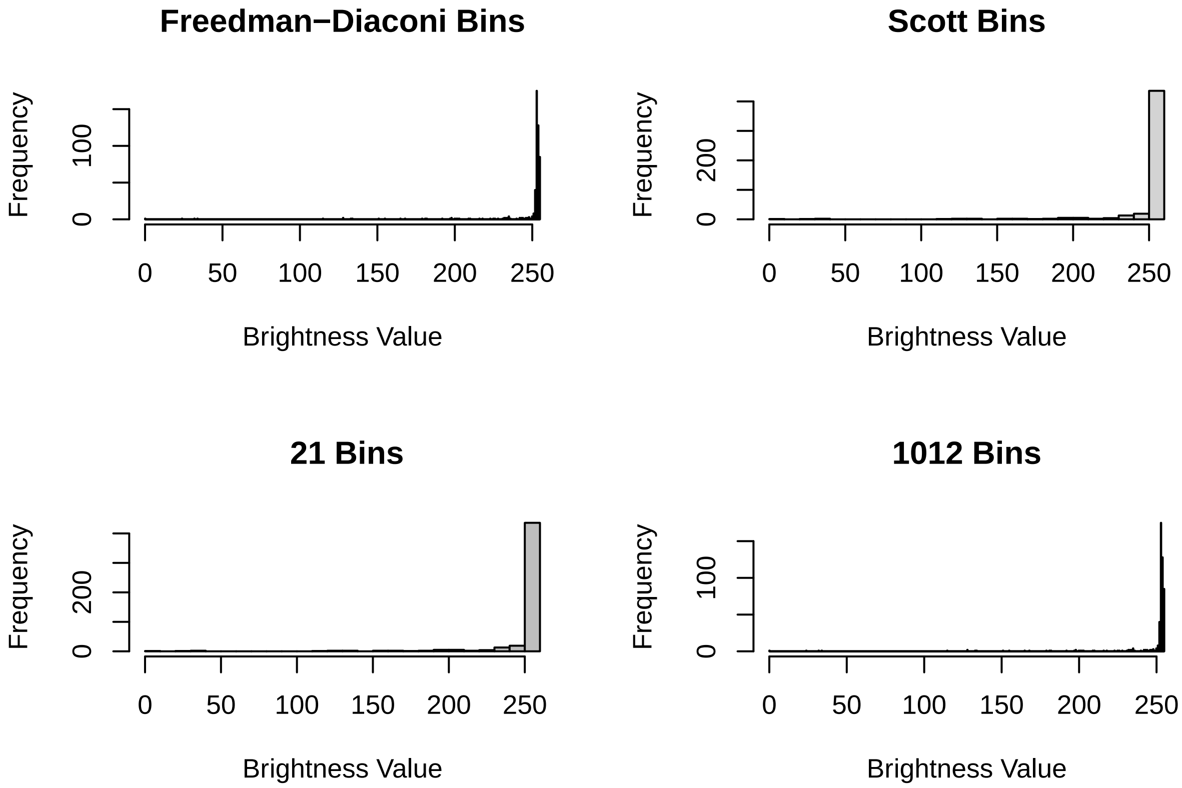

Figure 9

(iii)

```
hist(vector5,xlab="Brightness Values",main="Sturges' Number of Bins")
```

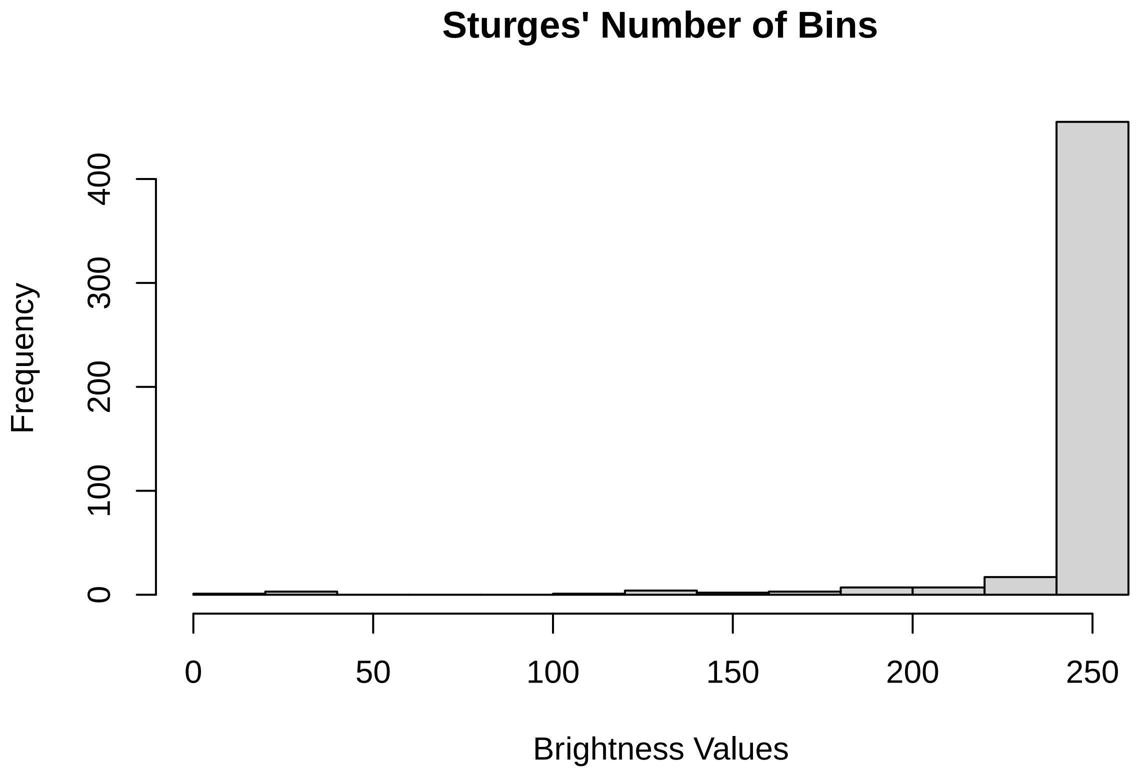

Figure 10

```
vector5<-na.omit(vector5)
N<-length(vector5)
vectorrank5 <- rank(vector5, ties.method = "first")
p<-vectorrank5/N
plot(p, vector5, pch=19, col =adjustcolor("grey",alpha =0.5),
     main="Digit 1, Pixel 406 Quantile Plot",
     xlab="Proportion p",ylab="Brightness Value, Q(p)")
```

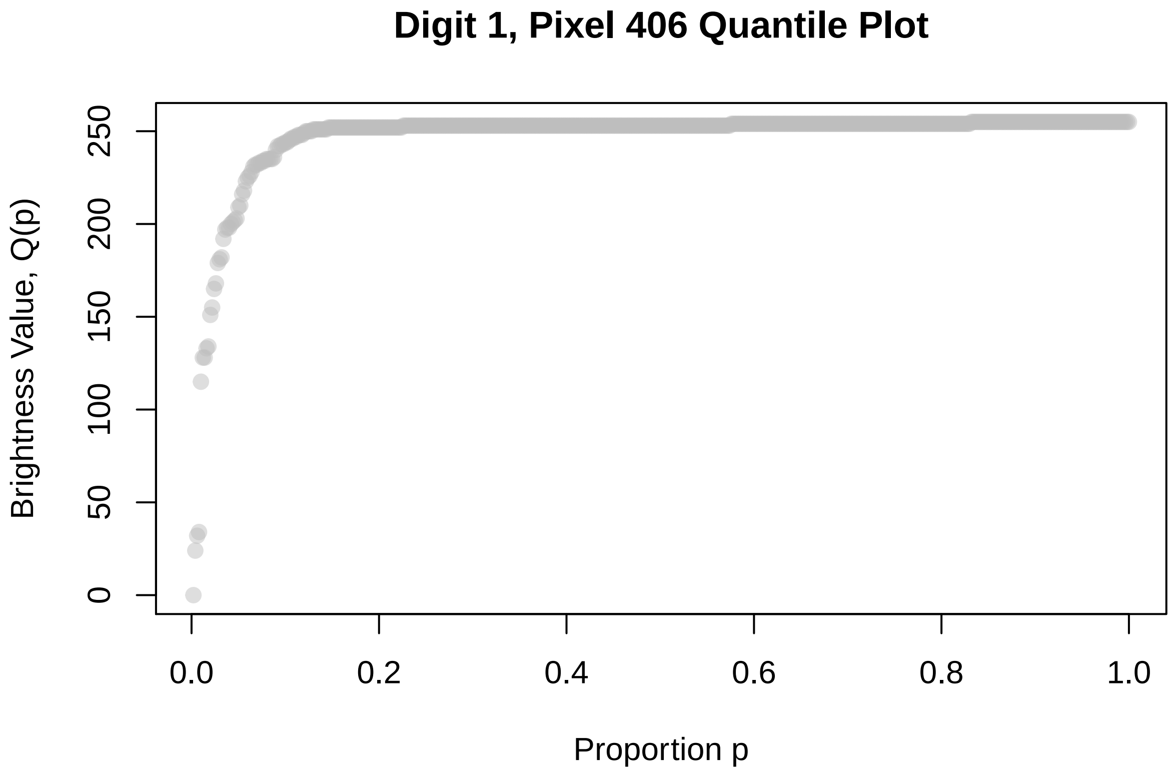

Figure 11

The quantile plot conveys that the first points, though near each other in rank, have significantly different y-values and can thus be considered rapidly rising spots. This goes alongside the histogram that reinforces that the values to the left of the plateau have low frequencies. To the right of the spots, the spots rapidly becomes a flat plateau, hence a dense concentration of points with similar y-values starting at 0.2. This indicates that the values have a very negative and left-skewed distribution.

Part 1e

```
#Populate empty vector
a<- as.matrix(df1)
b <- as.matrix(df2)
a_mean <- colMeans(df1)
b_mean <- colMeans(df2)
c <- abs(a_mean-b_mean)

 tail(sort(c),5)
```

```
##    V436    V437    V409    V407    V408
## 136.288 142.872 147.698 163.688 167.540
```

From the code block above, columns 436, 437, 409, 407, and 408, in ascending order, have the largest difference in averages between the group of one's and two's.