# A2Q1

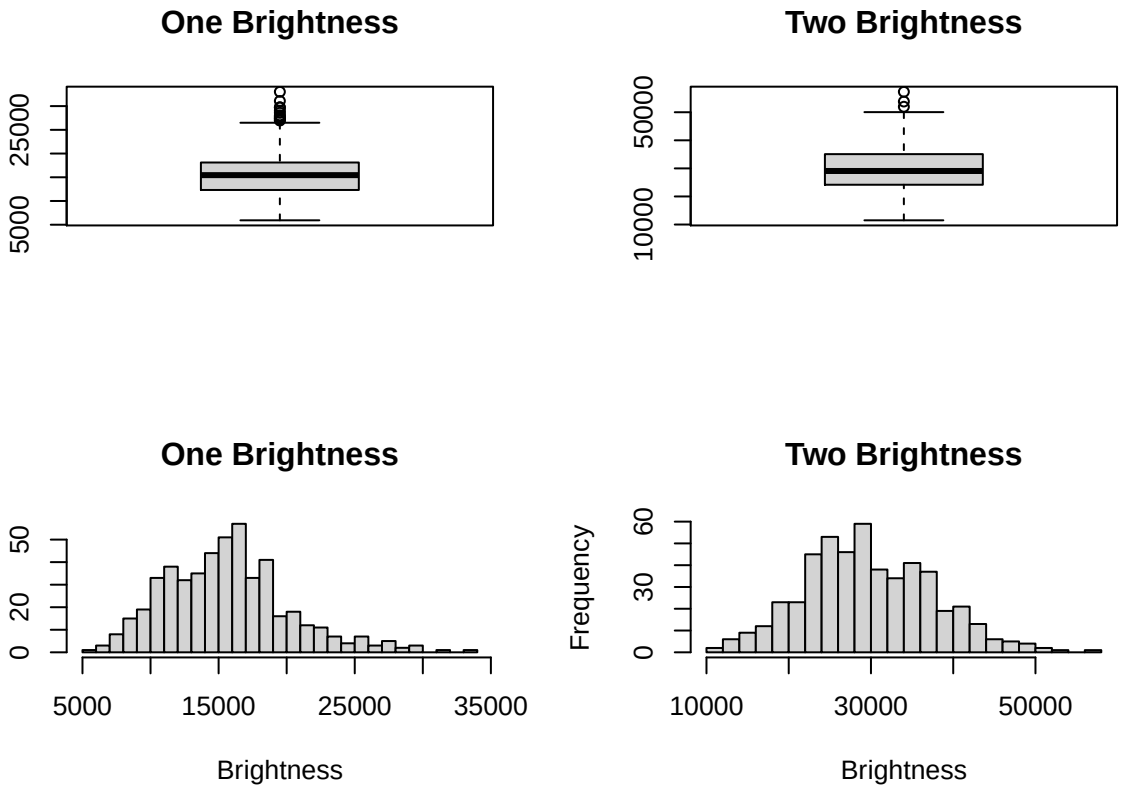
```
one <- read.csv("one500.csv",header=TRUE)
two <- read.csv("two500.csv",header=TRUE)
```

## a)

### i)

```
ink1 <- rowSums(one)
ink2 <- rowSums(two)
```

```
par(mfrow=c(2,2))
boxplot(ink1, main ="One Brightness")
boxplot(ink2, main ="Two Brightness")
hist(ink1, breaks = 25, main="One Brightness", xlab = "Brightness")
hist(ink2, breaks = 25, main="Two Brightness", xlab = "Brightness")
```



Comment:

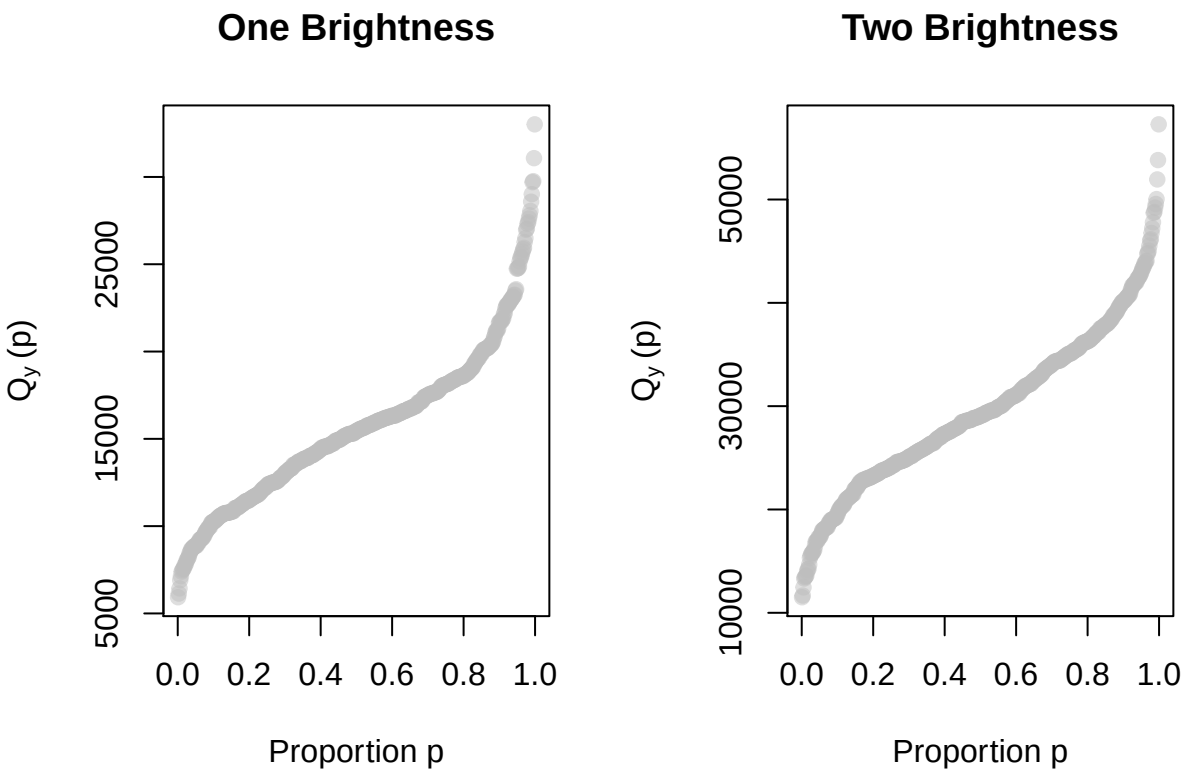
From the boxplot, we can see that there are more outliers in digit 1 than digit 2, which means that the distribution of digit 1 might be a little bit right skewed. The range of digit 2 is a little bigger than digit 1. Digit 2 has more dispersion. However, we see that medians are nearly in the middle of IQR in both boxes. Therefore, we can say that both distributions are roughly symmetric. For the histogram, we see that graph for digit 1 has a long right tail, which matches our conclusion in the boxplot. From the structure of both graphs, we can see that despite there are some outliers, most of the data group around the center. Therefore, we conclude that both distributions are roughly symmetric, but digit one has a longer right tail.

## ii)

Yes, both boxplots clearly summarize the structure of the distribution. They correctly show that there exists only one mode in both distribution and some outliers are located at the right.

## iii)

```
par(mfrow = c(1,2))
q1 <- sort(ink1)
q2 <- sort(ink2)
pvals1 <- ppoints(length(q1))
pvals2 <- ppoints(length(q2))
plot(pvals1,q1, pch = 19, col = adjustcolor("grey", alpha = 0.5),
     xlim = c(0, 1), xlab = "Proportion p", ylab = bquote("Q"["y"] ~ "(p)"),
     main = "One Brightness")
plot(pvals2,q2, pch = 19, col = adjustcolor("grey", alpha = 0.5),
     xlim = c(0, 1), xlab = "Proportion p", ylab = bquote("Q"["y"] ~ "(p)"),
     main = "Two Brightness")
```



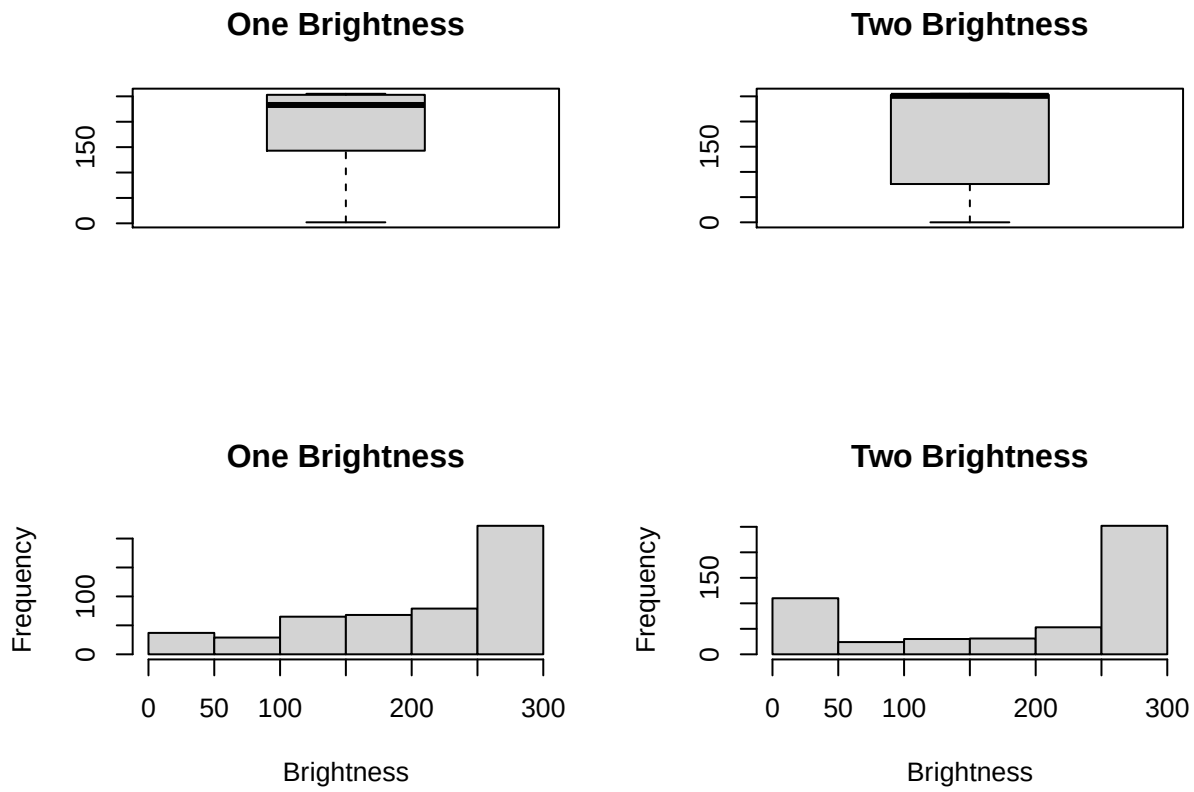

Comment:

For digit 1, we can see that the curve is generally smooth, but there is a break and a heavy tail at the top. For digit 2, we can see that the curve is smooth with a light tail at the top.

## b)

### i)

```
par(mfrow=c(2,2))
one403 <- one[, 403]
two403 <- two[, 403]
boxplot(one403, main ="One Brightness")
boxplot(two403, main ="Two Brightness")
hist(one403, breaks = 5, main="One Brightness", xlab = "Brightness")
hist(two403, breaks = 5, main="Two Brightness", xlab = "Brightness")
```




Comment:

For boxplot, we see that there are nearly no outliers for both plots. Both medians are close to upper quartile and minimum values are away from the box. This means that both distributions are asymmetric and left skewed. Since median of digit 2 is closer to upper quartile, the distribution is more skewed. They have the same range, but digit 2 has a larger IQR. For histogram, we can see that both graphs have long left tail, which means that they are left skewed. there are two modes in the distribution. One is near zero, and another is near 300. Digit 2 has more data at the left than digit 1.

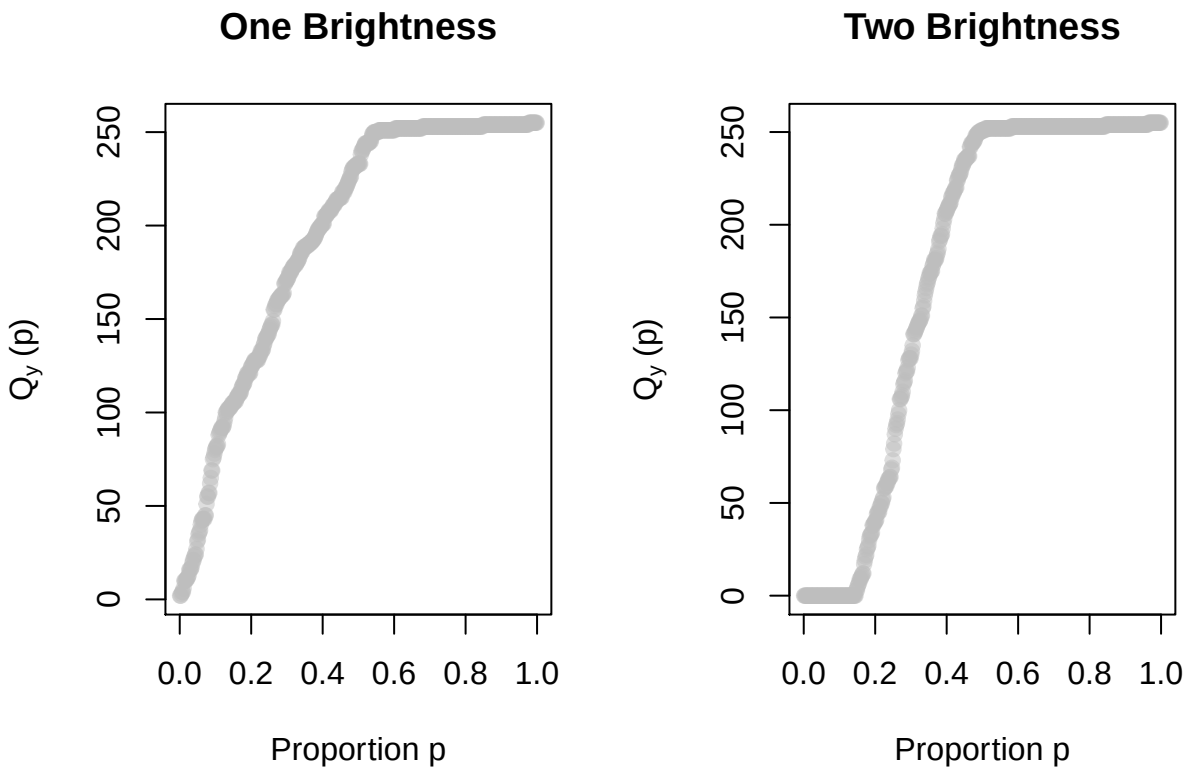
### ii)

No, both boxplots summarize the basic characteristics of the distributions. However, there exists two modes in the graph, but they implicitly assume the data is unimodel. The histogram gives more structure and has a better representation.

### iii)

```
par(mfrow = c(1,2))
q1 <- sort(one403)
q2 <- sort(two403)
pvals1 <- ppoints(length(q1))
pvals2 <- ppoints(length(q2))
plot(pvals1,q1, pch = 19, col = adjustcolor("grey", alpha = 0.5),
     xlim = c(0, 1), xlab = "Proportion p", ylab = bquote("Q"["y"] ~ "(p)"),
```

```
     main = "One Brightness")
plot(pvals2,q2, pch = 19, col = adjustcolor("grey", alpha = 0.5),
     xlim = c(0, 1), xlab = "Proportion p", ylab = bquote("Q"["y"] ~ "(p)"),
     main = "Two Brightness")
```




Comment:

For digit 1, the quartile looks like a straight line except for at the top. For digit 2, the quartile looks like a straight line except for at both ends. Both graphs show that there exists significant amount of data at the end.

## c)

Boxplot implicitly assumes the data is unimodel. So for those data which is uninmodel, boxplot will be a good representation of the population.

## d)

### i)

Freedman-Diaconis Rule:

$$\text{Bin Size} = 2\frac{IQR(x)}{N^{\frac{1}{3}}}$$

```
one406 <- one[, 406]
width <- 2*IQR(one406)/length(one406)^(1/3)
ceiling((max(one406)-min(one406))/width)
```

```
## [1] 1012
```

```
length(hist(one406, breaks = "FD", plot = FALSE)$breaks)-1
```

```
## [1] 1275
```

Scott's Rule:

$$\text{Bin Size} = 3.5\frac{\sigma}{N^{\frac{1}{3}}}$$

```
width <- 3.5*sd(one406)/length(one406)^(1/3)
ceiling((max(one406)-min(one406))/width)
```

```
## [1] 22
```

```
length(hist(one406, breaks = "Scott", plot = FALSE)$breaks)-1
```

```
## [1] 26
```

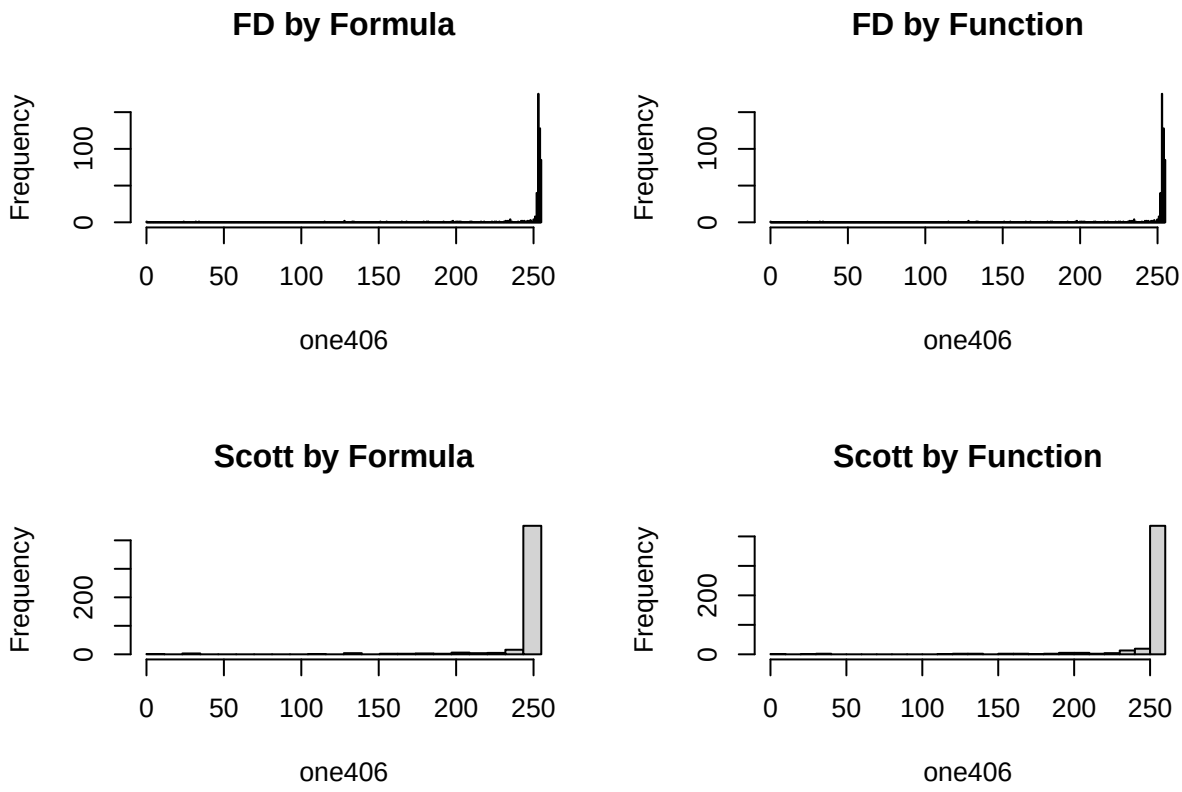
Comment:

By comparing both formula breaks and function breaks, we conclude that hist function creates more breaks than formula in this distribution.

| | By Formula | Hist Function |
|---|---|---|
| FD | 1012 | 1275 |
| Scott | 22 | 26 |

## ii)

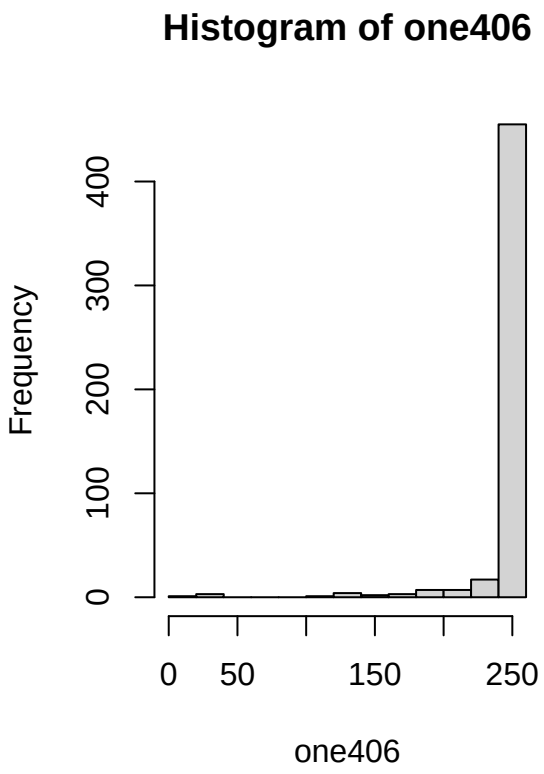
```
par(mfrow= c(2, 2))
hist(one406, breaks = seq(min(one406), max(one406), length.out = 1013),
     main = "FD by Formula")
hist(one406, breaks = "FD", main = "FD by Function")
hist(one406, breaks = seq(min(one406), max(one406), length.out = 23),
     main = "Scott by Formula")
hist(one406, breaks = "Scott", main = "Scott by Function")
```

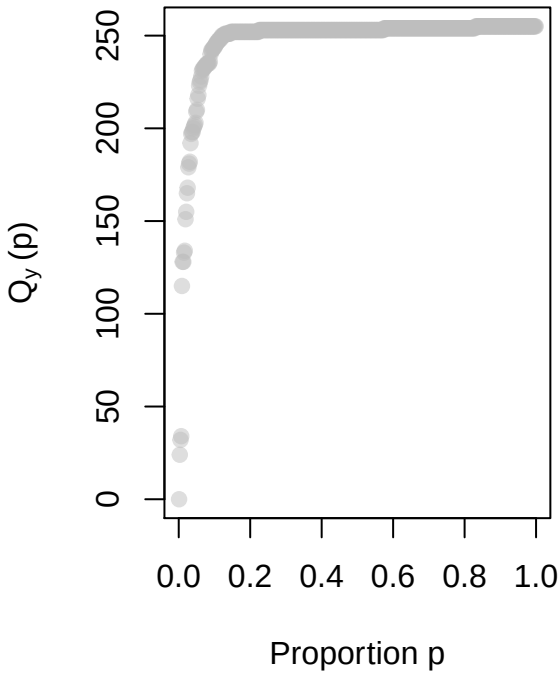



iii)

```
par(mfrow= c(1,2))
hist(one406, breaks = c(0, 20, 40, 60, 80, 100, 120, 140, 160, 180,
                        200, 220, 240, 260))

qvals <- sort(one406)
pvals <- ppoints(length(one406))
plot(pvals,qvals, pch = 19, col = adjustcolor("grey", alpha = 0.5),
     xlim = c(0, 1), xlab = "Proportion p", ylab = bquote("Q"["y"] ~ "(p)"), )
```




Comment:

There are many breaks throughout the quartile. Most of the data are located at the top of the graph.

## e)

```
difference <- apply(one, 2, mean) - apply(two, 2, mean)
sort(abs(difference), decreasing = TRUE)[1: 5]
```

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

Top 5 Pixels are:

408, 407, 409, 437, 436.