Lab 7

Nur Izzati Binti Shalahudin

4/29/2022

1. Load packages from library

```
library(mlbench)
library(caret)
data("BreastCancer")
```

2.Display the dimension of the "BreastCancer" dataset.

```
dim(BreastCancer)
```

```
## [1] 699  11
```

3.Displays the first 20 recods of the "BreastCancer" dataset

```
head(BreastCancer,20)
```

```
##         Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 1  1000025            5         1          1             1            2
## 2  1002945            5         4          4             5            7
## 3  1015425            3         1          1             1            2
## 4  1016277            6         8          8             1            3
## 5  1017023            4         1          1             3            2
## 6  1017122            8        10         10             8            7
## 7  1018099            1         1          1             1            2
## 8  1018561            2         1          2             1            2
## 9  1033078            2         1          1             1            2
## 10 1033078            4         2          1             1            2
## 11 1035283            1         1          1             1            1
## 12 1036172            2         1          1             1            2
## 13 1041801            5         3          3             3            2
## 14 1043999            1         1          1             1            2
## 15 1044572            8         7          5            10            7
## 16 1047630            7         4          6             4            6
```

```
## 17 1048672            4         1          1            1           2
## 18 1049815            4         1          1            1           2
## 19 1050670           10         7          7            6           4
## 20 1050718            6         1          1            1           2
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses     Class
## 1            1           3               1       1    benign
## 2           10           3               2       1    benign
## 3            2           3               1       1    benign
## 4            4           3               7       1    benign
## 5            1           3               1       1    benign
## 6           10           9               7       1 malignant
## 7           10           3               1       1    benign
## 8            1           3               1       1    benign
## 9            1           1               1       5    benign
## 10           1           2               1       1    benign
## 11           1           3               1       1    benign
## 12           1           2               1       1    benign
## 13           3           4               4       1 malignant
## 14           3           3               1       1    benign
## 15           9           5               5       4 malignant
## 16           1           4               3       1 malignant
## 17           1           2               1       1    benign
## 18           1           3               1       1    benign
## 19          10           4               1       2 malignant
## 20           1           3               1       1    benign
```

4.Displays the data types for the attributes in the "BreastCancer" dataset

```
#change data type of Id from character to integer
BreastCancer$Id<- as.integer(BreastCancer$Id)
sapply(BreastCancer, class)
```

```
## $Id
## [1] "integer"
##
## $Cl.thickness
## [1] "ordered" "factor"
##
## $Cell.size
## [1] "ordered" "factor"
##
## $Cell.shape
## [1] "ordered" "factor"
##
## $Marg.adhesion
## [1] "ordered" "factor"
##
## $Epith.c.size
## [1] "ordered" "factor"
##
```

```
## $Bare.nuclei
## [1] "factor"
##
## $Bl.cromatin
## [1] "factor"
##
## $Normal.nucleoli
## [1] "factor"
##
## $Mitoses
## [1] "factor"
##
## $Class
## [1] "factor"
```

5.Removes the “Id” attribute from the “BreastCancer” dataset

```
NoId_BreastCancer = subset(BreastCancer,select = -(Id))
```

Convert the data types of the remaining attributes to “numeric”

```
#EXTRA: convert class into characters first
breast=NoId_BreastCancer
breast$Class<-as.character(breast$Class)
breast$Class<- replace(breast$Class,breast$Class=='benign', "1")
breast$Class<- replace(breast$Class,breast$Class=='malignant', "2")

#convert the data types of the remaining attributes to "numeric"
indx <- sapply(breast, is.factor)
breast[indx] <- lapply(breast[indx], function(x) as.numeric(as.character(x)))
indx1 <- sapply(breast, is.character)
breast[indx1] <- lapply(breast[indx1], function(x) as.numeric(as.character(x)))


#check if the data type is changed
sapply(breast,class)
```

```
##    Cl.thickness       Cell.size      Cell.shape   Marg.adhesion    Epith.c.size
##       "numeric"       "numeric"       "numeric"       "numeric"       "numeric"
##     Bare.nuclei     Bl.cromatin Normal.nucleoli         Mitoses           Class
##       "numeric"       "numeric"       "numeric"       "numeric"       "numeric"
```

6.Displays the summary information of the data in the “Breast-Cancer”dataset

```
summary(breast)
```

```
##   Cl.thickness      Cell.size        Cell.shape     Marg.adhesion
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.000
##  Mean   : 4.418   Mean   : 3.134   Mean   : 3.207   Mean   : 2.807
##  3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##   Epith.c.size     Bare.nuclei      Bl.cromatin     Normal.nucleoli
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000
##  Median : 2.000   Median : 1.000   Median : 3.000   Median : 1.000
##  Mean   : 3.216   Mean   : 3.545   Mean   : 3.438   Mean   : 2.867
##  3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##                   NA's   :16
##     Mitoses          Class
##  Min.   : 1.000   Min.   :1.000
##  1st Qu.: 1.000   1st Qu.:1.000
##  Median : 1.000   Median :1.000
##  Mean   : 1.589   Mean   :1.345
##  3rd Qu.: 1.000   3rd Qu.:2.000
##  Max.   :10.000   Max.   :2.000
##
```

Comment on the summary information

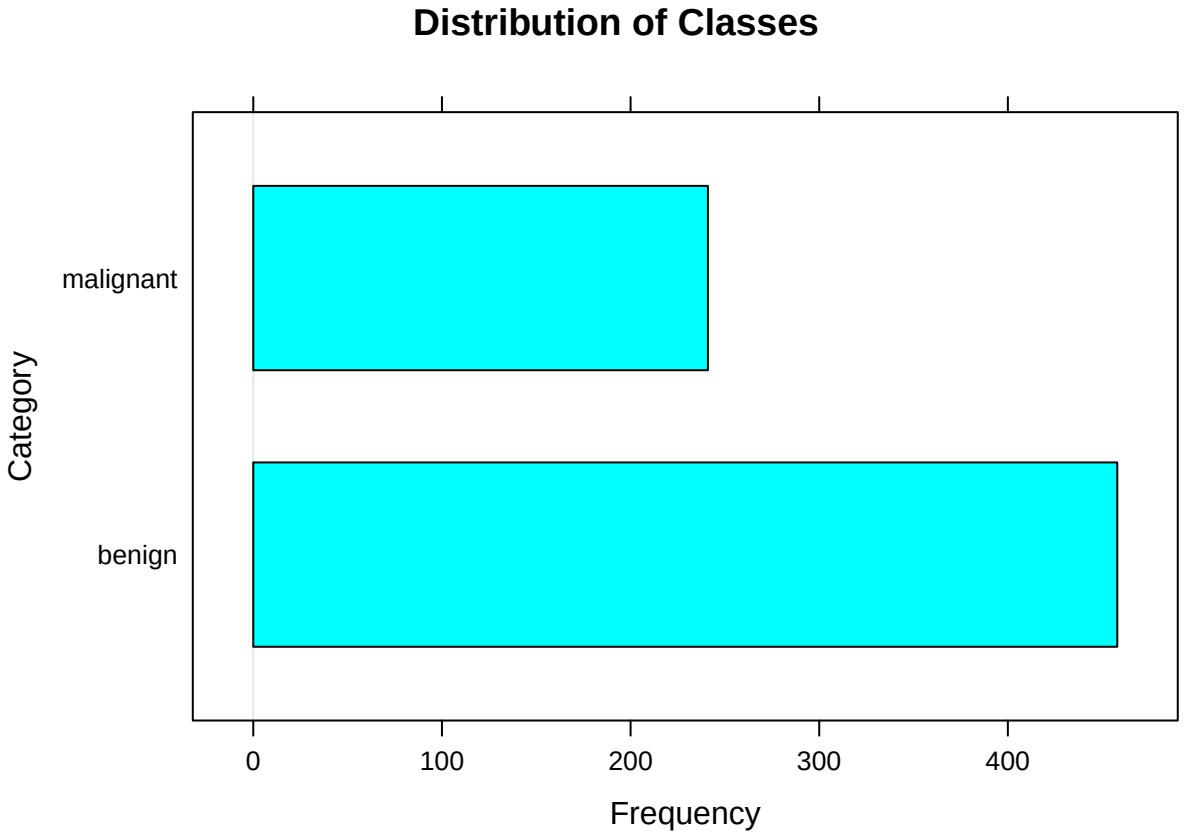
Based on the summary information given, which shows us the 3 quadrants, minimum, median, mean and maximum; a few information can be glimpse from it.

For instance, for the clump thickness, the mean of it is 4.418 meaning that average of clump thickness in the dataset is 4. This observation of mean values applies to all attributes of the dataset.

There summary table also note the attribute that possesses NA values which is Bare.nuclei. With this information, we can plan how to handle the NA values (e.g: removal/exclusion/imputation/etc.)

7.Displays the distribution of classes in the "BreastCancer" dataset

```
barchart(BreastCancer$Class, xlab="Frequency", ylab="Category", main="Distribution of Classes")
```



Comment on the distribution

Since this column contains categorical data of malignant and benign category, bar chart is the best to show its distribution. Based on the distribution, benign classes exceeds malignant classes in frequency. We can conclude that in this dataset, there are more benign classes data.

8.Displays the correlation of attributes in the "BreastCancer" dataset.

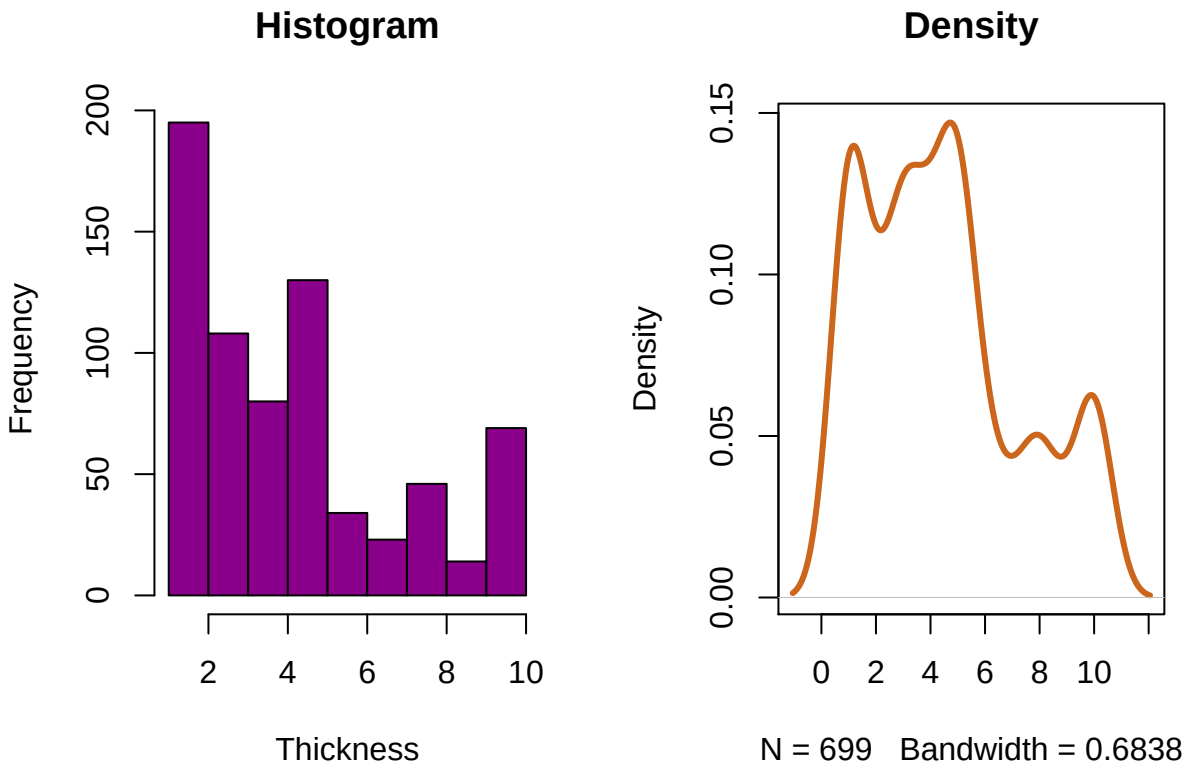
```
breast_correlation<-round(cor(breast),2)
```

9.Displays the distribution of individual attributes in the "Breast-Cancer" dataset.

Comments on graphs From the graphs shown, all the attributes shows the tendency towards skewed right distribution (positive skew).

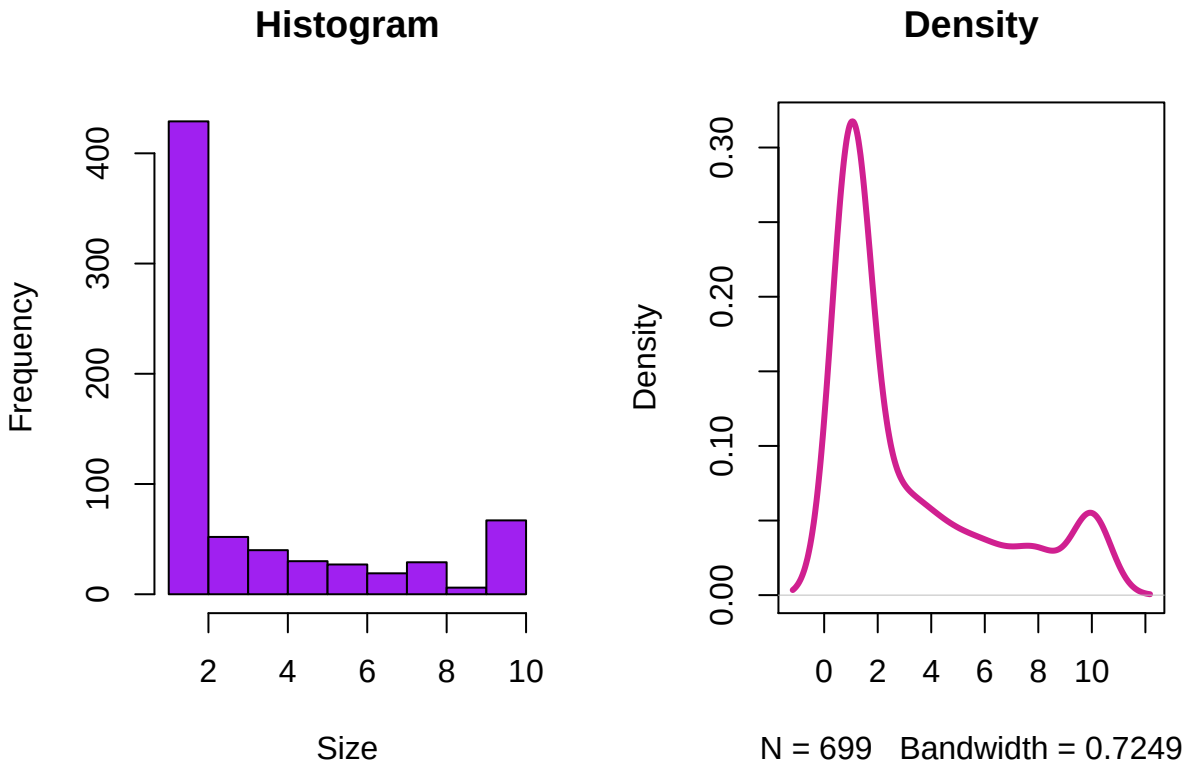
Clump Thickness

```
par(mfrow=c(1,2))
hist(breast$Cl.thickness,
     main="Histogram",
     xlab="Thickness",
     xlim=c(1,10),
     col="darkmagenta")
plot(density(breast$Cl.thickness),
     lwd=3,col = "chocolate3", main="Density")
```



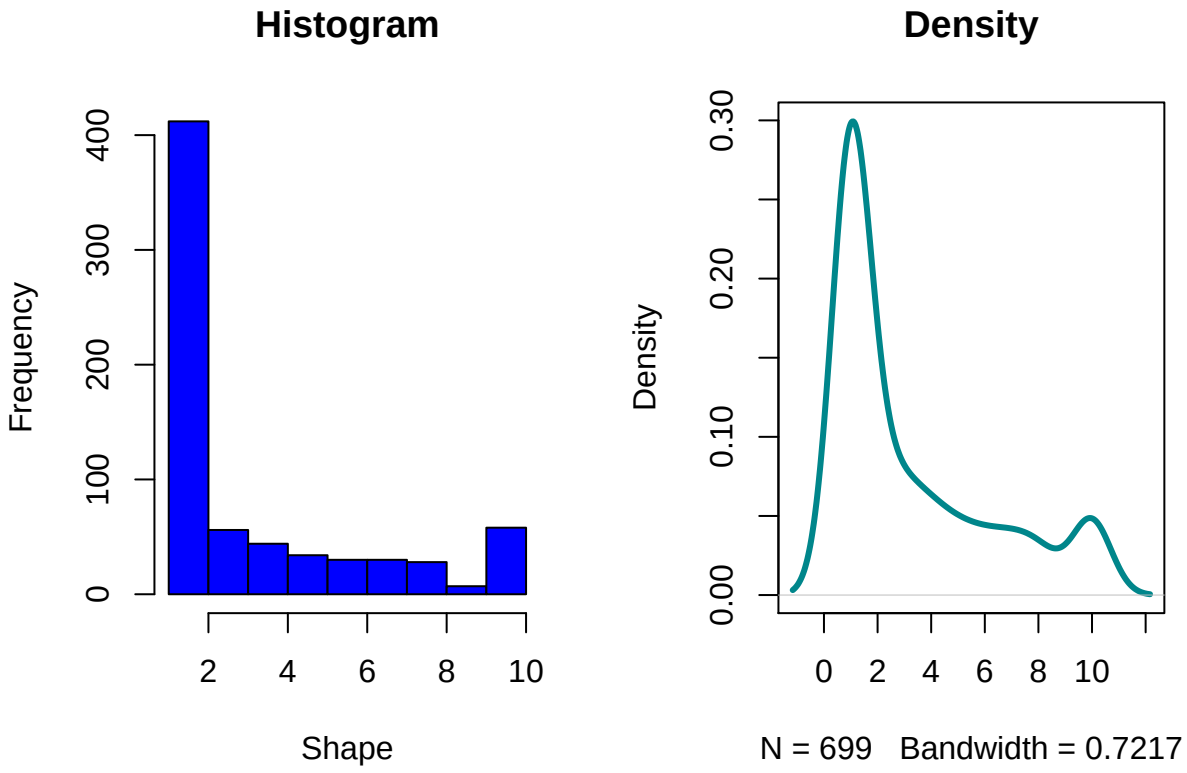
Uniformity of Cell Size.

```
par(mfrow=c(1,2))
hist(breast$Cell.size,
     main="Histogram",
     xlab="Size",
     xlim=c(1,10),
     col="purple")
plot(density(breast$Cell.size),
     lwd=3,col = "violetred", main="Density")
```



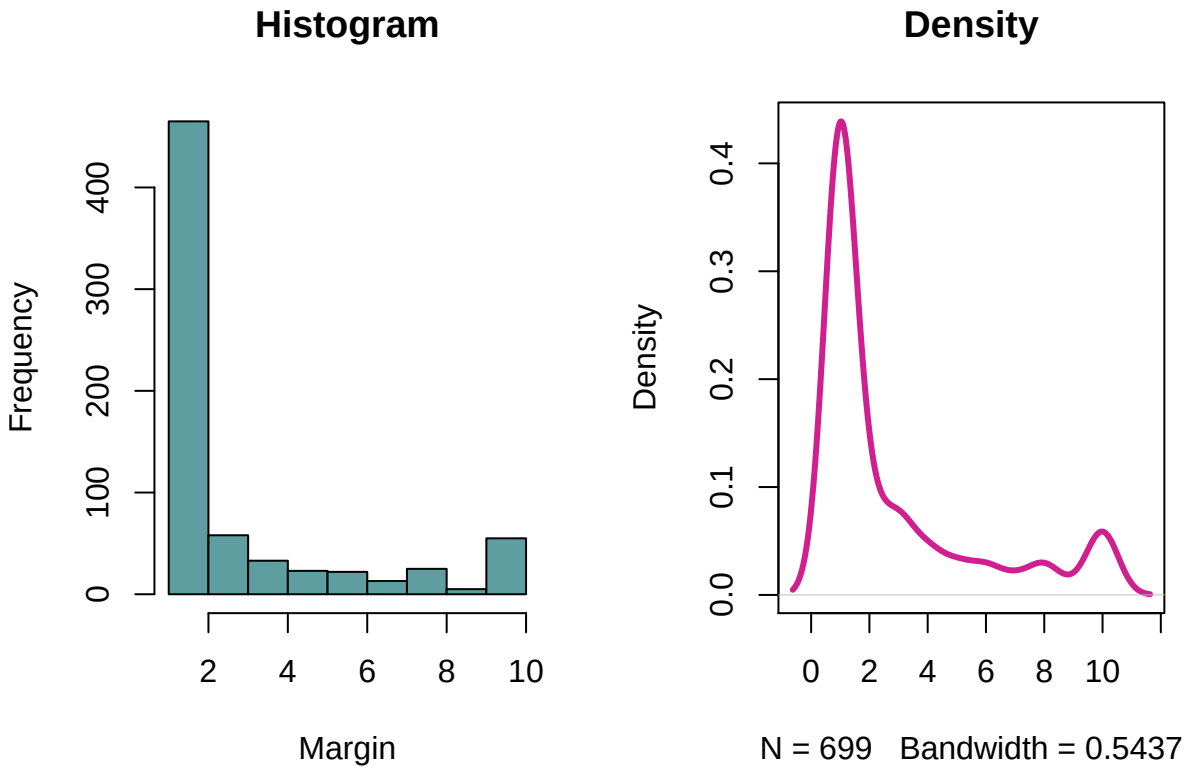
Uniformity of Cell Shape.

```
par(mfrow=c(1,2))
hist(breast$Cell.shape,
     main="Histogram",
     xlab="Shape",
     xlim=c(1,10),
     col="blue")
plot(density(breast$Cell.shape),
     lwd=3,col = "turquoise4", main="Density")
```



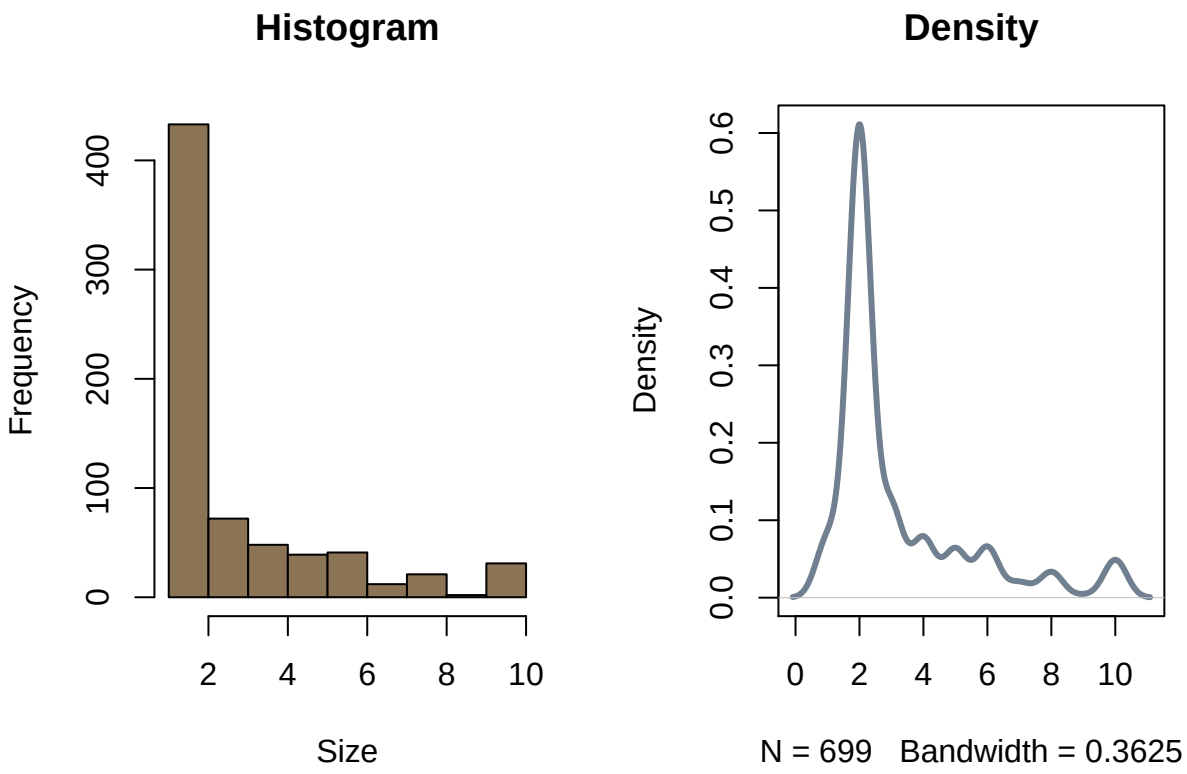
Marginal Adhesion.

```
par(mfrow=c(1,2))
hist(breast$Marg.adhesion,
     main="Histogram",
     xlab="Margin",
     xlim=c(1,10),
     col="cadetblue")
plot(density(breast$Marg.adhesion),
     lwd=3,col = "violetred", main="Density")
```



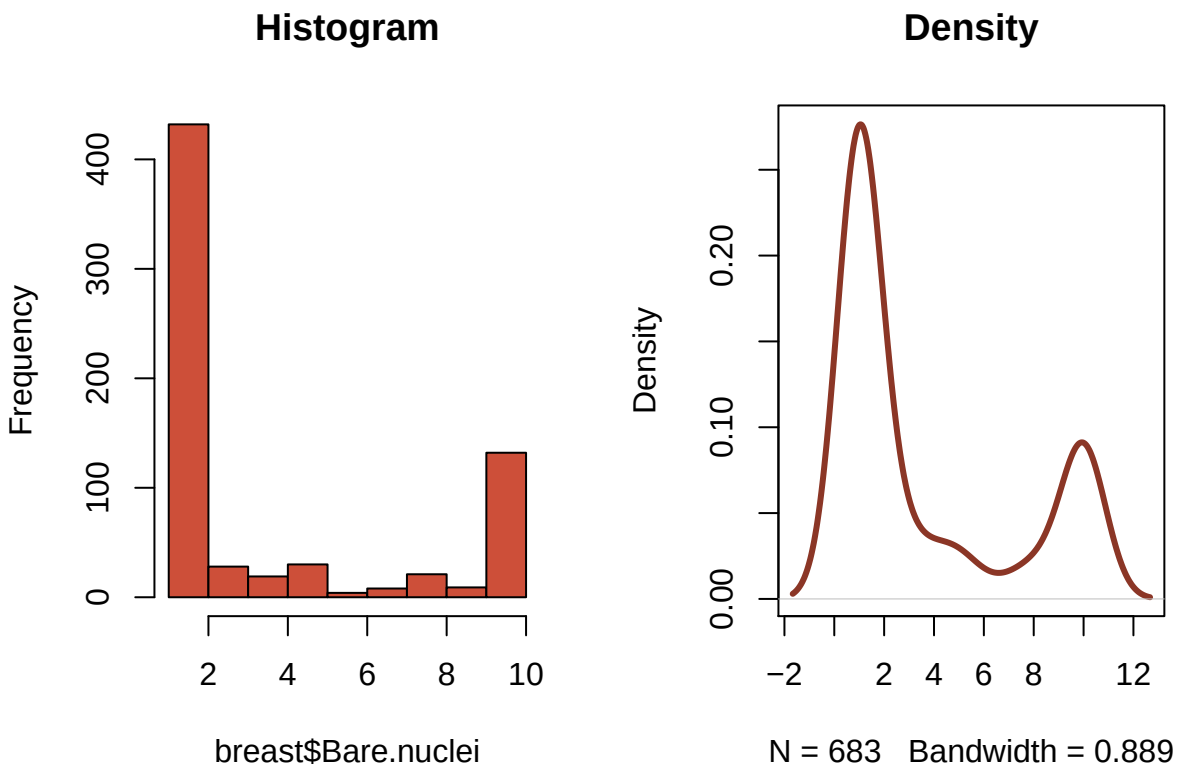
Single Epithelial Cell Size.

```
par(mfrow=c(1,2))
hist(breast$Epith.c.size,
     main="Histogram",
     xlab="Size",
     xlim=c(1,10),
     col="burlywood4")
plot(density(breast$Epith.c.size),
     lwd=3,col = "slategrey", main="Density")
```



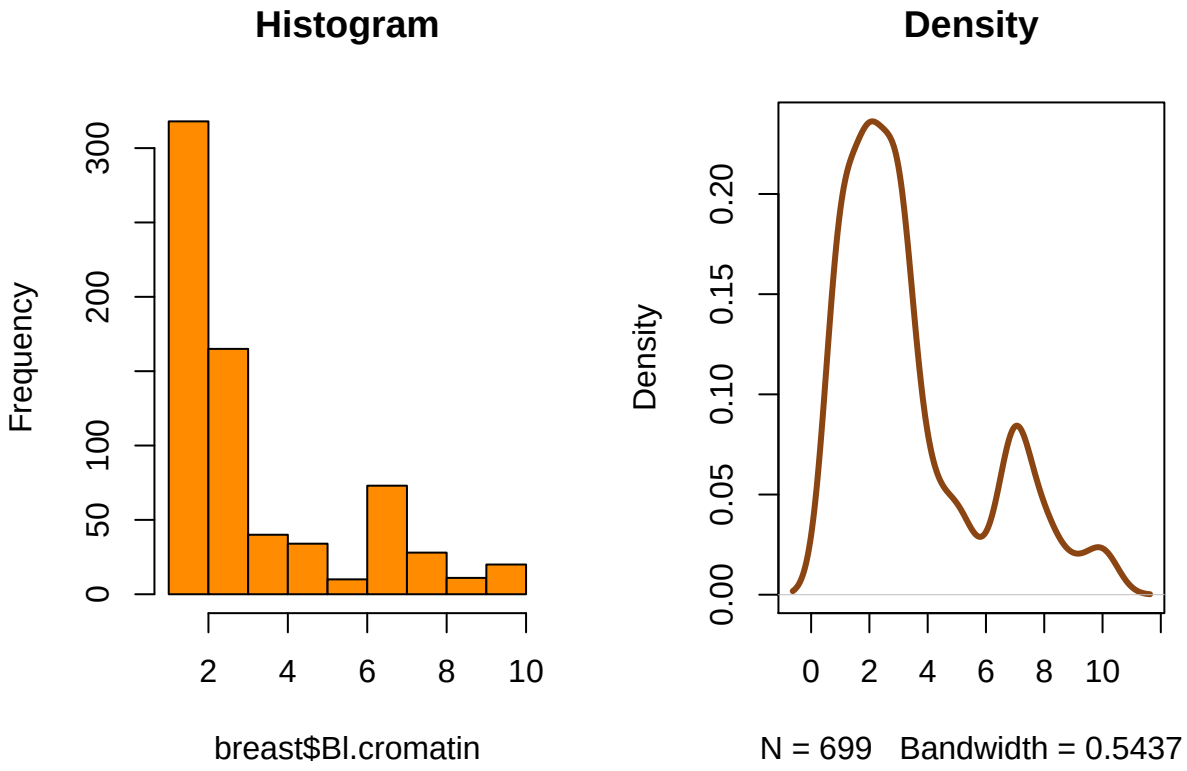
Bare Nuclei.

```
par(mfrow=c(1,2))
hist(breast$Bare.nuclei,
     main="Histogram",
     xlim=c(1,10),
     col="tomato3")
plot(density(breast$Bare.nuclei,na.rm = TRUE),
     lwd=3,col = "tomato4", main="Density")
```



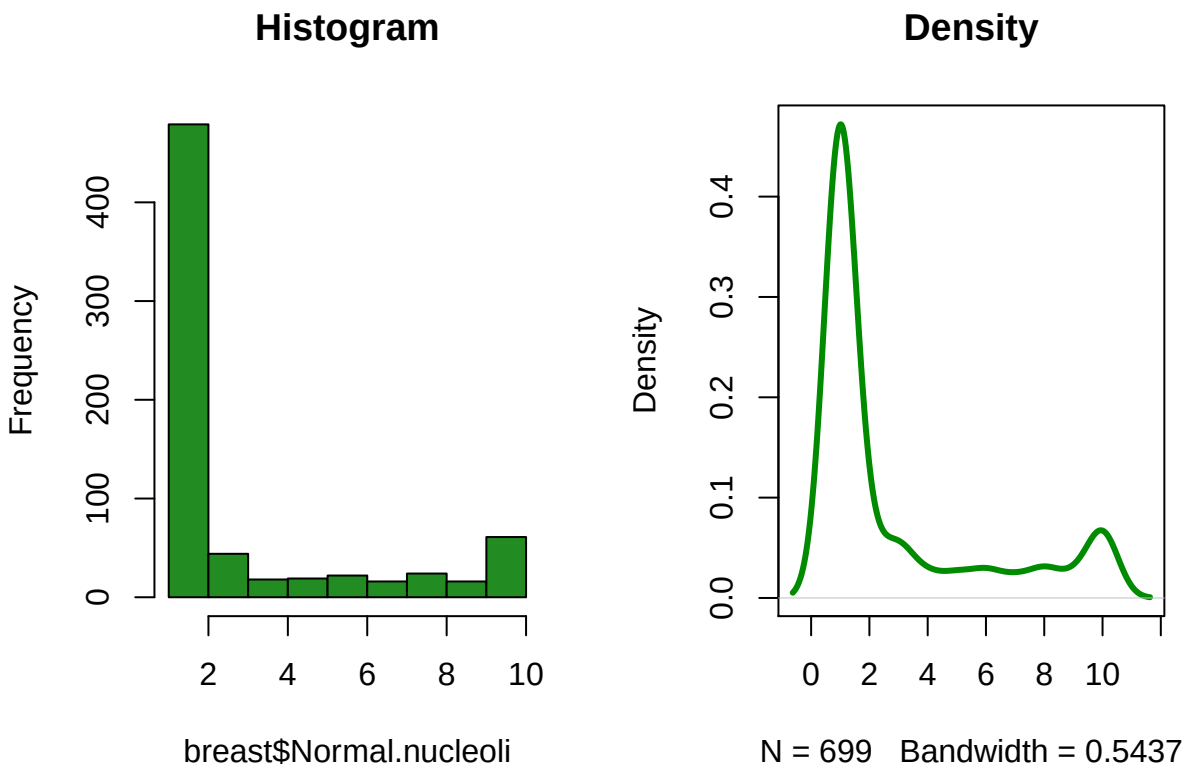
Bland Chromatin.

```
par(mfrow=c(1,2))
hist(breast$Bl.cromatin,
     main="Histogram",
     xlim=c(1,10),
     col="darkorange")
plot(density(breast$Bl.cromatin),
     lwd=3,col = "saddlebrown", main="Density")
```



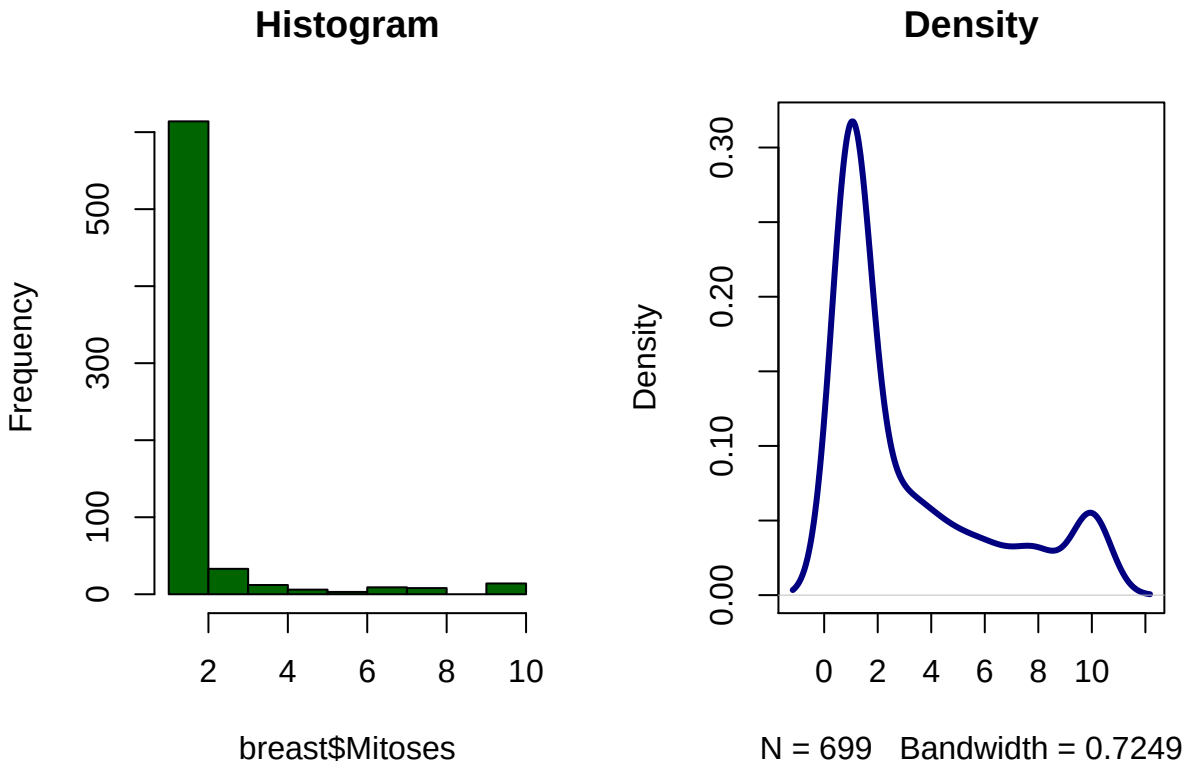
Normal Nucleoli.

```
par(mfrow=c(1,2))
hist(breast$Normal.nucleoli,
     main="Histogram",
     xlim=c(1,10),
     col="forestgreen")
plot(density(breast$Normal.nucleoli),
     lwd=3,col = "green4", main="Density")
```



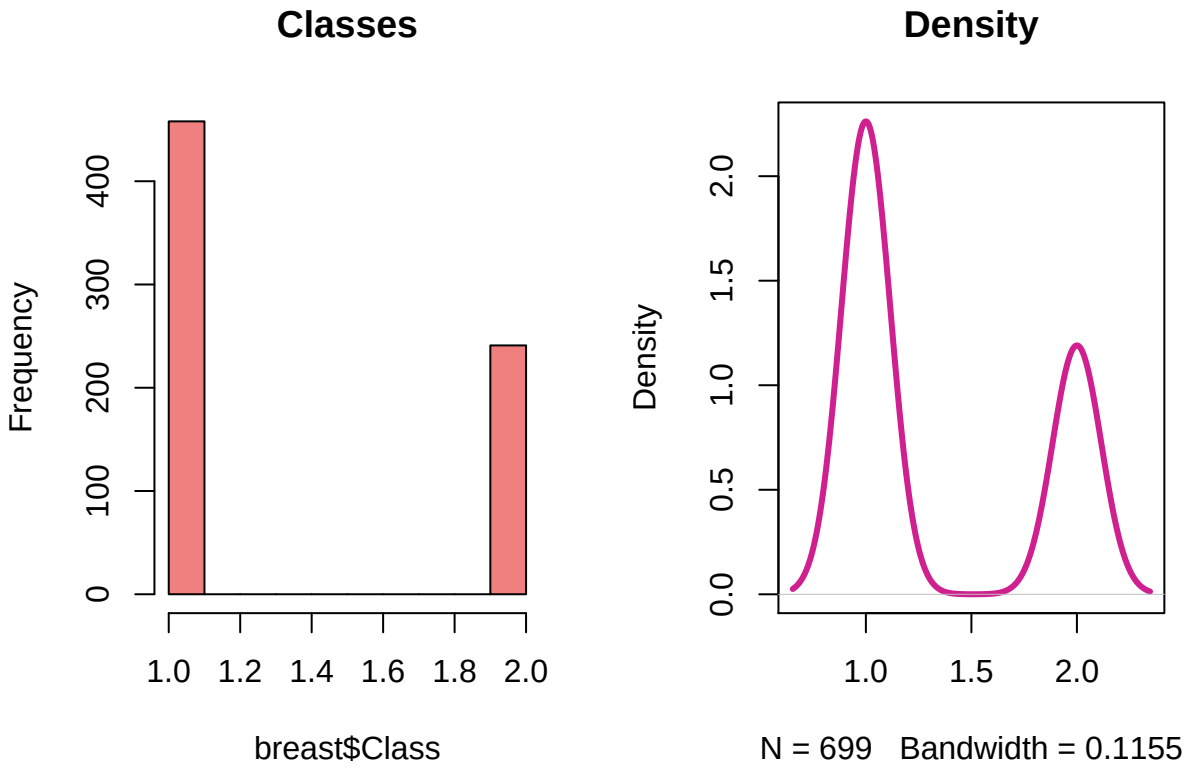
Mitoses.

```
par(mfrow=c(1,2))
hist(breast$Mitoses,
     main="Histogram",
     xlim=c(1,10),
     col="darkgreen")
plot(density(breast$Cell.size),
     lwd=3,col = "navyblue", main="Density")
```



Class.

```
par(mfrow=c(1,2))
hist(breast$Class,
     main="Classes",
     col="lightcoral")
plot(density(breast$Class),
     lwd=3,col = "violetred", main="Density")
```


Classes
Frequency
0
100
200
300
400
1.0
1.2
1.4
1.6
1.8
2.0
breast$Class
Density
Density
0.0
0.5
1.0
1.5
2.0
1.0
1.5
2.0
N = 699 Bandwidth = 0.1155