ds202_lab2

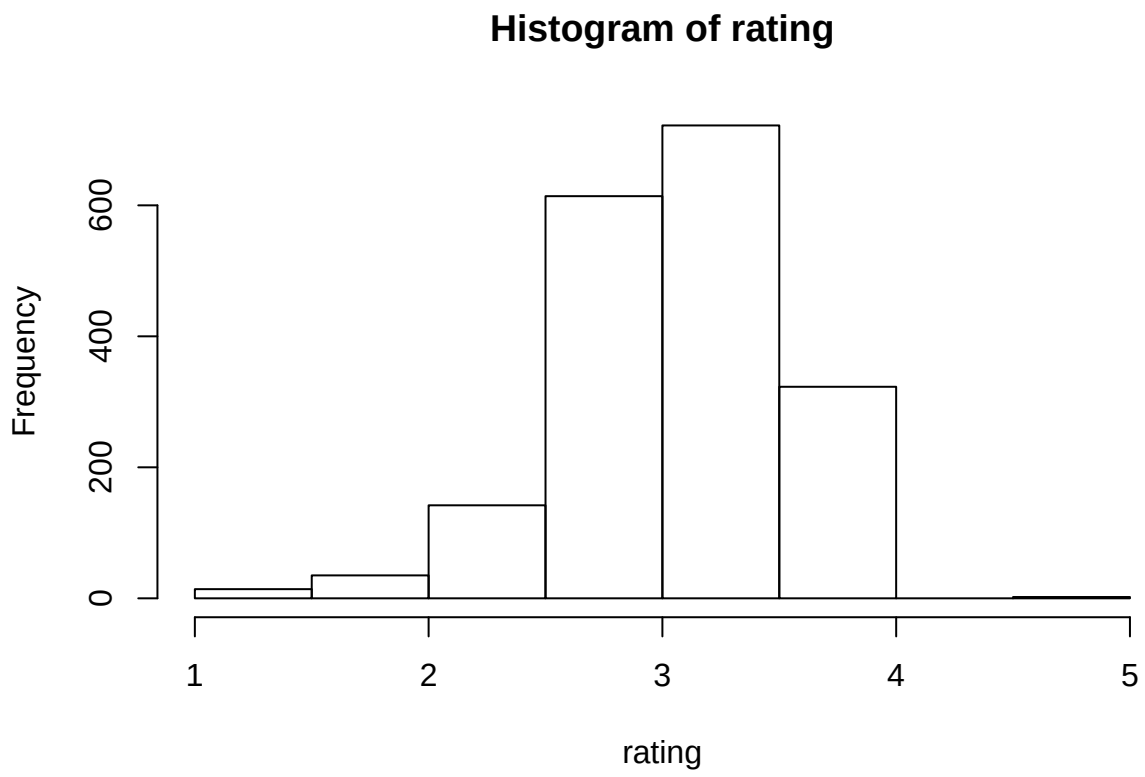
Anyesha Ray

February 5, 2020

```
choco <- read.csv("https://xdaiisu.github.io/ds202materials/hwlabs/choco.csv")
head(choco)
```

```
##    Company Specific.Bean.Origin  REF Review.Date Cocoa.Pct
## 1 A. Morin          Agua Grande 1876        2016        63
## 2 A. Morin                Kpime 1676        2015        70
## 3 A. Morin               Panama 1011        2013        70
## 4 A. Morin           Madagascar 1011        2013        70
## 5 A. Morin               Brazil 1011        2013        70
## 6 A. Morin             Equateur 1011        2013        70
##   Company.Location Rating Bean.Type Broad.Bean.Origin
## 1           France   3.75                   Sao Tome
## 2           France   2.75                       Togo
## 3           France   2.75                     Panama
## 4           France   3.00   Criollo        Madagascar
## 5           France   3.25                     Brazil
## 6           France   3.75                    Ecuador
```

```
rating=choco$Rating
hist_rating= hist(rating)
```

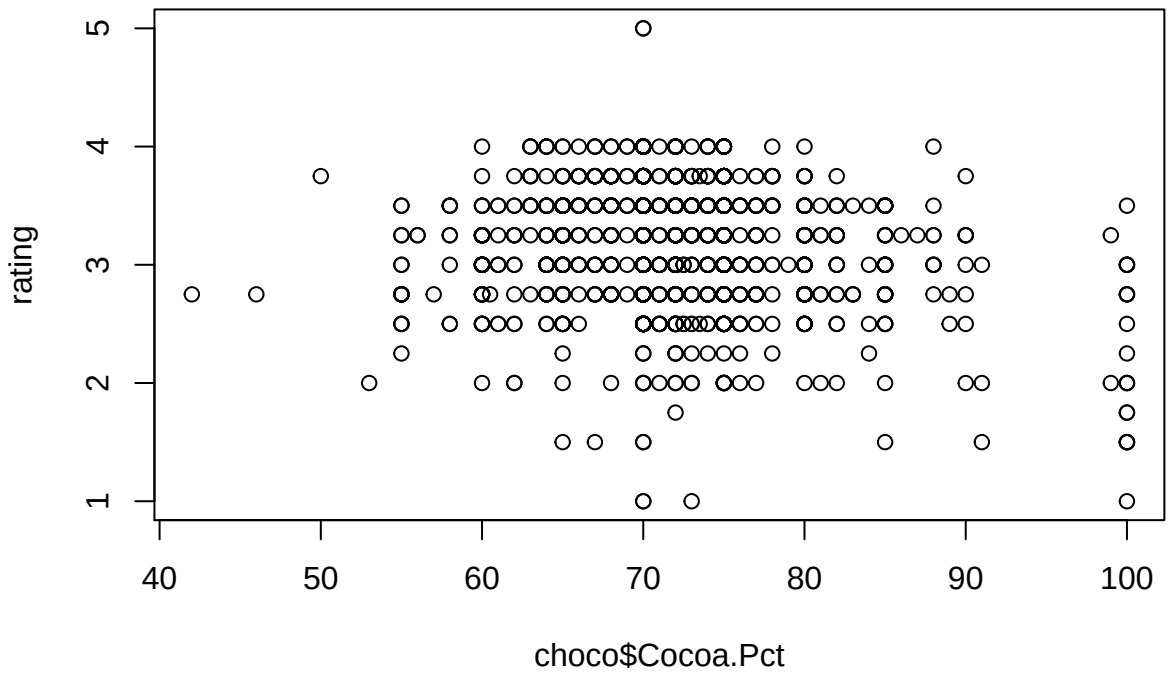


```
hist_rating
```

```
## $breaks
## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
##
## $counts
## [1]  14  35 142 614 722 323   0   2
##
## $density
## [1] 0.015118790 0.037796976 0.153347732 0.663066955 0.779697624 0.348812095
## [7] 0.000000000 0.002159827
##
## $mids
## [1] 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75
##
## $xname
## [1] "rating"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

The distribution, according to the histogram, is centered a little above 3. Most of the ratings are between 2.5 and 3.5. It looks fairly normal with a slight skew. I would say there are no outlier because if the mean is around 3 the values do not go high enough to be considered an outlier by the outlier test. However there only seems to be a few ratings above 4 so those could be considered outliers from just a visual perspective.

```
plot(choco$Cocoa.Pct, rating)
```




```
cor(choco$Cocoa.Pct, rating)
```

```
## [1] -0.1618855
```

I would say thetho variables do not depend on each other. There seems to be no correlation and they are all scattered about with no pattern visible in the scatter plot. The correlation is also very close to 0.