STAT 8010 R Lab 3: Data Summary/Visualization II

Whitney Huang

8/27/2020

Load the dataset

There are several ways to load a dataset into R:

- Importing data over the Internet

```
sport <- read.table("https://whitneyhuang83.github.io/STAT8010/Data/sport.txt", header = TRUE)
```

Let's take a look at the data

```
#sport
head(sport) # print the first 6 observations
```

```
##        sport
## 1     Others
## 2     Others
## 3   Football
## 4 Volleyball
## 5 Volleyball
## 6 Basketball
```

- Read the dataset from you computer

```
# Set working directory
setwd("/Users/wkhuang/Desktop/Desktop - mass-mini19-huang/Teaching/R/20Fall")
# This is the path  of the folder (in your computer).
getwd()
```

```
## [1] "/Users/wkhuang/Desktop/Desktop - mass-mini19-huang/Teaching/R/20Fall"
```

```
dir()
```

```
##  [1] "maxHeartRate.csv"   "SLR.Rmd"            "sport.txt"
##  [4] "STAT8010_RLab1.pdf" "STAT8010_RLab1.Rmd" "STAT8010_RLab2.pdf"
##  [7] "STAT8010_RLab2.Rmd" "STAT8010_RLab3.Rmd" "STAT8020_RLab1.pdf"
## [10] "STAT8020_RLab1.Rmd" "STAT8020_RLab2.pdf" "STAT8020_RLab2.Rmd"
## [13] "STAT8020_RLab3.pdf" "STAT8020_RLab3.Rmd"
```

```
sport1 <- read.table("sport.txt", header = TRUE)
```

Frequency Table

```
tab1 <- table(sport)
tab1 # print the table
```

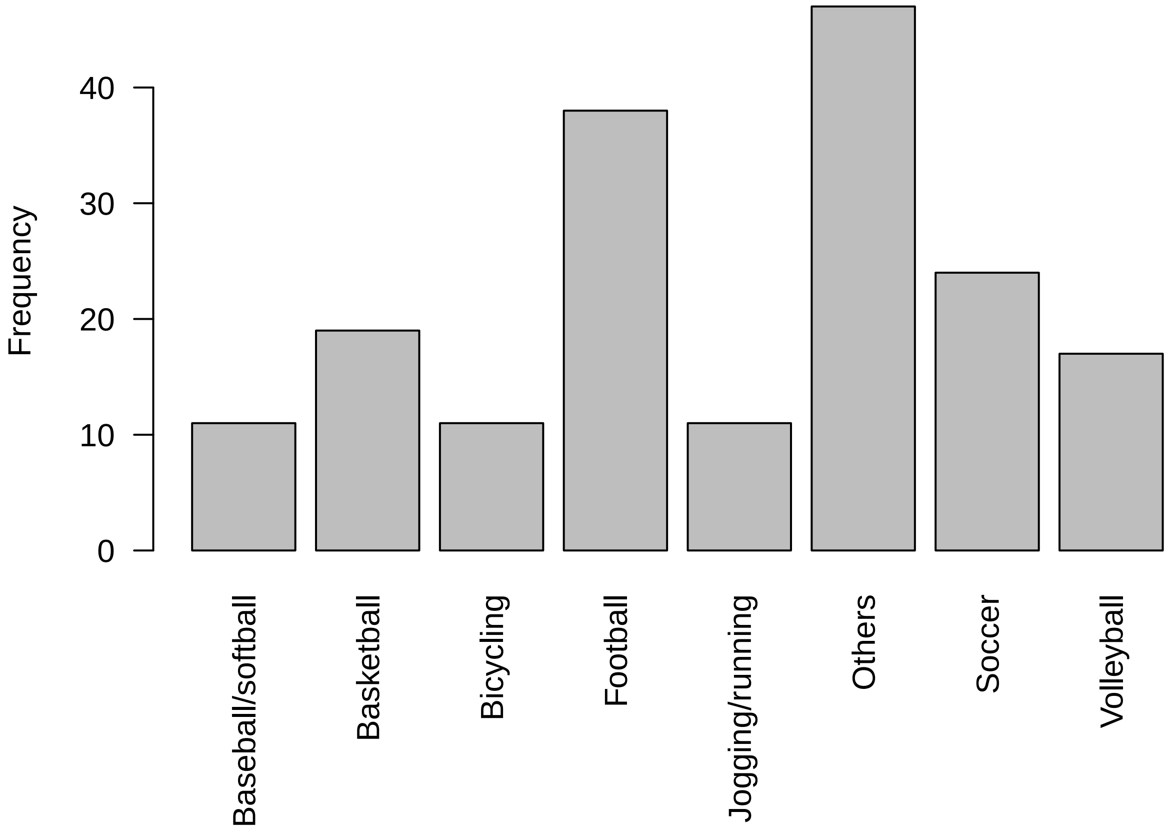
```
## sport
## Baseball/softball        Basketball         Bicycling          Football
##                11                19                11                38
##   Jogging/running            Others            Soccer        Volleyball
##                11                47                24                17
```

```
# Relative frequency
n <- dim(sport)[1] # sample size
tab2 <- table(sport) / n
tab2
```

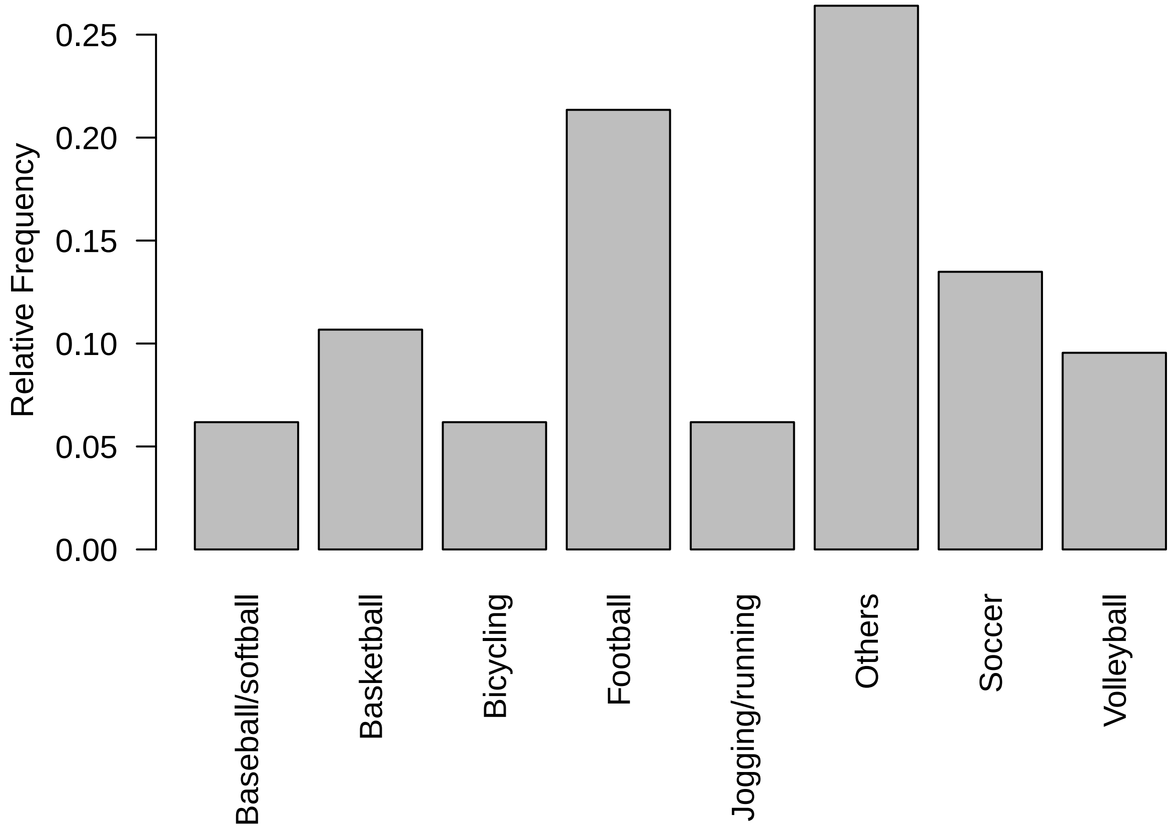
```
## sport
## Baseball/softball        Basketball         Bicycling          Football
##        0.06179775        0.10674157        0.06179775        0.21348315
##   Jogging/running            Others            Soccer        Volleyball
##        0.06179775        0.26404494        0.13483146        0.09550562
```

Bar Chart

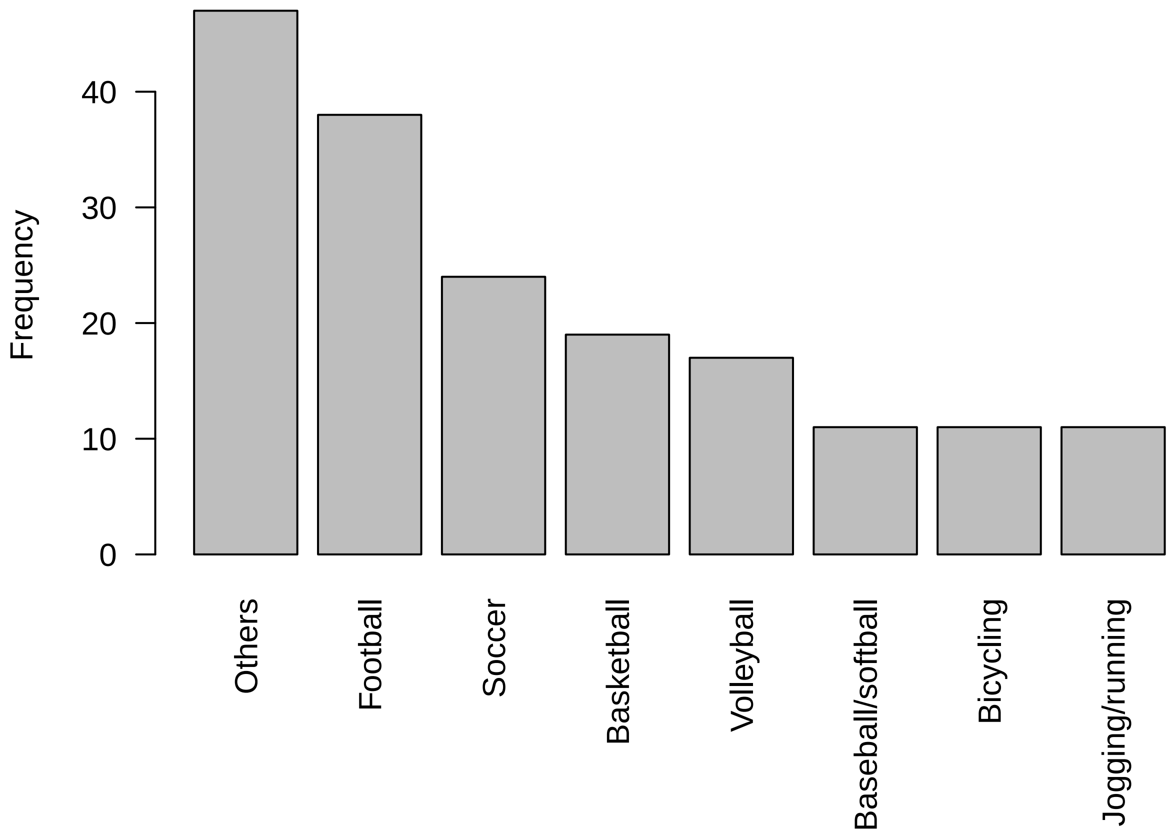
```
# Bart chart for the frequency
par(las = 2, mar = c(7.1, 4.1, 1.1, 1.1))
barplot(tab1, ylab = "Frequency")
```



```
# Bart chart for the relative frequency
par(las = 2, mar = c(7.1, 4.1, 1.1, 1.1))
barplot(tab2, ylab = "Relative Frequency")
```

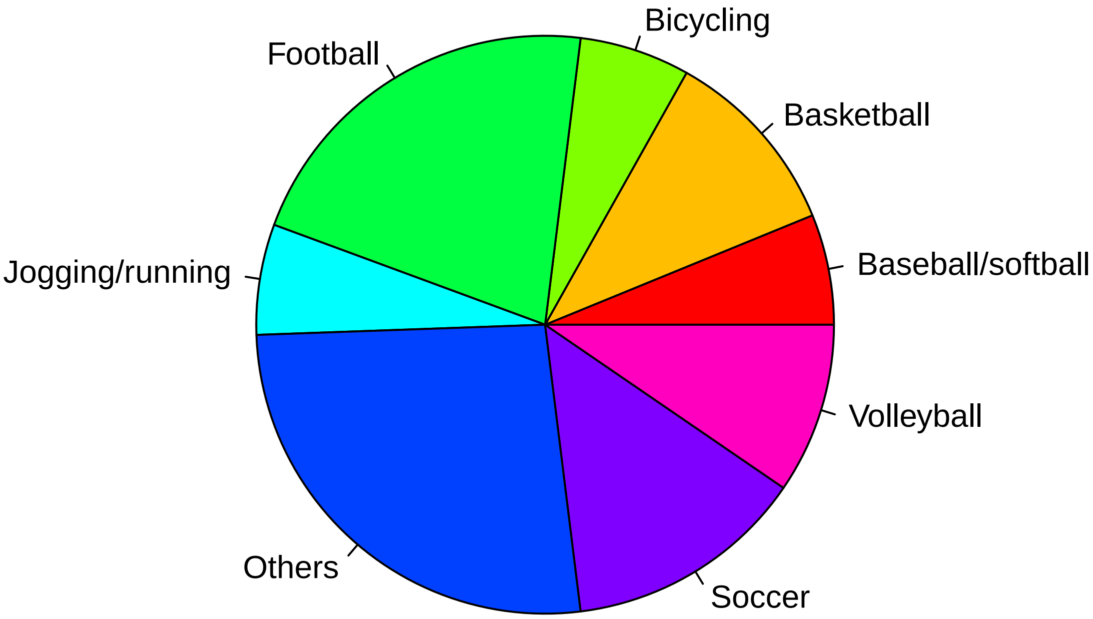


```
# Pareto chart
par(las = 2, mar = c(7.1, 4.1, 1.1, 1.1))
barplot(sort(tab1, decreasing = T), ylab = "Frequency")
```



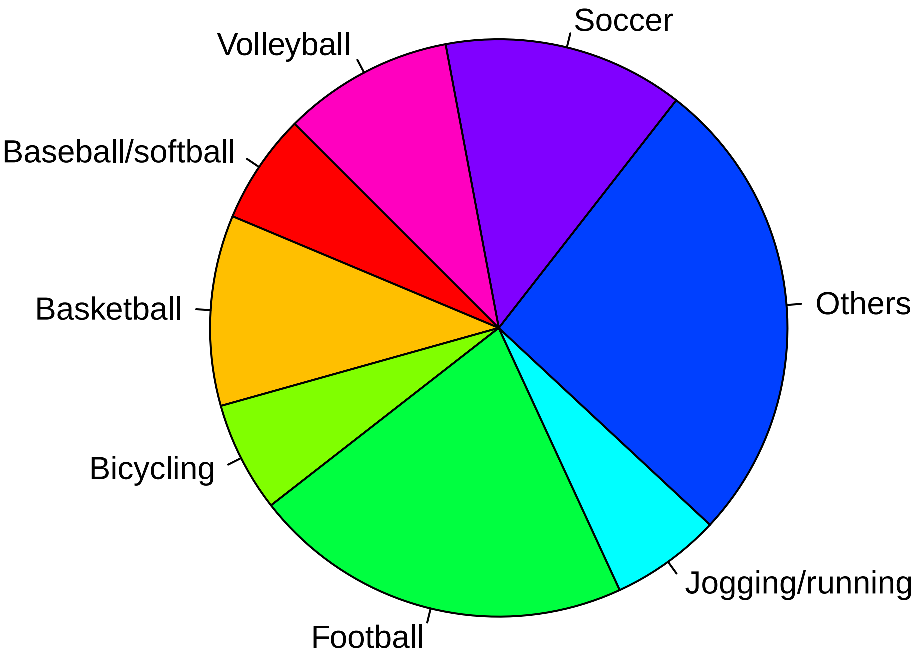
Pie Chart

```
par(mar = c(1.1, 3.1, 1.1, 3.1))
pie(tab1, col = rainbow(8))
```



```
# rotate the pie
par(mar = c(1.1, 3.1, 1.1, 3.1))
```

```
pie(table(sport), col = rainbow(8), init.angle = 135)
```



Violent Crime Rates by US State

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

```
data(USArrests) # this is a bulit-in data in R
dim(USArrests)
```

```
## [1] 50  4
```

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Stem-and-Leaf Plot

```
stem(USArrests$Murder)
```

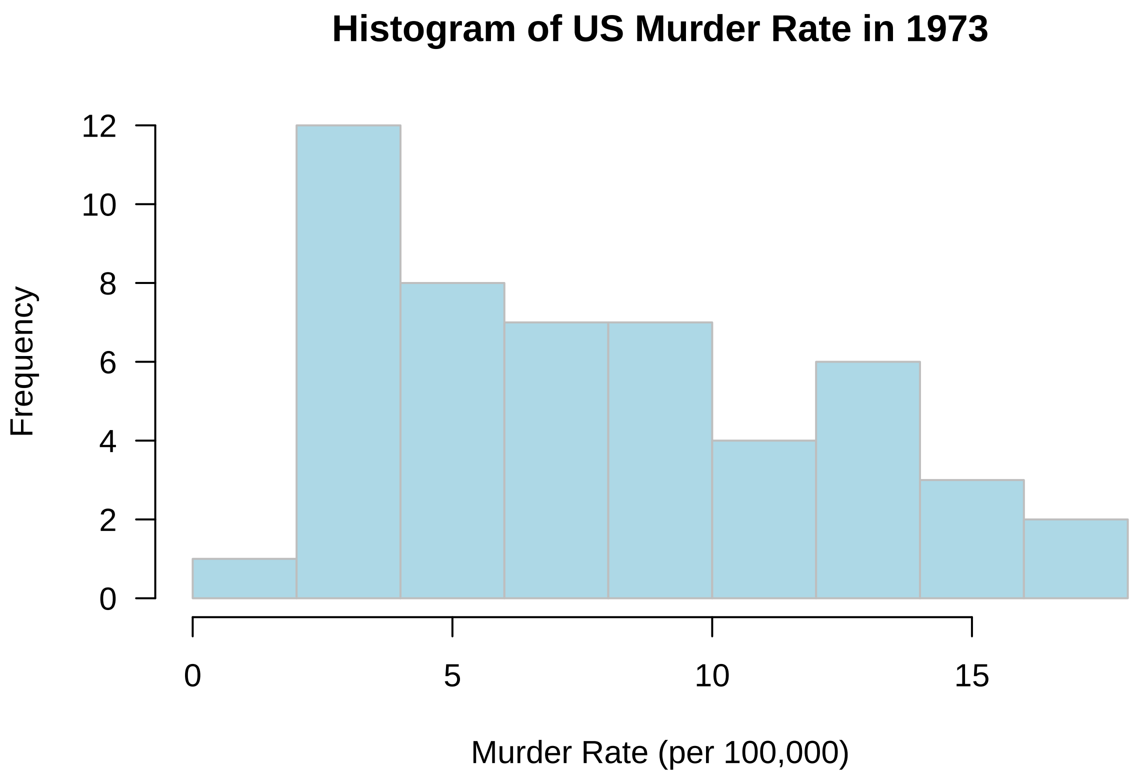
```
##
##   The decimal point is at the |
##
##    0 | 8
##    2 | 11226672348
##    4 | 0349379
##    6 | 003682349
##    8 | 158007
##   10 | 04134
##   12 | 127022
##   14 | 444
##   16 | 14
```

```
stem(USArrests$Murder, scale = 2)
```

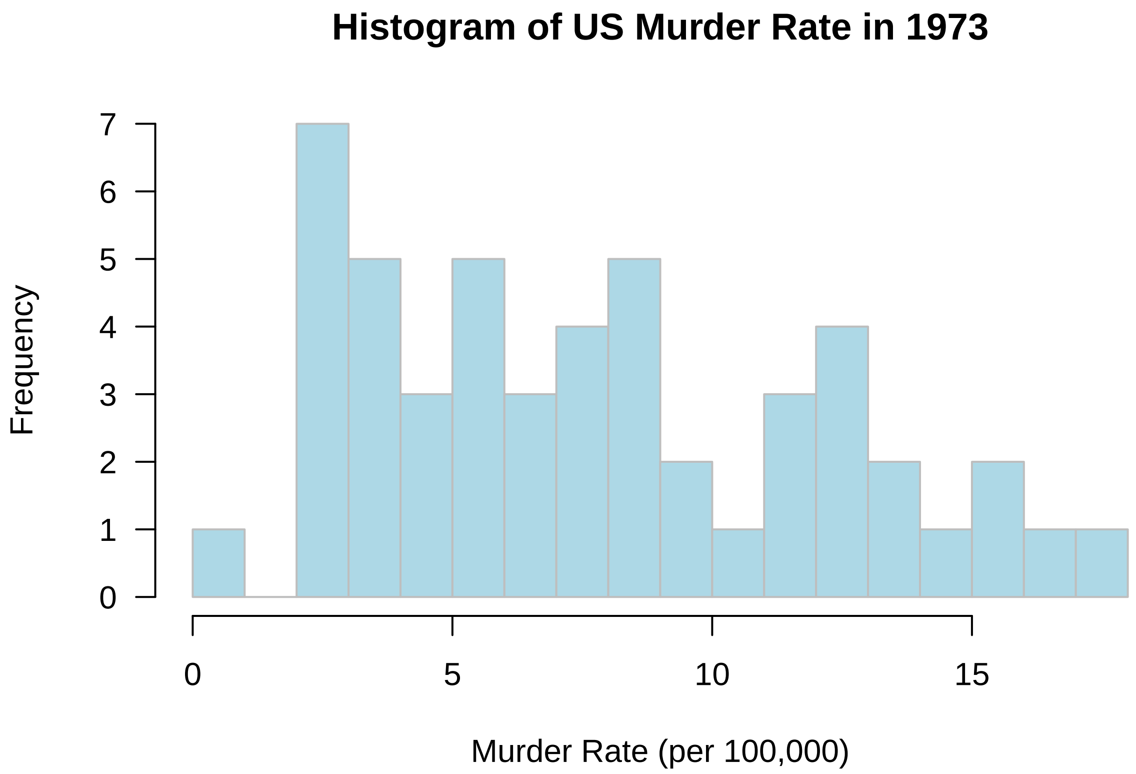
```
##
##   The decimal point is at the |
##
##    0 | 8
##    1 |
##    2 | 1122667
##    3 | 2348
##    4 | 0349
##    5 | 379
##    6 | 00368
##    7 | 2349
##    8 | 158
##    9 | 007
##   10 | 04
##   11 | 134
##   12 | 127
##   13 | 022
##   14 | 4
##   15 | 44
##   16 | 1
##   17 | 4
```

Histogram

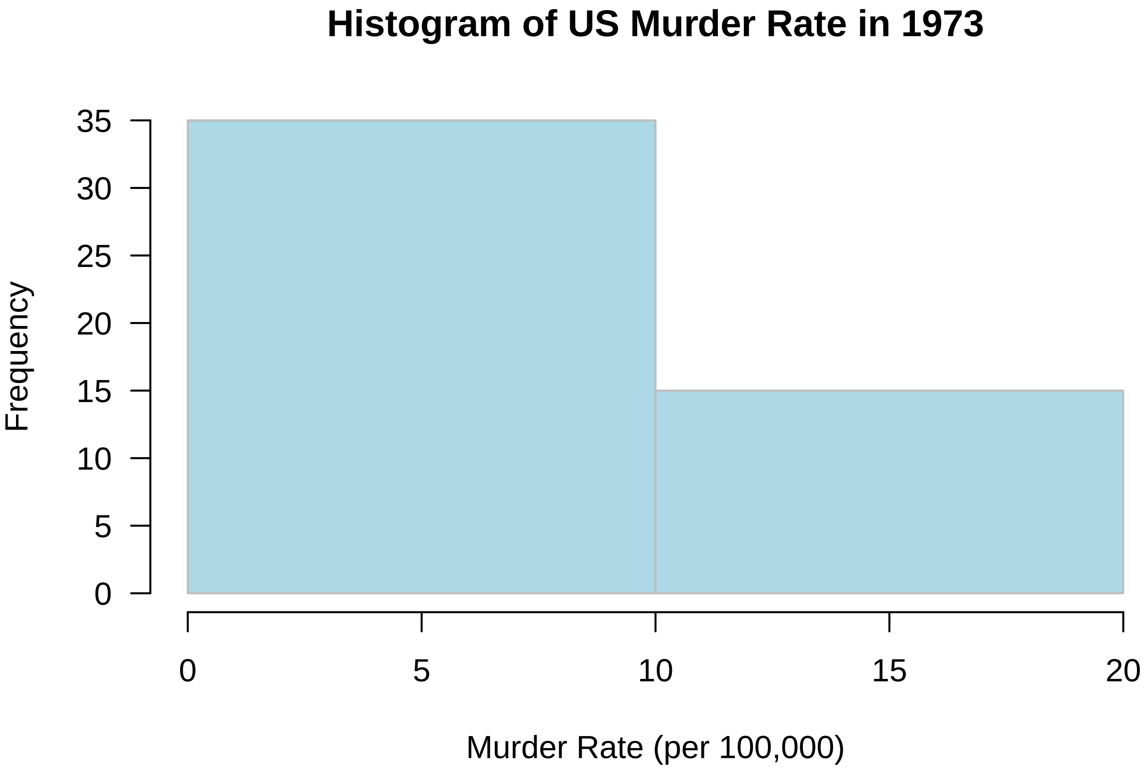
```
par(las = 1)
hist(USArrests$Murder, main = "Histogram of US Murder Rate in 1973",
     col = "lightblue", border = "gray", xlab = "Murder Rate (per 100,000)")
```



```r
# Let's change the bin size
par(las = 1)
hist(USArrests$Murder, nclass = 15,
     main = "Histogram of US Murder Rate in 1973", col = "lightblue",
     border = "gray", xlab = "Murder Rate (per 100,000)")
```

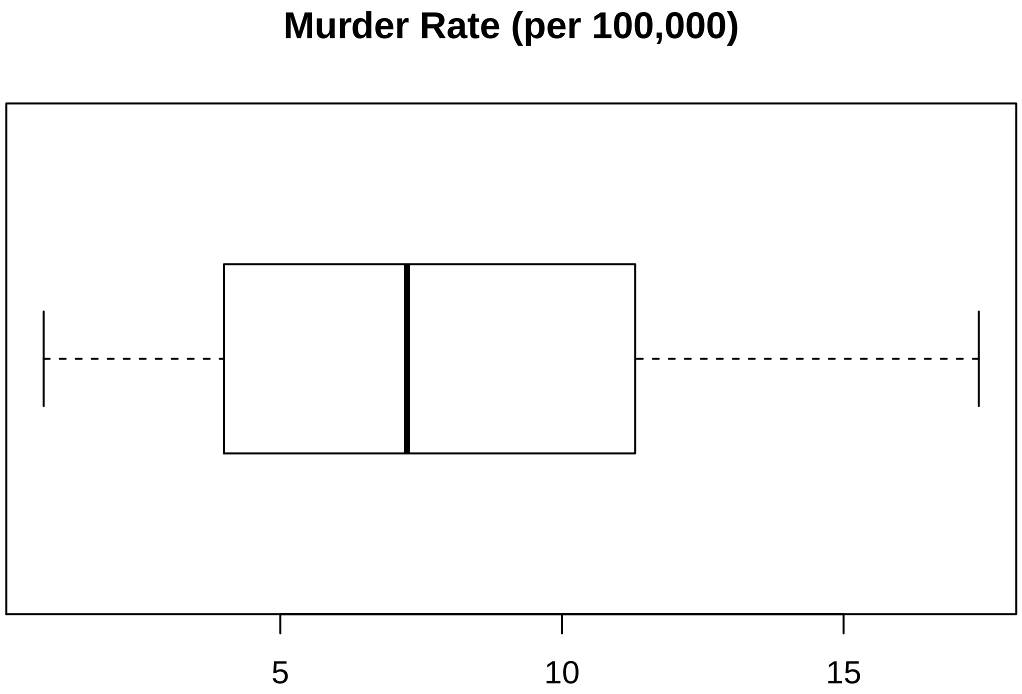


```
# Let's change the bin size again
par(las = 1)
hist(USArrests$Murder, nclass = 2,
     main = "Histogram of US Murder Rate in 1973", col = "lightblue",
     border = "gray", xlab = "Murder Rate (per 100,000)")
```

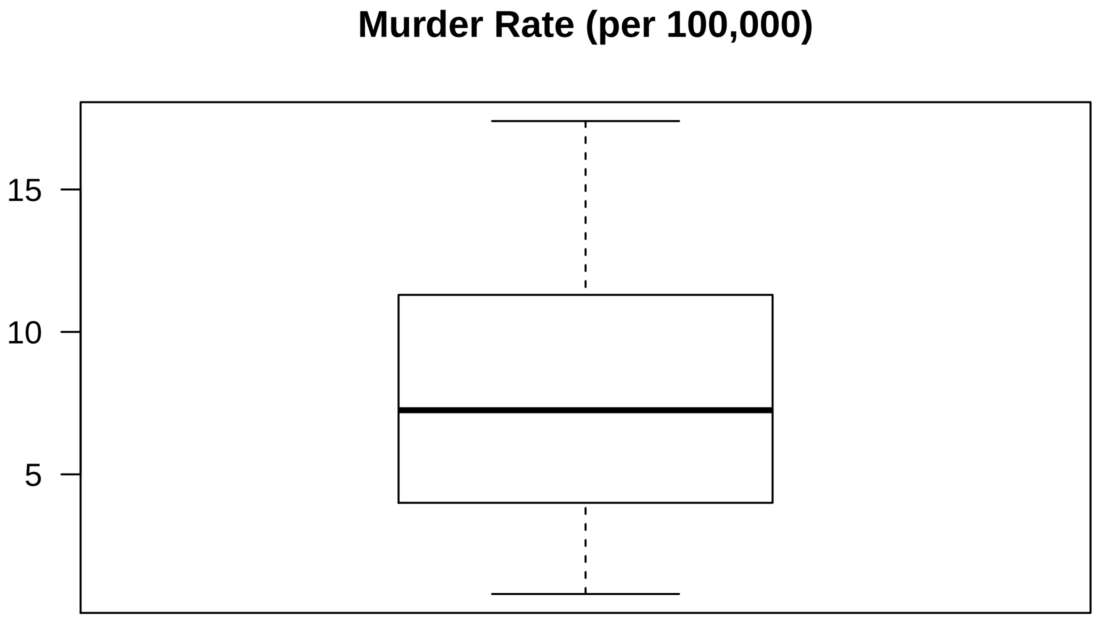


Boxplot

```
# Horizontal boxplot
par(las = 1)
boxplot(USArrests$Murder, main = "Murder Rate (per 100,000)", horizontal = T)
```



```r
# Vertical boxplot
par(las = 1)
boxplot(USArrests$Murder, main = "Murder Rate (per 100,000)")
```



Numerical summary of central tendency and variability

```r
mean(USArrests$Murder)
```

```
## [1] 7.788
```

```r
median(USArrests$Murder)
```

```
## [1] 7.25
```

```r
sort(table(USArrests$Murder), decreasing = T)
```

```
## 
##  2.1  2.2  2.6    6    9 13.2 15.4  0.8  2.7  3.2  3.3  3.4  3.8    4  4.3  4.4 
##    2    2    2    2    2    2    2    1    1    1    1    1    1    1    1    1 
##  4.9  5.3  5.7  5.9  6.3  6.6  6.8  7.2  7.3  7.4  7.9  8.1  8.5  8.8  9.7   10 
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1 
## 10.4 11.1 11.3 11.4 12.1 12.2 12.7   13 14.4 16.1 17.4 
##    1    1    1    1    1    1    1    1    1    1    1
```

```r
var(USArrests$Murder)
```

```
## [1] 18.97047
```

```r
sd(USArrests$Murder)
```

```
## [1] 4.35551
```

```r
IQR(USArrests$Murder)
```

```
## [1] 7.175
```

```r
range(USArrests$Murder)
```

```
## [1]  0.8 17.4
```

```r
diff(range(USArrests$Murder))
```

```
## [1] 16.6
```

Load the ORD flight dataset

```r
url <- "https://whitneyhuang83.github.io/STAT8010/Data/flights.csv"
ORD <- read.csv(url, header = TRUE)
```

Let’s take a look at the data

```r
dim(ORD)
```

```
## [1] 12678     4
```

```r
n <- dim(ORD)[1]
head(ORD)
```

```
##   month carrier origin arr_delay
## 1     1      UA    EWR        12
## 2     1      AA    LGA         8
## 3     1      AA    LGA        14
## 4     1      AA    LGA         4
## 5     1      UA    LGA        20
## 6     1      UA    EWR        21
```

2 way Frequency Table

```r
tab3 <- table(ORD[, c("carrier", "origin")])
tab3
```

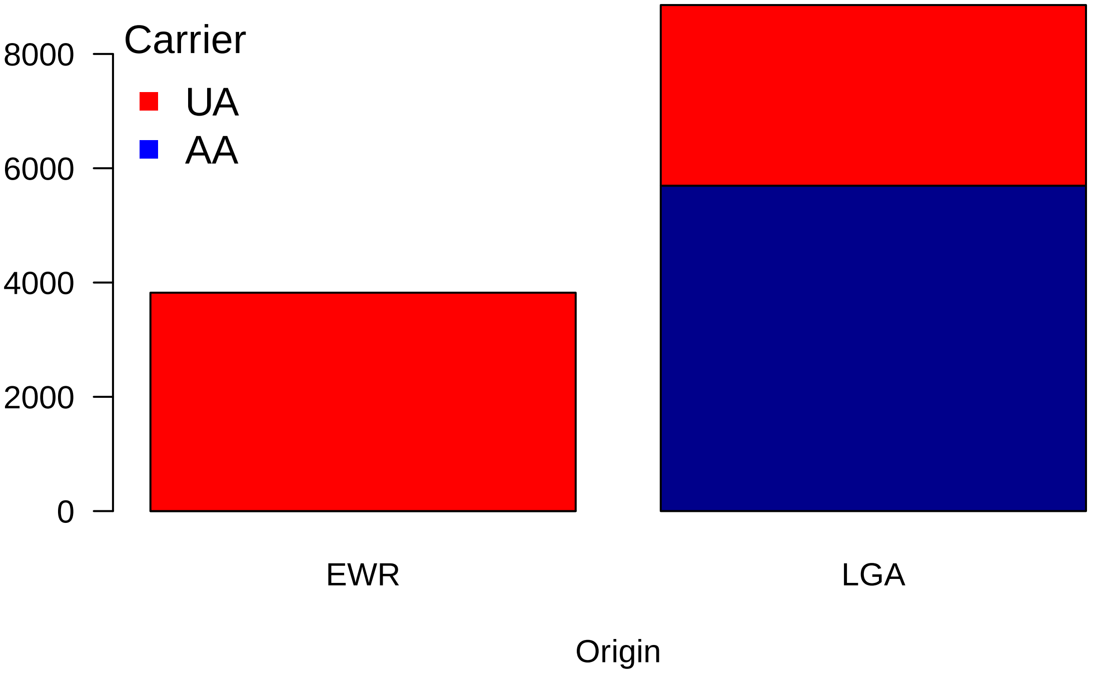
```
##        origin
## carrier  EWR  LGA
##      AA    0 5694
##      UA 3822 3162
```

```
tab4 <- table(ORD[, c("carrier", "origin")]) / n
tab4
```

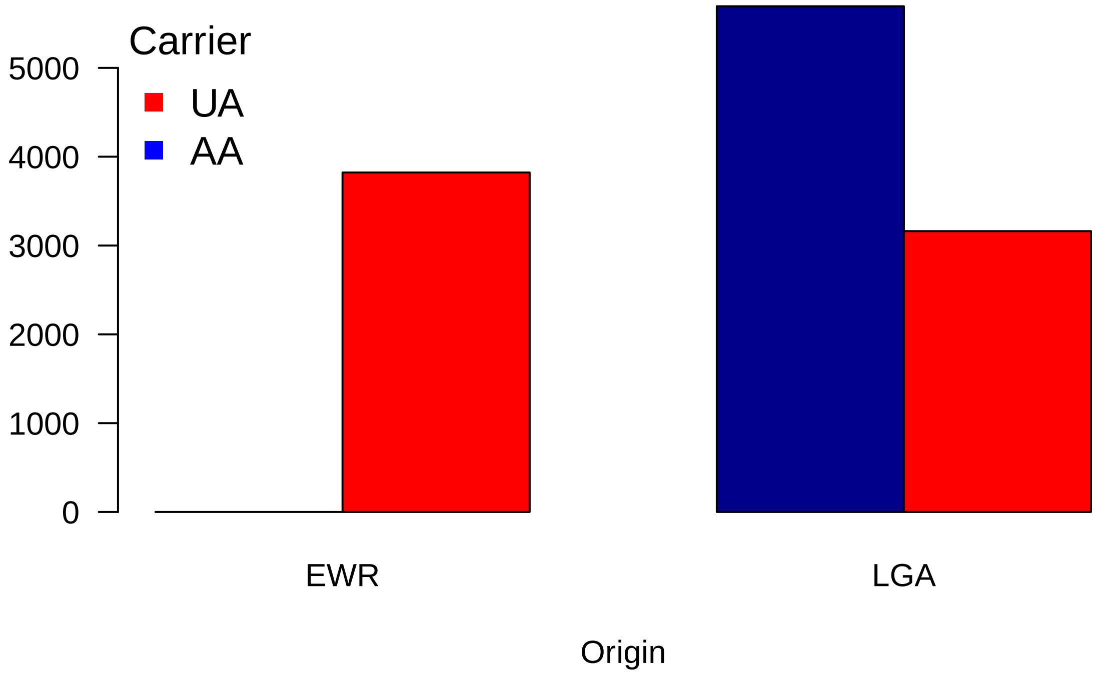
```
##        origin
## carrier       EWR       LGA
##      AA 0.0000000 0.4491245
##      UA 0.3014671 0.2494084
```

Stacked/dodged bar chart

```
## Stacked bar chart
barplot(tab3, xlab = "Origin", col = c("darkblue","red"), args.legend = list(x = "topleft"), las = 1)
legend("topleft", legend = c("UA", "AA"),
       pch = 15, col = c("red", "blue"), bty = "n", cex = 1.25, title = "Carrier")
```



```
## Dodged bar chart
barplot(tab3, xlab = "Origin", col = c("darkblue","red"), args.legend = list(x = "topleft"), las = 1, b
legend("topleft", legend = c("UA", "AA"),
       pch = 15, col = c("red", "blue"), bty = "n", cex = 1.25, title = "Carrier")
```



Qualitative vs Quantitative: Side by Side Boxplots

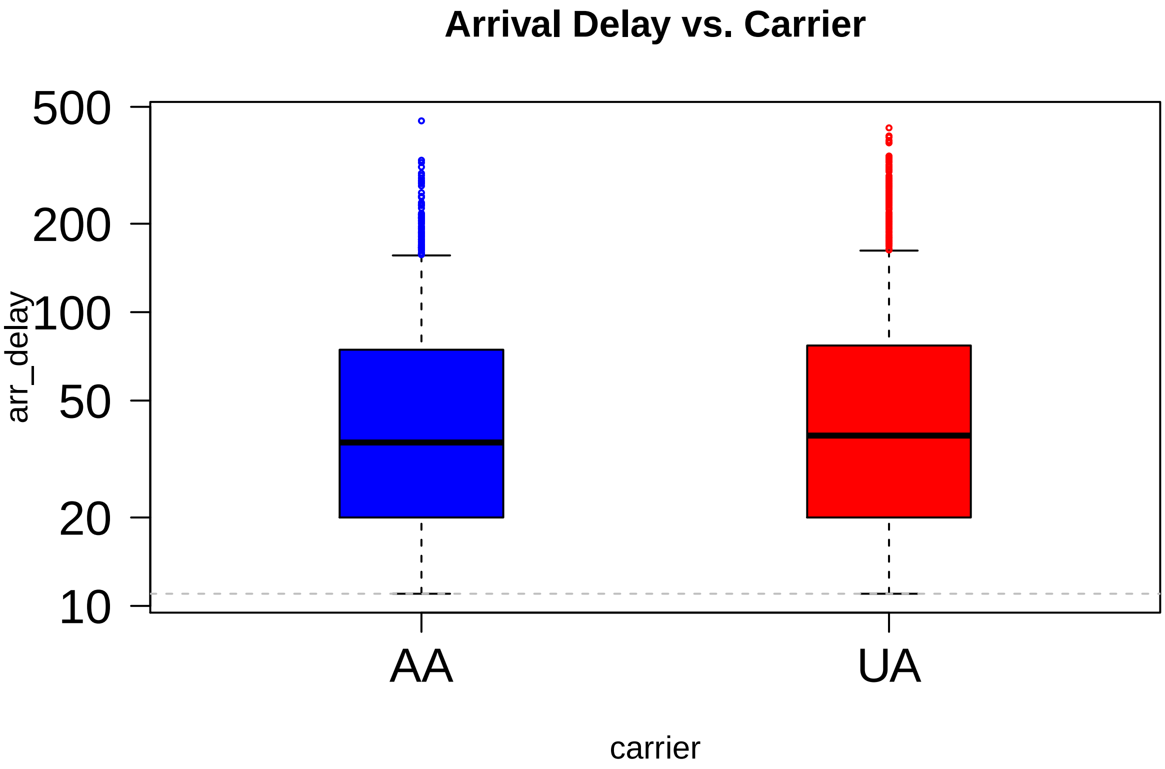
```
attach(ORD)
library(tidyverse)
```

```
## -- Attaching packages ------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ---------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

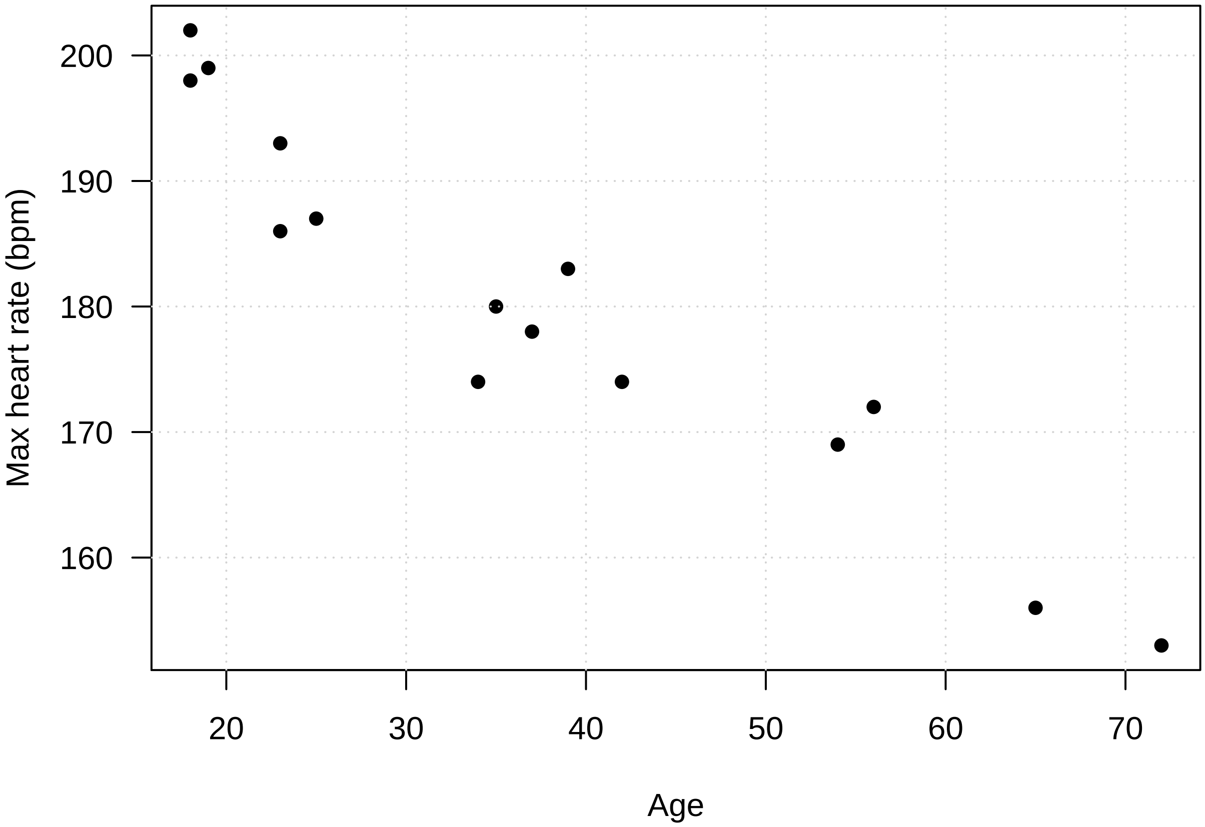
```
boxplot(arr_delay ~ carrier, filter(ORD, arr_delay > 10), boxwex = 0.35,
        col = c("blue", "red"),
        staplewex = 0.35, outwex = 0.35,
        cex.axis = 1.5, las = 1, log = "y",
        outcol = c("blue", "red"),
        outcex = 0.35, main = "Arrival Delay vs. Carrier")
abline(h = 11, lty = 2, col = "gray")
```



Quantitative vs Quantitative: Scatter Plot

```
url <- "https://whitneyhuang83.github.io/STAT8010/Data/maxHeartRate.csv"
dat <- read.csv(url, header = TRUE)

par(las = 1, mar = c(4.1, 4.1, 1.1, 1.1))
plot(dat$Age, dat$MaxHeartRate, pch = 16, xlab = "Age", ylab = "Max heart rate (bpm)")
grid()
```


Max heart rate (bpm)
200
190
180
170
160
20
30
40
50
60
70
Age