STAT 8010 R Session 1: Exploratory Data Analysis

Whitney Huang

5/19/2023

Contents

Lab Objective

- To gain experience with R, a programming language and free software environment for statistical computing and graphics.
- To read data into R.
- To perform exploratory data analysis using `R`

Setup

- You should have R installed, if not, open a web browser and go to (http://cran.r-project.org) and download and install R. It also helpful to install RStudo (http://rstudio.com).
- Create a folder for this R lab. Download the Maximum Heart Rate dataset at (http://whitneyhuang83.github.io/STAT8010/Data/maxHeartRate.csv) and save it in the folder you just created.

Load a max heart rate dataset

There are several ways to load a dataset into R:

Importing Data over the Internet

```
dat <- read.csv('http://whitneyhuang83.github.io/STAT8010/Data/maxHeartRate.csv', header = T)
ls()
```

```
## [1] "dat"
```

Let's take a look at the data

```
dat
```

```
##    Age MaxHeartRate
## 1   18          202
## 2   23          186
## 3   25          187
## 4   35          180
## 5   65          156
## 6   54          169
## 7   34          174
## 8   56          172
## 9   72          153
## 10  19          199
```

```
## 11  23           193
## 12  42           174
## 13  18           198
## 14  39           183
## 15  37           178
```

```
head(dat, 3)
```

```
##   Age MaxHeartRate
## 1  18          202
## 2  23          186
## 3  25          187
```

```
tail(dat, 3)
```

```
##    Age MaxHeartRate
## 13  18          198
## 14  39          183
## 15  37          178
```

Read the dataset from you computer

Note that you will need to either place the data file in the same folder as the Rmd file or specify the path to the data file if it is located in a different folder.

```
dat <- read.csv('maxHeartRate.csv', header = T)
```

Type the data into R

```
age <- c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37)
maxHeartRate <- c(202, 186, 187, 180, 156, 169, 174, 172, 153,
                  199, 193, 174, 198, 183, 178)
dat <- data.frame(cbind(age, maxHeartRate))
head(dat, 5)
```

```
##   age maxHeartRate
## 1  18          202
## 2  23          186
## 3  25          187
## 4  35          180
## 5  65          156
```

Loading a built-in R data

```
data("mtcars")
head(mtcars, 6)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
?mtcars
```

Exploratory Data Analysis

Load the dataset

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

Let's take a look at the data

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

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

Frequency Table

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

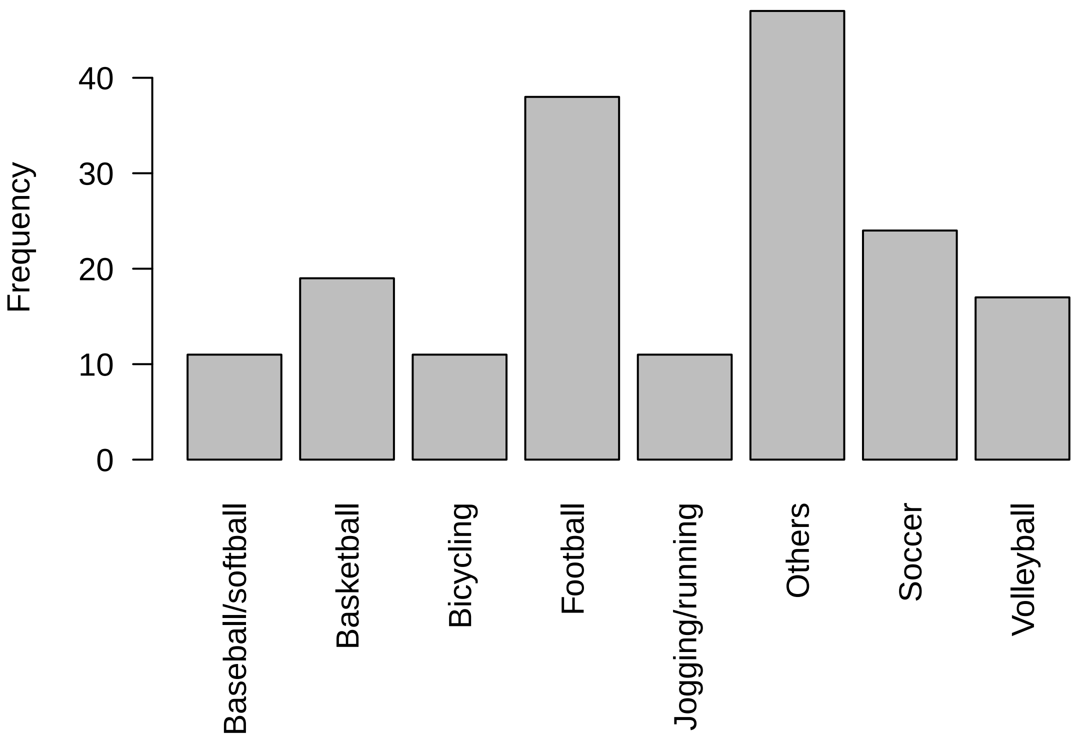
```
## sport
## Baseball/softball        Basketball         Bicycling          Football
##                11                19                11                38
##   Jogging/running            Others            Soccer        Volleyball
##                11                47                24      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

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

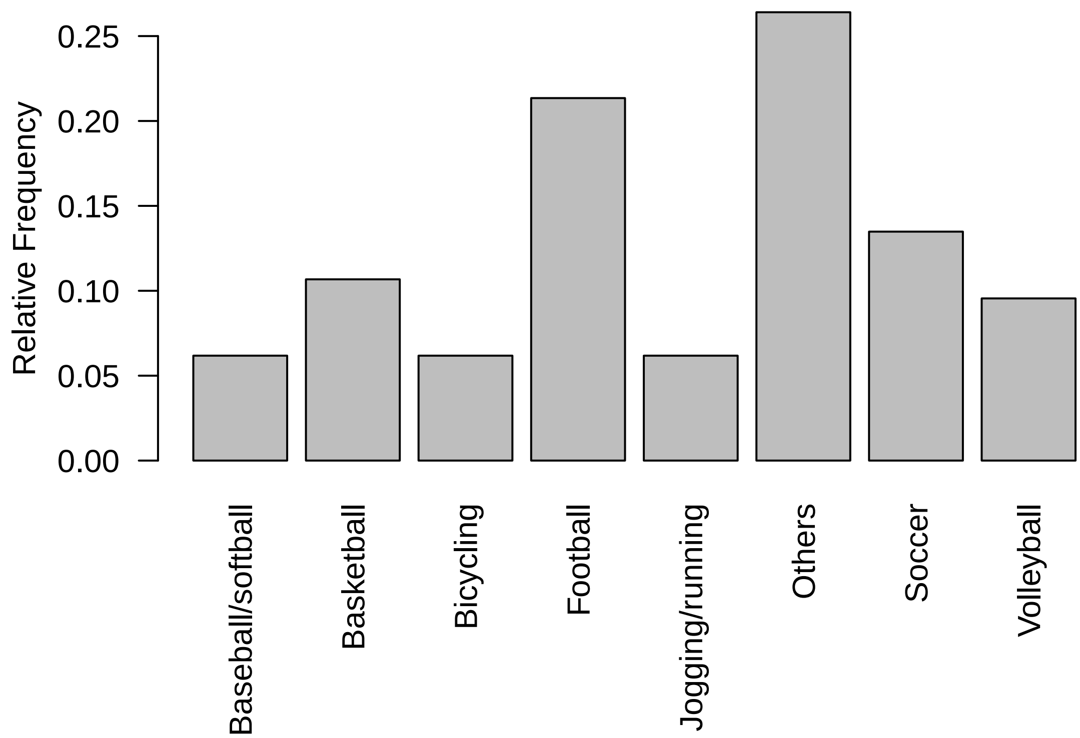
```
## sport
## Baseball/softball        Basketball         Bicycling          Football
##        0.06179775        0.10674157        0.06179775        0.21348315
##   Jogging/running            Others            Soccer        Volleyball
##        0.06179775        0.26404494        0.13483146      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

Bar Chart

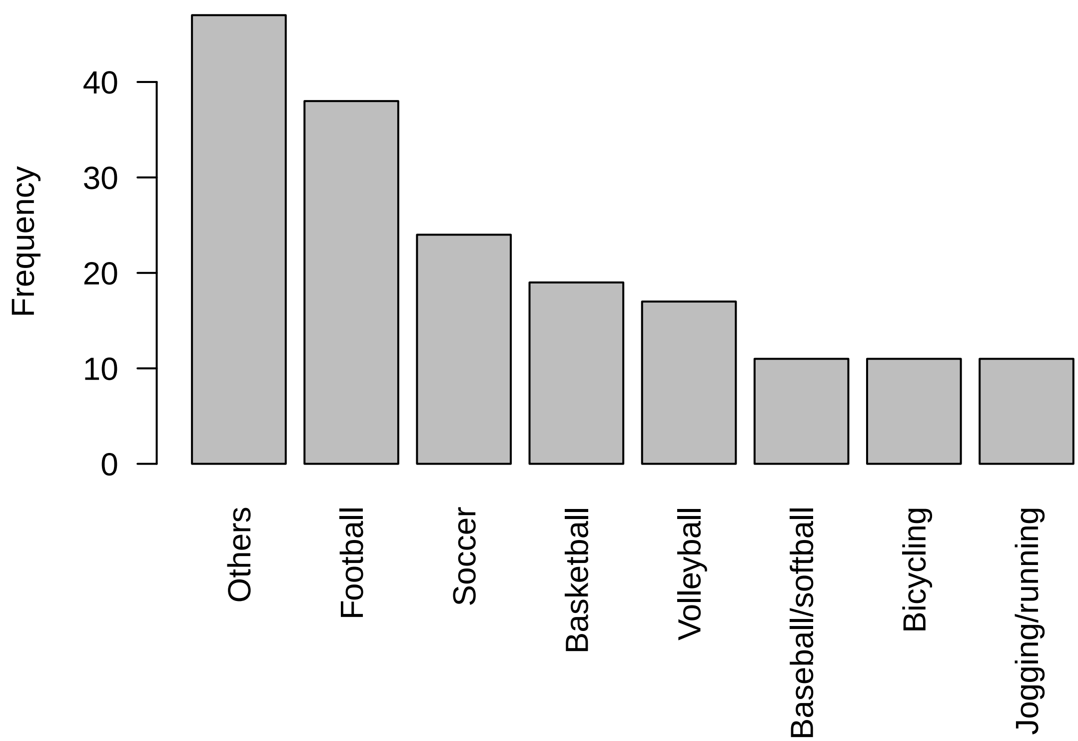
```r
# Bart chart for the frequency
par(las = 2, mar = c(7.1, 4.1, 1.1, 1.1))
barplot(tab1, ylab = "Frequency")
```




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

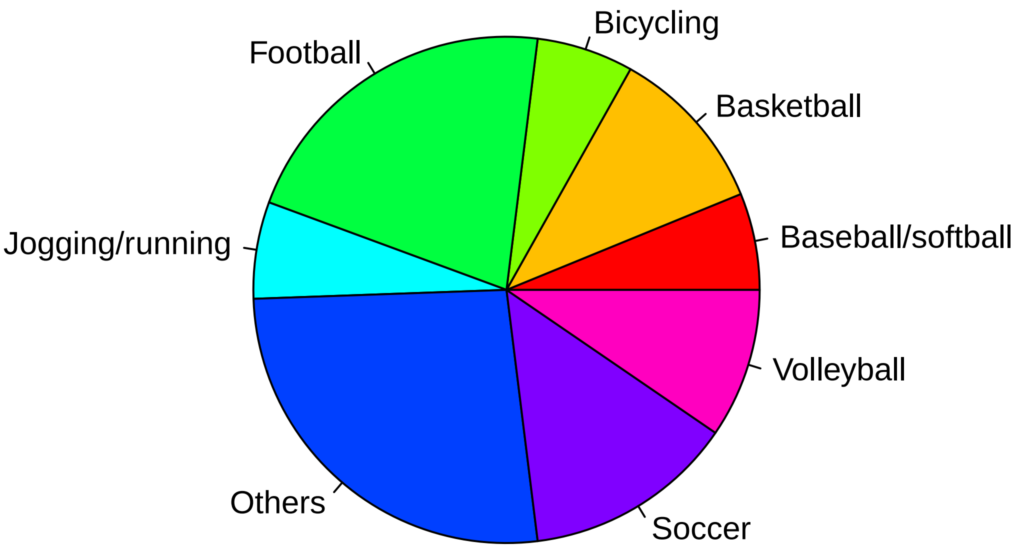



```r
# Pareto chart
par(las = 2, mar = c(7.1, 4.1, 1.1, 1.1))
barplot(sort(tab1, decreasing = T), ylab = "Frequency")
```

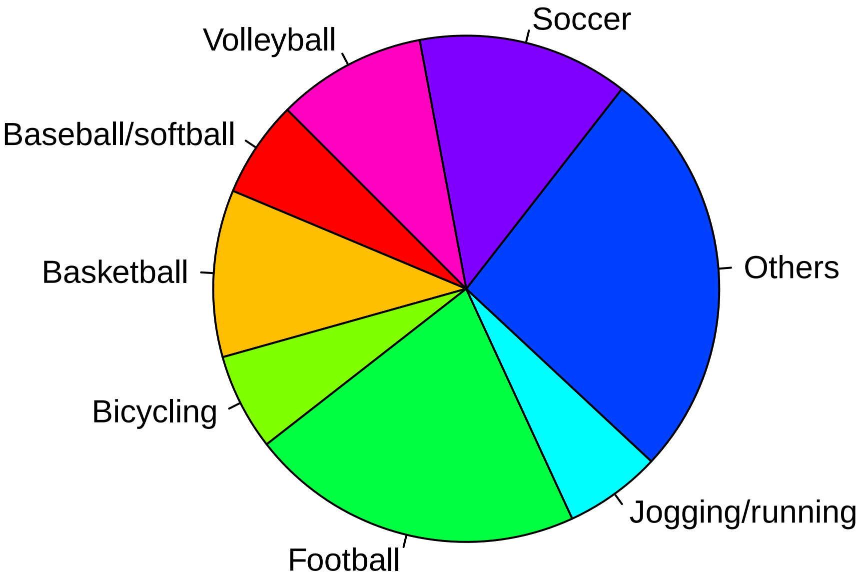


Pie Chart

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




```r
# rotate the pie
par(mar = c(1.1, 3.1, 1.1, 3.1))
pie(table(sport), col = rainbow(8), init.angle = 135)
```



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

```
dim(ORD)
```

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

```
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
```

```
summary(ORD)
```

```
##      month          carrier             origin            arr_delay
##  Min.   : 1.000   Length:12678       Length:12678       Min.   :  0.00
##  1st Qu.: 4.000   Class :character   Class :character   1st Qu.:  0.00
##  Median : 7.000   Mode  :character   Mode  :character   Median :  0.00
##  Mean   : 6.751                                         Mean   : 14.93
##  3rd Qu.:10.000                                         3rd Qu.: 10.00
##  Max.   :12.000                                         Max.   :448.00
##                                                         NA's   :443
```

2 way Frequency Table

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

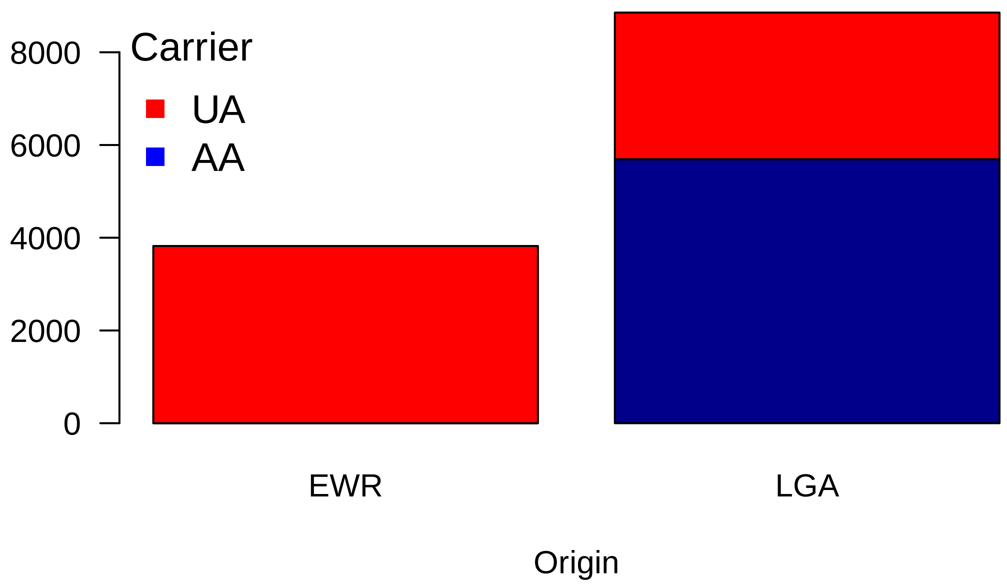
```
##        origin
## carrier  EWR  LGA
##      AA    0 5694
##      UA 3822 3162
```

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

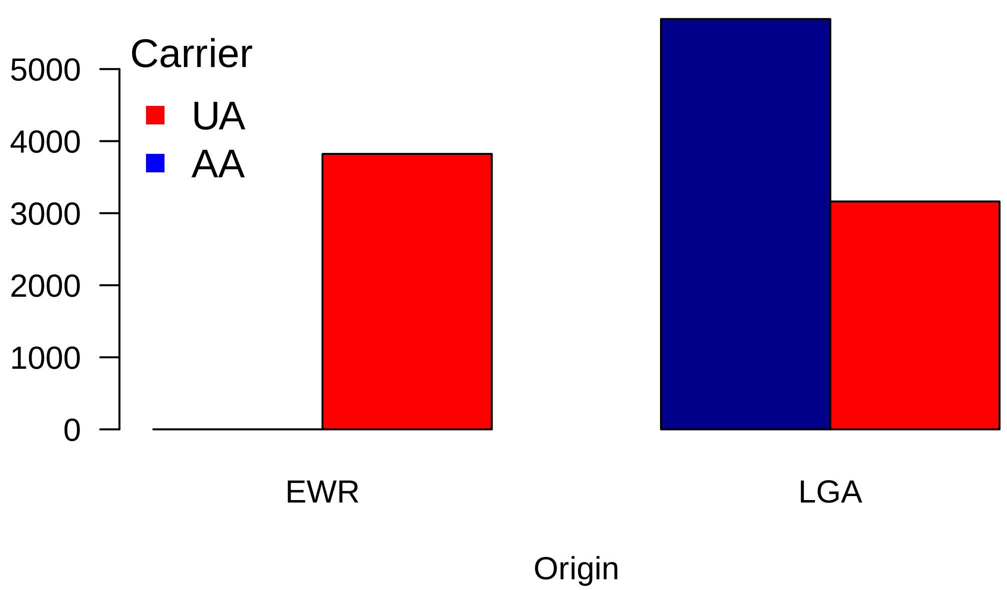
```
##        origin
## carrier       EWR       LGA
##      AA 0.0000000 0.4491245
##      UA 0.3014671 0.2494084
```

Stacked/dodged bar chart

```r
## 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")
```




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



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.

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

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

```r
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

```r
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
```

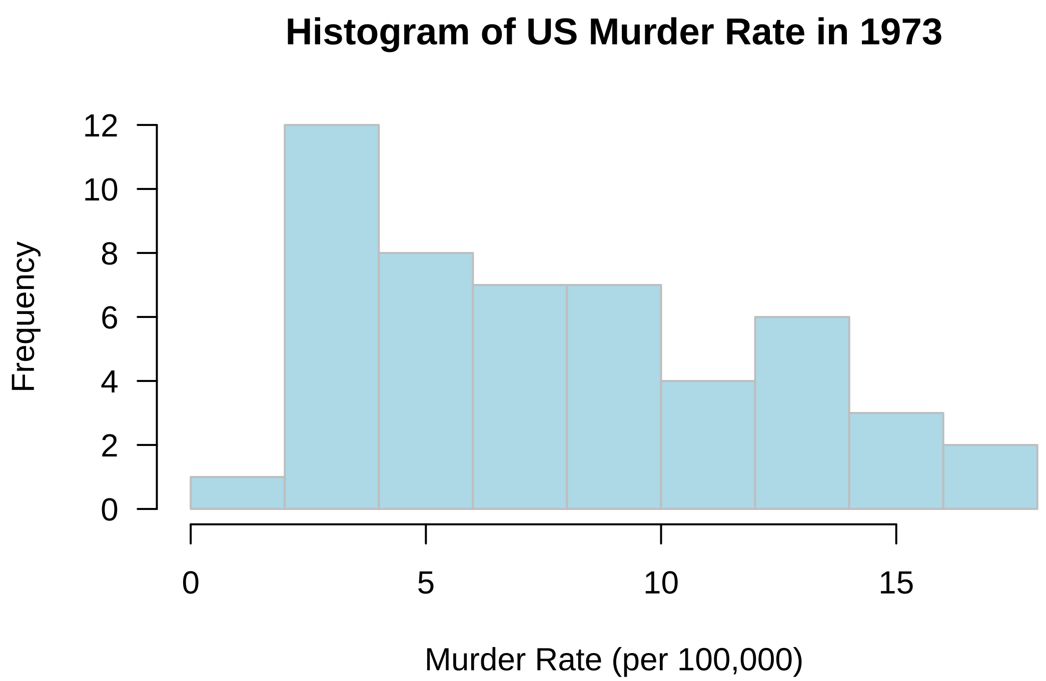
```r
stem(USArrests$Murder, scale = 2)
```

```
##
##   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
```

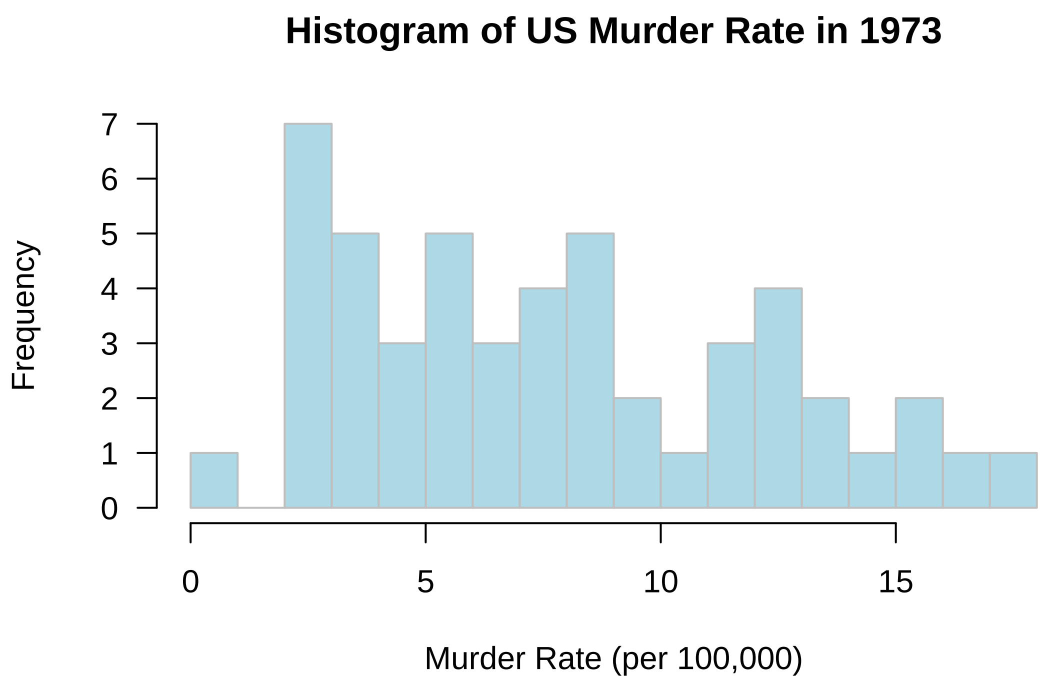
```
##    12 | 127
##    13 | 022
##    14 | 4
##    15 | 44
##    16 | 1
##    17 | 4
```

Histogram

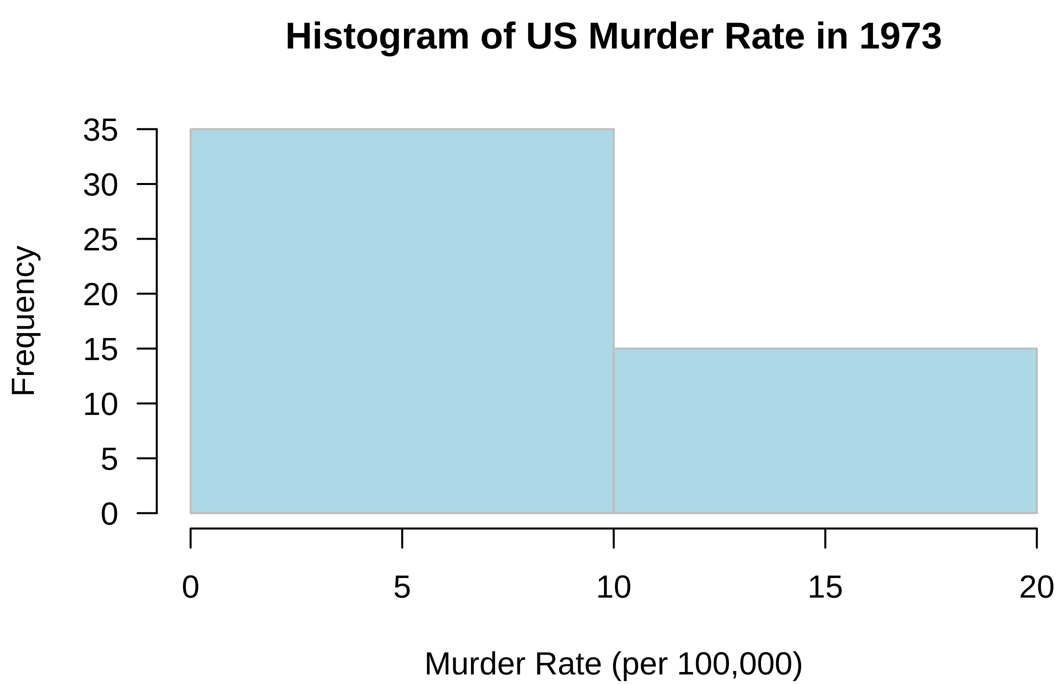
```
par(las = 1)
hist(USArrests$Murder, main = "Histogram of US Murder Rate in 1973",
     col = "lightblue", border = "gray", xlab = "Murder Rate (per 100,000)")
```




```
# 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)")
```

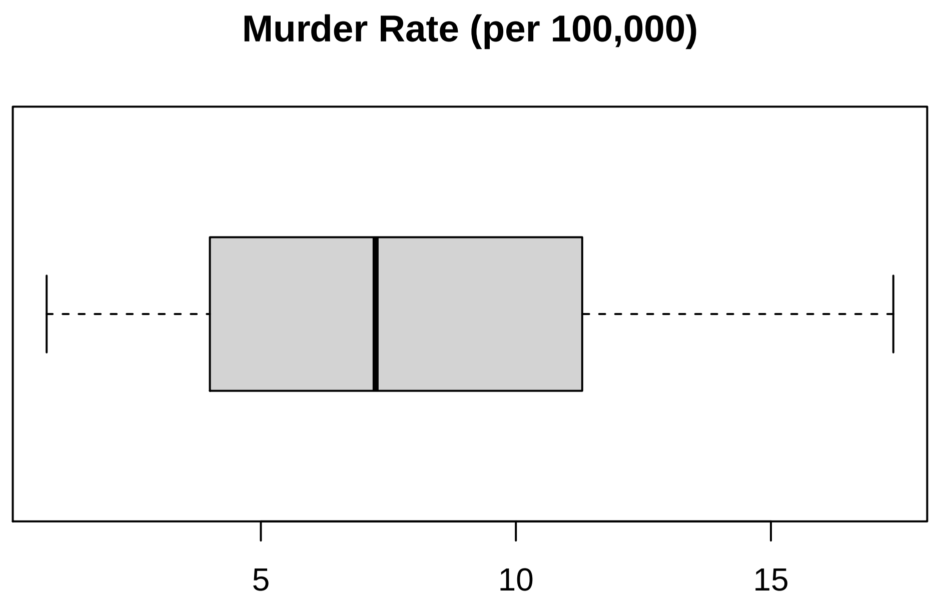



```
# Let's change the bin size again
par(las = 1)
hist(USArrests$Murder, nclass = 2,
     main = "Histogram of US Murder Rate in 1973", col = "lightblue",
     border = "gray", xlab = "Murder Rate (per 100,000)")
```

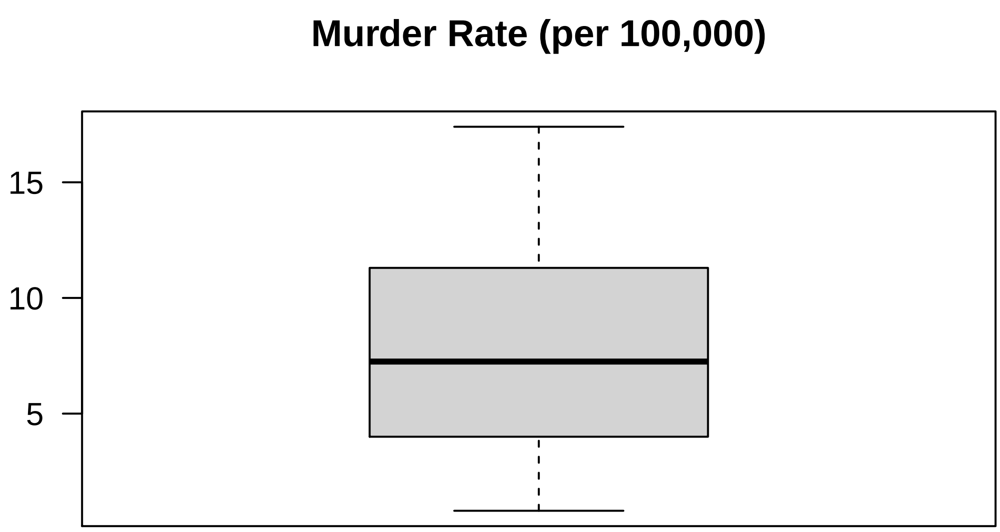


Boxplot

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




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



Numerical summary of central tendency and variability

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

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

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

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

```
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
```

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

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

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

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

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

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

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

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

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

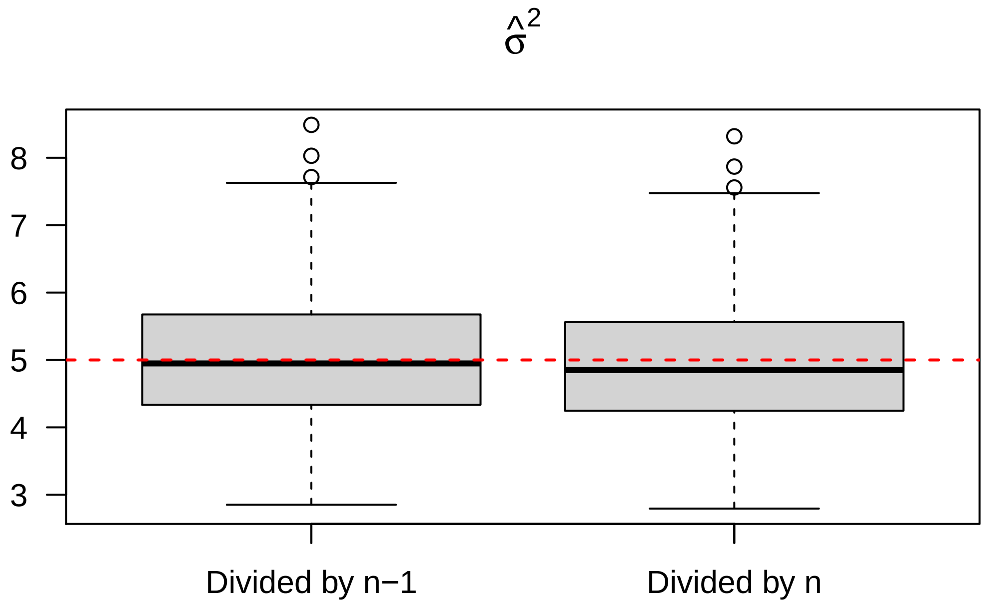
```
## [1] 16.6
```

Sample variance

$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$.

Why divided by $n-1$ instead of n? Let's conduct a simulation!

```
set.seed(123)
sim <- replicate(500, rnorm(50, mean = 10, sd = sqrt(5)))
## True "population" variance is 5
varEst <- apply(sim, 2, var)
varEst1 <- apply(sim, 2, function(x) var(x) * (49 / 50))
boxplot(varEst, varEst1, las = 1, main = expression(hat(sigma)^2))
axis(1, at = 1:2, labels = c("Divided by n-1", "Divided by n"))
abline(h = 5, lty = 2, col = "red", lwd = 1.5)
```



Interquartile range (IQR)

```
data1 <- c(13, 18, 13, 14, 13, 16, 14, 21, 13)
IQR(data1, type = 1)
```

```
## [1] 3
```

```
data2 <- c(13, 18, 13, 14, 13, 16, 14, 210, 13)
IQR(data2, type = 1)
```

```
## [1] 3
```

Percentiles

```
#Q1
quantile(data1, 0.25, type = 1)
```

```
## 25%
##  13
```

```
#Q2 aka median
quantile(data1, 0.5, type = 1)
```

```
## 50%
##  14
```

```r
#Q3
quantile(data1, 0.75, type = 1)
```

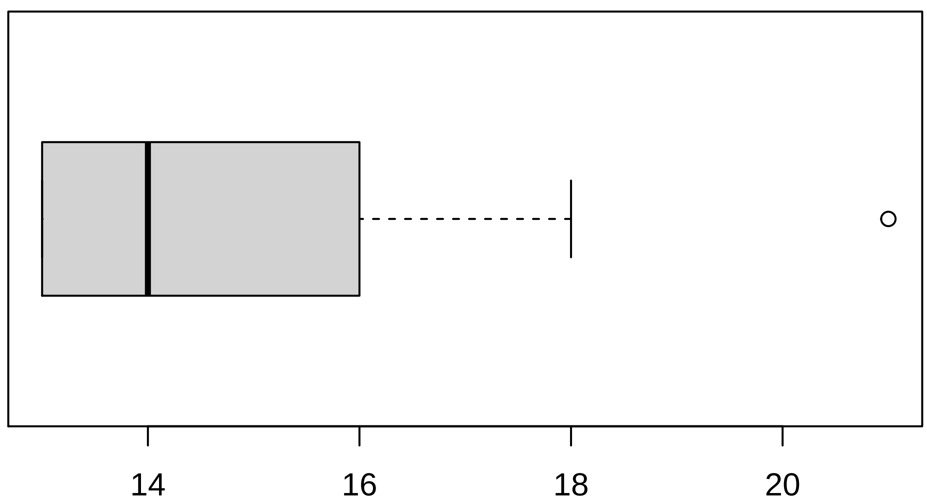
```
## 75%
##  16
```

```r
data3 <- c(13, 18, 13, 14, 13, 16, 14, 21, 13, 9,
           27, 18, 25, 20, 6)
quantile(data3, c(0.35, 0.65), type = 1)
```

```
## 35% 65%
##  13  18
```

Boxplot

```r
boxplot(data1, horizontal = T)
```

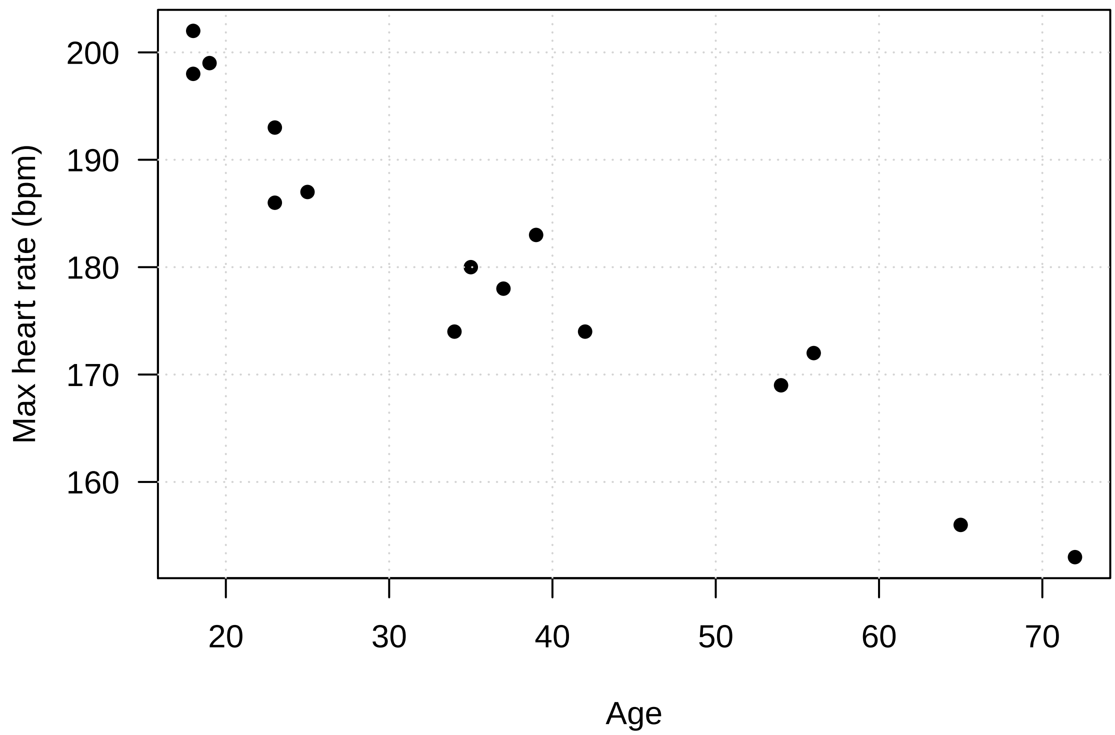


Qualitative vs Quantitative: Side by Side Boxplots

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

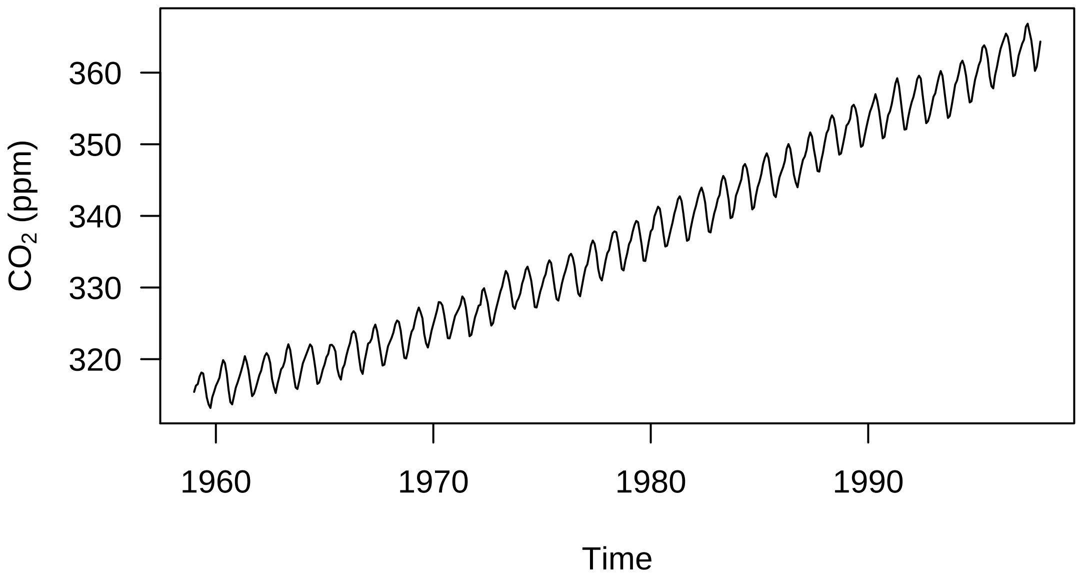
```
url <- "https://whitneyhuang83.github.io/STAT8010/Data/maxHeartRate.csv"
dat <- read.csv(url, header = TRUE)

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()
```



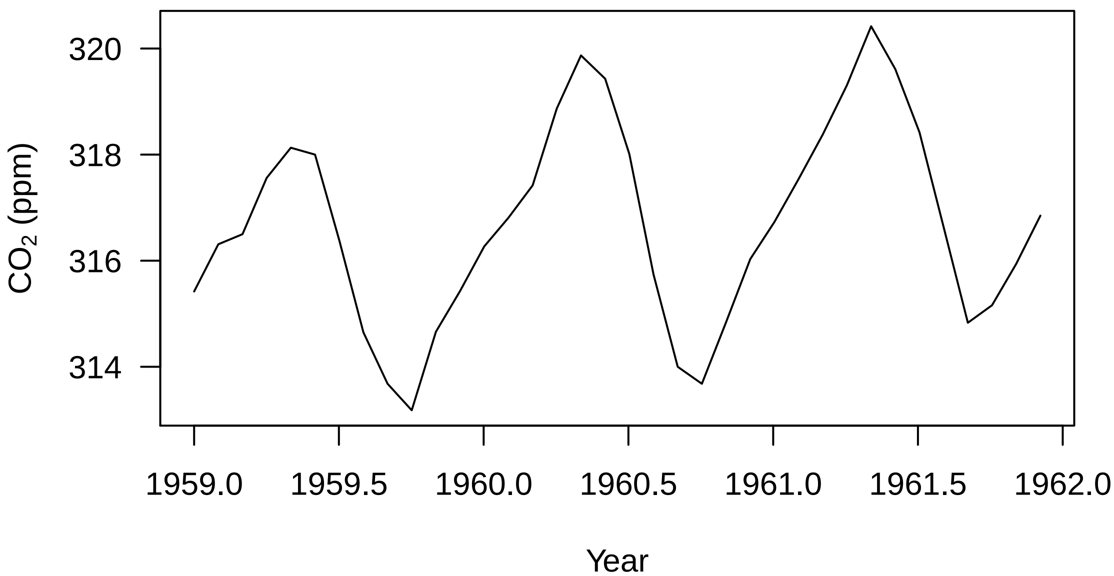
Visualizing Time Series Data: Mauna Loa Atmospheric CO2 Concentration

Atmospheric concentrations of CO_2 are expressed in parts per million (ppm) and reported in the preliminary 1997 SIO manometric mole fraction scale.

```
data("co2")
par(las = 1)
ts.plot(co2, ylab = expression(paste(CO[2], " (ppm)")))
```




```
time <- seq(1959, 1998, len = 468)
plot(time[1:36], co2[1:36], type = "l", xlab = "Year",
     ylab = expression(paste(CO[2], " (ppm)")),
     las = 1)
```



Visualizing Cross-Sectional Data

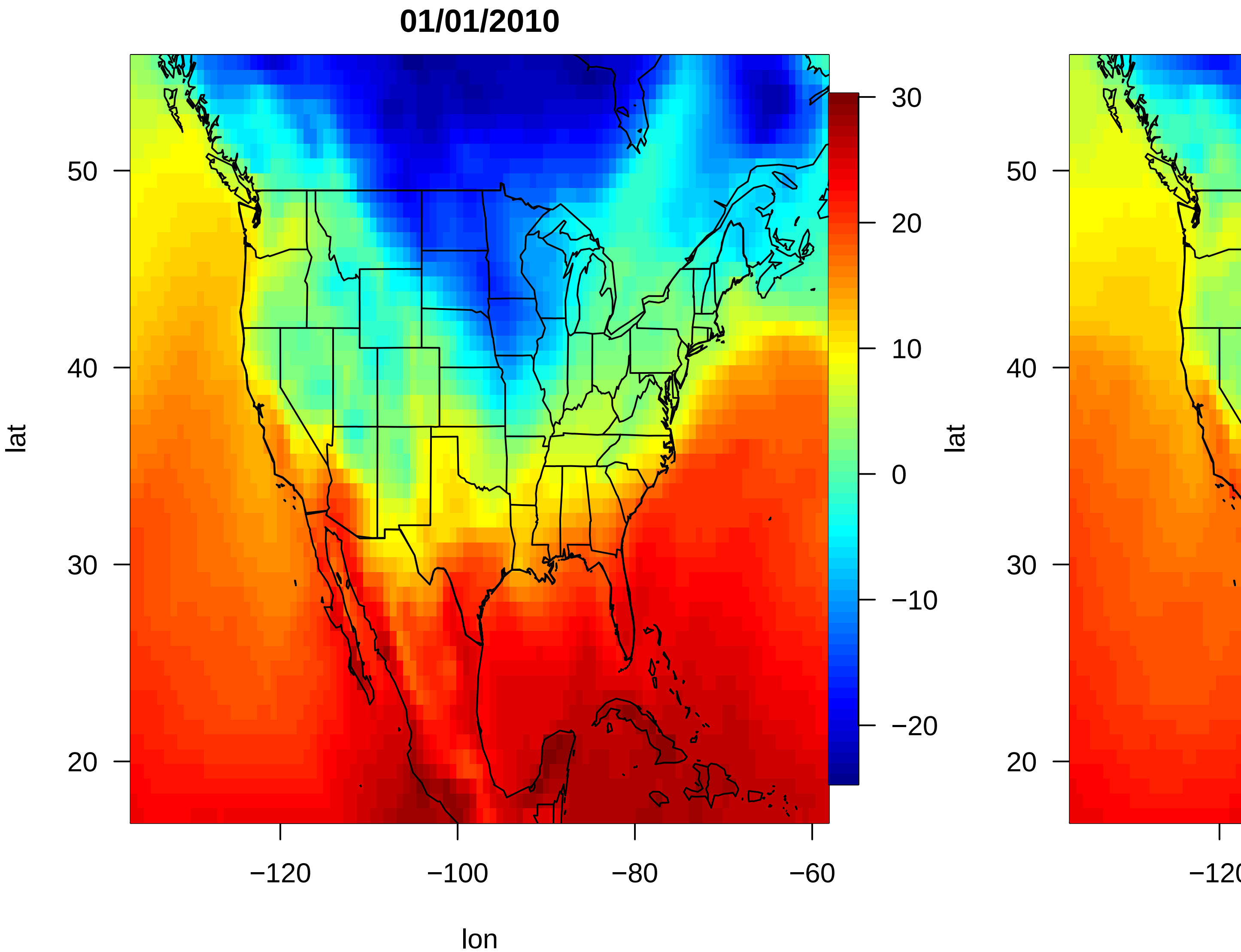
```
library(maps)
library(ggmap)
data("USArrests")
USArrests$region <- tolower(row.names(USArrests))
statesMap <- map_data("state")
str(statesMap)
murderMap <- merge(statesMap, USArrests, by = "region")
```

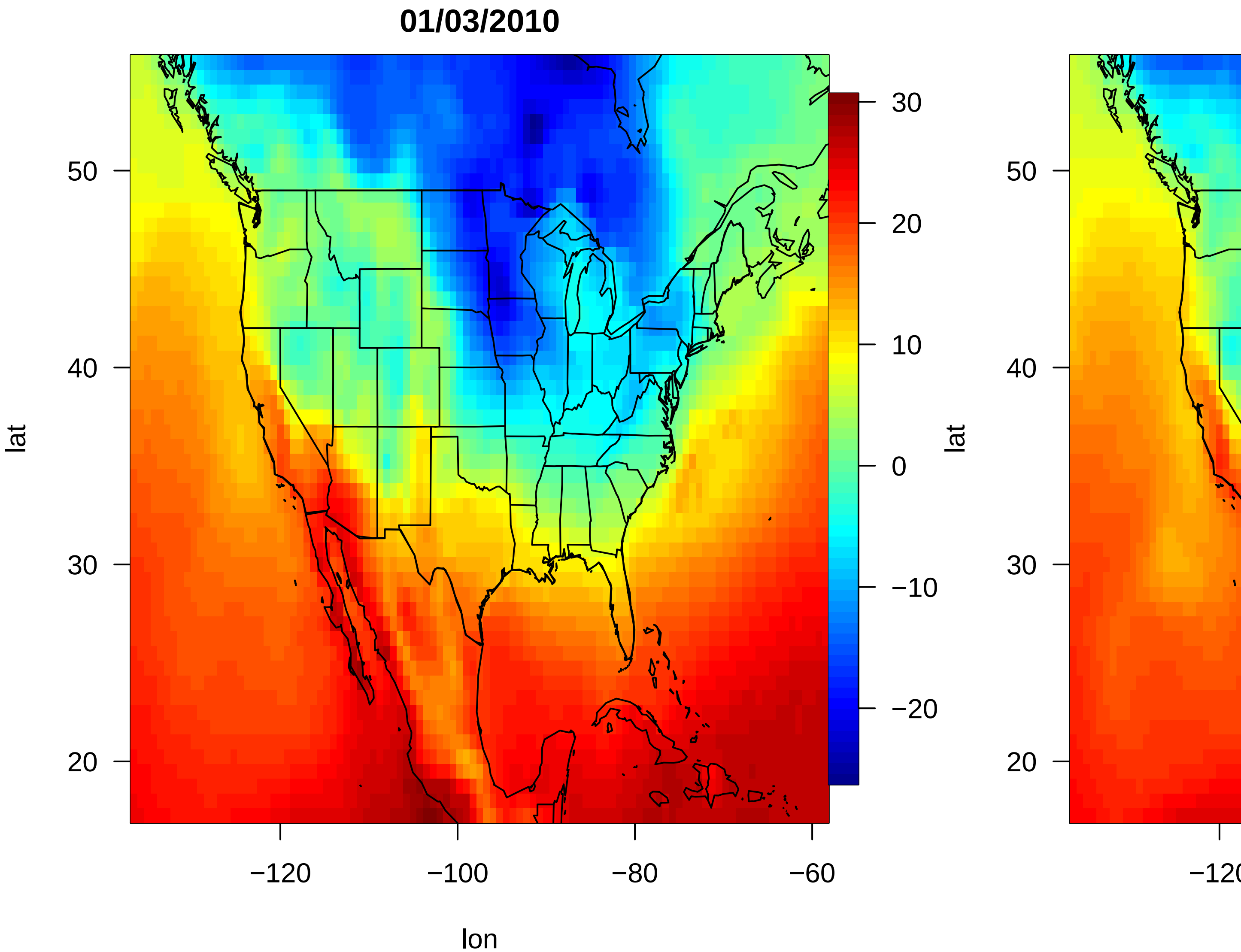
```
str(murderMap)

p1 <- ggplot(murderMap, aes(x = long, y = lat, group = group, fill = Murder))
p2 <- geom_polygon(color = "black")
p3 <- scale_fill_gradient(low = "lightblue",high = "red", guide = "legend")
p1 + p2 + p3
```

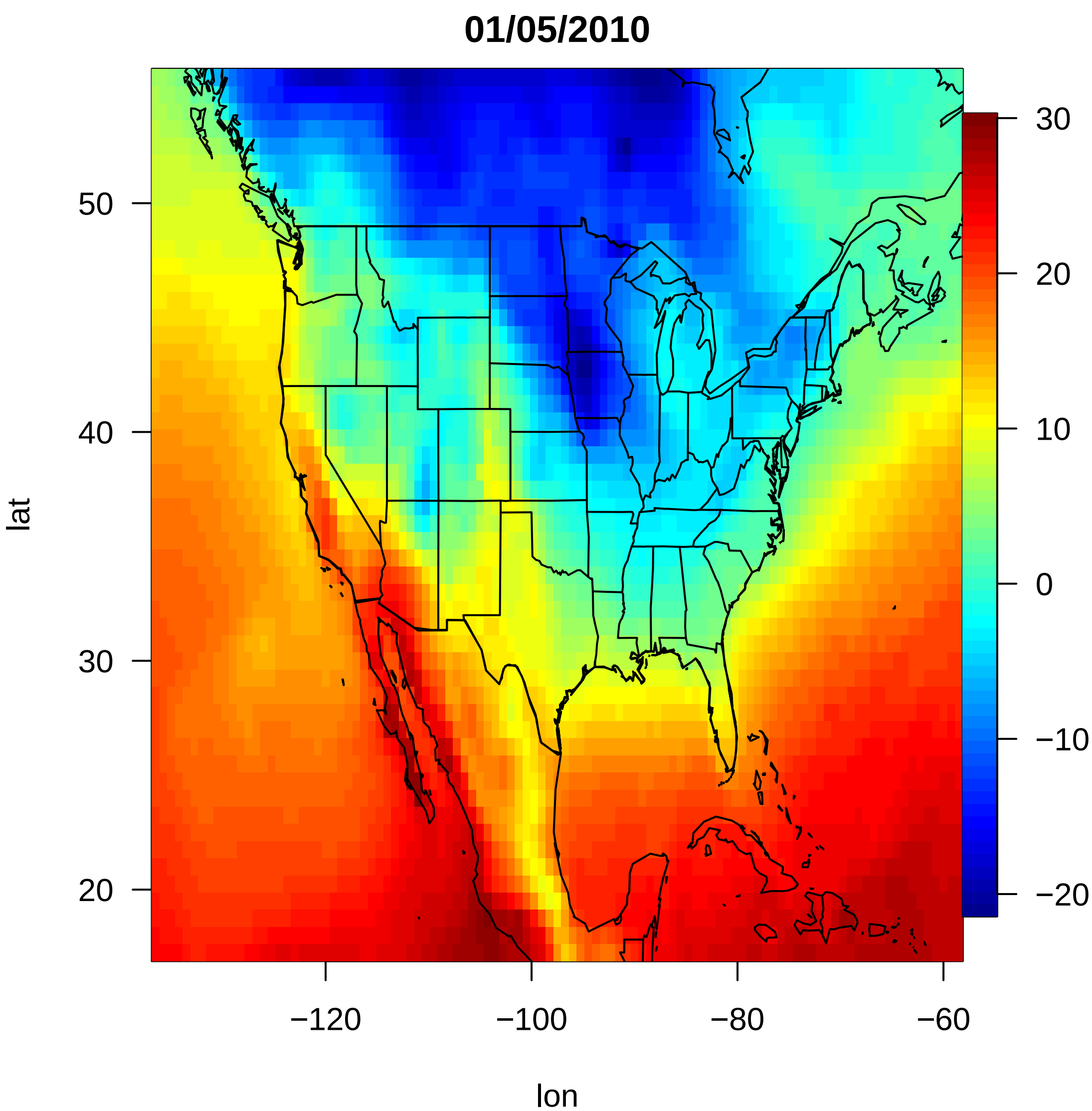
Visualizing Spatio-Temporal Data: ERA-Interim

The ERA-Interim is a global atmospheric reanalysis dataset. Reanalysis is an approach to produce spatially and temporally gridded datasets via data assimilation for climate monitoring and analysis.

```
library(maps)
load("ERA_tmx_2010_JanFeb.RData")
library(fields)
par(mar = c(4.6, 4.1, 2.1, 0))
for (i in seq(1:5)){
        image.plot(lon, lat, tmx_dat[,, i], las = 1, main = format(day[i], "%m/%d/%Y"))
        map("state", xlim = range(lon), ylim = range(lat),
            add = T)
        map("world", xlim = range(lon), ylim = range(lat),
            add = T)
}
```


01/01/2010
lat
50
40
30
20
−120
−100
−80
−60
lon
30
20
10
0
−10
−20
lat
50
40
30
20
−120


01/03/2010
lat
lon
50
40
30
20
−120
−100
−80
−60
30
20
10
0
−10
−20
lat
50
40
30
20
−120




```
library(animation)
saveLatex({
  for (i in 1:58){
          image.plot(lon, lat, tmx_dat[,, i], las = 1, main = format(day[i], "%m/%d/%Y"),
                     zlim = range(tmx_dat))
          map("state", xlim = range(lon), ylim = range(lat), add = T)
          map("world", xlim = range(lon), ylim = range(lat), add = T)
  }
}, img.name = "ERA_Tmax", ani.opts = "controls,width=0.975\\textwidth",
latex.filename = ifelse(interactive(), "ERA_TMX_JanFeb.tex", ""),
interval = 0.5, nmax = 58, ani.dev = "pdf", ani.type = "pdf", ani.width = 8,
ani.height = 6, documentclass = paste("\\documentclass{article}",
                                     "\\usepackage[papersize={8in,6in},margin=0.1in]{geometry}",
                                     sep = "\n"))
```