Descriptive Statistics: the Rainfall edition

M P Gururajan, Hina A Gokhale and Dayadeep Monder

Indian Institute of Technology Bombay, Mumbai

In this session, we are going to use another real world dataset. The dataset is called `RainfallData.csv` and is stored in `csv` format. The data lists the rain fall over different subdivisions of India for the period 1901 to 2016 – monthwise in every year. Of course, like many real word data, sometimes, for some months, the data is not available and is marked as `NA` in the data. We have to find a way of dealing with that!

As usual, it is a good idea to take a look at the data before processing them in R. As compared to the marital status data, this data is much more amenable for processing in R. For example, the formatting is nice; the first line is the header and the rest are data. So, reading the data is easy!

Let us first load the data.

```
unlink("~/.RData")
X <- read.csv("../Data/RainfallData.csv", header=TRUE,
              na.strings = TRUE)
str(X)
```

```
## 'data.frame': 4187 obs. of  19 variables:
##  $ SUBDIVISION: chr  "Andaman & Nicobar Islands" "Andaman & Nicobar Islands" "Andaman & Nicobar Islar
##  $ YEAR       : int  1901 1902 1903 1904 1905 1906 1907 1908 1910 1911 ...
##  $ JAN        : chr  "49.2" "0" "12.7" "9.4" ...
##  $ FEB        : chr  "87.1" "159.8" "144" "14.7" ...
##  $ MAR        : chr  "29.2" "12.2" "0" "0" ...
##  $ APR        : chr  "2.3" "0" "1" "202.4" ...
##  $ MAY        : chr  "528.8" "446.1" "235.1" "304.5" ...
##  $ JUN        : chr  "517.5" "537.1" "479.9" "495.1" ...
##  $ JUL        : chr  "365.1" "228.9" "728.4" "502" ...
##  $ AUG        : chr  "481.1" "753.7" "326.7" "160.1" ...
##  $ SEP        : chr  "332.6" "666.2" "339" "820.4" ...
##  $ OCT        : chr  "388.5" "197.2" "181.2" "222.2" ...
##  $ NOV        : chr  "558.2" "359" "284.4" "308.7" ...
##  $ DEC        : chr  "33.6" "160.5" "225" "40.1" ...
##  $ ANNUAL     : chr  "3373.2" "3520.7" "2957.4" "3079.6" ...
##  $ JF         : chr  "136.3" "159.8" "156.7" "24.1" ...
##  $ MAM        : chr  "560.3" "458.3" "236.1" "506.9" ...
##  $ JJAS       : chr  "1696.3" "2185.9" "1874" "1977.6" ...
##  $ OND        : chr  "980.3" "716.7" "690.6" "571" ...
```

Of course, once we read the data, we realise that all the data is stored as character type – including those which are numbers - except for the year! We need to turn all the columns into numbers for easy processing! We do that by identifying the columns, and turning those columns into numerical values.

```
Columns <- c("JAN","FEB","MAR","APR","MAY","JUN",
             "JUL","AUG","SEP","OCT","NOV","DEC",
             "ANNUAL","JF","MAM","JJAS","OND")
X[Columns] <- sapply(X[Columns],as.numeric)
```

```
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
```

```
str(X)
```

```
## 'data.frame':    4187 obs. of  19 variables:
##  $ SUBDIVISION: chr  "Andaman & Nicobar Islands" "Andaman & Nicobar Islands" "Andaman & Nicobar Islan
##  $ YEAR       : int  1901 1902 1903 1904 1905 1906 1907 1908 1910 1911 ...
##  $ JAN        : num  49.2 0 12.7 9.4 1.3 ...
##  $ FEB        : num  87.1 159.8 144 14.7 0 ...
##  $ MAR        : num  29.2 12.2 0 0 3.3 ...
##  $ APR        : num  2.3 0 1 202.4 26.9 ...
##  $ MAY        : num  529 446 235 304 280 ...
##  $ JUN        : num  518 537 480 495 629 ...
##  $ JUL        : num  365 229 728 502 369 ...
##  $ AUG        : num  481 754 327 160 330 ...
##  $ SEP        : num  333 666 339 820 297 ...
##  $ OCT        : num  388 197 181 222 261 ...
##  $ NOV        : num  558.2 359 284.4 308.7 25.4 ...
##  $ DEC        : num  33.6 160.5 225 40.1 344.7 ...
##  $ ANNUAL     : num  3373 3521 2957 3080 2567 ...
##  $ JF         : num  136.3 159.8 156.7 24.1 1.3 ...
##  $ MAM        : num  560 458 236 507 310 ...
##  $ JJAS       : num  1696 2186 1874 1978 1625 ...
##  $ OND        : num  980 717 691 571 631 ...
```

Exploration

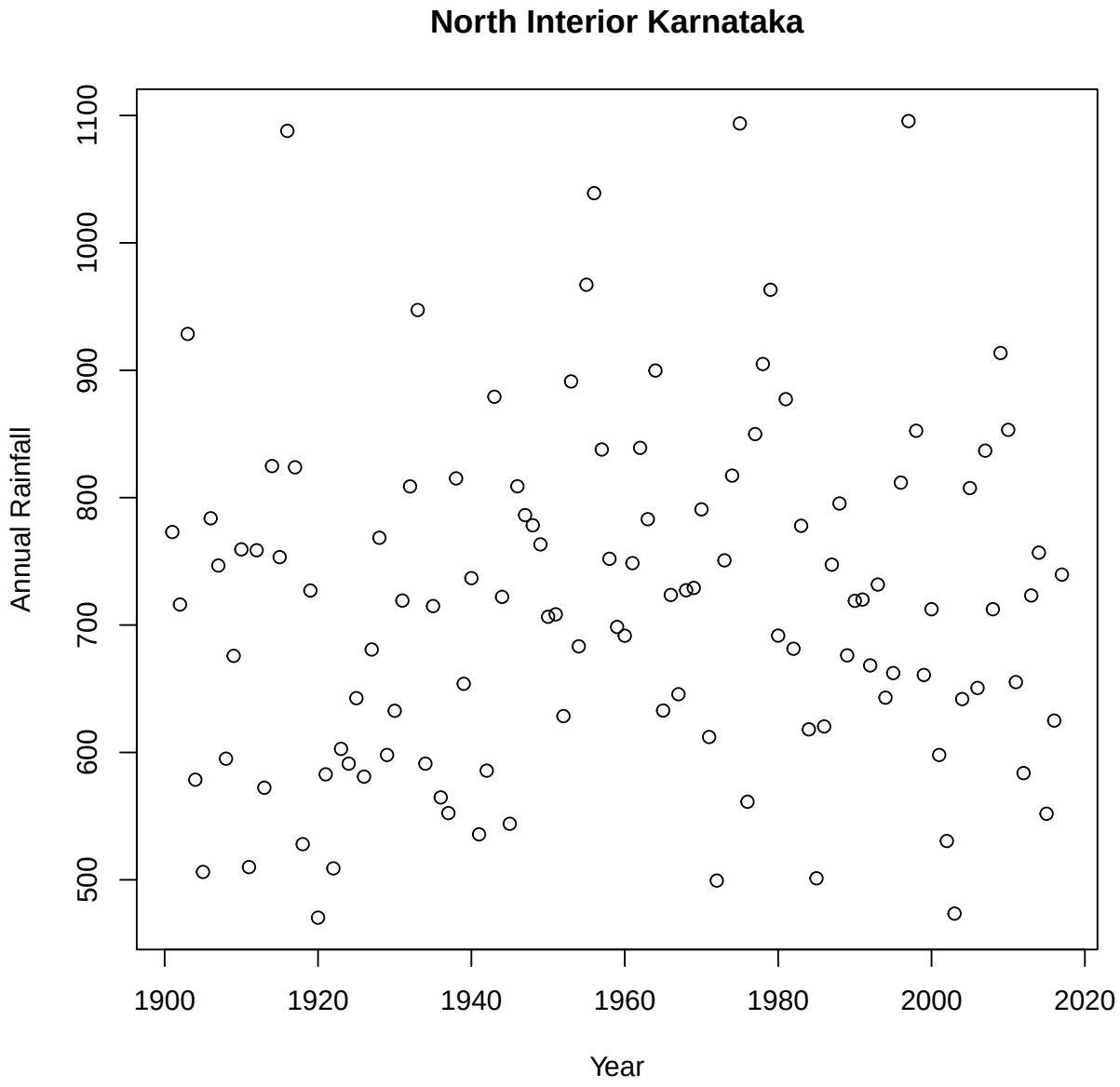
1. Is there any other way for you to turn these characters into numbers? Check!!

Now that we have the data, let us start the descriptive statistics!

1 The graphical measures!

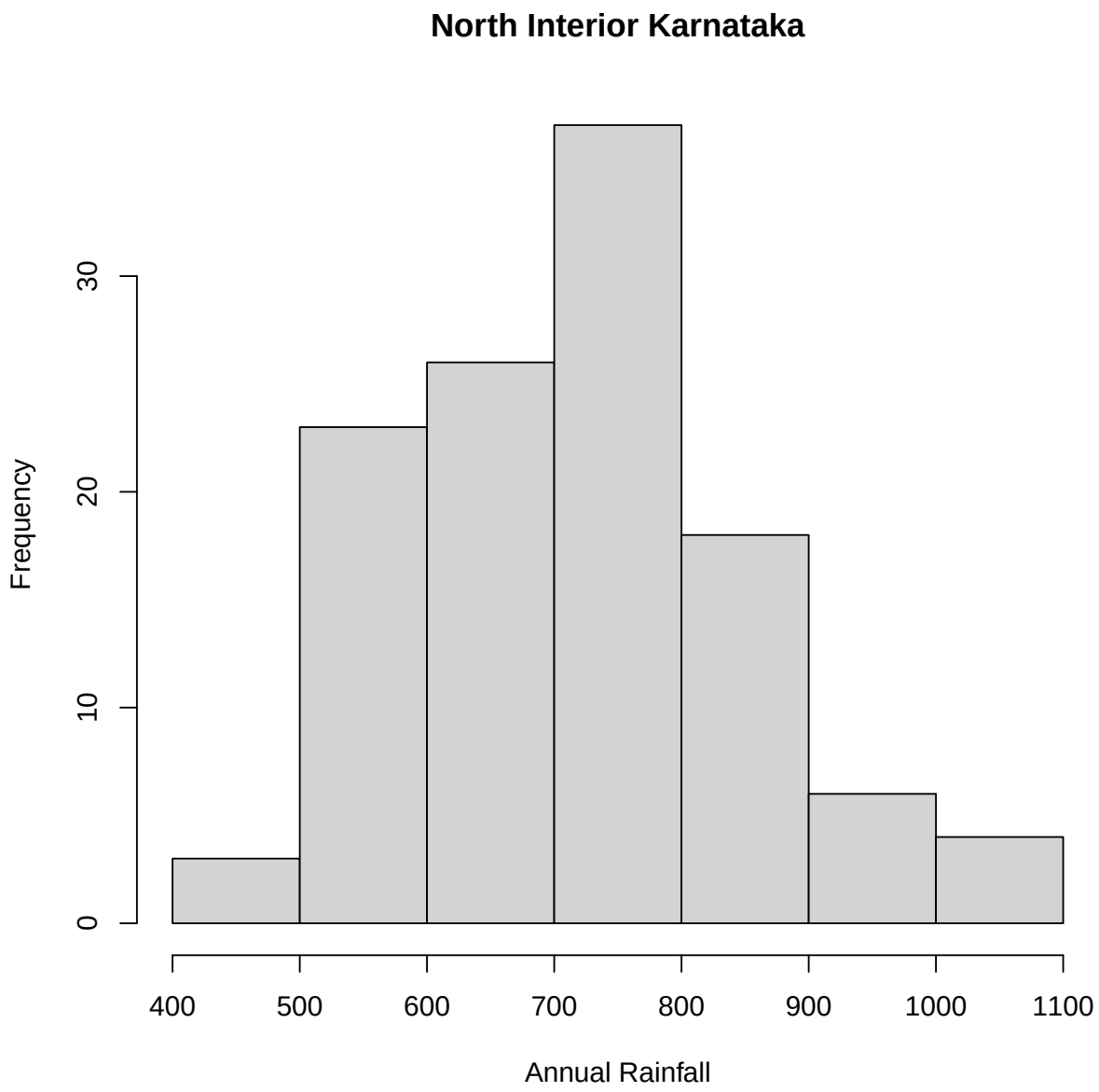
The first and easiest thing to do is to plot the data. For example, let us plot the annual rainfall over north interior Karnataka over the hundred odd years!

```
NIK <- subset(X,SUBDIVISION=="North Interior Karnataka")
plot(NIK$YEAR,NIK$ANNUAL,xlab="Year",ylab="Annual Rainfall",main="North Interior Karnataka")
```



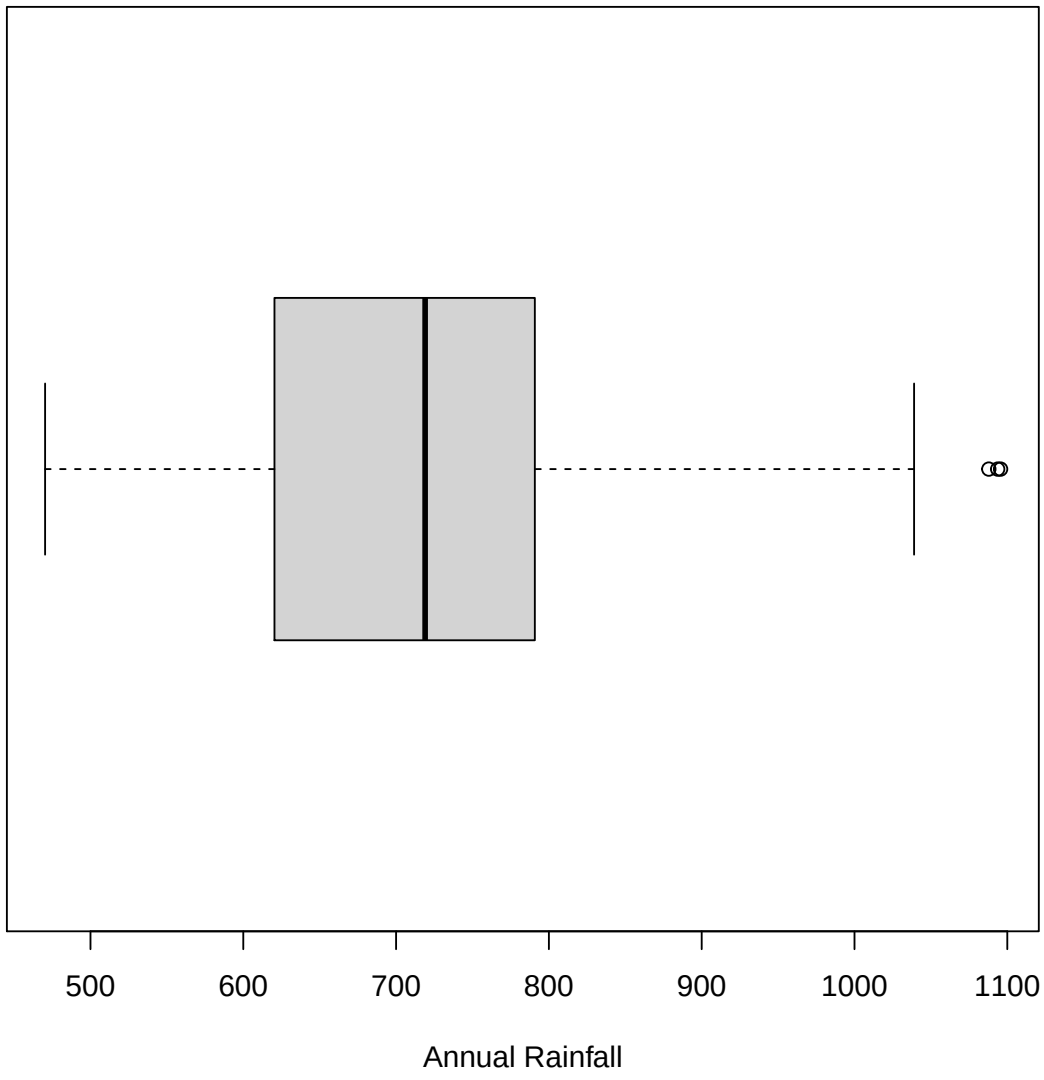
Of course, one can plot histograms!

```
hist(NIK$ANNUAL,main="North Interior Karnataka",xlab="Annual Rainfall")
```



One can plot the box plot too!

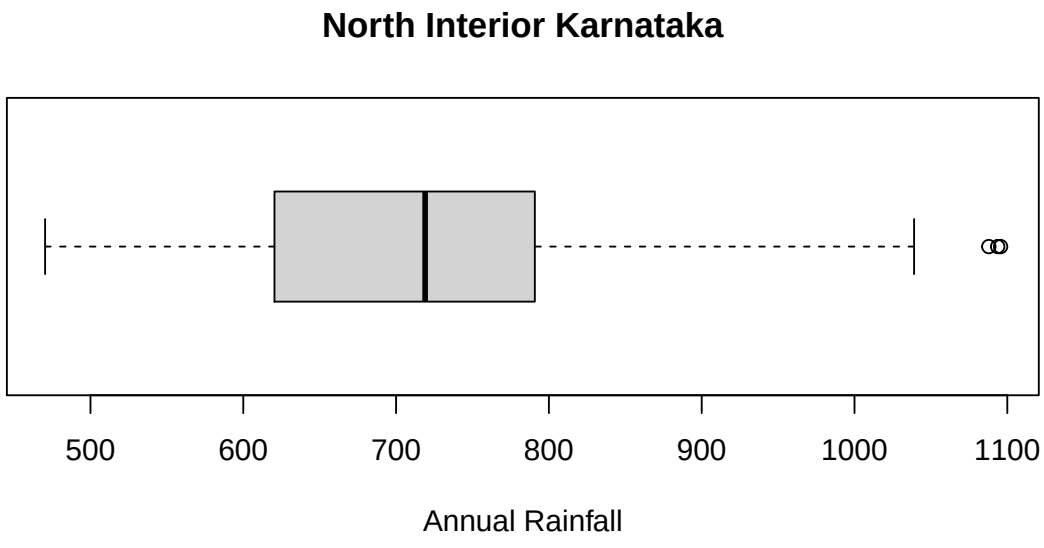
```
boxplot(NIK$ANNUAL,main="North Interior Karnataka",xlab="Annual Rainfall",horizontal = TRUE)
```

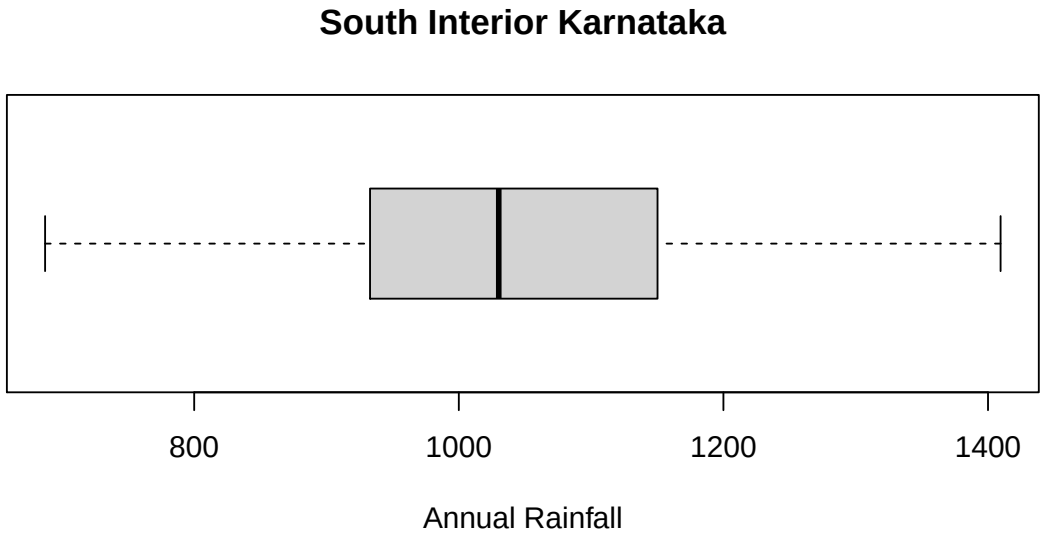


The dark line is the median – which is about 720 mm. The box represents the second and third quartile. The whisker represents 1.5 times the standard deviation. The data about 1100 mm are the outliers.

Let us compare this with South Interior Karnataka!

```
SIK <- subset(X,SUBDIVISION=="South Interior Karnataka")
par(mfrow=c(2,1))
boxplot(NIK$ANNUAL,main="North Interior Karnataka",xlab="Annual Rainfall",horizontal = TRUE)
boxplot(SIK$ANNUAL,main="South Interior Karnataka",xlab="Annual Rainfall",horizontal = TRUE)
```





Well! If you pay attention to the x-axis tick marks, the outliers of North Interior Karnataka are about the median of South Interior Karnataka!!

Let us do a stem plot!

```
stem(NIK$ANNUAL)
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##    4 | 77
##    5 | 001113344
##    5 | 556678888999
##    6 | 00001223333444
##    6 | 55566678888899
```

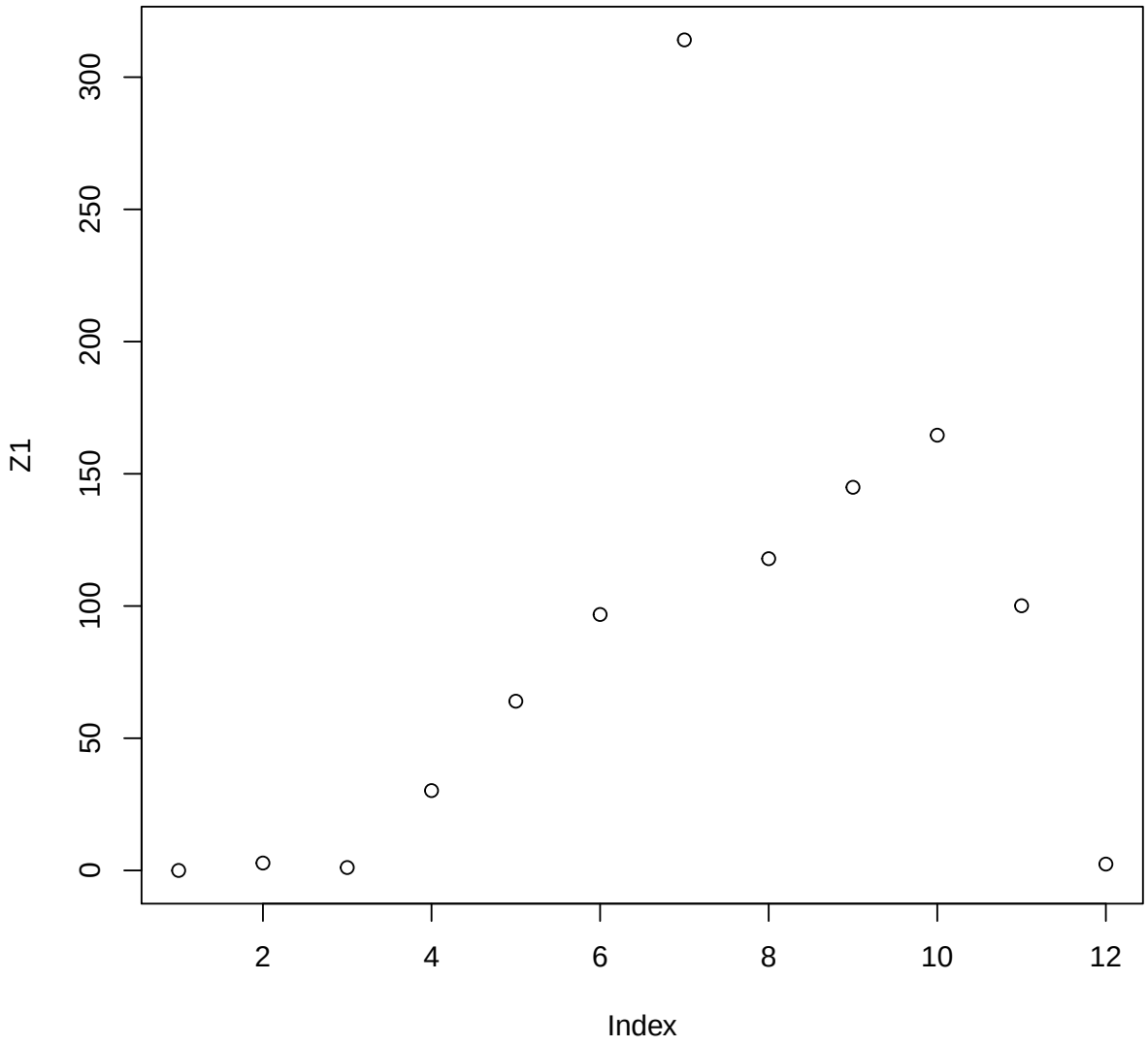
```
##    7 | 0111112222222333344
##    7 | 555555666677888899
##    8 | 011112222444
##    8 | 555889
##    9 | 0013
##    9 | 567
##   10 | 4
##   10 | 99
##   11 | 0
```

Which are the years that give above 1000 mm rain fall in North Interior Karnataka? Let us separate that data and plot these years!

```
Y <- subset(NIK, ANNUAL > 1000)
View(Y)
```

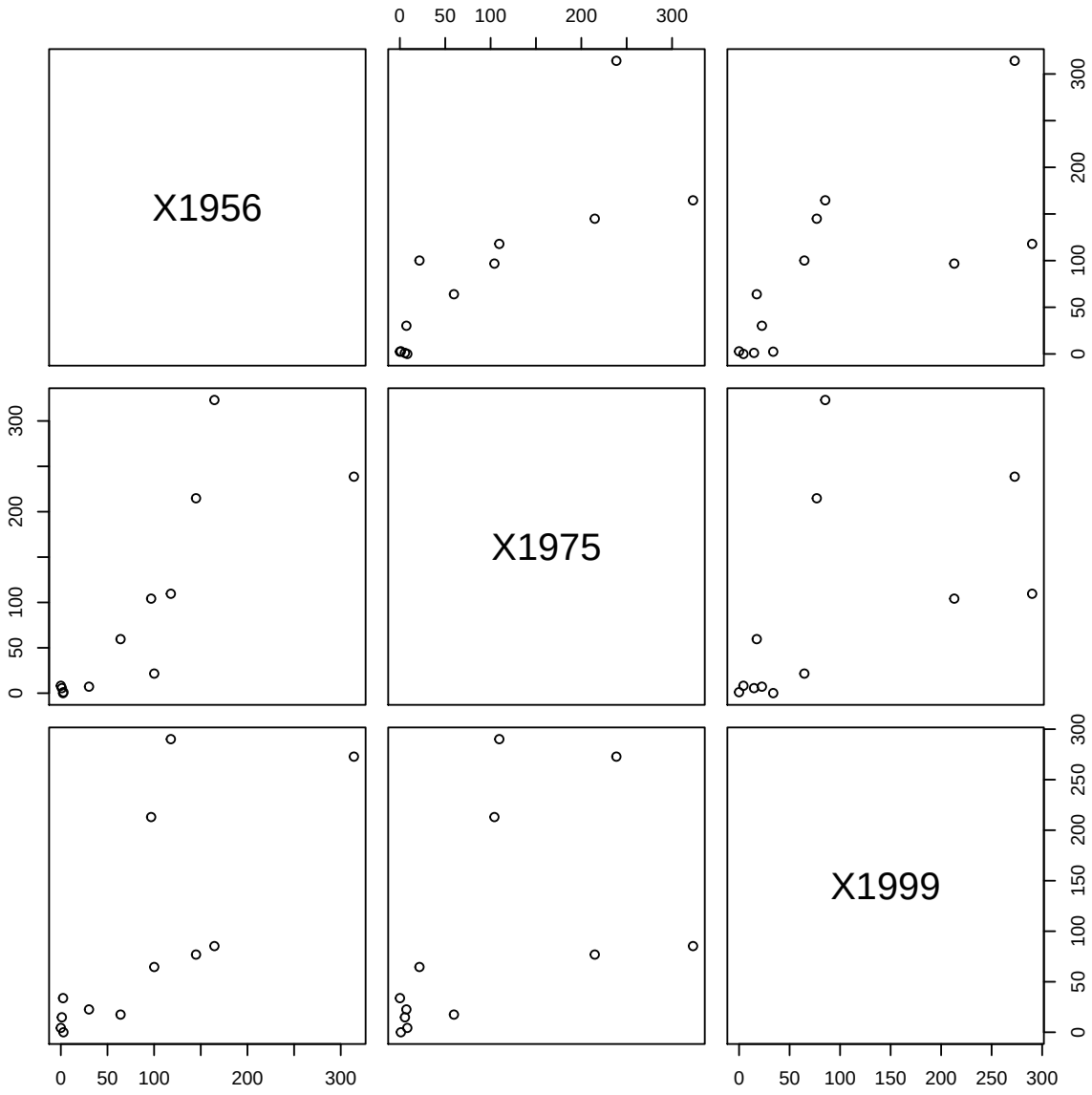
Let us pick one of the years, say, 1956, and plot the month-wise rainfall!

```
Z1 <- as.numeric(X[3777,Columns[1:12]])
plot(Z1)
```



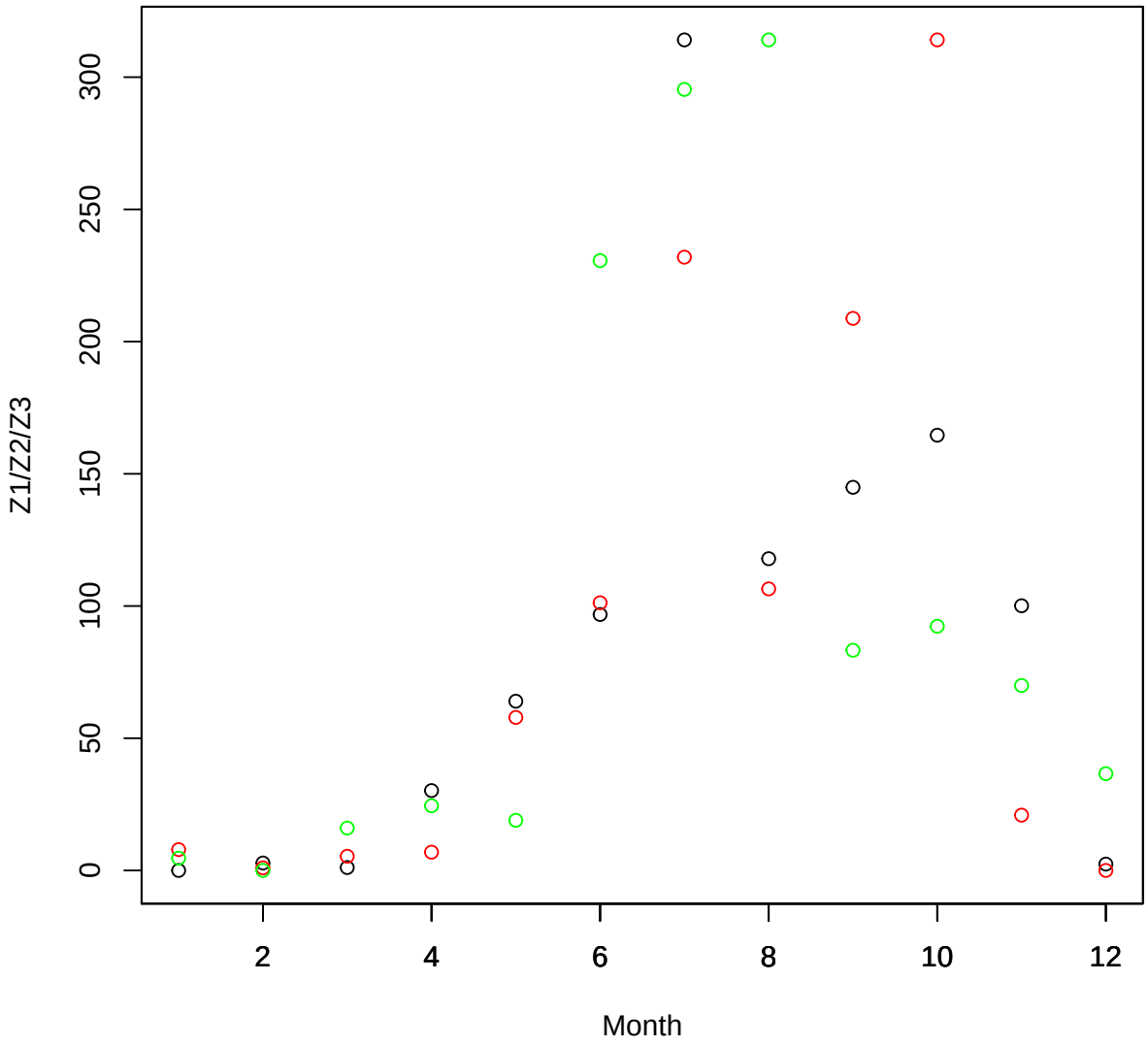
Let us compare it with 1975 and 1999.

```
Z1 <- as.numeric(X[3777,Columns[1:12]])
Z2 <- as.numeric(X[3796,Columns[1:12]])
Z3 <- as.numeric(X[3818,Columns[1:12]])
Z <- data.frame("1956"=Z1,"1975"=Z2,"1999"=Z3)
plot(Z)
```



As you can see, using plot, you can plot the entire dataframe! Sometimes, this is useful. But we are interested in monthly rainfall in these three years. How to do that? Here is a solution!

```
plot(Z1,ylab="Z1/Z2/Z3",xlab="Month")
par(new=TRUE)
plot(Z2,col="red",yaxt="n",ylab="",xlab="")
par(new=TRUE)
plot(Z3,col="green",yaxt="n",ylab="",xlab="")
```



Of course, from the figure, it is clear that the outliers probably come from excess rains in the months June to October! You can confirm this by plotting any one year data (say 1999) with the data of the preceding and following (1998 and 2000, respectively) years.

Home work

- Come up with different solutions to the problems mentioned above!
- Pick a state or region of your preference and carry out the analysis!
- Do a comparison between two different regions of your choice!
- Are all outliers in rainfall in every state or region due to the vagaries of monsoon rainfall?

- Are there any outliers in rainfall in Tamilnadu? Rain in which months contribute to the excess rainfaull? How does it compare with the outliers of North Interior Karanataka? Are they the same months or different? Why?
- Of course, the possibilities are endless! So, keep playing with the data!

2 Numerical description

One can proceed to the calculation of some numerical descriptors of the data. Of course, if you feed the data to the commands that we learnt, such as `mean`, `median`, `sd`, `var` and so on, you will get the numbers.

```
summary(NIK$ANNUAL)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   470.3   620.4   719.0   717.2   790.8  1095.6

sd(NIK$ANNUAL)

## [1] 133.1747
```

So, I will stop this tutorial at this point! This brings us to the end of our descriptive statistics lecture – Lectures 3 and 4 in the course.

Hunt for different data (stock market, bullion rates, employment status, sports data, and so on), do the analysis and try to understand!

Have fun!!