**STAT 40001/MA59800 Statistical Computing Fall 2017**

**Lab-7**

1. The National Highway System Designation Act was signed into law in 1995. It abolished the federal mandate of 55 mph speed limits. The data ***speed*** shows percentage changes in interstate highway traffic fatalities from 1995 to 1996 are provided in the Blackboard.

Important note: Please note that it is a tab-delimited file.

1. Print first 5 lines of the data

> data = read.table("C:\\Users\\wu1114\\Desktop\\speed.txt",sep="\t",header=T)

> head(data,5)

STATE INCREASe FATALITIESCHANGE

1 Alaska No -29.0

2 Connecticut No -4.4

3 Dist. of Columbia No -80.0

4 Hawaii No -25.0

5 Indiana No -13.2

1. Draw the histogram of the percentage changes in interstate highway traffic fatalities from 1995 to 1996

> attach(data)

> hist(FATALITIESCHANGE,col=c(1,2,3,4,5,6,7))

1. Compare the speed limit and traffic fatalities by displaying a side by side boxplots

The box plot should look as below

> x = factor(INCREASe)

> plot(FATALITIESCHANGE~x)

> x = factor(INCREASe,c("Yes","No"))

> plot(FATALITIESCHANGE~x)

> plot(FATALITIESCHANGE~x,col=c(2,3))



1. The link below includes the crime rates for 50 states in 2005.

<http://datasets.flowingdata.com/crimeRatesByState2005.tsv>

1. Import the dataset in R and name it *crime*.

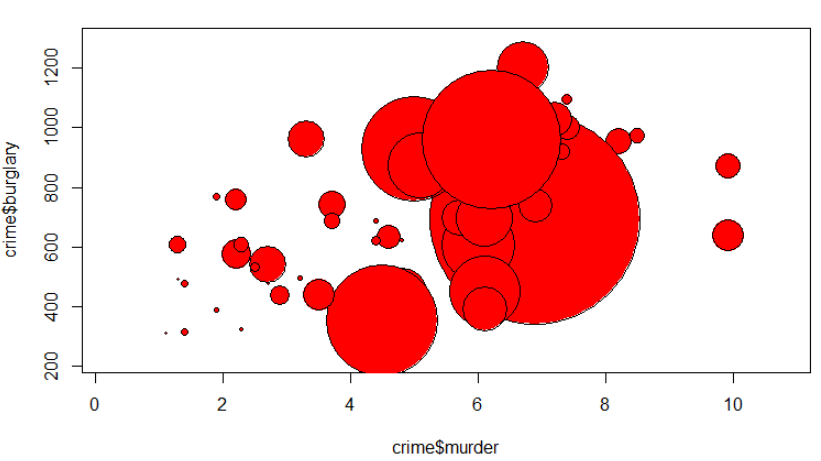
> crime = read.table("http://datasets.flowingdata.com//crimeRatesByState2005.tsv",header=T,sep="\t")

1. How many variables are included in the data

> length(names(crime))

[1] 9

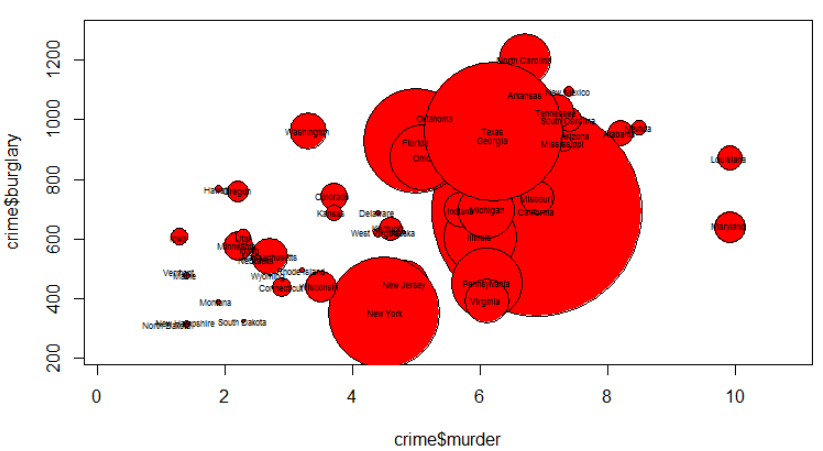
1. Use code below to draw bubble plots



1. Add the name of the states using code:

text(crime$murder, crime$burglary, crime$state, cex=0.5)

You may add option: bg=”red” etc.



1. Daily Observations on Exchange Rates of the US Dollar Against Other Currencies are provided as *Garch* in the link below

<https://vincentarelbundock.github.io/Rdatasets/datasets.html>

1. Import the data in R  
   > data = read.table("https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/Ecdat/Garch.csv",header=T,sep=',')

> head(data,5)

X date day dm ddm bp cd dy sf

1 1 800102 wednesday 0.5861 NA 2.2490 0.8547 0.004206 0.6365

2 2 800103 thursday 0.5837 -0.0041032713 2.2365 0.8552 0.004187 0.6357

3 3 800104 friday 0.5842 0.0008562377 2.2410 0.8566 0.004269 0.6355

4 4 800107 monday 0.5853 0.0018811463 2.2645 0.8538 0.004315 0.6373

5 5 800108 tuesday 0.5824 -0.0049670394 2.2560 0.8553 0.004257 0.6329

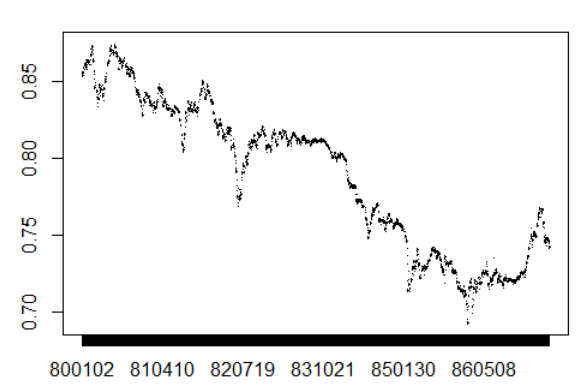
1. How many variables are included in the dataset?

> ncol(data)

[1] 9

1. Choose the Canadian dollar (cd) and display its distribution daily by creating parallel box plot.

> attach(data)  
> date = factor(date)

> plot(date,cd,col=c(1,2,3,4,5,6,7))  


1. The dataset concerning hepatitis are provided in the link below

<https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data>

1. Import the data in R

> data = read.table("https://archive.ics.uci.edu//ml//machine-learning-databases//hepatitis/hepatitis.data",sep = ',')

> head(data,5)

V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20

1 2 30 2 1 2 2 2 2 1 2 2 2 2 2 1.00 85 18 4.0 ? 1

2 2 50 1 1 2 1 2 2 1 2 2 2 2 2 0.90 135 42 3.5 ? 1

3 2 78 1 2 2 1 2 2 2 2 2 2 2 2 0.70 96 32 4.0 ? 1

4 2 31 1 ? 1 2 2 2 2 2 2 2 2 2 0.70 46 52 4.0 80 1

5 2 34 1 2 2 2 2 2 2 2 2 2 2 2 1.00 ? 200 4.0 ? 1

1. Note that the dataset doesn’t have variable names.

Below are the name of the variables

1. Class:   
   2. AGE: 3. SEX: male, female   
   4. STEROID: no, yes   
   5. ANTIVIRALS: no, yes   
   6. FATIGUE: no, yes   
   7. MALAISE: no, yes   
   8. ANOREXIA: no, yes   
   9. LIVER BIG: no, yes   
   10. LIVER FIRM: no, yes   
   11. SPLEEN PALPABLE: no, yes   
   12. SPIDERS: no, yes   
   13. ASCITES: no, yes   
   14. VARICES: no, yes   
   15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00   
   -- see the note below   
   16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250   
   17. SGOT: 13, 100, 200, 300, 400, 500,   
   18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0   
   19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90   
   20. HISTOLOGY: no, yes

Insert these variable names to the dataset  
>names(data)=c("Class","AGE","SEX","STEROID","ANTIVIRALS","FATIGUE","MALAISE","ANOREXIA","LIVER BIG","LIVER FIRM","SPLEEN PALPABLE","SPIDERS","ASCITES","VARICES","BILIRUBIN","ALK PHOSPHATE","SGOT","ALBUMIN","PROTIME","HISTOLOGY")

1. Missing values are marked as “? “ Replace them with NA and delete them.

> data = read.table("https://archive.ics.uci.edu//ml//machine-learning-databases//hepatitis/hepatitis.data",sep = ',',na.strings = '?')

>names(data)=c("Class","AGE","SEX","STEROID","ANTIVIRALS","FATIGUE","MALAISE","ANOREXIA","LIVER BIG","LIVER FIRM","SPLEEN PALPABLE","SPIDERS","ASCITES","VARICES","BILIRUBIN","ALK PHOSPHATE","SGOT","ALBUMIN","PROTIME","HISTOLOGY")

> head(data)

Class AGE SEX STEROID ANTIVIRALS FATIGUE MALAISE ANOREXIA LIVER BIG LIVER FIRM SPLEEN PALPABLE

1 2 30 2 1 2 2 2 2 1 2 2

2 2 50 1 1 2 1 2 2 1 2 2

3 2 78 1 2 2 1 2 2 2 2 2

4 2 31 1 NA 1 2 2 2 2 2 2

5 2 34 1 2 2 2 2 2 2 2 2

6 2 34 1 2 2 2 2 2 2 2 2

SPIDERS ASCITES VARICES BILIRUBIN ALK PHOSPHATE SGOT ALBUMIN PROTIME HISTOLOGY

1 2 2 2 1.0 85 18 4.0 NA 1

2 2 2 2 0.9 135 42 3.5 NA 1

3 2 2 2 0.7 96 32 4.0 NA 1

4 2 2 2 0.7 46 52 4.0 80 1

5 2 2 2 1.0 NA 200 4.0 NA 1

6 2 2 2 0.9 95 28 4.0 75 1  
> data1 = na.omit(data)

> dim(data1)

[1] 80 20

1. How many observations contain missing information?  
   > nrow(data) - nrow(na.omit(data))

[1] 75

1. Eleven golfers in a golf tournament play 18 holes each day. The scores for each of the golfers on the four days are given below. Display this data using parallel boxplots. Make sure you have labels the box plots in the order of the days

Friday Saturday Sunday Monday

70 77 81 70

71 78 83 81

75 81 84 81

79 82 84 88

80 83 86 88

81 83 86 89

83 85 87 90

83 85 87 90

84 85 87 91

85 88 88 93

90 89 89 94

> data = read.table("C:\\Users\\Administrator\\Desktop\\5.txt",header=T,sep="\t")

> head(data,3)

Friday Saturday Sunday Monday

1 70 77 81 70

2 71 78 83 81

3 75 81 84 81

> attach(data)

> boxplot(data.frame(Sunday,Monday,Friday,Saturday),col=c(1,2,3,4),main="box plot",xlab="date",ylab="value")

