STAT S 670 - Exploratory Data Analysis - Homework #2

Ganesh Nagarajan

September 15, 2015

1. Following is the R code for analyzing the Colorado Department of Transport Data.

library(aplpack)

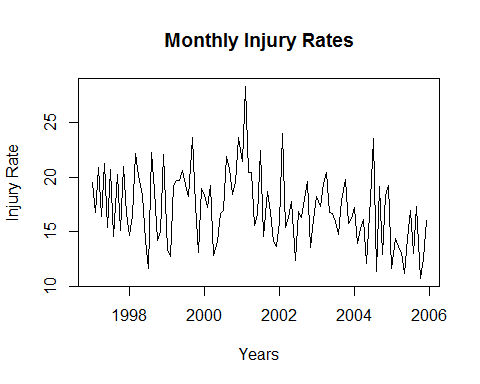
## Warning: package 'aplpack' was built under R version 3.2.2

## Loading required package: tcltk

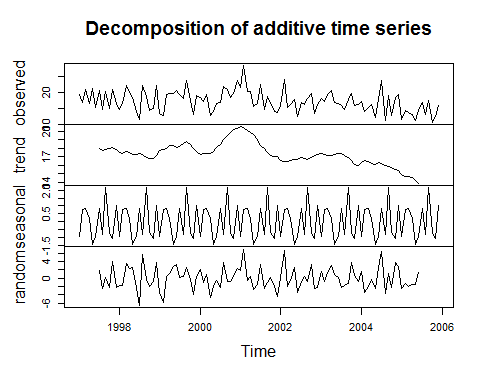
library(outliers)

## Warning: package 'outliers' was built under R version 3.2.2

A<-c(19.50,16.72,20.92,16.42,21.22,15.40,20.68,14.55,20.23,  
 15.11,20.95,16.68,14.67,16.50,22.15,20.14,18.33,14.20,  
 11.61,22.24,18.75,14.22,15.03,22.07,13.34,12.73,19.23,  
 19.74,19.74,20.60,19.29,18.22,23.65,17.44,13.07,19.00,  
 18.44,17.25,19.19,12.77,14.10,16.69,16.92,21.92,20.84,  
 18.43,19.54,23.61,21.40,28.34,20.43,20.43,15.58,16.58,  
 22.44,14.59,18.70,16.79,14.12,13.67,15.94,24.04,15.42,  
 16.26,17.74,12.37,16.87,16.28,17.97,19.56,13.56,16.13,  
 18.20,17.29,19.38,20.47,16.75,16.69,15.93,14.73,17.83,  
 19.78,15.78,16.17,17.18,13.90,15.33,16.10,12.03,17.92,  
 23.56,11.35,19.10,12.91,18.32,19.24,11.57,14.33,13.60,  
 13.12,11.19,14.33,16.91,13.03,17.32,10.70,12.56,16.04)  
B<- ts(A,frequency = 12,start=c(1997,1))  
plot(B,main="Monthly Injury Rates",xlab="Years",ylab="Injury Rate")



tsDecompose <- decompose(B)  
plot(tsDecompose)



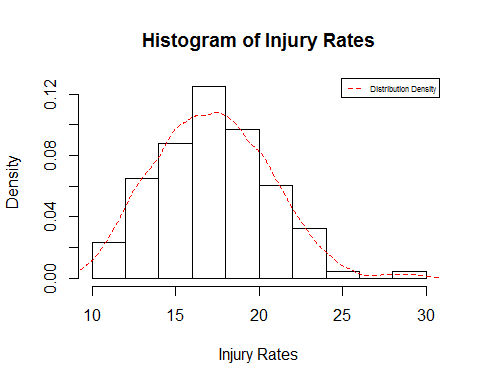
stem.leaf(B,rule.line = "Dixon")

## 1 | 2: represents 1.2  
## leaf unit: 0.1  
## n: 108  
## 1 10 | 7  
## 5 11 | 1356  
## 11 12 | 035779  
## 19 13 | 00135669  
## 29 14 | 1122335567  
## 38 15 | 013445799  
## 56 16 | 011122455666777899  
## (9) 17 | 122347899  
## 43 18 | 22334477  
## 35 19 | 0112223555777  
## 22 20 | 1244466899  
## 12 21 | 249  
## 9 22 | 0124  
## 5 23 | 566  
## 2 24 | 0  
## HI: 28.34

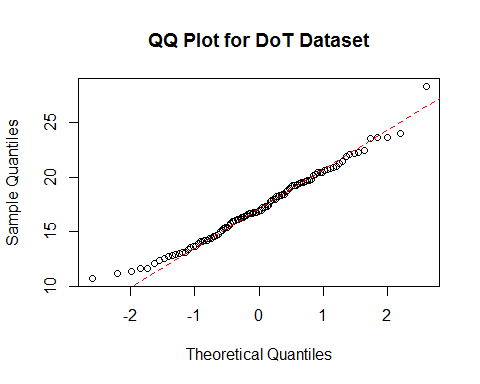
source("./lvalprogs.r")  
lval(B)

## Depth Lower Upper Mid Spread pseudo-s  
## M 54.5 16.890 16.890 16.8900 0.000 0.0000  
## F 27.5 14.630 19.520 17.0750 4.890 3.6250  
## E 14.0 13.120 20.920 17.0200 7.800 3.3903  
## D 7.5 12.465 22.195 17.3300 9.730 3.1712  
## C 4.0 11.570 23.610 17.5900 12.040 3.2318  
## B 2.5 11.270 23.845 17.5575 12.575 2.9192  
## A 1.5 10.945 26.190 18.5675 15.245 3.1530  
## Z 1.0 10.700 28.340 19.5200 17.640 3.3157

hist(B,main = "Histogram of Injury Rates",xlab="Injury Rates",freq = F)  
lines(density(B,kernel = "epanechnikov"),col=2,lty=2)  
legend("topright",legend = c("Distribution Density"),  
 col = c(2),lty = 2,cex = 0.5)



qqnorm(B,main="QQ Plot for DoT Dataset")  
qqline(B,lty=2,col=2)



outlier(B)

## [1] 28.34

1. Time Series Plot for the Data Set is given above Since the dataframe is decomposed, the decomposed trends plor is as follows,
2. From the stem-leaf plot, since the stems below the median is greater than stems above the median, this suggests that mean is to the right of the median. Also the distribution seems to be skewed to right.
3. Letter Value Display ///

d)QQ Plot From the QQ plot it can be seen that there is a light left tail and light right tail, the distribution seems to be almost normal, however there is an outlier at the top right corner. e) Yes, There are Outliers, This can be verified using the histogram aswell as the QQ Plot. Quantitatively, outliers package can be used to identify the outliers,

outlier(B)

## [1] 28.34

1. Outliers based on Gaussian Theory

According to Tukey, the outliers of any single batch of observation can be estimated by the function, outliers = 0.4+0.07n.

Below is the R program finding the outliers,

calculateOutliers<-function (sampleSize){  
 i = length(sampleSize);  
 outlier=0;  
 for ( i in seq(1:i)){  
 outlier=outlier+0.4+(0.00698\*sampleSize[i])  
 }  
 print(outlier)  
}  
sampleSize<-c(120)  
calculateOutliers(sampleSize)

## [1] 1.2376

sampleSize<-c(60,60)  
calculateOutliers(sampleSize)

## [1] 1.6376

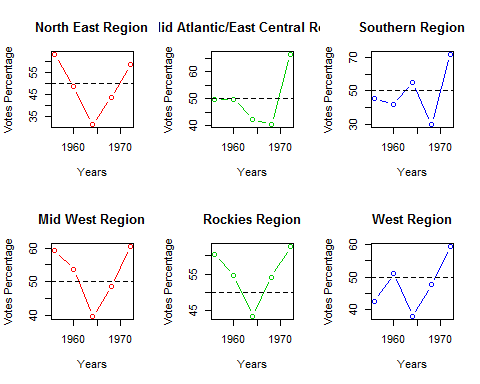
sampleSize<-c(40,30,20,10,5,5,5)  
calculateOutliers(sampleSize)

## [1] 3.6027

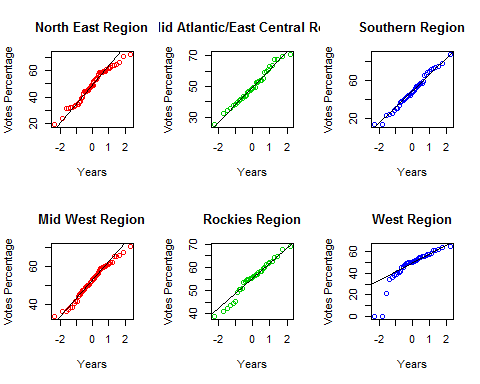
Hence the outliers are, 1) 1.2376 for n=120 2) 1.6376 for n=60,60 3) 3.6027 gor n=40,30,20,10,5,5,5,5

1. R code for the voters dataset analysis is as follows,

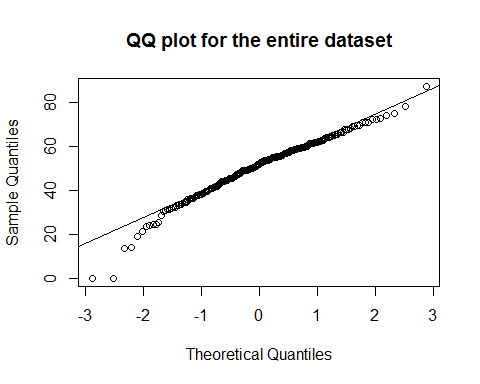
library(cluster)  
library(reshape2)  
library(ggplot2)  
plotAggregate <- function(statesList,df1,xlabc,ylabc,mainc,colc){  
 dataFrame1<-df1[which(df1$StateName %in% statesList),2:3]  
 dataFrame1<-aggregate(dataFrame1,by=list("yearTrend"=dataFrame1$variable),mean)  
 plot(dataFrame1$yearTrend,dataFrame1$value,type="b",xlab=xlabc,ylab = ylabc,main=mainc,col=colc)  
 abline(h=50,lty=2)  
}  
plotqq <- function(statesList,df1,xlabc,ylabc,mainc,colc){  
 dataFrame1<-df1[which(df1$StateName %in% statesList),2:3]  
 #dataFrame1<-aggregate(dataFrame1,by=list("yearTrend"=dataFrame1$variable),mean)  
 qqnorm(dataFrame1$value,xlab=xlabc,ylab = ylabc,main=mainc,col=colc)  
 qqline(dataFrame1$value)  
 #boxplot(dataFrame1$value,xlab=xlabc,ylab = ylabc,main=mainc,col=colc)  
}  
data("votes.repub")  
df<-as.data.frame(votes.repub[,26:30])  
colnames(df)<-c(1956,1960,1964,1968,1972)  
df["StateName"]<-rownames(df)  
df1<-melt(df,id.vars = "StateName")  
df1$StateName<-as.factor(df1$StateName)  
df1$variable<-as.integer(as.character(df1$variable))  
df1$value[is.na(df1$value)]<-0  
NorthEast=c("Connecticut","Delaware","Maine","Massachusetts","New Hampshire","New Jersey","New York","Pennsylvania","Rhode Island","Vermont")  
MaEc<-c("Kentucky","Maryland","North Carolina","South Carolina","Tennessee","Virginia","West Virginia")  
South<-c("Alabama","Arkansas","Florida","Georgia","Louisiana","Mississippi","Oklahoma","Texas")  
midwest<-c("Illinois","Indiana","Iowa","Kansas","Michigan","Minnesota","Missouri","Nebraska","Ohio","Wisconsin")  
rockies<-c("Colorado","Idaho","Montana","North Dakota","South Dakota","Utah","Wyoming")  
west<-c("Alaska","Arizona","California","Hawaii","Nevada","New Mexico","Oregon","Washington")  
par(mfrow=c(2,3))  
plotAggregate(NorthEast,df1,"Years","Votes Percentage","North East Region",2)  
plotAggregate(MaEc,df1,"Years","Votes Percentage","Mid Atlantic/East Central Region",3)  
plotAggregate(South,df1,"Years","Votes Percentage","Southern Region",4)  
plotAggregate(midwest,df1,"Years","Votes Percentage","Mid West Region",2)  
plotAggregate(rockies,df1,"Years","Votes Percentage","Rockies Region",3)  
plotAggregate(west,df1,"Years","Votes Percentage","West Region",4)



par(mfrow=c(2,3))  
plotqq(NorthEast,df1,"Years","Votes Percentage","North East Region",2)  
plotqq(MaEc,df1,"Years","Votes Percentage","Mid Atlantic/East Central Region",3)  
plotqq(South,df1,"Years","Votes Percentage","Southern Region",4)  
plotqq(midwest,df1,"Years","Votes Percentage","Mid West Region",2)  
plotqq(rockies,df1,"Years","Votes Percentage","Rockies Region",3)  
plotqq(west,df1,"Years","Votes Percentage","West Region",4)



par(mfrow=c(1,1))  
qqnorm(df1$value,main="QQ plot for the entire dataset")  
qqline(df1$value)



#qplot(variable,value,data=df1,facets = Region~.,color=StateName,geom="line")+theme(legend.position="bottom")+guides(colour=guide\_legend(nrow=5))+ggtitle("Votes by Country by Region")