Assignment3

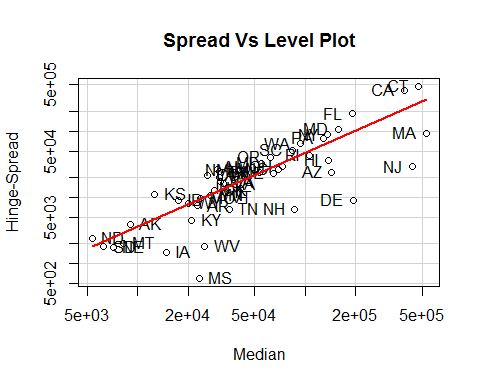
FNU Anirudh

September 30, 2015

Question 1

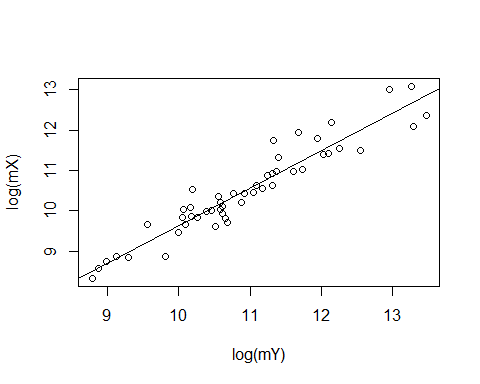
#1 a)  
library(noncensus)  
data(counties)  
library(car)  
  
completeCounties<-counties[complete.cases(counties),]  
stateAggre<-with(counties, aggregate(counties$state, list(counties$state), FUN=unique))  
med<-with(counties, aggregate(counties$population, list(counties$state),FUN=median))  
median<-as.vector(med$x)  
state<-as.matrix(stateAggre$x)  
stateMedianCombine <- cbind(median, state)  
completeMedian<-stateMedianCombine[complete.cases(stateMedianCombine),]  
  
quantiles<-with(counties, aggregate(counties$population, list(counties$state),FUN=quantile, na.rm = TRUE ))  
  
completeQuantiles <- quantiles[complete.cases(quantiles),]  
Forths<-matrix()  
for(i in 1:nrow(quantiles$x)){  
 Forths<-c(Forths,(quantiles$x[i,'75%'] - quantiles$x[i,'25%']))  
}  
transForths<- t(t(Forths))  
completeForths <- transForths[complete.cases(transForths)]  
matrixMH<- cbind(completeMedian,completeForths)  
mX<-as.numeric(matrixMH[-c(8),1])  
mY<-as.numeric(matrixMH[-c(8),3])  
matrixMH2 <- cbind(mX,mY)  
  
#Level vs Spread Plot  
spreadLevelPlot(matrixMH2,by=matrixMH[-c(8),2],main="Spread Vs Level Plot")

## Warning in spreadLevelPlot.default(matrixMH2, by = matrixMH[-c(8), 2], main  
## = "Spread Vs Level Plot"): NAs ignored



## LowerHinge Median UpperHinge Hinge-Spread  
## ND 4153.00 5347.500 6542.00 2389.00  
## SD 5369.50 6269.375 7169.25 1799.75  
## NE 6274.00 7152.500 8031.00 1757.00  
## MT 7198.00 8218.125 9238.25 2040.25  
## AK 7029.00 8999.000 10969.00 3940.00  
## KS 7053.00 12674.000 18295.00 11242.00  
## IA 14200.50 14939.750 15679.00 1478.50  
## ID 13014.00 17491.875 21969.75 8955.75  
## WY 15885.00 19987.250 24089.50 8204.50  
## KY 18751.00 20992.375 23233.75 4482.75  
## AR 19019.00 22763.000 26507.00 7488.00  
## MS 22989.50 23290.625 23591.75 602.25  
## MO 18956.00 23816.250 28676.50 9720.50  
## WV 24069.00 24983.250 25897.50 1828.50  
## CO 15083.50 26006.250 36929.00 21845.50  
## MN 21676.00 27043.000 32410.00 10734.00  
## OK 22119.00 28499.500 34880.00 12761.00  
## NV 16528.00 30000.500 43473.00 26945.00  
## TX 18381.00 30535.375 42689.75 24308.75  
## UT 20802.00 30776.000 40750.00 19948.00  
## GA 22598.00 31214.500 39831.00 17233.00  
## VT 26781.75 31877.375 36973.00 10191.25  
## VA 24544.00 32585.375 40626.75 16082.75  
## IL 27315.50 33364.500 39413.50 12098.00  
## TN 31807.00 35074.250 38341.50 6534.50  
## NM 27213.00 39976.000 52739.00 25526.00  
## IN 33844.00 40806.375 47768.75 13924.75  
## LA 33685.50 44425.500 55165.50 21480.00  
## AL 34339.00 48512.000 62685.00 28346.00  
## WI 41384.00 53248.000 65112.00 23728.00  
## MI 38520.00 55003.750 71487.50 32967.50  
## OR 41536.50 61540.500 81544.50 40008.00  
## ME 53323.00 64776.625 76230.25 22907.25  
## NC 55621.50 68953.125 82284.75 26663.25  
## OH 58185.50 72728.500 87271.50 29086.00  
## SC 57750.00 83540.250 109330.50 51580.50  
## NH 83117.50 86425.750 89734.00 6616.50  
## WA 60699.00 93059.500 125420.00 64721.00  
## RI 83270.00 105124.500 126979.00 43709.00  
## PA 88880.00 127905.000 166930.00 78050.00  
## NY 91301.00 135558.875 179816.75 88515.75  
## HI 117988.00 136411.000 154834.00 36846.00  
## AZ 131346.00 143223.500 155101.00 23755.00  
## MD 103129.50 156613.125 210096.75 106967.25  
## FL 98786.00 190115.000 281444.00 182658.00  
## DE 188084.50 192614.750 197145.00 9060.50  
## CA 179140.50 386866.250 594592.00 415451.50  
## NJ 419669.00 434201.500 448734.00 29065.00  
## CT 231991.00 469961.250 707931.50 475940.50  
## MA 479204.50 525250.000 571295.50 92091.00  
##   
## Suggested power transformation: -0.113459

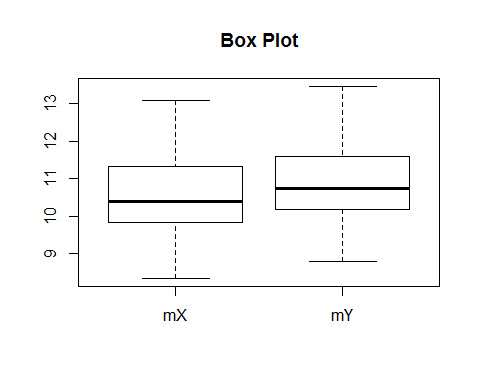
#Scatter plot with log transform  
#scatterplotMatrix(log(matrixMH))  
#abline(lm(log(matrixMH[-c(7),1])~log(matrixMH[-c(7),2])))  
  
#matrixMHState<-cbind(matrixMH,as.matrix(stateAggre$x))  
  
  
plot(log(mX)~log(mY))  
abline(mod<-lm(log(mX)~log(mY)))



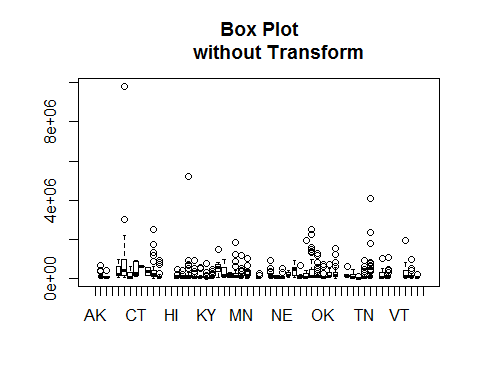
slope<- coef(mod)[2]  
slope

## log(mY)   
## 0.9290134

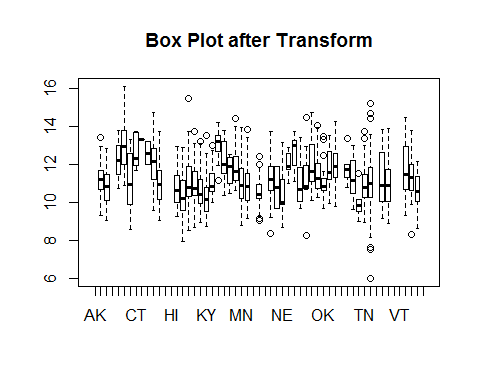
# Equation of line that fits is y=0.93x+c where Slope m=0.93  
# Substituting (10,9.5) we get c=0.2  
# Equation of line is y=0.93x+0.2  
# b)  
#The slope of the line is is 0.93, so p = 1-b = 1-0.93 = 0.07 = 0 (approx)  
#T(x) = log(x)  
boxplot(log(matrixMH2),main="Box Plot")



# c)  
# Without Transform  
boxplot(completeCounties$population~completeCounties$state,main="Box Plot  
 without Transform")



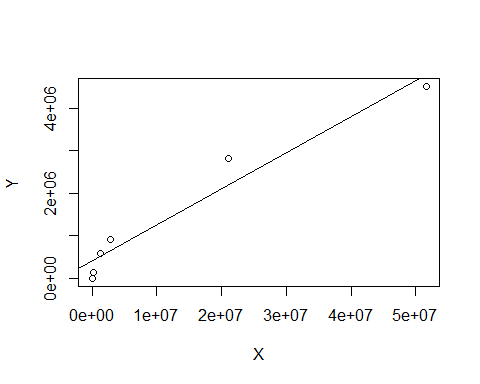
# After Transform  
boxplot((log(completeCounties$population)~completeCounties$state),  
main="Box Plot after Transform")



# d)  
source("C:/Users/lenovo/Documents/lvalprogs.R")  
CAsubset<- completeCounties[completeCounties["state"] == "CA",]  
letterValues <- lval(CAsubset$population)  
#root<-letterValues^(1/30)  
VectorXL <- as.vector(letterValues[,2])  
VectorXU <- as.vector(letterValues[,3])  
  
M <- letterValues[1,"Lower"]   
  
Y<- (VectorXL + VectorXU)/2 - M  
X<- ((VectorXL-M)^2 + (VectorXU - M)^2) / (4\*M)  
letterValue<-c("M","F","E","D","C","B");  
p<- (1-Y/X)  
p

## [1] NaN 0.3345424 0.5727270 0.6801925 0.8663948 0.9126492

Table<- data.frame(letterValue,VectorXL,VectorXU,X,Y,p)  
  
# e)  
plot(Y~X)  
abline(mod<-lm(Y~X))



slope <- coef(mod)[2]  
slope

## X   
## 0.08475095

# b=0 after rounding slope to nearest 0.5  
# Power transform p= 1-0= 1.  
  
# f)  
# T(x) = log(x), then T'(x) = 1/x i.e. z = x0\*(1 + log(x) - log(x0)) = 427761.5 \* log(x) - 5118731  
#a = 427761.5, b = - 5118731

Question 2

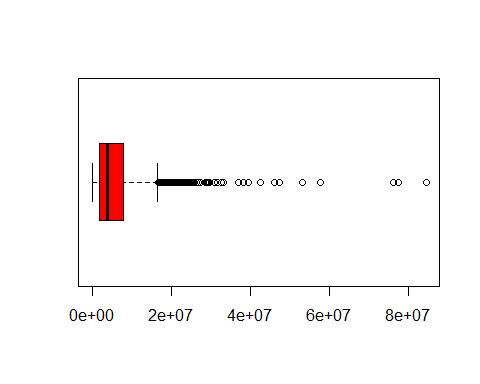
data = read.table("ceo.txt", header = T)  
# a) Number of CEO's and Highest Paid CEO's  
n=length(data$TotalCompensation)  
max\_sal=max(data$TotalCompensation)  
print(max\_sal)

## [1] 84515000

print(n)

## [1] 1835

boxplot(data,horizontal = TRUE,col="red")



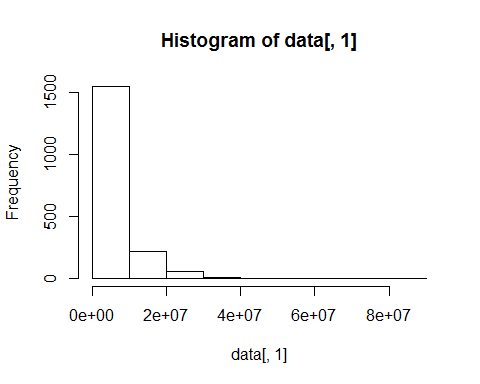
# There are many Outliers which can be said to be unusual values.  
# b) Graphical Display for Data  
hist(data[,1])  
# Distribution is skewed to the right and I would like to transform the   
#data.  
  
# c) Cube root transform will be more appropriate as it will make data  
#symmetric and can resemble normal distribution.  
  
# d) We need to remove low valued data also few Ceo with salary 0 as it   
# affect mean.  
summary(data)

## TotalCompensation   
## Min. : 0   
## 1st Qu.: 1987500   
## Median : 4011000   
## Mean : 6010907   
## 3rd Qu.: 7857000   
## Max. :84515000

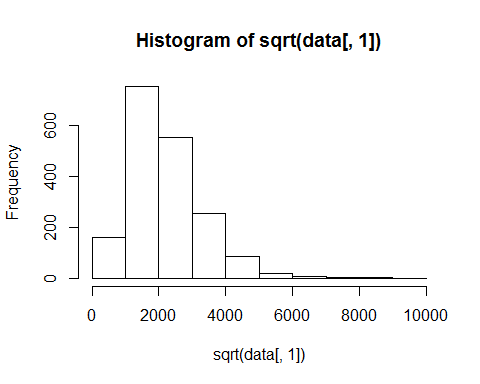
clear\_data<- data[data[,1]> 10000,]  
summary(clear\_data)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 14000 2011000 4044000 6041000 7870000 84520000

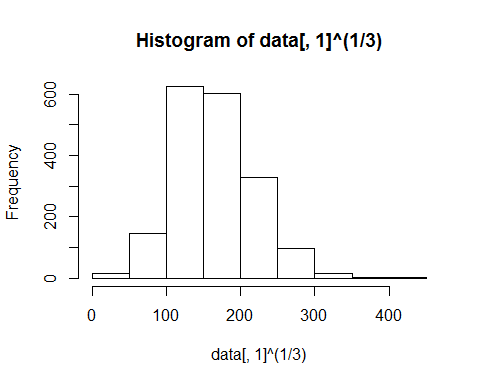
# e)   
hist(data[,1])



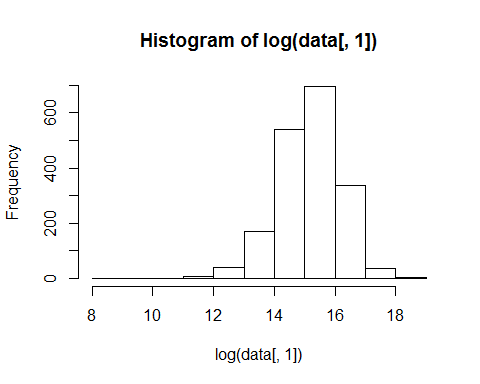
# Square root Transform  
hist(sqrt(data[,1]))



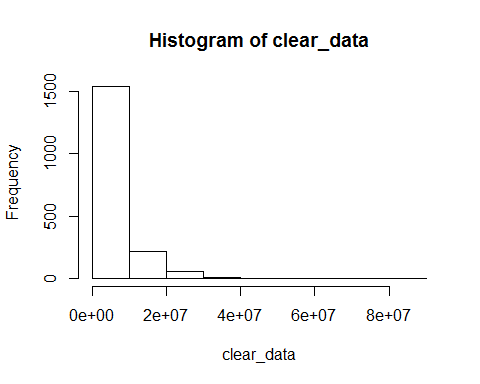
# Cube root Transform  
hist(data[,1]^(1/3))



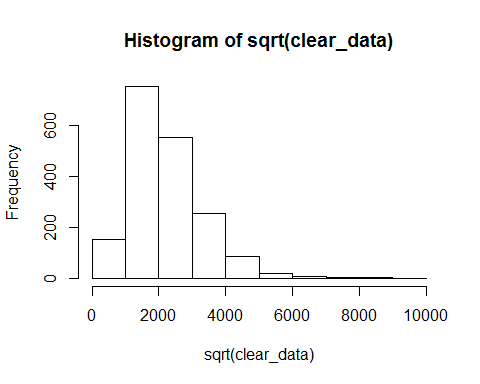
# Log Transform  
hist(log(data[,1]))



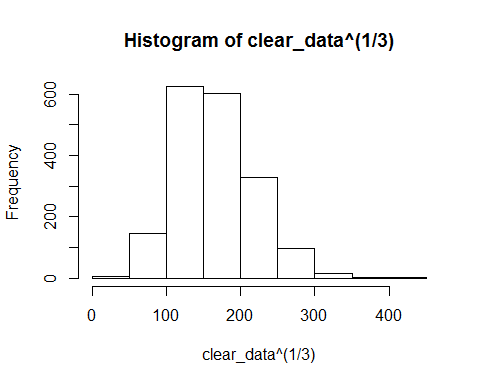
# Transform after removing outliers  
hist(clear\_data)



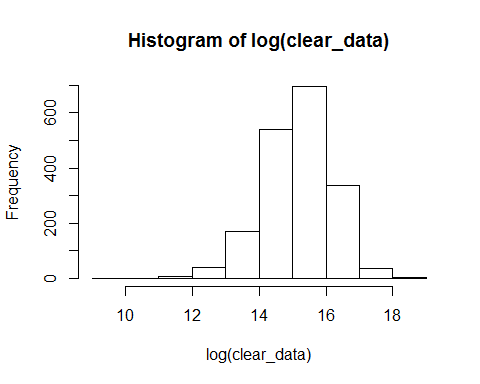
hist(sqrt(clear\_data))



hist(clear\_data^(1/3))



hist(log(clear\_data))



# As you can see data has become more symmetric when cube root transform  
#is applied.  
  
# f) I would choose cube root transform as it makes data symmetric and   
#less skewed compared to log or square root transform

Question 3

#3 a)  
LVhouseHold\_data <- cbind(c(rep(0, 10)),c(0, 2412, 1788, 1517, 1248, 963.5, 727.5, 579,345,114),c(3480,3678,4115.5,4400.5,4799,4978.75,5241,5394.5,5510.25,5494),c(0,4944,6443,7284,8350,8994,9754.5,10210,10675.5,10874));  
  
#x0 = 3480, z = a\*x^(1/3) + b, dz/dx = (1/3)a\*x^(-2/3), Now, at x0 = 3480 , a = 688.9284 , b = -6960 Hence, z = 688.9284\* x^(1/3) - 6960  
  
# b)  
Transformed\_value <- 688.9284\* (LVhouseHold\_data)^(1/3) - 6960  
Transformed\_value

## [,1] [,2] [,3] [,4]  
## [1,] -6960 -6960.0000 3480.000 -6960.000  
## [2,] -6960 2279.1749 3674.359 4776.364  
## [3,] -6960 1401.7387 4080.319 5859.472  
## [4,] -6960 955.9422 4329.501 6394.595  
## [5,] -6960 457.2961 4660.487 7016.644  
## [6,] -6960 -155.5766 4803.797 7367.103  
## [7,] -6960 -763.9000 5006.821 7760.043  
## [8,] -6960 -1217.9504 5122.528 7985.691  
## [9,] -6960 -2128.1462 5208.336 8209.461  
## [10,] -6960 -3619.5187 5196.362 8302.905

# c)  
# Mids are almost same as compared to data in table 4-5.But mids are very  
# far seperated compared to log, square root, fourth root transform,  
# By Comparing 25th and 75th quantiles, it has moved away from orignal  
# data. Spread has increased compared to other transforms hence we can  
# easily identify any outliers.