ASSIGNMENT 01

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

Date: Nov 08, 2019 Group Mates: Hamid, Mateen & Saadoon

Note: *R Commands are provided at the end of each respective answer*

Question No 01:

Present a complete description using the concepts learned in Chapter 2 in Everitt of the dataset: usair.dat

General Information: General information related to usair dataset is provided below:

No. of Columns	No. of Rows
07	41

Column Name	Description	Unit
SO_2	Sulphur dioxide content of air in micrograms	per m^3
Temp	Annual average temperature	fahrenheit
Manuf	Number of manufacturing enterprises employing 20+ workers	unit
Pop	Population size (1970 census	thousands
Wind	Average annual wind speed	miles/hour
Precip	Average annual precipitation	inches
Days	Average number of days with precipitation	per year

Summary Statistics for Multivariate Data:

Column wise Mean

SO_2	Temp	Manuf	Pop	Wind	Precip	Days
30.04	-55.76	463.09	608.60	9.44	36.76	113.90

Correlation Matrix

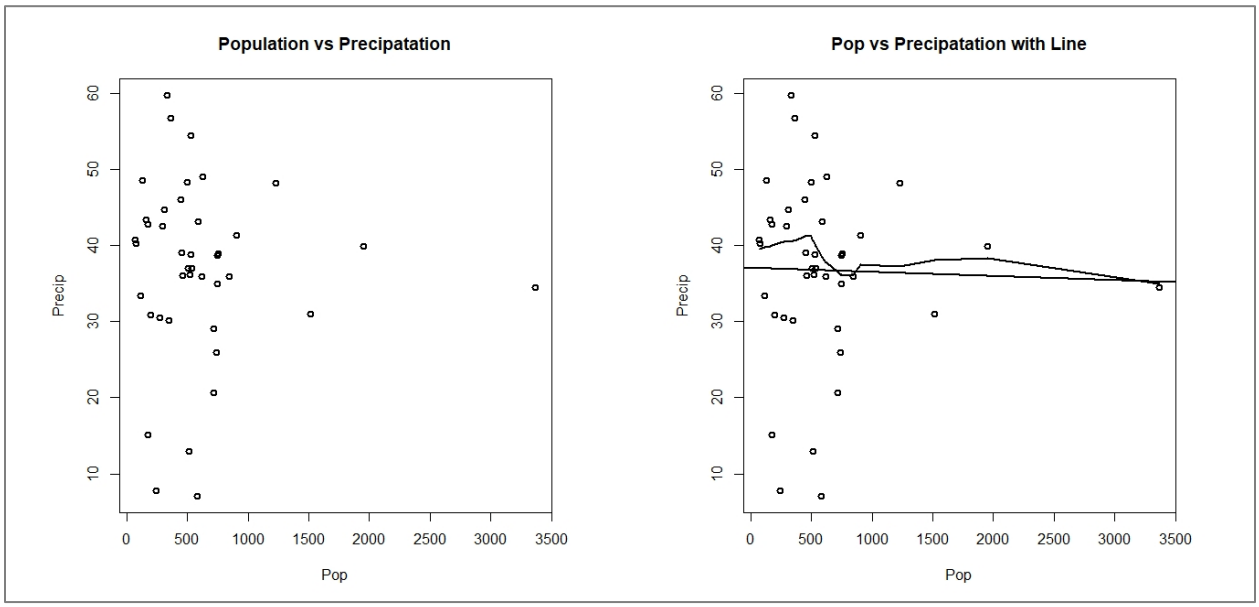
	SO_2	Neg.Temp	Manuf	Pop	Wind	Precip	Days
SO_2	1.00	0.43	0.64	0.49	0.09	0.05	0.37
Neg.Temp	0.43	1.00	0.19	0.06	0.35	-0.39	0.43
Manuf	0.64	0.19	1.00	0.96	0.24	-0.03	0.13
Pop	0.49	0.06	0.96	1.00	0.21	-0.03	0.04
Wind	0.09	0.35	0.24	0.21	1.00	-0.01	0.16
Precip	0.05	-0.39	-0.03	-0.03	-0.01	1.00	0.50
Days	0.37	0.43	0.13	0.04	0.16	0.50	1.00

ASSIGNMENT 01

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

Note: In order to provide description of data using concept learned, we are taking two variables as (1) Population (2) Precipitation.

CONCEPT 01: SCATTER PLOT

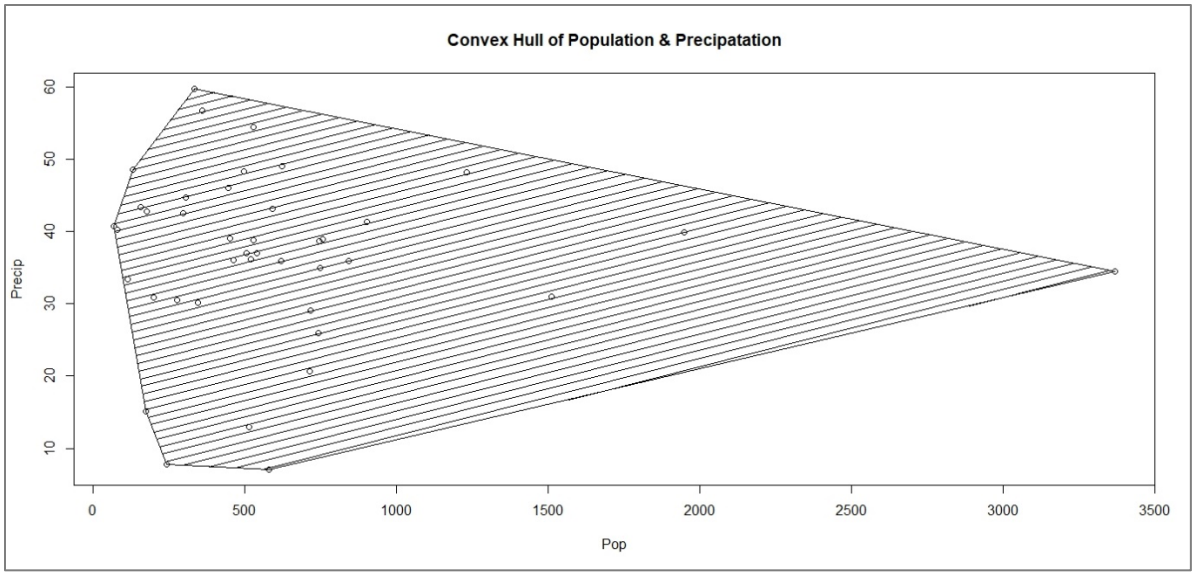


By just looking at the scatter plot (left hand side), we can not identify any particular trend. We cannot highlight a clear dependency. As we apply linear line to the plot, it appears as horizontally straight which confirms that there is No Dependency.

Lowess (local weighted regression) is a technique which is applied locally to identify non-linear behaviors. If lowess differs from the linear model, then it means data is overall Not Linear. In our case, lowess differs from the straight line, hence we can say that data is Not Linear.

Jitter is a graphical technique to slightly shake the data to highlight patterns. In our case as data is already clearly not dependent and not linear. We have not applied jitter as we feel it won't highlight much more insights about our observations.

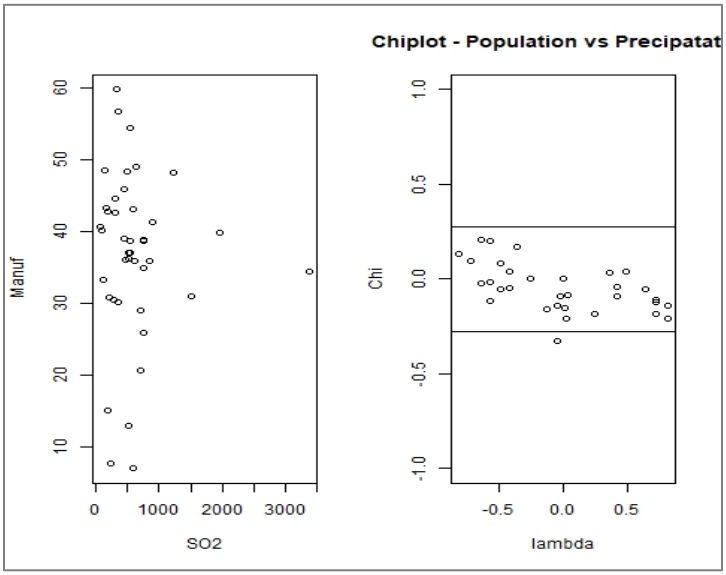
CONCEPT 02: CONVEX HULL



Convex Hull is a technique to identify outliers. Convex Hull highlights extremes which may be potential outliers. We cannot say with assurance that it is an outlier unless we apply other technique to confirm our inference.

In our case, Population value of 3000+ is far away from all other observations and can be considered as possible outlier. We may ignore this observation as it affects our data. Correlation [cor(Pop,Precip)] for data is -0.02 whereas we compute correlation after removing convex hull [cor(SO2[-hull],Precip[-hull])], correlation value is 0.04, which supports our conclusion.

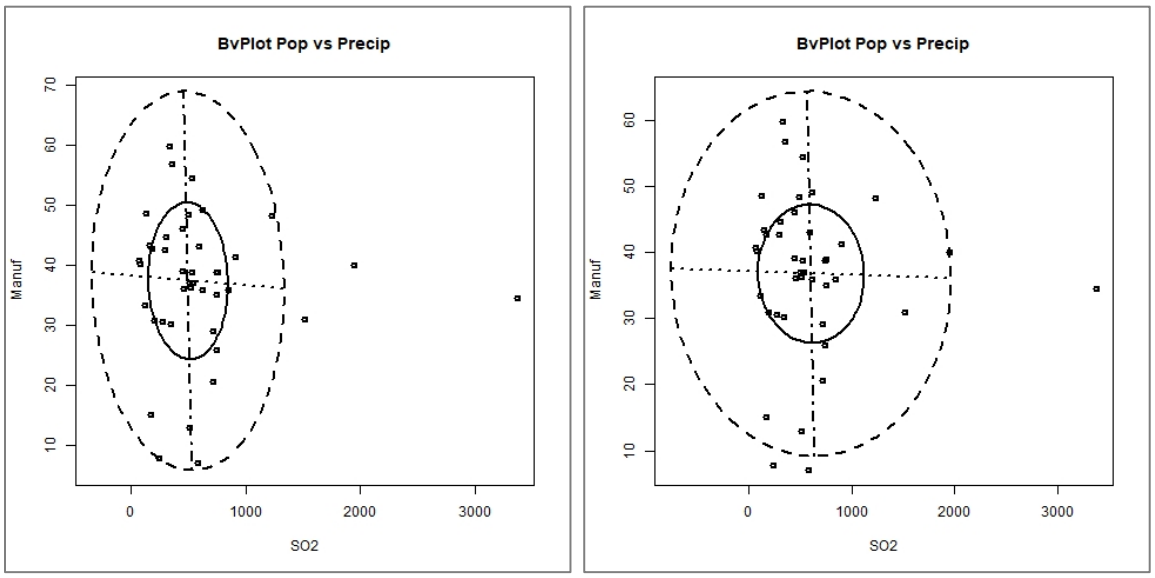
CONCEPT 03: CHI PLOT



Chi plot is a graphical procedure to check for dependence between variables (not necessarily linear). It is in form of a band which depicts Confidence Interval. Plot is interpreted as so that if all observations lie inside the band then variables are Independent whereas if most of the observations are inside and some are outside that means there is a dependency.

In our case, as almost all observation falls inside the horizontal band, we can say that variables are Independent.

CONCEPT 04: BIVARIATE BOX PLOT

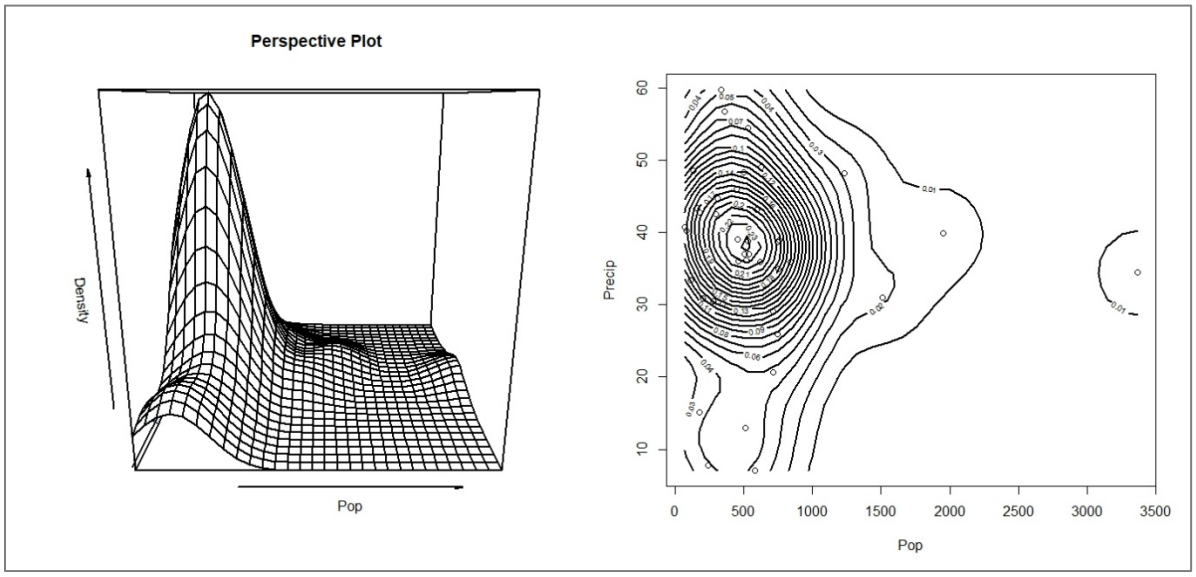


With method = 'O'

Bivariate plot is for multivariate analysis like Box plot for a variable. Both are used to identify outliers by considering quantiles of data with center at median.

In our case, we can state that there are very few outliers. Same was depicted in Convex Hull. Most of the data is concentrated near median and within first two quantiles.

CONCEPT 05: PERSPECTIVE AND CONTOUR PLOT

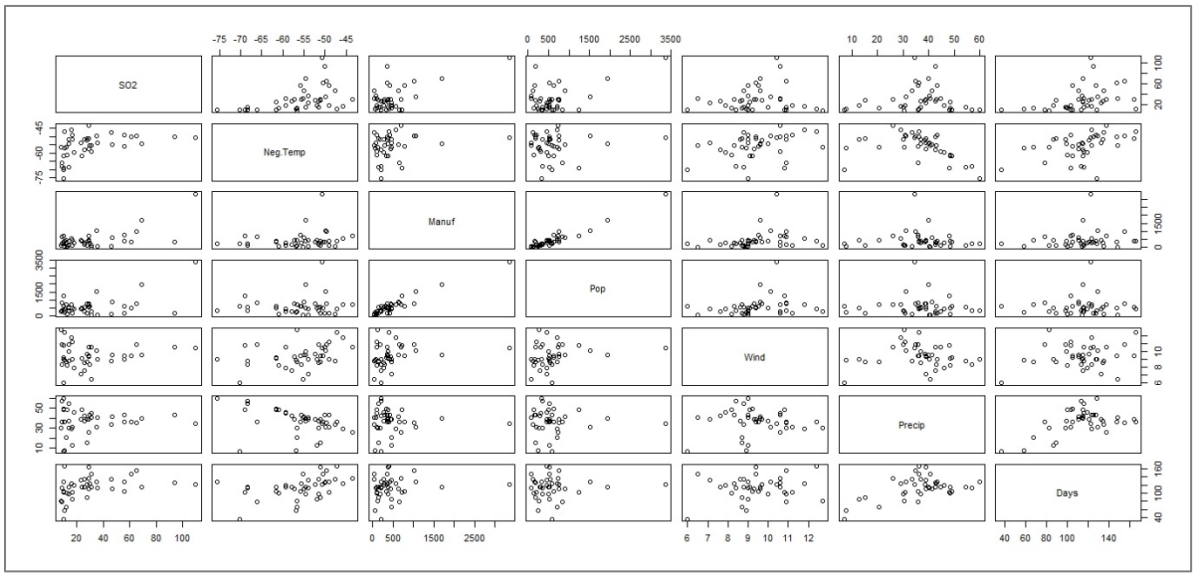


Perspective plot is a kind of three-dimensional representation of a histogram depicting observations plot. It gives a clear idea about the skewness of bivariate density of data.

Contour plot is a technique similar in aim to perspective plot, with a representation in form of contour lines which are concentric in nature and connected on the basis of bivariate density.

In our case, both plots inform us that data is skewed, with few possible outliers.

CONCEPT 06: SCATTER PLOT MATRIX



Scatterplot matrix is a quick glance over the data in form of a collection of scatterplots between all variables. This quickly provide us with observation related to relationships among variables.

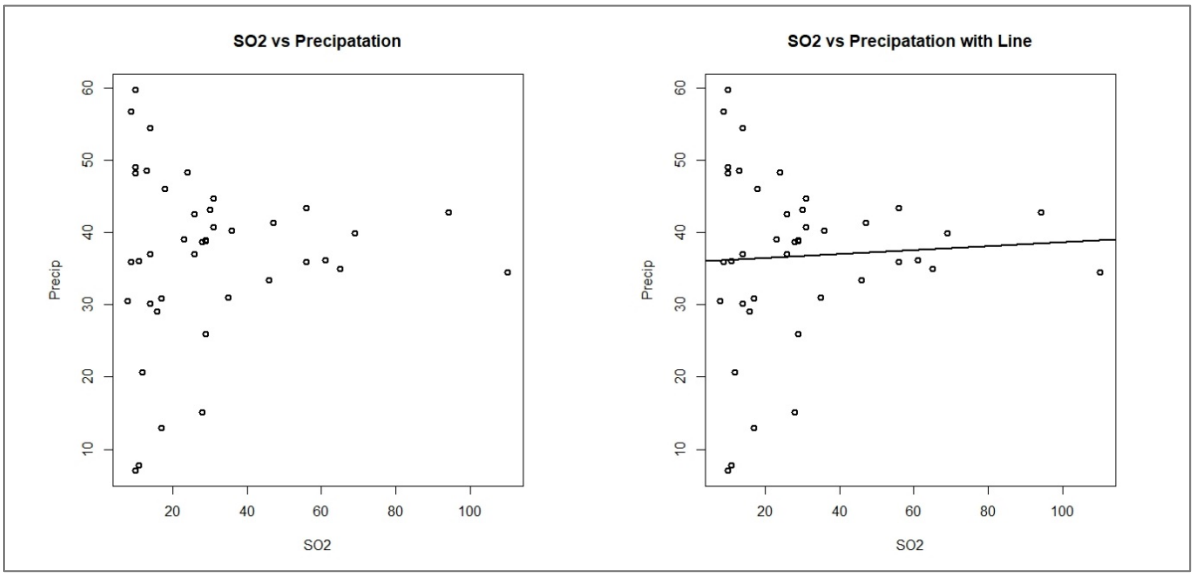
In our case, we can say that Manuf and Pop are positively dependent, Precipitation and Neg-Temperature are negatively dependent and no clear dependency among many of the variables.

If we consider one variable of SO_2, and observes all other variables related to it, we can say:

- SO_2 and Negative temperature are dependent.
- SO_2 and Manuf's plot is not that clear and, in this case, jitter will help highlight more.
- SO_2 and Pop – as above
- SO_2 and Wind, Precip and Days and all are kind of dependent

Note: But we need to apply more techniques to say with assurance.

(one example related to SO_2 is illustrated below with a scatterplot)



After adding a trend line, we can see that variables are slightly positively dependent.

R Commands:

```
center<-colMeans(usair)

var(usair)

cor(usair)


par(mfrow=c(1,2))

par(pty="s")

plot(Pop,Precip,pch=1,lwd=2)

title("Population vs Precipatation",lwd=2)

plot(Pop,Precip,pch=1,lwd=2)

abline(lm(Precip~Pop),lwd=2)

lines(lowess(Pop,Precip),lwd=2)

title("Pop vs Precipatation with Line",lwd=2)


par(mfrow=c(1,2))

hist(Precip,lwd=2)

boxplot(Precip,lwd=2)


hull<-chull(Pop,Precip)

plot(Pop,Precip,pch=1)

title("Convex Hull of Population & Precipatation",lwd=2)

polygon(Pop[hull],Precip[hull],density=15,angle=30)

cor(Pop,Precip)
```

```
cor(SO2[-hull],Precip[-hull])


jpeg(file="E:\\Chiplot SO2 vs Manuf.jpg",quality = 100)
par(mfrow=c(1,1))
chiplot(SO2,Manuf,vlabs=c("SO2","Manuf"))
title("Chiplot - SO2 vs Manufacturing",lwd=2)
dev.off()


jpeg(file="E:\\Chiplot Pop vs Percp.jpg",quality = 100)
par(mfrow=c(1,1))
chiplot(Pop,Precip,vlabs=c("SO2","Manuf"))
title("Chiplot - Population vs Precipatation",lwd=2)
dev.off()


jpeg(file="E:\\BVP Pop vs Precip.jpg",quality = 100)
par(mfrow=c(1,1))
bvbox(cbind(Pop,Precip),xlab="SO2",ylab="Manuf")
title("BvPlot Pop vs Precip",lwd=2)
dev.off()


jpeg(file="E:\\BVP SO2 vs Manuf.jpg",quality = 100)
par(mfrow=c(1,1))
bvbox(cbind(SO2,Manuf),xlab="SO2",ylab="Manuf")
title("BvPlot SO2 vs Manuf",lwd=2)
```

```
dev.off()


jpeg(file="E:\\BVPO Pop vs Precip.jpg",quality = 100)
par(mfrow=c(1,1))
bvbox(cbind(Pop,Precip),xlab="SO2",ylab="Manuf", method="O")
title("BvPlot Pop vs Precip",lwd=2)
dev.off()


jpeg(file="E:\\BVPO SO2 vs Manuf.jpg",quality = 100)
par(mfrow=c(1,1))
bvbox(cbind(SO2,Manuf),xlab="SO2",ylab="Manuf", method="O")
title("BvPlot SO2 vs Manuf",lwd=2)
dev.off()


jpeg(file="E:\\Perspective Plot Pop vs Precip.jpg",quality = 100)
par(mfrow=c(1,1))
h2d<-hist2d(Pop,Precip)
persp(h2d,xlab="Pop",ylab="Precipitation",zlab="Frequency")


jpeg(file="E:\\Perspective Plot Pop vs Precip.jpg",quality = 100)
par(mfrow=c(1,1))
den1<-bivden(Pop,Precip)
persp(den1$seqx,den1$seqy,den1$den,xlab="Pop",ylab="Precip",zlab="Density",lwd=2)
dev.off()
```

```
jpeg(file="E:\\Perspective Plot SO2 vs Manuf.jpg",quality = 100)
par(mfrow=c(1,1))
den1<-bivden(SO2,Manuf)
persp(den1$seqx,den1$seqy,den1$den,xlab="SO2",ylab="Manuf",zlab="Density",lwd=2)
dev.off()

par(mfrow=c(1,2))
den1<-bivden(Pop,Precip)
persp(den1$seqx,den1$seqy,den1$den,xlab="Pop",ylab="Precip",zlab="Density",lwd=2)
plot(Pop,Precip)
contour(den1$seqx,den1$seqy,den1$den,lwd=2,nlevels=20,add=TRUE)

par(mfrow=c(1,2))
den1<-bivden(SO2,Manuf)
persp(den1$seqx,den1$seqy,den1$den,xlab="SO2",ylab="Manuf",zlab="Density",lwd=2)
plot(SO2,Manuf)
contour(den1$seqx,den1$seqy,den1$den,lwd=2,nlevels=20,add=TRUE)
```

Question No 02:

Using the matrix data of usair.dat, calculate the set of following distances:

a) Euclidean distances for original data
b) Euclidean distances for normalized data
c) Mahalanobis distances for the original data

Euclidean Distance in General:

Euclidean distance provide difference between two observations (rows) in terms of a distance value. But important to note is that it does not consider that datasets are in different scales.

Mahalanobis Distance in General:

Mahalanobis distance gives a measure of distance as above but considers the distribution of data and gives a measure of distance from the mean vector.

a) Euclidean Distance for Original Data:

	Phoenix	Little Rock	San Francisco	Denver	Hartford
Phoenix	0.00	472.54	277.30	255.93	481.19
Little Rock	472.54	0.00	688.46	529.16	326.28
San Francisco	277.30	688.4656	0.00	202.16	564.93
Denver	255.93	529.1679	202.16	0.00	365.16
Hartford	481.19	326.2879	564.93	365.16	0.00

b) Euclidean Distance for Normalized Data:

In order to normalize the data, we divide each column by its respective standard deviation.

	Phoenix	Little Rock	San Francisco	Denver	Hartford
Phoenix	0.00	4.79	3.17	3.87	6.23
Little Rock	4.79	0.00	3.01	3.49	2.82
San Francisco	3.17	3.01	0.00	1.26	3.80
Denver	3.87	3.49	1.26	0.00	3.52
Hartford	6.23	2.85	3.80	3.52	0.00

We can note that data has been normalized and is in a common scale.

c) Mahalanobis Distance for Original Data:

First, we compute column wise means and covariance and then use them to calculate required distance.

Phoenix	Little Rock	San Francisco
20.258912	6.207429	4.53921
Denver	Hartford	Wilmington
5.400965	7.760199	2.15683
Washington	Jacksonville	Miami
1.5383	5.286583	14.277037
Atlanta	Chicago	Indianapolis
1.448871	26.89145	4.2701
Des Moines	Wichita	Louisville
4.310154	9.060638	3.060861
New Orleans	Baltimore	Detroit
5.394046	3.421276	7.222633
Minneapolis-St. Paul	Kansas City	St. Louis
4.94583	3.06065	4.767204
Omaha	Alburquerque	Albany
3.052545	8.063093	4.081128
Buffalo	Cincinnati	Cleveland
12.880983	7.265585	11.489013
Columbus	Philadelphia	Pittsburgh
3.14576	6.722708	7.955753
Providence	Memphis	Nashville
18.17604	3.573384	2.564959
Dallas	Houston	Salt Lake City
6.36829	8.684843	4.43138
Norfolk	Richmond	Seattle
3.75473	2.492672	6.535345
Charleston	Milwaukee	
7.888784	5.593824	

Table provides information of distance of an observation from mean of the data. Chicago has the highest value of 26.89 which indicates that it is most distant than others whereas many observations have low values with Atlanta having the smallest as 1.44.

R Commands:

```
dis1<-dist(usair)
distance1<-dist2full(dis1)
round(distance1,digits=2)

write(distance1,file="E:\\question02-1-distance1.csv",ncolumns=ncol(distance1),sep=",")
subsetdistance1<-cbind(distance1[1:5,1:5])

write(subsetdistance1,file="E:\\question02-1-
subset.csv",ncolumns=ncol(subsetdistance1),sep=",")

---------------

dataWithoutColumn<-cbind(usair[,1],usair[,2],usair[,3],usair[,4],usair[,5],usair[,6],usair[,7])

sd(dataWithoutColumn)
columnWiseSD<-sapply(usair,sd)
normalizedData<-sweep(usair,2,columnWiseSD,FUN='/')

dis2<-dist(normalizedData)
distance2<-dist2full(dis2)
round(distance2,digits=2)

write(distance2,file="E:\\question02-2-distance2.csv",ncolumns=ncol(distance2),sep=",")
subsetdistance2<-cbind(distance2[1:5,1:5])

write(subsetdistance2,file="E:\\question02-2-
subset.csv",ncolumns=ncol(subsetdistance2),sep=",")

---------------

center<-colMeans(usair)
covariance<-cov(usair)

distance3<-mahalanobis(usair,center,covariance)
write(distance3,file="E:\\question02-3.csv")
```

Question No 03:

Show the Q-Q plots of individual variables for usair dataset and explain the results.

Q-Q Plot in General: Plots such as quantile-quantile or probability-probability involves ordering the observations and then plotting them against values of assumed Cumulative Distribution Function. QQ plot is used to get insights of data about its Normality. If the plot follows or aligns with the line, then it means that variable may follow a normal distribution. It is important to note that it is just a graphical procedure not a formal one to state normality of data with assurance.

Usair dataset contains seven variables. QQ plot for each is available on next page. Explanation is provided below:

SO2: We can say that data is not normally distributed completely. Major part of data aligns with the straight line however part of it deviates as well. As possible outliers are points at the end of line, distanced from the bulk of the observation, in interpretation we may ignore their consideration.

Neg Temp: We can say that data follows the normal distribution as almost all the data is on the straight line after ignoring the outliers.

Manuf: Data strictly follows the normal distribution. As 75% of the data lies before 1 (quantile 3), which is normally distributed we can apply the same inference for the whole data.

Population: Same as manufacturing enterprises, data follows the normal distribution as 75% of the data strictly follows straight line.

Wind: We can say that is generally normal as it runs along the straight line.

Precipitation: 50% of the data strictly follows straight line. Ignoring the outliers, we may consider data as normally distributed in general.

Days: Same as Precipitation, we may consider data as normally distributed in general.

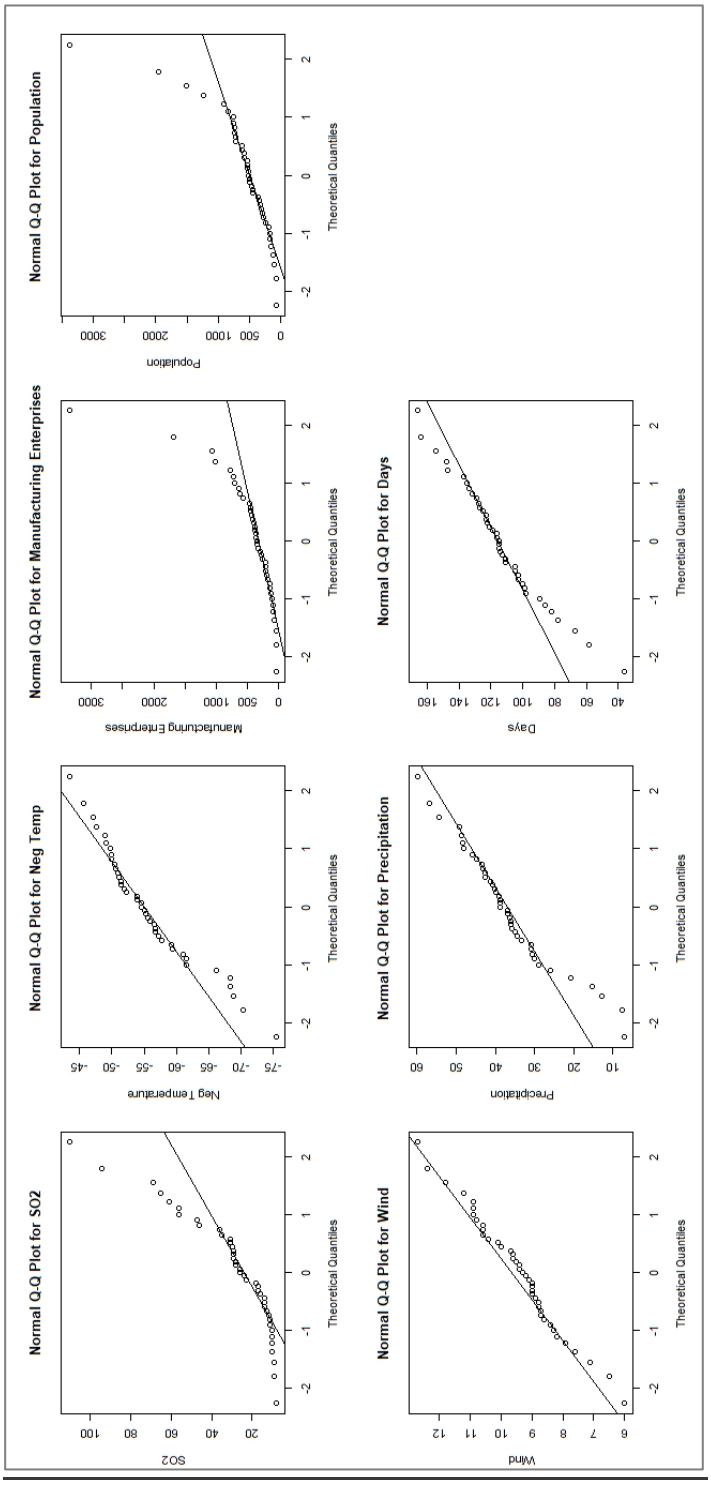
Important:

In order to highlight the deviation from straight line, another approach is to take log of data and then plot the same. We have provided those plots as well which in this case does not identify anything new however, deviation from straight line is slightly highlighted.

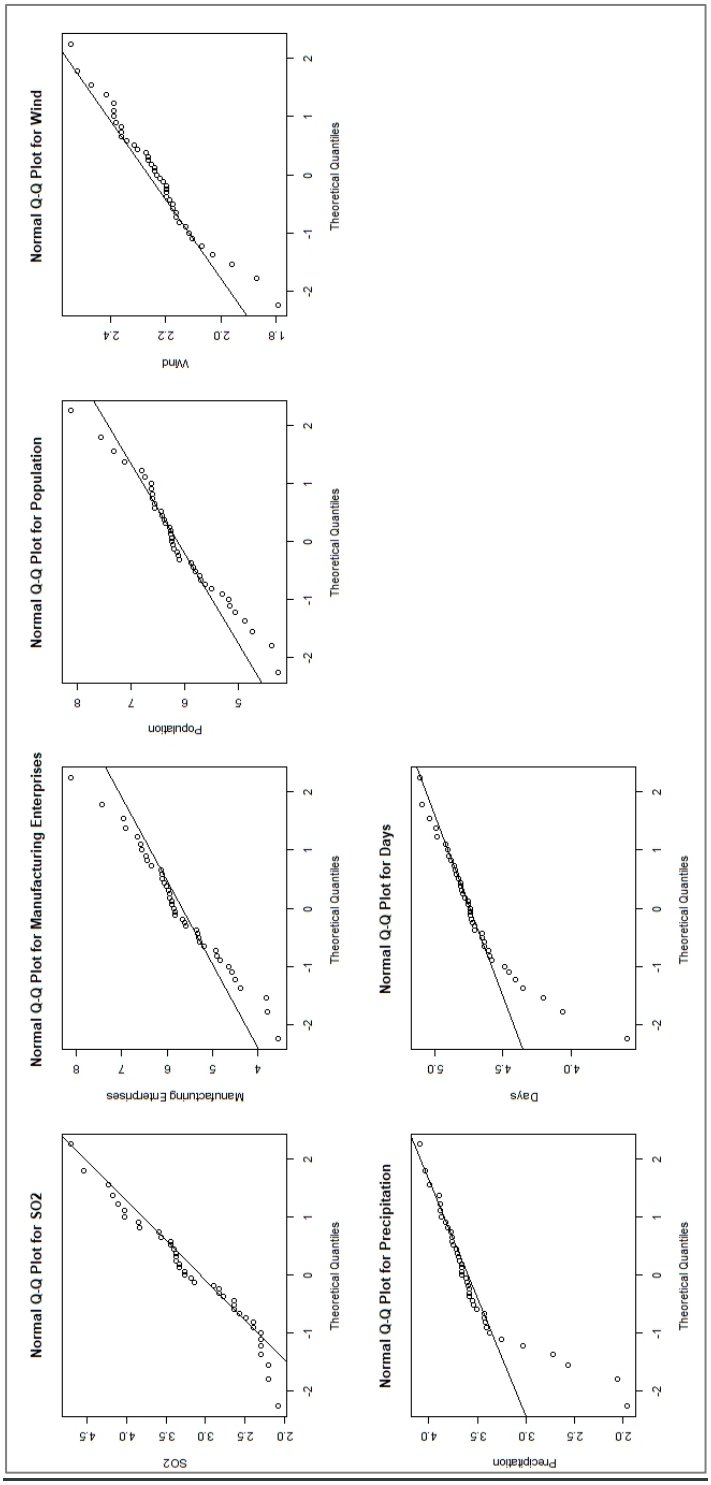
ASSIGNMENT 01

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

QQ Plot:



QQ Plot (after log of data):



R Commands:

```
jpeg(file="E:\\question03.jpg")
par(mfrow=c(2,4))
qqnorm(usair[,1],ylab="SO2",main = "Normal Q-Q Plot for SO2")
qqline(usair[,1])
qqnorm(usair[,2],ylab="Neg Temperature",main = "Normal Q-Q Plot for Neg Temp")
qqline(usair[,2])
qqnorm(usair[,3],ylab="Manufacturing Enterprises",main = "Normal Q-Q Plot for Manufacturing
Enterprises")
qqline(usair[,3])
qqnorm(usair[,4],ylab="Population",main = "Normal Q-Q Plot for Population")
qqline(usair[,4])
qqnorm(usair[,5],ylab="Wind",main = "Normal Q-Q Plot for Wind")
qqline(usair[,5])
qqnorm(usair[,6],ylab="Precipitation",main = "Normal Q-Q Plot for Precipitation")
qqline(usair[,6])
qqnorm(usair[,7],ylab="Days",main = "Normal Q-Q Plot for Days")
qqline(usair[,7])
dev.off()
```

Related to Log:

```
qqnorm(log(usair[,1]),ylab="SO2",main = "Normal Q-Q Plot for SO2")
qqline(log(usair[,1]))
```

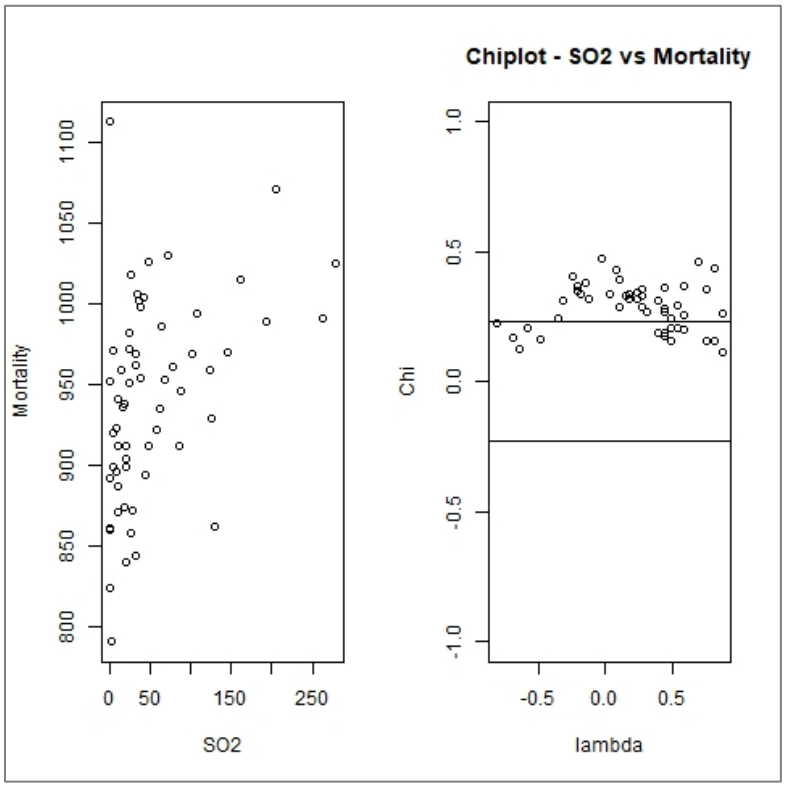
Question No 04:

Investigate the use of Chi Plot function on all pairs of variables in the air pollution data.

Chi Plot in General: Chi plot is a graphical procedure to check for dependence between variables (not necessarily linear). It is in form of a band which depicts Confidence Interval. Plot is interpreted as so that if all observations lie inside the band then variables are Independent whereas if most of the observations are inside and some are outside that means there is a dependency.

For solution to this question, we have tried five pairs and their explanation is as follows:

Chi Plot - SO2 vs Mortality



Description:

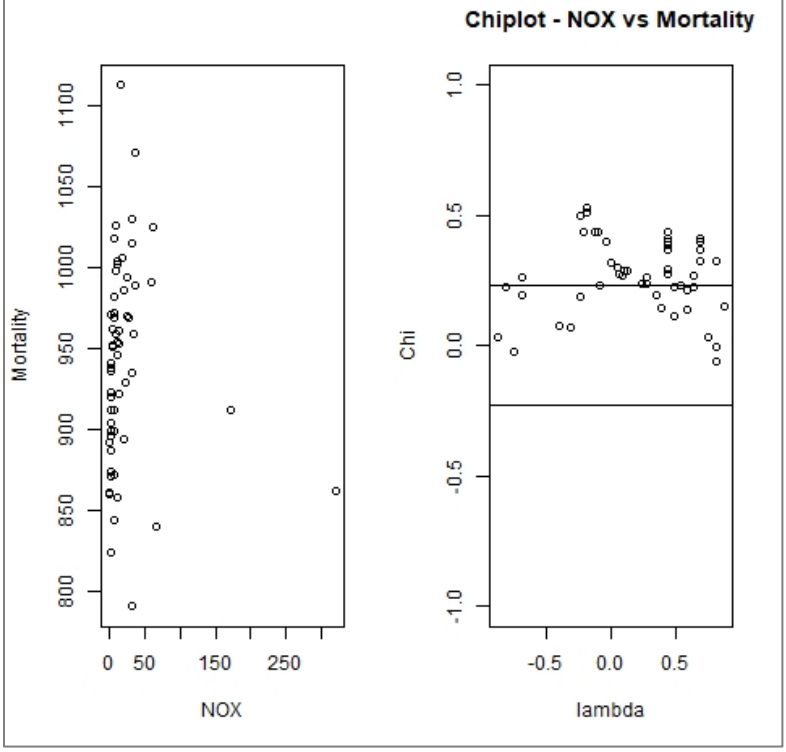
Lack of points in the horizontal band shows departure from independence. As most of the observations are outside the band, this chi plot suggests that variables are dependent.

Same inference can be made by looking at the scatterplot on the left-hand side. We can see that as values of SO2 are increased, Mortality is also increased.

ASSIGNMENT 01

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

Chi Plot - NOX vs Mortality

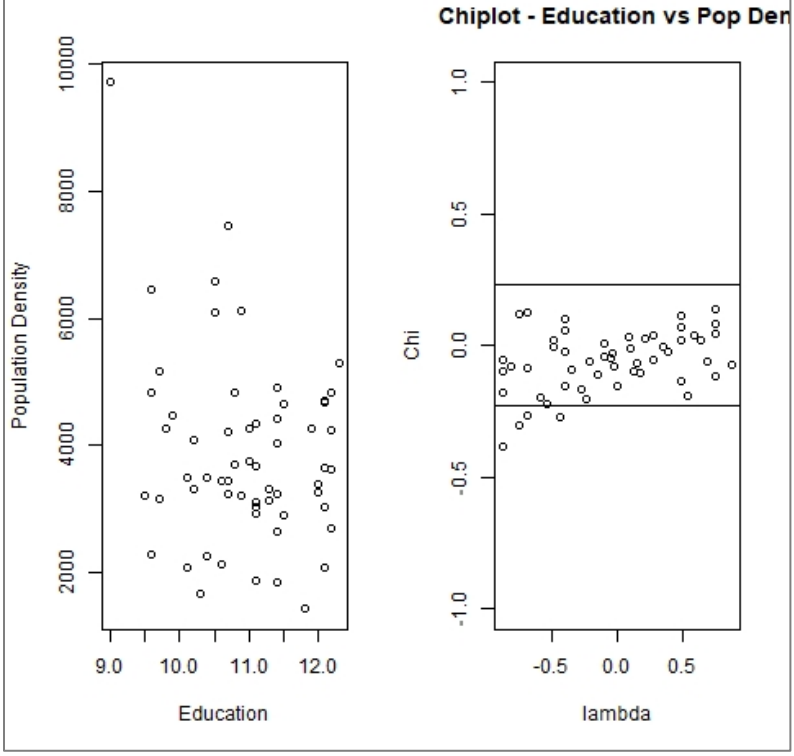


Description:

Like SO2, NOX's plot against Mortality also shows tendency of being dependent as most of the observations departs from the horizontal band. We can say that NOX and Mortality are dependent.

Scatter plot does not clearly support the above argument of dependency. Most of the observations are clustered along the y-axis and shows that there are high number of mortalities even when SO2 count is small.

Chi Plot – Education vs Population Density



Description:

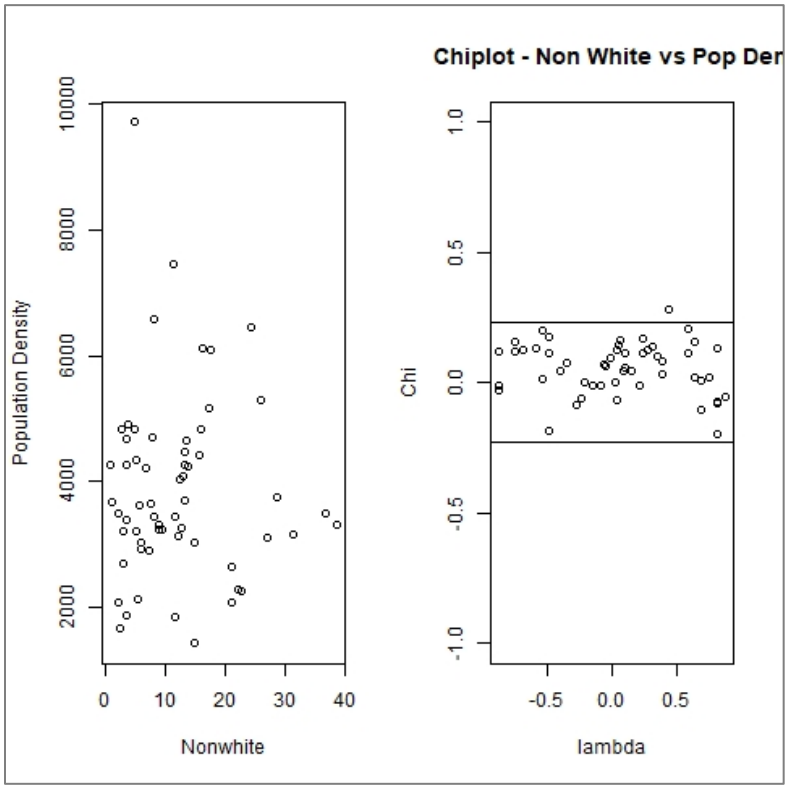
In Chi plot of Education and Population density, almost all the observations are inside the horizontal band which confirms independence of variables.

Scatter plot also shows a random distribution of observations with no trend tendency.

ASSIGNMENT 01

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

Chi Plot – Non-White vs Population Density

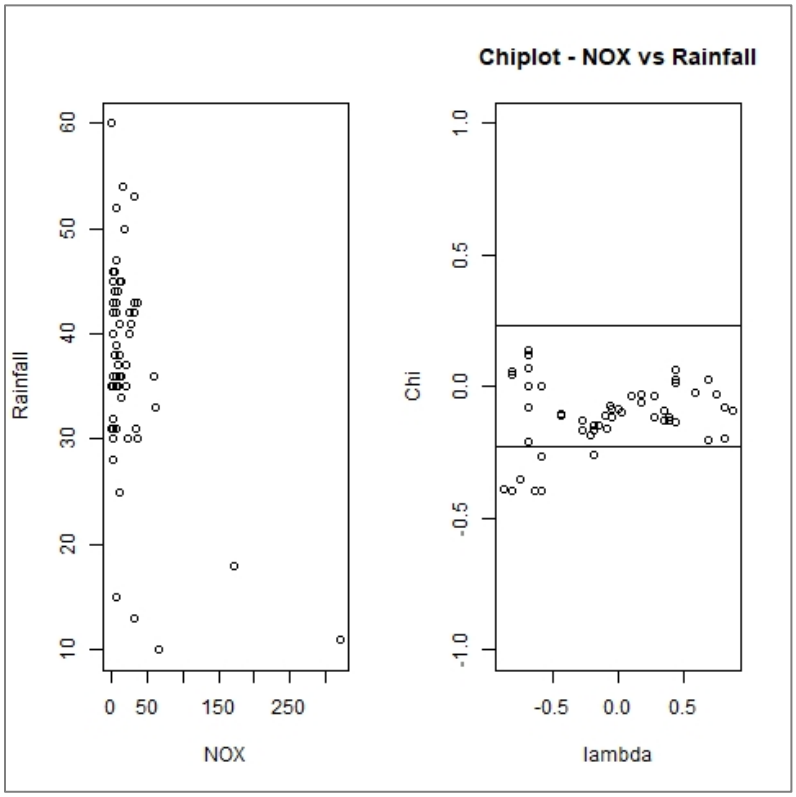


Description:

Similar to above Chi plot of (Education vs Population Density), Chi plot of Non-White vs Population have all observations inside the horizontal band which illustrates independence.

Scatterplot on the left also supports the argument as distribution is random with no dependency trend.

Chi Plot – NOX vs Rainfall



Description:

Chi plot of NOX vs Rainfall plots most of the observations inside the horizontal band which highlights independence however considerable number of observations also tends to depart from horizontal band from which we can say that variables are somehow dependent as well.

Scatterplot confirms that even when there is low amount of NOX, rainfall still have a higher number.

R Commands:

```
jpeg(file="E:\\question04-1.jpg",quality = 100)
par(mfrow=c(1,1))
chiplot(SO2,Mortality,vlabs=c("SO2","Mortality"))
title("Chiplot - SO2 vs Mortality",lwd=2)
dev.off()

jpeg(file="E:\\question04-2.jpg",quality = 100)
par(mfrow=c(1,1))
chiplot(NOX,Mortality,vlabs=c("NOX","Mortality"))
title("Chiplot - NOX vs Mortality",lwd=2)
dev.off()

jpeg(file="E:\\question04-3.jpg",quality = 100)
par(mfrow=c(1,1))
chiplot(Education,Popden,vlabs=c("Education","Population Density"))
title("Chiplot - Education vs Population Density",lwd=2)
dev.off()

jpeg(file="E:\\question04-4.jpg",quality = 100)
par(mfrow=c(1,1))
chiplot(Nonwhite,Popden,vlabs=c("Nonwhite","Population Density"))
title("Chiplot - Non White vs Population Density",lwd=2)
dev.off()

jpeg(file="E:\\question04-5.jpg",quality = 100)
par(mfrow=c(1,1))
chiplot(NOX,Rainfall,vlabs=c("NOX","Rainfall"))
title("Chiplot - NOX vs Rainfall",lwd=2)
dev.off()
```