Project 1

Code

JONATHAN FENG

PART 1

DATA IN

Hide

```
file <- "C:\\Users\\Taterthot\\Desktop\\da 410\\project1\\airpoll.txt"
airpol.full <- read.table(file, header = TRUE, sep = "", dec = ".")

airpol.full
```

City <chr>	Rainfall <int>	Education <dbl>	Popden <int>	Nonwhite <dbl>	N… <int>	SO2 <int>	Mortality <dbl>
akronOH	36	11.4	3243	8.8	15	59	921.9
albanyNY	35	11.0	4281	3.5	10	39	997.9
allenPA	44	9.8	4260	0.8	6	33	962.4
atlantGA	47	11.1	3125	27.1	8	24	982.3
baltimMD	43	9.6	6441	24.4	38	206	1071.0
birmhmAL	53	10.2	3325	38.5	32	72	1030.0
bostonMA	43	12.1	4679	3.5	32	62	934.7
bridgeCT	45	10.6	2140	5.3	4	4	899.5
bufaloNY	36	10.5	6582	8.1	12	37	1002.0
cantonOH	36	10.7	4213	6.7	7	20	912.3

1-10 of 60 rows Previous **1** 2 3 4 5 6 Next

Hide

```
city.names <-as.character(airpol.full[1:16,1])
airpol.data.sub <-airpol.full[1:16,2:8]

city.names
```

```
 [1] "akronOH"  "albanyNY" "allenPA"  "atlantGA" "baltimMD" "birmhmAL"
 [7] "bostonMA" "bridgeCT" "bufaloNY" "cantonOH" "chatagTN" "chicagIL"
[13] "cinnciOH" "clevelOH" "colombOH" "dallasTX"
```

DATA TO AIRPOL

Hide

```
city.names <-as.character(airpol.full[1:16,1])
airpol.data.sub <-airpol.full[1:16,2:8]

city.names
```

```
 [1] "akronOH"  "albanyNY" "allenPA"  "atlantGA" "baltimMD" "birmhmAL"
 [7] "bostonMA" "bridgeCT" "bufaloNY" "cantonOH" "chatagTN" "chicagIL"
[13] "cinnciOH" "clevelOH" "colombOH" "dallasTX"
```

PART 2

A

SAMPLE COVARIANCE MATRIX

Hide

```
airpol.cov <- round(var(airpol.data.sub),digits=2)
airpol.cov
```

```
          Rainfall Education     Popden Nonwhite      NOX      SO2
Rainfall     39.72     -2.46   -2766.77    34.61    -8.30   -84.69
Education    -2.46      0.61    -224.44    -2.25    -1.14   -16.09
Popden    -2766.77   -224.44 2229957.93 -1665.31 14294.73 71437.88
Nonwhite     34.61     -2.25   -1665.31   102.69    51.96   198.87
NOX          -8.30     -1.14   14294.73    51.96   268.60  1169.95
SO2         -84.69    -16.09   71437.88   198.87  1169.95  5981.26
Mortality   101.49    -24.96   47577.01   294.06   513.19  2502.85
          Mortality
Rainfall     101.49
Education    -24.96
Popden     47577.01
Nonwhite     294.06
NOX          513.19
SO2         2502.85
Mortality   3030.05
```

SAMPLE CORRELATION MATRIX

Hide

```
airpol.cor <- round(cor(airpol.data.sub),digits=2)
airpol.cor
```

```
          Rainfall Education Popden Nonwhite   NOX   SO2 Mortality
Rainfall      1.00     -0.50  -0.29     0.54 -0.08 -0.17      0.29
Education    -0.50      1.00  -0.19    -0.28 -0.09 -0.27     -0.58
Popden       -0.29     -0.19   1.00    -0.11  0.58  0.62      0.58
Nonwhite      0.54     -0.28  -0.11     1.00  0.31  0.25      0.53
NOX          -0.08     -0.09   0.58     0.31  1.00  0.92      0.57
SO2          -0.17     -0.27   0.62     0.25  0.92  1.00      0.59
Mortality     0.29     -0.58   0.58     0.53  0.57  0.59      1.00
```

RELATIONSHIP [STRENGTH X DIRECTION]

Hide

```
D.minus.12 <- diag( 1/sqrt(diag(airpol.cov) ) )
my.R <- D.minus.12 %*% airpol.cov %*% D.minus.12
my.R
```

```
           [,1]        [,2]       [,3]       [,4]        [,5]
[1,]  1.0000000 -0.49976469 -0.2939816  0.5419172 -0.08035640
[2,] -0.4997647  1.00000000 -0.1924363 -0.2842847 -0.08906087
[3,] -0.2939816 -0.19243630  1.0000000 -0.1100481  0.58408307
[4,]  0.5419172 -0.28428472 -0.1100481  1.0000000  0.31286146
[5,] -0.0803564 -0.08906087  0.5840831  0.3128615  1.00000000
[6,] -0.1737525 -0.26637581  0.6185629  0.2537516  0.92303455
[7,]  0.2925457 -0.58057023  0.5787939  0.5271655  0.56885330
           [,6]       [,7]
[1,] -0.1737525  0.2925457
[2,] -0.2663758 -0.5805702
[3,]  0.6185629  0.5787939
[4,]  0.2537516  0.5271655
[5,]  0.9230345  0.5688533
[6,]  1.0000000  0.5879137
[7,]  0.5879137  1.0000000
```

I would describe [5,6] only having a strong correlation in terms of pairs, and those variables would be NOX and SO2.The direction is strong and the direction is positive. The rest of the variables either have mostly negative correlation and weak strength, and a few having positive correlation and weak strength. NOX and SO2 was the only notable one having a correlation above 0.80.

B

DISTANCE MATRIX

Hide

```
dis <- dist(airpol.data.sub)

dist2full<-function(ds){
n<-attr(ds,"Size")
 full<-matrix(0,n,n)
full[lower.tri(full)]<-ds
full+t(full)
}

std <- sapply(airpol.data.sub, sd)
airpol.std <- sweep(airpol.data.sub,2,std,FUN="/")

dis <- dist(airpol.std)
dis.matrix<-dist2full(dis)
round(dis.matrix,digits=2)
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
 [1,] 0.00 1.76 2.80 2.84 5.14 4.81 2.09 2.21 2.92  1.34  4.15
 [2,] 1.76 0.00 2.22 3.13 4.53 4.88 2.62 2.90 1.74  1.67  3.97
 [3,] 2.80 2.22 0.00 3.23 4.51 4.56 3.42 2.18 2.45  2.03  2.98
 [4,] 2.84 3.13 3.23 0.00 4.53 2.61 3.42 2.82 3.57  3.08  2.29
 [5,] 5.14 4.53 4.51 4.53 0.00 3.62 5.05 5.87 3.74  5.11  4.40
 [6,] 4.81 4.88 4.56 2.61 3.62 0.00 4.91 4.75 4.81  5.02  2.49
 [7,] 2.09 2.62 3.42 3.42 5.05 4.91 0.00 3.25 3.21  2.72  4.77
 [8,] 2.21 2.90 2.18 2.82 5.87 4.75 3.25 0.00 3.86  2.03  3.23
 [9,] 2.92 1.74 2.45 3.57 3.74 4.81 3.21 3.86 0.00  2.32  4.25
[10,] 1.34 1.67 2.03 3.08 5.11 5.02 2.72 2.03 2.32  0.00  4.01
[11,] 4.15 3.97 2.98 2.29 4.40 2.49 4.77 3.23 4.25  4.01  0.00
[12,] 5.00 4.85 5.69 5.73 3.14 5.46 4.63 6.54 4.56  5.46  6.37
[13,] 2.40 2.38 2.40 2.94 3.00 3.62 2.98 3.17 2.57  2.51  3.29
[14,] 1.42 1.59 2.92 2.46 4.27 4.02 2.64 2.88 2.68  2.09  3.60
[15,] 1.41 1.71 3.16 2.50 5.13 4.69 2.12 2.87 2.57  1.87  4.24
[16,] 2.16 3.57 4.21 3.52 6.96 5.77 3.71 2.55 4.73  2.68  4.95
      [,12] [,13] [,14] [,15] [,16]
 [1,]  5.00  2.40  1.42  1.41  2.16
 [2,]  4.85  2.38  1.59  1.71  3.57
 [3,]  5.69  2.40  2.92  3.16  4.21
 [4,]  5.73  2.94  2.46  2.50  3.52
 [5,]  3.14  3.00  4.27  5.13  6.96
 [6,]  5.46  3.62  4.02  4.69  5.77
 [7,]  4.63  2.98  2.64  2.12  3.71
 [8,]  6.54  3.17  2.88  2.87  2.55
 [9,]  4.56  2.57  2.68  2.57  4.73
[10,]  5.46  2.51  2.09  1.87  2.68
[11,]  6.37  3.29  3.60  4.24  4.95
[12,]  0.00  3.60  4.38  5.25  6.88
[13,]  3.60  0.00  1.94  2.99  4.23
[14,]  4.38  1.94  0.00  1.74  3.05
[15,]  5.25  2.99  1.74  0.00  2.68
[16,]  6.88  4.23  3.05  2.68  0.00
```

Hide

```
city.names
```

```
 [1] "akronOH"  "albanyNY" "allenPA"  "atlantGA" "baltimMD"
 [6] "birmhmAL" "bostonMA" "bridgeCT" "bufaloNY" "cantonOH"
[11] "chatagTN" "chicagIL" "cinnciOH" "clevelOH" "colombOH"
[16] "dallasTX"
```

Let me say that any distance that is less than 1.5 would be considered most similar and any distance above 6 would be considered most different.

Thus we get similar: [10, 1], [14, 1], [15, 1] and those being the cities of [Canton, OH | Akron, OH], [Cleveland, Akron], [Colombia, Akron].

For the most different: [5, 16], [11, 12], [12, 10], [16, 12]. And to those respectively [Baltimore, Dallas], [Chatag, Chicago], [Chicago, Canton], and [Dallas, Chicago].

C

NORMALITY CHECK CHI-SQUARED PLOT

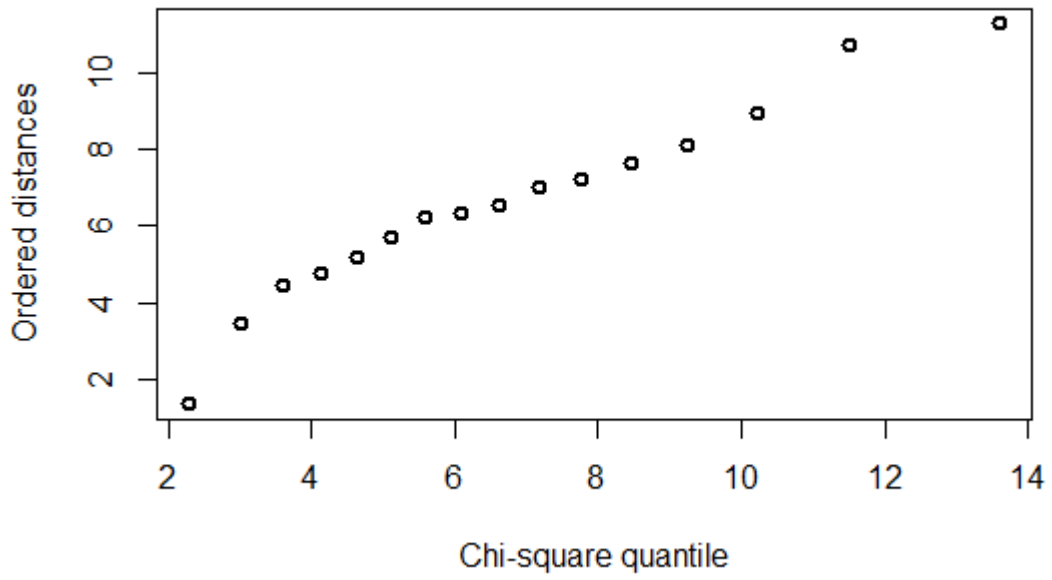
Hide

```
chisplot <- function(x) {
if (!is.matrix(x)) stop("x is not a matrix")

 n <- nrow(x)
 p <- ncol(x)
 xbar <- apply(x, 2, mean)
 S <- var(x)
 S <- solve(S)
 index <- (1:n)/(n+1)
 xcent <- t(t(x) - xbar)
 di <- apply(xcent, 1, function(x,S) x %*% S %*% x,S)

 quant <- qchisq(index,p)
 plot(quant, sort(di), ylab = "Ordered distances",
 xlab = "Chi-square quantile", lwd=2,pch=1)
}

chisplot(as.matrix(airpol.data.sub))
```



Since we have the chi-square quantiles being displayed on the horizontal, we are checking if there are some bends or uneven parts to the data, straight line meaning that it is normally distributed. For the most part besides the second quantile and maybe the 14th, we can say that it follows a pretty straight line with little to no deviation meaning the ordered distances are normal, thus we can say the data used is approximately normal and comes from a multivariate normal distribution.