Project 1: Air Pollution

DA 410

Marjorie Blanco

Part 1:

Download `airpoll.txt` in this problem, we will only focus on the first 16 observations (cities). Read the data into R (as a data frame) and name the data as `airpol.full`

```
airpol.full <- read_table2("MV/airpoll.txt")
```

Then use the following code to "extract" the first 16 observations.

```
airpol.full <- airpol.full %>% remove_rownames %>% column_to_rownames(var="City")
airpol.data.sub <- airpol.full[1:16,1:7]
```

Display the subset data `airpol.data.sub`

	Rainfall	Education	Popden	Nonwhite	NOX	SO2	Mortality
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
chatagTN	52	9.6	2302	22.2	8	27	1018.0
chicagIL	33	10.9	6122	16.3	63	278	1025.0
cinnciOH	40	10.2	4101	13.0	26	146	970.5
clevelOH	35	11.1	3042	14.7	21	64	986.0
colombOH	37	11.9	4259	13.1	9	15	958.8
dallasTX	35	11.8	1441	14.8	1	1	860.1

Part 2:

Use R to perform the following analysis on the subset data `airpol.data.sub`. Make sure you include clear headings, command lines, and relevant output/results.

a) Calculate the sample covariance matrix and the sample correlation matrix. Identify which pairs of variables seem to be strongly associated, and describe the nature (strength and direction) of the relationship between these variable pairs.

Sample covariance matrix

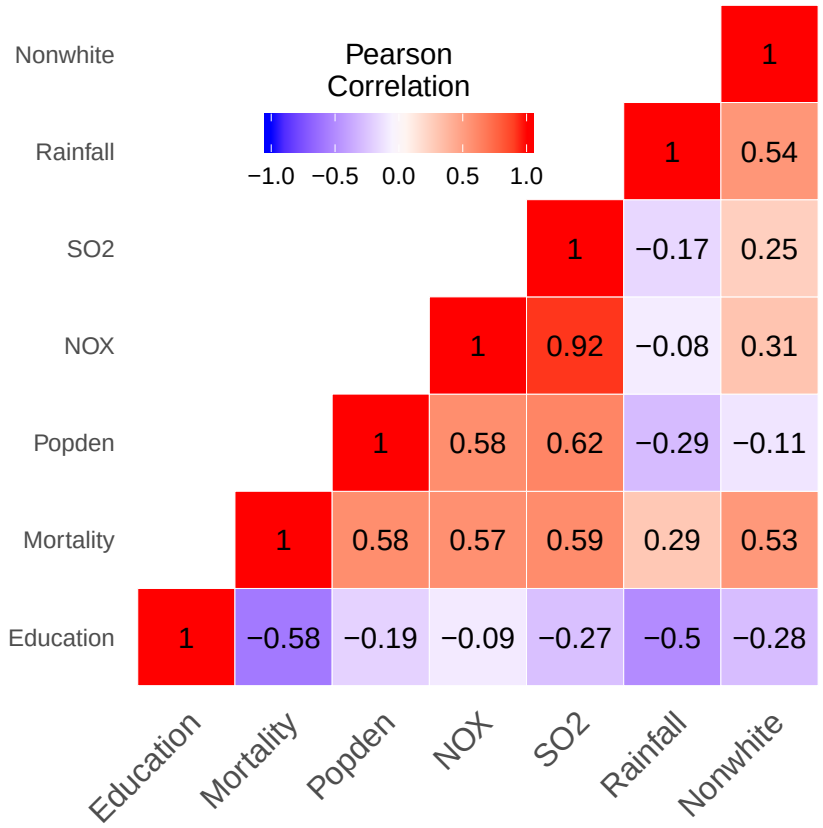
	Rainfall	Education	Popden	Nonwhite	NOX	SO2	Mortality
Rainfall	39.72	-2.46	-2766.77	34.61	-8.30	-84.69	101.49
Education	-2.46	0.61	-224.44	-2.25	-1.14	-16.09	-24.96
Popden	-2766.77	-224.44	2229957.93	-1665.31	14294.73	71437.88	47577.01
Nonwhite	34.61	-2.25	-1665.31	102.69	51.96	198.87	294.06
NOX	-8.30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

Sample correlation matrix

Range	Relationship
–1	A perfect downhill (negative) linear relationship
-0.99 to -0.70	A strong downhill (negative) linear relationship
-0.69 to -0.50	A moderate downhill (negative) relationship
-0.49 to -0.30	A weak downhill (negative) linear relationship
-0.29 to 0.29	No linear relationship
0.30 to 0.49	A weak uphill (positive) linear relationship
0.50 to 0.69	A moderate uphill (positive) relationship
0.70 to 0.99	A strong uphill (positive) linear relationship
1	A perfect uphill (positive) linear relationship

```
cormat <- round(cor(airpol.data.sub), 2)
melted_cormat <- melt(cormat)
```

	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



Var1	Var2	Freq	Relationship	Strength
SO2	NOX	0.92	Positive	strong
SO2	Popden	0.62	Positive	moderate
Mortality	SO2	0.59	Positive	moderate
NOX	Popden	0.58	Positive	moderate
Mortality	Popden	0.58	Positive	moderate
Mortality	NOX	0.57	Positive	moderate
Nonwhite	Rainfall	0.54	Positive	moderate
Mortality	Nonwhite	0.53	Positive	moderate
NOX	Nonwhite	0.31	Positive	weak
Mortality	Rainfall	0.29	Positive	None
SO2	Nonwhite	0.25	Positive	None
NOX	Rainfall	-0.08	Negative	None
NOX	Education	-0.09	Negative	None
Nonwhite	Popden	-0.11	Negative	None
SO2	Rainfall	-0.17	Negative	None
Popden	Education	-0.19	Negative	None
SO2	Education	-0.27	Negative	None
Nonwhite	Education	-0.28	Negative	None
Popden	Rainfall	-0.29	Negative	None
Education	Rainfall	-0.50	Negative	moderate
Mortality	Education	-0.58	Negative	moderate

`NOX` and `SO2` are the most positively strongly associated. The correlation coefficient is 0.92.

b) Calculate the distance matrix for these observations (after scaling the variables by dividing each variable

by its standard deviation). Describe some of the most similar pairs of cities and some of the most different pairs of cities, giving evidence from the distance matrix.

```
# finding standard deviations of variables
std <-sapply(airpol.data.sub, sd)

# dividing each variable by its standard deviation
airpol.data.sub.std <-sweep(airpol.data.sub, 2, std, FUN = "/")

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

rownames(dis.matrix) <- rownames(airpol.data.sub)
colnames(dis.matrix) <- rownames(airpol.data.sub)
```

```
## pdf
##   2
```

```
kable(dis.matrix, digits = 2)  %>%
  kable_styling("striped", full_width = T, font_size = 8) %>%
  row_spec(0, angle = -45)
```

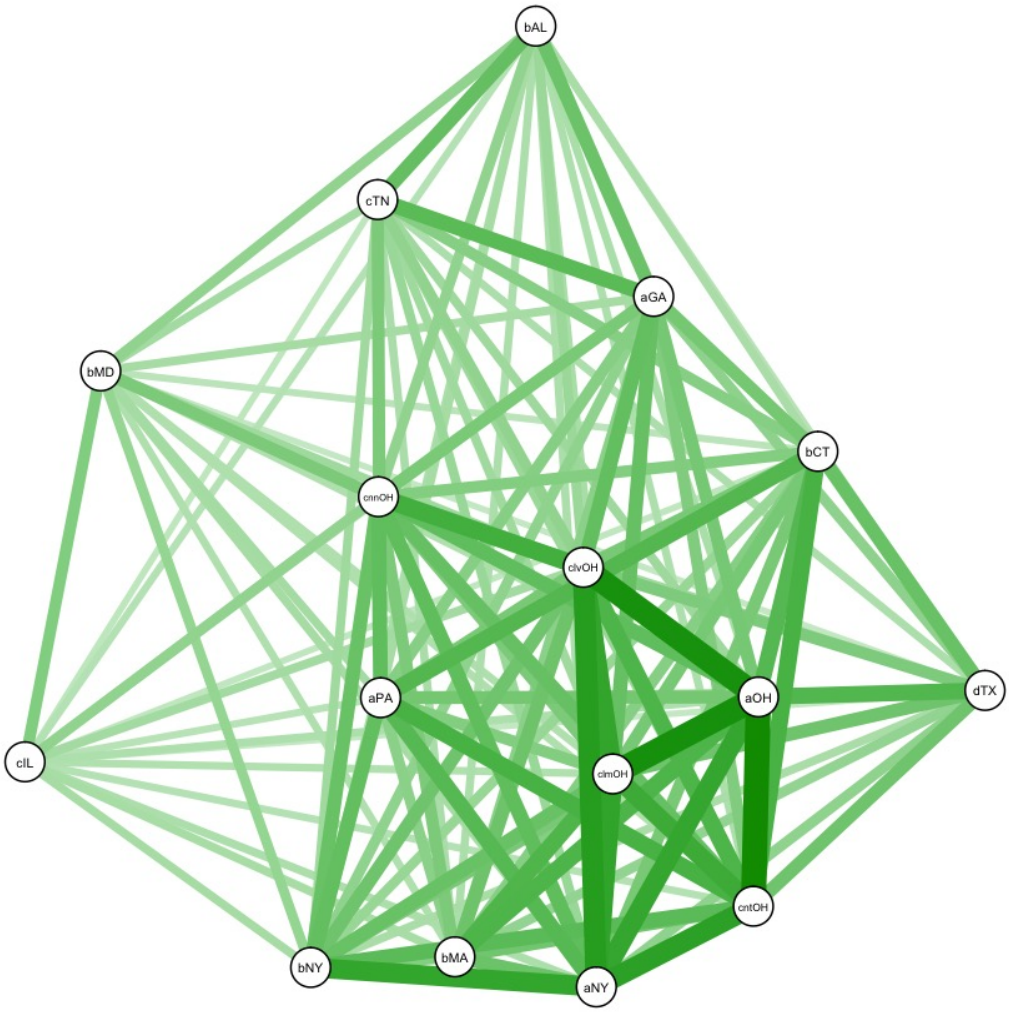
	akronOH	albanyNY	allenPA	atlantGA	baltimMD	birmhmAL	bostonMA	bridgeCT	bufaloNY	cantonOH	chatagTN	chicagIL	cinnciOH	clevelOH	colombOH	dallasTX
akronOH	0.00	1.76	2.80	2.84	5.14	4.81	2.09	2.21	2.92	1.34	4.15	5.00	2.40	1.42	1.41	2.16
albanyNY	1.76	0.00	2.22	3.13	4.53	4.88	2.62	2.90	1.74	1.67	3.97	4.85	2.38	1.59	1.71	3.57
allenPA	2.80	2.22	0.00	3.23	4.51	4.56	3.42	2.18	2.45	2.03	2.98	5.69	2.40	2.92	3.16	4.21
atlantGA	2.84	3.13	3.23	0.00	4.53	2.61	3.42	2.82	3.57	3.08	2.29	5.73	2.94	2.46	2.50	3.52
baltimMD	5.14	4.53	4.51	4.53	0.00	3.62	5.05	5.87	3.74	5.11	4.40	3.14	3.00	4.27	5.13	6.96
birmhmAL	4.81	4.88	4.56	2.61	3.62	0.00	4.91	4.75	4.81	5.02	2.49	5.46	3.62	4.02	4.69	5.77
bostonMA	2.09	2.62	3.42	3.42	5.05	4.91	0.00	3.25	3.21	2.72	4.77	4.63	2.98	2.64	2.12	3.71
bridgeCT	2.21	2.90	2.18	2.82	5.87	4.75	3.25	0.00	3.86	2.03	3.23	6.54	3.17	2.88	2.87	2.55
bufaloNY	2.92	1.74	2.45	3.57	3.74	4.81	3.21	3.86	0.00	2.32	4.25	4.56	2.57	2.68	2.57	4.73
cantonOH	1.34	1.67	2.03	3.08	5.11	5.02	2.72	2.03	2.32	0.00	4.01	5.46	2.51	2.09	1.87	2.68
chatagTN	4.15	3.97	2.98	2.29	4.40	2.49	4.77	3.23	4.25	4.01	0.00	6.37	3.29	3.60	4.24	4.95
chicagIL	5.00	4.85	5.69	5.73	3.14	5.46	4.63	6.54	4.56	5.46	6.37	0.00	3.60	4.38	5.25	6.88
cinnciOH	2.40	2.38	2.40	2.94	3.00	3.62	2.98	3.17	2.57	2.51	3.29	3.60	0.00	1.94	2.99	4.23
clevelOH	1.42	1.59	2.92	2.46	4.27	4.02	2.64	2.88	2.68	2.09	3.60	4.38	1.94	0.00	1.74	3.05
colombOH	1.41	1.71	3.16	2.50	5.13	4.69	2.12	2.87	2.57	1.87	4.24	5.25	2.99	1.74	0.00	2.68
dallasTX	2.16	3.57	4.21	3.52	6.96	5.77	3.71	2.55	4.73	2.68	4.95	6.88	4.23	3.05	2.68	0.00

```
dis.matrix.m <- dis.matrix.m %>% arrange(value) #%>% select(-rescale)
dis.matrix.m <- dis.matrix.m[seq(1, nrow(dis.matrix.m), 2),]
dis.matrix.m <- dis.matrix.m %>% filter(value != 0) %>% filter(value <= 1.67 | value >= 5.87)
colnames(dis.matrix.m) <- c("City 1", "City 2", "Distance", "Scale")
dis.matrix.m$Note <- c(rep("Most similar", 5),  rep("Most different", 5))
kable(dis.matrix.m, digits = 2)  %>%
  kable_styling("striped", full_width = T)
```

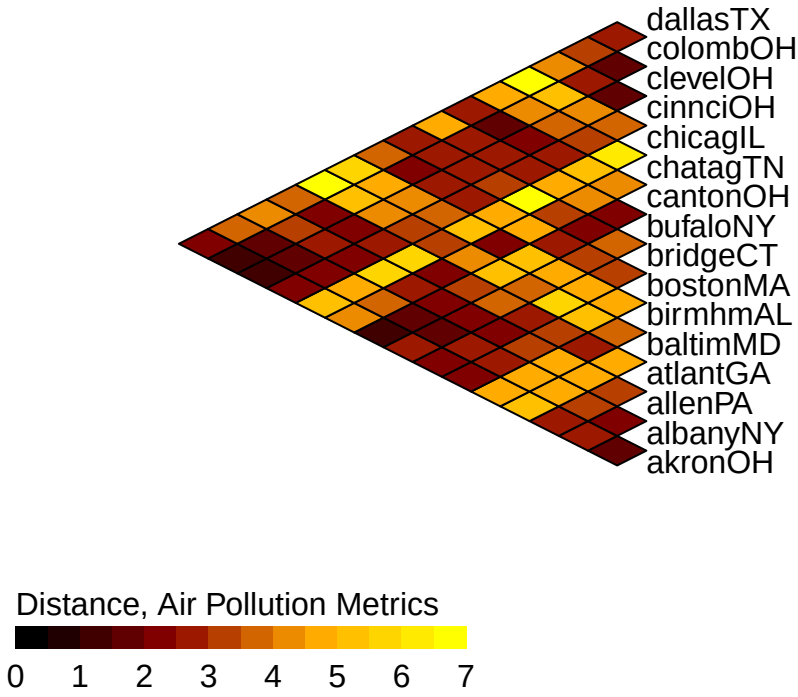
City 1	City 2	Distance	Scale	Note
cantonOH	akronOH	1.34	0.26	Most similar
colombOH	akronOH	1.41	0.27	Most similar
clevelOH	akronOH	1.42	0.28	Most similar
clevelOH	albanyNY	1.59	0.33	Most similar
cantonOH	albanyNY	1.67	0.34	Most similar
bridgeCT	baltimMD	5.87	0.84	Most different
chicagIL	chatagTN	6.37	1.00	Most different
chicagIL	bridgeCT	6.54	1.00	Most different
dallasTX	chicagIL	6.88	1.00	Most different
dallasTX	baltimMD	6.96	1.00	Most different

```
dist_m <- as.matrix(dist(airpol.data.sub.std))
dist_mi <- 1/dist_m # one over, as qgraph takes similarity matrices as input
library(qgraph)
jpeg('airpol_forcedraw.jpg', width=1000, height=1000, unit='px')
qgraph(dist_mi, layout='spring', vsize=3, label.cex = 1)
dev.off()
```

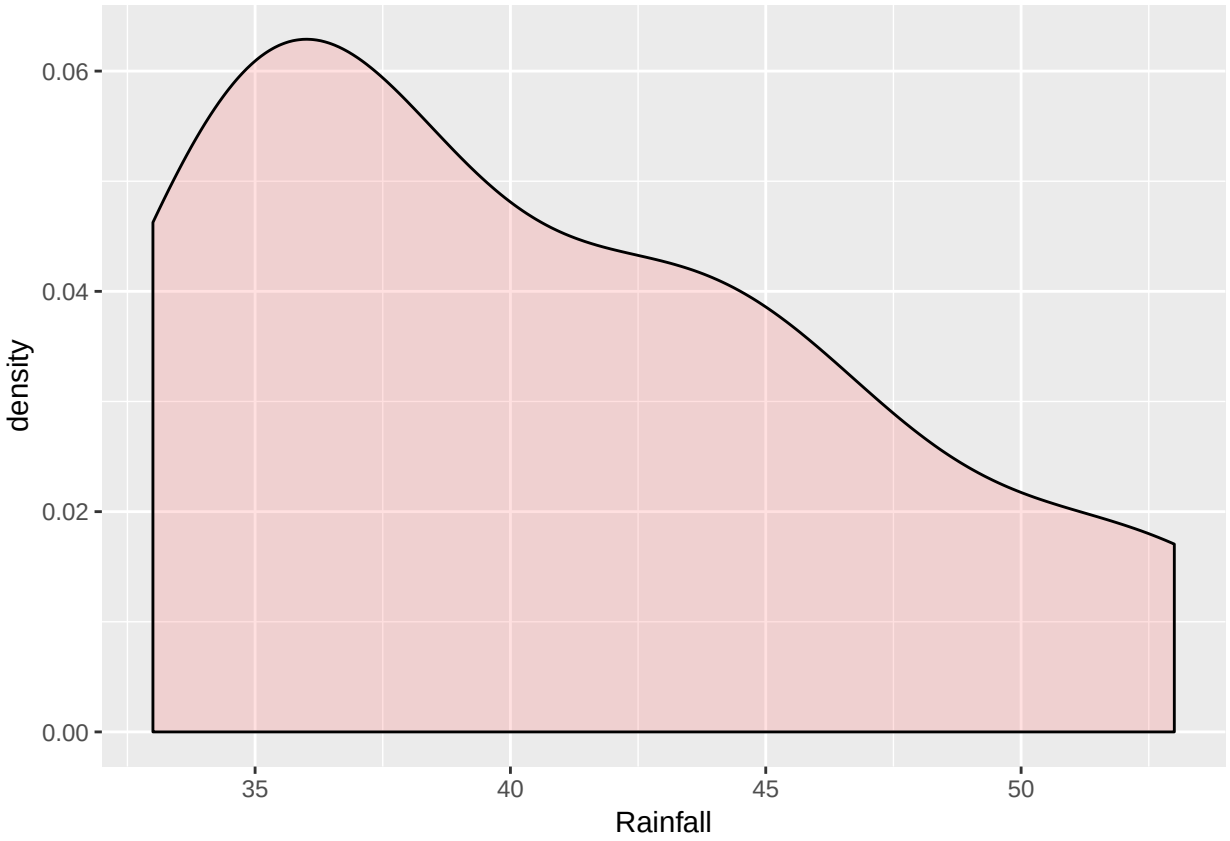
```
## pdf
##   2
```

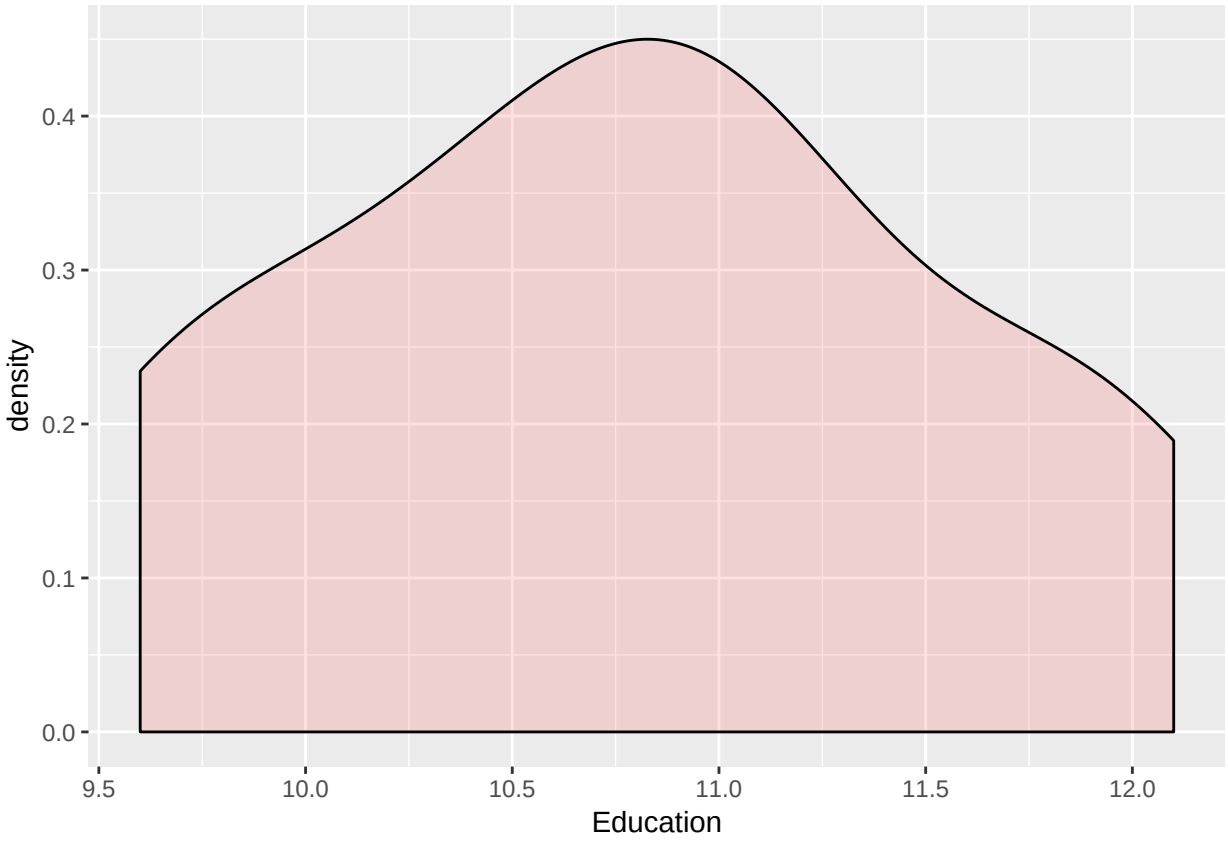


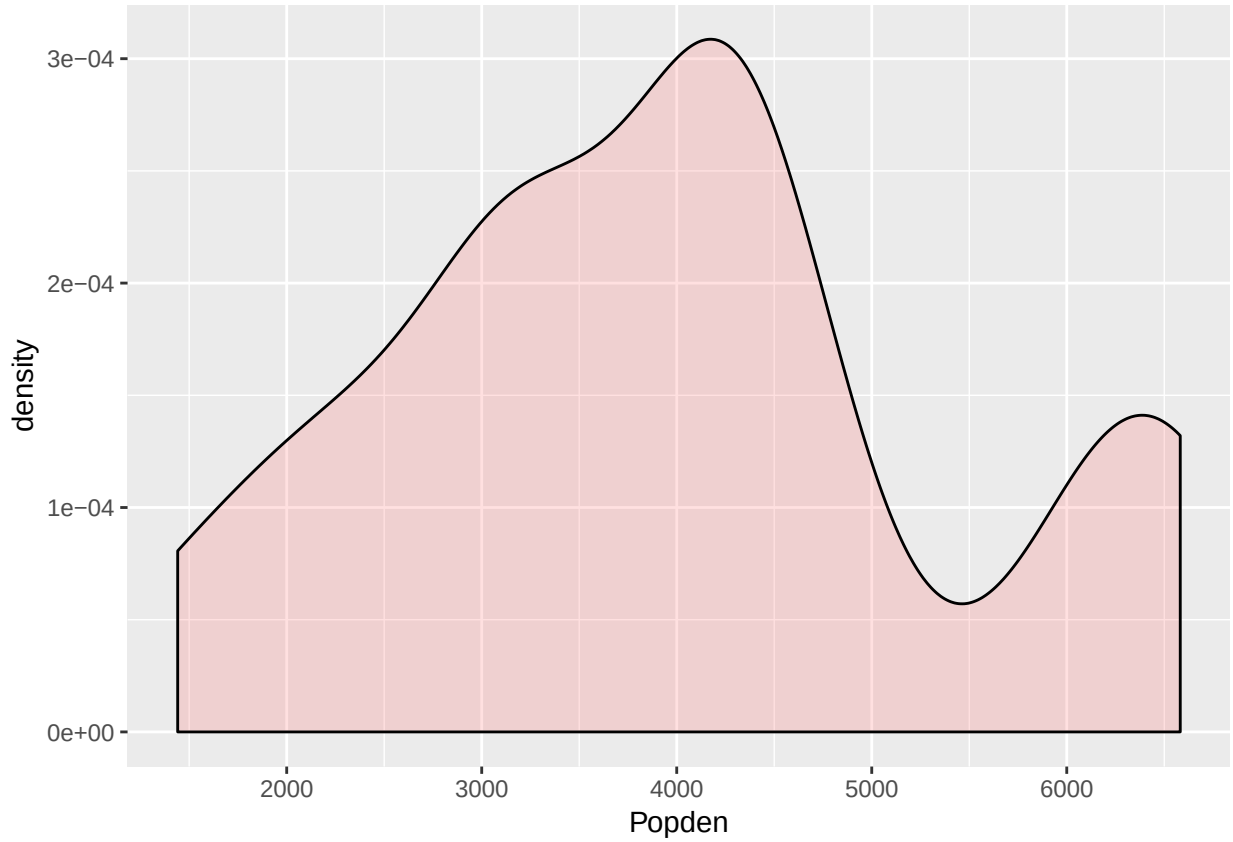

```
distogram(dist_m, title = 'Distance, Air Pollution Metrics')
```

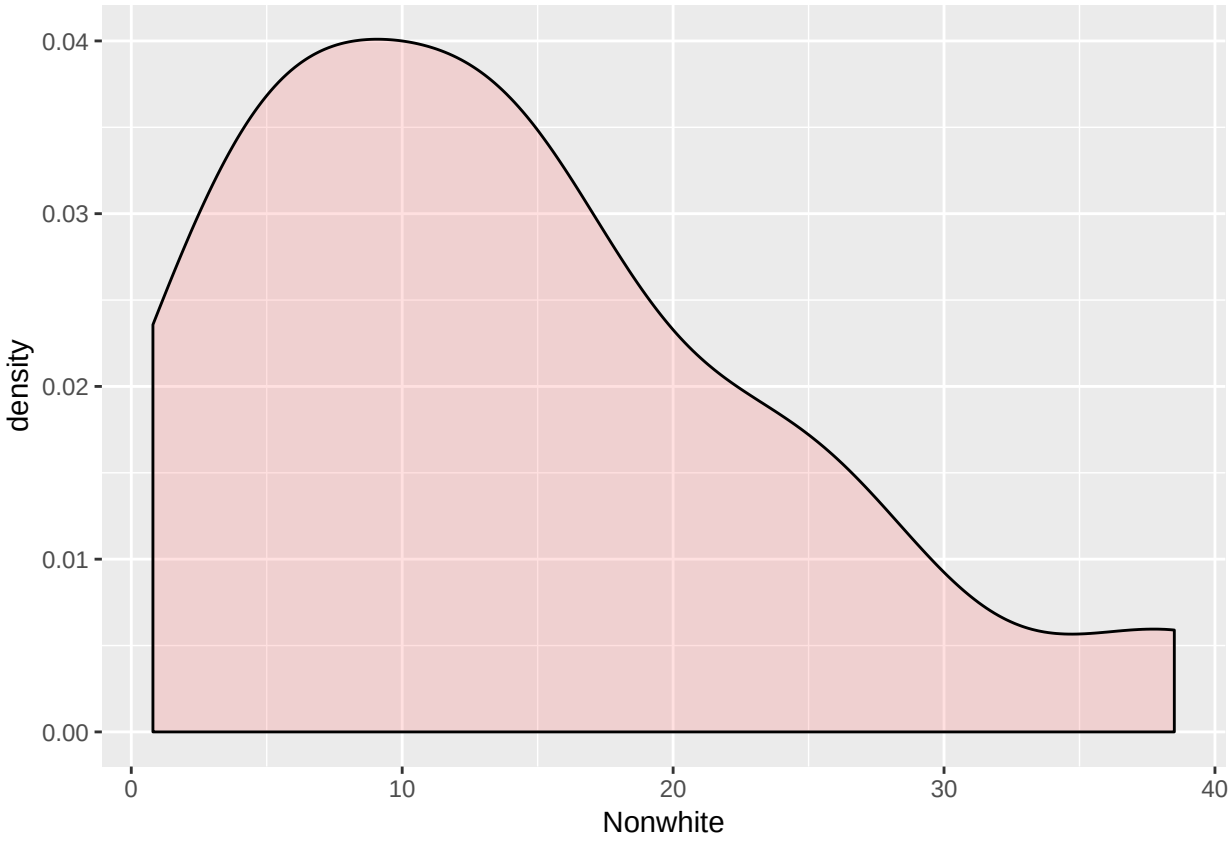


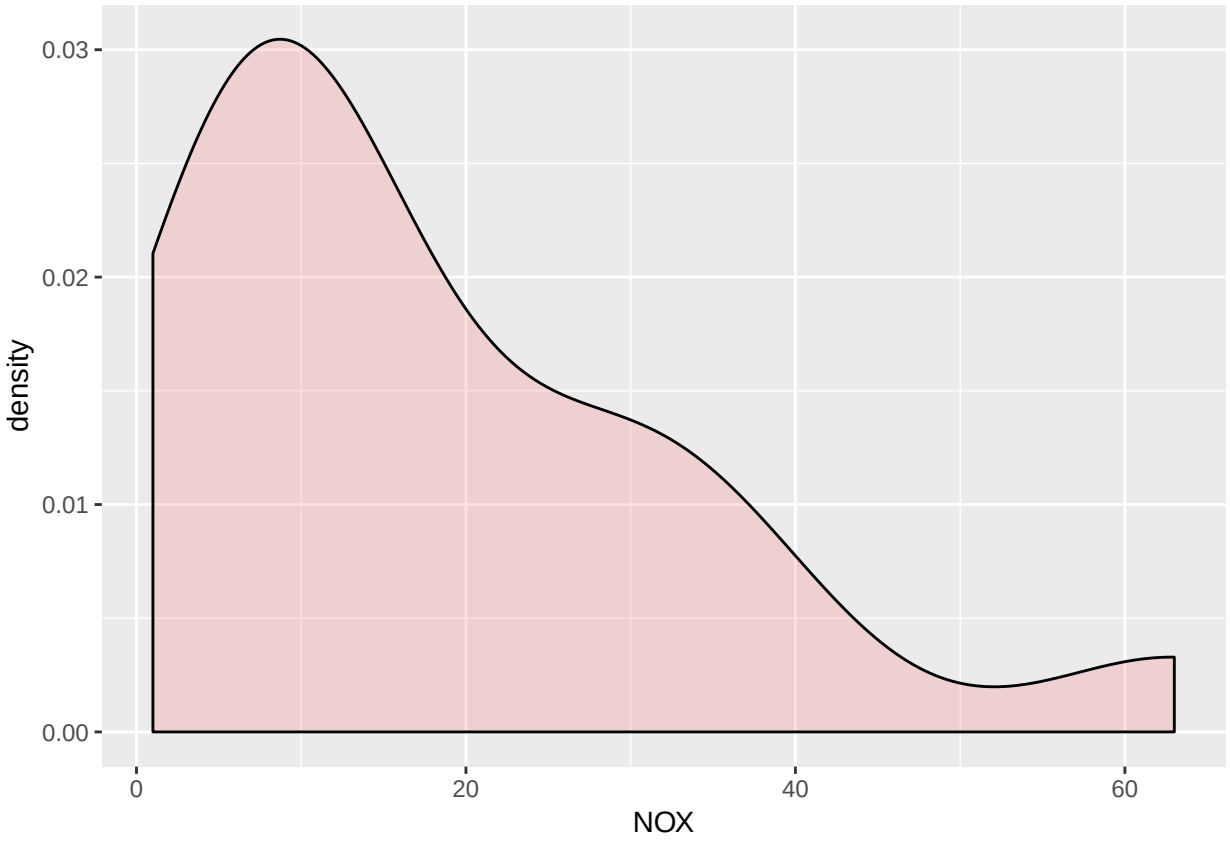
c) Display a plot that will help assess whether this data set comes from a multivariate normal distribution. What is your conclusion based on the plot?

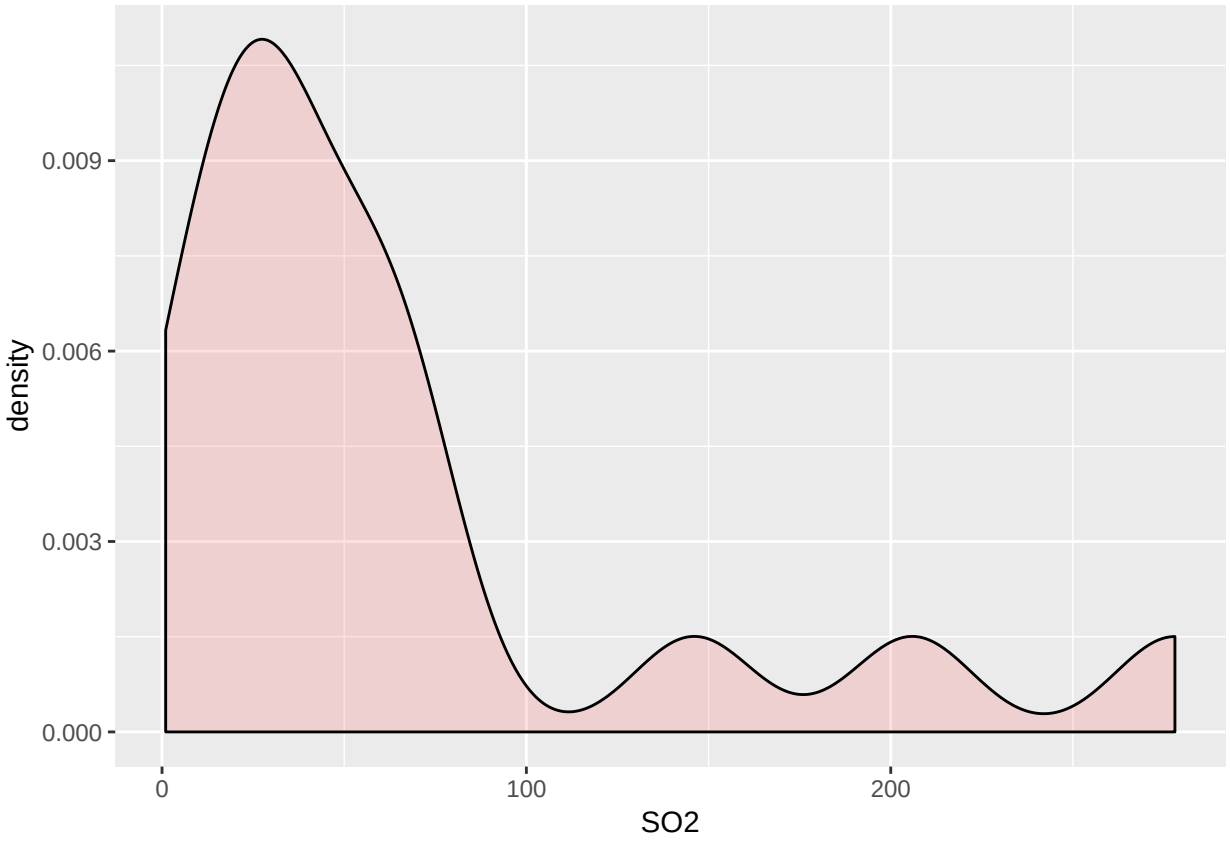

0.06
0.04
0.02
0.00
density
35
40
45
50
Rainfall


density
0.4
0.3
0.2
0.1
0.0
9.5
10.0
10.5
11.0
11.5
12.0
Education


3e−04
2e−04
1e−04
0e+00
density
2000
3000
4000
5000
6000
Popden


0.04
0.03
0.02
0.01
0.00
density
0
10
20
30
40
Nonwhite


0.03
0.02
0.01
0.00
density
0
20
40
60
NOX

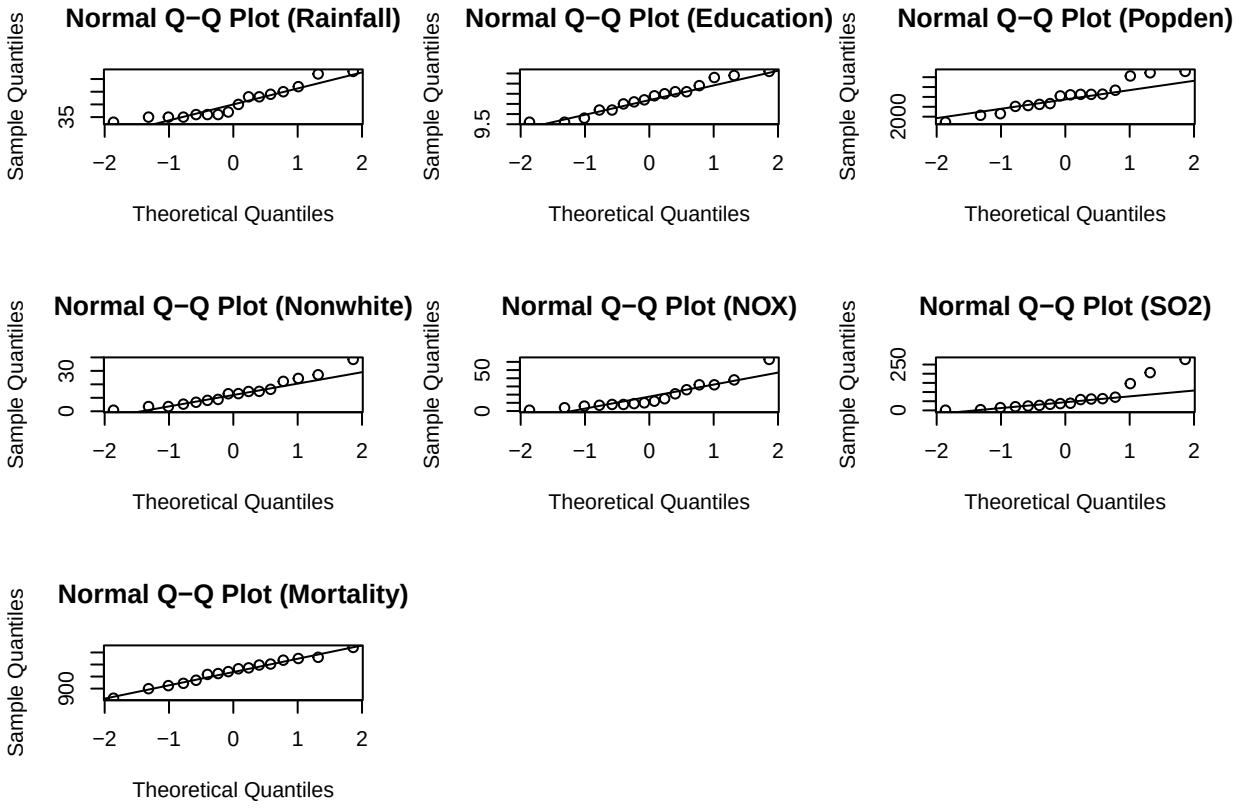

density
0.009
0.006
0.003
0.000
0
100
200
SO2

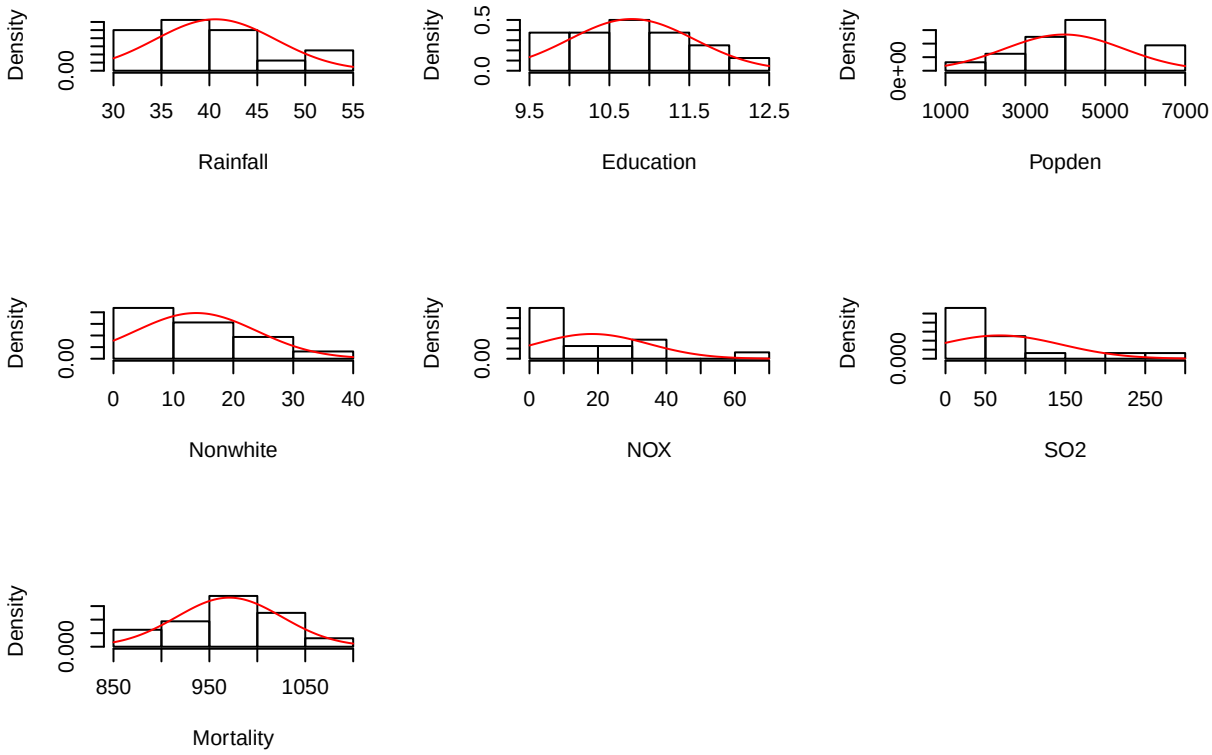

```
result <- mvn(data = airpol.data.sub, mvnTest = "royston")
result
```

```
## $multivariateNormality
##      Test        H     p value MVN
## 1 Royston 21.84198 0.001431724  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk Rainfall     0.8885    0.0527    YES
## 2 Shapiro-Wilk Education    0.9599    0.6607    YES
## 3 Shapiro-Wilk  Popden      0.9448    0.4119    YES
## 4 Shapiro-Wilk Nonwhite     0.9268    0.2170    YES
## 5 Shapiro-Wilk    NOX       0.8420    0.0104    NO
## 6 Shapiro-Wilk    SO2       0.7565    0.0008    NO
## 7 Shapiro-Wilk Mortality    0.9880    0.9975    YES
##
## $Descriptives
##             n       Mean      Std.Dev  Median    Min    Max    25th
## Rainfall   16   40.62500    6.3021160   38.50   33.0   53.0   35.75
## Education  16   10.78125    0.7841928   10.80    9.6   12.1   10.20
## Popden     16 3972.25000 1493.3043673 4157.00 1441.0 6582.0 3104.25
## Nonwhite   16   13.80000   10.1338377   13.05    0.8   38.5    6.35
## NOX        16   18.25000   16.3890207   11.00    1.0   63.0    7.75
## SO2        16   67.93750   77.3386223   38.00    1.0  278.0   23.00
## Mortality  16  970.77500   55.0459142  976.40  860.1 1071.0  931.50
```

```
##                75th        Skew   Kurtosis
## Rainfall     44.250  0.61253293 -1.0056510
## Education    11.175  0.05209694 -1.1994737
## Popden     4380.500  0.26464949 -0.9011197
## Nonwhite     17.775  0.81420631 -0.1455478
## NOX          27.500  1.24841241  0.8477042
## SO2          66.000  1.52756033  1.2053294
## Mortality 1006.000 -0.20570580 -0.7988297
```

```
# create univariate Q-Q plots
result <- mvn(data = airpol.data.sub, mvnTest = "royston", univariatePlot = "qqplot")
# create univariate histograms
result <- mvn(data = airpol.data.sub, mvnTest = "royston", univariatePlot = "histogram")
```





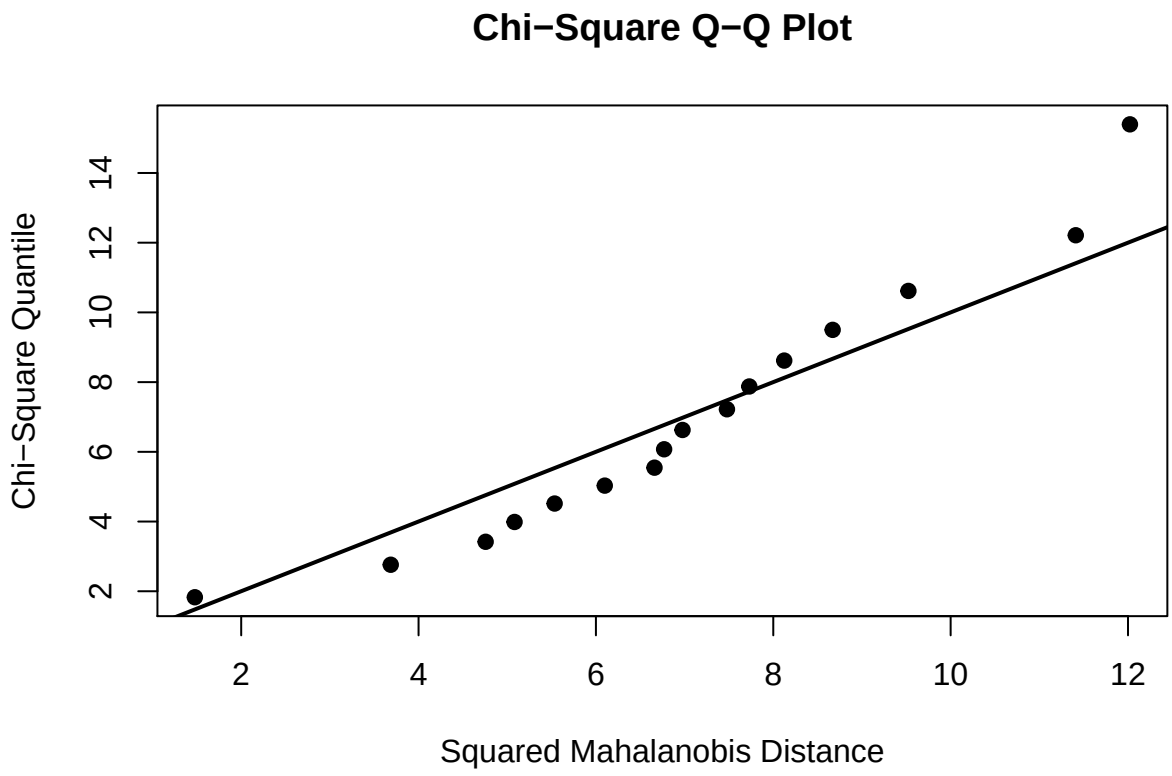

```
result <- mvn(data = airpol.data.sub, mvnTest = "hz")
result$multivariateNormality
```

```
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.8959774 0.2633274 YES
```

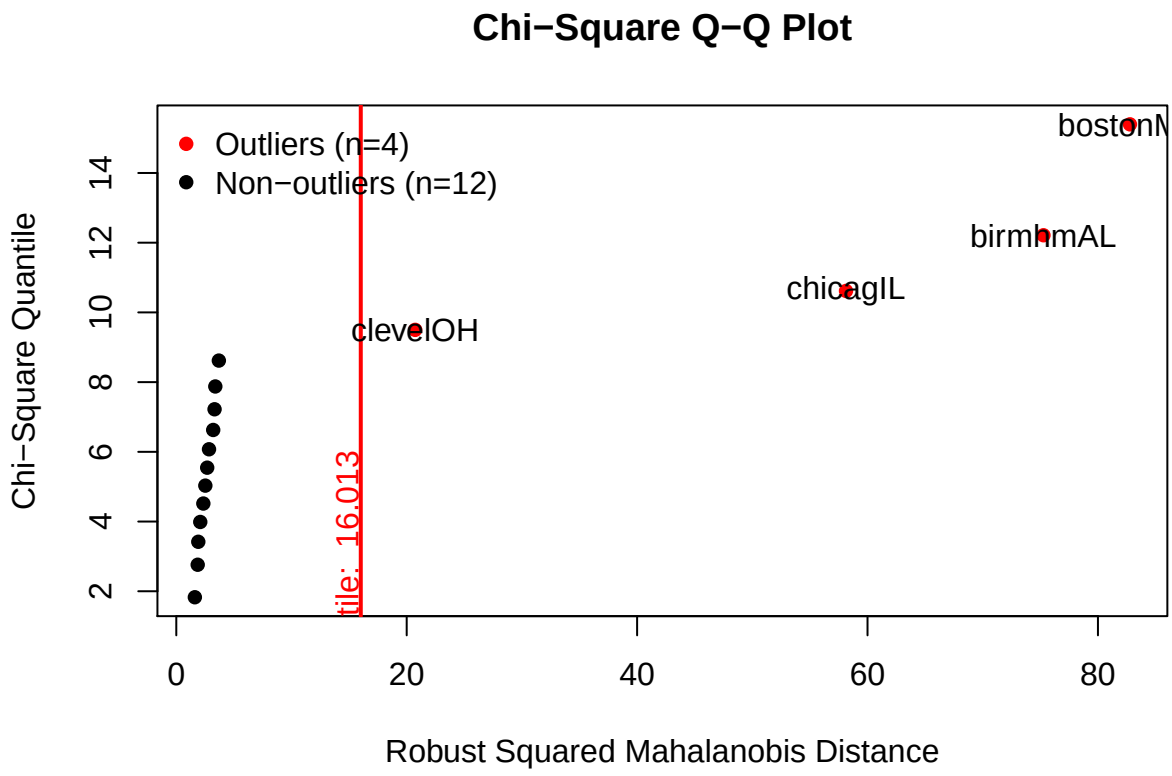
```
mn <- mult.norm(airpol.data.sub, chicrit=0.001)
mn$mult.test
```

```
##          Beta-hat    kappa      p-val
## Skewness 22.70107 60.53619 0.97502033
## Kurtosis 49.06386 -2.48306 0.01302591
```

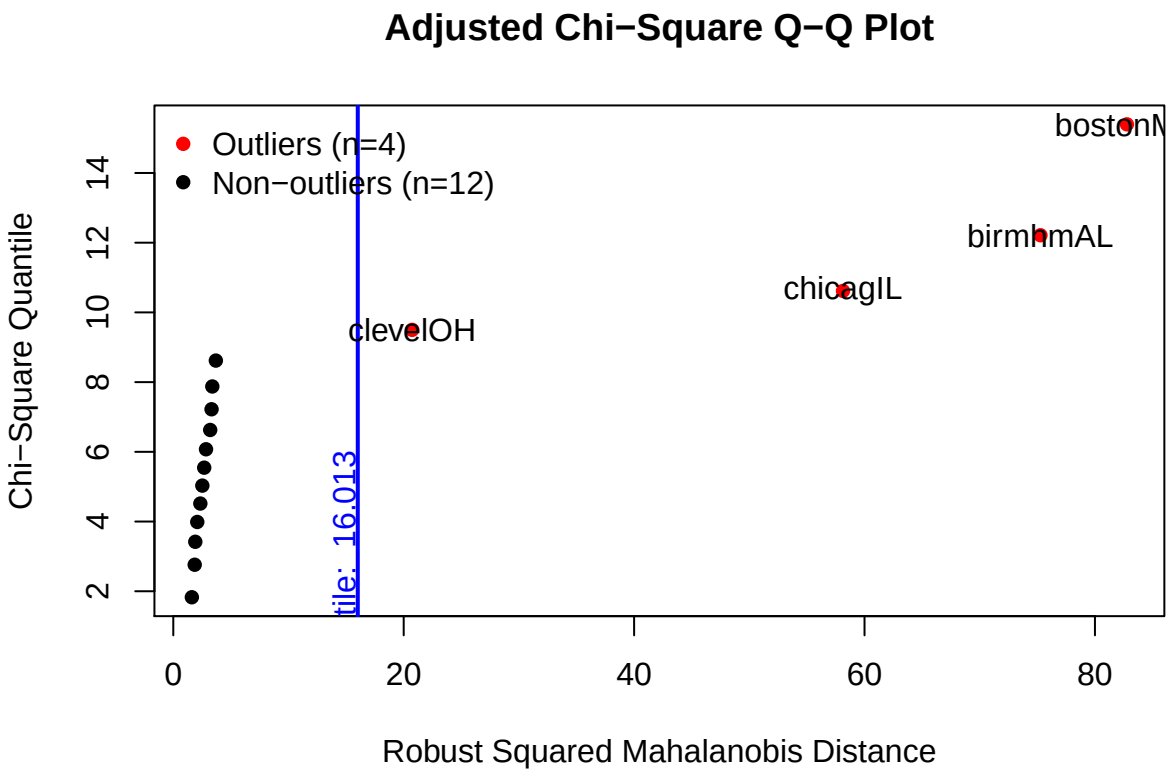
```
result <- mvn(data = airpol.data.sub, mvnTest = "hz", multivariatePlot = "qq")
```




```
# Mahalanobis distance
result <- mvn(data = airpol.data.sub, mvnTest = "hz", multivariateOutlierMethod = "quan")
```




```
# Adjusted Mahalanobis distance
result <- mvn(data = airpol.data.sub, mvnTest = "hz", multivariateOutlierMethod = "adj")
```



From the figure, Mahalanobis distance and adjusted Mahalanobis distance declares 4 observations as multivariate outlier.

According to the Henze-Zirkler's test results, dataset for airpol does not follow a multivariate normal distribution.