I. Objective:

This project aims to produce a lower dimensional representation of the dataset using two methods; Principal components analysis and Multidimensional Scaling. Principal Components Analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. We reduce our variables to the first m components that account for most of the variability. Multidimensional Scaling is also a dimension reduction technique but works on reverse order. If we have a matrix X of the dataset, we compute a matrix D whose elements are the distances (dissimilarities) between each pair of observations, then from this matrix we create another matrix Y by spectral decomposition that have less number of variables. I will begin first by plotting the data, then identify the outliers if there is any, then run both methods on my dataset, before and after removing the outliers.

II. Data Source:

Regression Analysis by Example, 5th Edition, Samprit Chatterjee and Ali S. Hadi, John Wiley & Sons. 2012

Data Description:

McDonald and Schwing (1 973) present a study that relates total mortality to climate, socioeconomic, and pollution variables. Fifteen predictor variables selected for the study are listed in Table 11.1 1 and there are 60 observations in the dataset. The response variable is the total age-adjusted mortality from all causes.

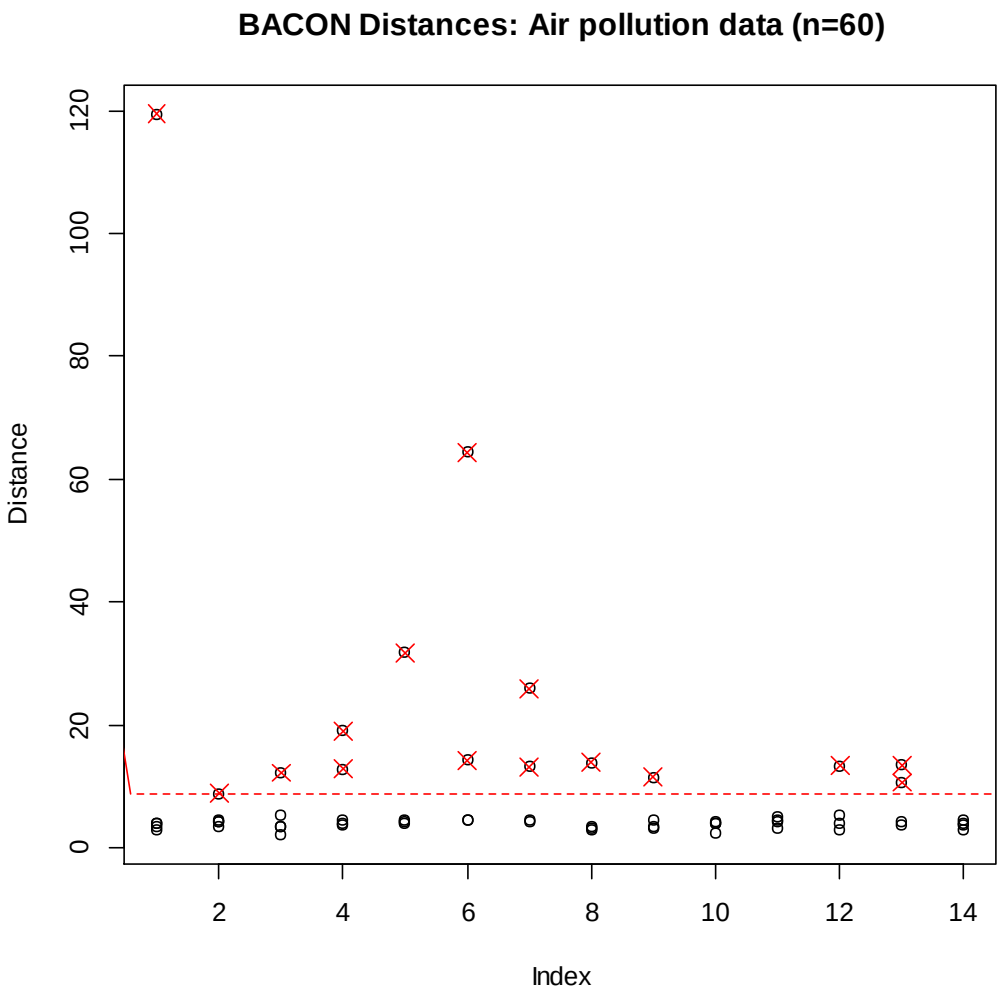
Table 11.11 Description of Variables, Means, and Standard Deviations, SD ($n = 60$)

Variable	Description	Mean	SD
X_1	Mean annual precipitation (inches)	37.37	9.98
X_2	Mean January temperature (degrees Fahrenheit)	33.98	10.17
X_3	Mean July temperature (degrees Fahrenheit)	74.58	4.76
X_4	Percent of population over 65 years of age	8.80	1.46
X_5	Population per household	3.26	0.14
X_6	Median school years completed	10.97	0.85
X_7	Percent of housing units that are sound	80.92	5.15
X_8	Population per square mile	3876.05	1454.10
X_9	Percent of nonwhite population	11.87	8.92
X_{10}	Percent employment in white-collar jobs	46.08	4.61
X_{11}	Percent of families with income under $3000	14.37	4.16
X_{12}	Relative pollution potential of hydrocarbons	37.85	91.98
X_{13}	Relative pollution potential of oxides of nitrogen	22.65	46.33
X_{14}	Relative pollution potential of sulfur dioxide	53.77	63.39
X_{15}	Percent relative humidity	57.67	5.37
Y	Total age-adjusted mortality from all causes.	940.36	62.21

III. Plot of the pairs of X

IV. Identification of outliers (BACON Method)

The BACON algorithm produced a basic subset of 45 observations, thus resulting in the appearance of 15 outlier observations 6,7,18,29,30,31,32,37,40,41,47,48,49,50,55 and the adjusted C(alpha) is 8.65.



I created a code that would allow me to see the index of the outlying observations and let me be able to remove them all at once from the dataset when applying the analysis after removing outliers.

Code:

```
#identifying indices of outliers
T=c()
for(i in output$dis)
  if(i > output$limit)
  {
      print(match(i, output$dis))
  }

#creating a vector for indices of outliers
T=c()
for(i in output$dis)
  if(i > output$limit)
  {
    T <- c(T, match(i, output$dis))
  }
```

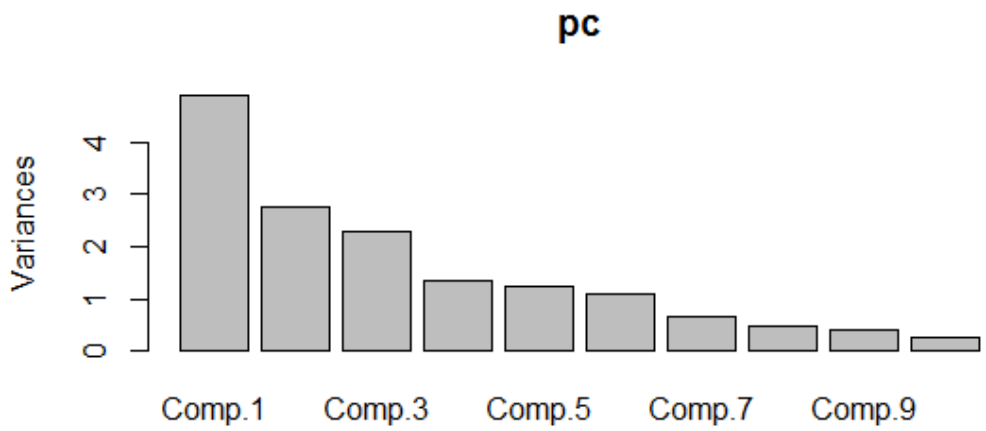
V. Analysis before removing the outliers:

Principal Components Method (PCA)

When we run the PCA code in R we will get the cumulative proportion of the components which implies the total variability explained by the components, and we will also get the scree plot which visually shows the contribution of each component in the total variability.

```
> print(summary(pc,loadings=T))
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     2.2087547 1.6633019 1.5140940 1.16260883 1.10612286 1.04246592
Proportion of Variance 0.3049123 0.1729108 0.1432800 0.08447871 0.07646924 0.06792095
Cumulative Proportion  0.3049123 0.4778231 0.6211032 0.70558188 0.78205112 0.84997207
                           Comp.7     Comp.8     Comp.9    Comp.10    Comp.11     Comp.12
Standard deviation     0.81331213 0.69240511 0.63835494 0.4947953 0.44056621 0.395475499
Proportion of Variance 0.04134229 0.02996405 0.02546856 0.0153014 0.01213116 0.009775054
Cumulative Proportion  0.89131436 0.92127841 0.94674697 0.9620484 0.97417953 0.983954589
                          Comp.13     Comp.14     Comp.15      Comp.16
Standard deviation     0.341774878 0.298806124 0.214387491 0.0683331270
Proportion of Variance 0.007300629 0.005580319 0.002872625 0.0002918385
Cumulative Proportion  0.991255218 0.996835537 0.999708161 1.0000000000
Error in if (loadings) { : argument is of length zero
>
```

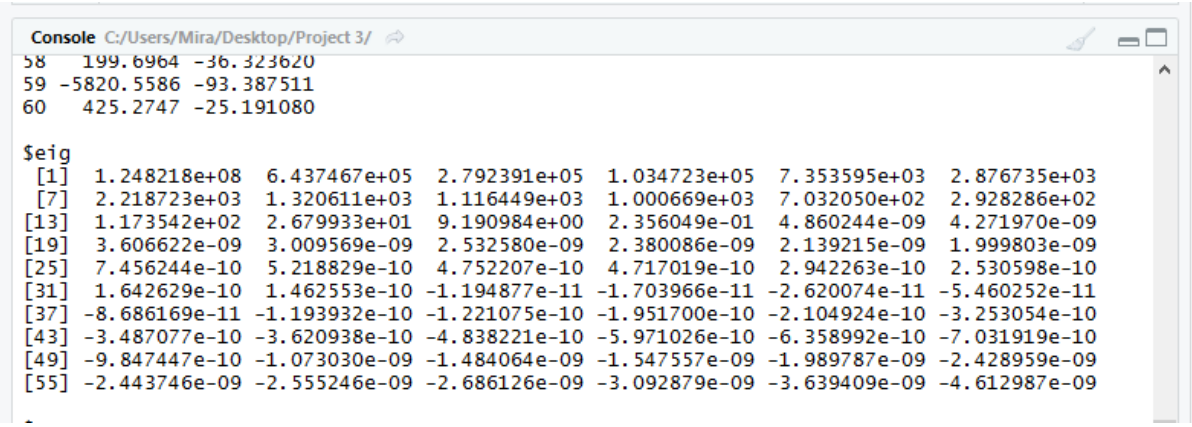
As shown in the output above, the first component on its own contains 30% of the total variability in the data. If we take the first 8 components only, they will account for 92% of the total variability in the data.



According to the scree plot, we could observe that the components decrease relatively in importance starting from component 5 or 6. To make a compromise between explanation of total variability and the reduction of number of variables I can choose to keep the first 6 components.

Multidimensional Scaling (MDS)

Using the method of MDS, I tried first to run the code setting m (the reduced number of variables) to 2.



Looking at the eigen values we can notice that only the first 15 eigen values are positive and the rest are nearly zeros.

Next, we take a look at the goodness of fit test:

The adequacy of the m-dimensional representation of X can be judged by this test. Since it is equal to 99.7% this means that m=2 is a reasonable MDS representation.

VI. Analysis after removing the outliers:

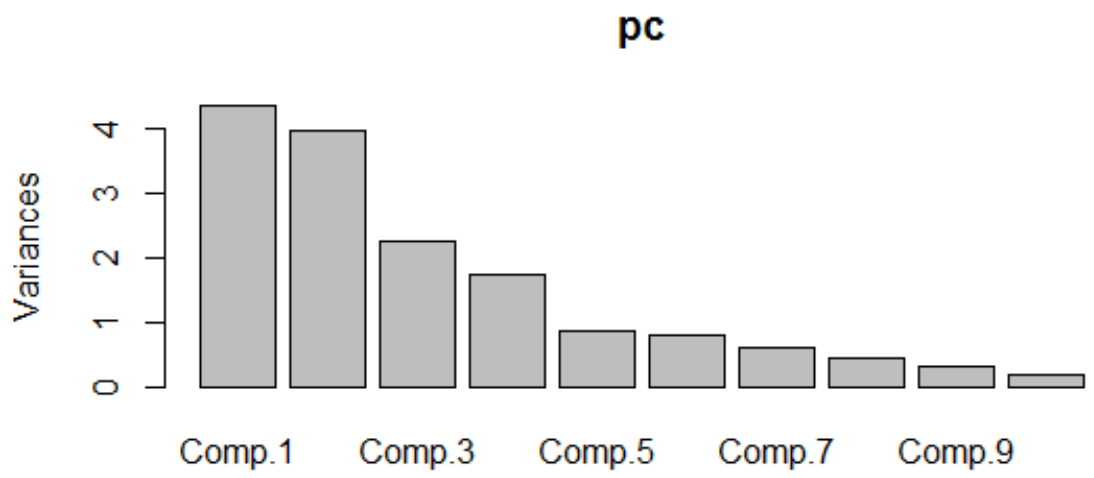
PCA

```
Console C:/Users/Mira/Desktop/Project 3/
 16  variables and  45 observations.
> print(summary(pc,loadings=T))
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6
Standard deviation     2.0882613 1.9919308 1.5056021 1.3178147 0.93705040 0.90269826
Proportion of Variance 0.2725522 0.2479868 0.1416774 0.1085397 0.05487897 0.05092901
Cumulative Proportion  0.2725522 0.5205390 0.6622163 0.7707561 0.82563503 0.87656404
                           Comp.7    Comp.8     Comp.9    Comp.10   Comp.11     Comp.12
Standard deviation     0.77862609 0.6663849 0.56699334 0.43844761 0.3610983 0.310819339
Proportion of Variance 0.03789116 0.0277543 0.02009259 0.01201477 0.0081495 0.006038041
Cumulative Proportion  0.91445520 0.9422095 0.96230209 0.97431686 0.9824664 0.988504404
                          Comp.13     Comp.14     Comp.15      Comp.16
Standard deviation     0.287489581 0.238639247 0.196059074 0.0767556643
Proportion of Variance 0.005165641 0.003559293 0.002402448 0.0003682145
Cumulative Proportion  0.993670045 0.997229338 0.999631785 1.0000000000

Loadings:
    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
```

As we can see in the cumulative proportions here that after removing the outliers, component 2 became more significant than before.

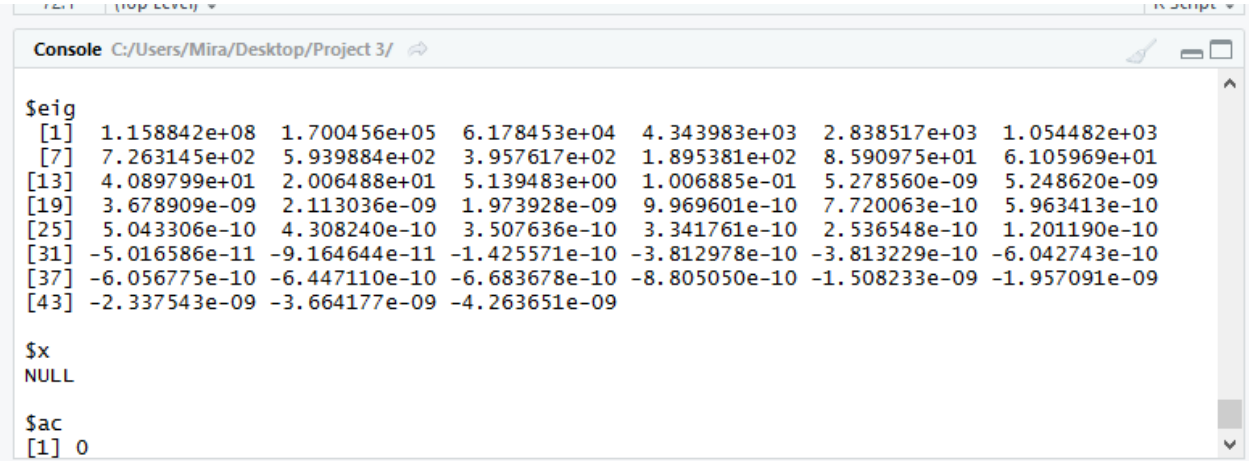
Scree plot:



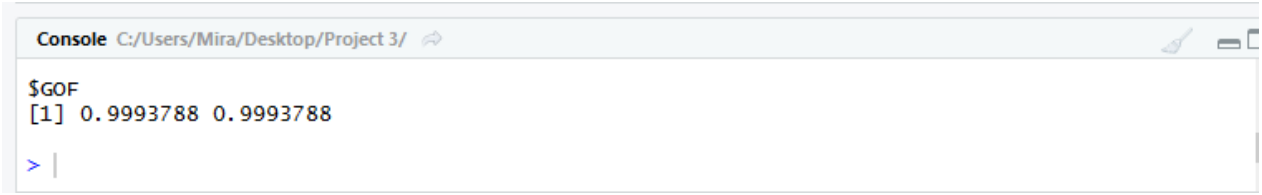
Regarding the scree plot, one could decide to stop at the 4th component according to the elbow rule, this will allow me to capture 77% of the variability in my data.

MDS

I reduced the variable number into two only, these are the eigen values. We have the first 15 are positive.



GOF test:



The goodness of fit test became better as the number changed from 99.3% to 99.9% almost 100%.