Mira Baraka

ID# 900141601

MACT 4233: Applied Multivariate Analysis, Professor Ali Hadi

Project 3

Due Date: May 6th

1. *Objective:*

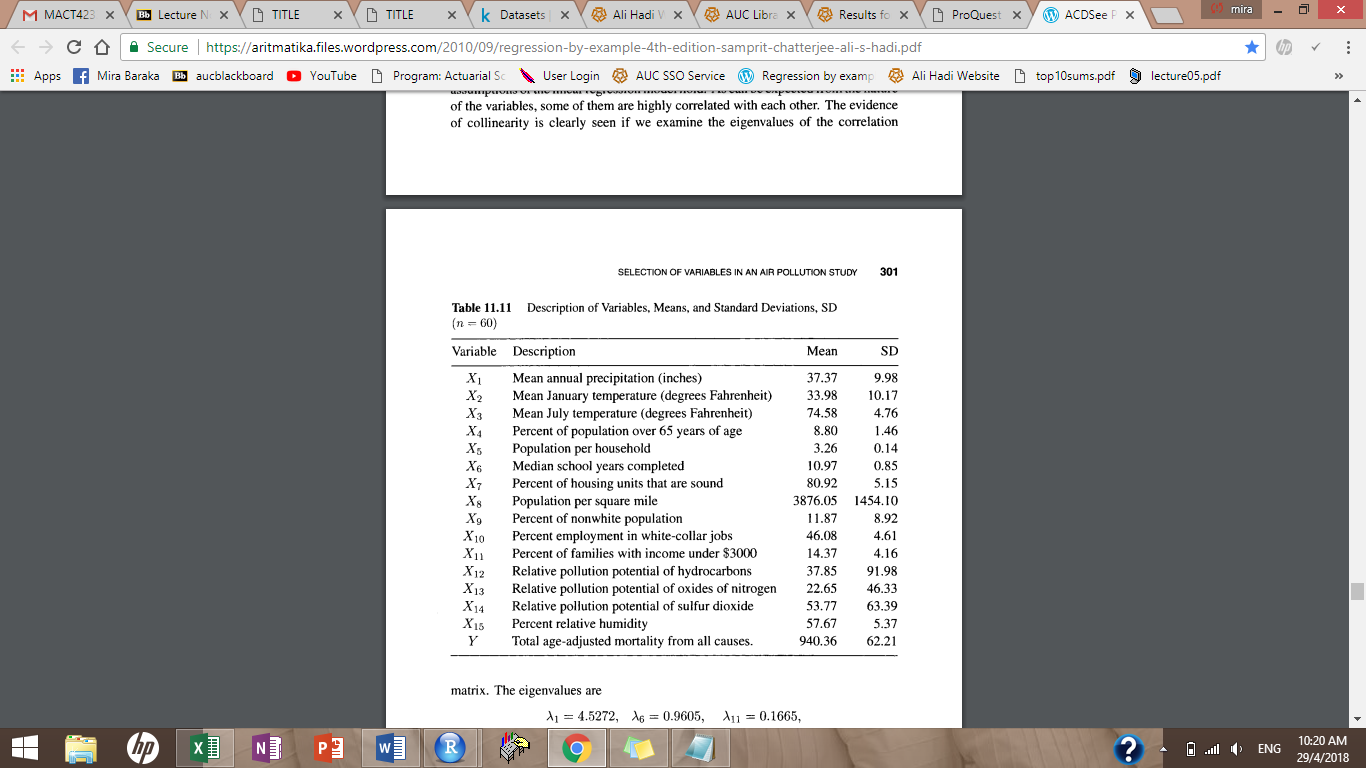
This project aims to produce a lower dimensional representation of the dataset using two methods; Principal components analysis and Multidimensional Scaling. Principal Components Analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number 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.

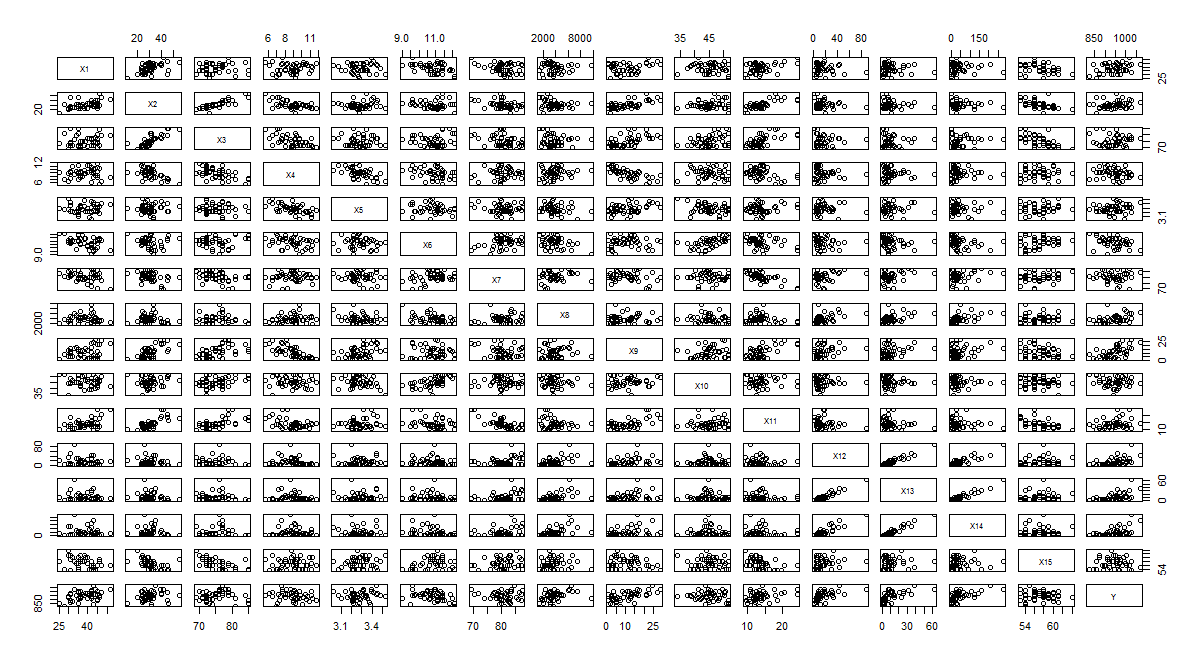
1. *Data Source:*

Regression Analysis by Example, 5th Edition, Samprit Chatterjee and Ali S. Hadi, [John Wiley & Sons.](http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470905840.html) 2012

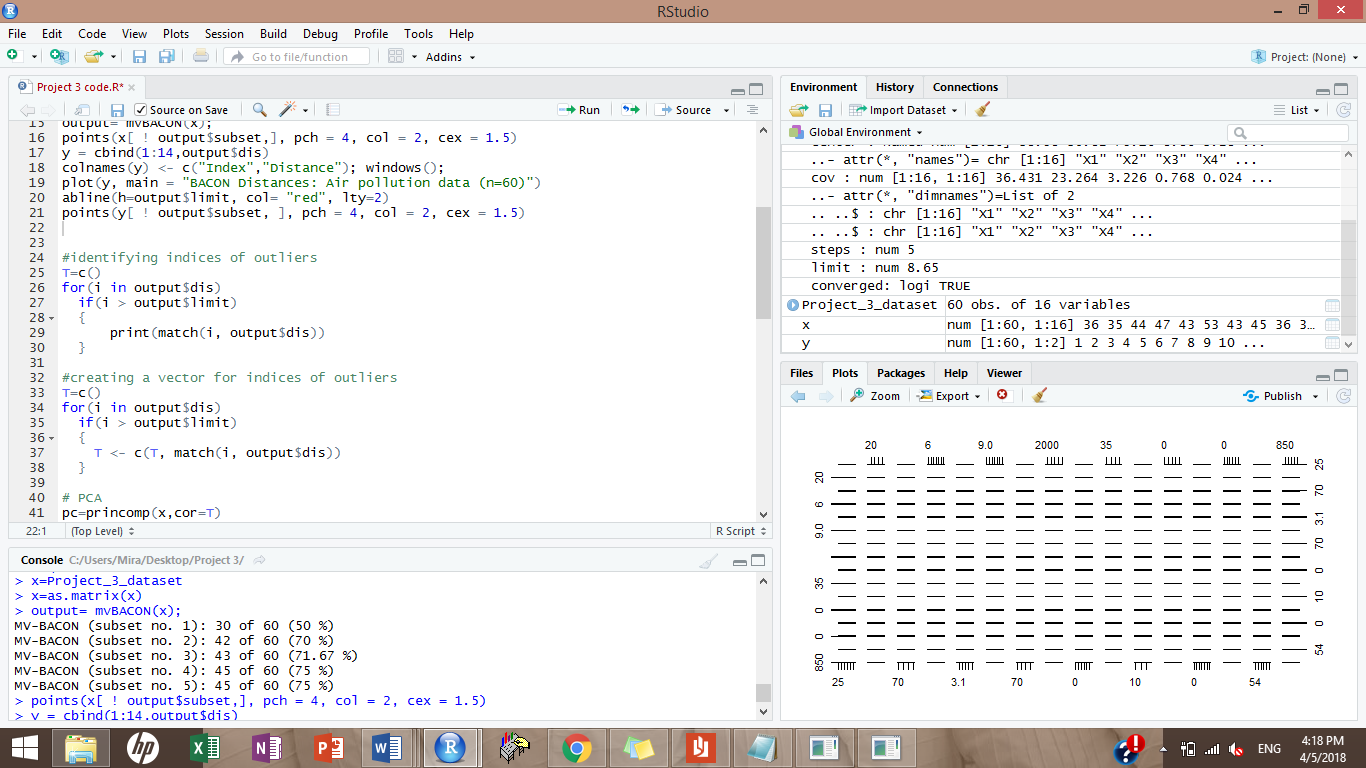
*Data Description:*

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.



1. *Plot of the pairs of X*
2. *Identification of outliers (BACON Method)*

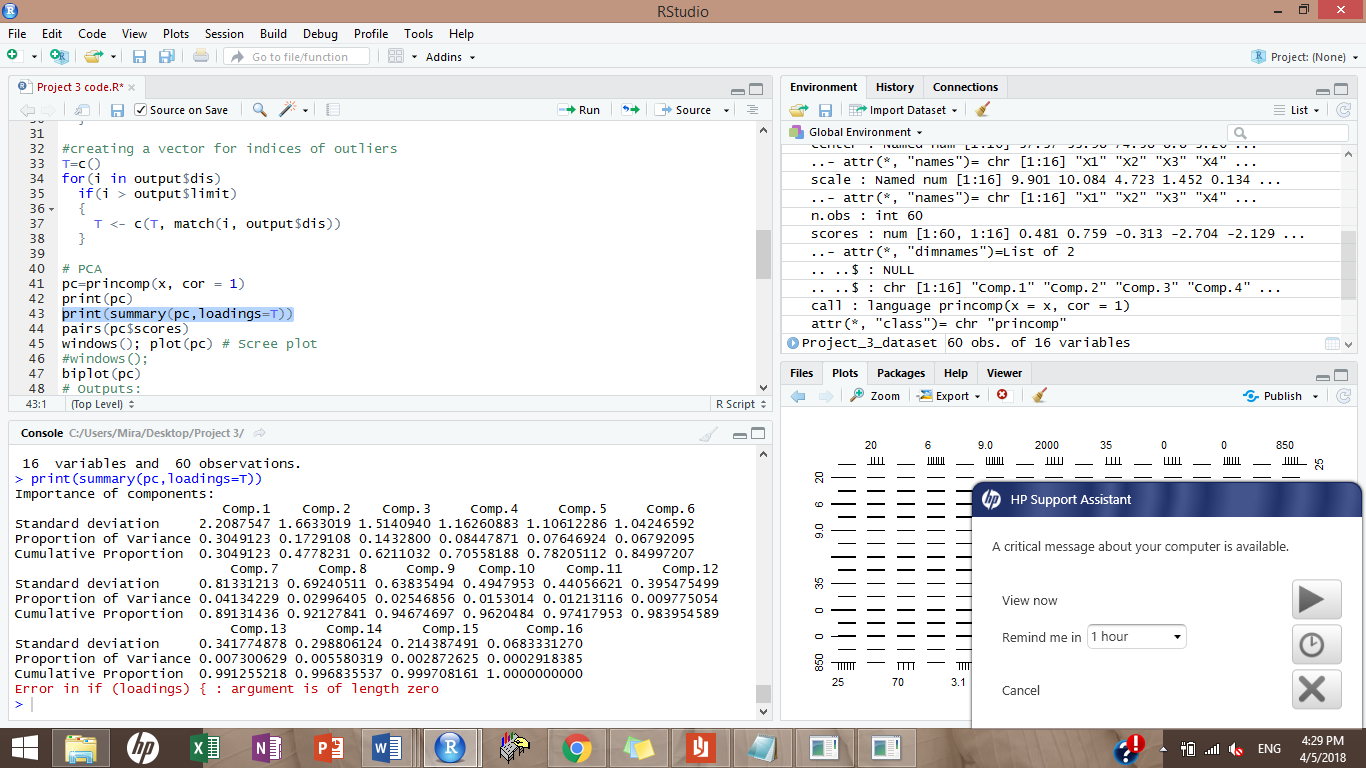
The BACON algorithm produced a basic subset of 45 observations, thus resulting in the appearance of 15 outlier observations 6,7,18,29,30,31,32,37,40,41,47,48,49,50,55 and the adjusted C(alpha) is 8.65.

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.

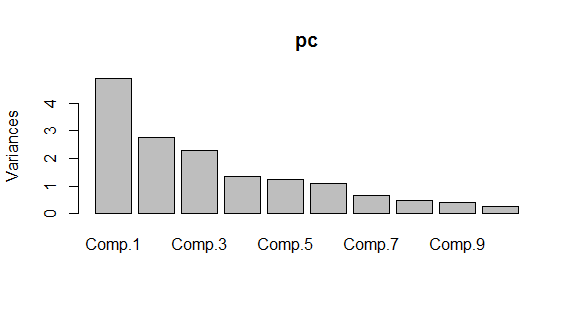
Code:

1. *Analysis before removing the outliers:*

*Principal Components Method (PCA)*

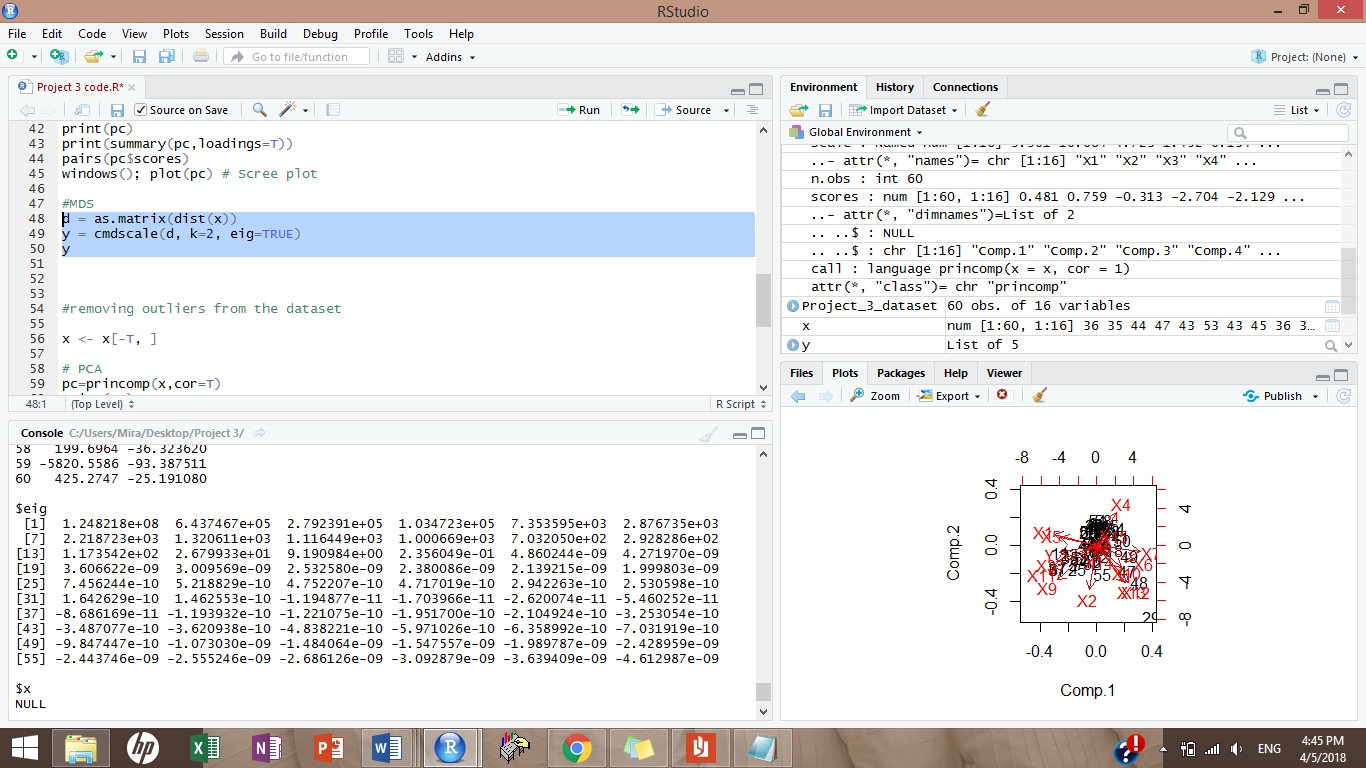
When we run the PCA code in R we will get the cumulative proportion of the components which implies the total variability explained by the components, and we will also get the scree plot which visually shows the contribution of each component in the total variability.

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.



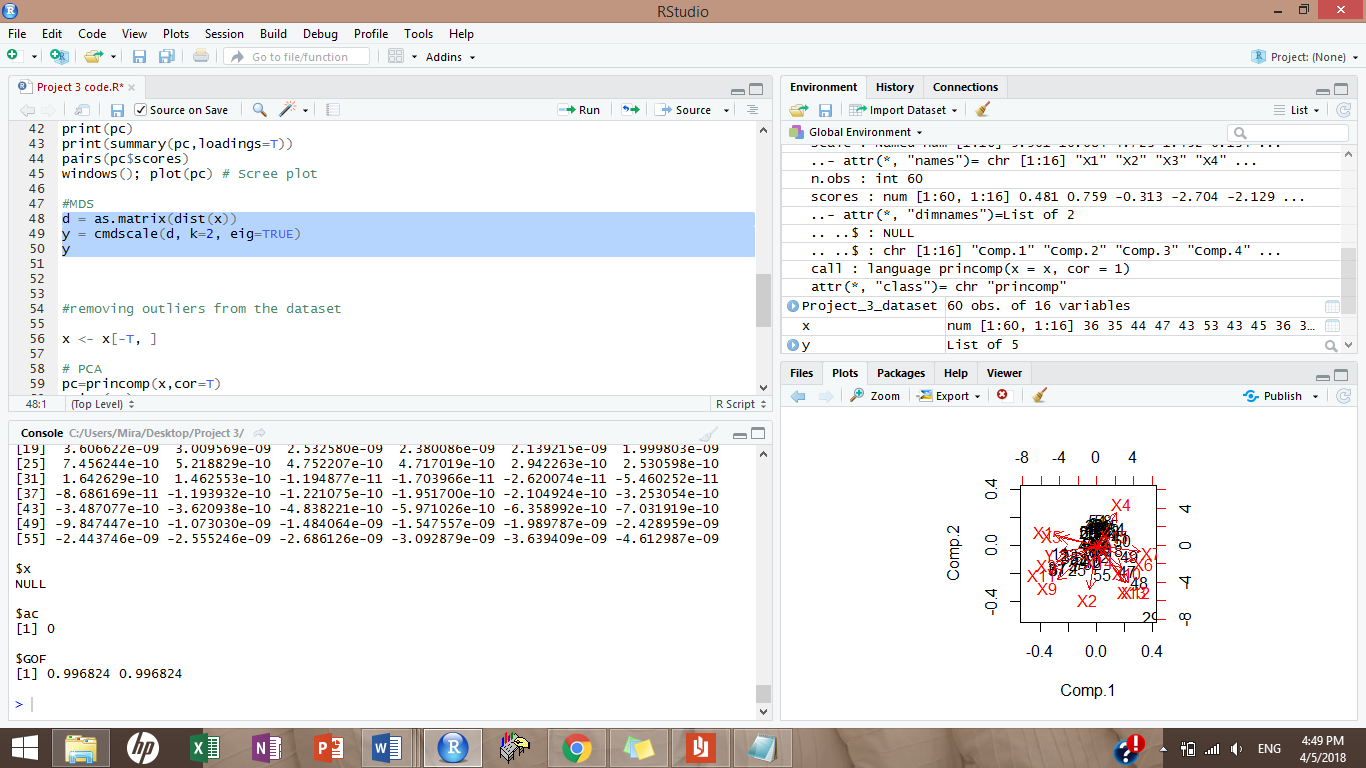
According to the scree plot, we could observe that the components decrease relatively in importance starting from component 5 or 6. To make a compromise between explanation of total variability and the reduction of number of variables I can choose to keep the first 6 components.

*Multidimensional Scaling (MDS)*

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

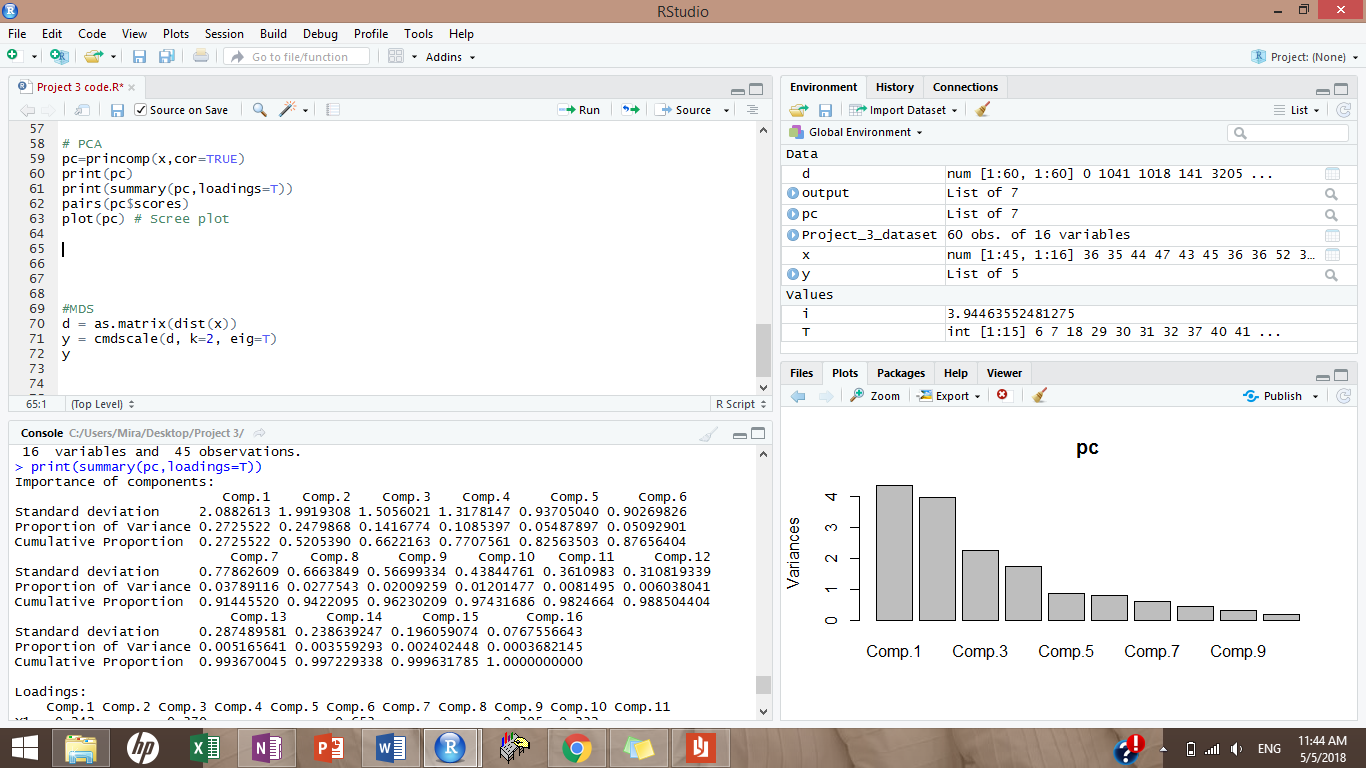
Looking at the eigen values we can notice that only the first 15 eigen values are positive and the rest are nearly zeros.

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

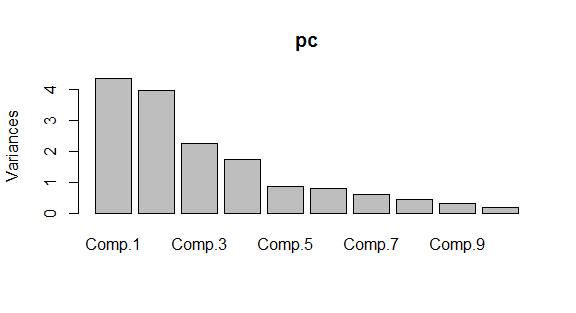


The adequacy of the m-dimensional representation of X can be judged by this test. Since it is equal to 99.7% this means that m=2 is a reasonable MDS representation.

1. *Analysis after removing the outliers:*

*PCA*

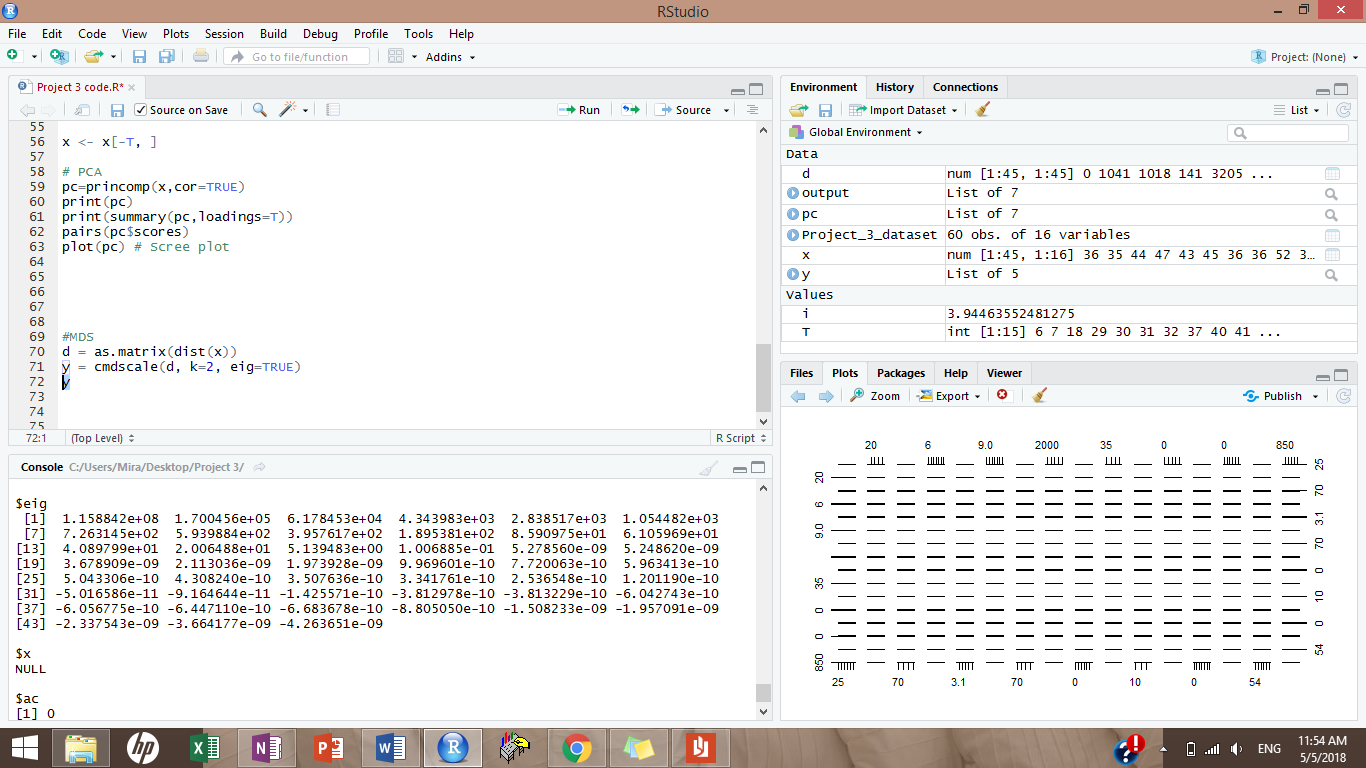
As we can see in the cumulative proportions here that after removing the outliers, component 2 became more significant than before.

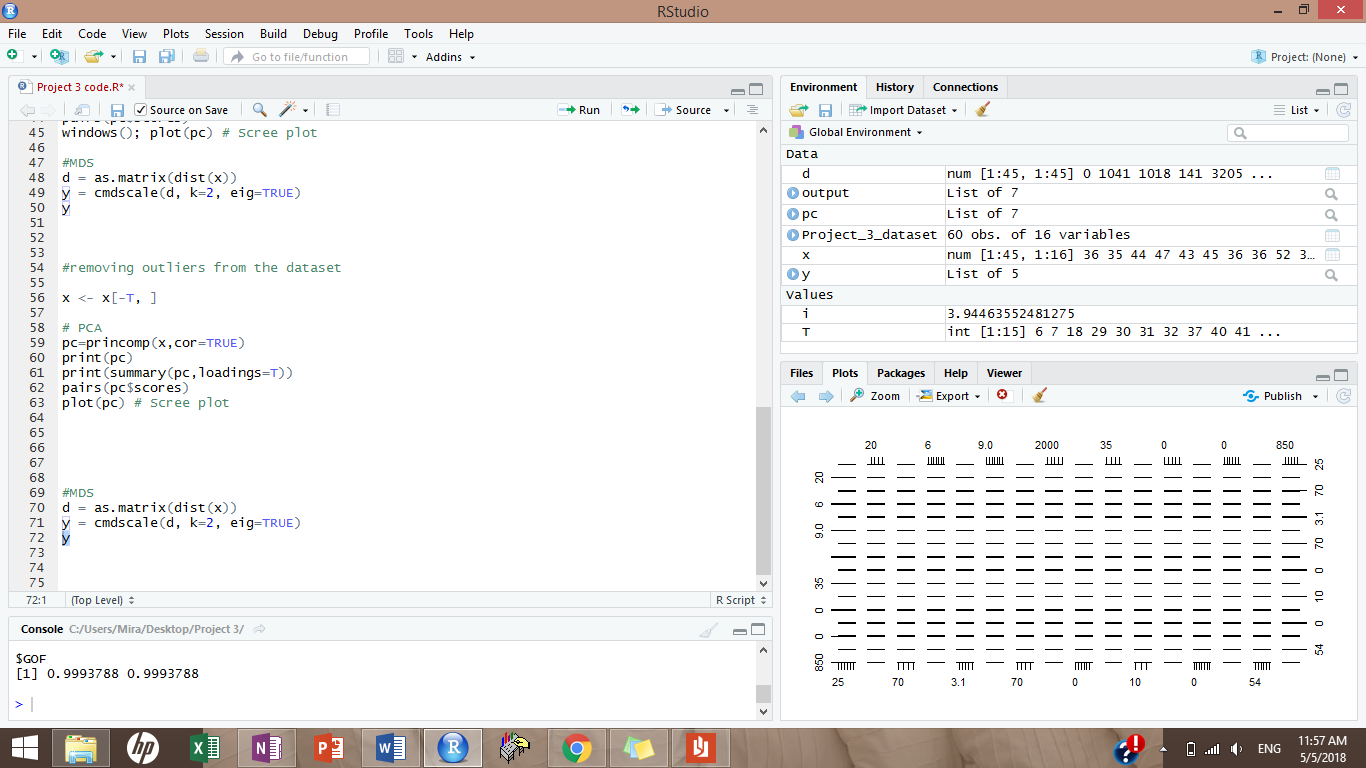
Scree plot:

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%.