**Multivariate Statistical Analysis 2**

**Assignment 1**

**PRN – 21060641054**

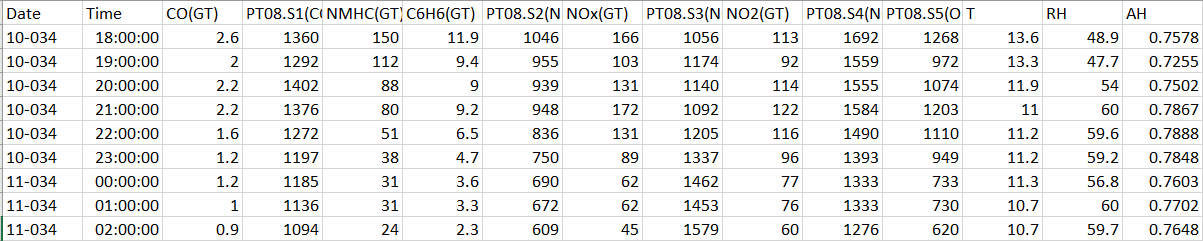
**Name- Vaidehi Rane**

**Dataset** – Air Quality Data Set from UCI ML Repository

**Pre-processing:**

1. **Change the decimal point from , (comma) to . (dot)**

After making data in proper format, the data looks like



1. **Change -200 to missing value**
2. Using excel:

Selection of find and replace tab, putting -200 in ‘find what’ and selecting replace all option in order to replace -200 value to null/ missing values.

1. Using R:



Using above command, -200 is replaced to null values

1. **Remove column numbers 1, 2 and 5**



As columns 16 and 17 are also eliminated as they are giving null values

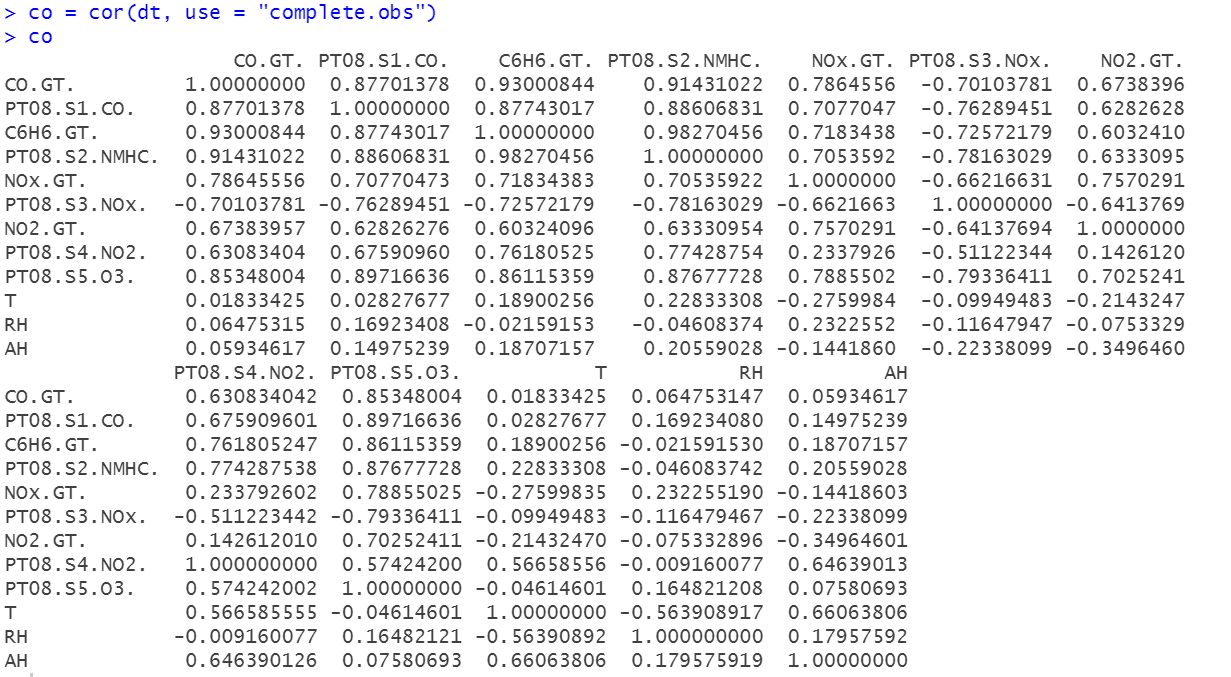
OR



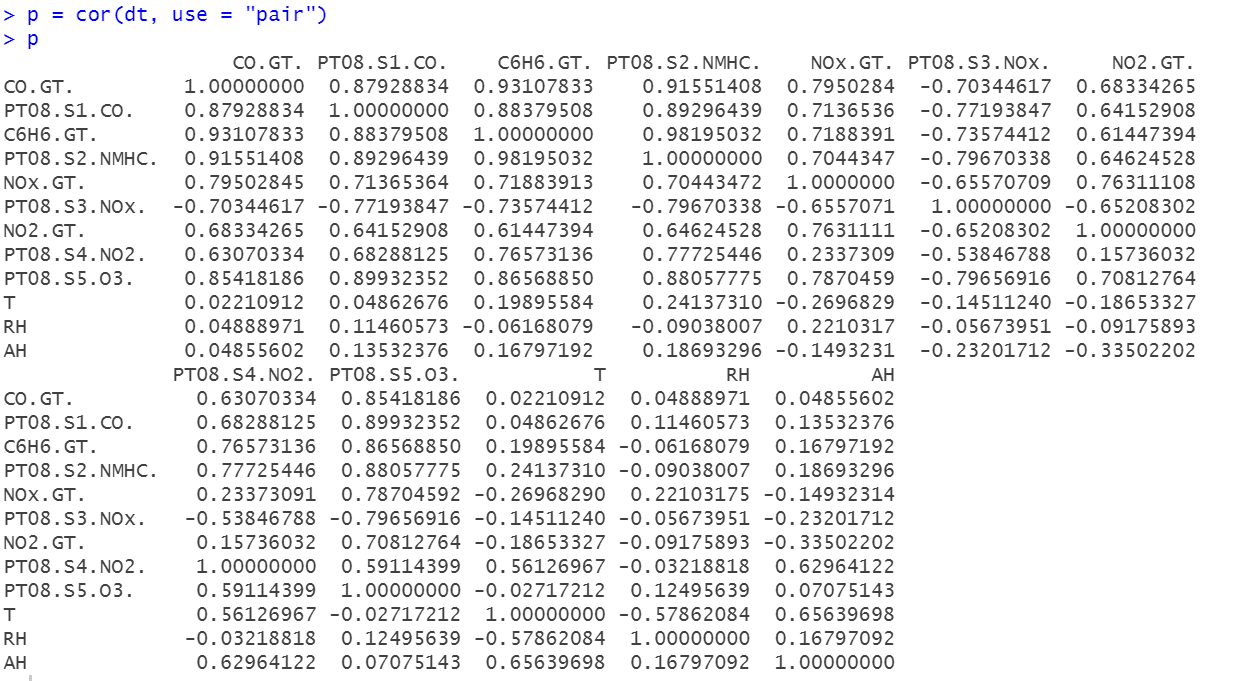
**Analysis:**

1. **Compute the correlation matrix in two ways using the options use = “complete.obs” and use = “pair”.**

Case 1:

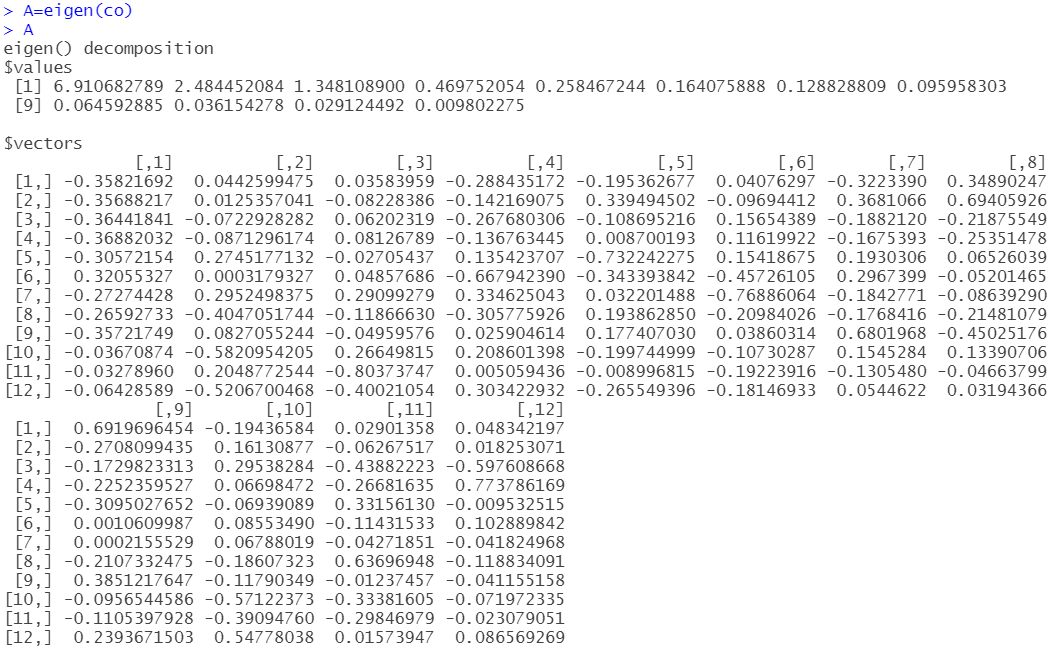


Case 2:

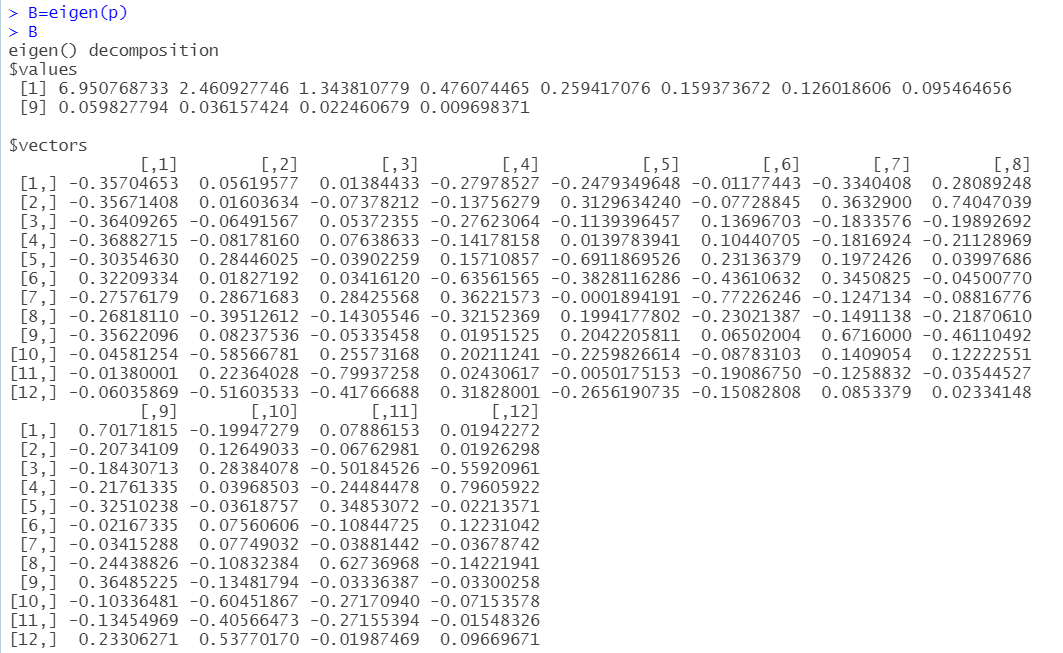


1. **Find the eigenvalues and eigenvectors of the correlation matrix in each case.**

Case1:



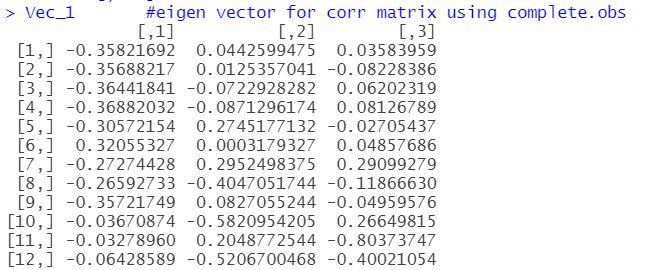
Case 2:



1. **In each case, identify eigenvalues greater than unity and corresponding eigenvectors.**

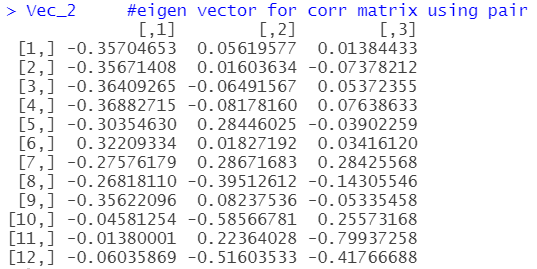
Case 1:





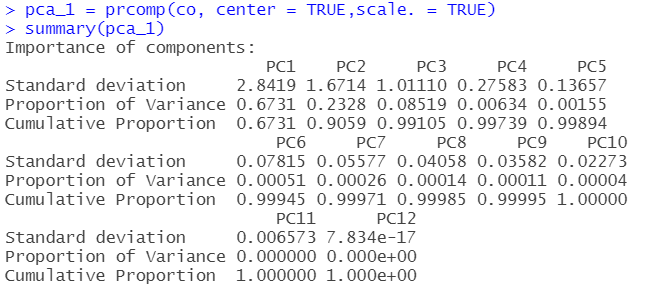
Case 2:





1. **In each case, find the proportion of variation explained by the selected eigenvalues.**

Case 1:



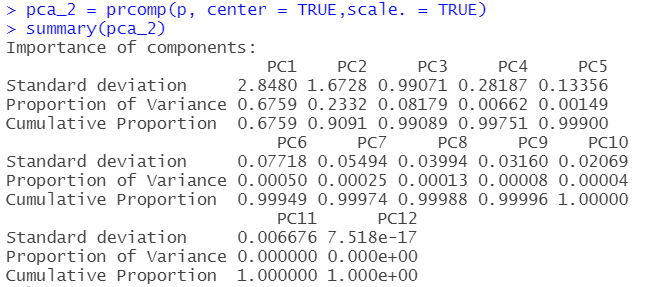
From above output, we conclude that proportion of variance explained by eigenvalues greater than unity are 0.6731, 0.2328, 0.08519 for 6.910683, 2.484452, 1.348109 respectively.

i.e. The 1st component explains 67.3% variation, 2nd component explains 23.8% variation and 3rd component explains 8.5% variation.

As eigenvalue of a factor divided by the sum of the eigen values is the proportion of variance explained by that factor, so proportion of variance explained by eigenvalues greater than unity are



Case 2:



From above output, we conclude that proportion of variance explained by eigenvalues greater than unity are 0.6759, 0.2332, 0.08179 for 6.950769, 2.460928, 1.343811 respectively.

i.e. The 1st component explains 67.6% variation, 2nd component explains 23.3% variation and 3rd component explains 8.18% variation.

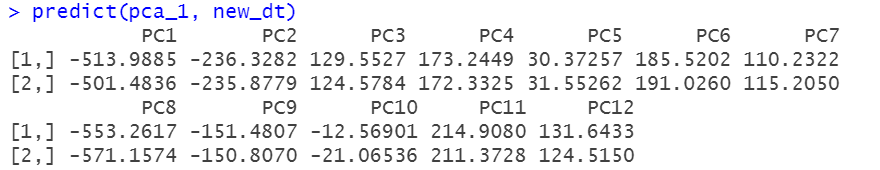
As eigenvalue of a factor divided by the sum of the eigen values is the proportion of variance explained by that factor, so proportion of variance explained by eigenvalues greater than unity are



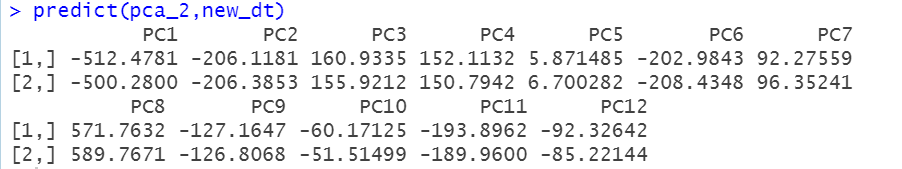
1. **Compute principal component scores in each case.**

In PCA the relationships between a group of scores is analyzed such that an equal number of new "imaginary" variables (aka principle components) are created.

Case 1:

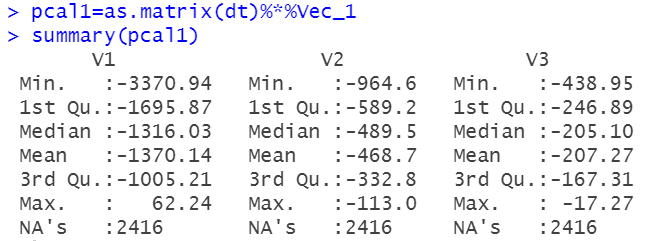


Case 2:

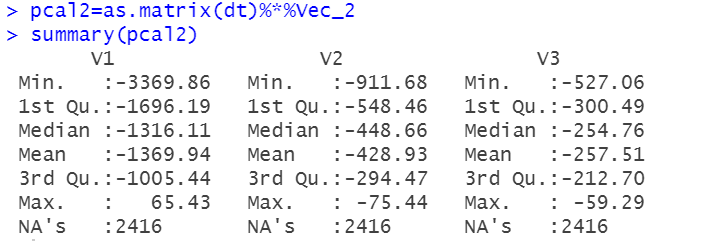


OR

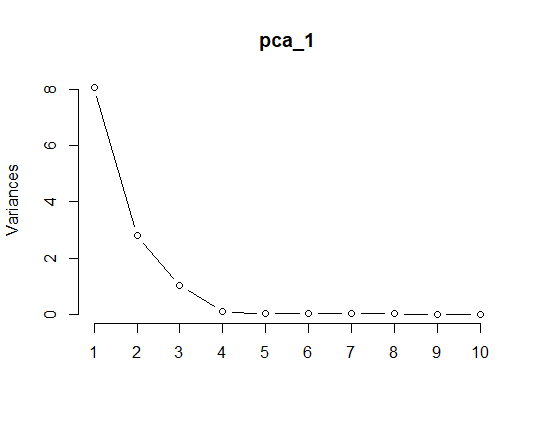
Case 1:

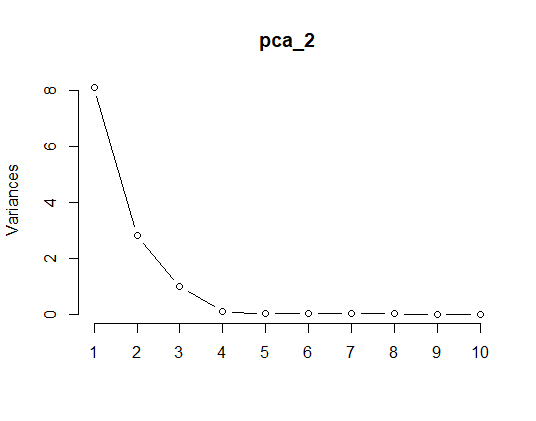


Case 2:



1. **Make a comparison between the two results and comment on your findings.**





The Scree Plot has two lines: the lower line shows the proportion of variance for each principal component, while the upper line shows the cumulative variance explained by the first 3 components. The principal components are sorted in decreasing order of variance, so the most important principal component is always listed first.

Both the graph is similar and with the naked eye it is impossible to detect the difference or to say the best between these model.