Statistical Structures in Data - Assignment 2

Name – Suman Pal (19BM6JP22)
PGDBA, First Year, 2019–2021

Ques. 1: From the *Concrete Compressive Strength Data Set* in the UCI Machine Learning Repository, (https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength), use the observations on the 9 variables (*Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age, Concrete compressive strength*) to compute the dispersion matrix **S** and the correlation matrix **R**. Perform Principal Component Analysis (PCA) with **S** and **R** separately and provide the following in each case:

i. The coefficients of the PCs
ii. The variances of the PCs
iii. The scree plot
iv. The number of PCs which explain 90% of the variation

Solution: The dataset was downloaded from the said repository and PCA was performed for both the dispersion matrix and the correlation matrix.

Case 1: The Dispersion Matrix (S):

The Dispersion Matrix is given as follows:

	Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2) (kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5) (kg in a m^3 mixture)	Coarse Aggregate (component 6) (kg in a m^3 mixture)	Fine Aggregate (component 7) (kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
Cement (component 1)(kg in a m^3 mixture)	10921.74265	-2481.35943	-2658.3508	-181.98979	57.91462	-888.60851	-1866.1511	540.99182	869.1476
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	-2481.35943	7444.08373	-1786.6076	197.67855	22.35531	-1905.21057	-1947.9113	-241.15038	194.3294
Fly Ash (component 3)(kg in a m^3 mixture)	-2658.35075	-1786.60759	4095.5481	-351.29712	144.25026	-49.64420	405.7364	-624.06475	-113.0614
Water (component 4)(kg in a m^3 mixture)	-181.98979	197.67855	-351.2971	456.06024	-83.87096	-302.72431	-771.5735	374.49650	-103.3223
Superplasticizer (component 5)(kg in a m^3 mixture)	57.91462	22.35531	144.2503	-83.87096	35.68260	-123.68745	106.5620	-72.72060	36.5338
Coarse Aggregate (component 6)(kg in a m^3 mixture)	-888.60851	-1905.21057	-49.6442	-302.72431	-123.68745	6045.65623	-1112.7952	-14.81127	-214.2298
Fine Aggregate (component 7)(kg in a m^3 mixture)	-1866.15111	-1947.91126	405.7364	-771.57347	106.56203	-1112.79516	6428.0992	-790.56558	-224.0107
Age (day)	540.99182	-241.15038	-624.0647	374.49650	-72.72060	-14.81127	-790.5656	3990.43773	347.0626
Concrete compressive strength(MPa, megapascals)	869.14762	194.32935	-113.0614	-103.32229	36.53380	-214.22975	-224.0107	347.06265	279.0797

Fig: Dispersion Matrix (S)

The coefficients of the PCs for the Dispersion Matrix(S) are provided below in order of their importance (i.e. in order of their eigen values):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Cement (component 1)(kg in a m^3 mixture)	-0.904445813	-0.023040685	-0.15203754	0.01346216	-0.15375901	0.2767427	-0.18387377	0.1548744	0.011235218
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	0.254633567	-0.788837027	-0.07143657	0.20066782	-0.10135129	0.4338754	-0.18288729	0.1881842	0.012089533
Fly Ash (component 3)(kg in a m^3 mixture)	0.239384888	0.299039739	0.04888714	-0.68625730	-0.18776084	0.4954283	-0.19442343	0.2478254	-0.003146049
Water (component 4)(kg in a m^3 mixture)	-0.005428255	-0.075493692	0.04205631	-0.07576154	0.09397864	-0.4679254	-0.07068464	0.8325592	0.246628008
Superplasticizer (component 5)(kg in a m^3 mixture)	0.001101790	0.004857144	-0.02419471	-0.02038522	-0.02279838	0.1013839	0.05590572	-0.2224131	0.967255259
Coarse Aggregate (component 6)(kg in a m^3 mixture)	0.012822425	0.276099110	0.75984686	0.47859604	-0.06207670	0.2752859	-0.07624081	0.1732152	0.041610727
Fine Aggregate (component 7)(kg in a m^3 mixture)	0.212270079	0.446453833	-0.61270296	0.48052792	0.14558390	0.2556211	-0.10237149	0.2270640	0.027084378
Age (day)	-0.100313043	-0.069996987	0.11771897	-0.14656970	0.94573844	0.2043009	-0.11288137	-0.0281684	0.001245791
Concrete compressive strength(MPa, megapascals)	-0.067216106	-0.040078608	-0.02018436	-0.03166726	0.04484212	0.2786040	0.92616718	0.2328213	-0.029035081

Fig: A table of loading factors of all the PCs for Dispersion Matrix(S)

The eigenvalues of the individual PCs are mentioned in the given table which is also their variances. The percentage and cumulative variance explained by each one of them is also given in the attached document below.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	12897.94299	32.49147577	32.49148
Dim.2	9825.43415	24.75145501	57.24293
Dim.3	7287.26315	18.35749579	75.60043
Dim.4	4247.63405	10.70030307	86.30073
Dim.5	3986.92184	10.04353751	96.34427
Dim.6	1268.12213	3.19455277	99.53882
Dim.7	102.07298	0.25713416	99.79595
Dim.8	69.74592	0.17569840	99.97165
Dim.9	11.25295	0.02834753	100.00000

Fig: A table of eigenvalues and the corresponding variances for Dispersion Matrix(S)

A scree plot for the PCA was generated to understand the amounts of variances explained by each one of the PCs. It was found that **a total of 5 PCs** alone could explain almost **96%** of the variances in the data.

The scree plot for the PCA is also attached herewith.

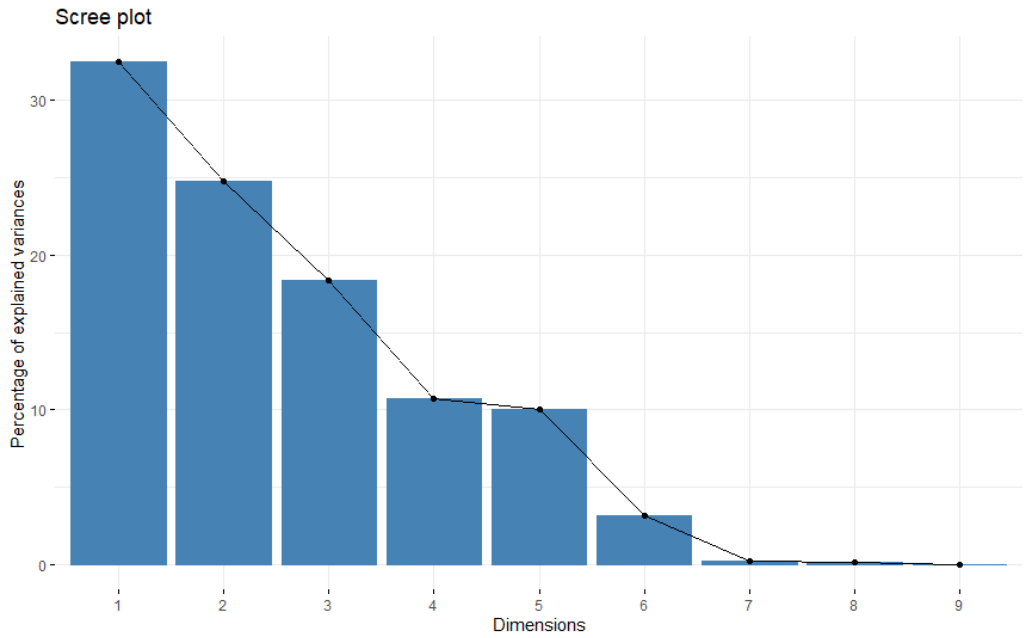

Fig: Scree plot generated after PCA of Dispersion Matrix

Case 2: The Correlation Matrix (R):

The Correlation Matrix is given as follows:

	Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2) (kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5) (kg in a m^3 mixture)	Coarse Aggregate (component 6) (kg in a m^3 mixture)	Fine Aggregate (component 7) (kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
Cement (component 1) (kg in a m^3 mixture)	1.00000000	-0.27519344	-0.397475440	-0.08154361	0.09277137	-0.109356039	-0.22272017	0.081947264	0.4978327
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	-0.27519344	1.00000000	-0.323569468	0.10728594	0.04337574	-0.283998230	-0.28159326	-0.044245801	0.1348244
Fly Ash (component 3) (kg in a m^3 mixture)	-0.39747544	-0.32356947	1.000000000	-0.25704400	0.37733956	-0.009976788	0.07907635	-0.154370165	-0.1057533
Water (component 4) (kg in a m^3 mixture)	-0.08154361	0.10728594	-0.257043997	1.00000000	-0.65746444	-0.182311668	-0.45063498	0.277604429	-0.2896135
Superplasticizer (component 5) (kg in a m^3 mixture)	0.09277137	0.04337574	0.377339559	-0.65746444	1.00000000	-0.266302755	0.22250149	-0.192716518	0.3661023
Coarse Aggregate (component 6) (kg in a m^3 mixture)	-0.10935604	-0.28399823	-0.009976788	-0.18231167	-0.26630276	1.000000000	-0.17850575	-0.003015507	-0.1649278
Fine Aggregate (component 7) (kg in a m^3 mixture)	-0.22272017	-0.28159326	0.079076351	-0.45063498	0.22250149	-0.178505755	1.00000000	-0.156094049	-0.1672490
Age (day)	0.08194726	-0.04424580	-0.154370165	0.27760443	-0.19271652	-0.003015507	-0.15609405	1.000000000	0.3288770
Concrete compressive strength(MPa, megapascals)	0.49783272	0.13482445	-0.105753348	-0.28961348	0.36610230	-0.164927821	-0.16724896	0.328876976	1.0000000

Fig: Correlation Matrix (R)

The coefficients of the PCs for the Correlation Matrix(R) are provided below in order of their importance (i.e. in order of their eigen values):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Cement (component 1)(kg in a m^3 mixture)	-0.04106772	0.5364728	-0.359699156	-0.30976550	-0.0546888297	-0.3898607	-0.133772604	0.2983673	-0.47251654
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	-0.16299258	0.1363006	0.698988731	0.07626221	-0.3625936171	0.2703252	0.004823084	0.2287723	-0.45115907
Fly Ash (component 3)(kg in a m^3 mixture)	0.36981124	-0.2684064	-0.019777961	0.60066828	0.2275943912	-0.3202302	0.247179667	0.2553441	-0.38647003
Water (component 4)(kg in a m^3 mixture)	-0.56408457	-0.1181279	0.120302488	0.04692072	0.2960859177	-0.3061956	-0.009807539	-0.5855720	-0.35604330
Superplasticizer (component 5)(kg in a m^3 mixture)	0.53605897	0.2482312	0.187959037	0.16585735	-0.0369894532	-0.0827810	-0.613878507	-0.4475792	-0.05281921
Coarse Aggregate (component 6)(kg in a m^3 mixture)	-0.06045806	-0.2248310	-0.549499919	0.22164550	-0.5454645518	0.3475894	-0.059842857	-0.2430661	-0.33720421
Fine Aggregate (component 7)(kg in a m^3 mixture)	0.38168340	-0.1870620	-0.001233946	-0.52781995	0.3844845216	0.4091099	0.174686900	-0.1403404	-0.41871225
Age (day)	-0.26191454	0.2518295	-0.169597266	0.35951724	0.5285269011	0.5097793	-0.343644186	0.2260221	-0.03967220
Concrete compressive strength(MPa, megapascals)	0.10723462	0.6301150	-0.033524610	0.22526326	0.0003062718	0.1539897	0.625981489	-0.3469226	0.06055837

Fig: A table of loading factors of all the PCs for Correlation Matrix(R)

The eigenvalues of the individual PCs are mentioned in the given table which is also their variances. The percentage and cumulative variance explained by each one of them is also given in the attached document below.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.28771185	25.4190206	25.41902
Dim.2	1.93651535	21.5168373	46.93586
Dim.3	1.40892580	15.6547311	62.59059
Dim.4	1.04278807	11.5865342	74.17712
Dim.5	1.01415431	11.2683812	85.44550
Dim.6	0.84741063	9.4156736	94.86118
Dim.7	0.28695777	3.1884197	98.04960
Dim.8	0.14678093	1.6308992	99.68050
Dim.9	0.02875528	0.3195031	100.00000

Fig: A table of eigenvalues and the corresponding variances for Correlation Matrix(R)

A scree plot for the PCA was generated to understand the amounts of variances explained by each one of the PCs. It was found that **a total of 6 PCs** alone could explain almost **94%** of the variances in the data.

The scree plot for the PCA is also attached herewith.

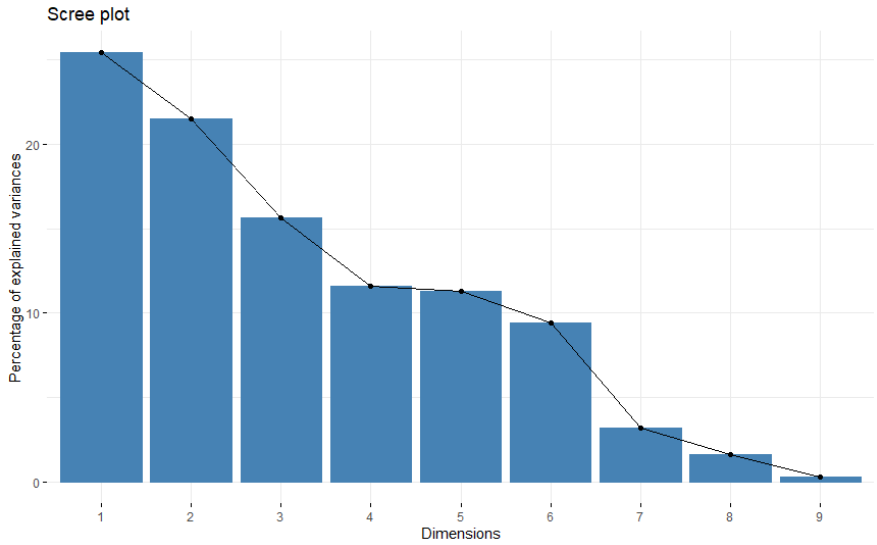

Fig: Scree plot generated after PCA of Correlation Matrix(R)

For Ques. 2-6: The data from the provided excel file was loaded into an R environment by using the R package **readxl**.

The code syntax used to read sheet wise data is as presented in the attached image below.

```
library("readxl")
```

```
novel <- read_excel("Assignment_2_data.xlsx", sheet = "author")
air <- read_excel("Assignment_2_data.xlsx", sheet = "USairpollution")
pot <- read_excel("Assignment_2_data.xlsx", sheet = "pottery")
tea <- read_excel("Assignment_2_data.xlsx", sheet = "tea")
flowers <- read_excel("Assignment_2_data.xlsx", sheet = "gardenflowers")
```

I have used the above generated objects like **novel, air, pot, tea** and **flowers** to deal with the problems in Ques.2-6.

Ques. 2: The dataset **author** provided in the first sheet of the attached MS-Excel file, *Assignment_2_data.xlsx*, contains the counts of the 26 letters of the alphabet (columns of matrix) for 12 different novels (rows of matrix). Each row contains letter counts in a sample of text from each work, excluding proper nouns.

i. Use any appropriate function from any R package to perform correspondence analysis on the data.

ii. Visualize the data in a two-dimensional space using the first two extracted coordinates from both rows and columns.
iii. Comment on the information provided by the 2-D CA plot regarding the association between them.

Solution: Since, the data provided is in the form of a contingency table having novels on vertical axis and the alphabets on the horizontal axis, hence **normal CA** must be performed on this data. By using R package, **FactoMineR**, a correspondence analysis on this data was performed and the results were noted. The 2D-CA plot / biplot obtained is as attached below:

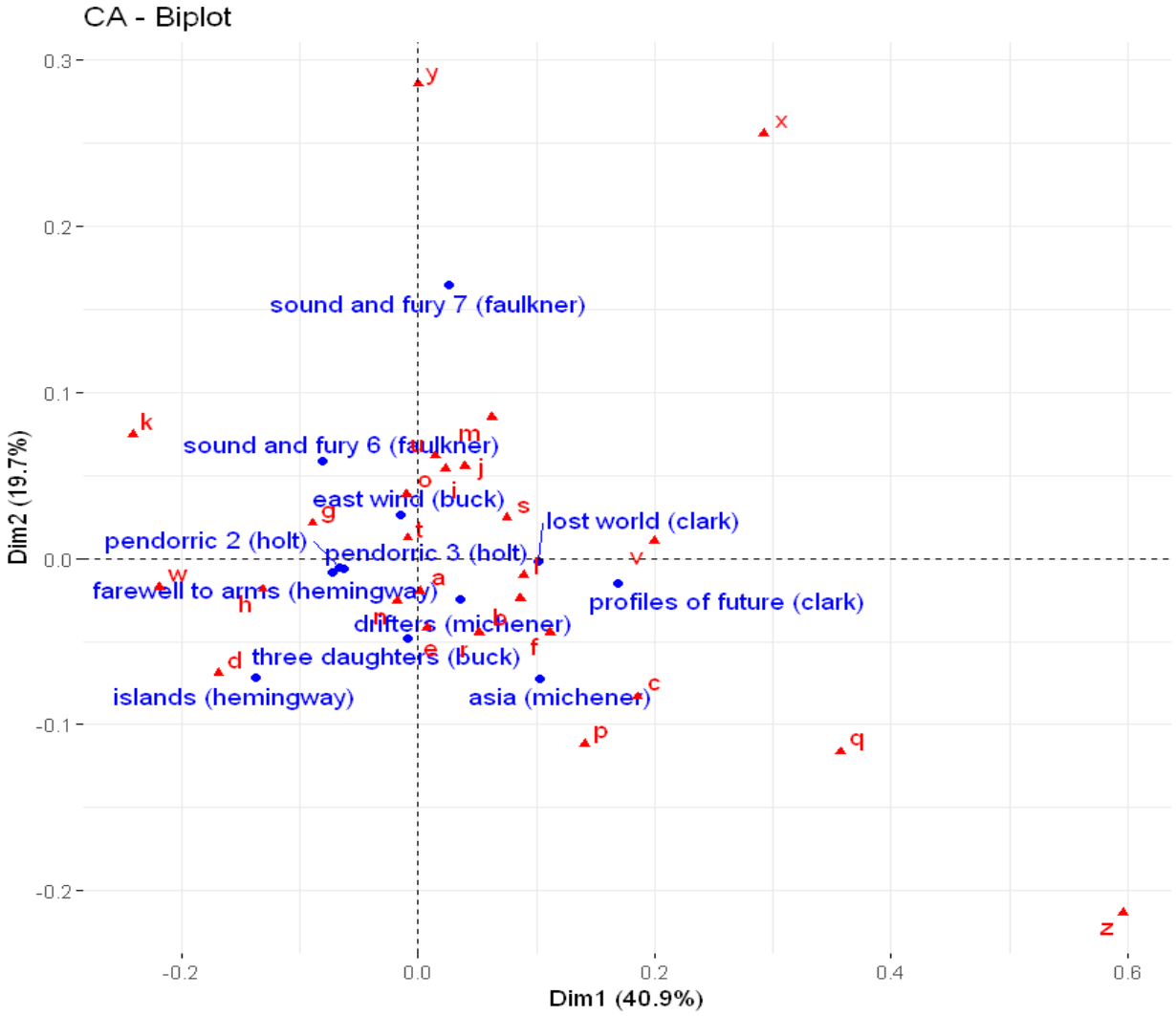

Fig: 2D CA plot for the author data

Comments:

1. The alphabets **x, y, q and z** are highly uncorrelated with the novel they have been used in. It indicates that these must have been a part of some specific novels like a word *(q)ueen* will be associated only with the fictions and/or historical documents.

2. Since, the usages of **vowels** can not be omitted in any literature, the dependence of each of these vowels with any novel is high.
3. The profile of **vowels a, e, i, o, u** is around the origin and hence suggesting an average profile for these alphabets.
4. The novel series **Pendorric** has both its novels very near to each other and hence suggesting equivalent usages of all the alphabets involved. On the other hand, **Sound and fury** part 6 and part 7 are strikingly different.
5. The usages of alphabets by the writer **Buck** in novels *East wind* and *Three Daughters* are drastically different indicating the different writing styles by the same author in 2 different novels. Whereas the alphabets used by **Michener** are quite similar in both of his 2 novels, *Drifters* and *Asia* indicating an adaptation of similar writing practices for both the books. Similar comments can be made for other authors as well.

Ques. 3: Consider the dataset **tea**, that is provided in the second sheet of the attached MS-Excel file *Assignment_2_data.xlsx*. It is a data frame (of factors) containing the answers of a questionnaire on tea consumption for 300 individuals. Although the data contains 36 columns (i.e., variables), consider only the following six columns:

- What kind of tea do you drink (black, green, flavored)
- How do you drink it (alone, w/milk, w/lemon, other)
- What kind of presentation do you buy (tea bags, loose tea, both)
- Do you add sugar (yes, no)
- Where do you buy it (supermarket, shops, both)
- Do you always drink tea (always, not always)

i. Use any appropriate function from any R package to perform correspondence analysis (CA) on the data.
ii. Visualize the data in a two-dimensional space using the first two extracted coordinates from both rows and columns.
iii. Comment on the information provided by the 2-D CA plot regarding the association between them.
iv. Consider the data in the last five columns, which correspond to binary attributes. Treat these as observations as ordinal variables by assigning the value 0 to "not-A" and the value 1 to A, A being the attribute corresponding to the respective columns. Compute the tetrachoric correlations for these 5 variables and perform PCA with the tetrachoric correlation matrix. Identify the attributes that explain 90% of the variation.

Solution: For parts (i) - (iii):
The tea consumption data as provided is in the form of an indicator matrix with multiple levels. Hence **Multiple Correspondence Analysis (MCA)** must be applied on this data after transforming each level of the horizontal axis data into a binary variable. This will be done by **one-hot encoding** method. However, we do not need to do this step explicitly as the R package, **FactoMineR** takes care of this. A multiple correspondence analysis on this data was performed and the results were noted. The 2D-MCA plot / biplot obtained is as attached below:

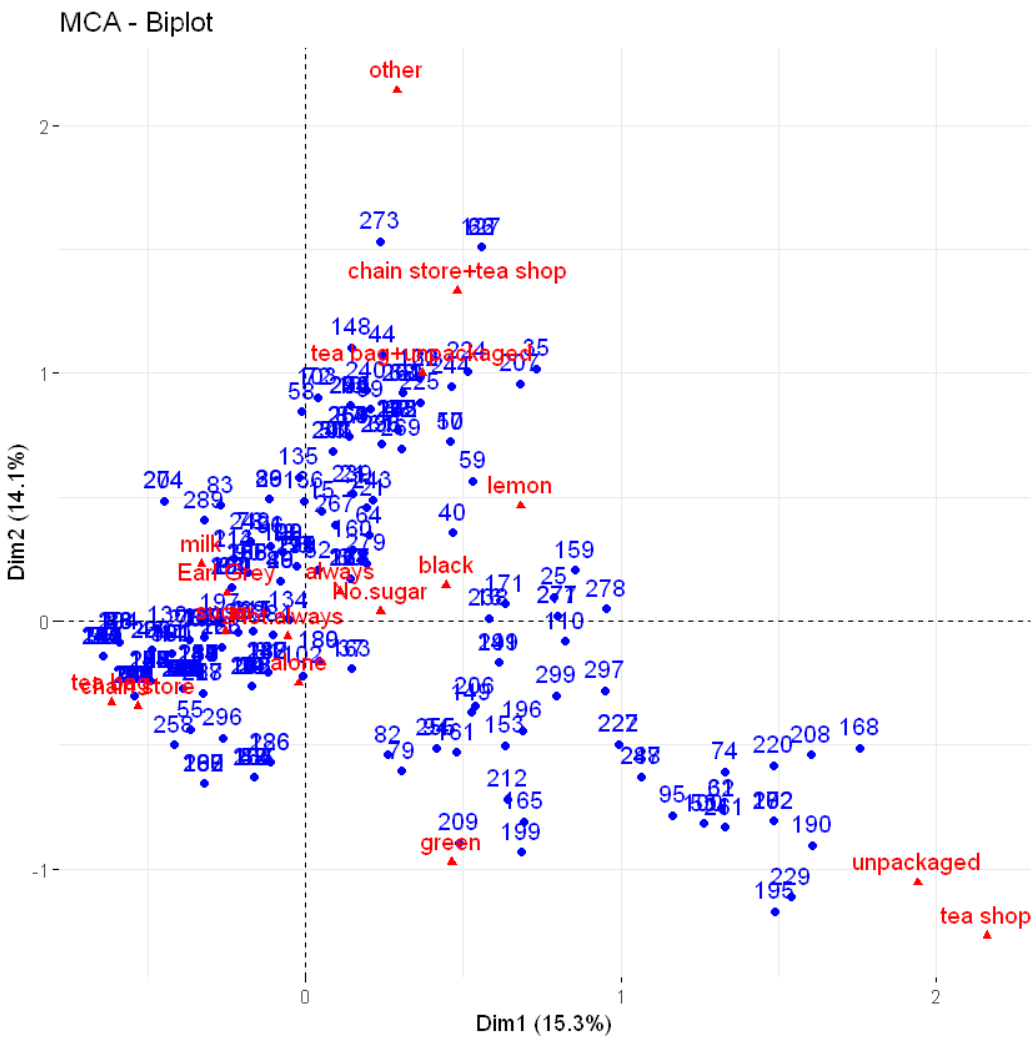

Fig: 2D- MCA plot for the tea data for assigned 6 attributes

Comments:

1. The individuals prefer to buy tea packs over loose packs of tea.
2. The individuals prefer to buy their tea packs from supermarket rather than from designated tea shops.
3. Not many individuals drink green tea.
4. Not many individuals drink their tea with both the milk and lemon which is expected.
5. Many individuals drink their tea in a similar fashion by using sugar and milk which is obvious.

Part iv):

The data given in the last five columns of the file was all binary. Each of these columns were converted to numeric digits by encoding them as either 0 or 1 as instructed. Ifelse() statement has been used to transform each of these attributes. After the transformation, a tetrachoric correlation matrix was obtained from the data.

```
tea_data2$sophisticated <- ifelse(tea_data2$sophisticated== 'sophisticated',1,0)
tea_data2$slimming <- ifelse(tea_data2$slimming== 'slimming',1,0)
tea_data2$exciting <- ifelse(tea_data2$exciting== 'exciting',1,0)
tea_data2$relaxing <- ifelse(tea_data2$relaxing== 'relaxing',1,0)
tea_data2$effect.on.health <- ifelse(tea_data2$effect.on.health== 'effect on health',1,0)
```

```
head(tea_data2)
```

sophisticated	slimming	exciting	relaxing	effect.on.health
0	0	0	0	0
0	0	1	0	0
0	0	0	1	0
1	0	0	1	0
0	0	0	1	0
0	0	0	1	0

Fig: Label Encoding of last 5 variables

```
tea_tetracorr= tetrachoric(tea_data2)$rho
tea_tetracorr
```

	sophisticated	slimming	exciting	relaxing	effect.on.health
sophisticated	1.000000000	0.12048370	0.17873799	0.12752601	-0.009948105
slimming	0.120483696	1.00000000	0.13188454	0.07432462	0.184223072
exciting	0.178737987	0.13188454	1.00000000	-0.40209463	0.014108792
relaxing	0.127526008	0.07432462	-0.40209463	1.00000000	-0.177777961
effect.on.health	-0.009948105	0.18422307	0.01410879	-0.17777796	1.000000000

Fig: Tetrachoric Correlation Matrix obtained from the given set of columns

Thereafter PCA was performed on the matrix mentioned above. Following are the code snippets used to develop these steps:

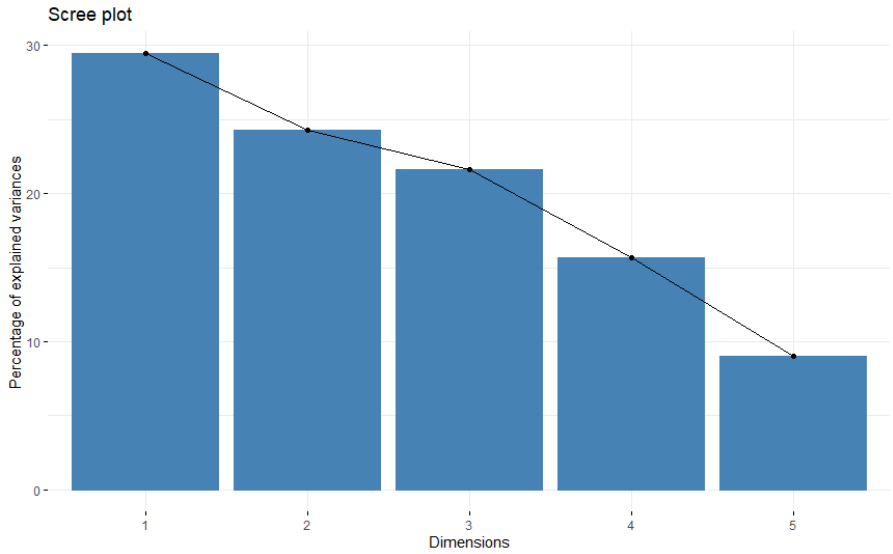

Fig: Scree plot obtained after the PCA on the tetrachoric correlation matrix

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.2133655 1.1024474 1.0394257 0.8846188 0.67037129
Proportion of Variance 0.2944512 0.2430781 0.2160811 0.1565101 0.08987953
Cumulative Proportion  0.2944512 0.5375292 0.7536104 0.9101205 1.00000000
```

Fig: Importance of each PC obtained for last 5 column of tea data

From the above results, it is evident to say that the first **4 principal components** identified almost 91% of the data variation.

Ques. 4: The third sheet of the attached MS-Excel file *Assignment_2_data.xlsx,* labeled **pottery**, contains the results of chemical analysis on 45 pots of Romano-British origin, made in five different kilns located in three different regions, in the form of observations on nine different chemical constituents.

i. Compute the distance matrix for the 45 pots.

ii. Perform metric multidimensional scaling to ascertain to what extent the chemical profiles of the pots suggest similarity among them, examining the 2-dimensional MDS plot corresponding to the data.

iii. If you are given additional information that

- the first 21 pots are from kiln no. 1, the next 12 are from kiln no. 2, followed by 2, 5 and 5 pots from kiln nos. 3, 4 and 5 respectively
- region 1 contains kiln 1, region 2 contains kilns 2 and 3, and region 3 contains kilns 4 and 5

do your conclusions in (ii) appear to reflect similarity in respect of kiln and/or region? Explain with the help of a modified version of the MDS plot in which pots from different kilns are shown in different colours.

Solution: The pottery data contained mineral content of individual pots. These different minerals can be treated as multiple dimensions that can describe a given pot from another. The distance matrix was calculated by using **dist** function for the data. Further, metric multidimensional scaling was performed from the distance matrix obtained. The result obtained was plotted for 2 dimensions as below:

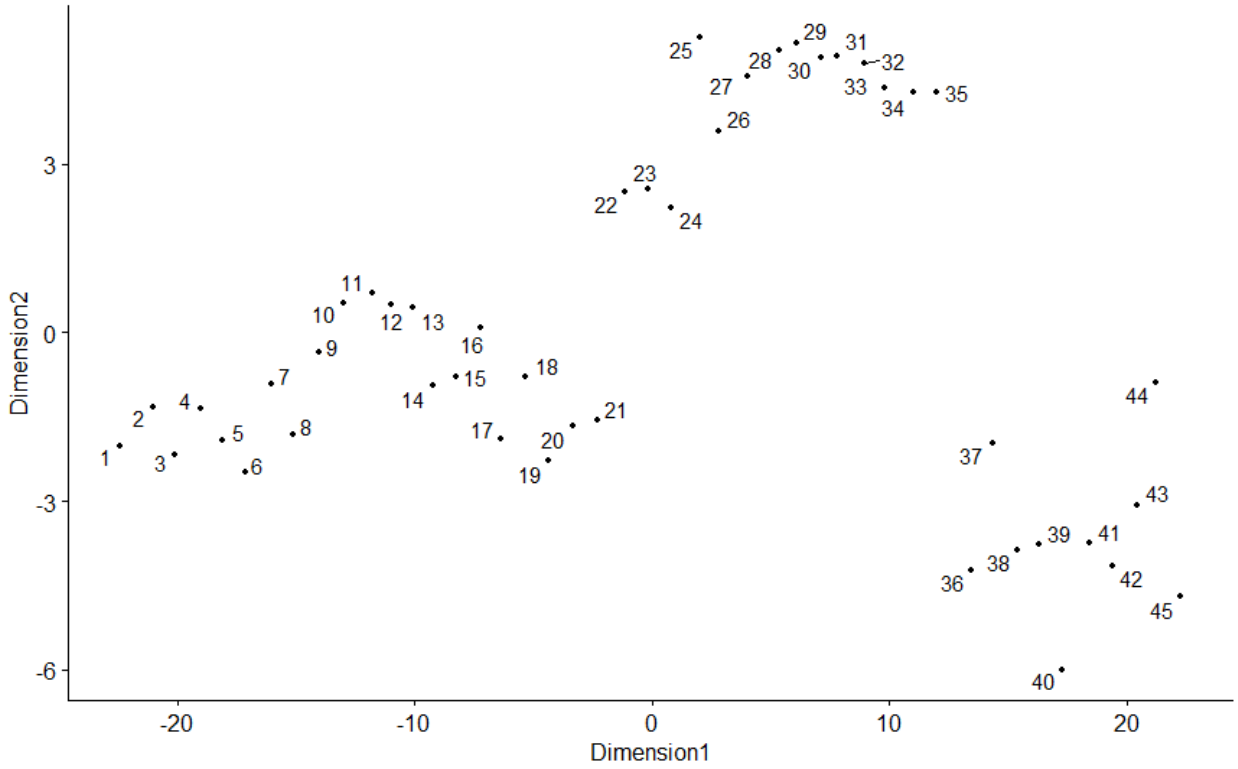

Fig: 2D MDS for pottery data using the arrived distance matrix

Comments:

1. We can observe the clusters easily by seeing the plot.
2. Pots # (1-21), (22-35), (38-43) seem to be different clusters indicating similarity in their chemical constituents.
3. Pots # (37, 44) are far from others indicating that they are very different from other pots or an outlier to the scaling process.

To understand the difference in kilns and the regions to which they belong, each kiln was marked with a specific color. The color code used is as follows:
Kiln 1: **Black** points |Kiln 2: **Blue** points |Kiln 3: **Green** points |Kiln 4: **Red** points | Kiln 5: **Orange** points

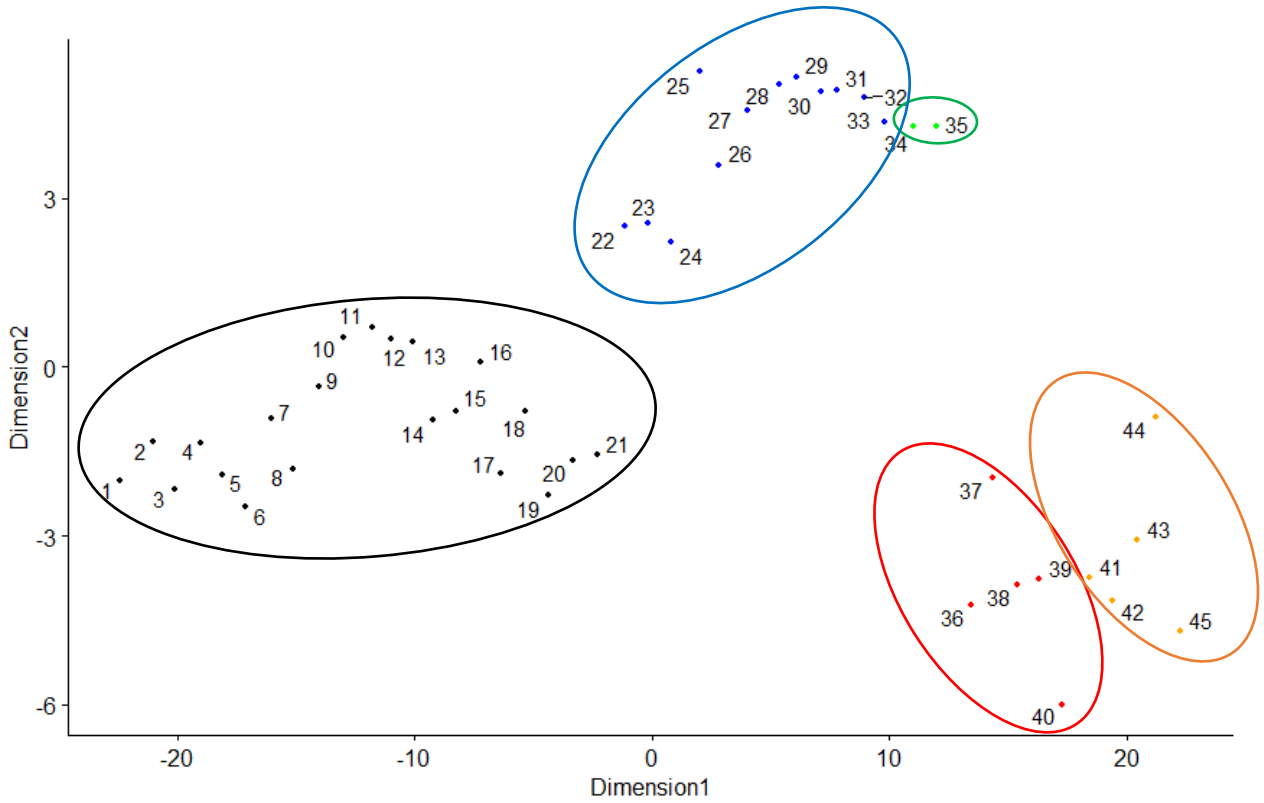

Fig: 2D MDS for pottery data using the arrived distance matrix and the colors for each kiln

Comments:

1. As interpreted from previous plot, we can see that all the pots from kiln 1 (1-21) are similar to each other
2. Similarly, for the kiln 2,3,4 and 5 all the pots of same kiln are close to each other compared to others indicating their similarity in chemical constituents.
3. We can also kiln 2 and 3 are not much far from each other indicating that kiln 2 and kiln 3 has similar chemical constituents.
4. Similarly plot indicates that kiln 4 and kiln 5 has similar chemical constituents.

Ques. 5: The fourth sheet of the attached MS-Excel file *Assignment_2_data.xlsx,* labeled **gardenflowers**, contains the dissimilarity matrix of 18 species of garden flowers.
i. Use some form of non-metric multidimensional scaling to investigate which species share common properties.
ii. Compute Kruskal's stress measure for a number of dimensions and generate a scree plot with the values.
iii. According to Kruskal's guidelines what is the assessment of fit in 2 dimensions?

Solution:

From the dissimilarity matrix of the garden flowers, we can execute a non-metric multidimensional scaling to find out the relation among the different species of the flowers. This can be done by using the R package **MASS** and the function **isoMDS()** that it contains. The non-metric MDS is done after finding out the distance matrix from this dissimilarity matrix. The 2D- Non MDS plot was obtained after this step which is presented below:

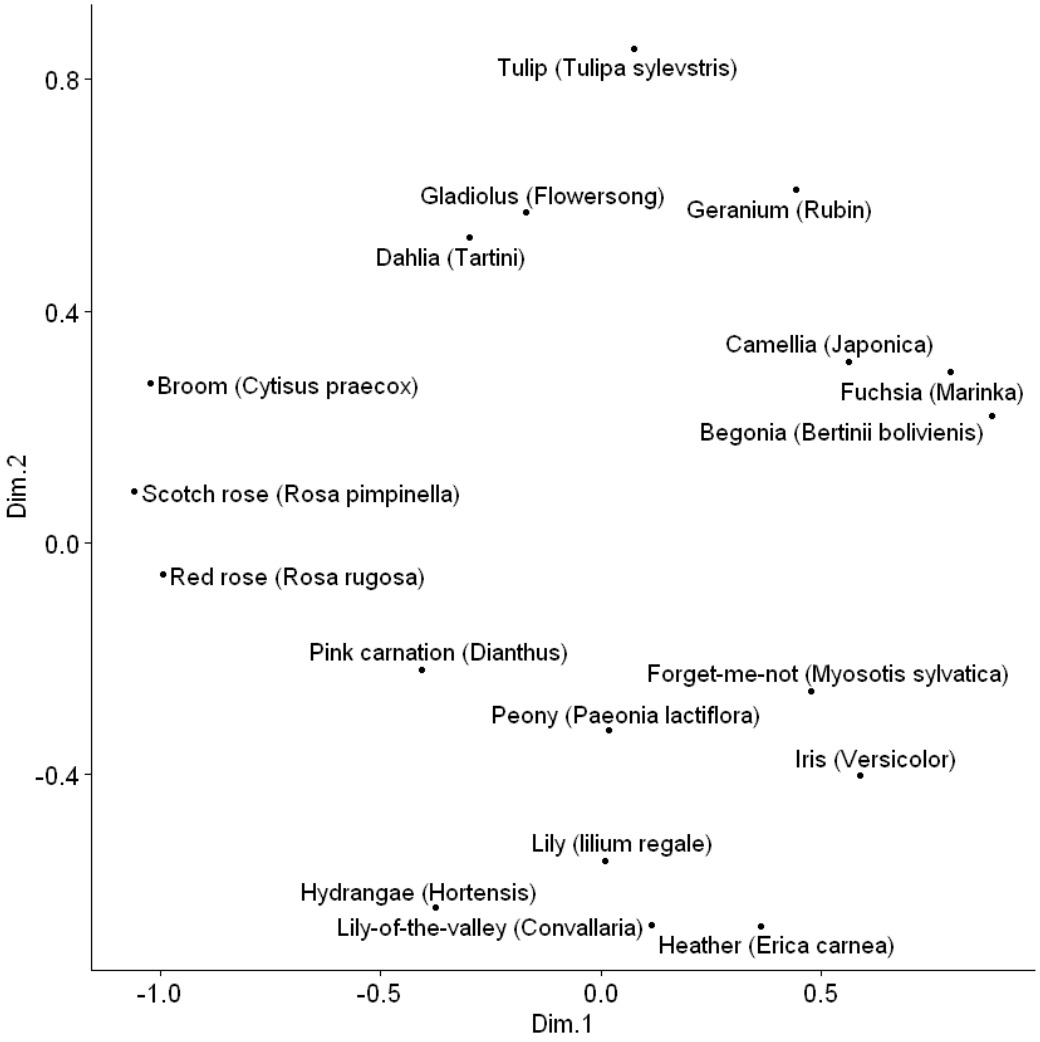

Fig: 2D- Non MDS plot for garden flowers

Comments:

1. As evident from the plot, flowers like Dahlia, Gladiolus, Geranium and Tulip share common properties. This can be found from the extreme upper part of the graph.
2. Similar to 1, flowers like Lily, Hydrangae, Lily-of-the-valley and Heather share common properties. This is indicated in the bottom part of the plot.
3. Species of Rose like Scotch Rose and Pink Rose are similar in nature.
4. Camelia, Fuchsia, Begonia are very similar in properties.

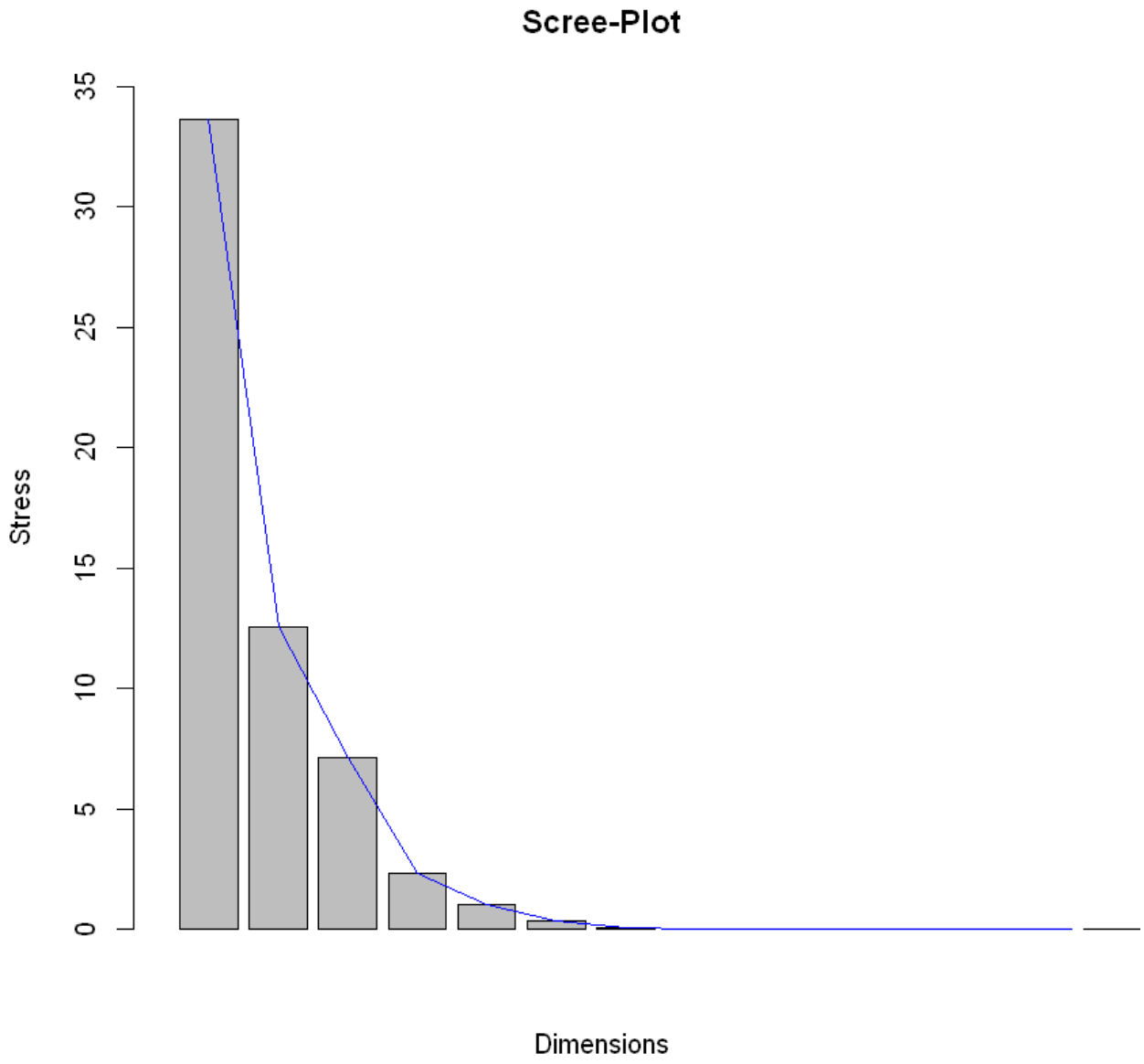

Fig: Scree plot for Kruskal's stress for garden flowers data

A scree plot was generated having stress for multiple dimensions for this data. The stress kept on decreasing with increasing dimensions as expected. The stress related to the 2-dimensional non-MDS was found to be **12.57%**. According to Kruskal's measure of fit, the fit is **Good**.

Ques. 6: The last sheet in the attached MS-Excel file *Assignment_2_data.xlsx,* labeled **USairpollution**, contains observations on seven variables, collected in a study of air pollution in 41 cities in the USA. The variables are:

i. *SO2*: SO2 content of air in micrograms per cubic metre
ii. *temp*: average annual temperature in degrees Fahrenheit
iii. *manu*: number of manufacturing enterprises employing 20 or more workers
iv. *popul*: population size (1970 census) in thousands
v. *wind*: average annual wind speed in miles per hour

vi. *precip*: average annual precipitation in inches
vii. *predays*: average number of days with precipitation per year

a) Using sulphur dioxide content (SO2) as the response variable and the remaining six variables as explanatory variables, fit a linear regression model by least squares.
b) Generate the residual plot and comment.
c) Test whether the regression is significant.
d) Perform appropriate tests of hypotheses to infer the significance of each explanatory variable in the regression model.
e) Obtain 95% confidence intervals for the regression coefficients that were found to be significantly different from 0 in part (c).
f) Obtain the 95% confidence interval for the mean sulphur dioxide content when the vector of observations on the predictors is

$$\mathbf{x}_0 = (20,55,440,500,10.0,11.75,80)^T$$

g) Obtain the 95% prediction interval for the mean sulphur dioxide content when the vector of observations on the predictors is $\mathbf{x}_0$ as given in part (f).
h) Use appropriate regression diagnostic tools to identify influential observations.
i) Repeat the regression analysis of parts (a)-(d) above after removing whatever cities you think should be regarded as outliers.

Solution:

The data on US air pollution contains city wise data of air quality parameters. Sulphur Dioxide (SO2) content is to treated as a response variable when treated against other explanatory variables. A linear model was fit onto the given data and the result/summary obtained is as follows:

```
Call:
lm(formula = SO2 ~ temp + manu + popul + wind + precip + predays,
    data = air)

Residuals:
    Min      1Q  Median      3Q     Max
-23.004  -8.542  -0.991   5.758  48.758

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.72848   47.31810   2.361 0.024087 *
temp         -1.26794    0.62118  -2.041 0.049056 *
manu          0.06492    0.01575   4.122 0.000228 ***
popul        -0.03928    0.01513  -2.595 0.013846 *
wind         -3.18137    1.81502  -1.753 0.088650 .
precip        0.51236    0.36276   1.412 0.166918
predays      -0.05205    0.16201  -0.321 0.749972
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.64 on 34 degrees of freedom
Multiple R-squared:  0.6695,    Adjusted R-squared:  0.6112
F-statistic: 11.48 on 6 and 34 DF,  p-value: 5.419e-07
```

Fig: Initial linear fit on air pollution data with all variables

As evident from the above result, the adjusted r-squared value is **0.6112** but only a few explanatory variables are significant as can be seen from their respective t-values. The residual plot is given as follows:

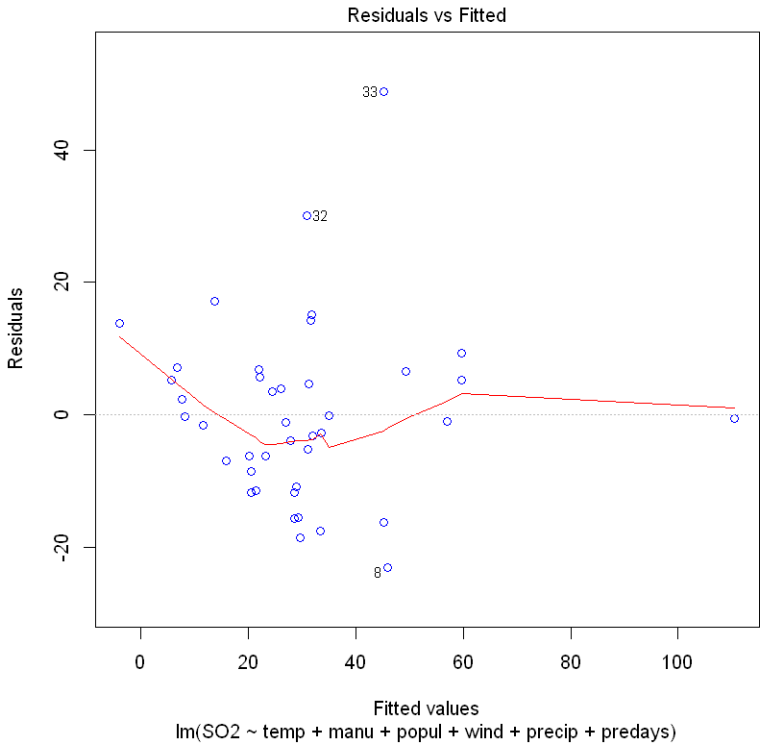

Fig: Residual plot for linear fit

The residuals seems to be homogeneously scattered around the x-axis. Hence, our initial implicit assumption that the system is **homoscedastic**, is valid.

Please note that the **p-value** of this fit is **5.419e-07**. This means that our null hypothesis that none of the explanatory variables can explain the variation in the response variable must be rejected. Hence, the regression is significant.

For testing significance of each explanatory variables in the regression, a t-test can be performed separately but the results of this linear fit already have the t-value calculated for these variables.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.72848   47.31810   2.361 0.024087 *
temp         -1.26794    0.62118  -2.041 0.049056 *
manu          0.06492    0.01575   4.122 0.000228 ***
popul        -0.03928    0.01513  -2.595 0.013846 *
wind         -3.18137    1.81502  -1.753 0.088650 .
precip        0.51236    0.36276   1.412 0.166918
predays      -0.05205    0.16201  -0.321 0.749972
```

Fig: Results from t-test of individual variables

As from the above result, the variables **temp, manu** and **popul** are very significant (with significance level of **5%**) in determining the level of SO2 in a city in US.

Confidence Interval: The confidence interval of the regression coefficients was estimated as below. The confidence intervals of the significant variables are highlighted.

	2.5 %	97.5 %
(Intercept)	15.56653024	207.890431035
temp	-2.53032976	-0.005552425
manu	0.03291387	0.096922472
popul	-0.07003016	-0.008523320
wind	-6.86992838	0.507196811
precip	-0.22484804	1.249565962
predays	-0.38130196	0.277201580

Fig: Confidence Interval of coefficient of significant variables

Prediction for new data point: The new data point was fed into the linear model and the expected value was generated and cross checked against the actual observed value for SO2 content. The predicted value with upper and lower limits of 95% confidence interval for mean SO2 content and the corresponding prediction interval are as follows:
Fit: 20.96 **Actual Value:** 20
Confidence Interval: (7.63, 34.29)
Prediction Interval: (-11.63, 53.55)

As expected theoretically, the prediction interval is found to be **wider** than the confidence interval for a new data point.

Influential Observations:

To identify the influential observations, Cook's distance and Studentized residuals plot was plotted for each observation. Studentized residuals plot have a threshold of **3,** whereas for Cook's Distance any observation falling above (**4/total_observation**) was treated as influential. The Studentized residuals plot and Cook's distance plot is as given below:

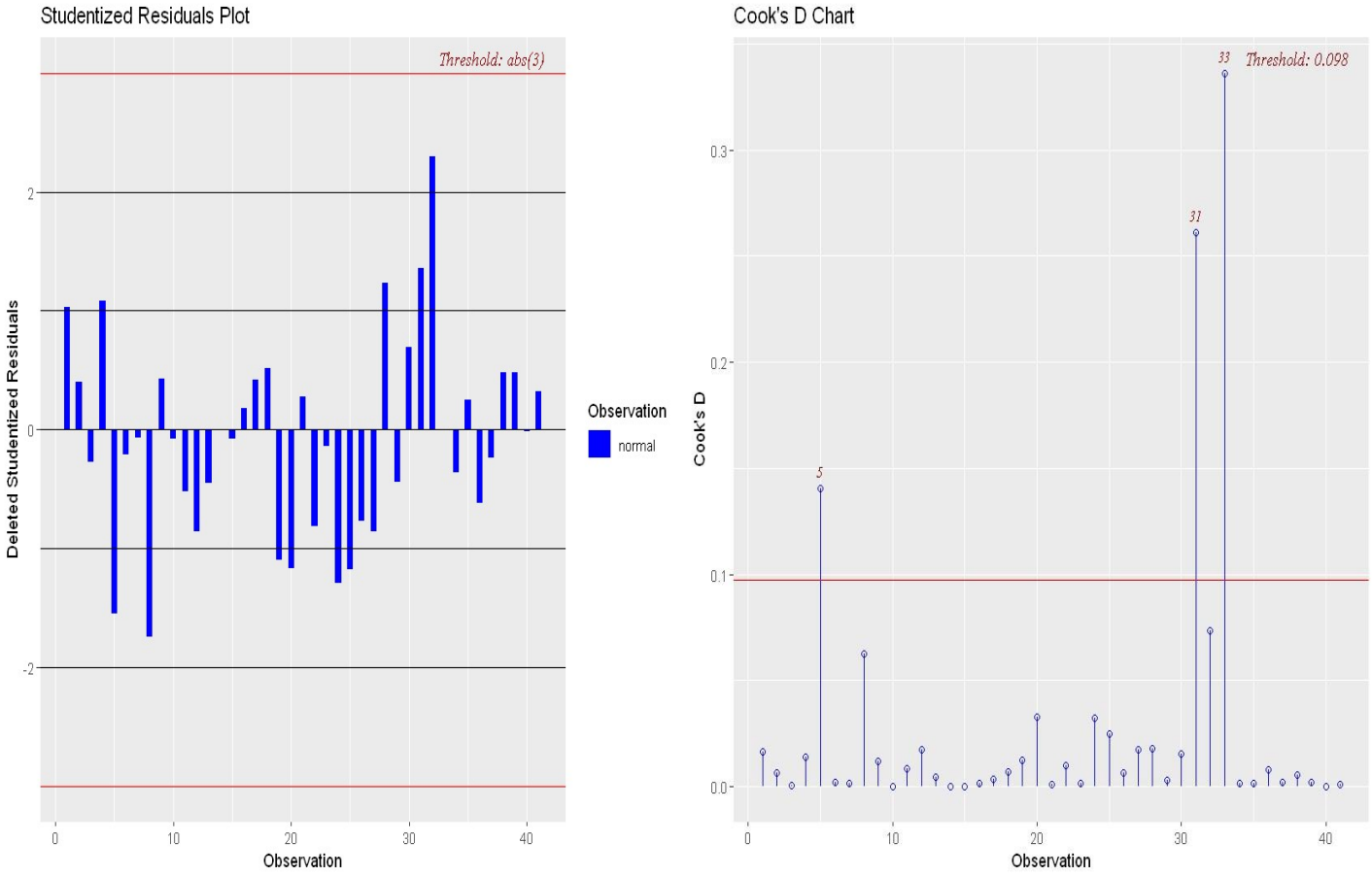

Fig: Stundentized residuals plot and Cook's distance chart for the US air pollution data

As can be observed from the above charts, observation **#5, #31** and **#33** are influential as compared to other observations as per Cook's Distance chart. Hence, these observations must be dropped from the data list and the model must be fit again.

New Model: A linear regression model was fit again on the remaining data and the results that were obtained from the fit are as below:

```
Call:
lm(formula = SO2 ~ temp + manu + popul + wind + precip + predays,
    data = air_new)

Residuals:
    Min      1Q  Median      3Q     Max
-19.695  -7.717  -1.569   6.620  26.303

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.30214   39.49851   1.831 0.076804 .
temp        -1.00086    0.54114  -1.850 0.073931 .
manu         0.05172    0.01244   4.159 0.000234 ***
popul       -0.02634    0.01206  -2.184 0.036652 *
wind        -2.15003    1.60830  -1.337 0.191007
precip       0.28885    0.33744   0.856 0.398558
predays      0.12473    0.13864   0.900 0.375226
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.24 on 31 degrees of freedom
Multiple R-squared:  0.7719,    Adjusted R-squared:  0.7277
F-statistic: 17.48 on 6 and 31 DF,  p-value: 1.005e-08
```

Fig: linear fit on air pollution data with non-influential variables

The above result has r-squared value as **0.7277** (**> 0.6112**) and also the p-value (**1.005e-08**) signifies that the regression is significant. Hence, removing the influential cities from the data improved the model performance.

The residual plot analysis for this fit is as follows:

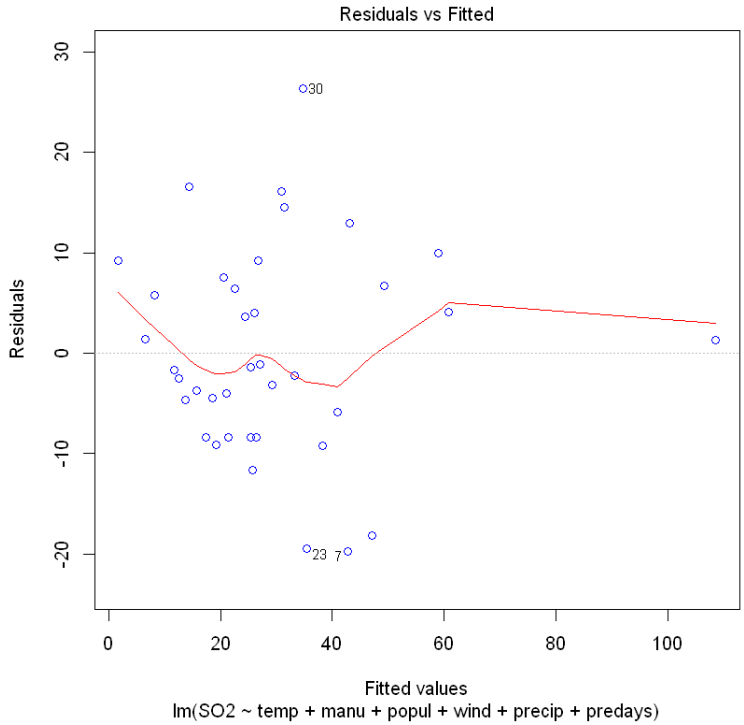

Fig: Residual plot for linear fit without influential observations

There are still some outliers as can be seen from this graph. But the system went more **homoscedastic**.

The significance of each regression variable was tested by using a t-test and the results are as below:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.30214   39.49851   1.831 0.076804 .
temp        -1.00086    0.54114  -1.850 0.073931 .
manu         0.05172    0.01244   4.159 0.000234 ***
popul       -0.02634    0.01206  -2.184 0.036652 *
wind        -2.15003    1.60830  -1.337 0.191007
precip       0.28885    0.33744   0.856 0.398558
predays      0.12473    0.13864   0.900 0.375226
```

Fig: Results from t-test of individual variables

As can be seen from the results, variables **manu** and **popul** are significant again (at 5% significance level) but the significance of the variable **temp** has decreased from before because its significance was primarily due to the influence of some cities.

Therefore, it can be concluded that the removing influential cities from data can increase model performance.
