Statistical Structures in Data

Assignment 2

Anurag Shukla (22BM6JP08)

Solution to question 1:

(i) The loadings of Principal Components for Dispersion Matrix (S) is –

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Cement	-0.904	-0.023	-0.152	0.013	-0.154	0.277	-0.184	0.155	0.011
Blast Furnace Slag	0.255	-0.789	-0.071	0.201	-0.101	0.434	-0.183	0.188	0.012
Fly Ash	0.239	0.299	0.049	-0.686	-0.188	0.495	-0.194	0.248	-0.003
Water	-0.005	-0.075	0.042	-0.076	0.094	-0.468	-0.071	0.833	0.247
Superplasticizer	0.001	0.005	-0.024	-0.020	-0.023	0.101	0.056	-0.222	0.967
Coarse Aggregate	0.013	0.276	0.760	0.479	-0.062	0.275	-0.076	0.173	0.042
Fine Aggregate	0.212	0.446	-0.613	0.481	0.146	0.256	-0.102	0.227	0.027
Age	-0.100	-0.070	0.118	-0.147	0.946	0.204	-0.113	-0.028	0.001
Concrete compressive strength	-0.067	-0.040	-0.020	-0.032	0.045	0.279	0.926	0.233	-0.029

The loadings of Principal Components for Correlation Matrix (R) is –

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Cement	-0.041	0.536	-0.360	-0.310	-0.055	-0.390	-0.134	0.298	-0.473
Blast Furnace Slag	-0.163	0.136	0.699	0.076	-0.363	0.270	0.005	0.229	-0.451
Fly Ash	0.370	-0.268	-0.020	0.601	0.228	-0.320	0.247	0.255	-0.386
Water	-0.564	-0.118	0.120	0.047	0.296	-0.306	-0.010	-0.586	-0.356
Superplasticizer	0.536	0.248	0.188	0.166	-0.037	-0.083	-0.614	-0.448	-0.053
Coarse Aggregate	-0.060	-0.225	-0.549	0.222	-0.545	0.348	-0.060	-0.243	-0.337
Fine Aggregate	0.382	-0.187	-0.001	-0.528	0.384	0.409	0.175	-0.140	-0.419
Age	-0.262	0.252	-0.170	0.360	0.529	0.510	-0.344	0.226	-0.040
Concrete compressive strength	0.107	0.630	-0.034	0.225	0.000	0.154	0.626	-0.347	0.061

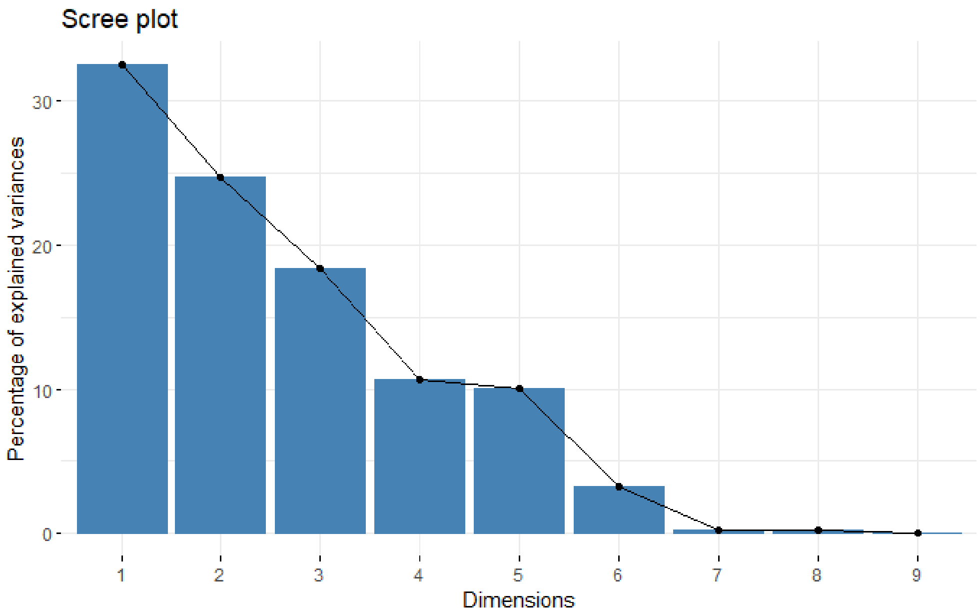
(ii) Variance of Principal Components for Dispersion Matrix (S) is –

PC	1	2	3	4	5	6	7	8	9
Variance	12897.94	9825.43	7287.26	4247.63	3986.92	1268.12	102.07	69.75	11.25

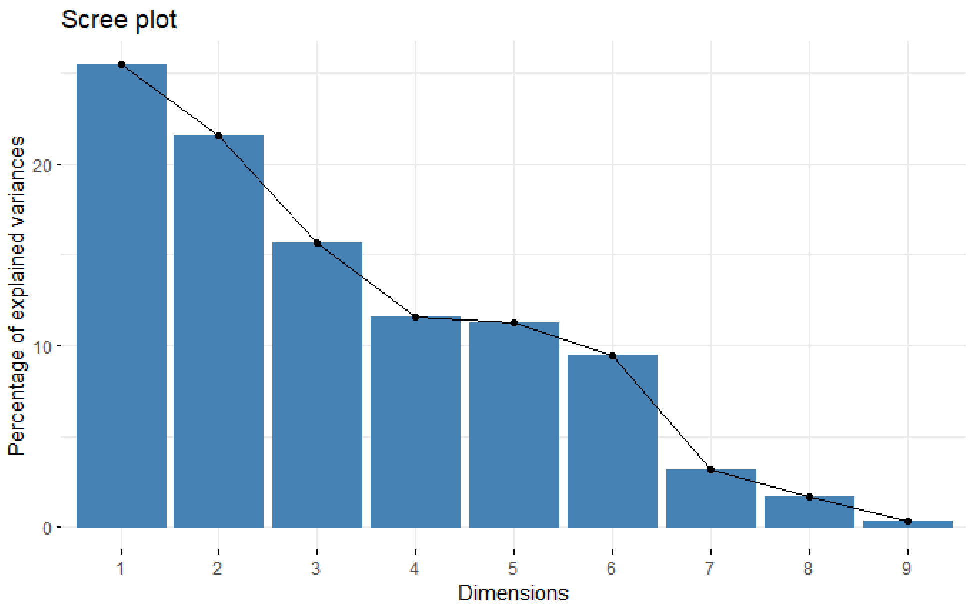
Variance of Principal Components for Correlation Matrix (R) is –

PC	1	2	3	4	5	6	7	8	9
Variance	2.2877	1.9365	1.4089	1.0427	1.0141	0.8474	0.2869	0.1467	0.0287

(iii) Scree Plot for Dispersion Matrix (S) –



Scree Plot for Correlation Matrix (R) –

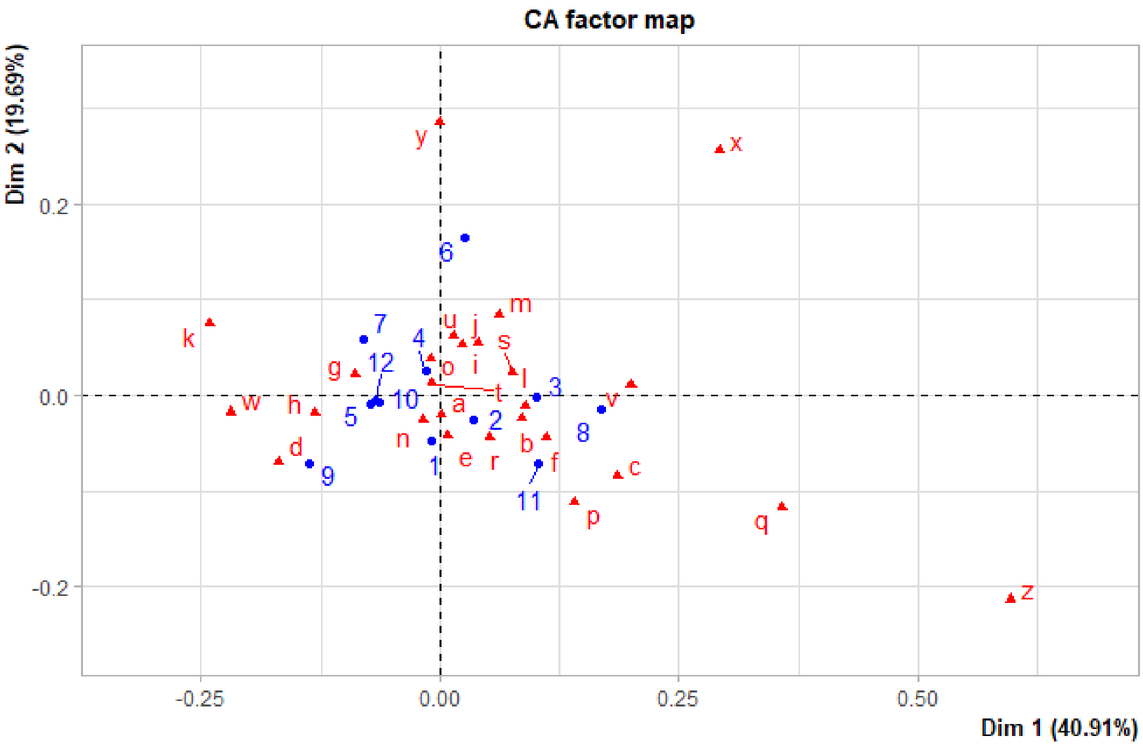


(iv) For Dispersion Matrix, first 5 principal components explain more than 90% variance of the data (96.34%).
For Dispersion Matrix, first 6 principal components explain more than 90% variance of the data (94.86%).

Solution to question 2:

(i) The Correspondence Analysis is done using CA() function from FactoMineR library in R.

(ii) The 2-Dimensional plot for the CA is given below –



(iii) The percentage proportion explained by dimensions of CA is calculated from the eigen values of C matrix. That is,

$$\%Variance = \frac{\lambda 1 + \lambda 2}{\sum_{i=1}^{R} \lambda i}$$

Where, λ_1 and λ_2 are first two-dimension eigen values and R is total number of dimensions.

	Eigenvalue	Percentage of Variance	Cumulative Percentage of Variance
dim 1	0.007664	40.90704	40.90704
dim 2	0.003688	19.687	60.59403
dim 3	0.002411	12.87016	73.46419
dim 4	0.001383	7.381117	80.84531
dim 5	0.001002	5.34651	86.19182
dim 6	0.000723	3.860898	90.05271
dim 7	0.000659	3.51538	93.56809

dim 8	0.000455	2.427824	95.99592
dim 9	0.000374	1.995822	97.99174
dim 10	0.000263	1.404109	99.39585
dim 11	0.000113	0.604151	100

For given dataset, first two dimension explain 60.6% variance in the data. If usually want to have this value above 80%. Therefore, the plot is not very reliable in describing the association between rows and columns.

(iv) Following are the observations from the 2-D CA plot –

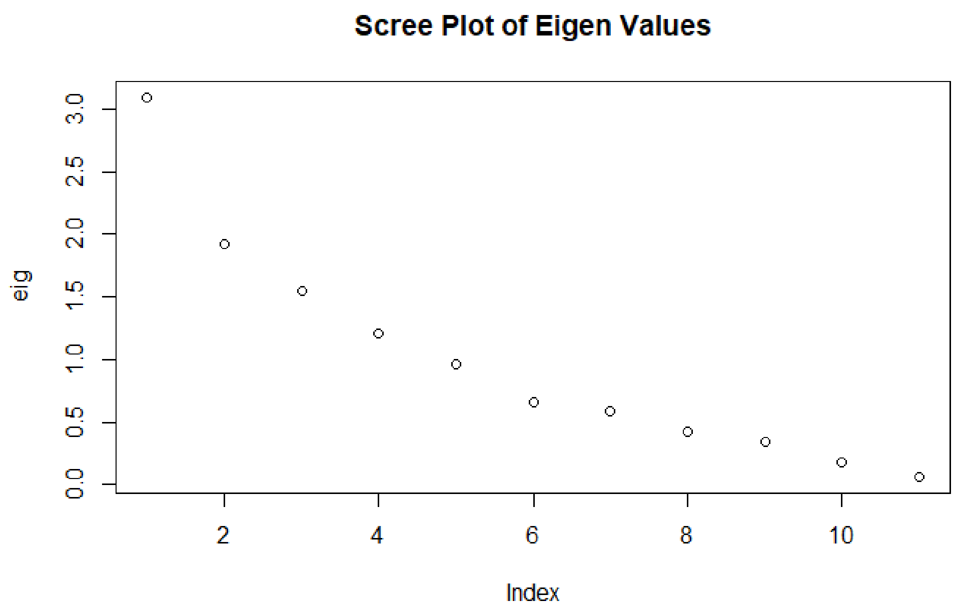
1. Letters ‘z’, ‘q’ and ‘x’ are least associated with any other letters.
2. Vowels ‘a’, ‘e’, ‘I’, ‘o’ and ‘u’ have high association.
3. Novels “Farewell to Arms (5)”, “Pendorric 3(10)” and “Pendorric 2(12)” are highly associated.
4. Novels “Sound and Fury 7(6)” and “Islands(9)” are less associated with other novels.
5. High association is present between novel “Three Daughters (1)” and letters ‘e’ and ‘a’.

Solution to question 3:

(i) For a factor model we use the following equation to determine the feasible number of factors that can be accommodated for the given dataset.

$$s = \frac{1}{2}[(p - k)^2 - (p + k)]$$

Here, ‘p’ is the number of features/columns in the data and ‘k’ is number of factors, For s = 0, we get exact solution and s > 0 overdetermined solution. For given problem, we have, p = 11. Therefore, possible values of ‘k’ are {1,2,3,4,5,6}.



Two ways to determine number of factors are –
1. Take all factors with eigen values more than 1.
2. From scree plot, choose value for which elbow bend is seen in the curve

Here, the first method is used. First 4 eigen values are more than 1. Therefore, the factor model with k = 4 is chosen.

(ii) Factor Loadings without rotation are –

	Factor1	**Factor2**	**Factor3**	**Factor4**
fixed.acidity	0.438	-0.521	0.628	0.203
volatile.acidity	0.123	0.328	-0.202	-0.218
citric.acid	0.162	-0.430	0.499	0.373
residual.sugar	0.185	0.325	0.379	0.125
chlorides	0.245	-0.093	-0.020	0.099
free.sulfur.dioxide	0.027	0.451	-0.104	0.575
total.sulfur.dioxide	0.160	0.467	-0.148	0.750
density	0.871	0.031	0.486	-0.014
pH	-0.329	0.632	-0.168	-0.439
sulphates	0.037	-0.075	0.245	0.167
alcohol	-0.856	0.016	0.512	0.000

Factor interpretations for un-rotated model –
1. Factor 1 load on density variable so could be an indicator "density of wine".
2. Factor 2 load on fixed.acidity, citric.acidity and pH variables so could be an indicator of "Acidity of wine."
3. Factor 3 load on fixed.acidity and acohol variables so could be an indicator of "alcohol in wine."
4. Factor 4 load on free.sulfur.dioxide and total.sulfur.dioxide variables so could be an indicator of "Sulphides of wine."

Factor Loadings with rotation (varimax) are –

	Factor1	**Factor2**	**Factor3**	**Factor4**
fixed.acidity	0.790	-0.273	-0.227	0.386
volatile.acidity	-0.438	-0.107	0.024	0.087
citric.acid	0.747	-0.064	-0.009	0.193
residual.sugar	0.043	0.042	0.195	0.508
chlorides	0.112	-0.252	0.035	0.037
free.sulfur.dioxide	-0.049	0.001	0.733	0.085
total.sulfur.dioxide	0.015	-0.142	0.894	0.098
density	0.240	-0.587	-0.101	0.763
pH	-0.756	0.388	-0.005	0.086
sulphates	0.265	0.036	0.053	0.144
alcohol	0.227	0.968	-0.082	0.013

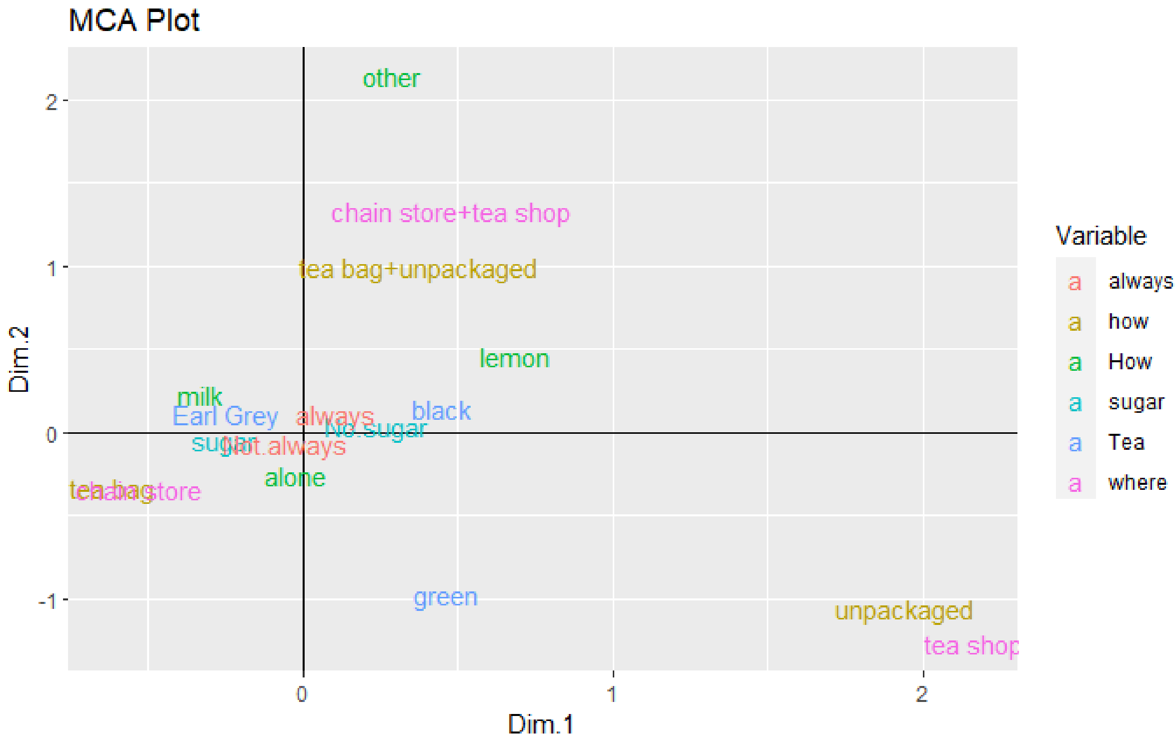
Factor interpretations for rotated model –

1. Factor 1 is heavily loaded on pH, fixed.acidity and citric.acidity so can be classified as “Wine Acidity factor.”
2. Factor 2 is heavily loaded on alcohol variable so can be classified as “Wine Alcohol factor.”
3. Factor 3 is heavily loaded on free.sulfur.dioxide and total.sulfur.dioxide so can be classified as “Wine Sulphide factor.”
4. Factor 4 is heavily loaded on density so can be classified as “Wine Thickness Factor.”

Solution to question 4:

(i) The Correspondence Analysis is done using CA() function from FactoMineR library in R.

(ii) The 2-Dimensional plot for the CA is given below –



(iii) The percentage proportion explained by dimensions of CA is calculated from the eigen values of C matrix. That is,

$$\%Variance = \frac{\lambda 1 + \lambda 2}{\sum_{i=1}^{R} \lambda i}$$

Where, λ_1 and λ_2 are first two-dimension eigen values and R is total number of dimensions.

Dimensions	**Eigenvalue**	**Percentage of Variance**	**Cumulative Percentage of Variance**
dim 1	0.279762	15.25973	15.25973
dim 2	0.257748	14.05897	29.3187
dim 3	0.220138	12.00752	41.32622
dim 4	0.18793	10.25071	51.57693

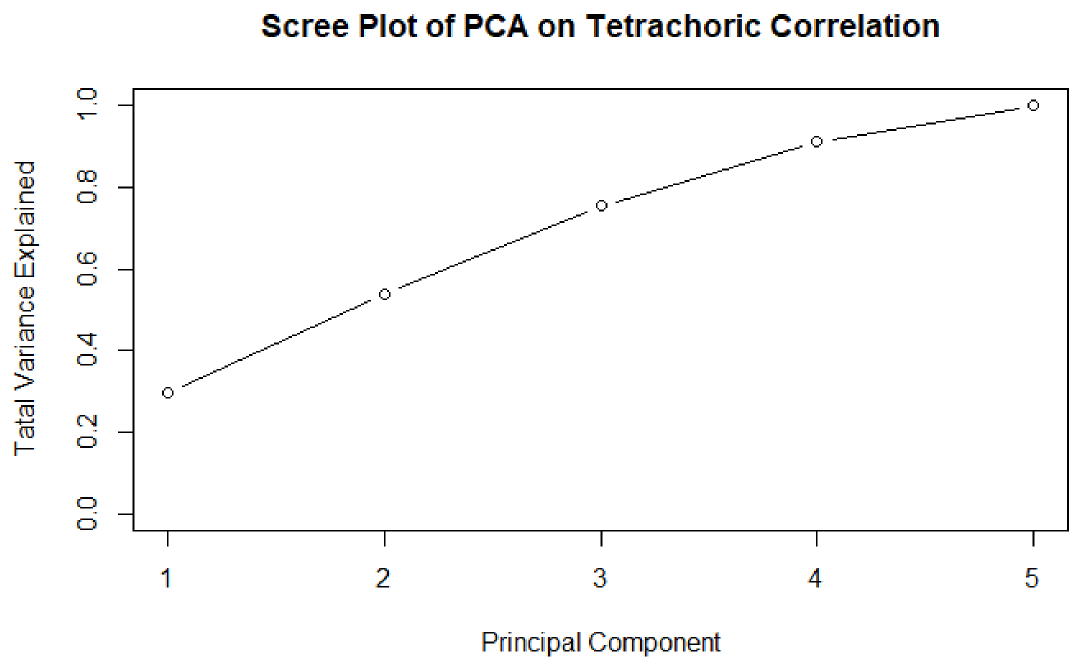
dim 5	0.168765	9.205361	60.78229
dim 6	0.163687	8.928363	69.71065
dim 7	0.152888	8.339364	78.05002
dim 8	0.138387	7.548372	85.59839
dim 9	0.115692	6.310455	91.90885
dim 10	0.086126	4.697802	96.60665
dim 11	0.062211	3.393353	100

The cumulative variance explained by first two dimensions is 29.32% which is very low compared to usual standard of 80%. So, the 2-Dimension plot is not reliable for depicting the association between rows and columns of the dataset.

(iv) The tetrachoric correlation matrix of the transformed data is –

	sophisticated	**slimming**	**exciting**	**relaxing**	**effect.on.health**
sophisticated	1.000	0.120	0.179	0.128	-0.010
slimming	0.120	1.000	0.132	0.074	0.184
exciting	0.179	0.132	1.000	-0.402	0.014
relaxing	0.128	0.074	-0.402	1.000	-0.178
effect.on.health	-0.010	0.184	0.014	-0.178	1.000

The Scree Plot for the principal components with Cumulative Variance –



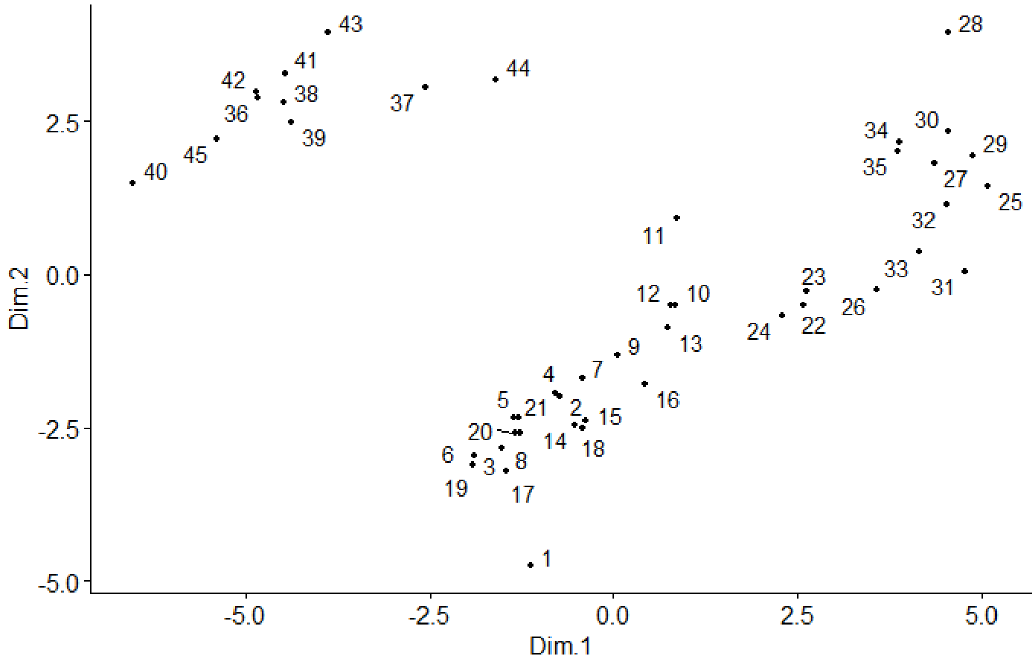
First 4 principal components explain more than 90% of the variance.

Principal Component	**1**	**2**	**3**	**4**	**5**
% Variance	0.294	0.243	0.216	0.157	0.090
Cumulative	0.294	0.538	0.754	0.910	1.000

Solution to question 5:

(i) The distance matrix is computed for the 45 pots using dist() function in R.

(ii) Metric MDS was performed on the generated distance matrix for the dataset. From the 2-Dimension plot of the MDS we can identify 3 clusters in the data.

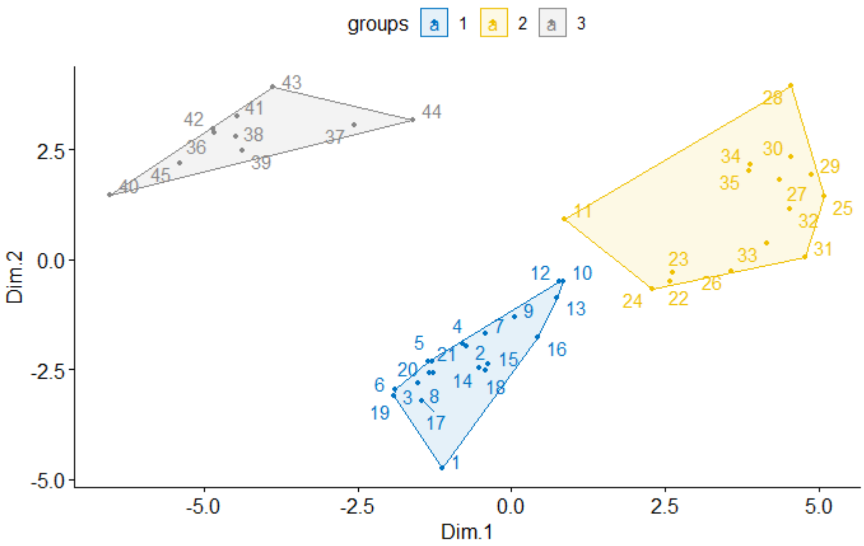


From initial observation, we can say that the pots 1 to 21 (except for 11) form a cluster, pots 22 to 35 form another cluster and rest of the pots from a cluster. That is a total of 3 clusters with pot 11 being an outlier point.

(iii) Combining the information on kilns and regions we get following observation –

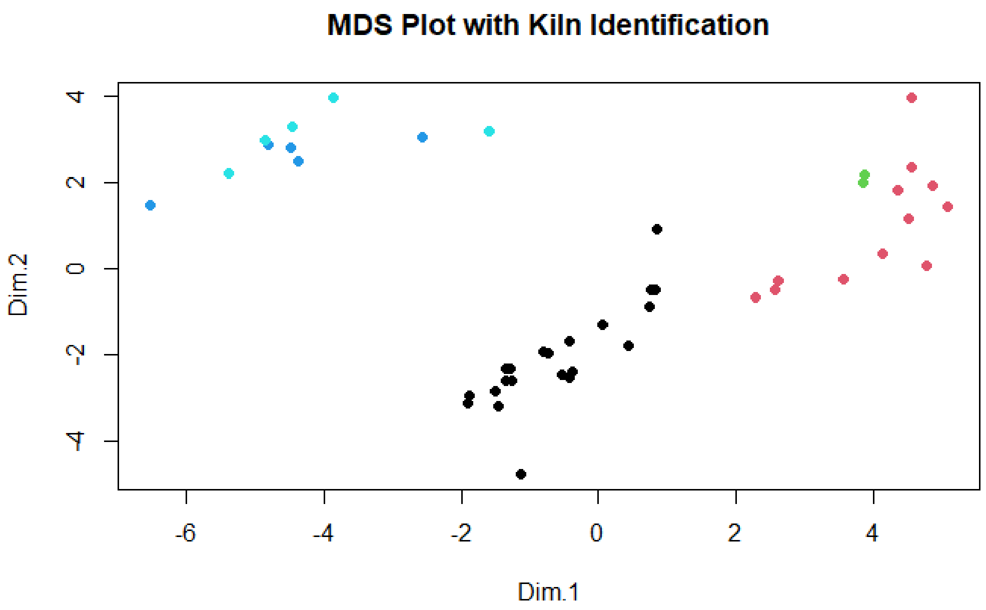
Pots	Kiln	Region
1-21	1	1
22-33	2	2
34-35	3	2
36-40	4	3
41-45	5	3

Given this information, we use K Means Clustering to visualize the data with K = 3 for three regions present. The following is the result of clustering.



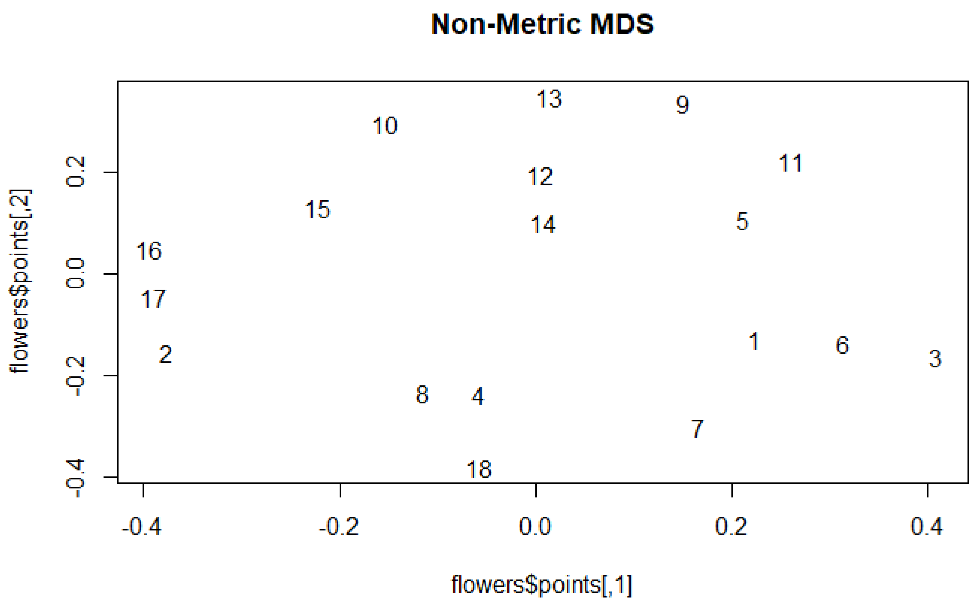
We observe that except for Pot 11, all pots made in same region follow a general clustering. The observation is similar as previous case. That is, pot behaviour is based on region of manufacturing rather than kiln used.

Following plot shows MDS with different colours for kiln of pots –



Solution to question 6:

(i) To perform non-metric MDS, the isoMDS() function from "MASS" library is used. The 2-Dimension plot of result is given below.



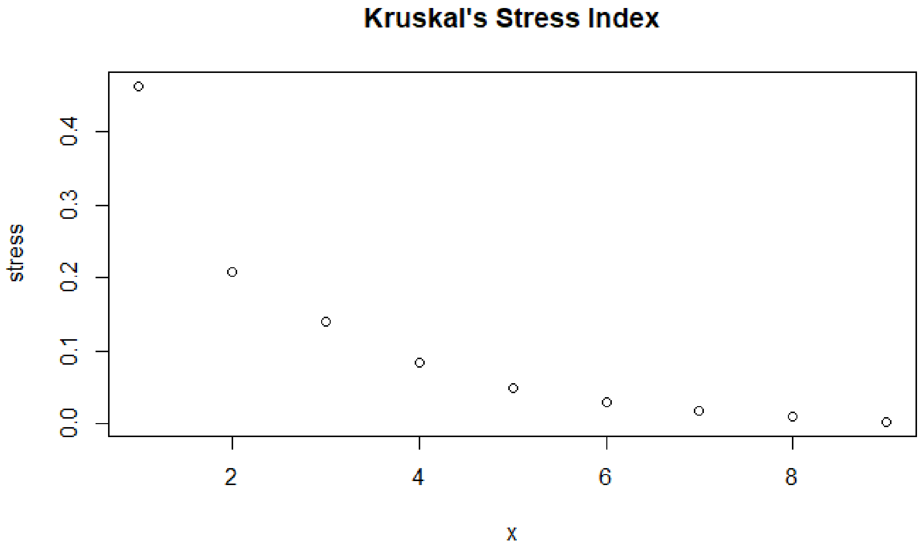
Following associations can be seen from the plot –

1. Broom (2), Red rose (16) and Scotch rose (17) are closely associated.
2. Begonia (1), Camellia (3) and Fuchsia (6) are closely associated.
3. Dahlia (4), Gladiolus (8) and Tulip (18) are closely associated.
4. Lily (12), Lily-of-valley (13) and Peony (14) are closely associated.
5. Forget-me-not (5) and Iris (11) are closely associated.

(ii) The Kruskal's Stress for different dimensions is as follow –

Dimension	Stress Value	Proportion
1	42.09	0.46
2	18.88	0.20
3	12.64	0.14
4	7.58	0.83
5	4.43	0.05
6	2.63	0.03
7	1.66	0.02
8	0.86	0.01
9	0.17	0.01

The Scree plot for Kruskal's Stress Value for different dimensions –



(iii) The Kruskal's Index for 2-dimension MDS of the given data is around 0.207 which is considered to be bad. General norm is –

Kruskal's Index	Quality
> 0.20	Bad
0.05	Good
0	Best

Solution to question 7:

a) The lm() function in R base package is used to carry out multivariate linear regression. Following is the summary of the regression output.

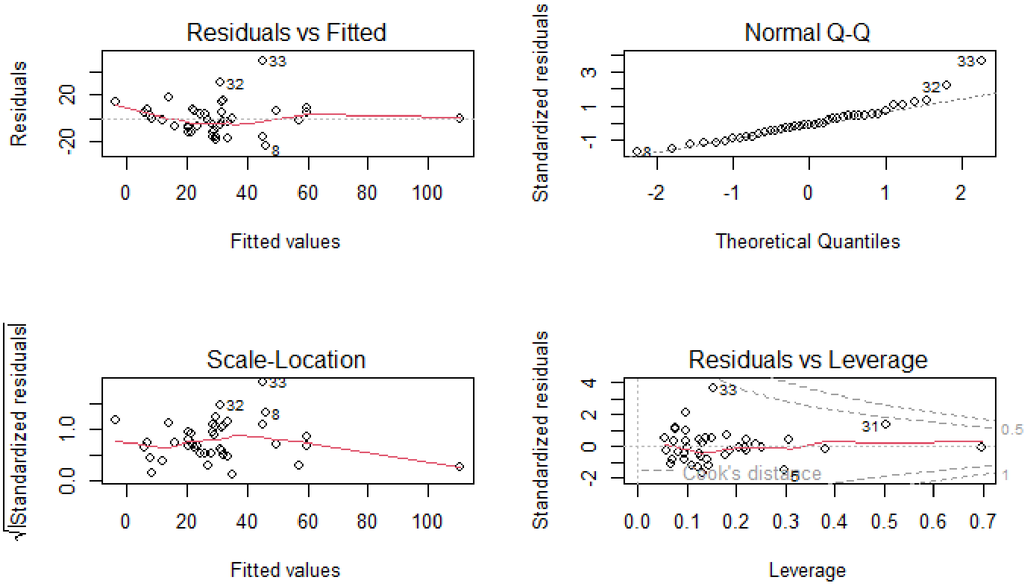
```
lm(formula = SO2 ~ temp + manu + popul + wind + precip + predays,
    data = df)

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
```

b) The residual plots for the fitted model is as follows –



Following can be inferred from residual plots –

1. The Normal QQ plot shows a close resemblance to normality of residuals.
2. Both Residual and standardized residual plots show a decent fit of the data with residuals unbiased and homoscedastic.
3. The leverage plot shows presence of few outliers/influential points in the dataset.
4. Overall, we get a good fit as per residual plots.

c) To check the goodness of fit of regression model, F-Test can be used. The hypothesis test using F-statistic is given by –

$H_0 : \beta_1 = \beta_2 = \beta_3 ..... \beta_k = 0$ $\quad$ H_1 : Atleast one of them is non-zero

Where, β_i's are linear regression coefficients. The parameters of F-statistic are 'k' and 'n-k-1' where 'n' is the number of datapoints and 'k' is number of attributes.

For given dataset we have,

$$n = 41 \qquad k= 6 \qquad F \sim (6, 34)$$

The value of F-statistic is, $F_0 = 11.48$

The p-value for $F_0 = 5.4 \times 10^{-7}$

Therefore, the null hypothesis can be rejected as p-value for the F-statistic is very low. Hence the regression is significant.

d) To check the significance of each explanatory variable, T-test can be used. The hypothesis test using T-statistic is given by –

$$H_0 : \beta_i = 0 \qquad\qquad H_1 : \beta_i \neq 0$$

The test results are as below –

Variable	T-Statistic	p-value
temp	2.361	0.049
manu	-2.041	0.0002
popul	4.122	0.014
wind	-2.595	0.089
precip	-1.753	0.167
predays	1.412	0.75
(intercept)	-0.321	0.024

At 5% significance level, the variables temp, manu and popul are significant with p-value less than 0.05. The variable wind is significant at 10% significant level. The variables precip and predays are insignificant with high p-values.

e) The 95% confidence interval of the significant variables are –

Variable	Value	Lower Limit	Upper Limit
temp	-1.268	-2.530	-0.006
manu	0.065	0.033	0.097
popul	-0.039	-0.070	-0.009
wind	-3.181	-6.870	0.507

f)

g) For the given X_0, the predicted value by the model is 20.96.

Interval	Lower Limit	Upper Limit
Confidence	7.632	34.288
Prediction	-11.634	53.554

h) To identify the influential points, the Difference in Fits (DFFITS) is used. The threshold value for DFFITS is given by –

$$Threshold = 2\sqrt{\frac{k}{n}} = 0.765$$

Where 'n = 41' is the number of datapoints and 'k = 6' is number of parameters. With this, the points that crosses the threshold are Buffalo (5), Phoenix (31) and Providence (33).

i) The linear regression is performed after removing the above influential points. The summary of the model is given below –

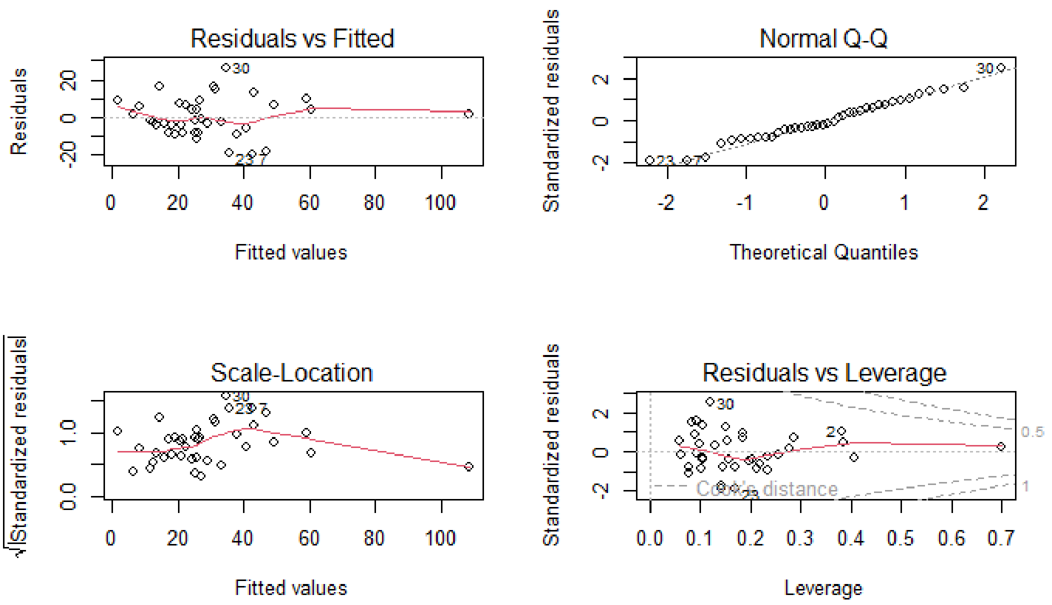
```
lm(formula = SO2 ~ temp + manu + popul + wind + precip + predays,
    data = df1)

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
```

The residual plots are –



Following can be inferred from residual plots –

1. The Normal QQ plot shows a close resemblance to normality of residuals.
2. Both Residual and standardized residual plots show a decent fit of the data with residuals unbiased and homoscedastic.
3. The leverage plot shows that no outliers/influential points are present.

Overall, we get a good fit as per residual plots.

For given dataset we have,

$$n = 38 \qquad k= 6 \qquad F \sim (6, 31)$$

The value of F-statistic is, $F_0 = 17.48$
The p-value for $F_0 = 1.005 \times 10^{-8}$
Therefore, the null hypothesis can be rejected as p-value for the F-statistic is very low. Hence the regression is significant. Note that the F-statistic and p-value has improved after removing influential points compared to previous case.

The T-test is used to check for significance of each of the explanatory variables.

Variable	T-Statistic	p-value
temp	-1.85	0.074
manu	4.159	0.0002
popul	-2.184	0.036
wind	-1.337	0.191
precip	0.856	0.398
predays	0.9	0.37
(intercept)	1.831	0.07

At 5% significance level, only “manu” and “popul” are significant in the linear regression model.