Introduction

K-Means Clustering

K-means clustering is an iterative algorithm that tries to divide the dataset into predefined K different clusters where each data point belongs to only one cluster. It tries to make the inter-cluster data points as close as possible while keeping the clusters as far as possible. Data points are assigned to a clusters using the sum of the squared distance between the data points using centroids. The less cluster variation means, the more homogeneous the data points are in the same cluster.

Elbow Method

Elbow method applies k-means clustering over loop and output the total sum of square over number of clusters graph where it is possible to observe changes.

Silhouette Method

First for the data point we find the silhouette coefficient which is calculated using the mean intra-cluster distance and the mean nearest-cluster distance, then we take the mean of it and based on the graph which also shows the number of clusters, we decide on the optimal cluster size.

PCA

Principle component analysis allows us to extract principle components which explains the variation in the dataset without having to use all of the other components.

Agglomerative Clustering

Agglomerative clustering is a bottom-up strategy where each observation begins in its own cluster and when one goes up the hierarchy, pairs of clusters are combined. It has quadratic time and space complexity which makes it a bad choice if we work with high amount of data.It outputs a dendrogram which then be used for clustering based on observation.

Dataset

TASK1 - Dataset in this assignment contains comprehensive information about student evaluations of teachers from Gazi University. It involves data regarding to instructor identifier, course code which represented as class attribute, nb.repeat as the students course repetition counter, attendance as the level of attendance and level of difficulty student perceived.

```
student_eval… 5820 obs. of 33 variables
```

We have total of 5820 observations.

```
> str(student_evaluation)
List of 33
 $ instr     : Factor w/ 3 levels
 $ class     : Factor w/ 13 levels
 $ nb.repeat : Factor w/ 3 levels
 $ attendance: Factor w/ 5 levels
 $ difficulty: Factor w/ 5 levels
```

There are total of 28 questions with 5 levels which are all Likert-type. Since they are likert-type we can threat them as numeric for extracting certain metrics as-well, that's why I have used the as numeric data for the rest of the assignment.

```
$ Q1        : Factor w/ 5 levels
$ Q2        : Factor w/ 5 levels
$ Q3        : Factor w/ 5 levels
$ Q4        : Factor w/ 5 levels
$ Q5        : Factor w/ 5 levels
$ Q6        : Factor w/ 5 levels
$ Q7        : Factor w/ 5 levels
$ Q8        : Factor w/ 5 levels
$ Q9        : Factor w/ 5 levels
$ Q10       : Factor w/ 5 levels
$ Q11       : Factor w/ 5 levels
$ Q12       : Factor w/ 5 levels
$ Q13       : Factor w/ 5 levels
$ Q14       : Factor w/ 5 levels
```

```
$ Q15       : Factor w/ 5 levels
$ Q16       : Factor w/ 5 levels
$ Q17       : Factor w/ 5 levels
$ Q18       : Factor w/ 5 levels
$ Q19       : Factor w/ 5 levels
$ Q20       : Factor w/ 5 levels
$ Q21       : Factor w/ 5 levels
$ Q22       : Factor w/ 5 levels
$ Q23       : Factor w/ 5 levels
$ Q24       : Factor w/ 5 levels
$ Q25       : Factor w/ 5 levels
$ Q26       : Factor w/ 5 levels
$ Q27       : Factor w/ 5 levels
$ Q28       : Factor w/ 5 levels
```

- Q1: The semester course content, teaching methods and evaluation system were provided at the
start.
- Q2: The course aims, and objectives were clearly stated at the beginning of the period.

- Q3: The course was worth the amount of credit assigned to it.
Q4: The course was taught according to the syllabus announced on the first day of class.

- Q5: The class discussions, homework assignments, applications and studies were satisfactory.

- Q6: The textbook and other courses resources were sufficient and up to date.

- Q7: The course allowed field work, applications, laboratory, discussion and other studies.

- Q8: Quizzes, assignments, projects and exams contributed to helping the learning.

- Q9: I greatly enjoyed the class and was eager to participate during the lectures actively.

- Q10: My initial expectations about the course were met at the end of the period or year.

- Q11: The course was relevant and beneficial to my professional development.

- Q12: The course helped me look at life and the world with a new perspective.

- Q13: The Instructor's knowledge was relevant and up to date.

- Q14: The Instructor came prepared for classes.

- Q15: The Instructor taught in accordance with the announced lesson plan.

- Q16: The Instructor was committed to the course and was understandable.

- Q17: The Instructor arrived on time for classes.

- Q18: The Instructor has a smooth and easy to follow delivery/speech.

- Q19: The Instructor made effective use of class hours.

- Q20: The Instructor explained the course and was eager to be helpful to students.

- Q21: The Instructor demonstrated a positive approach to students.

- Q22: The Instructor was open and respectful of the views of students about the course.

- Q23: The Instructor encouraged participation in the course.

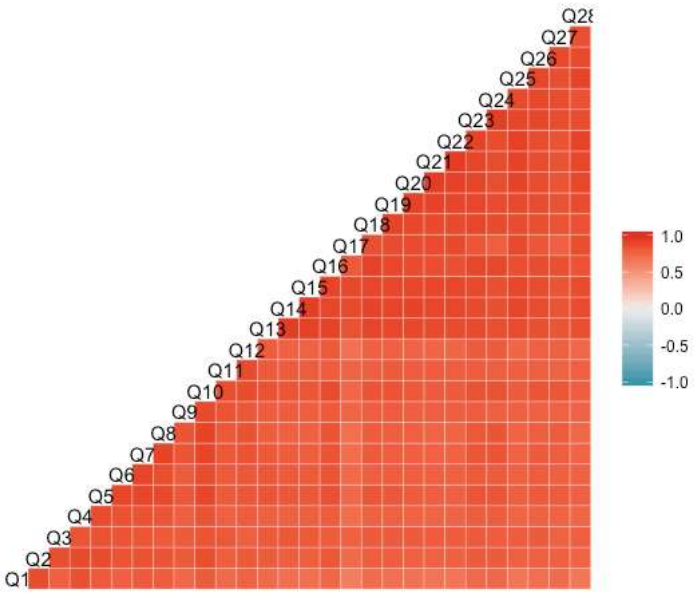
- Q24: The Instructor gave relevant homework assignments/projects and helped/ guided students.

- Q25: The Instructor responded to questions about the course inside and outside of the course.

- Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.)
effectively measured the course objectives.

- Q27: The Instructor provided solutions to exams and discussed them with students.

- Q28: The Instructor treated all students in a right and objective manner.



Running the below line results in correlation matrix as displayed in the above figure.

```
ggcorr(dataset)
```

Which indicates there is medium to strong positive correlation between the features. Below figure gives insight about features distribution, all of them do have 3 likert-type value for their respected mean.

```
       Q1              Q2              Q3              Q4              Q5              Q6       
 Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.00   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :3.00   Median :3.000   Median :3.000   Median :3.000   Median :3.000   Median :3.000  
 Mean   :2.93   Mean   :3.074   Mean   :3.179   Mean   :3.082   Mean   :3.106   Mean   :3.107  
 3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
       Q7              Q8              Q9              Q10             Q11             Q12       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :3.000   Median :3.000   Median :3.000   Median :3.000   Median :3.000   Median :3.000  
 Mean   :3.066   Mean   :3.042   Mean   :3.166   Mean   :3.091   Mean   :3.184   Mean   :3.036  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
      Q13             Q14             Q15             Q16             Q17             Q18       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00    Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.00    1st Qu.:3.000   1st Qu.:2.000  
 Median :3.000   Median :3.000   Median :3.000   Median :3.00    Median :4.000   Median :3.000  
 Mean   :3.243   Mean   :3.291   Mean   :3.287   Mean   :3.17    Mean   :3.398   Mean   :3.223  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.00    Max.   :5.000   Max.   :5.000  
      Q19             Q20             Q21             Q22             Q23             Q24       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :3.000   Median :3.000   Median :3.000   Median :3.000   Median :3.000   Median :3.000  
 Mean   :3.262   Mean   :3.285   Mean   :3.307   Mean   :3.318   Mean   :3.202   Mean   :3.167  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
      Q25             Q26             Q27             Q28       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000  
 Median :3.000   Median :3.000   Median :3.000   Median :3.000  
 Mean   :3.313   Mean   :3.222   Mean   :3.155   Mean   :3.308  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
```

Preprocessing

TASK 2 - In order to apply clustering, I have selected the features (Q1 to Q28) which are Likert-type.

```
dataset = student_evaluation[,c(6:33)]
```

Results

1) K-Means Clustering

When we run K-Means Clustering with different cluster sizes on the data set, results were as follows.

Cluster Size 3

```
K-means clustering with 3 clusters of sizes 2358, 2223, 1239

Cluster means:
        Q1       Q2       Q3       Q4       Q5       Q6       Q7       Q8       Q9      Q10      Q11
1 2.631891 2.831637 3.003817 2.849449 2.868533 2.898643 2.814249 2.786684 2.966497 2.836302 3.000848
2 4.100765 4.261808 4.294647 4.247413 4.313540 4.275753 4.264507 4.239316 4.304543 4.320288 4.334683
3 1.396287 1.403551 1.509282 1.435835 1.390638 1.408394 1.396287 1.379338 1.502825 1.368846 1.467312
       Q12      Q13      Q14      Q15      Q16      Q17      Q18      Q19      Q20      Q21      Q22
1 2.784563 3.116200 3.200170 3.188295 2.983036 3.371077 3.080153 3.141645 3.194656 3.223494 3.227311
2 4.224022 4.393162 4.413855 4.406658 4.377418 4.417454 4.389114 4.409357 4.418803 4.424202 4.436347
3 1.380952 1.419693 1.448749 1.467312 1.357546 1.622276 1.400323 1.430993 1.424536 1.463277 1.481840
       Q23      Q24      Q25      Q26      Q27      Q28
1 3.031807 2.979220 3.223919 3.074215 2.982188 3.223070
2 4.390463 4.358075 4.419253 4.377418 4.311741 4.409807
3 1.393059 1.386602 1.495561 1.430993 1.407587 1.493140
```

Cluster Size 5

```
K-means clustering with 5 clusters of sizes 1915, 729, 1608, 719, 849

Cluster means:
        Q1       Q2       Q3       Q4       Q5       Q6       Q7       Q8       Q9      Q10      Q11
1 2.699217 2.860574 3.024543 2.894517 2.904961 2.933681 2.869452 2.831854 2.966580 2.872585 3.012010
2 1.116598 1.093278 1.115226 1.091907 1.057613 1.052126 1.061728 1.056241 1.131687 1.027435 1.098765
3 3.568408 3.801617 3.853856 3.773010 3.854478 3.797264 3.769900 3.730100 3.858831 3.851990 3.910448
4 1.824757 1.930459 2.173853 1.990264 1.931850 1.994437 1.901252 1.910987 2.186370 1.913769 2.105702
5 4.733804 4.845701 4.870436 4.832744 4.893993 4.899882 4.885748 4.875147 4.879859 4.909305 4.898704
       Q12      Q13      Q14      Q15      Q16      Q17      Q18      Q19      Q20      Q21      Q22
1 2.824543 3.101828 3.175979 3.151958 2.984334 3.310183 3.058486 3.120627 3.156136 3.185901 3.182245
2 1.058985 1.049383 1.038409 1.042524 1.019204 1.111111 1.034294 1.042524 1.024691 1.048011 1.053498
3 3.733831 4.018035 4.054104 4.049129 3.988184 4.079602 4.026741 4.060323 4.078358 4.088930 4.103856
4 1.908206 2.100139 2.212796 2.257302 1.934631 2.603616 2.066759 2.129346 2.184979 2.242003 2.282337
5 4.840989 4.943463 4.951708 4.949352 4.929329 4.944641 4.926973 4.931684 4.948174 4.943463 4.954064
       Q23      Q24      Q25      Q26      Q27      Q28
1 3.012010 2.965535 3.177023 3.039687 2.964491 3.175979
2 1.026063 1.023320 1.049383 1.034294 1.028807 1.043896
3 4.016169 3.970771 4.074005 4.010572 3.923507 4.069652
4 2.038943 2.016690 2.333797 2.168289 2.068150 2.340751
5 4.941107 4.912839 4.948174 4.911661 4.873969 4.926973
```

Cluster size 7

```
K-means clustering with 7 clusters of sizes 840, 388, 579, 724, 459, 1329, 1501

Cluster means:
        Q1       Q2       Q3       Q4       Q5       Q6       Q7       Q8       Q9      Q10      Q11
1 4.720238 4.839286 4.863095 4.817857 4.897619 4.901190 4.890476 4.873810 4.882143 4.913095 4.900000
2 1.590206 1.927835 2.631443 2.018041 2.051546 2.213918 1.958763 1.943299 2.590206 1.976804 2.646907
3 1.889465 1.967185 2.160622 2.032815 1.968912 2.006908 1.970639 1.960276 2.117444 1.965458 2.075993
4 1.110497 1.085635 1.104972 1.087017 1.049724 1.044199 1.055249 1.048343 1.127072 1.022099 1.089779
5 2.307190 2.960784 3.274510 2.838780 2.997821 2.967320 2.784314 2.636166 3.126362 3.067538 3.263617
6 3.823928 3.951843 3.970655 3.967645 3.997743 3.940557 3.940557 3.937547 3.977427 3.975922 4.000000
7 2.952032 3.025316 3.039973 3.044637 3.049300 3.059294 3.036642 3.010660 3.035976 3.013991 3.052632
       Q12      Q13      Q14      Q15      Q16      Q17      Q18      Q19      Q20      Q21      Q22
1 4.839286 4.953571 4.960714 4.960714 4.940476 4.957143 4.938095 4.944048 4.960714 4.955952 4.967857
2 1.981959 2.981959 3.275773 3.250000 2.458763 3.979381 2.873711 3.054124 3.280928 3.474227 3.463918
3 1.948187 1.991364 2.062176 2.094991 1.867012 2.336788 1.960276 2.006908 2.006908 2.034542 2.077720
4 1.045580 1.046961 1.037293 1.041436 1.017956 1.104972 1.033149 1.042818 1.024862 1.048343 1.053867
5 2.793028 3.952070 4.095861 4.080610 3.840959 4.311547 4.050109 4.119826 4.257081 4.270153 4.278867
6 3.891648 4.002257 4.020316 4.015801 3.984951 4.018059 3.990971 4.015801 4.018059 4.031603 4.042889
7 2.994004 3.005330 3.029314 3.015989 2.975350 3.063957 2.962025 2.998001 2.986676 2.986676 2.990007
       Q23      Q24      Q25      Q26      Q27      Q28
1 4.955952 4.923810 4.958333 4.920238 4.882143 4.933333
2 2.693299 2.551546 3.422680 2.907216 2.621134 3.481959
3 1.929188 1.939551 2.131261 2.013817 1.948187 2.119171
4 1.026243 1.023481 1.049724 1.034530 1.029006 1.041436
5 3.928105 3.769063 4.202614 3.949891 3.697168 4.180828
6 3.995485 3.970655 4.025583 3.993228 3.938299 4.023326
7 2.967355 2.954031 3.006662 2.969354 2.957362 3.005330
```

Cluster Size 9

```
K-means clustering with 9 clusters of sizes 210, 1463, 306, 749, 720, 1170, 255, 525, 422

Cluster means:
        Q1       Q2       Q3       Q4       Q5       Q6       Q7       Q8       Q9      Q10      Q11
1 3.023810 3.776190 3.880952 3.490476 3.952381 3.795238 3.719048 3.566667 4.047619 4.033333 4.028571
2 2.941217 3.001367 3.028708 3.025974 3.037594 3.051948 3.036227 3.004785 3.023240 2.995899 3.038278
3 1.555556 1.833333 2.431373 1.941176 1.895425 2.039216 1.741830 1.803922 2.421569 1.856209 2.323529
4 4.877170 4.937250 4.954606 4.961282 4.966622 4.955941 4.962617 4.950601 4.942590 4.967957 4.962617
5 1.106944 1.083333 1.102778 1.083333 1.045833 1.043056 1.050000 1.044444 1.125000 1.019444 1.079167
6 3.928205 4.002564 4.007692 4.031624 4.024786 4.007692 3.991453 3.998291 3.999145 4.011966 4.017094
7 1.776471 2.286275 2.976471 2.349020 2.290196 2.392157 2.180392 2.105882 2.647059 2.262745 2.807843
8 1.918095 1.982857 2.129524 2.041905 2.001905 2.034286 2.017143 2.005714 2.110476 1.998095 2.099048
9 2.682464 3.222749 3.409953 3.085308 3.329384 3.244076 3.156398 3.030806 3.407583 3.353081 3.592417
       Q12      Q13      Q14      Q15      Q16      Q17      Q18      Q19      Q20      Q21      Q22
1 3.628571 4.709524 4.780952 4.752381 4.676190 4.880952 4.776190 4.847619 4.909524 4.880952 4.923810
2 2.991798 2.994532 3.014354 3.008886 2.967874 3.053999 2.959672 2.982912 2.976077 2.967874 2.970608
3 1.800654 2.653595 2.980392 2.934641 2.196078 3.709150 2.500000 2.709150 2.846405 3.058824 3.107843
4 4.915888 4.969292 4.967957 4.973298 4.951936 4.969292 4.951936 4.953271 4.961282 4.966622 4.971963
5 1.037500 1.038889 1.036111 1.038889 1.015278 1.108333 1.031944 1.041667 1.026389 1.045833 1.051389
6 3.952991 3.996581 4.018803 4.013675 3.987179 3.987179 3.980342 3.997436 3.995726 4.005983 4.017094
7 2.160784 3.635294 3.882353 3.886275 3.145098 4.450980 3.690196 3.792157 4.109804 4.215686 4.192157
8 2.000000 1.965714 1.980952 2.043810 1.881905 2.184762 1.940952 1.963810 1.954286 1.973333 2.009524
9 3.132701 3.758294 3.857820 3.793839 3.687204 3.990521 3.763033 3.879147 3.936019 3.969194 3.962085
       Q23      Q24      Q25      Q26      Q27      Q28
1 4.738095 4.514286 4.861905 4.704762 4.409524 4.900000
2 2.954887 2.953520 2.989747 2.961039 2.944634 2.978811
3 2.418301 2.307190 3.163399 2.705882 2.450980 3.241830
4 4.965287 4.939920 4.966622 4.935915 4.915888 4.941255
5 1.025000 1.020833 1.045833 1.034722 1.026389 1.041667
6 3.989744 3.986325 3.995726 3.979487 3.945299 3.994017
7 3.439216 3.172549 4.015686 3.450980 3.141176 4.019608
8 1.918095 1.935238 2.047619 1.975238 1.931429 2.030476
9 3.715640 3.630332 3.954976 3.767773 3.613744 3.931280
```

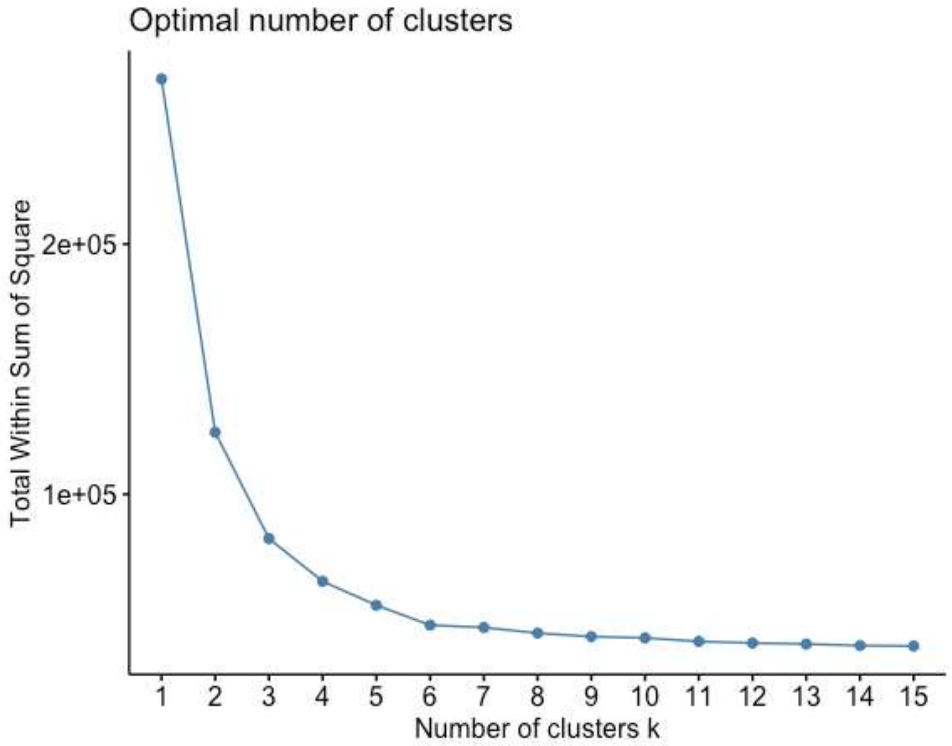
TASK 3 - Elbow Method

I have specified number of cluster size range using the below figure.

```
clusterRange = c(1:15)
maxCluster = max(clusterRange)
```

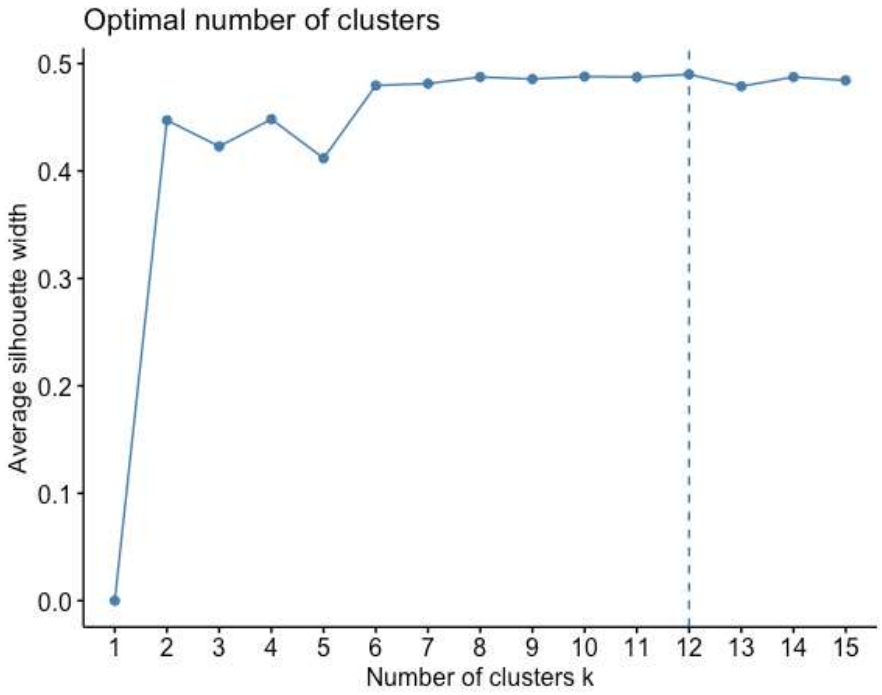
I have obtained the following graph by running the below line. Drastic change occurs when cluster size is incremented to 2 and 3. Since its instructed to choose odd number as the cluster size, I have chosen 3.

```
fviz_nbclust(dataset, kmeans, method = "wss",k.max = maxCluster)
```



TASK 4 - Silhouette Method

```
fviz_nbclust(dataset, kmeans , method = "silhouette",k.max = maxCluster)
```

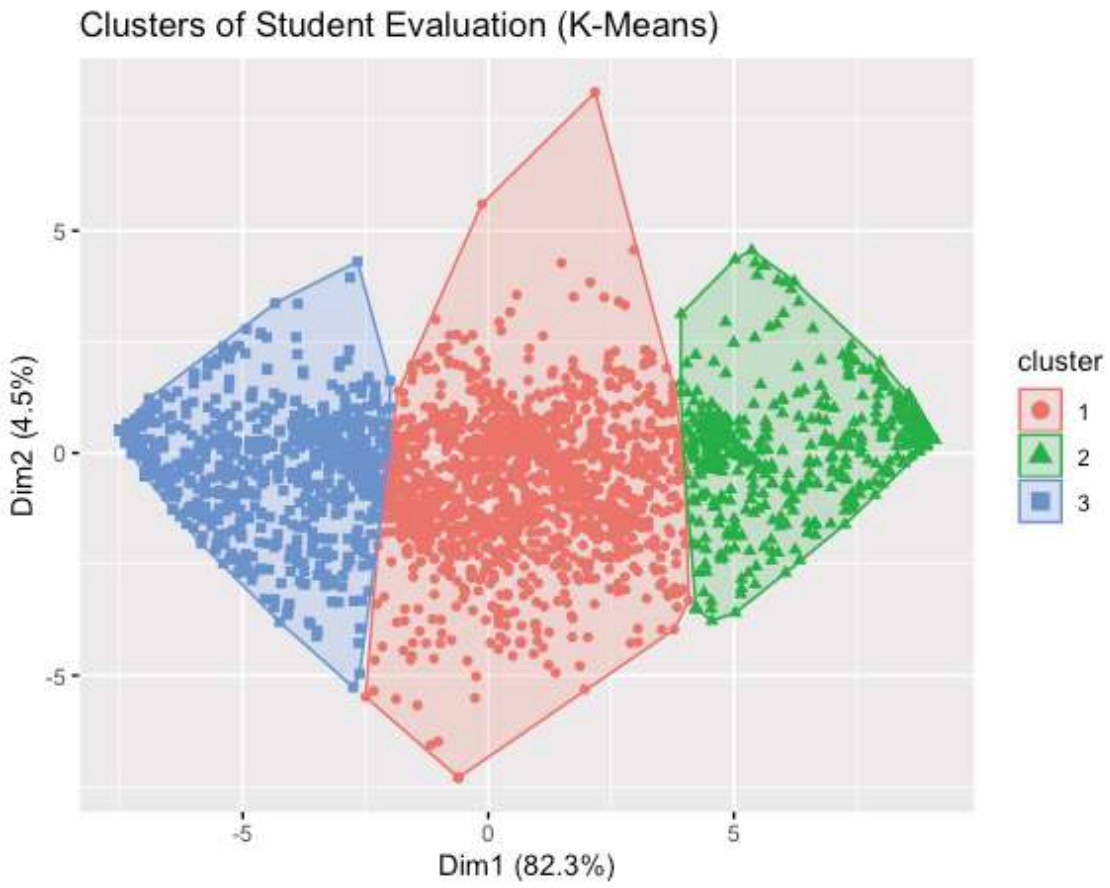


When I have applied the silhouette method and again, value 3 was the candidate for optimal cluster size.

Later, I have applied k-means clustering using cluster size as 3. Below figure represents the corresponding code block.

```
kMeansResult = kmeans(datasetPCA, centers = 3, nstart = 20)
kMeansResult # K-means clustering with 3 clusters of sizes 1239, 2358, 2223
```

```
K-means clustering with 3 clusters of sizes 2358, 1239, 2223
```



TASK 5 - PCA

I have executed the “prcomp” function which is responsible from finding the principle components in the dataset.

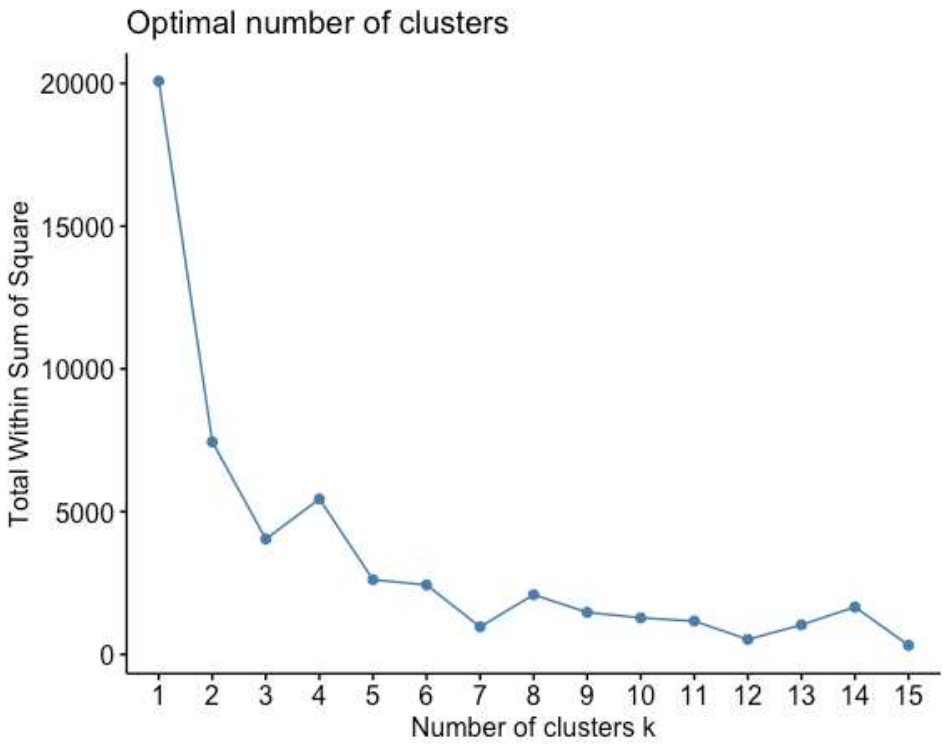
```
datasetPCA = prcomp(dataset, center = TRUE)
summary(datasetPCA)
```

```
Importance of components:
                          PC1     PC2    PC3     PC4     PC5     PC6    PC7     PC8     PC9   PC10    PC11
Standard deviation     6.129 1.43666 0.8169 0.76634 0.68817 0.65281 0.5777 0.54607 0.52703 0.4827 0.47764
Proportion of Variance 0.822 0.04516 0.0146 0.01285 0.01036 0.00932 0.0073 0.00652 0.00608 0.0051 0.00499
Cumulative Proportion  0.822 0.86714 0.8817 0.89459 0.90495 0.91427 0.9216 0.92810 0.93417 0.9393 0.94426
                          PC12    PC13    PC14   PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.47149 0.44491 0.43642 0.4328 0.42369 0.41829 0.40532 0.39378 0.38956 0.37073
Proportion of Variance 0.00486 0.00433 0.00417 0.0041 0.00393 0.00383 0.00359 0.00339 0.00332 0.00301
Cumulative Proportion  0.94913 0.95346 0.95763 0.9617 0.96565 0.96948 0.97307 0.97647 0.97979 0.98279
                          PC22    PC23    PC24   PC25   PC26    PC27    PC28
Standard deviation     0.36744 0.36181 0.35278 0.3379 0.3313 0.29799 0.28881
Proportion of Variance 0.00295 0.00286 0.00272 0.0025 0.0024 0.00194 0.00182
Cumulative Proportion  0.98575 0.98861 0.99133 0.9938 0.9962 0.99818 1.00000
```

Output was as follows, PC1 - Question 1 and PC2 - Question 2 combined, explains 86.71% variance of the dataset.

Then I have reduced the dataset to size in which only principle components are available and then applied elbow method to figure out the optimal cluster size.

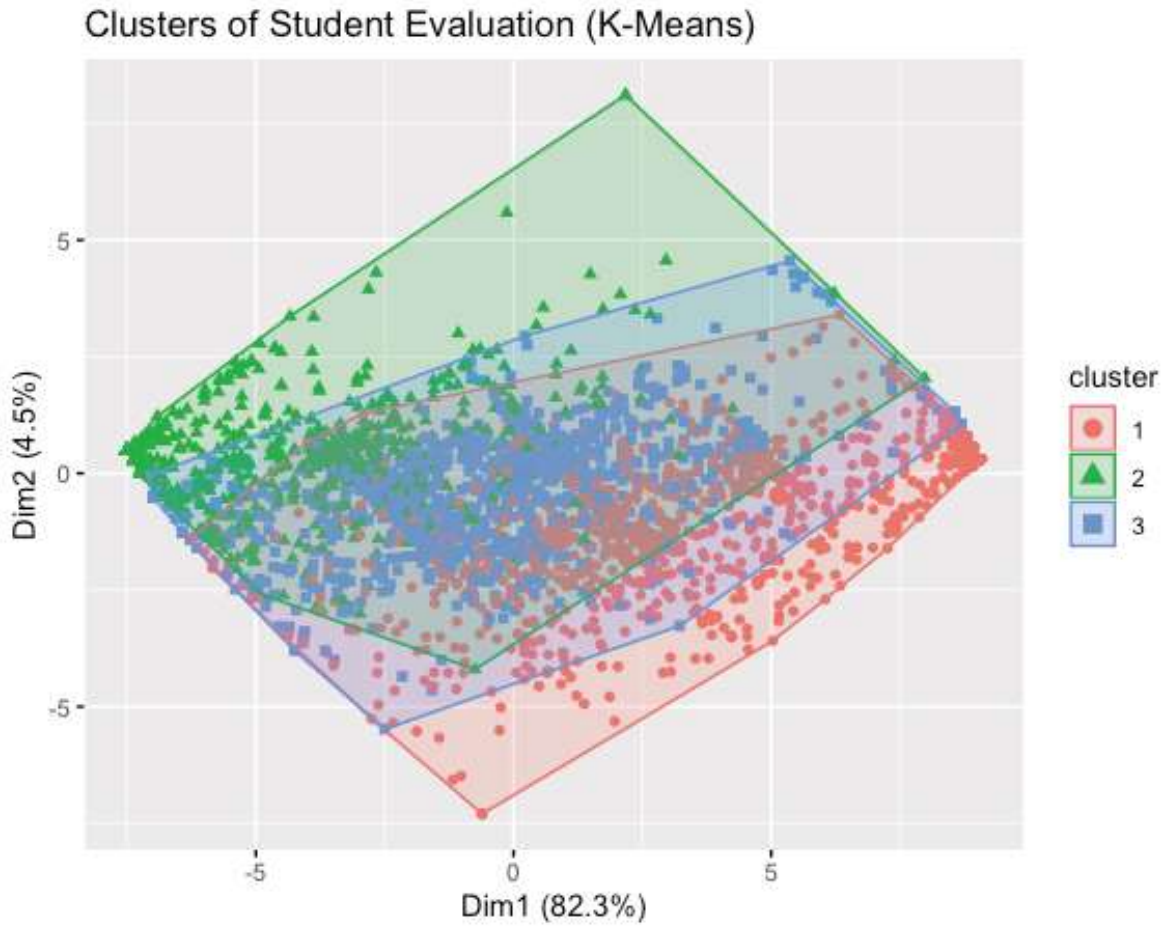
```
datasetPCA = dataset[,principleComps]
summary(prcomp(datasetPCA))
```



As previous results, again 3 was the optimal cluster size, so I have ran “kmeans” using dataset having only principle components and 3 as the number of clusters.

```
kMeansResult = kmeans(datasetPCA, centers = 3, nstart = 20)
kMeansResult # K-means clustering with 3 clusters of sizes 1767, 2043, 2010
```

Below figure also illustrates the result of the clustering using the dataset with only having principle components.

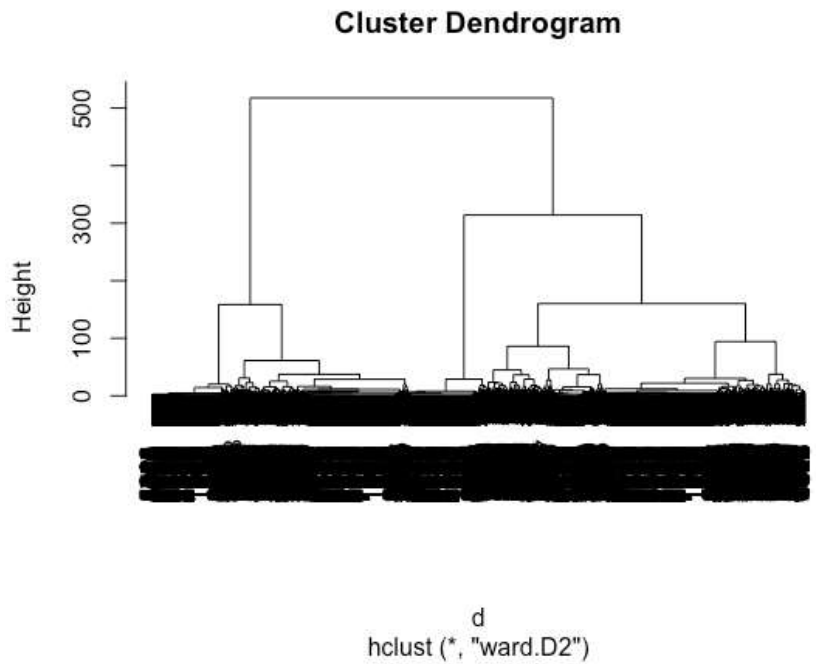

Clusters of Student Evaluation (K-Means)
Dim2 (4.5%)
5
0
-5
-5
0
5
Dim1 (82.3%)
cluster
1
2
3

3) Agglomerative Clustering

TASK 6 -

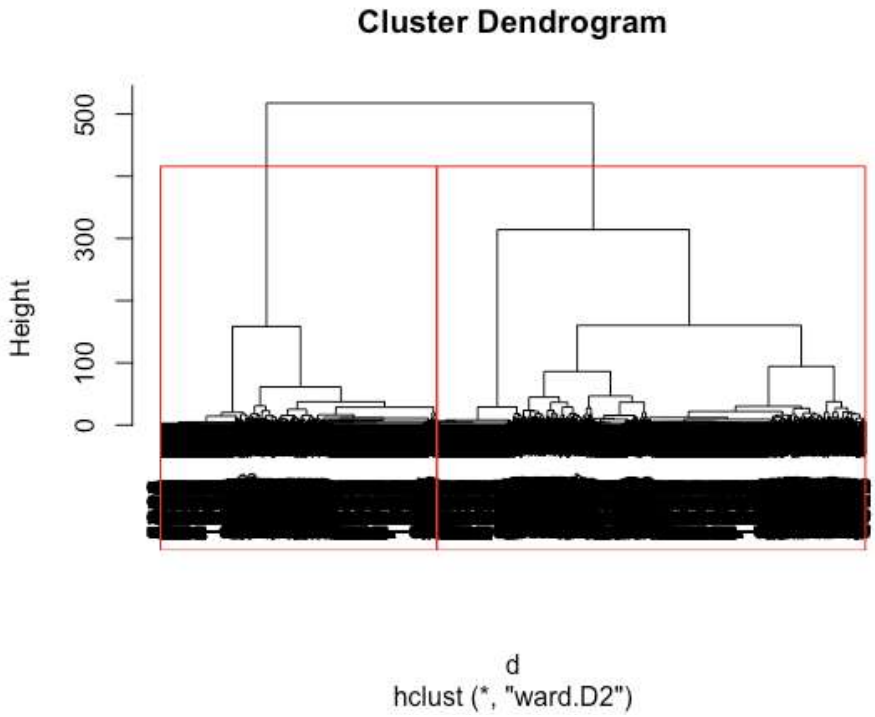
First I have created a distance matrix named as “d” and then I have applied the hierarchical clustering.

```
d = dist(dataset)
hc = hclust(d,method="ward.D2")
plot(hc)
```

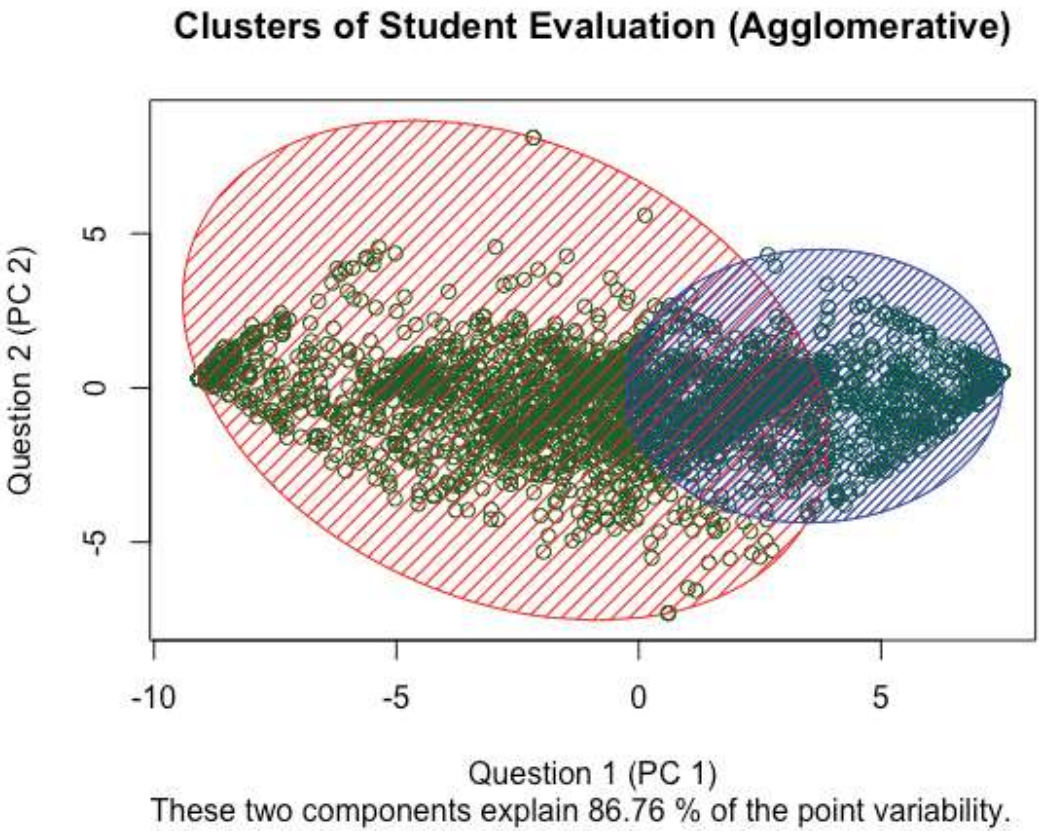


It resulted as in the above dendrogram. I have cut the tree to 2 using the below code.

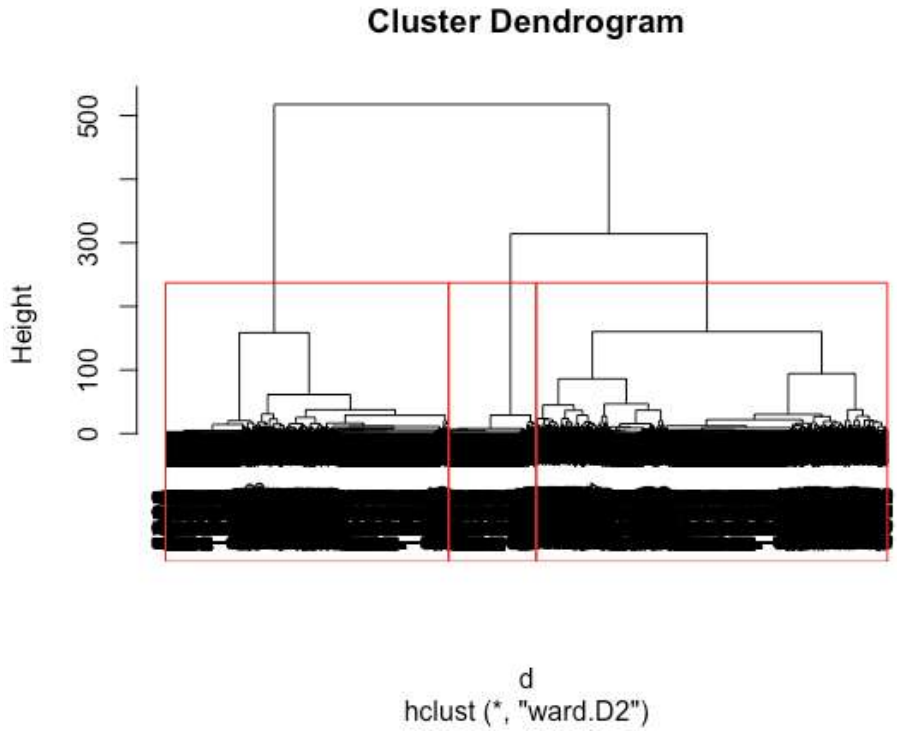
```
hc_2_cluster = cutree(hc,k=2)
plot(hc)
rect.hclust(hc, k=2, border="red")
visualiseAgglomerative(hc_2_cluster)
```

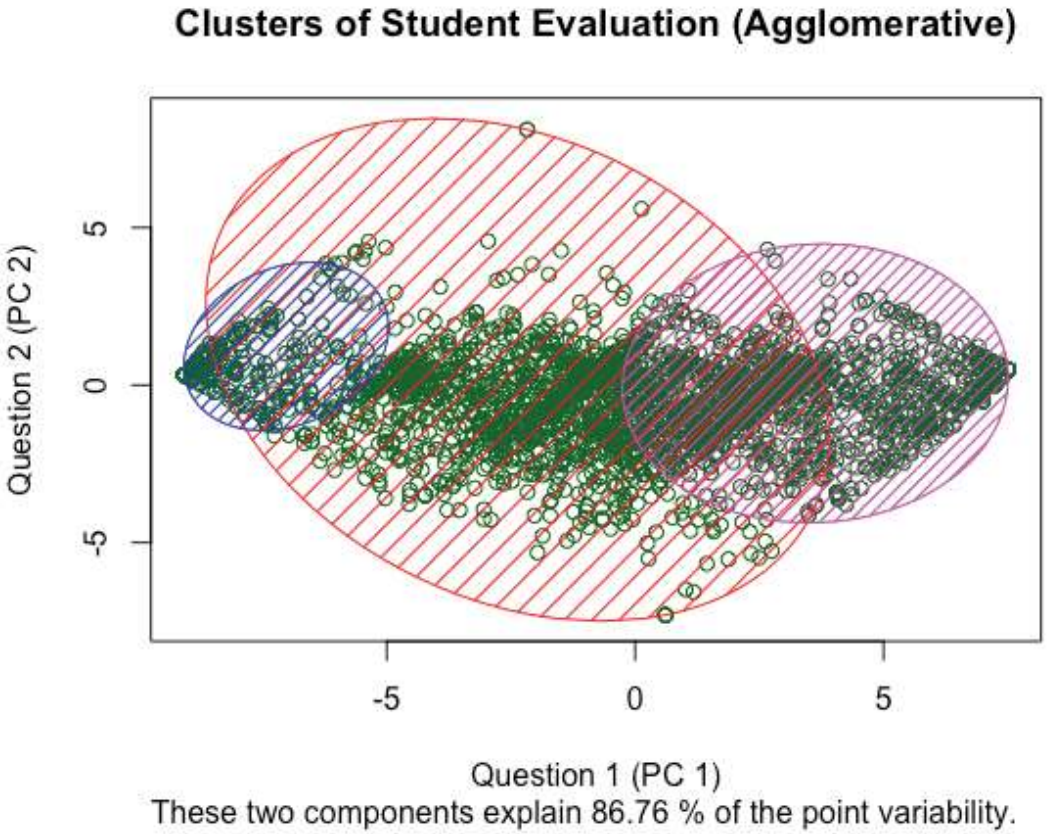


And then, I have visualized the result using the “visualiseAgglomerative” function. It resulted in the below figure.



I have applied the same routine but this time I have cut the tree to 3 parts. Results were as follows.





Shortcut for Task Answers

TASK 1 - I have shown the descriptive summary, statistics and correlation matrix with figures and I have written brief explanations about my findings.

TASK 2 - Since questions are likert-type, I have just separated the questions from the data set for pre-processing.

TASK 3 - Running elbow method resulted in optimal cluster size as 3.

TASK 4 - Yes, using silhouette method I had 3 as the optimal cluster size as-well.

TASK 5 - After applying PCA, dataset with only PC resulted in cluster size 3 (using elbow method)

TASK 6 - I have found 2 clusters using ward's method.

TASK 7 -

One of the biggest difference between hierarchical clustering and k-means clustering is the amount of data they can handle in a given time, hierarchical clustering have quadratic complexity whereas it's linear in k-means. Also, k-means might generate different results for different cluster sizes, however, in hierarchical clustering we have the fixed dendrogram. In k-means we need some information about the cluster size before the clustering to obtain a good result, elbow method and silhouette methods do help in this process whereas we do not require such analysis in the hierarchical clustering. K-means is sensitive to scaling whereas hierarchical is sensitive to large data, k-means can change the sample's cluster, however when hierarchical assigns a branch to the sample it cannot change it. Finally, since hierarchical clustering generates a dendrogram it is easier to interpret the results.

Conclusion

In this assignment, I have learnt the fundamental workflow that is required to be followed when applying clustering. I have used methods for determining the optimal cluster size for k-means clustering and interpreting dendrogram which was generated by applying the hierarchical clustering. Also, I have applied PCA on the dataset and re-run k-means clustering and observed the differences. Finally I have applied hierarchical clustering and I have visualized my findings for each task.