Homework 3B

Harshith Bondada
M10724854

Consider the data files used for Homework 3 for the following tasks.

1. **Consider the BCP dataset and its class variable with values "R" (Recurrence Occurred) and "N" (No Recurrence Occurred so far). Ignore the attribute that gives the number of years after which recurrence occurred or the number of years for which the patient has been free of recurrence. There are thirty other attribute values given as features measured for every patient. Use only these thirty attributes and perform the following:**

   a. **Run k-means algorithm with this dataset for k=4. Run it three different times and for each run show the cluster centers and the SSE values for each cluster and the total SSE value for the clustering.**

   First-time Execution:
   The below screen shot represents the Cluster centers of each attribute for the first-time execution of the algorithm. We can see that the SSE value is 92809449. The individual SSE value of the clusters are also provided in the below screenshot.

```
library(cluster)
library(ggplot2)
Data=read.csv("wpbc.data.csv")
Cluster <- kmeans(Data[,3:34], 4, nstart = 20, iter.max = 1)
```

```
Cluster means:
  radius.1 texture.1 perimeter.1    area.1 smoothness.1 compactness.1 concavity.1 concave.points.1
1 14.17236  21.67083    93.27083  625.3167    0.1058803     0.1391426   0.1312992       0.06651750
2 20.45229  22.46646   135.66042 1301.1583    0.1017167     0.1584346   0.1959981       0.11020292
3 17.70273  22.32470   116.35758  975.3288    0.1001609     0.1329308   0.1443685       0.08413061
4 23.55636  24.63000   156.11818 1745.9091    0.1017845     0.1585545   0.2216091       0.13449091
  symmetry.1 fractal.dimension.1  radius.2 texture.2 perimeter.2    area.2 smoothness.2 compactness.2
1  0.1949569          0.06651208 0.3985431  1.221642    2.878444  35.25028  0.006923750     0.03153642
2  0.1955979          0.06075146 0.7807312  1.222690    5.531646 101.52437  0.006686396     0.03278667
3  0.1897742          0.06052636 0.6096697  1.330680    4.251424  67.96833  0.006890712     0.03074527
4  0.1843727          0.05933455 1.1299364  1.272627    7.748636 176.05818  0.005470182     0.02633455
  concavity.2 concave.points.2 symmetry.2 fractal.dimension.2 radius.3 texture.3 perimeter.3
1  0.03867125       0.01383031 0.02087860         0.004129361 16.77750  30.28903    112.3356
2  0.04513104       0.01617663 0.02084437         0.003894438 25.00604  29.38542    167.8667
3  0.04057864       0.01580148 0.02049717         0.004011561 21.09258  29.94182    139.8258
4  0.03702909       0.01496727 0.01784727         0.003354727 30.93727  32.99909    206.8091
     area.3 smoothness.3 compactness.3 concavity.3 concave.points.3 symmetry.3 fractal.dimension.3
1  866.0806    0.1533246     0.3992064   0.4450233        0.1645104  0.3412611          0.10101014
2 1917.5833    0.1399794     0.3745333   0.4746812        0.2007725  0.3210000          0.08603062
3 1361.7879    0.1378444     0.3264879   0.3968244        0.1712867  0.3111485          0.08469061
4 2951.4545    0.1382455     0.3480818   0.4666273        0.2267455  0.2956091          0.08282364
  Tumor.size Lymph.node.status
1   2.318056          2.986111
2   3.481250          4.000000
3   2.943939          2.939394
4   2.772727          1.909091
```

```
> Cluster$totss
[1] 92809449
```

```
> cluster$withinss
[1] 2766961 2494725 2795136 3878790
```

Second-time Execution:
The below screen shot represents the Cluster centers of each attribute for the second-time execution of the algorithm. We can see that the SSE value is 92809449. The individual SSE value of the clusters are also provided in the below screenshot.

```
Cluster means:
  radius.1 texture.1 perimeter.1    area.1 smoothness.1 compactness.1 concavity.1 concave.points.1
1 20.45229  22.46646   135.66042 1301.1583    0.1017167     0.1584346   0.1959981       0.11020292
2 17.70273  22.32470   116.35758  975.3288    0.1001609     0.1329308   0.1443685       0.08413061
3 14.17236  21.67083    93.27083  625.3167    0.1058803     0.1391426   0.1312992       0.06651750
4 23.55636  24.63000   156.11818 1745.9091    0.1017845     0.1585545   0.2216091       0.13449091
  symmetry.1 fractal.dimension.1  radius.2 texture.2 perimeter.2    area.2 smoothness.2 compactness.2
1  0.1955979          0.06075146 0.7807312  1.222690    5.531646 101.52437  0.006686396     0.03278667
2  0.1897742          0.06052636 0.6096697  1.330680    4.251424  67.96833  0.006890712     0.03074527
3  0.1949569          0.06651208 0.3985431  1.221642    2.878444  35.25028  0.006923750     0.03153642
4  0.1843727          0.05933455 1.1299364  1.272627    7.748636 176.05818  0.005470182     0.02633455
  concavity.2 concave.points.2 symmetry.2 fractal.dimension.2 radius.3 texture.3 perimeter.3
1  0.04513104       0.01617663 0.02084437         0.003894438 25.00604  29.38542    167.8667
2  0.04057864       0.01580148 0.02049717         0.004011561 21.09258  29.94182    139.8258
3  0.03867125       0.01383031 0.02087860         0.004129361 16.77750  30.28903    112.3356
4  0.03702909       0.01496727 0.01784727         0.003354727 30.93727  32.99909    206.8091
     area.3 smoothness.3 compactness.3 concavity.3 concave.points.3 symmetry.3 fractal.dimension.3
1 1917.5833    0.1399794     0.3745333   0.4746812        0.2007725  0.3210000          0.08603062
2 1361.7879    0.1378444     0.3264879   0.3968244        0.1712867  0.3111485          0.08469061
3  866.0806    0.1533246     0.3992064   0.4450233        0.1645104  0.3412611          0.10101014
4 2951.4545    0.1382455     0.3480818   0.4666273        0.2267455  0.2956091          0.08282364
  Tumor.size Lymph.node.status
1   3.481250          4.000000
2   2.943939          2.939394
3   2.318056          2.986111
4   2.772727          1.909091
```

```
> cluster$totss
[1] 92809449
> cluster$withinss
[1] 2795136 2766961 3878790 2494725
```

Third-time Execution:
The below screen shot represents the Cluster centers of each attribute for the third-time execution of the algorithm. We can see that the SSE value is 92809449. The individual SSE value of the clusters are also provided in the below screenshot.

```
Cluster means:
  radius.1 texture.1 perimeter.1    area.1 smoothness.1 compactness.1 concavity.1 concave.points.1
1 17.70273  22.32470   116.35758  975.3288    0.1001609     0.1329308   0.1443685       0.08413061
2 14.17236  21.67083    93.27083  625.3167    0.1058803     0.1391426   0.1312992       0.06651750
3 23.55636  24.63000   156.11818 1745.9091    0.1017845     0.1585545   0.2216091       0.13449091
4 20.45229  22.46646   135.66042 1301.1583    0.1017167     0.1584346   0.1959981       0.11020292
  symmetry.1 fractal.dimension.1  radius.2 texture.2 perimeter.2    area.2 smoothness.2 compactness.2
1  0.1897742          0.06052636 0.6096697  1.330680    4.251424  67.96833  0.006890712     0.03074527
2  0.1949569          0.06651208 0.3985431  1.221642    2.878444  35.25028  0.006923750     0.03153642
3  0.1843727          0.05933455 1.1299364  1.272627    7.748636 176.05818  0.005470182     0.02633455
4  0.1955979          0.06075146 0.7807312  1.222690    5.531646 101.52437  0.006686396     0.03278667
  concavity.2 concave.points.2 symmetry.2 fractal.dimension.2 radius.3 texture.3 perimeter.3
1  0.04057864       0.01580148 0.02049717         0.004011561 21.09258  29.94182    139.8258
2  0.03867125       0.01383031 0.02087860         0.004129361 16.77750  30.28903    112.3356
3  0.03702909       0.01496727 0.01784727         0.003354727 30.93727  32.99909    206.8091
4  0.04513104       0.01617663 0.02084437         0.003894438 25.00604  29.38542    167.8667
     area.3 smoothness.3 compactness.3 concavity.3 concave.points.3 symmetry.3 fractal.dimension.3
1 1361.7879    0.1378444     0.3264879   0.3968244        0.1712867  0.3111485          0.08469061
2  866.0806    0.1533246     0.3992064   0.4450233        0.1645104  0.3412611          0.10101014
3 2951.4545    0.1382455     0.3480818   0.4666273        0.2267455  0.2956091          0.08282364
4 1917.5833    0.1399794     0.3745333   0.4746812        0.2007725  0.3210000          0.08603062
  Tumor.size Lymph.node.status
1   2.943939          2.939394
2   2.318056          2.986111
3   2.772727          1.909091
4   3.481250          4.000000
```

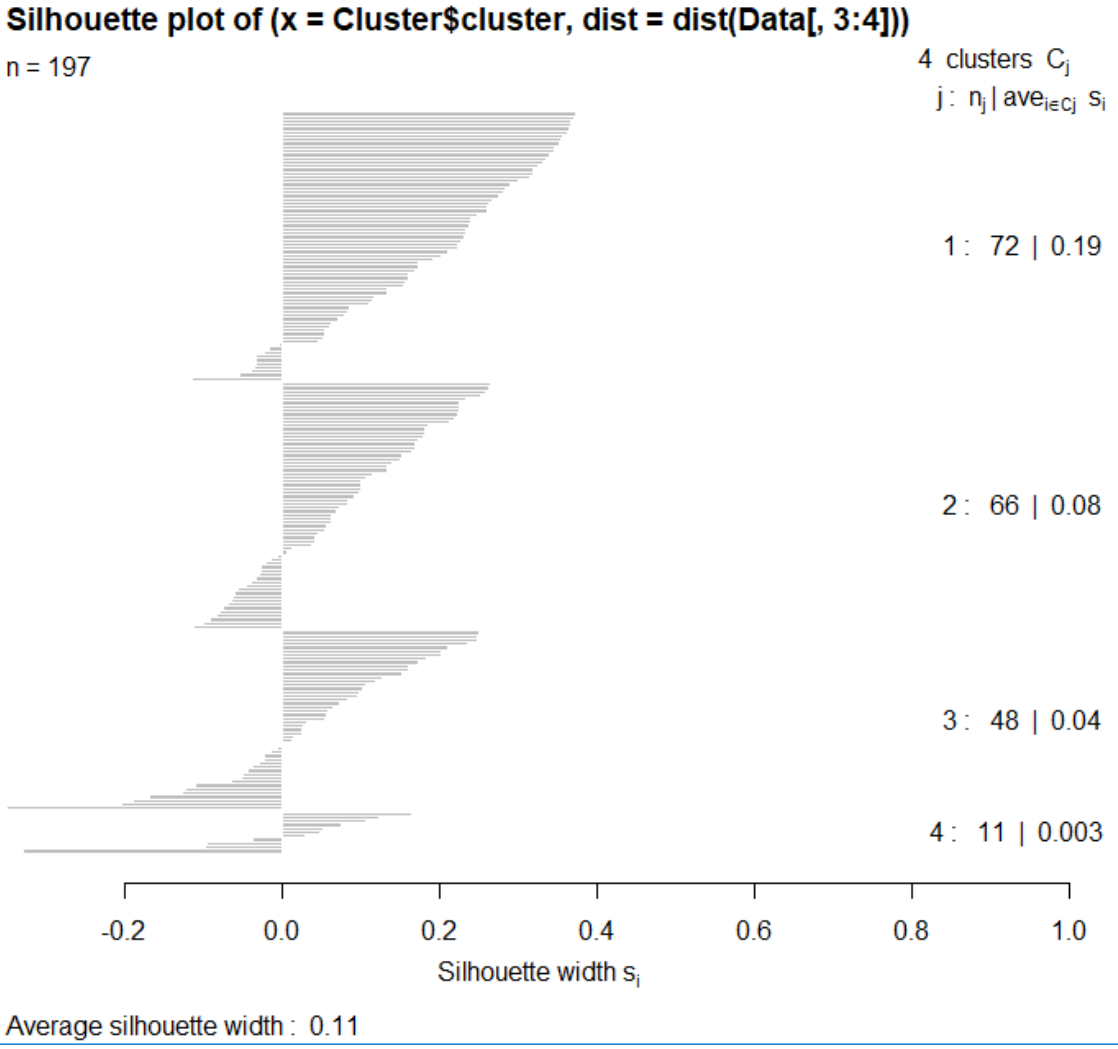
```
> Cluster$totss
[1] 92809449
> Cluster$withinss
[1] 2795136 3878790 2766961 2494725
```

b. Select the best of the above three clustering's and explain how you chose the best candidate.

In the above three clustering's, only total SSE values of the clusters are the same. But the centers of the clusters are different and the individual SSE errors are also different. Since, we are getting the same SSE value for all the three run's we can take any one of the run.

c. For the best candidate clustering chosen by you plot the Silhouette coefficient for the clustering. Compute and report the average Silhouette coefficient for each cluster of the chosen clustering.

The below Screen shot represents the plot for the Silhouette coefficient and for the first cluster the average Silhouette coefficient is 0.19, for the second cluster the Silhouette coefficient is 0.08, for the third cluster the Silhouette coefficient is 0.04, for the fourth cluster the Silhouette coefficient is 0.003, whereas, the average Silhouette width is 0.11



d. Now consider the class label for each data point in each cluster ("R" or "N"). To each cluster assign the label that belongs to most of the data points in that cluster. Report the cluster center, its SSE, and its class label, and the fraction of points that have the class label.

From the below screenshot we can infer that the Cluster 1 has N label since, it has 53 observations while, 19 in R. This is because, the cluster is assigned a label which has the highest majority.
Similarly, we can say that the Cluster 2 has label N since, it has 50 observations and only 16 in R.
Cluster 3 and Cluster 4 are also labelled N because they have the highest majority.

```
> table(cluster$cluster, Data$Outcome)

   N  R
1 53 19
2 50 16
3 38 10
4  9  2
```

From the below figure we can say that the Clusters are of size 72,66,48,11.
We now calculate the fraction of points in each cluster.

```
K-means clustering with 4 clusters of sizes 72, 66, 48, 11
```

Cluster1:
N-53/72=0.7336
R-19/72=0.2638

Cluster2:
N-50/66=0.7575
R-16/66=0.2424

Cluster3:
N-38/48=0.7916
R-10/48=0.2083

Cluster4:
N-9/11=0.8181
R-2/11=0.1818

The below screen shot represents the SSE value of each clusters:

```
> cluster$withinss
[1] 2766961 2494725 2795136 3878790
```

The below screenshot represents the Cluster Centers of each attribute:

```
Cluster means:
  radius.1 texture.1 perimeter.1    area.1 smoothness.1 compactness.1 concavity.1 concave.points.1
1 14.17236  21.67083    93.27083  625.3167    0.1058803     0.1391426   0.1312992       0.06651750
2 20.45229  22.46646   135.66042 1301.1583    0.1017167     0.1584346   0.1959981       0.11020292
3 17.70273  22.32470   116.35758  975.3288    0.1001609     0.1329308   0.1443685       0.08413061
4 23.55636  24.63000   156.11818 1745.9091    0.1017845     0.1585545   0.2216091       0.13449091
  symmetry.1 fractal.dimension.1  radius.2 texture.2 perimeter.2    area.2 smoothness.2 compactness.2
1  0.1949569          0.06651208 0.3985431  1.221642    2.878444  35.25028  0.006923750     0.03153642
2  0.1955979          0.06075146 0.7807312  1.222690    5.531646 101.52437  0.006686396     0.03278667
3  0.1897742          0.06052636 0.6096697  1.330680    4.251424  67.96833  0.006890712     0.03074527
4  0.1843727          0.05933455 1.1299364  1.272627    7.748636 176.05818  0.005470182     0.02633455
  concavity.2 concave.points.2 symmetry.2 fractal.dimension.2 radius.3 texture.3 perimeter.3
1  0.03867125       0.01383031 0.02087860         0.004129361 16.77750  30.28903    112.3356
2  0.04513104       0.01617663 0.02084437         0.003894438 25.00604  29.38542    167.8667
3  0.04057864       0.01580148 0.02049717         0.004011561 21.09258  29.94182    139.8258
4  0.03702909       0.01496727 0.01784727         0.003354727 30.93727  32.99909    206.8091
     area.3 smoothness.3 compactness.3 concavity.3 concave.points.3 symmetry.3 fractal.dimension.3
1  866.0806    0.1533246     0.3992064   0.4450233        0.1645104  0.3412611          0.10101014
2 1917.5833    0.1399794     0.3745333   0.4746812        0.2007725  0.3210000          0.08603062
3 1361.7879    0.1378444     0.3264879   0.3968244        0.1712867  0.3111485          0.08469061
4 2951.4545    0.1382455     0.3480818   0.4666273        0.2267455  0.2956091          0.08282364
  Tumor.size Lymph.node.status
1   2.318056          2.986111
2   3.481250          4.000000
3   2.943939          2.939394
4   2.772727          1.909091
```

e. Now, use the cluster centers and the class labels as a new classifier. Consider each data point again as belonging to your test set. For each data point predict its class label to be the one that belongs to the cluster center that is closest to the data point. Build the confusion matrix for this new classifier and compute its accuracy, precision and recall values.

```
> table(cluster$cluster, Data$Outcome)

    N  R
1  53 19
2  50 16
3  38 10
4   9  2
```

	N	R
N	150	47
R	0	0

The Accuracy: 0.7614
The Precision(N):0.7614
The Precision(R):0
The Recall(N):1
The Recall(R):0

f. Compare these performance results with those obtained by you in HW3 Q1. Comment on the possible causes for the differences between the two sets of performance values.

The below Screen shot represents the value for Q1 in HW3.

The accuracy of the tree is found to be 75.5%

For the Non-Recurrence class:
Precision:87%
Reacll:83%

For the Recurrence class:
Precision:40.5%
Recall:50.33%

The below details are obtained in HW 3B:

The Accuracy: 0.7614
The Precision(N):0.7614
The Precision(R):0
The Recall(N):1
The Recall(R):0

Upon comparing the three values, we find that the accuracy for the K means is a bit more than the one we got in HW3. This is because, the clustering is done a bit properly. The Recall value is 100% in clustering whereas 83% in HW3. This is because the model is not identifying the Recall values for the P class. Same goes with the precision values.

2. **Mix the datasets for the red and white wines in one dataset. Perform k-means clustering on this large dataset for the values of k to be: 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14. For each value of k report the lowest total SSE value after selecting the best of the 3-runs for each value of k. Plot the SSE value vs. the value of k. What can you infer from this plot?**

	k	SSE
1	3	4331572
2	4	3038740
3	5	2394008
4	6	2051827
5	7	1796423
6	8	1646336
7	9	1497247
8	10	1378249
9	11	1281993
10	12	1212691
11	13	1126838
12	14	1080529

The above fig. represents the k values against the least values of SSE. For a better viewing purposes, I have inserted the values in the data frame.

From the below plot, we can infer that the as the k value increases the SSE value Decreases. We ca say that this is because, there are 10 classes in the Wine Quality Data set. If we take only k=3, we say that the remaining 7 are not properly classified. Hence, a high SSE value. For example, if we take the k value as 10 the SSE value decreases since, the data set contains 10 classes, they would be properly clustered. Hence, a least SSE value compared to K=3. As the value of k increases the SSE value decreases. Therefore, we get the least SSE value for k=14.