Kmeans

Section a)

K-means cluster analysis is a tool designed to assign cases to a fixed number of groups (clusters) whose characteristics are not yet known but are based on a set of specified variables. It is most useful when you want to classify a large number (thousands) of cases.
A good cluster analysis is:

- Efficient - Uses as few clusters as possible.
- Effective - Captures all statistically and commercially important clusters. For example, a cluster with five customers may be statistically different but not very profitable.

The algorithm is called k-means, where k is the number of clusters you want, since a case is assigned to the cluster for which its distance to the cluster mean is the smallest. The action in the algorithm centers around finding the k-means.

1. First identify k clusters, it can be random and classify cases based on their distances to the centers.
2. Next, you compute the cluster means again, using the cases that are assigned to the cluster; then, you reclassify all cases based on the new set of means.
3. You keep repeating this step until cluster means don't change much between successive steps.
4. Finally, you calculate the means of the clusters once again and assign the cases to their permanent clusters.

In SPSS, the default algorithm for choosing initial cluster centers is not invariant to case ordering. SPSS finds k cases that are well-separated and use these values as initial cluster centers.

Section b)

Some methods determine the similarity between two objects by the distance between them. Such a distance can be defined on Euclidean space, a road network, a vector space, or any other space. In other methods, the similarity may be defined by connectivity based on density or contiguity, and may not rely on the absolute distance between two objects. Similarity measures play a fundamental role in the design of clustering methods. While distance-based methods can often take advantage of optimization techniques, density- and continuity-based methods can often find clusters of arbitrary shape.
A metric function or distance function is a function which defines a distance between elements/objects of a set. A set with a metric is known as metric space. This distance metric plays a very important role in clustering techniques. The numerous methods are available for clustering.

Minkowski

The Minkowski family includes Euclidean distance and Manhattan distance, which are particular cases of the Minkowski distance. The Minkowski distance is defined by

$$d_{min} = (\sum_{i=1}^{n} |x_i - y_i|^{m})^{\frac{1}{m}},\ m \geq 1,$$

where m is a positive real number and x_i and y_i are two vectors in n-dimensional space. The Minkowski distance performs well when the dataset clusters are isolated or compacted; if the dataset does not fulfil this condition, then the large-scale attributes would dominate the others.

Another problem with Minkowski metrics is that the largest-scale feature dominates the rest. Thus, normalizing the continuous features is the solution to this problem.

A modified version of the Minkowski metric has been proposed to solve clustering obstacles.

Manhattan distance

Manhattan distance is a special case of the Minkowski distance at m = 1. Like its parent, Manhattan is sensitive to outliers. When this distance measure is used in clustering algorithms, the shape of clusters is hyper-rectangular. A study by Perlibakas demonstrated that a modified version of this distance measure is among the best distance measures for PCA-based face recognition. This measure is defined as

$$d_{man} = \sum_{i=1}^{n} |x_i - y_i|.$$

Euclidean distance

The most well-known distance used for numerical data is probably the Euclidean distance. This is a special case of the Minkowski distance when m = 2. Euclidean distance performs well when deployed to datasets that include compact or isolated clusters.
Although Euclidean distance is very common in clustering, it has a drawback: if two data vectors have no attribute values in common, they may have a smaller distance than the other pair of data vectors containing the same attribute values.
Another problem with Euclidean distance as a family of the Minkowski metric is that the largest-scaled feature would dominate the others. Normalization of continuous features is a solution to this problem

Mahalanobis distance

Mahalanobis distance is a data-driven measure in contrast to Euclidean and Manhattan distances that are independent of the related dataset to which two data points belong. A regularized Mahalanobis distance can be used for extracting hyperellipsoidal clusters. On the other hand, Mahalanobis distance can alleviated distortion caused by linear correlation among features by applying a whitening transformation to the data or by using the squared Mahalanobis distance.

Mahalanobis distance is defined by $d_{mah} = \sqrt{(x-y)S^{-1}(x-y)^T}$ where *S* is the covariance matrix of the dataset.

In SPSS, distances are computed using simple Euclidean distance.

<u>Section c)</u>

When k =3

<u>Initial Clusters</u>

Initial Cluster Centers

	Cluster		
	1	2	3
area	20.16	13.20	11.23
perimeter	17.03	13.66	12.63
compactness	.8735	.8883	.8840
length	6.513	5.236	4.902
width	3.773	3.232	2.879
asymmetry	1.9100	8.3150	2.2690
length_kernel	6.185	5.056	4.703

Following are the initial values for the cluster centers.

Number of iterations

Iteration History[a]

	Change in Cluster Centers		
Iteration	1	2	3
1	2.628	2.307	2.384
2	.038	.941	.458
3	.116	.567	.607
4	.206	.234	.422
5	.082	.157	.239
6	.040	.049	.079
7	.086	.052	.113
8	.043	.035	.072
9	.044	.052	.084
10	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 10. The minimum distance between initial centers is 6.470.

There were 3 clusters had a total of 9 iterations.

Cluster Membership

kmeans_cluster_3.xlsx

Following is the cluster membership for each record

Final Cluster Centers

Final Cluster Centers

	Cluster		
	1	2	3
area	18.72	11.96	14.65
perimeter	16.30	13.27	14.46
compactness	.8851	.8522	.8792
length	6.209	5.229	5.564
width	3.723	2.873	3.278
asymmetry	3.6036	4.7597	2.6489
length_kernel	6.066	5.089	5.192

Following are the final values for cluster centers.

Distance between each cluster

Distances between Final Cluster Centers

Cluster	1	2	3
1		7.666	4.718
2	7.666		3.654
3	4.718	3.654	

Following is the distance between each cluster

ANOVA values

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
area	779.257	2	1.019	207	764.651	.000
perimeter	156.012	2	.215	207	726.839	.000
compactness	.022	2	.000	207	62.924	.000
length	16.562	2	.038	207	433.785	.000
width	12.309	2	.025	207	489.971	.000
asymmetry	83.292	2	1.478	207	56.363	.000
length_kernel	18.814	2	.062	207	302.894	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of records in each cluster

Number of Cases in each Cluster

Cluster	1	61.000
	2	77.000
	3	72.000
Valid		210.000
Missing		.000

Cluster 1 has 61 records while cluster 2 and cluster 3 have 72 and 77 records respectively.

Number of classes in each cluster

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
class * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%

Classification matrix

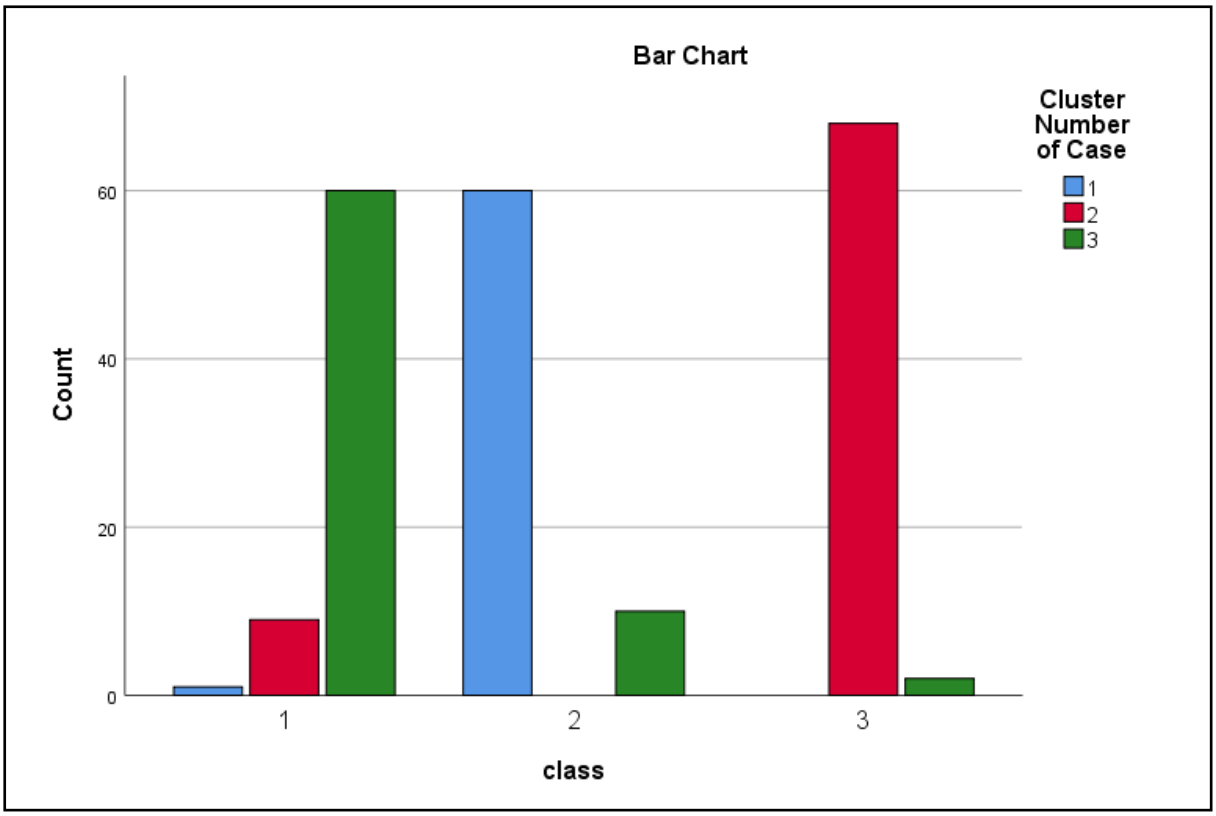
class * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case			
		1	2	3	Total
class	1	1	9	60	70
	2	60	0	10	70
	3	0	68	2	70
Total		61	77	72	210

Cluster 1 basically classifies class 2 with 60 records but it has a misclassification of 1 record from class 1
Cluster 2 separates record from class 3 but has a misclassification of 9 records from class 1.
Cluster 3 is defined for class 1 but a misclassification of 10 records from class 2 and 2 records from class 3.

Bar Chart based on number of clusters per class



Class 1 has records divided into all the clusters.
Class 2 has records split between cluster 1 and cluster 3.
Class 3 has records from cluster 2 and cluster 3.

When k =4

Initial Clusters

Initial Cluster Centers

	Cluster			
	1	2	3	4
area	12.70	13.16	18.88	21.18
perimeter	13.41	13.82	16.26	17.21
compactness	.8874	.8662	.8969	.8989
length	5.183	5.454	6.084	6.573
width	3.091	2.975	3.764	4.033
asymmetry	8.4560	.8551	1.6490	5.7800
length_kernel	5.000	5.056	6.109	6.231

Following are the initial values for the cluster centers.

Number of iterations

Iteration History[a]

	Change in Cluster Centers			
Iteration	1	2	3	4
1	2.848	2.135	1.725	1.584
2	.475	.467	.140	.347
3	.239	.313	.079	.278
4	.120	.222	.053	.177
5	.049	.083	.098	.225
6	.052	.058	.093	.132
7	.035	.040	.036	.051
8	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 8. The minimum distance between initial centers is 4.856.

There were 3 clusters had a total of 9 iterations.

Cluster Membership

kmeans_cluster_4.xlsx

Following is the cluster membership for each record

Final Cluster Centers

Final Cluster Centers

	Cluster			
	1	2	3	4
area	11.94	14.42	17.75	19.52
perimeter	13.27	14.35	15.88	16.65
compactness	.8515	.8795	.8840	.8844
length	5.229	5.524	6.048	6.350
width	2.867	3.253	3.614	3.812
asymmetry	4.8040	2.5904	3.1649	4.1641
length_kernel	5.095	5.127	5.921	6.184

Following are the final values for cluster centers.

Distance between each cluster

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		3.530	6.717	8.520
2	3.530		3.842	5.988
3	6.717	3.842		2.220
4	8.520	5.988	2.220	

Following is the distance between each cluster

ANOVA values

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
area	530.826	3	.859	206	617.832	.000
perimeter	106.703	3	.176	206	604.753	.000
compactness	.015	3	.000	206	43.551	.000
length	11.432	3	.033	206	349.915	.000
width	8.373	3	.023	206	367.140	.000
asymmetry	63.797	3	1.365	206	46.755	.000
length_kernel	13.338	3	.051	206	262.423	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of records in each cluster

Number of Cases in each Cluster

Cluster	1	75.000
	2	67.000
	3	40.000
	4	28.000
Valid		210.000
Missing		.000

Cluster 1 has 75 records while cluster 2 and cluster 3 has 67 and 40 records respectively.
Cluster 4 has 28 records.

Number of classes in each cluster

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
class * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%

Classification matrix

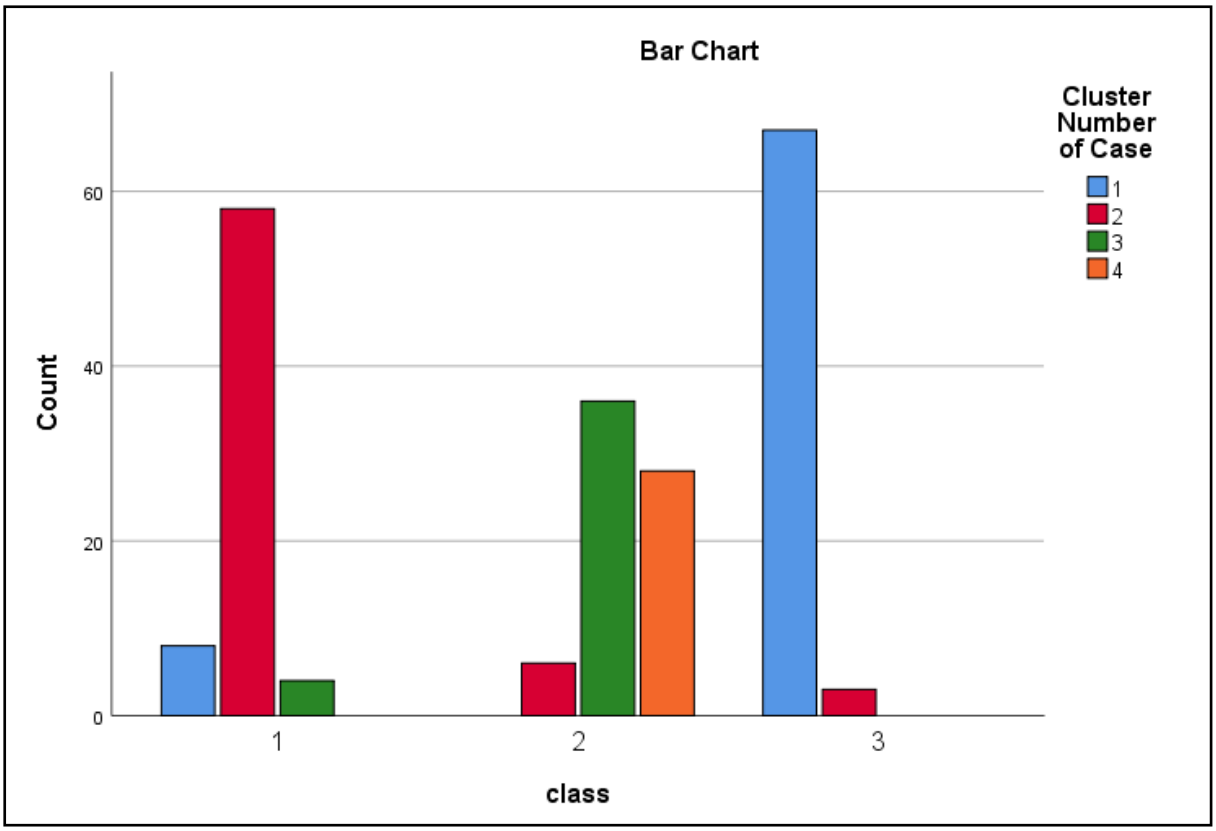
class * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case				
		1	2	3	4	Total
class	1	8	58	4	0	70
	2	0	6	36	28	70
	3	67	3	0	0	70
Total		75	67	40	28	210

Cluster 1 basically classifies class 2 with 67 records but it has a misclassification of 8 records from class 1
Cluster 2 identifies records from class 1 with 58 records but has a misclassification of 6 records from class 2 and 3 records class 3.
Cluster 3 is defined for class 2 with 36 records but a misclassification of 4 records from class 1.
Cluster 4 is independent with 28 records for class 2.

Bar Chart based on number of clusters per class



Class 1 has records divided into all the clusters 1, 2 and 3.
Class 2 has records split between cluster 2, 3 and 4.
Class 3 has records classified into cluster 1 and cluster 2.

When k = 5

Initial Clusters

Initial Cluster Centers

	Cluster				
	1	2	3	4	5
area	16.77	14.88	20.16	11.23	13.20
perimeter	15.62	14.57	17.03	12.88	13.66
compactness	.8638	.8811	.8735	.8511	.8883
length	5.927	5.554	6.513	5.140	5.236
width	3.438	3.333	3.773	2.795	3.232
asymmetry	4.9200	1.0180	1.9100	4.3250	8.3150
length_kernel	5.795	4.956	6.185	5.003	5.056

Following are the initial values for the cluster centers.

Number of iterations

Iteration History[a]

Iteration	Change in Cluster Centers 1	2	3	4	5
1	.589	1.219	1.578	.840	1.328
2	.089	.209	.214	.087	.427
3	.292	.031	.156	.140	.390
4	.111	.038	.057	.103	.163
5	.000	.000	.000	.138	.270
6	.000	.000	.000	.094	.143
7	.000	.000	.000	.126	.152
8	.000	.034	.000	.145	.108
9	.000	.066	.000	.162	.084
10	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 10. The minimum distance between initial centers is 4.540.

There were 3 clusters had a total of 9 iterations.

Cluster Membership

kmeans_cluster_5.xlsx

Following is the cluster membership for each record

Final Cluster Centers

Final Cluster Centers

	Cluster 1	2	3	4	5
area	16.56	14.69	19.15	12.09	11.98
perimeter	15.39	14.47	16.47	13.31	13.29
compactness	.8782	.8809	.8871	.8571	.8508
length	5.888	5.572	6.269	5.217	5.241
width	3.481	3.286	3.773	2.901	2.880
asymmetry	4.1095	2.4079	3.4604	3.3438	5.6733
length_kernel	5.725	5.159	6.127	5.005	5.122

Following are the final values for cluster centers.

Distance between each cluster

Distances between Final Cluster Centers

Cluster	1	2	3	4	5
1		2.773	2.946	5.121	5.380
2	2.773		5.160	3.050	4.435
3	2.946	5.160		7.936	8.319
4	5.121	3.050	7.936		2.335
5	5.380	4.435	8.319	2.335	

Following is the distance between each cluster

ANOVA values

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
area	410.587	4	.620	205	662.132	.000
perimeter	82.189	4	.135	205	608.304	.000
compactness	.011	4	.000	205	31.525	.000
length	8.815	4	.028	205	313.313	.000
width	6.409	4	.020	205	314.348	.000
asymmetry	65.304	4	1.031	205	63.366	.000
length_kernel	10.271	4	.046	205	223.982	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of records in each cluster

Number of Cases in each Cluster

Cluster	1	25.000
	2	51.000
	3	48.000
	4	44.000
	5	42.000
Valid		210.000
Missing		.000

Cluster 1 has 25 records while cluster 2 and cluster 3 has 51 and 48 records respectively.
Cluster 4 has 44 records and cluster 5 has 42 records

Number of classes in each cluster

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
class * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%

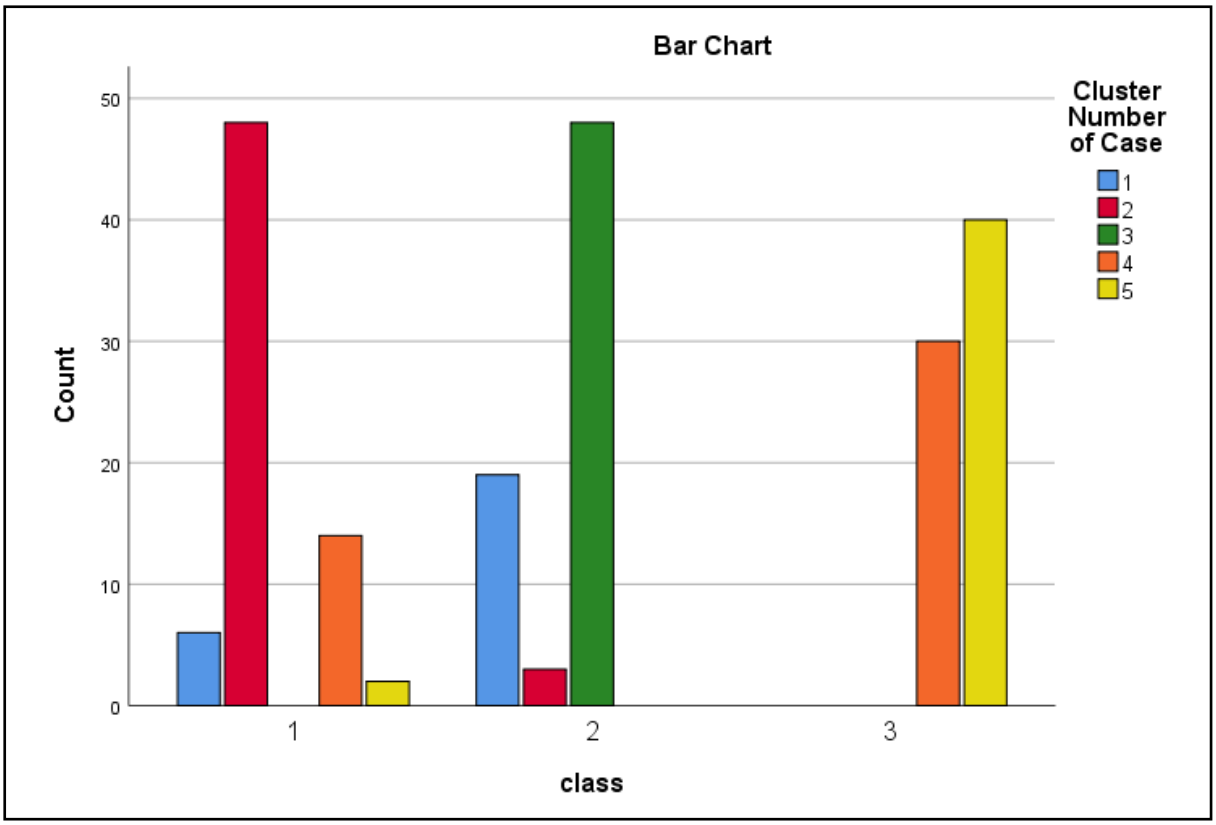
Classification matrix

class * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case					
		1	2	3	4	5	Total
class	1	6	48	0	14	2	70
	2	19	3	48	0	0	70
	3	0	0	0	30	40	70
Total		25	51	48	44	42	210

Cluster 1 basically classifies class 2 with 19 records but it has a misclassification of 6 records from class 1
Cluster 2 identifies records from class 1 with 48 records but has a misclassification of 3 records from class 2.
Cluster 3 entirely identifies 48 records for class 2.
Cluster 4 has 30 records from class 3 and 14 records from class 1.
Cluster 5 has 40 records from class 4 and 2 records from class 1.



Class 1 has records divided into all the clusters 1, 2, 4 and 5.
Class 2 has records split between cluster 1, 2 and 3.
Class 3 has records classified into cluster 4 and cluster 5.

When k = 6

Initial Clusters

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
area	11.23	14.34	17.55	20.16	13.20	21.18
perimeter	12.88	14.37	15.66	17.03	13.66	17.21
compactness	.8511	.8726	.8991	.8735	.8883	.8989
length	5.140	5.630	5.791	6.513	5.236	6.573
width	2.795	3.190	3.690	3.773	3.232	4.033
asymmetry	4.3250	1.3130	5.3660	1.9100	8.3150	5.7800
length_kernel	5.003	5.150	5.661	6.185	5.056	6.231

Following are the initial values for the cluster centers.

Number of iterations

Iteration History[a]

Iteration	Change in Cluster Centers					
	1	2	3	4	5	6
1	.746	1.083	.977	1.575	1.328	.804
2	.078	.075	.454	.095	.318	.551
3	.114	.033	.434	.076	.363	.448
4	.070	.089	.159	.090	.163	.207
5	.026	.100	.159	.044	.063	.095
6	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 6. The minimum distance between initial centers is 4.015.

There were 3 clusters had a total of 9 iterations.

Cluster Membership

kmeans_cluster_6.xlsx

Following is the cluster membership for each record

Final Cluster Centers

Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
area	11.83	14.24	16.41	18.95	12.32	19.58
perimeter	13.22	14.26	15.32	16.39	13.42	16.65
compactness	.8500	.8793	.8783	.8868	.8580	.8877
length	5.216	5.494	5.864	6.247	5.266	6.316
width	2.844	3.234	3.463	3.745	2.951	3.835
asymmetry	4.1684	2.3165	3.8501	2.7235	6.3367	5.0815
length_kernel	5.076	5.062	5.690	6.119	5.122	6.144

Following are the final values for cluster centers.

Distance between each cluster

Distances between Final Cluster Centers

Cluster	1	2	3	4	5	6
1		3.244	5.161	8.114	2.235	8.718
2	3.244		2.963	5.377	4.547	6.644
3	5.161	2.963		3.051	5.239	3.729
4	8.114	5.377	3.051		8.275	2.457
5	2.235	4.547	5.239	8.275		8.226
6	8.718	6.644	3.729	2.457	8.226	

Following is the distance between each cluster

ANOVA values

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
area	326.467	5	.672	204	485.654	.000
perimeter	65.254	5	.148	204	440.963	.000
compactness	.009	5	.000	204	26.554	.000
length	6.949	5	.031	204	225.593	.000
width	5.149	5	.020	204	257.918	.000
asymmetry	64.487	5	.736	204	87.675	.000
length_kernel	8.331	5	.043	204	192.458	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of records in each cluster

Number of Cases in each Cluster

Cluster	1	56.000
	2	54.000
	3	31.000
	4	33.000
	5	21.000
	6	15.000
Valid		210.000
Missing		.000

Cluster 1 has 56 records while cluster 2 and cluster 3 have 54 and 31 records respectively.
Cluster 4 has 33 records and cluster 5 and cluster 6 have 21 and 15 records respectively.

Number of classes in each cluster

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
class * Cluster Number of Case	210	100.0%	0	0.0%	210	100.0%

Classification matrix

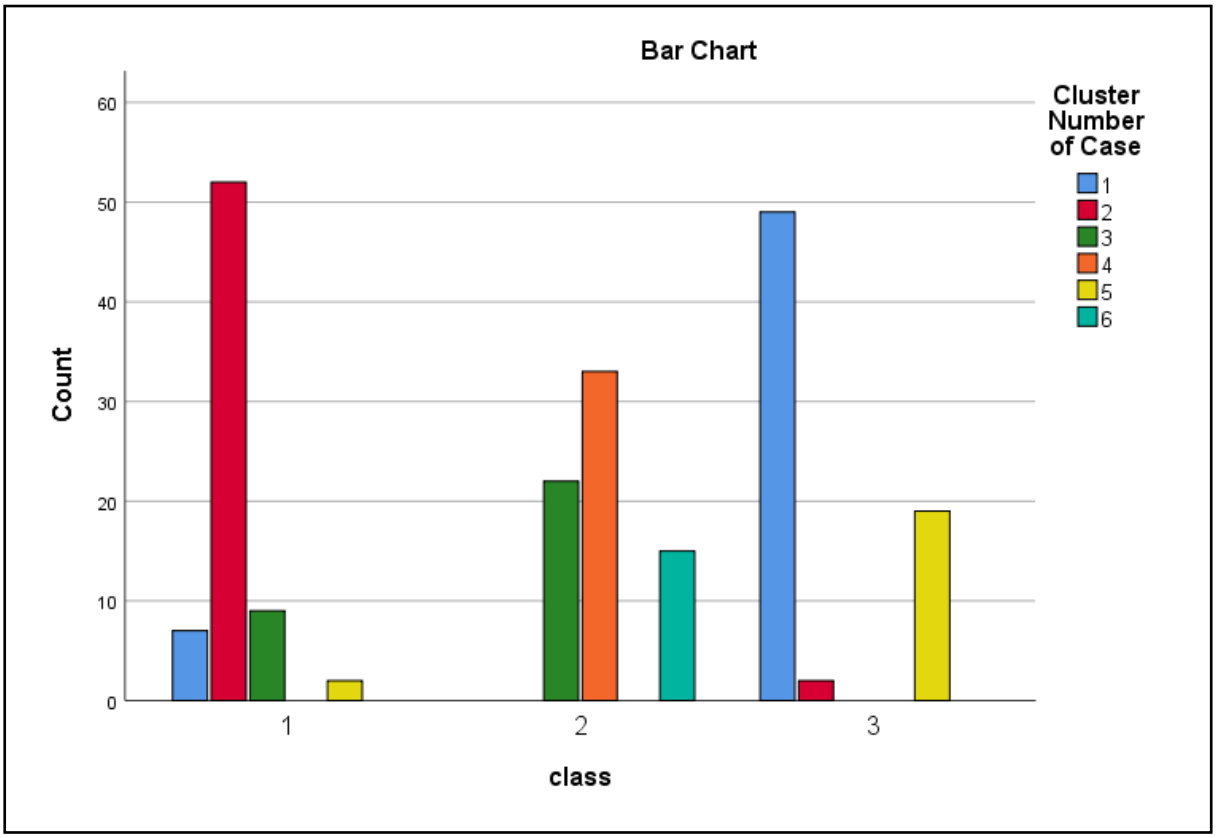
class * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case						
		1	2	3	4	5	6	Total
class	1	7	52	9	0	2	0	70
	2	0	0	22	33	0	15	70
	3	49	2	0	0	19	0	70
Total		56	54	31	33	21	15	210

Cluster 1 basically classifies class 3 with 49 records but it has a misclassification of 7 records from class 1
Cluster 2 identifies records from class 1 with 52 records but has a misclassification of 2 records from class 3.
Cluster 3 has 22 records from class 2 and 9 records from class 1.
Cluster 4 entirely identifies 33 records for class 2.

Cluster 5 has 19 records from class 3 and 2 records from class 1.
Cluster 6 entirely identifies 15 records for class 2.



Class 1 has records divided into all the clusters 1, 2, 3 and 5.
Class 2 has records split between cluster 3, 4 and 6.
Class 3 has records classified into cluster 1, cluster 2 and cluster 5.

Section c subsection iv)

Considering the dataset is divided into 3 different classes, I believe we must use the k = 3.
When k = 3, we have fewer misclassifications among other classes. The accuracy rate is the best amongst all the models.
Each cluster is easily dominated by one class helping us to easily identify the classes.
As we increase the value of k, the number of classes associated with each cluster gets cluttered making it difficult to classify.

Section c subsection v)

According to me, the normalisation of area, perimeter and compactness attributes will affect the clustering results.
Currently the values of these attributes area, perimeter and compactness have a different range compared to other variables with this dataset.
Normalisation of area, perimeter and compactness attributes can impact the clustering pattern and might improve the accuracy rate.

Problem 1 – ii

Hierarchical clustering: Single Linkage

Proximities

Case Processing Summary[a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
210	100.0%	0	0.0%	210	100.0%

a. Rescaled Squared Euclidean Distance used


Agglomeration_schedule_single.xlsx


hierarchical_cluster_membership_single.xl

Following are the agglomerative schedule and cluster membership for single linkage method for clusters are 3 to 6.

Dendrogram


dendrogram_single.jpg

Attached is the dendrogram diagram for single linkage method.

Class Distribution

Single Linkage

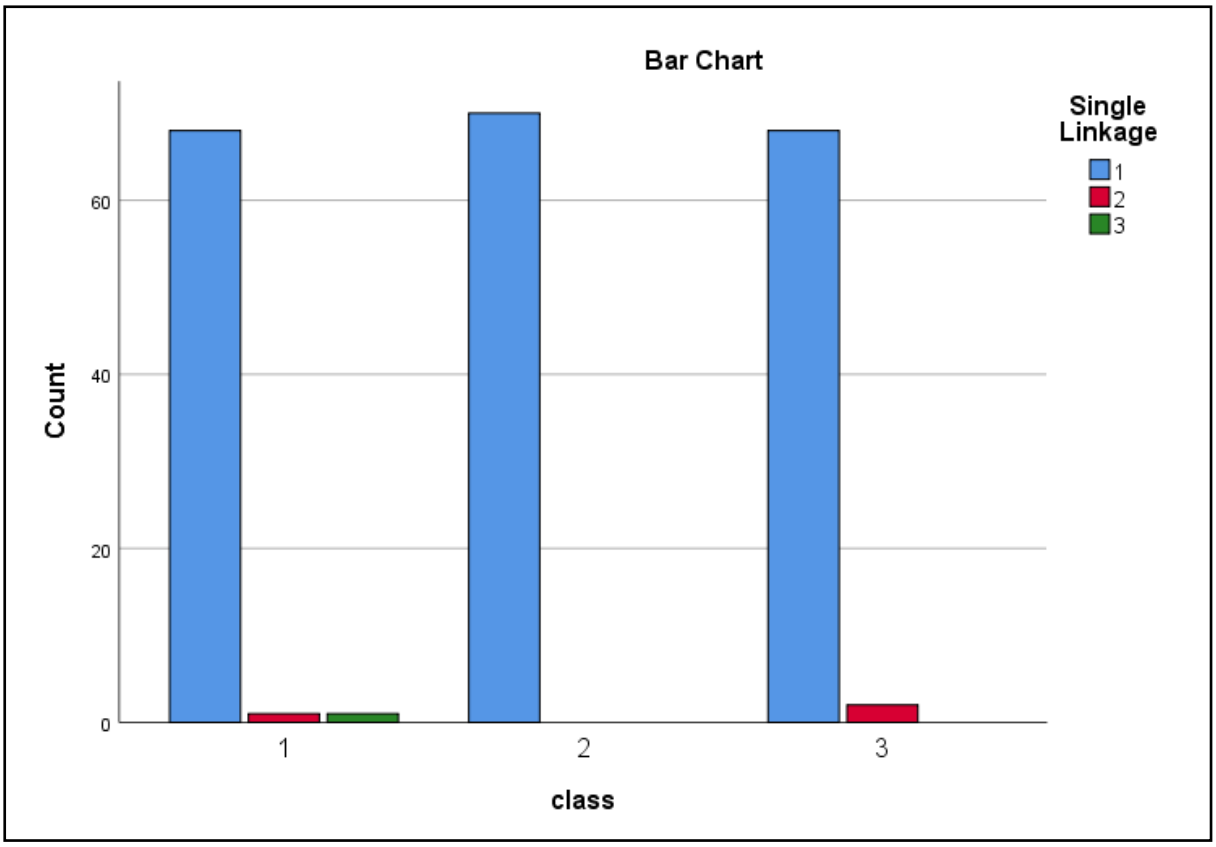
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	206	98.1	98.1	98.1
	2	3	1.4	1.4	99.5
	3	1	.5	.5	100.0
	Total	210	100.0	100.0	

The frequency table shows the 210 records divided into 3 clusters with cluster 1 having the majority records with 206 counts while cluster 2 and cluster 3 have 3 records and 1 record respectively.

class * Single Linkage Crosstabulation

Count

		Single Linkage			
		1	2	3	Total
class	1	68	1	1	70
	2	70	0	0	70
	3	68	2	0	70
Total		206	3	1	210



The bar chart shows class 1 is majorly identified by cluster 1 with a couple of records falling in cluster 2 and cluster 3.
Class 2 is allocated to cluster 1 and class 3 is divided into cluster 1 and cluster 2 with cluster 1 having the majority.

Hierarchical clustering: Complete Linkage

Proximities

Case Processing Summary[a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
210	100.0%	0	0.0%	210	100.0%

a. Rescaled Squared Euclidean Distance used


Agglomeration_schedule_complete.xlsx


hierarchical_cluster_membership_complete

Following are the agglomerative schedule and cluster membership for single linkage method for clusters are 3 to 6.

Dendrogram


dendrogram_complete.jpg

Attached is the dendrogram diagram for single linkage method.

Class Distribution

Complete Linkage

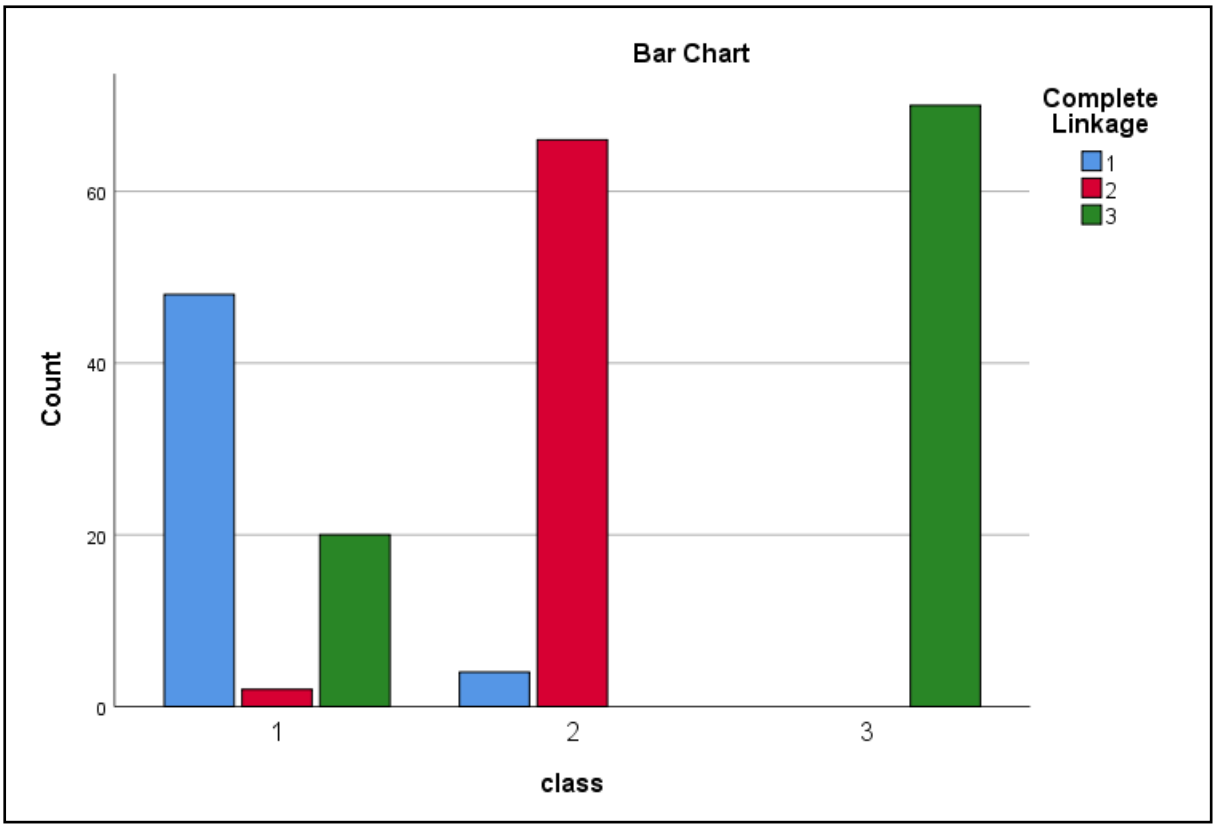
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	52	24.8	24.8	24.8
	2	68	32.4	32.4	57.1
	3	90	42.9	42.9	100.0
	Total	210	100.0	100.0	

The frequency table shows the 210 records divided into 3 clusters with cluster 1 having 52 records while cluster 2 and cluster 3 have 68 records and 90 records respectively.

class * Complete Linkage Crosstabulation

Count

		Complete Linkage			
		1	2	3	Total
class	1	48	2	20	70
	2	4	66	0	70
	3	0	0	70	70
Total		52	68	90	210

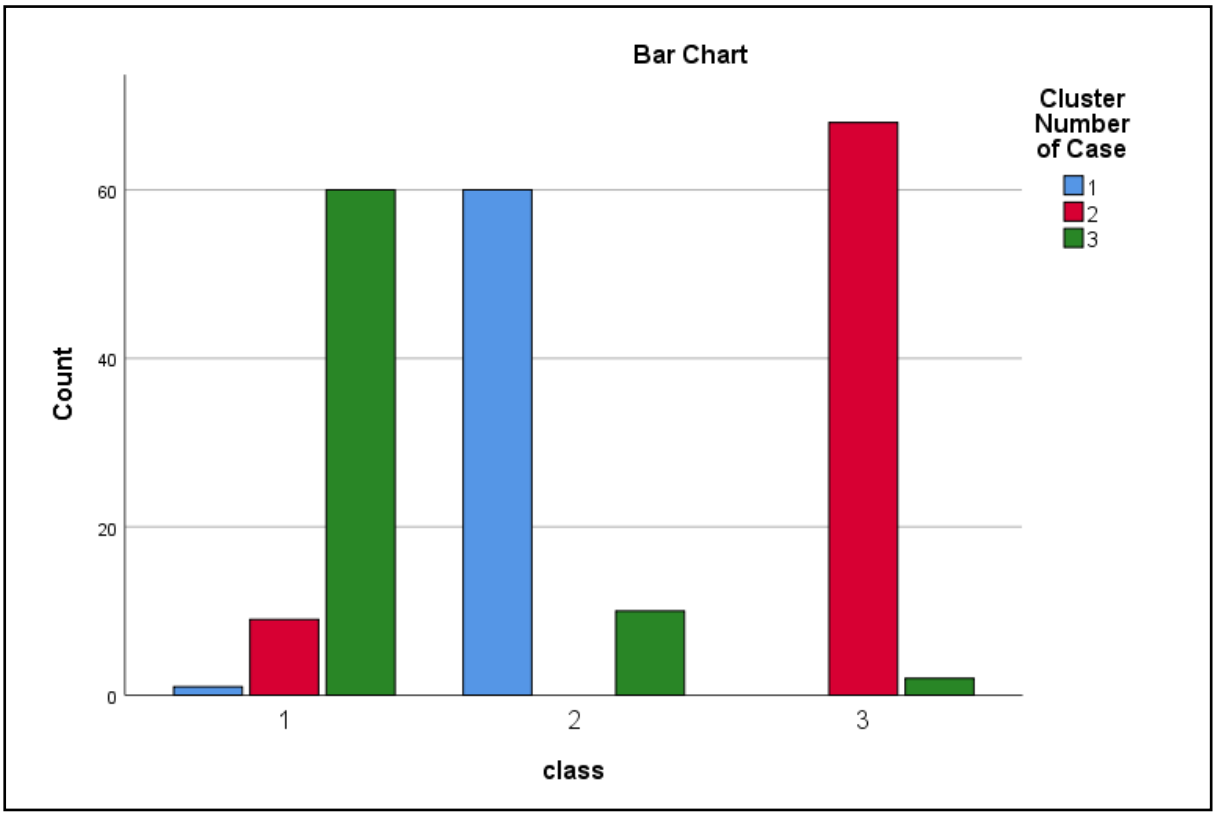


The bar chart shows class 1 is divided into all the clusters.
Class 2 is allocated to cluster 1 and cluster 2.
Class 3 is allocated to cluster 3.

Problem 1 – iii

Selecting the best model from k means is when k = 3.
We have few misclassifications among other classes. The accuracy rate is the best amongst all the models.
Each cluster is easily dominated by one class helping us to easily identify the classes.
The hierarchical clustering (complete linkage) produces the best model using the complete linkage model.

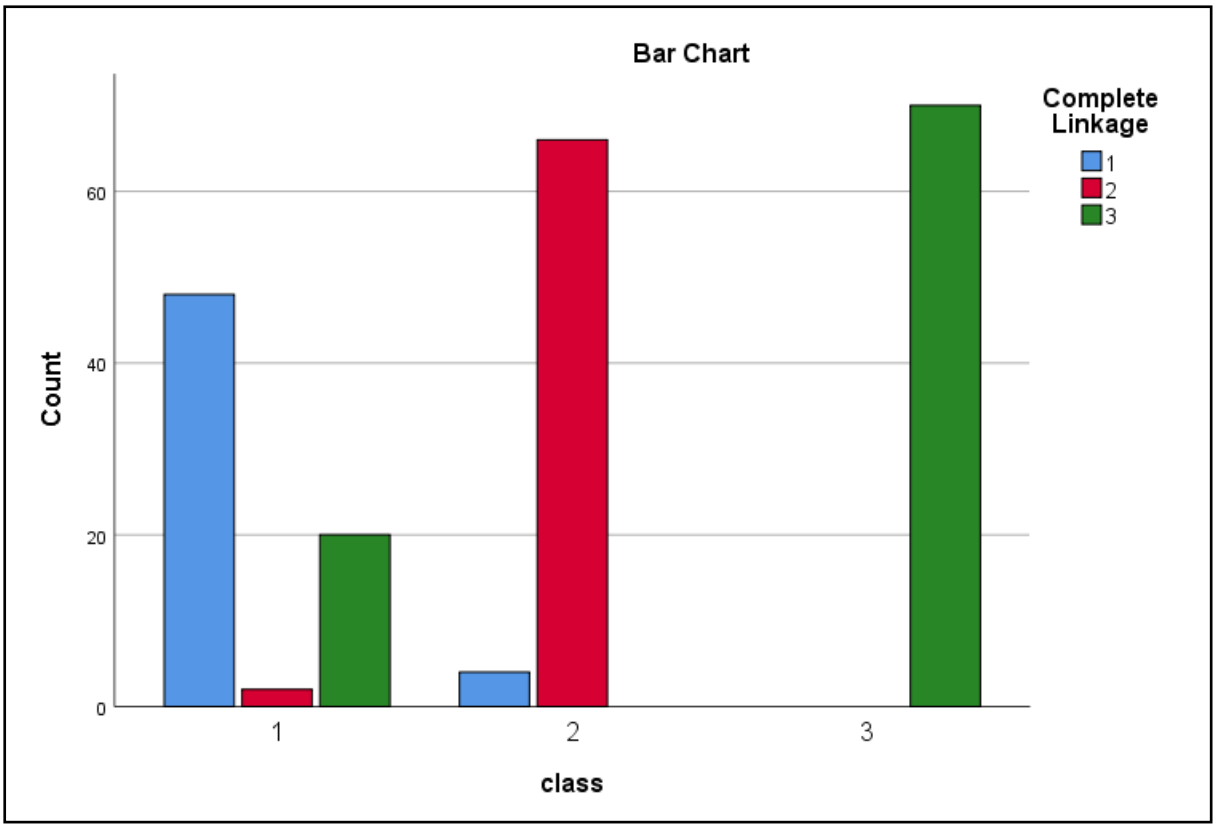
In K means produces have the following output.



The bar chart shows

- Class 1 has records divided into all the clusters.
- Class 2 has records split between cluster 1 and cluster 3.
- Class 3 has records from cluster 2 and cluster 3.

The hierarchical clustering (complete linkage) produces the following results.



The bar chart shows

- Class 1 is divided into all the clusters.
- Class 2 is allocated to cluster 1 and cluster 2.
- Class 3 is allocated to cluster 3.

The hierarchical clustering (complete linkage) produces a better a classification model.
There are fewer misclassifications within classes which can produce a better accuracy rate.

Problem 1 – iv

Clustering is basically a technique that groups similar data points such that the points in the same group are more similar to each other than the points in the other groups. The group of similar data points is called a **Cluster**. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.
Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples.

Kmeans Algorithm:
Kmeans algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points

that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.
The way kmeans algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).

Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Applications:
kmeans algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc. The goal usually when we undergo a cluster analysis is either:

1. Get a meaningful intuition of the structure of the data we're dealing with.
2. Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviours of different subgroups.
   An example of that is clustering patients into different subgroups and build a model for each subgroup to predict the probability of the risk of having heart attack.

Drawbacks:
Kmeans algorithm is good in capturing structure of the data if clusters have a spherical-like shape. It always try to construct a nice spherical shape around the centroid. That means, the minute the clusters have a complicated geometric shapes, kmeans does a poor job in clustering the data.

- Kmeans algorithm doesn't let data points that are far-away from each other share the same cluster even though the points belong to the same cluster.
- Suppose the data is generated from multivariate normal distributions with different means and standard deviations. So there are 3 groups of data where each group was generated from different multivariate normal distribution (different mean/standard deviation). One group will have a lot more data points than the other two combined.
- kmeans cannot figure out the correct clusters in complicated geometric shapes such as moons and circles within each other.

Hierarchical clustering Technique:
This clustering technique is divided into two types:

1. Agglomerative
2. Divisive

Agglomerative Hierarchical clustering Technique: In this technique, initially each data point is considered as an individual cluster.
At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

The basic algorithm of Agglomerative is straight forward.

- Compute the proximity matrix

- Let each data point be a cluster
- Repeat: Merge the two closest clusters and update the proximity matrix
- Until only a single cluster remains

Key operation is the computation of the proximity of two clusters.

The Hierarchical clustering Technique can be visualized using a **Dendrogram.**

A **Dendrogram** is a tree-like diagram that records the sequences of merges or splits.

Divisive Hierarchical clustering Technique: The Divisive Hierarchical clustering is exactly the opposite of the **Agglomerative Hierarchical clustering.**
In Divisive Hierarchical clustering, we consider all the data points as a single cluster and in each iteration, and then separate the data points from the cluster which are not similar. Each data point which is separated is considered as an individual cluster. In the end, they will be left with n clusters. As the process is dividing the single clusters into n clusters, it is named as **Divisive Hierarchical clustering.**

The similarity between two clusters is important to merge or divide the clusters. The different approaches which are used to calculate the similarity between two clusters:

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Ward's Method

Drawbacks:

1. There is no mathematical objective for Hierarchical clustering.
2. All the approaches to calculate the similarity between clusters has its own disadvantages.
3. High space and time complexity for Hierarchical clustering. Hence this clustering algorithm cannot be used when we have huge data.

The seeds dataset has 210 records with 3 classes. Each class has 70 records making the dataset well balanced.
There are 7 attributes with area and perimeter having a range of 10 – 20 and the compactness having a range of 0.5 – 1. The remaining attributes like length of kernel, width of kernel, asymmetry coefficient and length of kernel groove have their range value between 4 – 7.

As part of the exercise, we have performed the K means clustering and Hierarchical Single Linkage clustering and Hierarchical Complete Linkage clustering on the seeds dataset.

In K means clustering
The bar chart shows

- Class 1 has records divided into all the clusters.
- Class 2 has records split between cluster 1 and cluster 3.
- Class 3 has records from cluster 2 and cluster 3.

In Hierarchical complete linkage clustering
The bar chart shows

- Class 1 is divided into all the clusters.

- Class 2 is allocated to cluster 1 and cluster 2.
- Class 3 is allocated to cluster 3.

The hierarchical clustering (complete linkage) produces a better a classification model.
There are fewer misclassifications within classes which can produce a better accuracy rate.

For better results, we can normalise the values of area, perimeter and compactness attributes.
We can try using different similarity measures to check the classifications amongst different algorithms.
These experiments might affect the results from the clustering algorithm.