NAME: RISHABH POONIA
ID:20298657
PART 1: CLASSIFICATION

1)

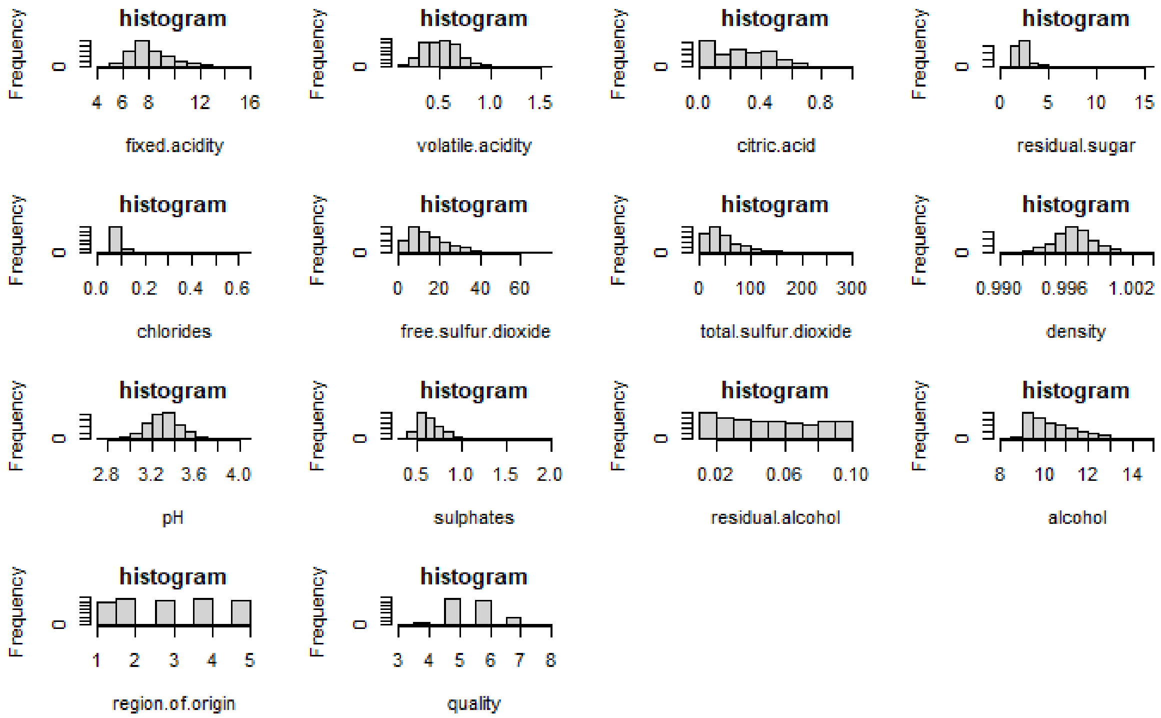
i)

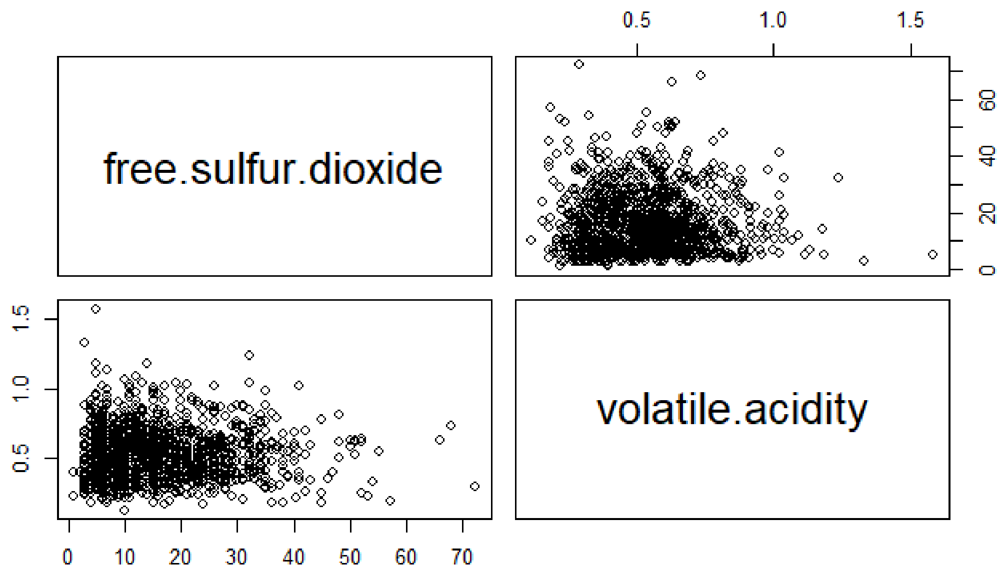
sample	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
Min. : 1.0	Min. : 4.60	Min. :0.1200	Min. :0.0000	Min. : 0.900	Min. :0.01200
1st Qu.: 400.8	1st Qu.: 7.10	1st Qu.:0.3900	1st Qu.:0.0900	1st Qu.: 1.900	1st Qu.:0.07000
Median : 800.5	Median : 7.90	Median :0.5200	Median :0.2600	Median : 2.200	Median :0.07900
Mean : 800.5	Mean : 8.32	Mean :0.5278	Mean :0.2709	Mean : 2.539	Mean :0.08738
3rd Qu.:1200.2	3rd Qu.: 9.20	3rd Qu.:0.6400	3rd Qu.:0.4200	3rd Qu.: 2.600	3rd Qu.:0.09000
Max. :1600.0	Max. :15.90	Max. :1.5800	Max. :1.0000	Max. :15.500	Max. :0.61100

free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	residual.alcohol
Min. : 1.00	Min. : 6.00	Min. :0.9901	Min. :2.740	Min. :0.3700	Min. :0.0100
1st Qu.: 7.00	1st Qu.: 22.00	1st Qu.:0.9956	1st Qu.:3.210	1st Qu.:0.5500	1st Qu.:0.0300
Median :14.00	Median : 38.00	Median :0.9968	Median :3.310	Median :0.6200	Median :0.0500
Mean :15.87	Mean : 46.49	Mean :0.9968	Mean :3.311	Mean :0.6597	Mean :0.0555
3rd Qu.:21.00	3rd Qu.: 62.00	3rd Qu.:0.9979	3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:0.0800
Max. :72.00	Max. :289.00	Max. :1.0037	Max. :4.010	Max. :2.0000	Max. :0.1000
		NA's :74		NA's :78	NA's :980

alcohol	region.of.origin	quality
Min. : 8.40	Min. :1.000	Min. :3.000
1st Qu.: 9.50	1st Qu.:2.000	1st Qu.:5.000
Median :10.20	Median :3.000	Median :6.000
Mean :10.42	Mean :3.029	Mean :5.636
3rd Qu.:11.10	3rd Qu.:4.000	3rd Qu.:6.000
Max. :14.90	Max. :5.000	Max. :8.000

ii.

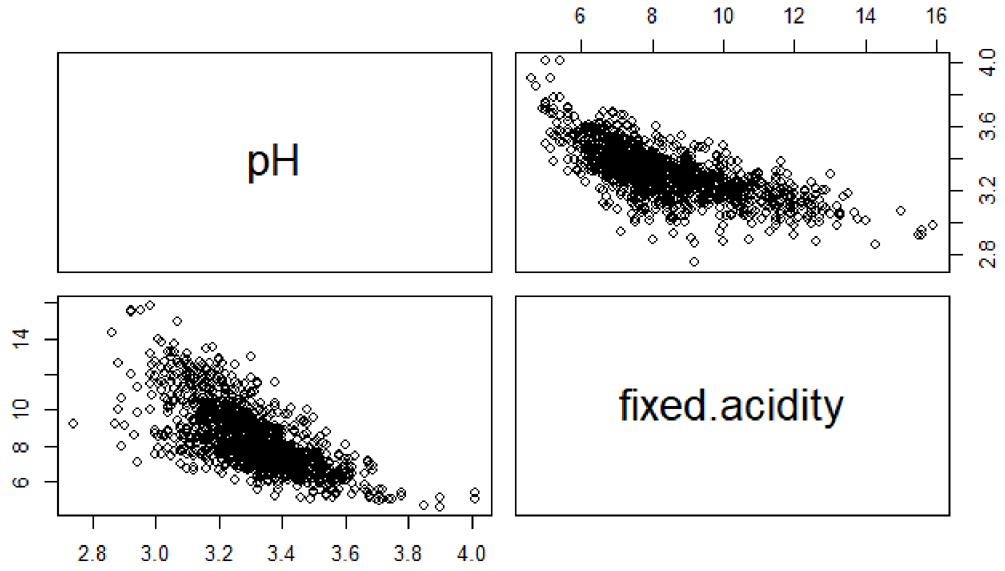




2 i) Scatterplot between free sulfur dioxide and volatile acidity

Correlation coefficient=-0.0105658

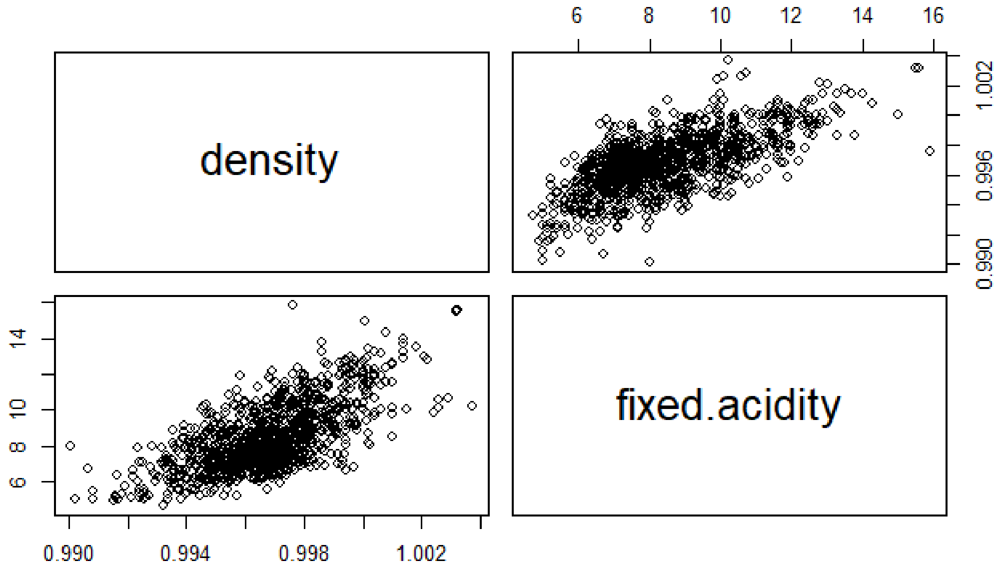
They are negatively lowly correlated which means increase in one value results in decrease in value of another.



Scatterplot between pH and fixed acidity

Correlation coefficient= -0.6829498

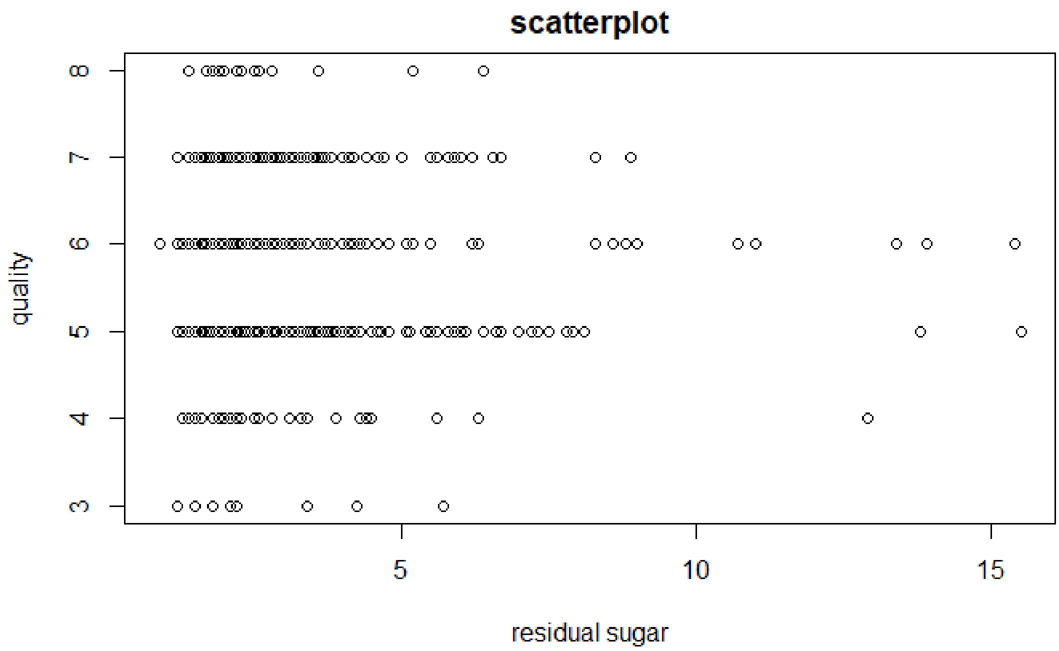
These are highly negatively correlated as it is close to value of -1 which means increase in value of one results in decrease in value of another.

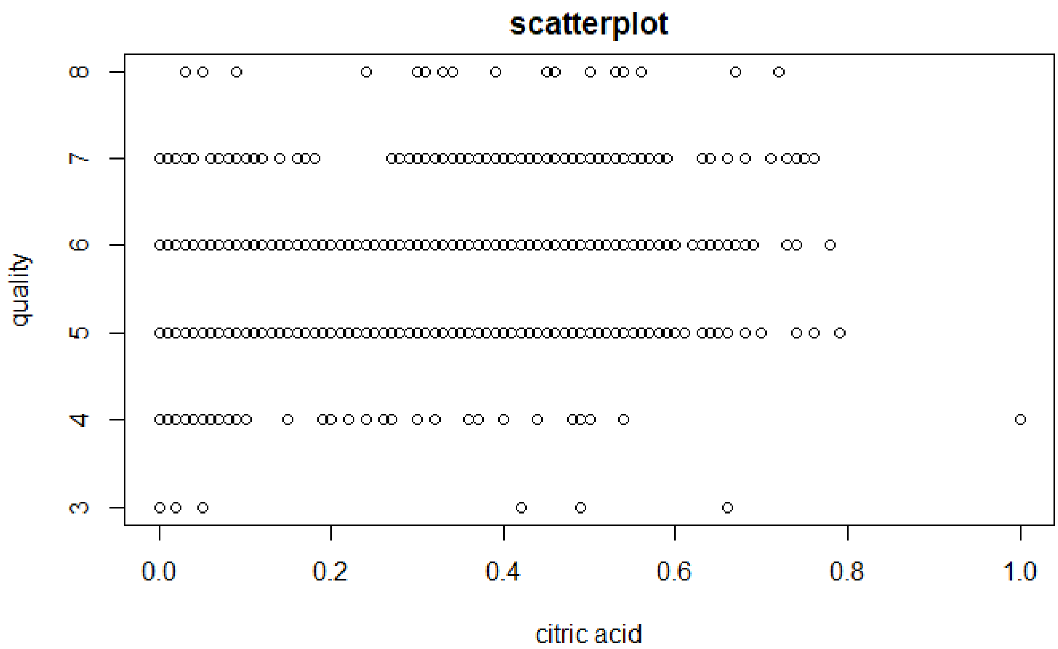


Scatterplot between density and fixed acidity

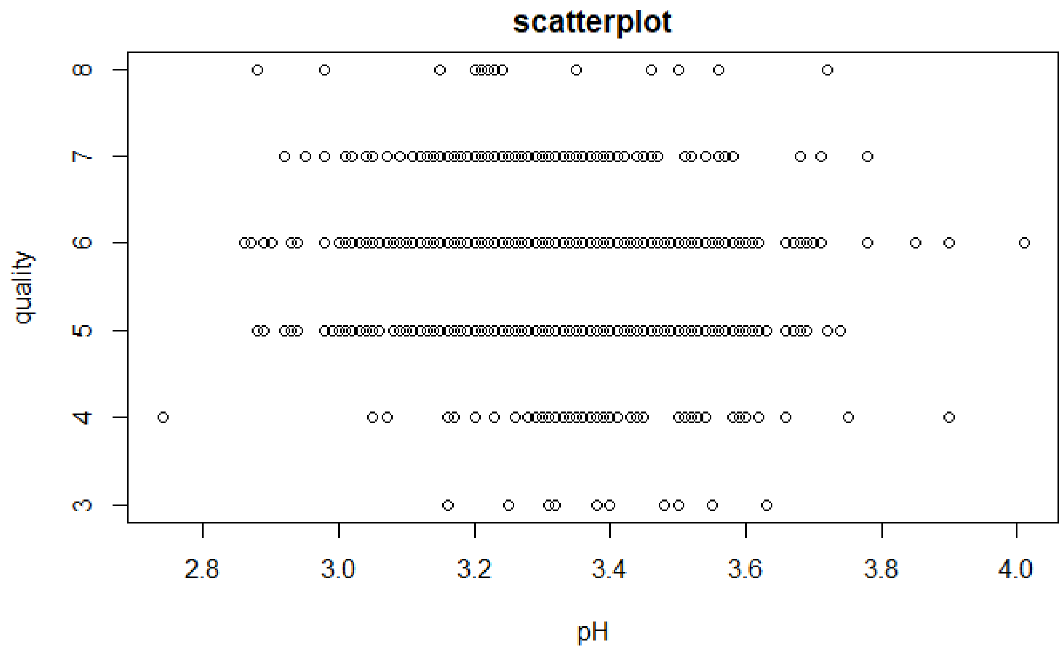
Correlation coefficient= 0.6545253

These are highly positively correlated as the value is close to 1 which means increase in value of one results in increase of another value.

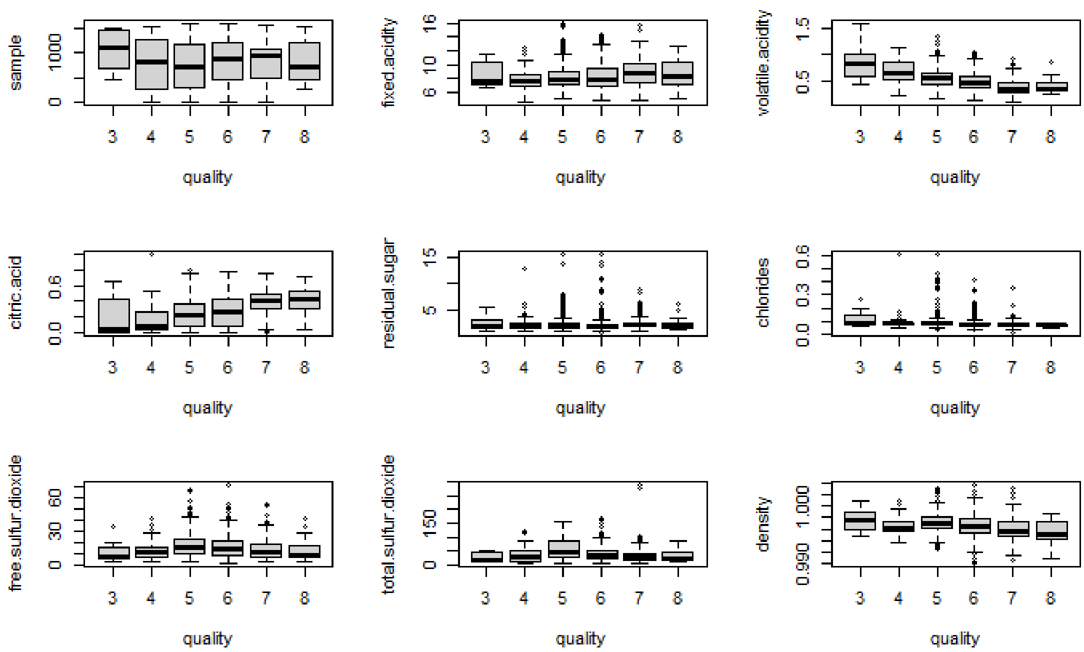


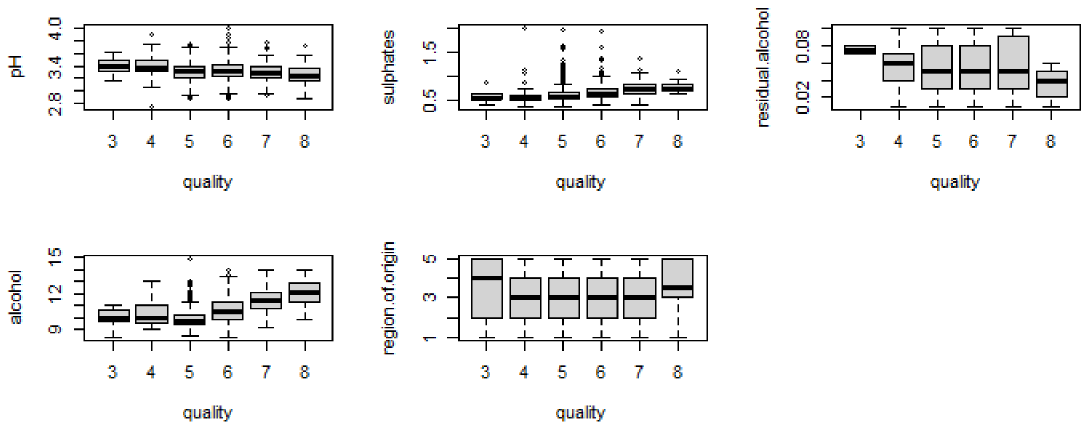


ii)


scatterplot
quality
pH
2.8
3.0
3.2
3.4
3.6
3.8
4.0
3
4
5
6
7
8

iii)


sample
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
quality



3)The attributes which seem to hold the insignificant information are residual alcohol(as there are a lot of missing values) and density(as the standard deviation is close to 0.i.e very less).

4) The replacement of missing values by mean and median is not different if the graph of the attribute is symmetric around its mean. But if the histogram of the data plotted is skewed then replacement with median is considered to be more suitable. replacement by zero or any other value is not going to make any difference.

5) The mean centering technique transforms the data values so that the mean value of the new dataset is zero. The standardisation technique transforms the data values to make the mean value zero and standard deviation of one. The normalization technique scales the values between -1 and 1.

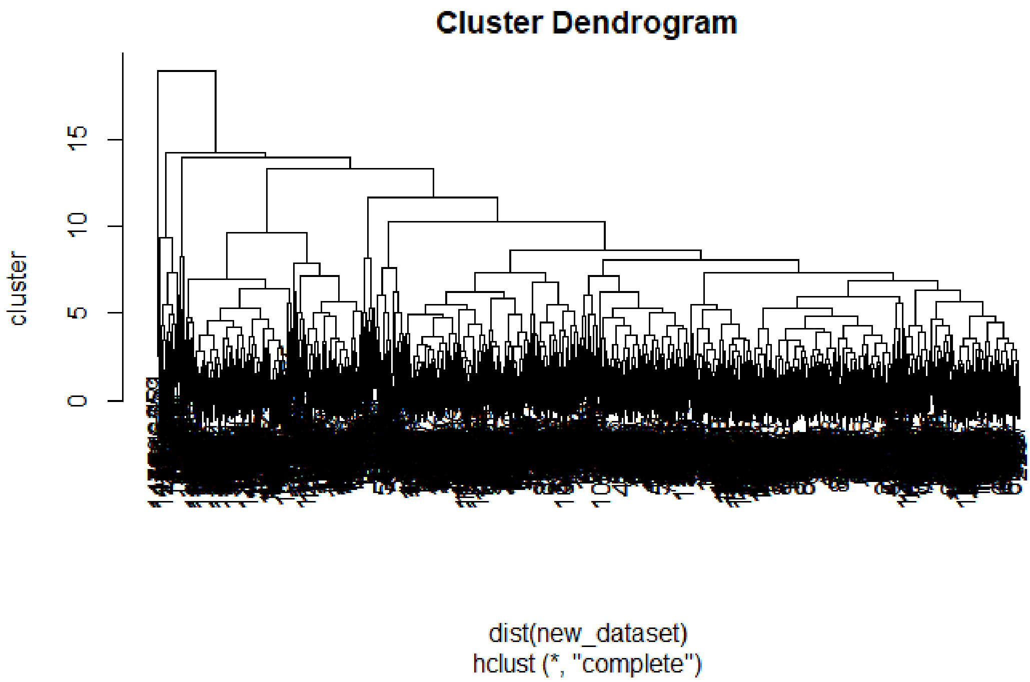
6)i)The residual sugar and density can be deleted because they seem to be insignificant. There are no instances which can be deleted.

ii)First the dataset is standardised and then sent through pca for data transformation. The original dataset matrix is then multiplied to the first eight attributes of the pca to get the reduced dataset. PCA decorrelates the attributes of the dataset.

PART 2: CLUSTERING

1.The dataset I have used is replacement of missing values by median and then transformation and reduction of dataset in part 1.

HCA:



	1	2	3	4	5	6
3	3	0	7	0	0	0
4	23	0	28	1	1	0
5	529	14	109	28	1	1
6	467	8	147	12	0	4
7	159	1	35	0	0	4
8	14	0	4	0	0	0

This is the confusion matrix we get if align it to maximise the diagonal we get an accuracy of 34%

Kmeans:

K-means clustering with 6 clusters of sizes 220, 279, 520, 294, 26, 261

Within cluster sum of squares by cluster:

[1] 2145.6883 1614.9508 2719.6607 1970.4427 329.3112 1308.5687

	1	2	3	4	5	6
3	0	0	7	3	0	0
4	5	15	26	5	1	1
5	172	66	319	92	16	17
6	41	159	160	137	8	133
7	2	35	8	52	1	101
8	0	4	0	5	0	9

External evaluation matrix

If we align this confusion matrix according to maximum diagonal the accuracy comes out to be 37% which is higher than the previous algorithm.

3.PAM

	Size	max_diss	av_diss	diameter	separation
[1,]	254	14.584771	2.874869	16.479467	0.5559814
[2,]	348	6.394477	2.225477	9.952182	0.2009535
[3,]	222	10.080527	2.730477	12.548375	0.2009535
[4,]	157	8.714495	2.506802	10.748211	0.5559814
[5,]	266	5.144978	2.450049	7.750098	0.5871599
[6,]	353	5.333479	2.246102	7.583971	0.3645779

INTERNAL EVALUATION MATRIX

	1	2	3	4	5	6
3	2	5	0	0	0	3
4	3	24	6	2	2	16
5	88	215	184	40	20	135
6	130	100	30	70	136	172
7	29	4	2	41	96	27

8 2 0 0 4 12 0

EXTERNAL EVALUATION MATRIX

The accuracy is 31% in this case.

K-means is the best method for classification because of its high accuracy rate.

2.In hierarchical clustering if we change the method from complete to single, it performs better in terms of accuracy but precision and recall rates become slightly worse.

In K-means the algorithm “Forgy” performs better than the “Lloyd” algorithm in terms of accuracy.

In “Forgy” algorithm the accuracy improves if we decrease the maximum number of iterations from 100 to 50 whearas in case of “Lloyd” algorithm the accuracy remains the same.

PAM

In PAM if we change the distance metric from “Euclidean” to “Manhattan” the accuracy increases.

3.I chose k-means clustering method because it is the best of all. It performs best in the case of reduced dataset followed by dataset obtained by deletion of instances and attributes and worst in the case of replacement of dataset of missing values.

PART 3: CLASSIFICATION

1.The evaluation protocol I used was stratified cross-validation to make sure that all the dataset is tested at some point. The data is split into k random subsets where k is taken as 10. The data is split into 10 folds and 9 folds are combined into a training set to test on 10th fold. The process is repeated for all the 10 folds.

a)**ZEROR**

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	682	42.625 %
Incorrectly Classified Instances	918	57.375 %
Kappa statistic	0	
Mean absolute error	0.2145	
Root mean squared error	0.3273	
Relative absolute error	100	%
Root relative squared error	100	%
Total Number of Instances	1600	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.500	0.006	3
0.000	0.000	?	0.000	?	?	0.480	0.032	4
1.000	1.000	0.426	1.000	0.598	?	0.498	0.425	5
0.000	0.000	?	0.000	?	?	0.498	0.398	6
0.000	0.000	?	0.000	?	?	0.497	0.124	7
0.000	0.000	?	0.000	?	?	0.455	0.011	8

Weighted Avg. 0.426 0.426 ? 0.426 ? ? 0.497 0.356

b)**ONER**

Correctly Classified Instances 801 50.0625 %

Incorrectly Classified Instances 799 49.9375 %

Kappa statistic 0.1685

Mean absolute error 0.1665

Root mean squared error 0.408

Relative absolute error 77.6184 %

Root relative squared error 124.6701 %

Total Number of Instances 1600

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	?	0.000	?	?	0.500	0.006	3
	0.000	0.000	?	0.000	?	?	0.500	0.033	4
	0.614	0.413	0.525	0.614	0.566	0.199	0.601	0.487	5
	0.549	0.391	0.482	0.549	0.513	0.155	0.579	0.444	6
	0.161	0.031	0.421	0.161	0.233	0.201	0.565	0.172	7
	0.000	0.000	?	0.000	?	?	0.500	0.011	8
Weighted Avg.	0.501	0.336	?	0.501	?	?	0.582	0.407	

c)**Naïve Bayes**

Correctly Classified Instances 856 53.5 %

Incorrectly Classified Instances 744 46.5 %

Kappa statistic 0.2862

Mean absolute error 0.1753

Root mean squared error 0.3229

Relative absolute error 81.763 %

Root relative squared error 98.6698 %

Total Number of Instances 1600

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.200	0.012	0.095	0.200	0.129	0.130	0.716	0.059	3
0.094	0.029	0.100	0.094	0.097	0.067	0.675	0.071	4
0.650	0.261	0.649	0.650	0.649	0.388	0.767	0.707	5
0.478	0.294	0.519	0.478	0.498	0.187	0.648	0.513	6
0.508	0.101	0.417	0.508	0.458	0.375	0.827	0.366	7
0.000	0.010	0.000	0.000	0.000	-0.011	0.787	0.047	8

Weighted Avg. 0.535 0.242 0.539 0.535 0.536 0.289 0.724 0.555

d)**IBk**

Correctly Classified Instances	868	54.25 %
Incorrectly Classified Instances	732	45.75 %
Kappa statistic	0.2498	
Mean absolute error	0.1755	
Root mean squared error	0.3178	
Relative absolute error	81.8287 %	
Root relative squared error	97.1031 %	
Total Number of Instances	1600	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.001	0.000	0.000	0.000	-0.002	0.594	0.023	3
0.019	0.005	0.125	0.019	0.033	0.036	0.580	0.045	4
0.717	0.389	0.578	0.717	0.640	0.325	0.730	0.616	5
0.497	0.317	0.510	0.497	0.503	0.181	0.625	0.485	6
0.307	0.044	0.496	0.307	0.379	0.325	0.787	0.370	7
0.000	0.000	?	0.000	?	?	0.586	0.020	8

Weighted 0.543 0.298 ? 0.543 ? ? 0.688 0.504

e)**J48**

Correctly Classified Instances	929	58.0625 %
Incorrectly Classified Instances	671	41.9375 %
Kappa statistic	0.3413	
Mean absolute error	0.1501	
Root mean squared error	0.3537	
Relative absolute error	69.9768 %	
Root relative squared error	108.0726 %	
Total Number of Instances	1600	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.007	0.000	0.000	0.000	-0.007	0.526	0.009	3
	0.057	0.027	0.068	0.057	0.062	0.033	0.520	0.037	4
	0.704	0.260	0.668	0.704	0.685	0.441	0.734	0.613	5
	0.563	0.279	0.573	0.563	0.568	0.285	0.636	0.510	6
	0.437	0.071	0.468	0.437	0.452	0.377	0.731	0.323	7
	0.000	0.008	0.000	0.000	0.000	-0.010	0.617	0.030	8
Weighted Avg.	0.581	0.232	0.573	0.581	0.577	0.349	0.685	0.506	

As we can see clearly the J48 method is the best one in terms of classification because it has the highest percentage of correctly classified instances and weighted average of Precision , Recall and F-measure is also relatively higher compared to the other algorithms.

2.In IBk we optimise k and the batch size. If we take k as 6 and batch size as 100 then we will get the results shown above in part a). But If we reduce k to 1 we get the results as follows:

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	822	51.375 %
Incorrectly Classified Instances	778	48.625 %
Kappa statistic	0.2372	
Mean absolute error	0.1626	
Root mean squared error	0.4018	

Relative absolute error 75.8022 %

Root relative squared error 122.7664 %

Total Number of Instances 1600

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.002	0.000	0.000	0.000	-0.003	0.499	0.006	3
	0.132	0.021	0.179	0.132	0.152	0.129	0.556	0.052	4
	0.607	0.321	0.584	0.607	0.595	0.284	0.643	0.522	5
	0.500	0.323	0.506	0.500	0.503	0.177	0.588	0.452	6
	0.412	0.086	0.406	0.412	0.409	0.324	0.663	0.240	7
	0.000	0.011	0.000	0.000	0.000	-0.011	0.495	0.011	8
Weighted Avg.	0.514	0.277	0.507	0.514	0.510	0.236	0.618	0.435	

As we can see, the number of correctly classified instances decrease from 54% to 51%. So This indicates that choosing k as a higher value results in more precise classifications .Now if we change the batch size from 100 to 10 in both the cases when k=1 and k=6,we observe that there is no change in any value of the matrix. The results remain the same irrespective of the batch size.

3) a) **when the dataset is reduced to 10 principal components**

Correctly Classified Instances 819 51.1875 %

Incorrectly Classified Instances 781 48.8125 %

Kappa statistic 0.2329

Mean absolute error 0.1701

Root mean squared error 0.3658

Relative absolute error 79.3369 %

Root relative squared error 111.7902 %

Total Number of Instances 1600

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.400	0.004	0.364	0.400	0.381	0.377	0.679	0.224	3

0.038	0.027	0.047	0.038	0.042	0.012	0.578	0.046	4
0.629	0.307	0.603	0.629	0.616	0.320	0.694	0.569	5
0.495	0.334	0.496	0.495	0.496	0.162	0.599	0.468	6
0.342	0.085	0.364	0.342	0.352	0.264	0.670	0.242	7
0.000	0.007	0.000	0.000	0.000	-0.009	0.555	0.016	8

Weighted Avg. 0.512 0.276 0.504 0.512 0.508 0.236 0.647 0.462

b) **The dataset after deletion of instances and attributes**

Correctly Classified Instances 942 58.875 %

Incorrectly Classified Instances 658 41.125 %

Kappa statistic 0.3551

Mean absolute error 0.1463

Root mean squared error 0.346

Relative absolute error 68.216 %

Root relative squared error 105.7293 %

Total Number of Instances 1600

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.007	0.000	0.000	0.000	-0.007	0.531	0.009	3
	0.075	0.028	0.083	0.075	0.079	0.049	0.524	0.039	4
	0.702	0.256	0.671	0.702	0.686	0.444	0.747	0.628	5
	0.575	0.270	0.585	0.575	0.580	0.306	0.658	0.540	6
	0.462	0.069	0.487	0.462	0.474	0.402	0.764	0.359	7
	0.000	0.007	0.000	0.000	0.000	-0.009	0.561	0.018	8
Weighted Avg.	0.589	0.227	0.583	0.589	0.585	0.363	0.703	0.529	

c)**Dataset replaced by mean**

Correctly Classified Instances 916 57.25 %

Incorrectly Classified Instances 684 42.75 %

Kappa statistic 0.3293

Mean absolute error 0.1515

Root mean squared error 0.3552

Relative absolute error 70.6206 %

Root relative squared error 108.5267 %

Total Number of Instances 1600

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.007	0.000	0.000	0.000	-0.007	0.531	0.009	3
0.094	0.032	0.093	0.094	0.093	0.062	0.520	0.041	4
0.695	0.271	0.656	0.695	0.675	0.421	0.721	0.590	5
0.545	0.283	0.561	0.545	0.553	0.264	0.626	0.506	6
0.447	0.067	0.486	0.447	0.466	0.394	0.722	0.341	7
0.000	0.006	0.000	0.000	0.000	-0.008	0.675	0.049	8

Dataset replaced by median

Weighted Avg. 0.573 0.238 0.567 0.573 0.569 0.336 0.675 0.497

Correctly Classified Instances 929 58.0625 %

Incorrectly Classified Instances 671 41.9375 %

Kappa statistic 0.3413

Mean absolute error 0.1501

Root mean squared error 0.3537

Relative absolute error 69.9768 %

Root relative squared error 108.0726 %

Total Number of Instances 1600

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.007	0.000	0.000	0.000	-0.007	0.526	0.009	3
0.057	0.027	0.068	0.057	0.062	0.033	0.520	0.037	4
0.704	0.260	0.668	0.704	0.685	0.441	0.734	0.613	5

0.563	0.279	0.573	0.563	0.568	0.285	0.636	0.510	6
0.437	0.071	0.468	0.437	0.452	0.377	0.731	0.323	7
0.000	0.008	0.000	0.000	0.000	-0.010	0.617	0.030	8

Weighted Avg. 0.581 0.232 0.573 0.581 0.577 0.349 0.685 0.506

Dataset replaced by 0

Correctly Classified Instances	933	58.3125 %
Incorrectly Classified Instances	667	41.6875 %
Kappa statistic	0.3422	
Mean absolute error	0.1477	
Root mean squared error	0.3488	
Relative absolute error	68.8623 %	
Root relative squared error	106.5845 %	
Total Number of Instances	1600	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.003	0.000	0.000	0.000	-0.004	0.638	0.036	3
	0.075	0.027	0.087	0.075	0.081	0.052	0.562	0.050	4
	0.695	0.284	0.645	0.695	0.669	0.408	0.732	0.610	5
	0.566	0.273	0.579	0.566	0.572	0.294	0.651	0.525	6
	0.472	0.061	0.525	0.472	0.497	0.431	0.755	0.358	7
	0.000	0.008	0.000	0.000	0.000	-0.009	0.554	0.016	8
Weighted Avg.	0.583	0.239	0.574	0.583	0.578	0.346	0.694	0.516	

We can conclude that the J48 works best when after deletion of instances and attributes because

Correctly classified rate is highest i.e. 58% and also Precision, Recall and F-measure rates are around 0.58(closest to 1 amongst all). When the dataset is replaced by median the figures are also similar but slightly on the lower side as compared to the former case.