CSC 400, Exam #2, At Home, Spring 2024 Name KEY

Instructions: This exam is in two parts: Part I is to be completed partly at home using the materials posted in the course for the at-home portion and you will answer questions about that work during the in-class portion of the exam; Part II is to be completed entirely in class. You may not use cell phones, and you may only access internet resources you are specifically directed to use.

At home, prepare for questions in Part I using R. Complete the calculations noted below. You will be asked for additional analysis and interpretation of this data in the in-class portion of the test. Print out the results of your analysis and code, and bring the pages with you to the exam. You will submit all this work along with the in-class exam.

Use the data on wine to complete the following tasks after importing the data **400exam1data.xlsx** in the file into R (this is the same dataset we used on the first exam.

Set seed to 999

1. Set aside the Type variable for now (save it before removing it from the dataset). Create a clustering model for the data set (you do not need to separate the data into test and training sets for this) using the following algorithms: K-means, Spectral, DBSCAN and mean shift. Compare models with k=2, k=3, k=4 clusters, and any other optimal model identified.

2. For each model identified above, create a confusion matrix (for the k=3 model using the Type data from the original dataset), and create appropriate model or diagnostic graphs.

3. Identify any outliers.

Import the dataset **beersales** from the TSA package.

4. Perform appropriate base time series analysis of the data such as differencing, acf and pacf graphs, decomposition, etc. Create appropriate graphs.

5. Create an ARIMA model of the time series. Create an exponential smoothing model of the time series. Use both to forecast 10 months into the future. Identify any margins of error on your predictions.

6. Redo the above analysis but remove the final 10 observations from the original data set as a test set, and using the remaining observations as a training set. Use your forecast model to predict those last 10 observations and compare the real data to your predictions. Perform appropriate diagnostics.

CSC 400, Exam #2, Spring 2024 Name KEY

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:
Use the work you did at home to answer these questions on the wine dataset.

1. How did rescaling your variables impact the accuracy of your predictions in the k-means model (or any of the clustering models)?

it improved the preformance by a lot, as much as 20%

2. What was the accuracy of your (best) k-means (semi-supervised) model with k=3?

96.63%

3. What was the accuracy of your (best) spectral clustering (semi-supervised) model with k=3?

96.63%

4. What was the accuracy of your (best) DBSCAN (semi-supervised) model with three clusters (this might be 2 clusters plus noise)?

67.98%

5. Describe the hyperparameters you had to set to obtain improvements in your models for spectral clustering and DBSAN.

Set $h = 2.8$

6. What was the accuracy of your (best) mean-shift clustering (semi-supervised) model with k=3?

89.32%

7. Given the four methods you tried here, which algorithm was the most successful (and efficient) and which was the least successful (or least efficient)?

K-means & spectral got the best scores, but K-means was easier to implement.

least was DBSCAN. it got the lowest results and was difficult to find the correct parameter settings

8. For your outlier analysis, which variables contained outliers based on looking at boxplots?

Malic Acid, Ash, Ash-Alcalinity, Magnesium, Proanthocyanins, Color Intensity and Hue

9. For the Ash variable, which observations were potential outliers (based on the boxplots)?

26, 60, 122

10. Which two variables appeared to be the most non-normal?

Malic Acid & OD280/OD315

other variables have long tails, but these are the least linear on the qqplots

11. Which statistical test did you perform on the data to test for possible outliers? Why this test?

answers may vary

Rosner Test

12. Based on the statistical test you chose, how many of the outliers flagged by the boxplots (for all variables), still considered outliers after the test? Give the observation numbers.

3

obs. 60, 95, 70

Ash, Magnesium

13. Based on the outlier analysis you did, how would you proceed? Would you remove the outliers? If so, based on which part of the analysis? If not, why not?

3 observations represent 1.7% of the data

I would consider removing them, but would want to compare model performance w/ and w/o them to see if it affects the outcome at all. If not, they can be left in.

answers may vary

For the questions that follow, use your analysis of the beersales data.

14. Is the original beersales data stationary? Or does it show a trend or seasonal pattern?

it is not. it shows both a slight trend and a seasonal cycle

15. After decomposing the beersales time series, describe the trend? Is it increasing or decreasing? Is it linear or non-linear?

increasing, nonlinear

16. Describe the differences between the standard decomposition output and the LOESS decomposition?

the display order is different
the residuals are not plotted as a line graph
the seasonal and trend components are similar

17. How many differences did you have to take to obtain a stationary series? Explain how you knew when to stop.

the histogram looked most normal after one set of differencing

18. How many lags should be included in your ARIMA model based on the ACF graph? Based on the PACF graph?

ACF = 3
PACF = 1

19. What ARIMA model did you settle on? What were the parameters? How did you determine the best fit?

(1, 1, 3)

Chose these from differencing & ACF / PACF graphs. Then experimented w/ nearby values to test for lowest AIC ~ auto.arima suggested a model of (2,0,2) ✗ (2,1,1)

20. For your exponential smoothing model, what was the (approximate) value of the best smoothing parameter?

approximately 0.10

Part II:

21. What are some potential drawbacks of using (non-linear) regression methods for irregular time series analysis (for interpolation).

interpolation can add bias
but traditional methods will not work at all

22. Why is cross validation essential for developing robust models?

it helps to test performance on data that did not build the model, helps to avoid overfitting

23. What is image segmentation? How can we use it to process images?

dividing up an image into subunits in a sensible way such as separating letters. They can be manipulated or interpreted separately

24. Give three advantages of being able to incorporate spatial information into your data mining analysis.

Some information may be spatially dependent (geographic regions).

You may be able to identify pattern in the data not immediately obvious from other variables. Outliers may come from a particular location

more kinds of data can be analyzed, such as spatiotemporal data

25. How can clustering be used to identify potential outliers or anomalies? How do these outliers differ from more traditional statistical methods?

Some methods, like density-based methods, may flag observations on the periphery of a cluster. They need not be on the outside of the data overall, but still somehow behaving differently from other points

26. In image processing, pixels in the image are typically turned into an array of pixel values (these may be in just one color or multiple colors). Describe some examples of feature engineering in this context. You may use a specific example, such as character recognition, as the basis for your answer.

in a character recognition context, one can engineer features such as distance measures to distinguish characters.

for example a 1 vs a 9 The distance to the one from the left side is nearly the same everywhere, but changes for a 9. A 0 vs 9. The minimum distance is higher for a 9 than for a 0. etc.

answers may vary

27. How does the hierarchical clustering algorithm work?

Start by grouping similar objects together.
then group those groups together w/ other similar groups
and so on until they are all in one group at the top.

28. In the at-home portion of the exam, we looked at how to use clustering methods in a semi-supervised way to create a classification model. How do clustering and classification differ?

Clustering does not use the labels (they may not exist)
to identify groups of points that are similar. In classification,
the labels are a part of the model itself.

29. In the at-home portion of the exam, we applied DBSCAN. Explain the general algorithm steps?

distances between points calculated
points w/in epsilon are identified
core points have more than min pts nearby
if core points not assigned, create new cluster
continue until all core points are assigned and any points w/in eps dist.
any points not assigned are noise

answers may vary some

30. Describe the algorithm steps for Fuzzy-C Means clustering.

Choose a # of clusters
assign coeffs. randomly to being in the cluster
repeat until it converges
compute centroid at center of cluster

Similar to k-means except for fuzzy boundaries and random assignment to coeffs.

31. Why does rescaling improve performance in algorithms that use a Euclidean distance metric?

if variables are very large values, they will dominate in any Euclidean distance calculation, such as income vs. # of children. But rescaling allows all variables to contribute on equal terms.

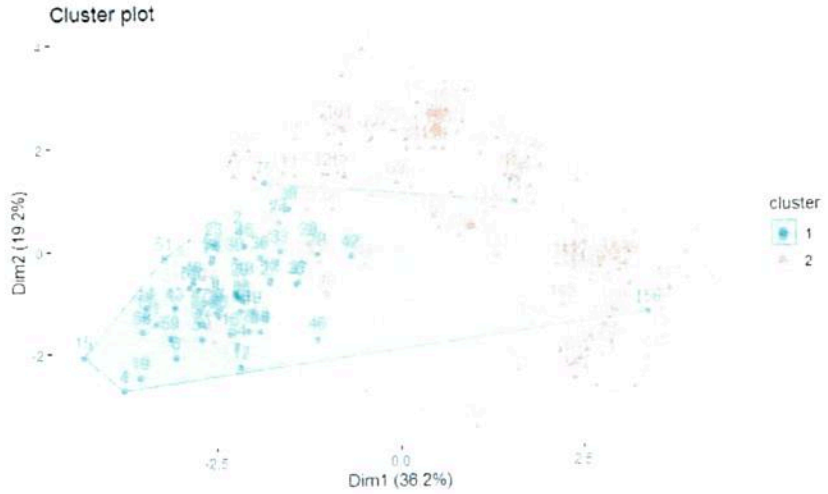
32. Give two examples of other distance metrics that could be employed instead of the standard Euclidean distance metric.

Cosine distance
absolute value (L1 distance)

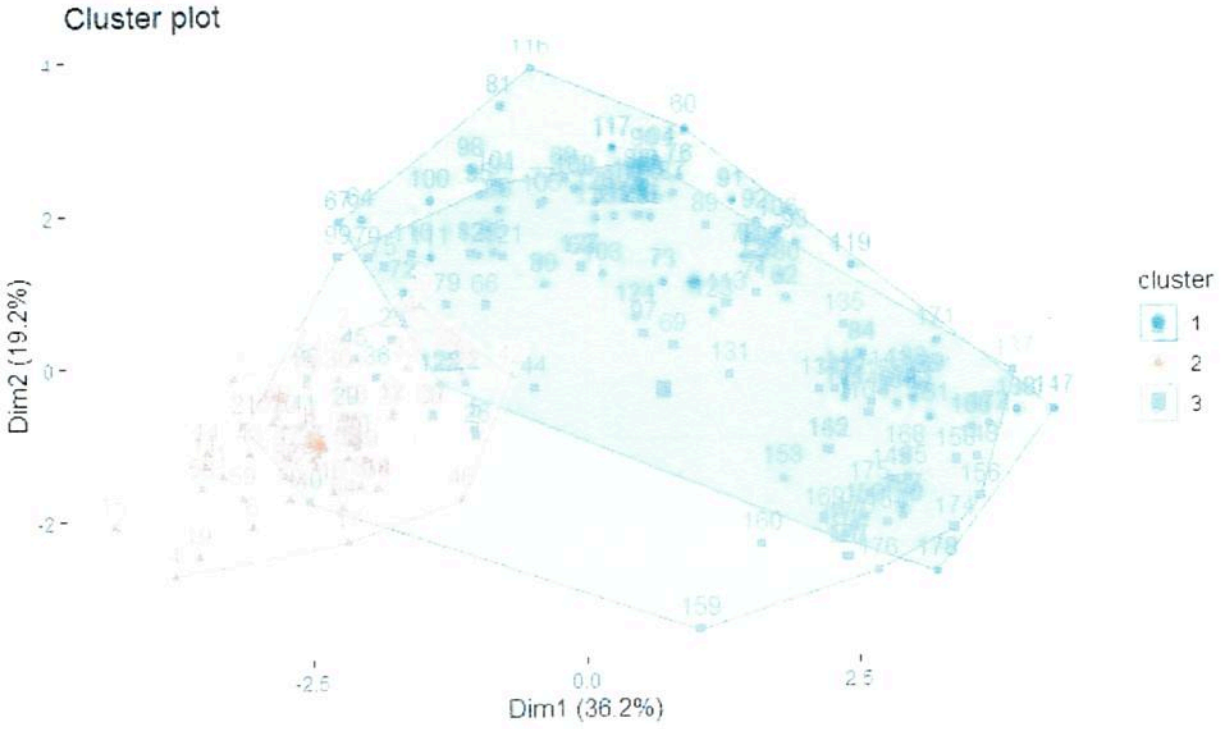
(There are other options)

CSC 400 Exam #2 At-home Analysis, Spring 2024

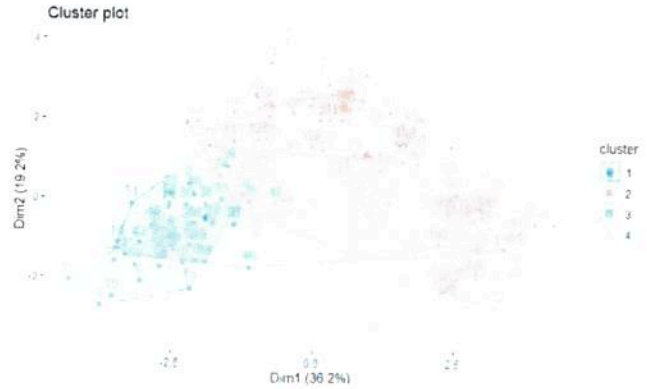
K-means, with k=2, unscaled



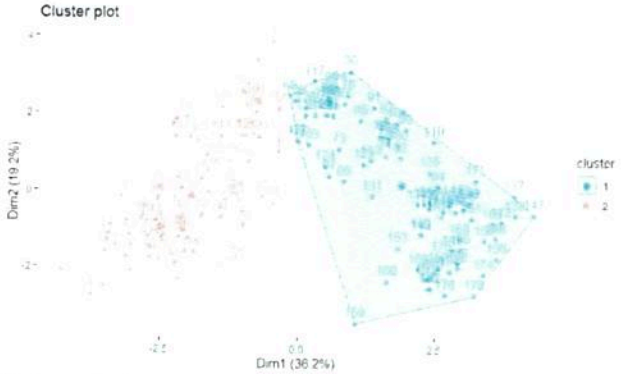
K=3, unscaled



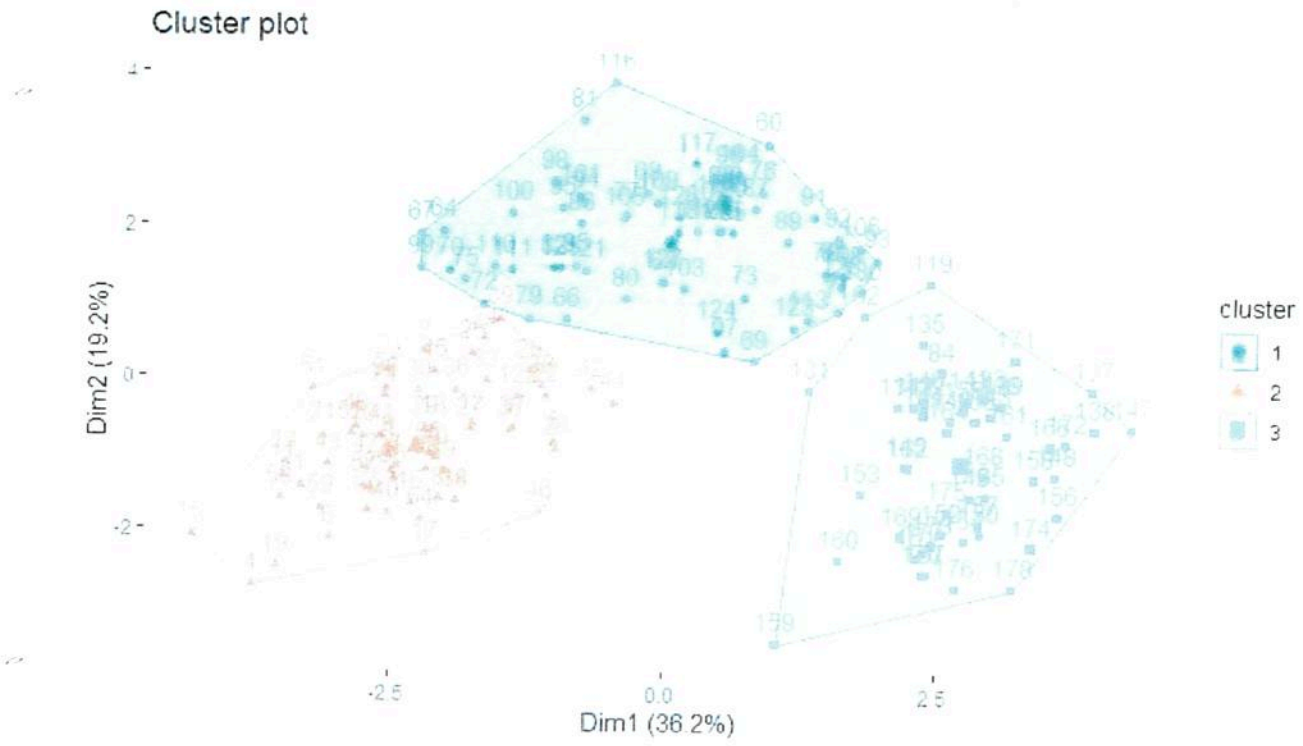
K=4, unscaled



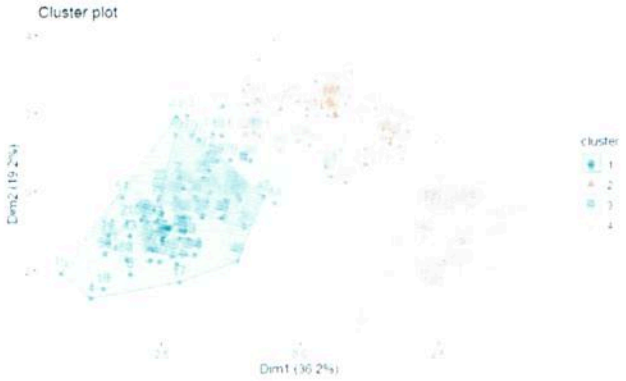
K=2, scaled



K=3, scaled



K=4, scaled



For k=3, unscaled, table:

Cell Contents
N
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total

Total Observations in Table: 178

dd1$Type	dd1$cluster 1	2	3	Row Total
A	0	46	13	59
	22.871	59.406	2.774	
	0.000	0.780	0.220	0.331
	0.000	0.979	0.210	
	0.000	0.258	0.073	
B	50	1	20	71
	18.357	16.801	0.905	
	0.704	0.014	0.282	0.399
	0.725	0.021	0.323	
	0.281	0.006	0.112	
C	19	0	29	48
	0.008	12.674	9.021	
	0.396	0.000	0.604	0.270
	0.275	0.000	0.468	
	0.107	0.000	0.163	
Column Total	69	47	62	178
	0.388	0.264	0.348	

Suggests, A=2, B=1, C=3
Accuracy is (46+50+29)/178 = 125/178 (0.7022...)

K=3, scaled, table:

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

Total Observations in Table: 178

dd2$Type	dd2$cluster 1	2	3	Row Total
A	0	59	0	59
	21.545	71.938	16.904	
	0.000	1.000	0.000	0.331
	0.000	0.952	0.000	
	0.000	0.331	0.000	
B	65	3	3	71
	58.885	19.094	14.785	
	0.915	0.042	0.042	0.399
	1.000	0.048	0.059	
	0.365	0.017	0.017	
C	0	0	48	48
	17.528	16.719	85.282	
	0.000	0.000	1.000	0.270
	0.000	0.000	0.941	
	0.000	0.000	0.270	
Column Total	65	62	51	178
	0.365	0.348	0.287	

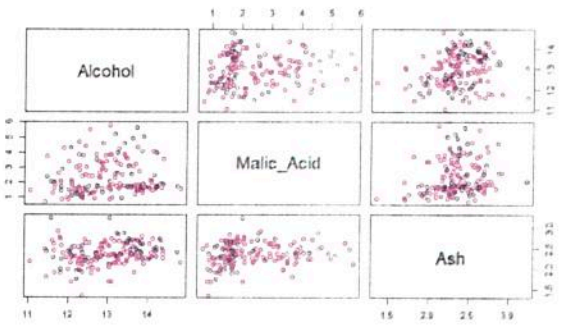
Type: A=2, B=1, C=3

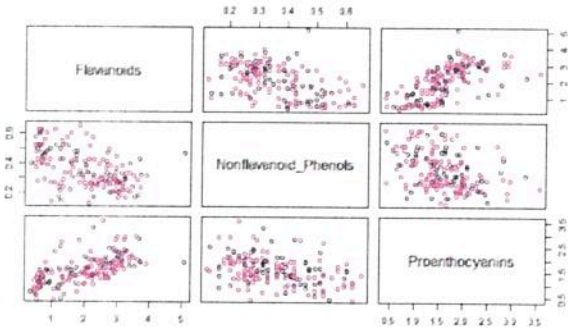
Accuracy = (59+65+48)/178=172/175 = (0.96629...)

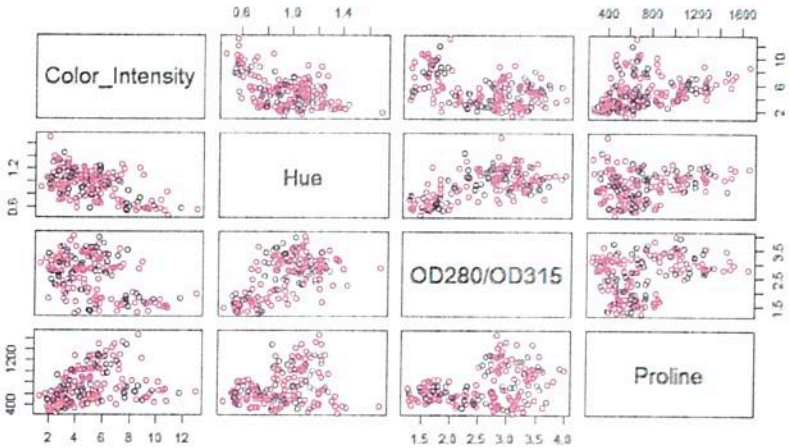
Spectral clustering

K=2, unscaled

```
Spectral Clustering object of class "specc"

 Cluster memberships:

1 2 2 1 1 2 2 2 2 2 2 2 1

String kernel function.  Type =  spectrum
 Hyperparameters : sub-sequence/string length =  4
 Normalized

Cluster size:
[1] 4 9
```

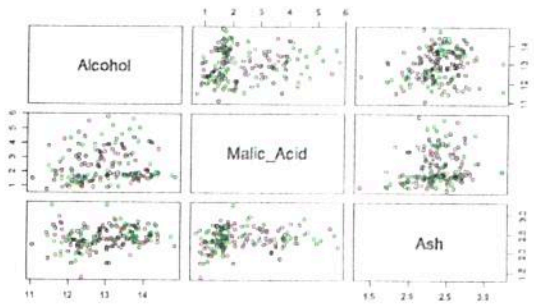


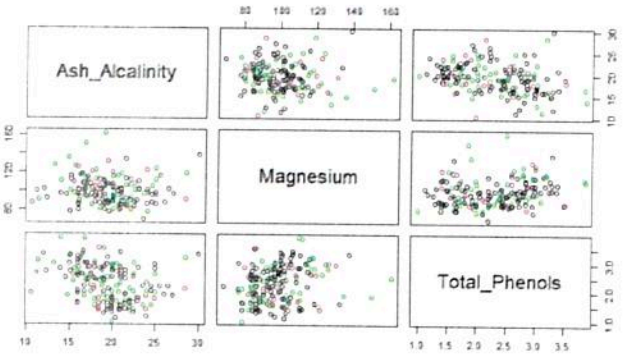


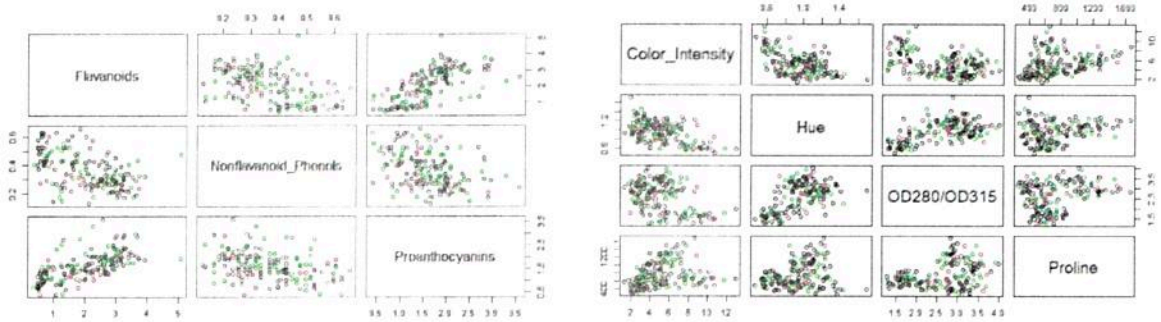


K=3, unscaled

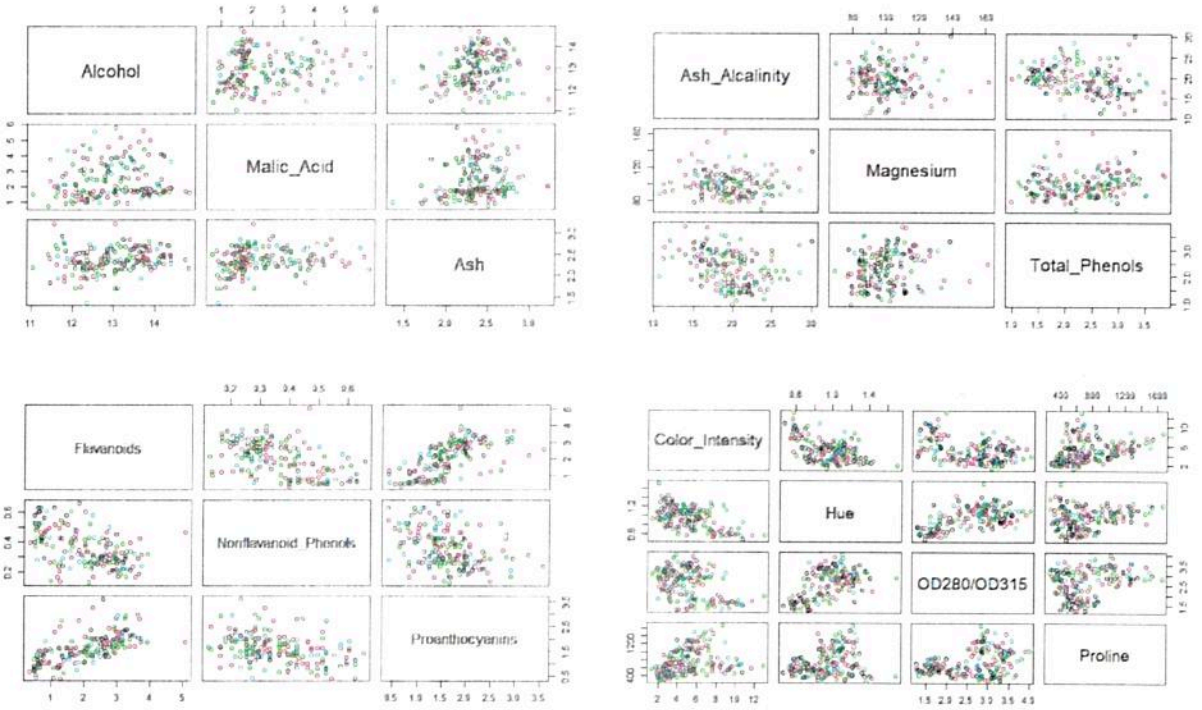
```
Spectral Clustering object of class "specc"

 Cluster memberships:

3 1 1 3 3 1 1 2 1 1 2 1 3

String kernel function.  Type =  spectrum
 Hyperparameters : sub-sequence/string length =  4
 Normalized

Cluster size:
[1] 7 2 4
```







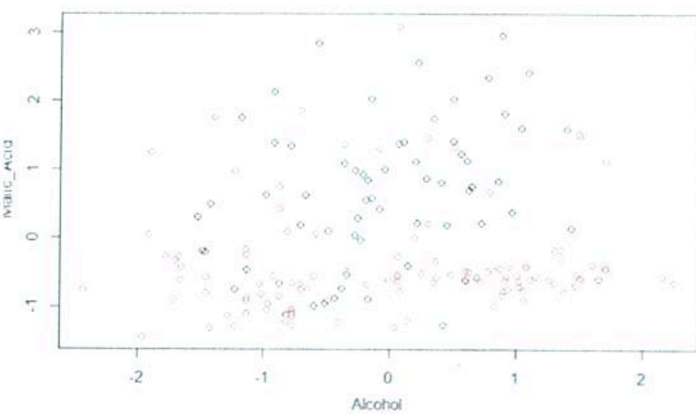

```
Spectral Clustering object of class "specc"

 Cluster memberships:

2 1 3 2 2 3 1 4 1 3 4 3 2

String kernel function.  Type =  spectrum
 Hyperparameters : sub-sequence/string length =  4
 Normalized

Cluster size:
[1] 3 4 4 2
```



With scaled data: k=2

```
Spectral Clustering object of class "specc"

 Cluster memberships:

2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 1 2 1 2 2 2 2 2 2 1
2 2 2 2 2 1 2 2 2 2 2 2 1 1 1 2 2 2 1 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 1 2 2 2 2
```

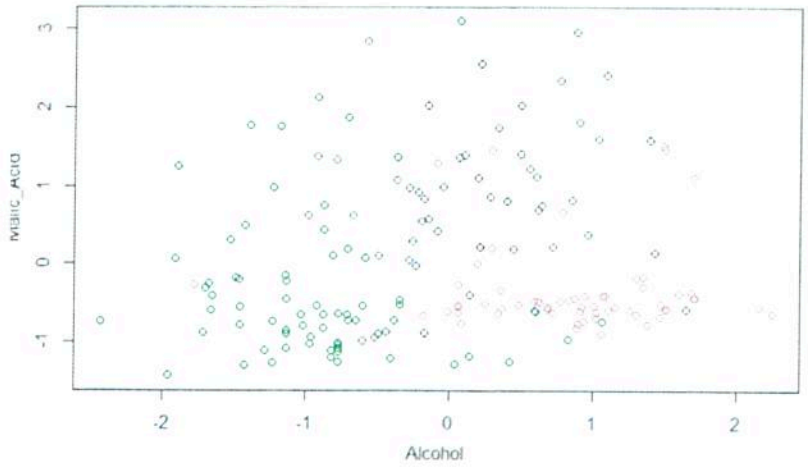
```
2 1 2 2 2 2 2 2 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.324195810029296

Centers:
            [,1]       [,2]       [,3]       [,4]        [,5]       [,6]
[,7]       [,8]
[1,] -0.05946372  0.6445513  0.1875452  0.4899760 -0.10443373 -0.9621197 -1.0
725250  0.7974058
[2,]  0.03338314 -0.3618533 -0.1052885 -0.2750743  0.05862946  0.5401374  0.6
021193 -0.4476664
           [,9]      [,10]      [,11]     [,12]      [,13]
[1,] -0.7228182  0.5684578 -0.8941923 -1.104942 -0.4427475
[2,]  0.4057926 -0.3191342  0.5020027  0.620318  0.2485600

Cluster size:
[1]  64 114

Within-cluster sum of squares:
[1] 1295.780 1592.424
```



K=3, scaled

```
Spectral Clustering object of class "specc"

 Cluster memberships:

2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3
3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 1 3 3 2 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.35412290150503

Centers:
           [,1]       [,2]       [,3]       [,4]       [,5]        [,6]
[,7]        [,8]
[1,]  0.1330775  0.8486578  0.2089687  0.5243829 -0.0276856 -0.97915417 -1.20
853360  0.67581638
[2,]  0.8572514 -0.2959421  0.3795923 -0.6160202  0.5139438  0.89185171  0.98
710283 -0.56418404
[3,] -0.9109594 -0.4011959 -0.5234077  0.1586050 -0.4601680 -0.05364519  0.04
046883 -0.01118808
           [,9]      [,10]      [,11]      [,12]      [,13]
[1,] -0.7635991  0.8996147 -1.1398021 -1.2735177 -0.4055800
[2,]  0.5397613  0.1907608  0.4658715  0.7893709  1.1305174
```

```
[3,]  0.1043340 -0.8987134  0.4746392  0.2780200 -0.7364831

Cluster size:
[1] 52 61 65

Within-cluster sum of squares:
[1] 1159.9931  843.5089  843.5645
```




```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  178
```

dd3$Type	dd3$sc5 1	2	3	Row Total
A	0	59	0	59
	17.236	74.383	21.545	
	0.000	1.000	0.000	0.331
	0.000	0.967	0.000	
	0.000	0.331	0.000	
B	4	2	65	71
	13.513	20.496	58.885	
	0.056	0.028	0.915	0.399
	0.077	0.033	1.000	
	0.022	0.011	0.365	
C	48	0	0	48
	82.330	16.449	17.528	
	1.000	0.000	0.000	0.270
	0.923	0.000	0.000	
	0.270	0.000	0.000	
Column Total	52	61	65	178
	0.292	0.343	0.365	

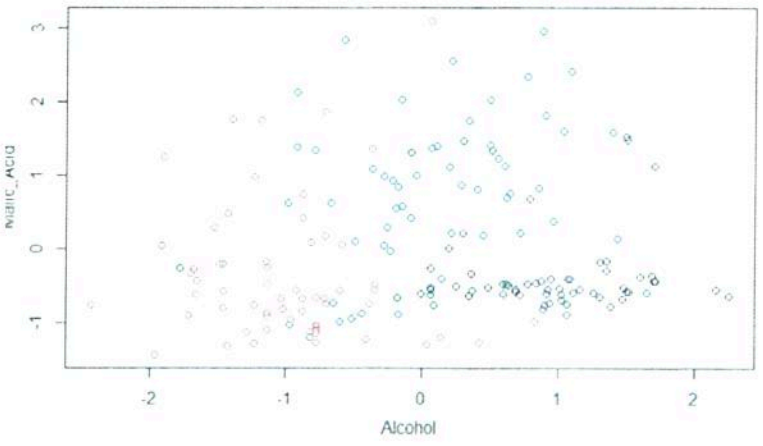
|---|---|---|---|---|

Classes: A=2, B=3, C=1

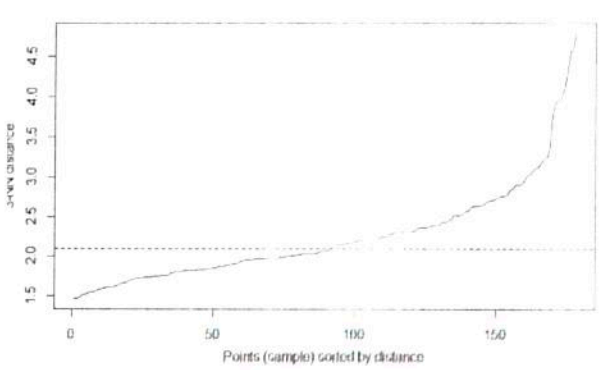
Accuracy: (59+65+48)/178 = 172/178 (0.96629...) same as k-means, scaled, k=3

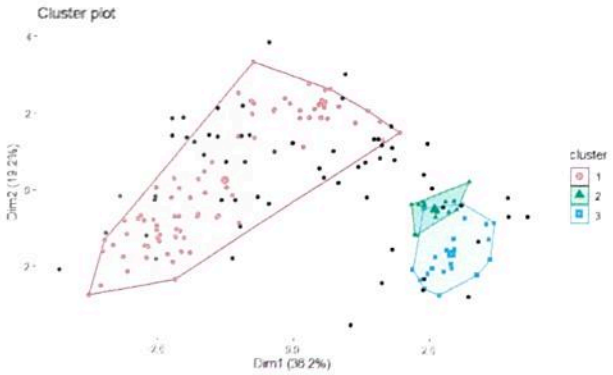
K=4,scaled:

```
Spectral Clustering object of class "specc"

 Cluster memberships:

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 3 1 1 3 1 1 1 1 1 1 1 1 3 3
1 1 3 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 4 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2 2
4 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 4 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 4 2 2 3 2 2 2 2 2 2 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.371044469262749

Centers:
            [,1]         [,2]       [,3]        [,4]       [,5]         [,6]
[,7]
[1,]  1.00788548 -0.363049278  0.3814080 -0.70101900  0.6064963  0.97602965
1.02747994
[2,] -0.94782617 -0.366819468 -0.5041850  0.18068732 -0.6464825 -0.05767501
0.04077688
[3,]  0.09408663  0.008639591  0.4428132  0.05241888 -0.1149361  0.46958081
0.76159956
[4,]  0.08120882  0.748511893  0.1180716  0.45285223  0.1644907 -0.94420317 -
1.15513349
            [,8]         [,9]       [,10]       [,11]       [,12]       [,13]
[1,] -0.57426596  0.634797356  0.2821160  0.4957738  0.7871818  1.2347064
[2,]  0.04437901 -0.004721085 -0.9010578  0.4318757  0.2795510 -0.7757933
[3,] -0.57735640  0.040361258 -0.3757474  0.4399084  0.7990454  0.3687011
[4,]  0.59869589 -0.602274343  0.8009449 -1.0277043 -1.1995731 -0.3739701

Cluster size:
[1] 52 61 10 55

Within-cluster sum of squares:
[1]  772.7500  769.6165  116.8056 1198.5754
```



DBSCAN





DBSCAN does poorly for this data. Calibrating epsilon is difficult, and setting it too high gives too many small clusters, and setting it too low produces one large cluster with lots of noise.

Cell Contents
N
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total

Total Observations in Table: 178

	dd4$`Dbscan1$cluster`				
dd4$Type	0	1	2	3	Row Total
A	10	49	0	0	59
	5.165	16.785	4.309	6.961	
	0.169	0.831	0.000	0.000	0.331
	0.164	0.590	0.000	0.000	
	0.056	0.275	0.000	0.000	
B	37	34	0	0	71
	6.596	0.024	5.185	8.376	
	0.521	0.479	0.000	0.000	0.399
	0.607	0.410	0.000	0.000	
	0.208	0.191	0.000	0.000	
C	14	0	13	21	48
	0.365	22.382	25.714	41.538	
	0.292	0.000	0.271	0.438	0.270
	0.230	0.000	1.000	1.000	
	0.079	0.000	0.073	0.118	
Column Total	61	83	13	21	178
	0.343	0.466	0.073	0.118	

Most As are in Cluster 1, and C's have a plurality in cluster 3, but B's are also mostly put in class 1, and only C's are in class 2. Most of the noise is actually class B.

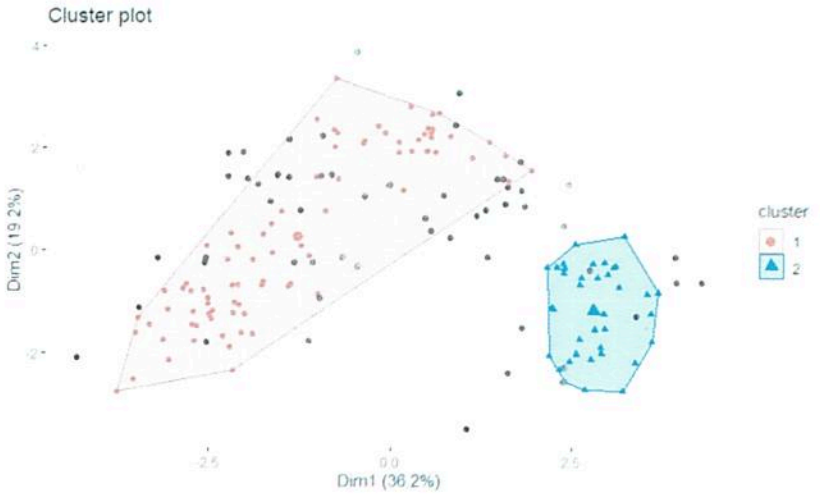
We can get 2 clusters plus the "noise" by adjusting the epsilon upward slightly. It seems to produce slightly better results, but B is still frequently misclassified.

Cell Contents
N
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total

Total Observations in Table: 178

	dd5$`Dbscan2$cluster`			
dd5$Type	0	1	2	Row Total
A	9	50	0	59
	4.926	16.908	12.264	
	0.153	0.847	0.000	0.331
	0.161	0.588	0.000	
	0.051	0.281	0.000	
B	35	35	1	71
	7.179	0.035	12.826	
	0.493	0.493	0.014	0.399
	0.625	0.412	0.027	
	0.197	0.197	0.006	
C	12	0	36	48
	0.637	22.921	67.869	
	0.250	0.000	0.750	0.270
	0.214	0.000	0.973	
	0.067	0.000	0.202	
Column Total	56	85	37	178
	0.315	0.478	0.208	

Class A=1, B=0 (equal in 1, but since A is 1, we'll use the noise), and C=2
Accuracy= (50+35+36)/178 = 121/178 (0.67977...)



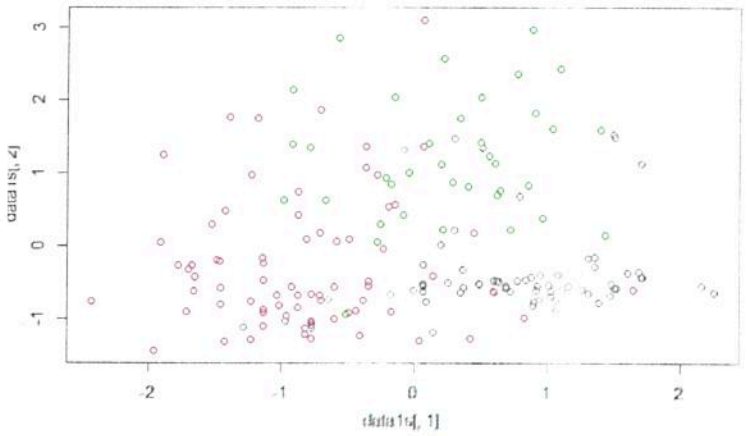
Mean Shift Clustering
After some finagling with h (here 2.8), I got three clusters.

Cell Contents
N
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total

Total Observations in Table: 178

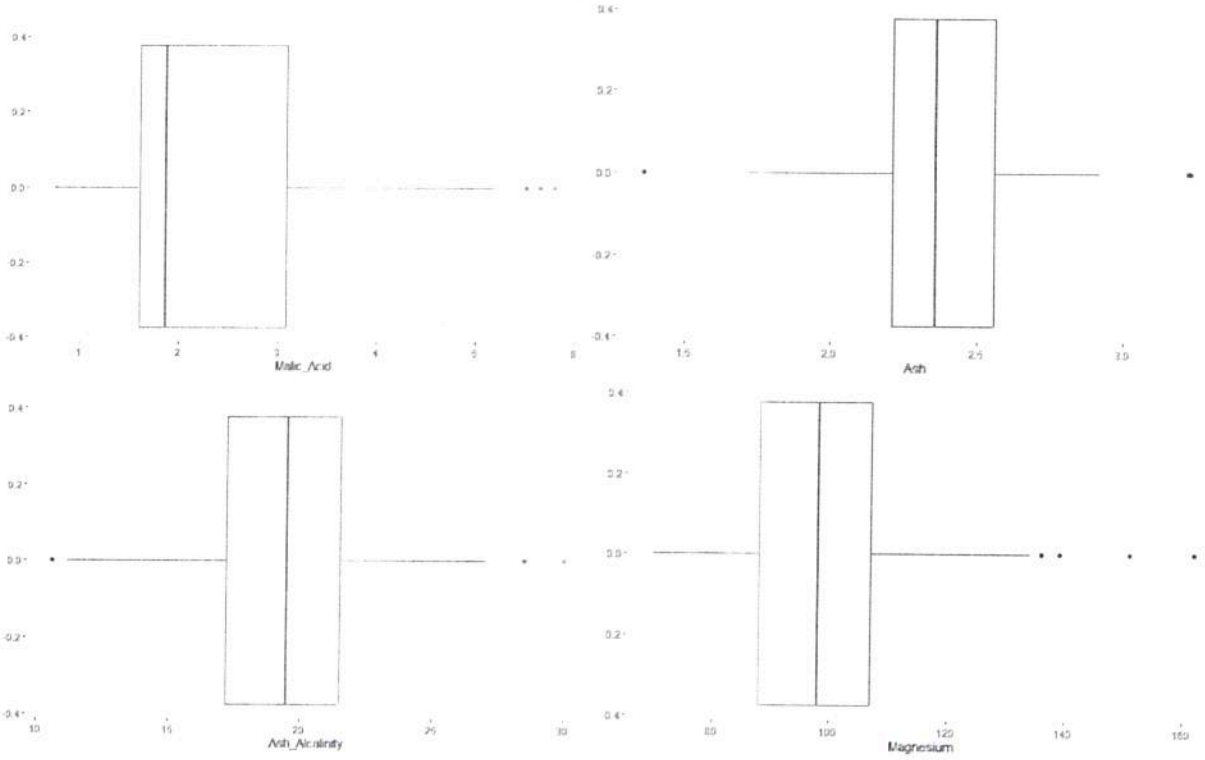
dd6$Type	dd6$meanShiftR_assignment 1	2	3	Row Total
A	59	0	0	59
	60.954	24.528	12.264	
	1.000	0.000	0.000	0.331
	0.881	0.000	0.000	
	0.331	0.000	0.000	
B	8	63	0	71
	13.120	37.982	14.758	
	0.113	0.887	0.000	0.399
	0.119	0.851	0.000	
	0.045	0.354	0.000	
C	0	11	37	48
	18.067	4.019	73.186	
	0.000	0.229	0.771	0.270
	0.000	0.149	1.000	
	0.000	0.062	0.208	
Column Total	67	74	37	178
	0.376	0.416	0.208	

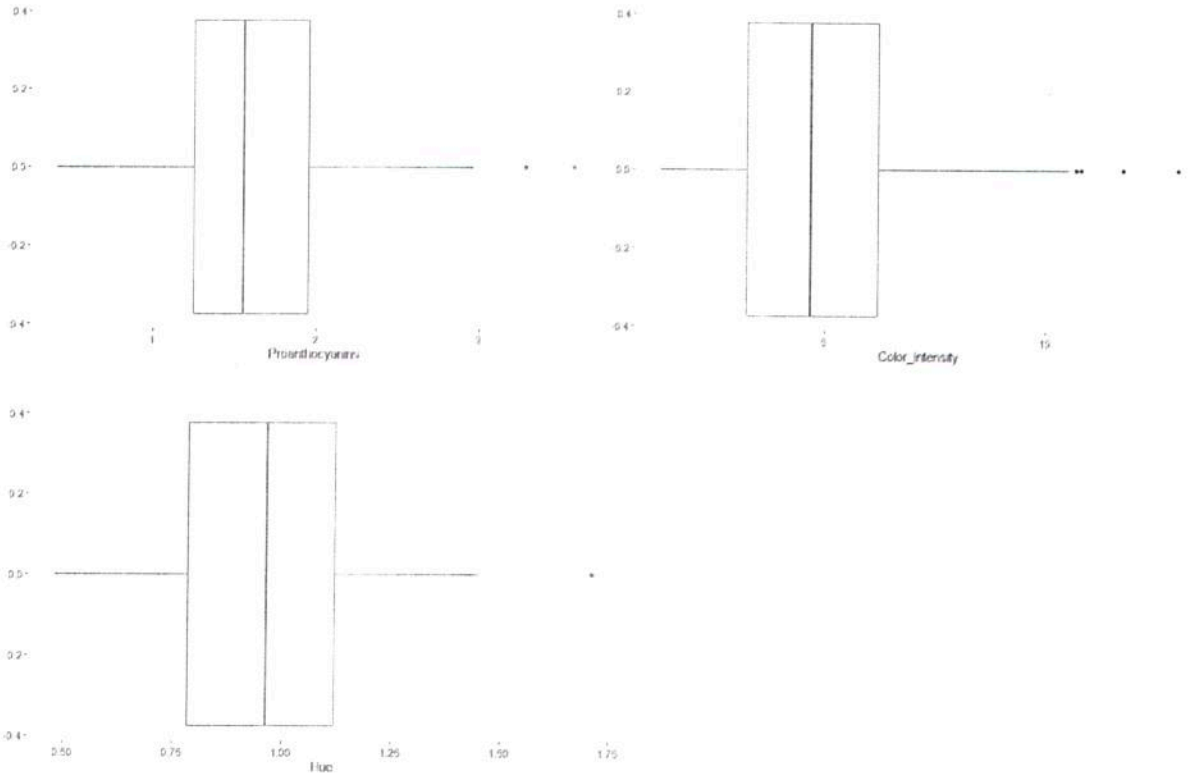
Accuracy = (59+63+37)/178 = 159/178 (0.8932..)



Outliers.

Boxplots are here for any variable that displayed potential outliers. Others were omitted



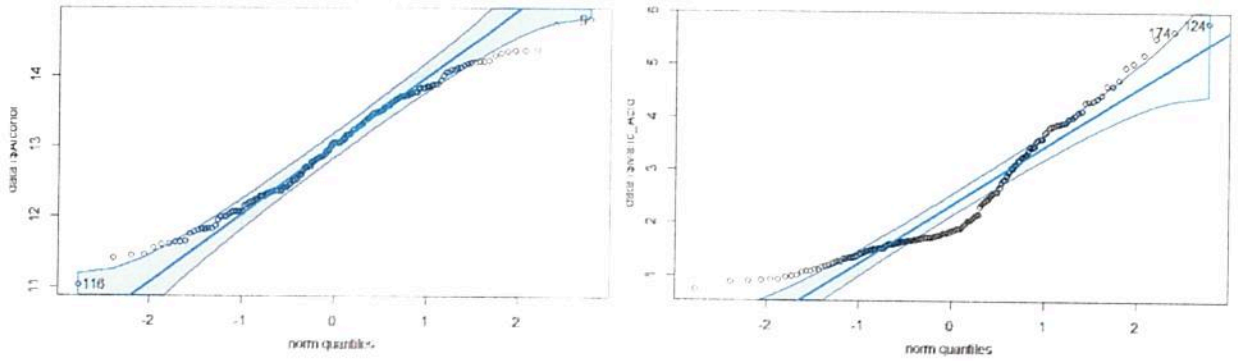


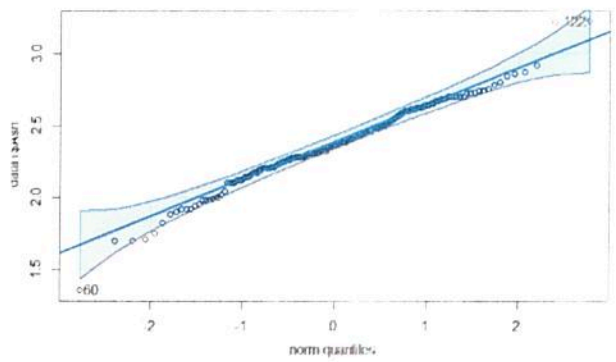
Observations that are potential outliers in boxplots:

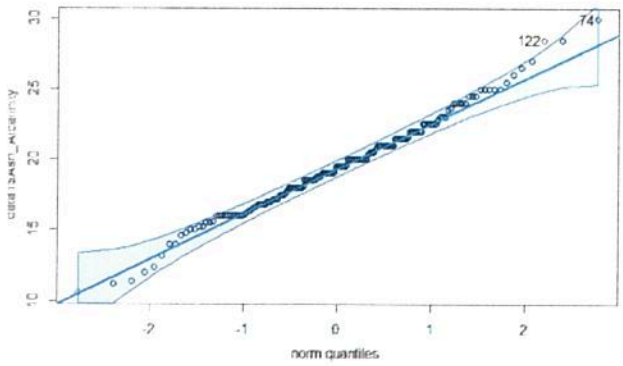
```
Malic_Acid: 124 138 174

Ash: 26  60 122
Ash_Alcalinity: 60  74 122 128
Magnesium: 70 74 79 96
Proanthocyanins: 96 111
Color_Intensity: 152 159 160
Hue: 116
```

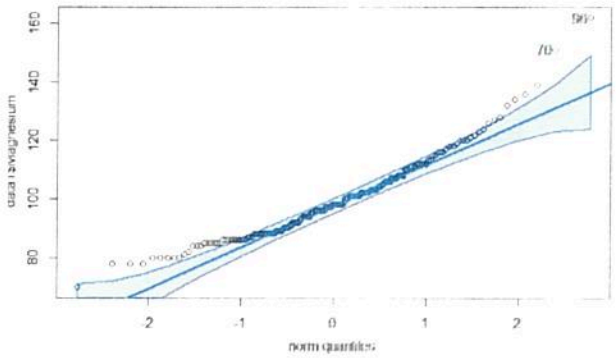
Some outliers are in common to more than one variable.

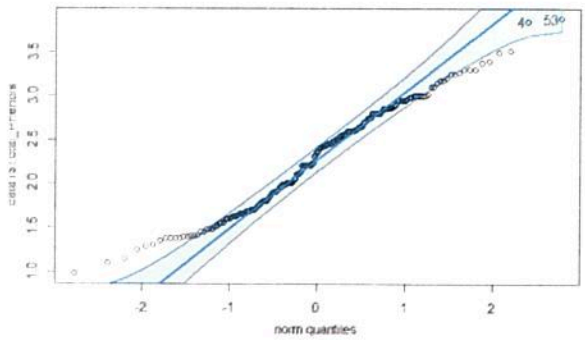
Normality?

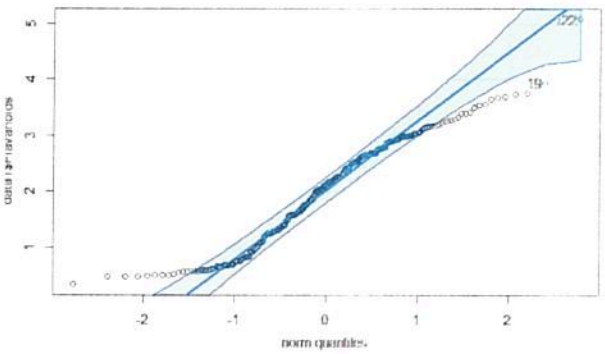


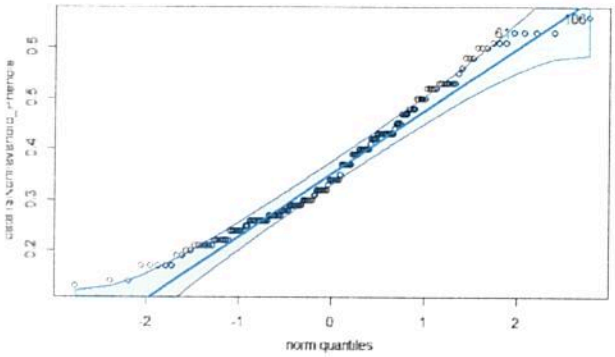

norm quantiles
122
60

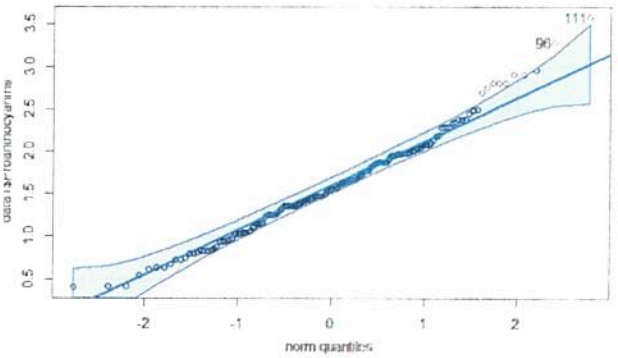

norm quantiles
74
122

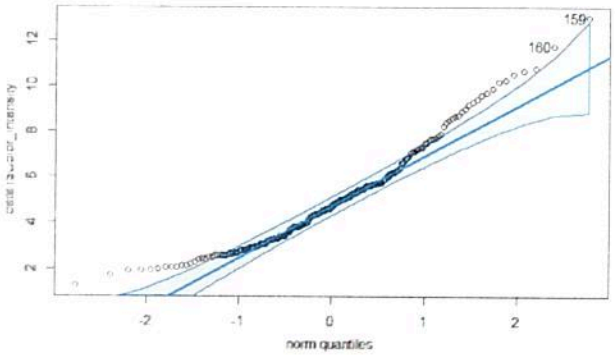

norm quantiles
99
70

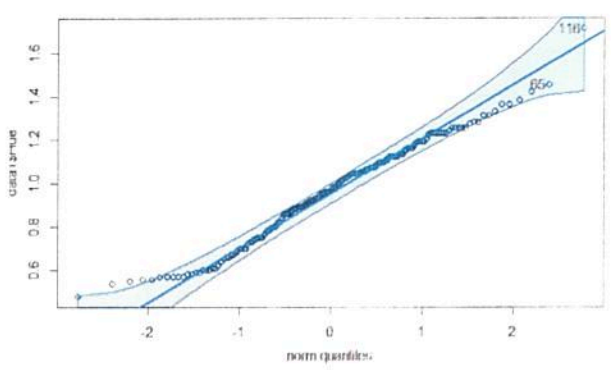

norm quantiles
40
50

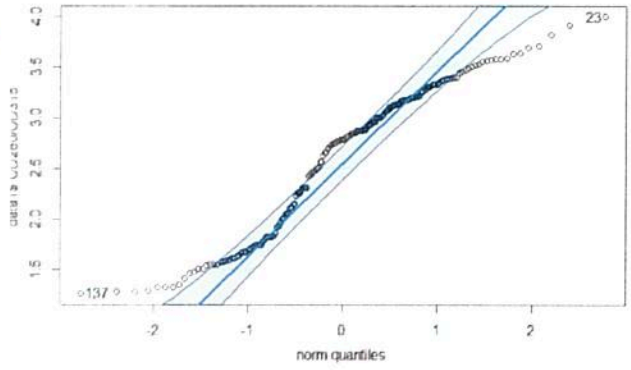

norm quantiles
122
19

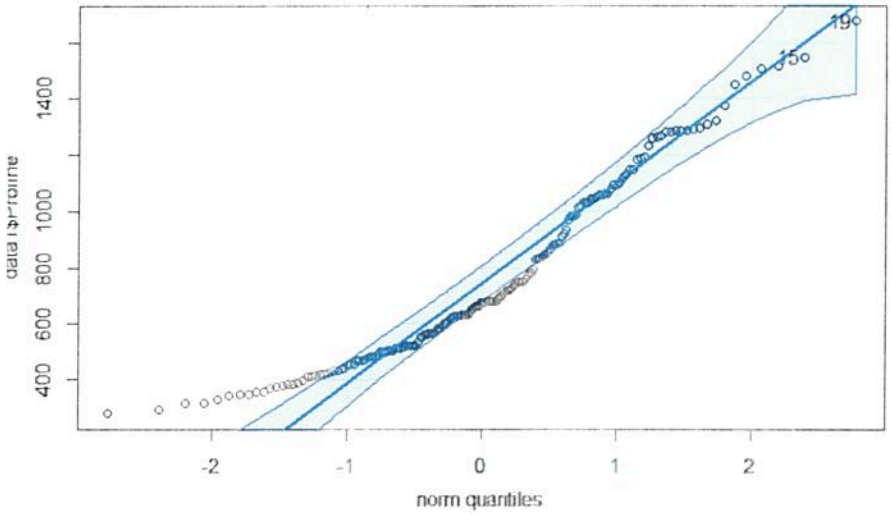

norm quantiles
61
104


norm quantiles
111
96


norm quantiles
159
160







Will use Rosner Test to identify outliers for each variable, if any. Will test k=4 for each variable, since as many as 4 outliers were identified in some variables in the box plots. Results below.

Results of Outlier Test

Test Method:	Rosner's Test for Outliers
Hypothesized Distribution:	Normal
Data:	data1$Alcohol
Sample Size:	178
Test Statistics:	R.1 = 2.427388 R.2 = 2.271714 R.3 = 2.211303 R.4 = 2.022822
Test Statistic Parameter:	k = 4
Alternative Hypothesis:	Up to 4 observations are not from the same Distribution.
Type I Error:	5%
Number of Outliers Detected:	0

```
  i    Mean.i       SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 13.00062 0.8118265 11.03     116 2.427388   3.570191   FALSE
2 1 13.01175 0.8003861 14.83       9 2.271714   3.568466   FALSE
3 2 13.00142 0.7907463 14.75      14 2.211303   3.566728   FALSE
4 3 12.99143 0.7817933 11.41     114 2.022822   3.564980   FALSE

Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            data1$Malic_Acid

Sample Size:                     178

Test Statistics:                 R.1 = 3.100446
                                 R.2 = 3.059988
                                 R.3 = 3.022429
                                 R.4 = 2.805398

Test Statistic Parameter:        k = 4

Alternative Hypothesis:          Up to 4 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i    Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 2.336348 1.117146  5.80     124 3.100446   3.570191   FALSE
2 1 2.316780 1.089292  5.65     174 3.059988   3.568466   FALSE
3 2 2.297841 1.062774  5.51     138 3.022429   3.566728   FALSE
4 3 2.279486 1.037469  5.19     156 2.805398   3.564980   FALSE

Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            data1$Ash

Sample Size:                     178

Test Statistics:                 R.1 = 3.668813
                                 R.2 = 3.244400
                                 R.3 = 3.317143
                                 R.4 = 2.655102

Test Statistic Parameter:        k = 4

Alternative Hypothesis:          Up to 4 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     1

  i    Mean.i       SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
```

```
  i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 2.366517 0.2743440  1.36      60 3.668813   3.570191    TRUE
2 1 2.372203 0.2643930  3.23     122 3.244400   3.568466   FALSE
3 2 2.367330 0.2570497  3.22      26 3.317143   3.566728   FALSE
4 3 2.362457 0.2495035  1.70      67 2.655102   3.564980   FALSE

Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            data1$Ash_Alcalinity

Sample Size:                     178

Test Statistics:                 R.1 = 3.145637
                                 R.2 = 2.786018
                                 R.3 = 2.857974
                                 R.4 = 2.796197

Test Statistic Parameter:        k = 4

Alternative Hypothesis:          Up to 4 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 19.49494 3.339564  30.0      74 3.145637   3.570191   FALSE
2 1 19.43559 3.253535  28.5     122 2.786018   3.568466   FALSE
3 2 19.38409 3.189640  28.5     128 2.857974   3.566728   FALSE
4 3 19.33200 3.122813  10.6      60 2.796197   3.564980   FALSE

Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            data1$Magnesium

Sample Size:                     178

Test Statistics:                 R.1 = 4.359076
                                 R.2 = 3.815127
                                 R.3 = 3.071864
                                 R.4 = 2.930870

Test Statistic Parameter:        k = 4

Alternative Hypothesis:          Up to 4 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     2

  i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 99.74157 14.28248   162      96 4.359076   3.570191    TRUE
2 1 99.38983 13.52777   151      70 3.815127   3.568466    TRUE
```

```
3 2 99.09659 12.98996   139       74 3.071864   3.566728   FALSE
4 3 98.86857 12.66908   136       79 2.930870   3.564980   FALSE

Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            data1$Total_Phenols

Sample Size:                     178

Test Statistics:                 R.1 = 2.532372
                                 R.2 = 2.538348
                                 R.3 = 2.139430
                                 R.4 = 2.058677

Test Statistic Parameter:        k = 4

Alternative Hypothesis:          Up to 4 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 2.295112 0.6258510  3.88      53 2.532372   3.570191   FALSE
2 1 2.286158 0.6160864  3.85       4 2.538348   3.568466   FALSE
3 2 2.277273 0.6063638  0.98     167 2.139430   3.566728   FALSE
4 3 2.284686 0.6000525  3.52      99 2.058677   3.564980   FALSE

Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            data1$Flavanoids

Sample Size:                     178

Test Statistics:                 R.1 = 3.054216
                                 R.2 = 1.967576
                                 R.3 = 1.809117
                                 R.4 = 1.821143

Test Statistic Parameter:        k = 4

Alternative Hypothesis:          Up to 4 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 2.029270 0.9988587  5.08     122 3.054216   3.570191   FALSE
2 1 2.012034 0.9747863  3.93      19 1.967576   3.568466   FALSE
3 2 2.001136 0.9666944  3.75      99 1.809117   3.566728   FALSE
4 3 1.991143 0.9603076  3.74      53 1.821143   3.564980   FALSE
```

Results of Outlier Test

Test Method: Rosner's Test for Outliers

Hypothesized Distribution: Normal

Data: data1$Nonflavanoid_Phenols

Sample Size: 178

Test Statistics: R.1 = 2.395645
R.2 = 2.198127
R.3 = 2.235402
R.4 = 2.274643

Test Statistic Parameter: k = 4

Alternative Hypothesis: Up to 4 observations are not from the same Distribution.

Type I Error: 5%

Number of Outliers Detected: 0

	i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
1	0	0.3618539	0.1244533	0.66	106	2.395645	3.570191	FALSE
2	1	0.3601695	0.1227547	0.63	61	2.198127	3.568466	FALSE
3	2	0.3586364	0.1213936	0.63	136	2.235402	3.566728	FALSE
4	3	0.3570857	0.1199832	0.63	138	2.274643	3.564980	FALSE

Results of Outlier Test

Test Method: Rosner's Test for Outliers

Hypothesized Distribution: Normal

Data: data1$Proanthocyanins

Sample Size: 178

Test Statistics: R.1 = 3.475269
R.2 = 3.069540
R.3 = 2.572366
R.4 = 2.536085

Test Statistic Parameter: k = 4

Alternative Hypothesis: Up to 4 observations are not from the same Distribution.

Type I Error: 5%

Number of Outliers Detected: 0

	i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
1	0	1.590899	0.5723589	3.58	112	3.475269	3.570191	FALSE
2	1	1.579661	0.5539393	3.28	96	3.069540	3.568466	FALSE
3	2	1.570000	0.5403586	2.96	13	2.572366	3.566728	FALSE
4	3	1.562057	0.5315054	2.91	51	2.536085	3.564980	FALSE

Results of Outlier Test

Test Method: Rosner's Test for Outliers

Hypothesized Distribution: Normal

Data: data1$Color_Intensity

Sample Size: 178

Test Statistics: R.1 = 3.425768
R.2 = 2.999436
R.3 = 2.655279
R.4 = 2.662772

Test Statistic Parameter: k = 4

Alternative Hypothesis: Up to 4 observations are not from the same Distribution.

Type I Error: 5%

Number of Outliers Detected: 0

	i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
1	0	5.058090	2.318286	13.00	159	3.425768	3.570191	FALSE
2	1	5.013220	2.246016	11.75	160	2.999436	3.568466	FALSE
3	2	4.974943	2.193764	10.80	152	2.655279	3.566728	FALSE
4	3	4.941657	2.155026	10.68	167	2.662772	3.564980	FALSE

Results of Outlier Test

Test Method: Rosner's Test for Outliers

Hypothesized Distribution: Normal

Data: data1$Hue

Sample Size: 178

Test Statistics: R.1 = 3.292407
R.2 = 2.237353
R.3 = 2.143184
R.4 = 2.149999

Test Statistic Parameter: k = 4

Alternative Hypothesis: Up to 4 observations are not from the same Distribution.

Type I Error: 5%

Number of Outliers Detected: 0

	i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
1	0	0.9574494	0.2285716	1.71	176	3.292407	3.570191	FALSE
2	1	0.9531977	0.2220492	1.45	65	2.237353	3.568466	FALSE
3	2	0.9503750	0.2194748	0.48	142	2.143184	3.566728	FALSE
4	3	0.9530629	0.2171801	1.42	100	2.149999	3.564980	FALSE

Results of Outlier Test

Test Method: Rosner's Test for Outliers

Hypothesized Distribution: Normal

Data: data1$`OD280/OD315`

Sample Size: 178

Test Statistics: R.1 = 1.955399
R.2 = 1.894048
R.3 = 1.890548
R.4 = 1.915688

Test Statistic Parameter: k = 4

Alternative Hypothesis: Up to 4 observations are not from the same Distribution.

Type I Error: 5%

Number of Outliers Detected: 0

	i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
1	0	2.611685	0.7099904	4.00	23	1.955399	3.570191	FALSE
2	1	2.603842	0.7042283	1.27	137	1.894048	3.568466	FALSE
3	2	2.611420	0.6989615	1.29	131	1.890548	3.566728	FALSE
4	3	2.618971	0.6937306	1.29	134	1.915688	3.564980	FALSE

Results of Outlier Test

Test Method: Rosner's Test for Outliers

Hypothesized Distribution: Normal

Data: data1$Proline

Sample Size: 178

Test Statistics: R.1 = 2.963114
R.2 = 2.616367
R.3 = 2.570869
R.4 = 2.611768

Test Statistic Parameter: k = 4

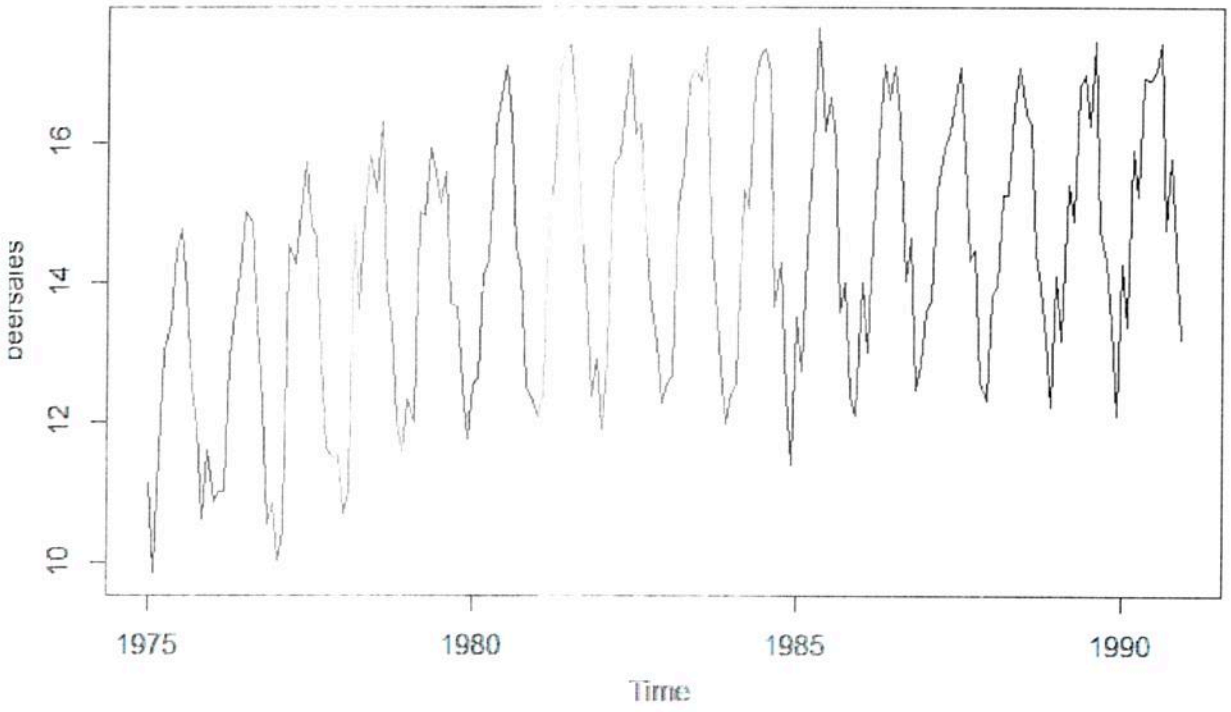
Alternative Hypothesis: Up to 4 observations are not from the same Distribution.

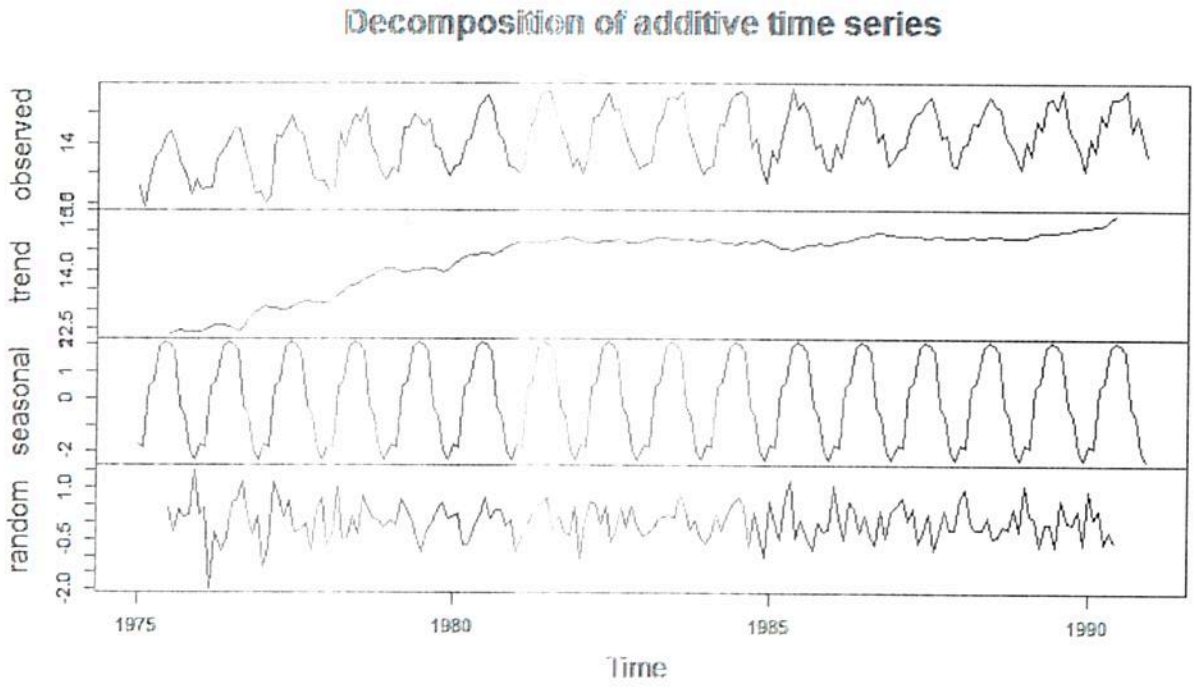
Type I Error: 5%

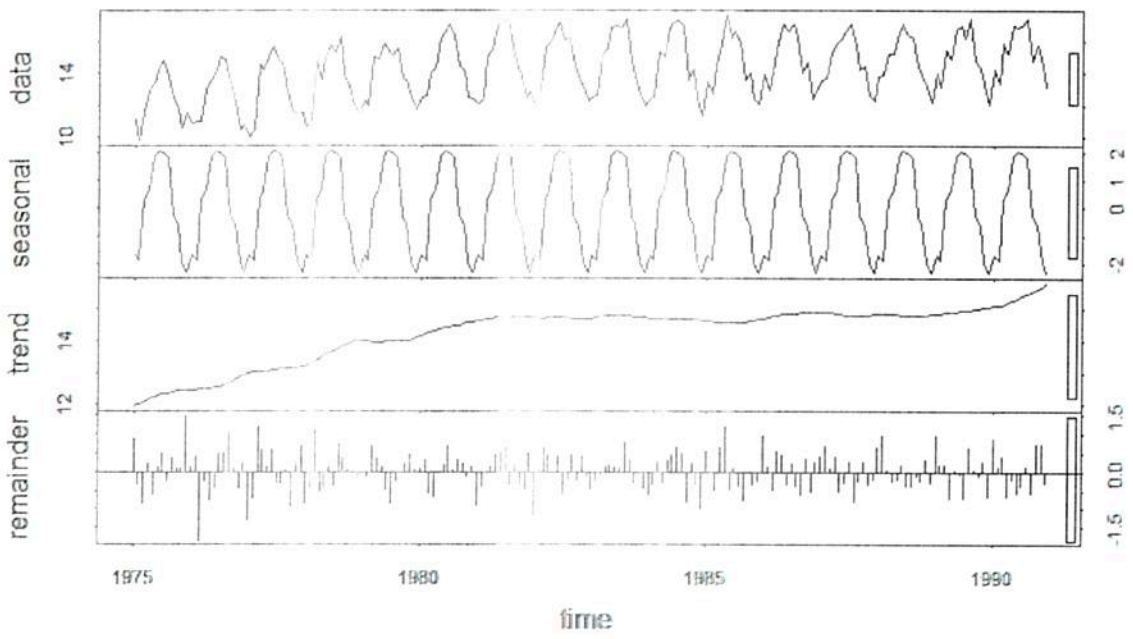
Number of Outliers Detected: 0

	i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
1	0	746.8933	314.9075	1680	59	2.963114	3.570191	FALSE
2	1	741.6215	307.8232	1547	35	2.616367	3.568466	FALSE
3	2	737.0455	302.6038	1515	57	2.570869	3.566728	FALSE
4	3	732.6000	297.6528	1510	11	2.611768	3.564980	FALSE

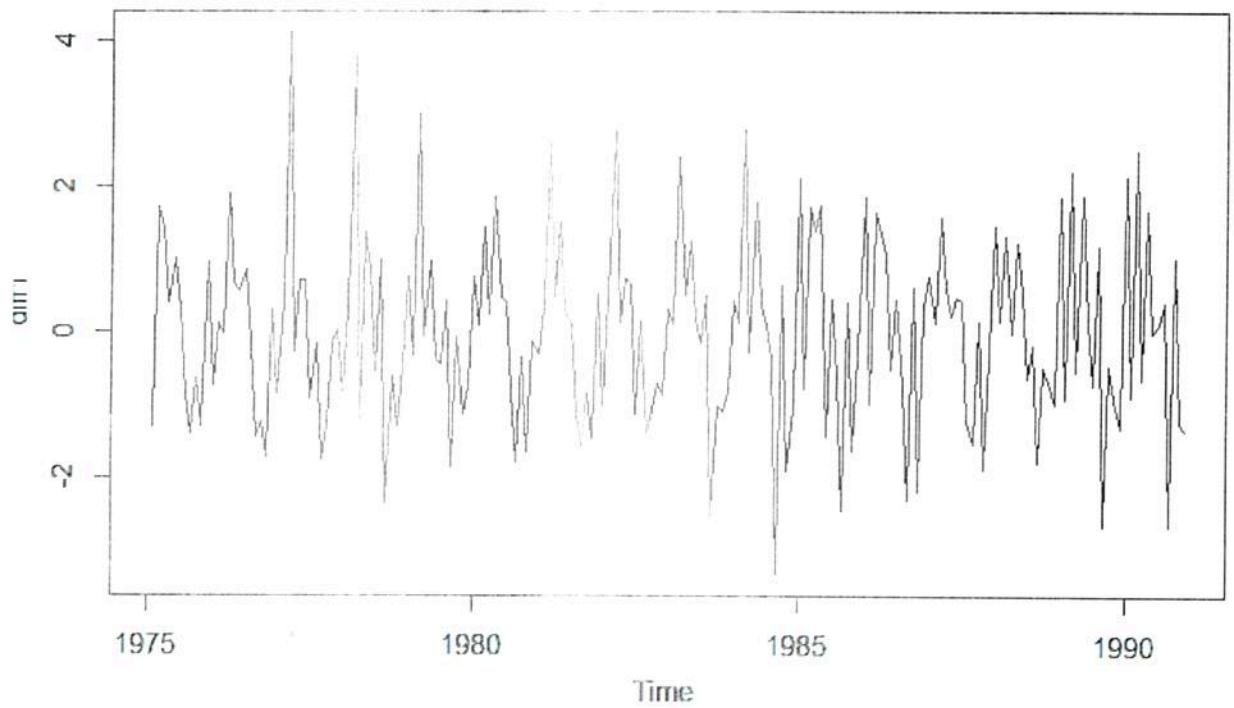
Beer sales

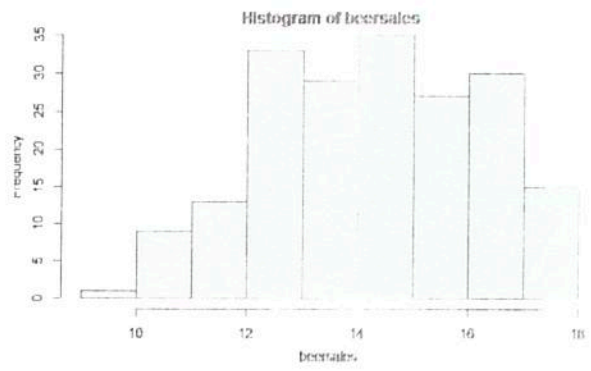

beersales
10
12
14
16
1975
1980
1985
1990
Time

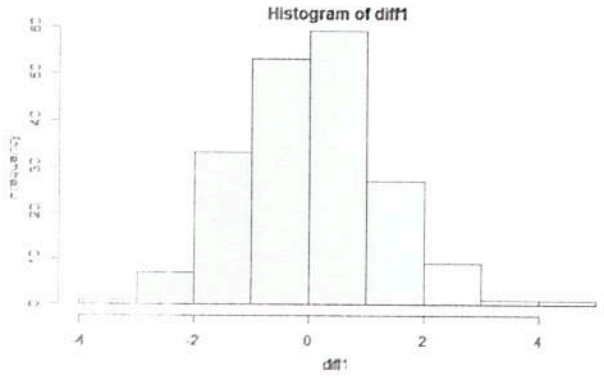

Decomposition of additive time series
observed
14
trend
14.0
seasonal
1
0
-2
random
1.0
-0.5
-2.0
1975
1980
1985
1990
Time

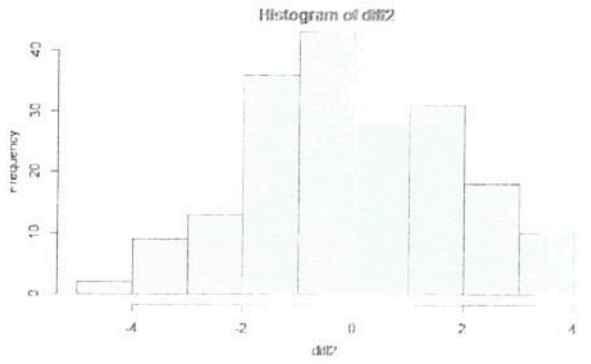


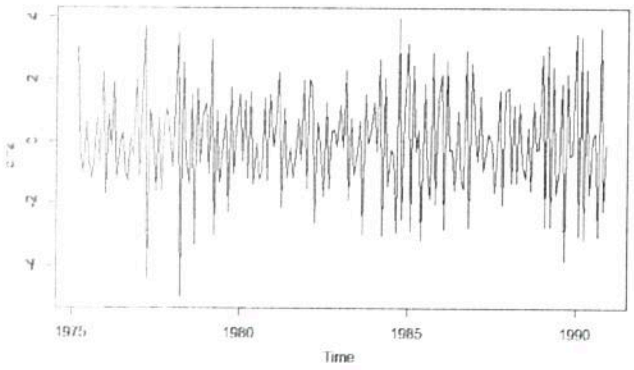
First difference

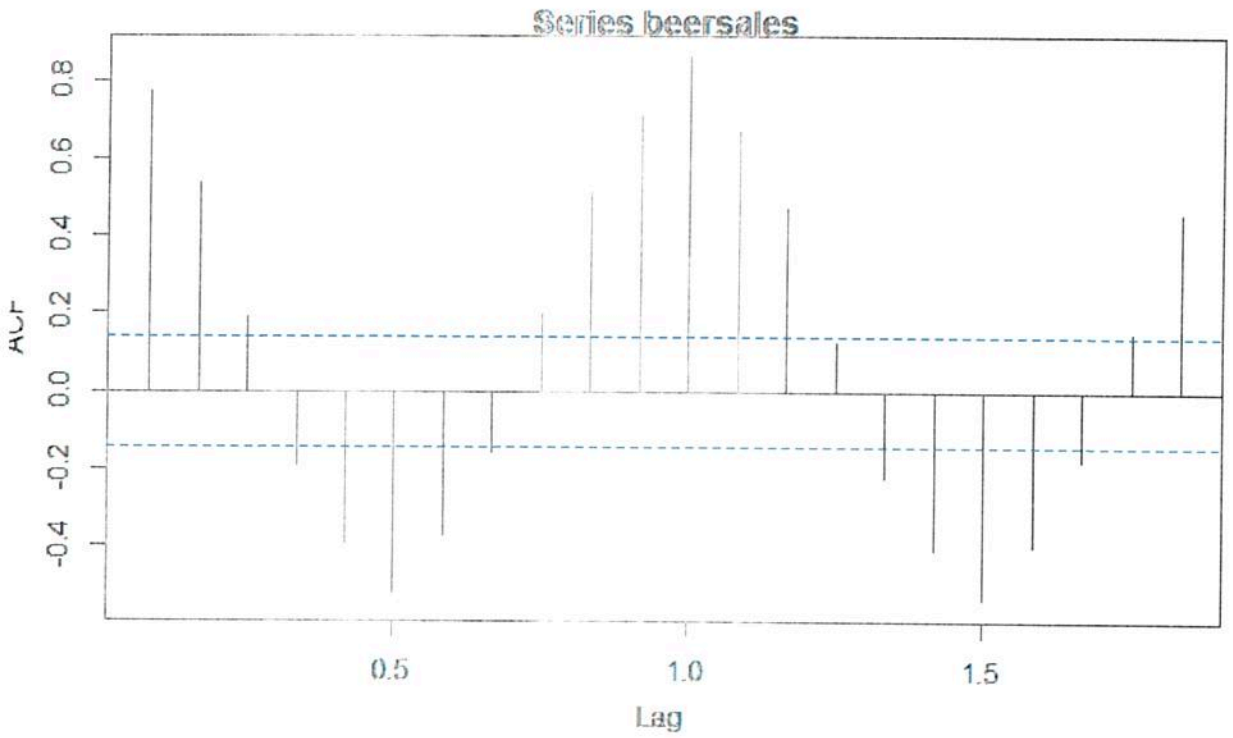


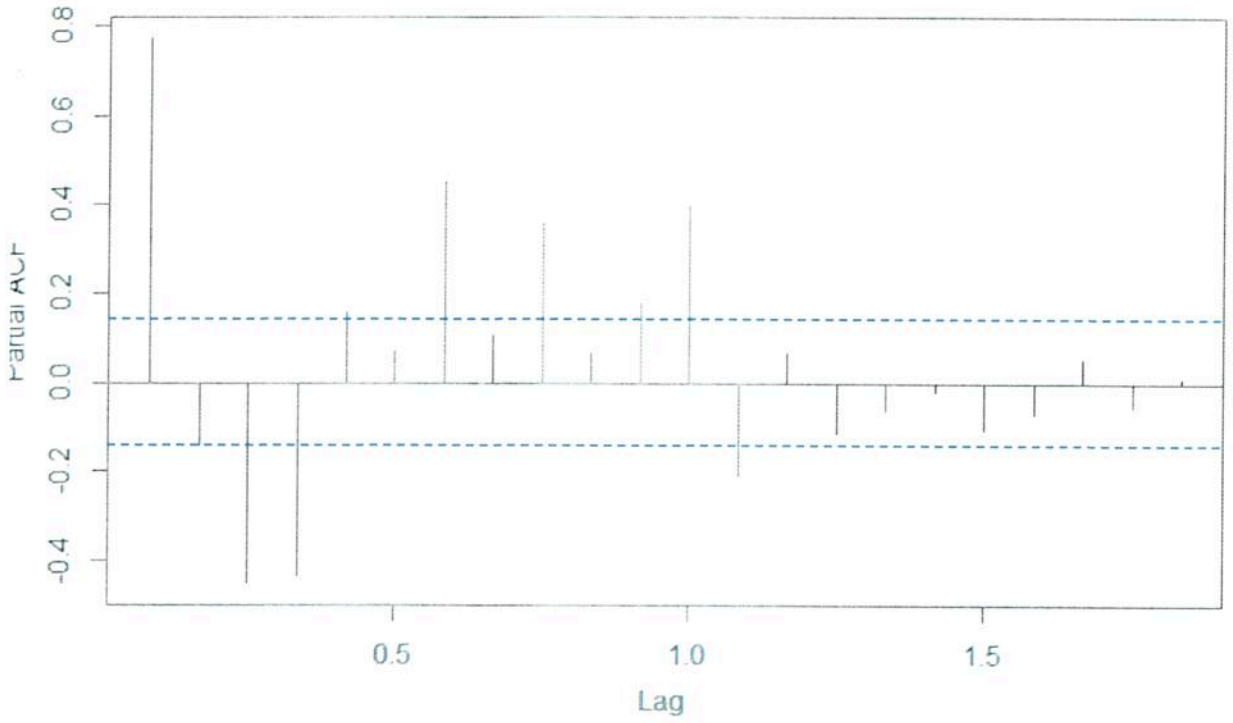

Histogram of beersales
Frequency
beersales


Histogram of diff1
diff1


Histogram of diff2
Frequency
diff2


1975
1980
1985
1990
Time

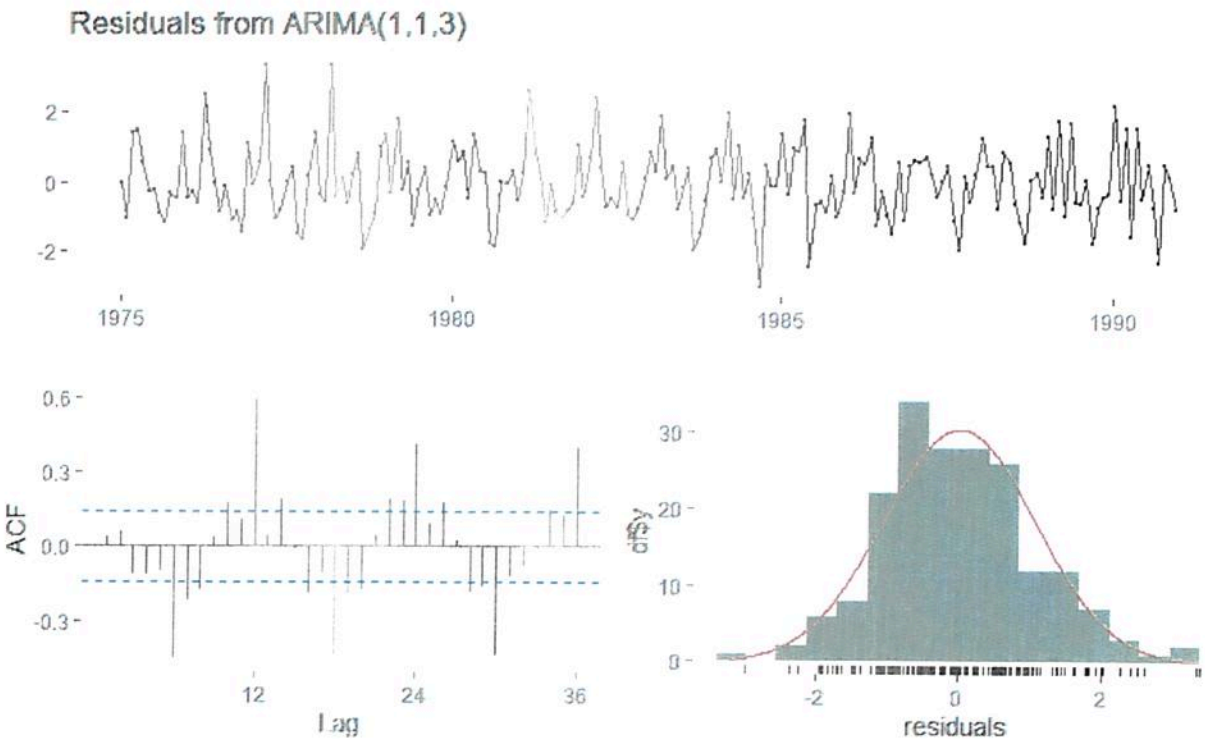

Series beersales
ACF
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
0.5
1.0
1.5
Lag

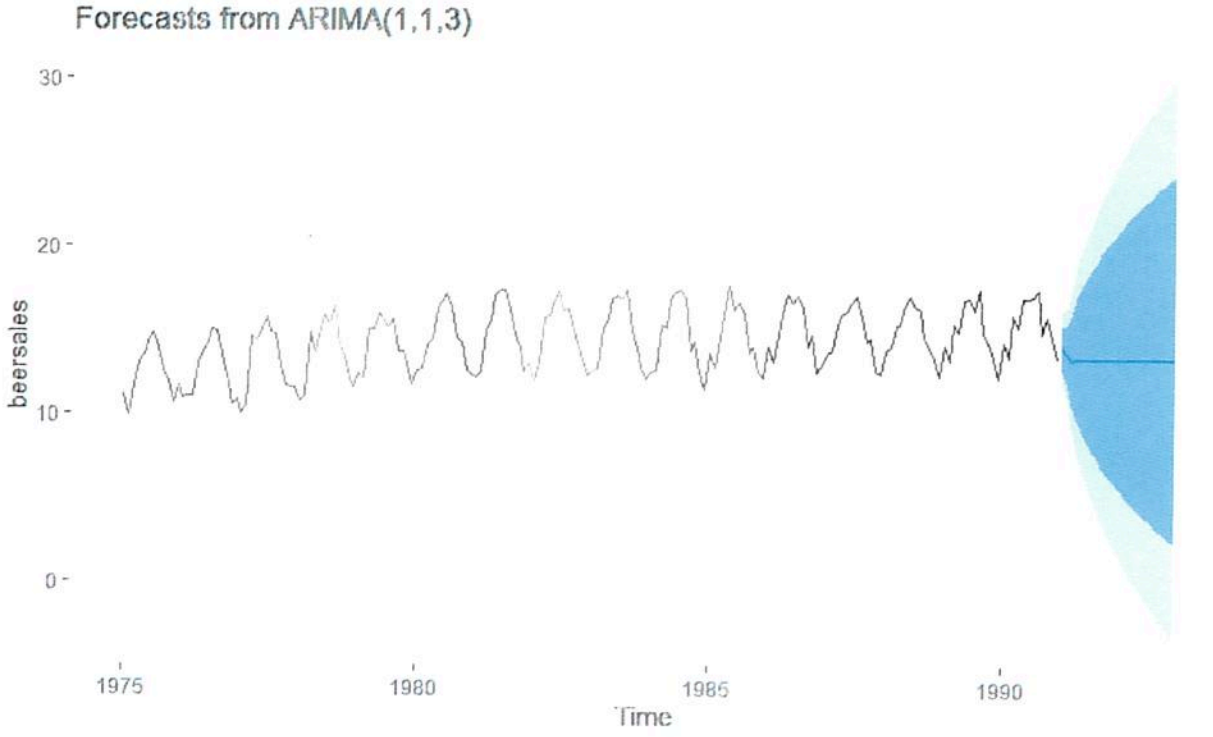


Starting with (1,1,3)

```
Call:
arima(x = beersales, order = c(1, 1, 3))

Coefficients:
          ar1     ma1     ma2     ma3
      -0.4099  0.3516  0.3724  0.6398
s.e.   0.1150  0.0913  0.0543  0.0601

sigma^2 estimated as 1.125:  log likelihood = -283.09,  aic = 574.18
```

After some experimenting, this seems to be the lowest AIC

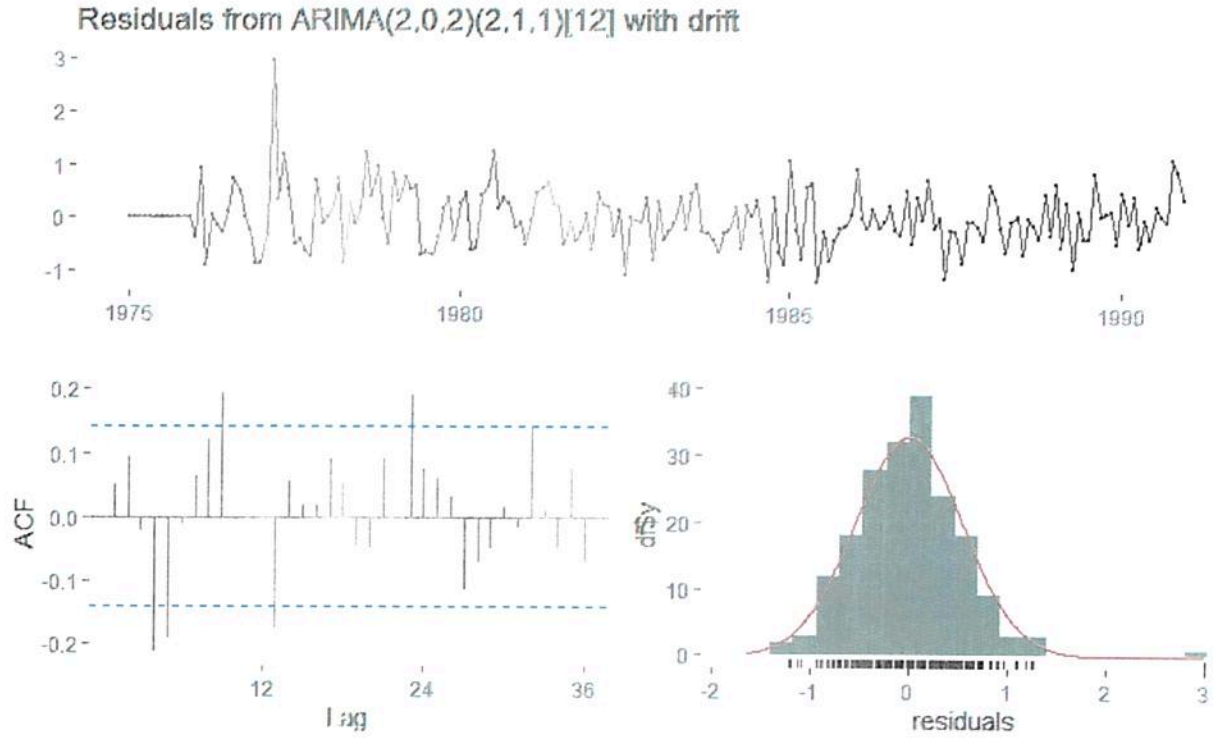


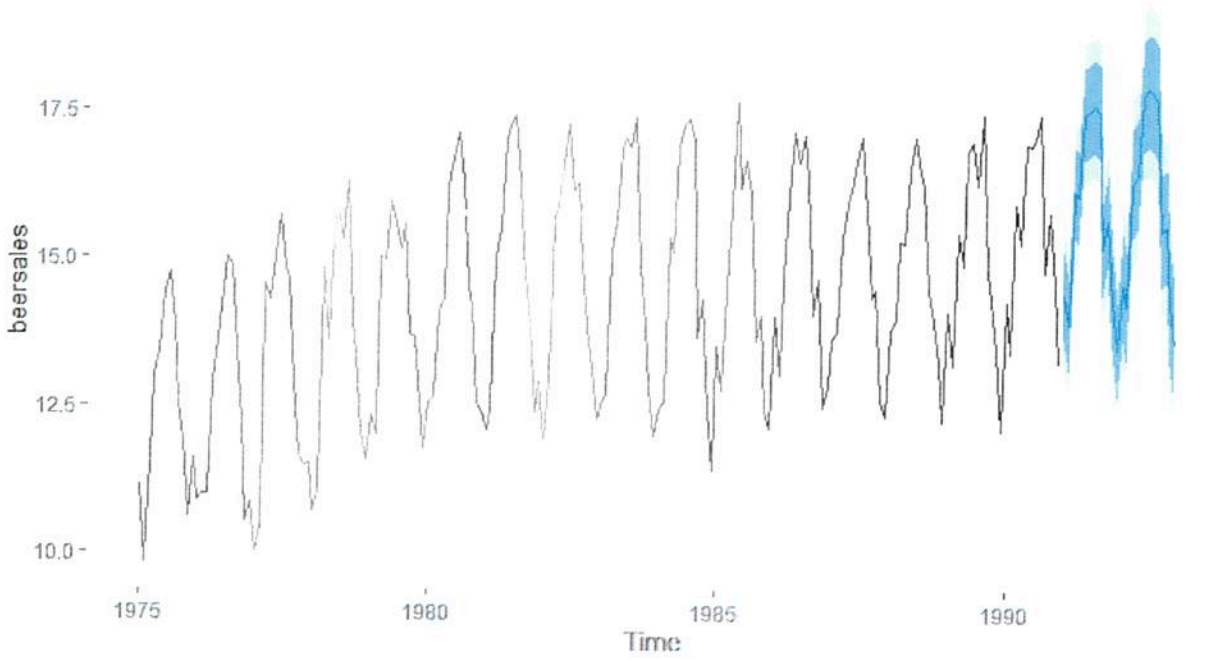


Auto.arima produced the following:

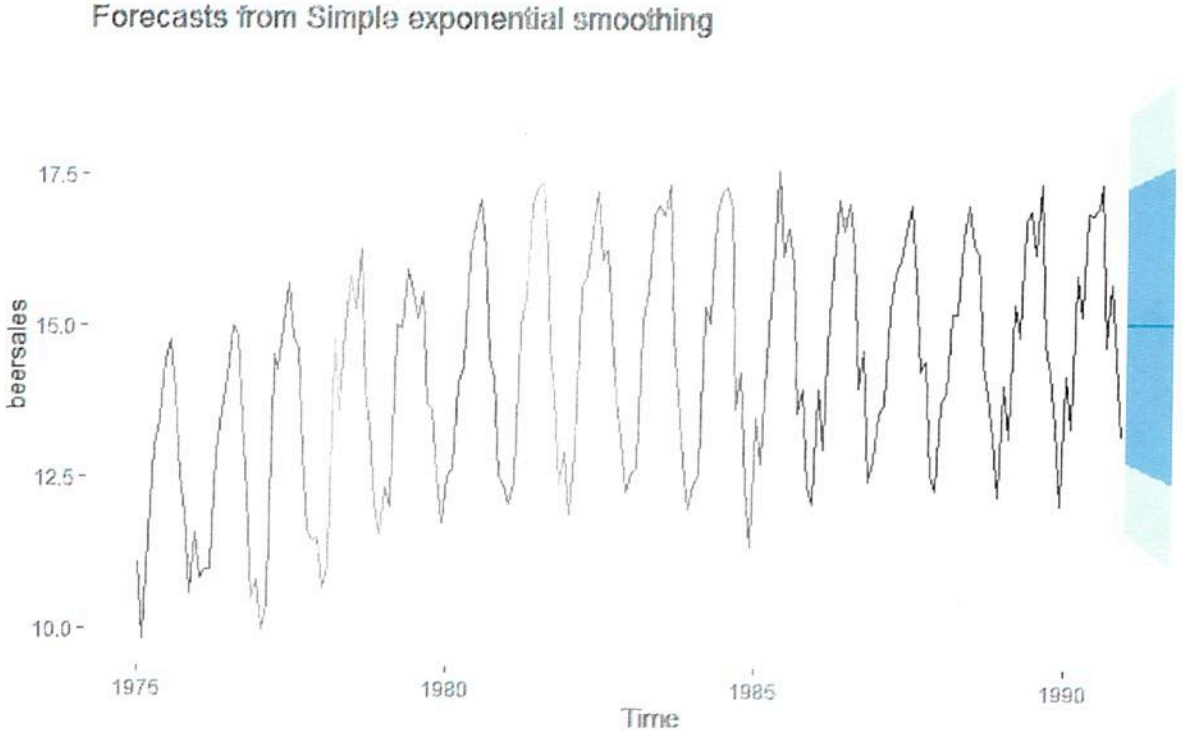
```
Series: beersales
ARIMA(2,0,2)(2,1,1)[12] with drift

Coefficients:
         ar1      ar2      ma1      ma2     sar1     sar2     sma1   drift
```

```
      0.4079  0.5658  -0.1300  -0.6782  0.3164  -0.1506  -0.7970  0.0174
s.e.  0.2188  0.2138   0.1788   0.1481  0.1319   0.1141   0.1296  0.0063

sigma^2 = 0.3291:  log likelihood = -155.78
AIC=329.57   AICc=330.63   BIC=358.3
```





Exponential smoothing model forecast



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	-0.007044044	1.345244	1.042633	101.86619	160.53603	1.698749	-0.02380294	NA
Test set	-0.064963856	1.406884	1.133598	91.38105	95.06294	1.846959	-0.38045342	0.9049192