7CCSMDM1 Data Mining Coursework 1

1.1 -

(i) number of instances - 48,842
(ii) number of missing values - 6,465
(iii) fraction of missing values over all attribute values - 6,465/633,108 = 1.02%
(iv) number of instances with missing values - 3,620
(v) fraction of instances with missing values over all instances - 3,620/48,842 = 7.41%

1.4 -

The results from when the missing results are simply deleted tend to have an increased error rate of around 3-5% with the data used in the coursework when compared to the other methods used. The method of replacing every Nan value for just the string ‘missing’ and the method of taking the mode for each attribute value and replacing the Nan’s with that value produce very close error rates of around 16% when tested multiple times. The two methods are so close that depending on the run, one can slightly edge the other with a margin of 0.1%. This shows that both methods are almost equally as effective as one another when being used in decision trees.

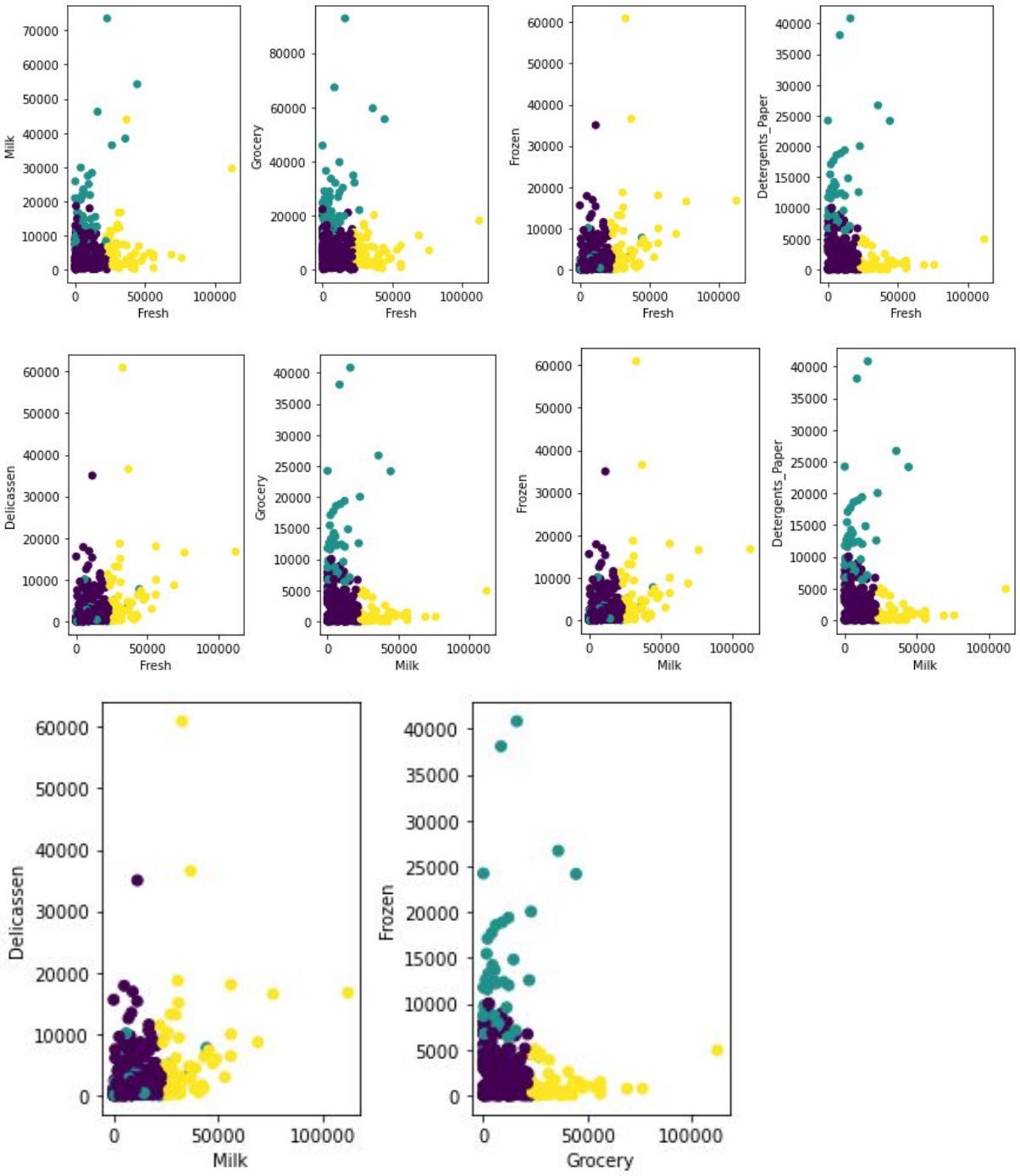
2.1 -

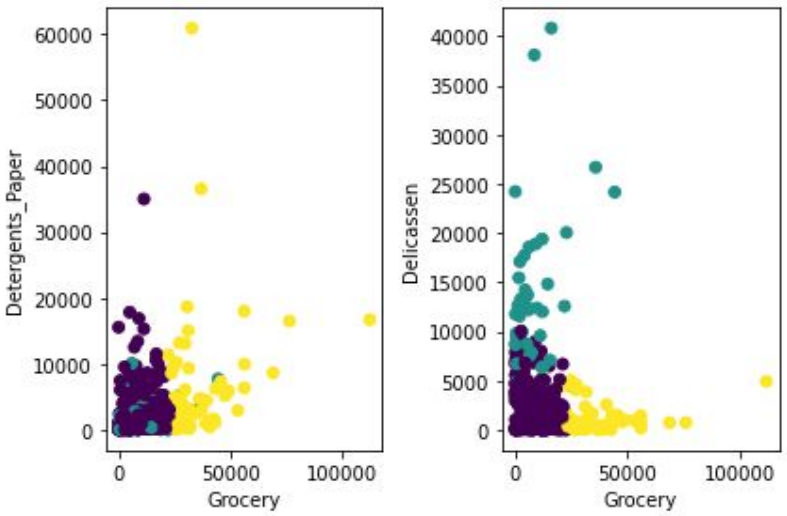
	Fresh	Milk	Grocery	Frozen	Detergents Paper	Delicatessen
Mean	12000.297727272728	5796.265909090909	7951.277272727273	3071.931818181818	2881.4931818181817	1524.8704545454545
Min	3	55	3	25	3	3
Max	112151	73498	92780	60869	40827	47943

2.1 -

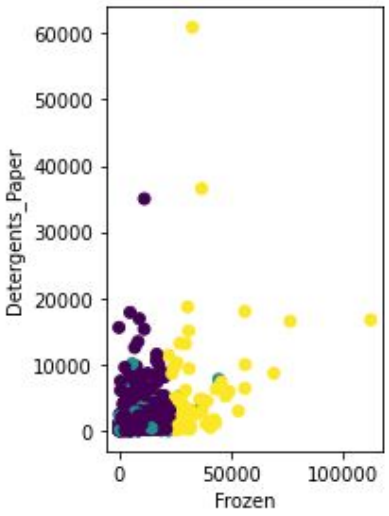
Below is all the scatter plots for all the different cluster pairs for the attributes in the wholesale customers data set.

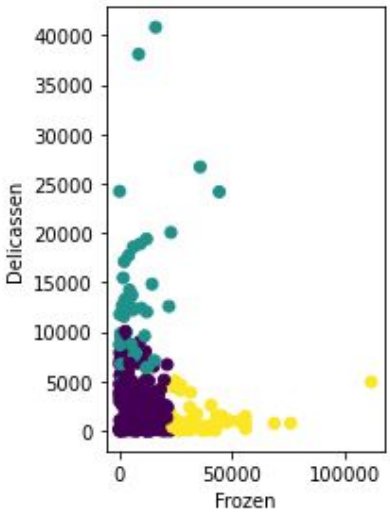
For the fresh attribute, there's a correlation all attributes except the following pairs:
Fresh/Frozen
Fresh/Delicatessen
Milk/Frozen
Milk/Delicatessen
Grocery/Detergents
Frozen/Detergents
Detergents/Delicatessen

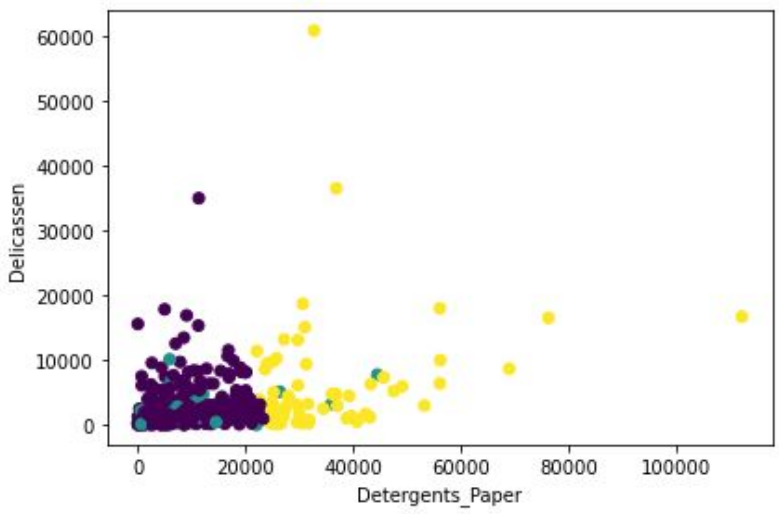

Milk
Fresh
Grocery
Fresh
Frozen
Fresh
Detergents_Paper
Fresh
Delicassen
Fresh
Grocery
Milk
Frozen
Milk
Detergents_Paper
Milk
Delicassen
Milk
Frozen
Grocery


Detergents_Paper
Grocery
0
10000
20000
30000
40000
50000
60000
0
50000
100000

Delicassen
Grocery
0
5000
10000
15000
20000
25000
30000
35000
40000
0
50000
100000


Detergents_Paper
Frozen
0
10000
20000
30000
40000
50000
60000
0
50000
100000


Delicassen
Frozen
0
5000
10000
15000
20000
25000
30000
35000
40000
0
50000
100000


Delicassen
Detergents_Paper
0
10000
20000
30000
40000
50000
60000
0
20000
40000
60000
80000
100000

2.3 -

As the clusters increased the within cluster scores decrease and the inverse happens with the between cluster scores, this would mean logically that, the clusters are more intense and the items are closer together within each cluster, and as the clusters increase, the distance between the clusters increase so the clusters are further apart. The ratio also increases as the number of clusters increases.

Below is the table of scores from the difference clusters.

	K = 3	K = 5	K = 10
BC	45239982498.877205	29454313223.728428	13756106907.279348
WC	1924917595.3908718	12710533365.635357	81065549205.12784
BC/WC	0.042549034925878	0.4315338561482987	5.893058970211275