7CCSMDM1 Data Mining

1 Classification

1.
(i) number of instances: 48842
(ii) number of missing values: 6465
(iii) fraction of missing values over all attributes: 0.0095
(iv) number of instances with missing values: 3620
(v) fraction of instances with missing values: 0.0741

2.
Attribute: age
Possible values: {0, 1, 2, 3, 4}
Attribute: workclass
Possible values: {0, 1, 2, 3, 4, 5, 6}
Attribute: education
Possible values: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}
Attribute: education-num
Possible values: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}
Attribute: marital-status
Possible values: {0, 1, 2, 3, 4, 5, 6}
Attribute: occupation
Possible values: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13}
Attribute: relationship
Possible values: {0, 1, 2, 3, 4, 5}
Attribute: race
Possible values: {0, 1, 2, 3, 4}
Attribute: sex
Possible values: {0, 1}
Attribute: capitalgain
Possible values: {0, 1, 2, 3, 4}
Attribute: capitalloss
Possible values: {0, 1, 2, 3, 4}
Attribute: hoursperweek
Possible values: {0, 1, 2, 3, 4}
Attribute: native-country
Possible values: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40}

3. Error rate: 0.1782

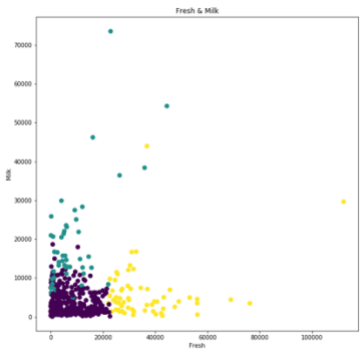
4.
Error rate of D1': 0.1727
Error rate of D2': 0.1664

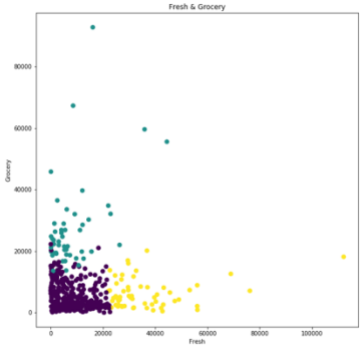
2 Clustering

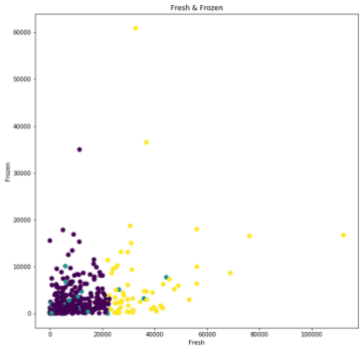
1. Mean and Range of each attribute

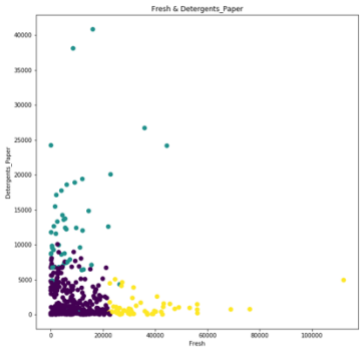
	Mean	Range
Fresh	12000.297727	[3, 112151]
Milk	5796.265909	[55, 73498]
Grocery	7951.277273	[3, 92780]
Frozen	3071.931818	[25, 60869]
Detergents_Paper	2881.493182	[3, 40827]
Delicassen	1524.870455	[3, 47943]

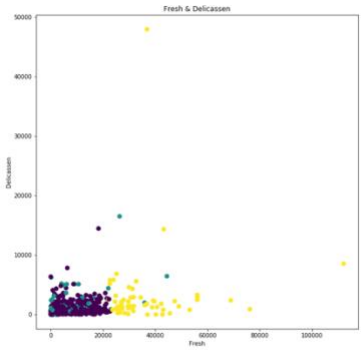
2. Pictures below demonstrate 15 scatter plots for each two criteria plots when kmeans = 3

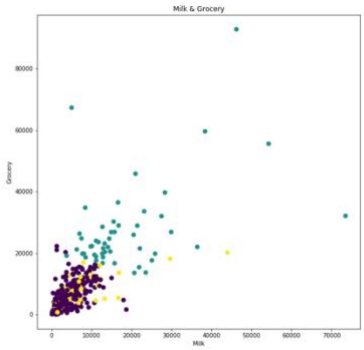


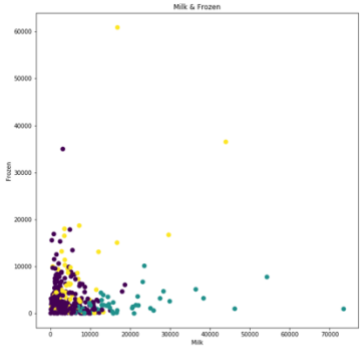


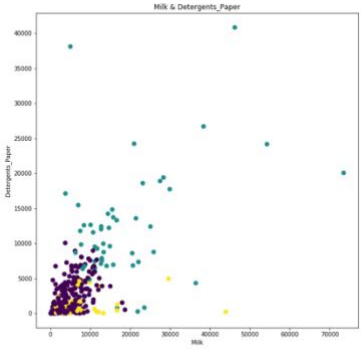


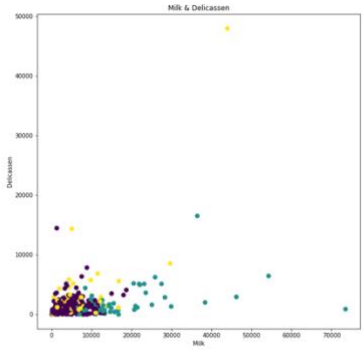

Fresh & Detergents_Paper
Detergents_Paper
Fresh

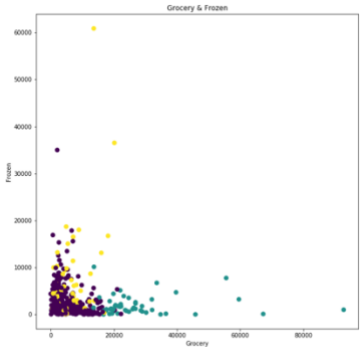

Fresh & Delicassen
Delicassen
Fresh

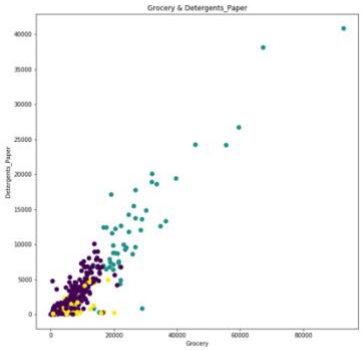

Milk & Grocery
Grocery
Milk

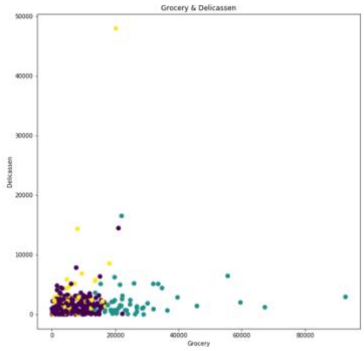

Milk & Frozen
Frozen
Milk

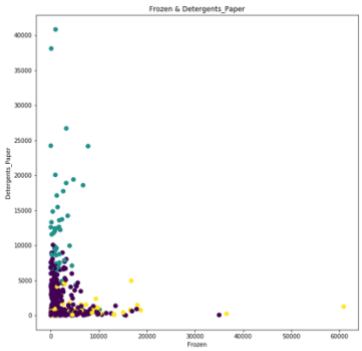

Milk & Detergents_Paper
Detergents_Paper
Milk

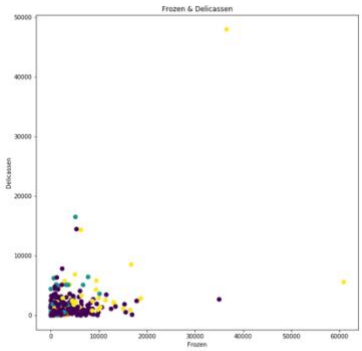

Milk & Delicassen
Delicassen
Milk

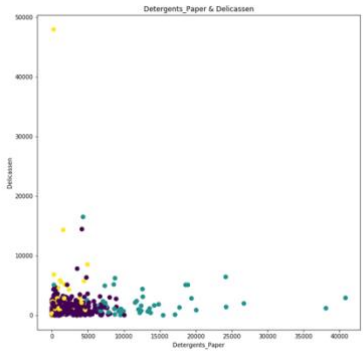

Grocery & Frozen
Frozen
Grocery


Grocery & Detergents_Paper
Detergents_Paper
Grocery


Grocery & Delicassen
Delicassen
Grocery


Frozen & Detergents_Paper
Detergents_Paper
Frozen


Frozen & Delicassen
Delicassen
Frozen


Detergents_Paper & Delicassen
Delicassen
Detergents_Paper

3. When k=[3,5,10], the following table with the between cluster distance BC, within cluster distance WC and ratio BC/WC of the set of clusters obtained for each k

	k=3	k=5	k=10
BC	1924917595	1206949957	1217404494
WC	4242007486	1839364265	3208260872
BC/WC	0.00453775	0.06561777	0.03794593