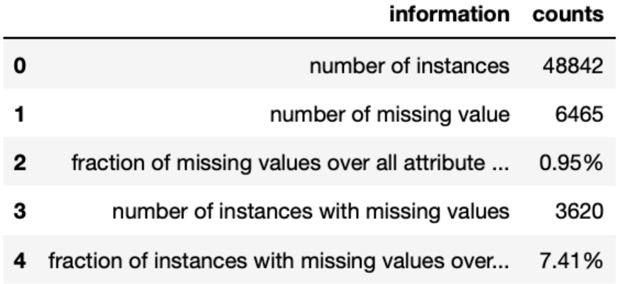
1920253
Hsin-Ju, Chan

7CCSMDM1 Data Mining

Coursework 1

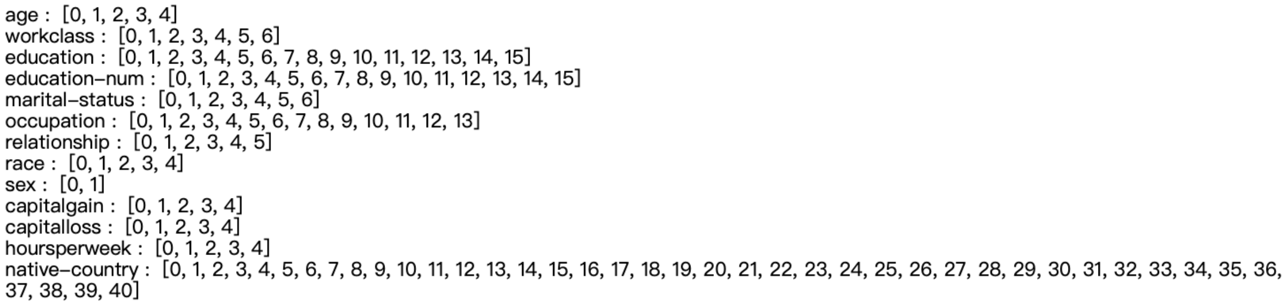
Classification

1. The table shows the counts of number of instances, number of missing values, fraction of missing values over all attribute values, number of instances with missing values and fraction of instances with missing values over all instances respectively.



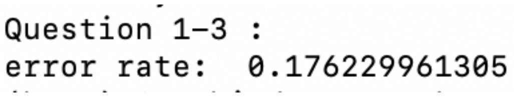
	information	counts
0	number of instances	48842
1	number of missing value	6465
2	fraction of missing values over all attribute ...	0.95%
3	number of instances with missing values	3620
4	fraction of instances with missing values over...	7.41%

2. The following shows the set of all possible discrete values for each attribute.



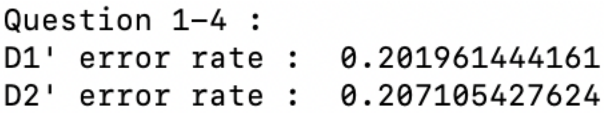
```
age :  [0, 1, 2, 3, 4]
workclass :  [0, 1, 2, 3, 4, 5, 6]
education :  [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
education-num :  [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
marital-status :  [0, 1, 2, 3, 4, 5, 6]
occupation :  [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]
relationship :  [0, 1, 2, 3, 4, 5]
race :  [0, 1, 2, 3, 4]
sex :  [0, 1]
capitalgain :  [0, 1, 2, 3, 4]
capitalloss :  [0, 1, 2, 3, 4]
hoursperweek :  [0, 1, 2, 3, 4]
native-country :  [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. The following show the error rate of the decision tree which has ignore the missing value.
   - The error rate is 17.623%



```
Question 1-3 :
error rate:  0.176229961305
```

4. The following picture shows the error rate of D1’ and D2’.
   - The error rate of D1’ is 20.1961%
   - The error rate of D2’ is 20.7105%

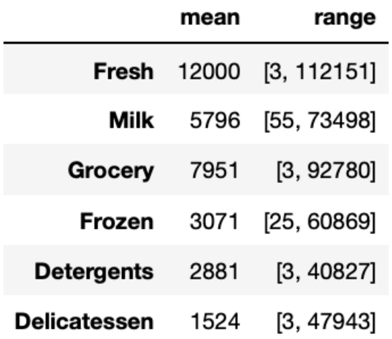


```
Question 1-4 :
D1' error rate :  0.201961444161
D2' error rate :  0.207105427624
```

1920253
Hsin-Ju, Chan

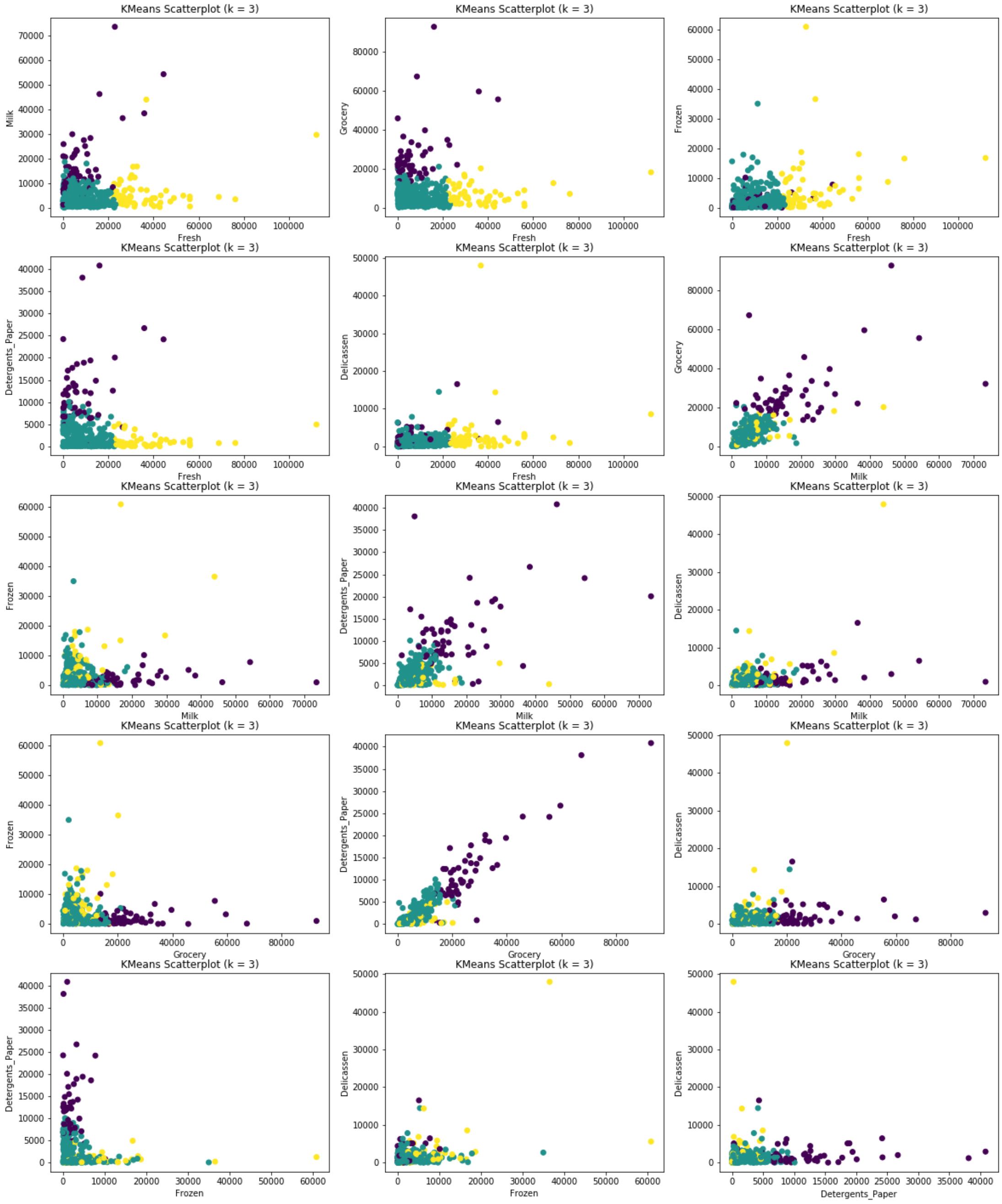
Clustering

1. This table shows the mean and range [$x_{j,min}$, $x_{j,max}$] of the data from wholesale_customers.csv.

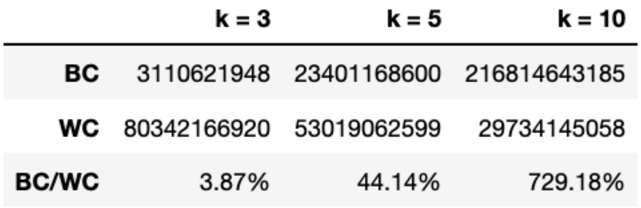


	mean	range
Fresh	12000	[3, 112151]
Milk	5796	[55, 73498]
Grocery	7951	[3, 92780]
Frozen	3071	[25, 60869]
Detergents	2881	[3, 40827]
Delicatessen	1524	[3, 47943]

2. The picture below is the 15 scatter plots for each pair of attributes using k-means (k=3).



3. The table below illustrates the between cluster distance BC, within cluster distance WC and ratio BC/WC of k = 3, 5, 10. As the table shows, when k is big, the between cluster distance is large and the within cluster distance is small; when k is small, the between cluster distance is small and the within cluster distance is large.



	k = 3	k = 5	k = 10
BC	3110621948	23401168600	216814643185
WC	80342166920	53019062599	29734145058
BC/WC	3.87%	44.14%	729.18%