7CCSMDM1 Data Mining

Coursework 1

Department of Informatics Semester 2 King's College London

1 Classification

D' = from the original data set D, containing (i) all instances with at least one missing value and (ii) an equal number of randomly selected instances without missing values. That is, if the number of instances with missing values is v in D, then D' should contain these v instances and additional v instances without any missing values, which are randomly selected from D.

D' 1 and D' 2 = using D' , constructed two modified data sets D' 1 and D' 2 to handle missing values.

Q1 :

(i) Number of instances = 48842

(ii) number of missing values =6465

(iii) fraction of missing values over all attribute values =(6465/683788)*100 = .95

(iv) number of instances with missing values = 3620

(v) fraction of instances with missing values over all instances.

Q2:

Convert all 13 attributes into nominal using a Scikit-learn LabelEncode

Age ['2', '3', '1', '0', '4']

Workclass ['State-gov', 'Self-emp-not-inc', 'Private', 'Federal-gov', 'Local-gov', 'nan', 'Self-emp-inc', 'Without-pay', 'Never-worked']

Education ['Bachelors', 'HS-grad', '11th', 'Masters', '9th', 'Some-college', 'Assoc-acdm', 'Assoc-voc', '7th-8th', 'Doctorate', 'Prof-school', '5th-6th', '10th', '1st-4th', 'Preschool', '12th']

Education-num ['13', '9', '7', '14', '5', '10', '12', '11', '4', '16', '15', '3', '6', '2', '1', '8']

Marital-status ['Never-married', 'Married-civ-spouse', 'Divorced', 'Married-spouse-absent', 'Separated', 'Married-AF-spouse', 'Widowed']

Occupation ['Adm-clerical', 'Exec-managerial', 'Handlers-cleaners', 'Prof-specialty', 'Other-service', 'Sales', 'Craft-repair', 'Transport-moving', 'Farming-fishing', 'Machine-op-inspct', 'Tech-support', 'nan', 'Protective-serv', 'Armed-Forces', 'Priv-house-serv']

Relationship ['Not-in-family', 'Husband', 'Wife', 'Own-child', 'Unmarried', 'Other-relative']

Race ['White', 'Black', 'Asian-Pac-Islander', 'Amer-Indian-Eskimo', 'Other']

Sex ['Male', 'Female']

Capitalgain ['1', '0', '4', '2', '3']

Capitalloss ['0', '3', '1', '2', '4']

Hoursperweek ['2', '0', '3', '4', '1']

Q3

From D testing

	precision	recall	f1-score	support
0	0.87	0.90	0.89	11100
1	0.66	0.59	0.62	3553
accuracy			0.83	14653
macro avg	0.76	0.74	0.75	14653
weighted avg	0.82	0.83	0.82	14653

```
print(confusion_matrix(y_test,predictions))
```

```
[[10018  1082]
 [ 1474  2079]]
```

Question 4

From D1 testing

```
predictions = dtree.predict(X_test)
from sklearn.metrics import classification_report,confusion_matrix
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.89	0.88	0.89	1766
1	0.51	0.54	0.53	406
accuracy			0.82	2172
macro avg	0.70	0.71	0.71	2172
weighted avg	0.82	0.82	0.82	2172

```
print(confusion_matrix(y_test,predictions))
```

```
[[1554  212]
 [ 185  221]]
```

Question 4

From D2 testing

```
predictions = dtree.predict(X_test)
from sklearn.metrics import classification_report,confusion_matrix
print(classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

           0       0.90      0.89      0.89      1766
           1       0.53      0.56      0.55       406

    accuracy                           0.83      2172
   macro avg       0.72      0.72      0.72      2172
weighted avg       0.83      0.83      0.83      2172
```

```
print(confusion_matrix(y_test,predictions))
```

```
[[1569  197]
 [ 180  226]]
```

Best model: Based on the performances above, I will choose d2 as a better model. Since the accuracy in d2 is 0.82 and d1 the accuracy is 0.83. In-addition,in d2 there is an improvement in the precision , recall and F1-score compared to d1.

Clustering 2

Question 1

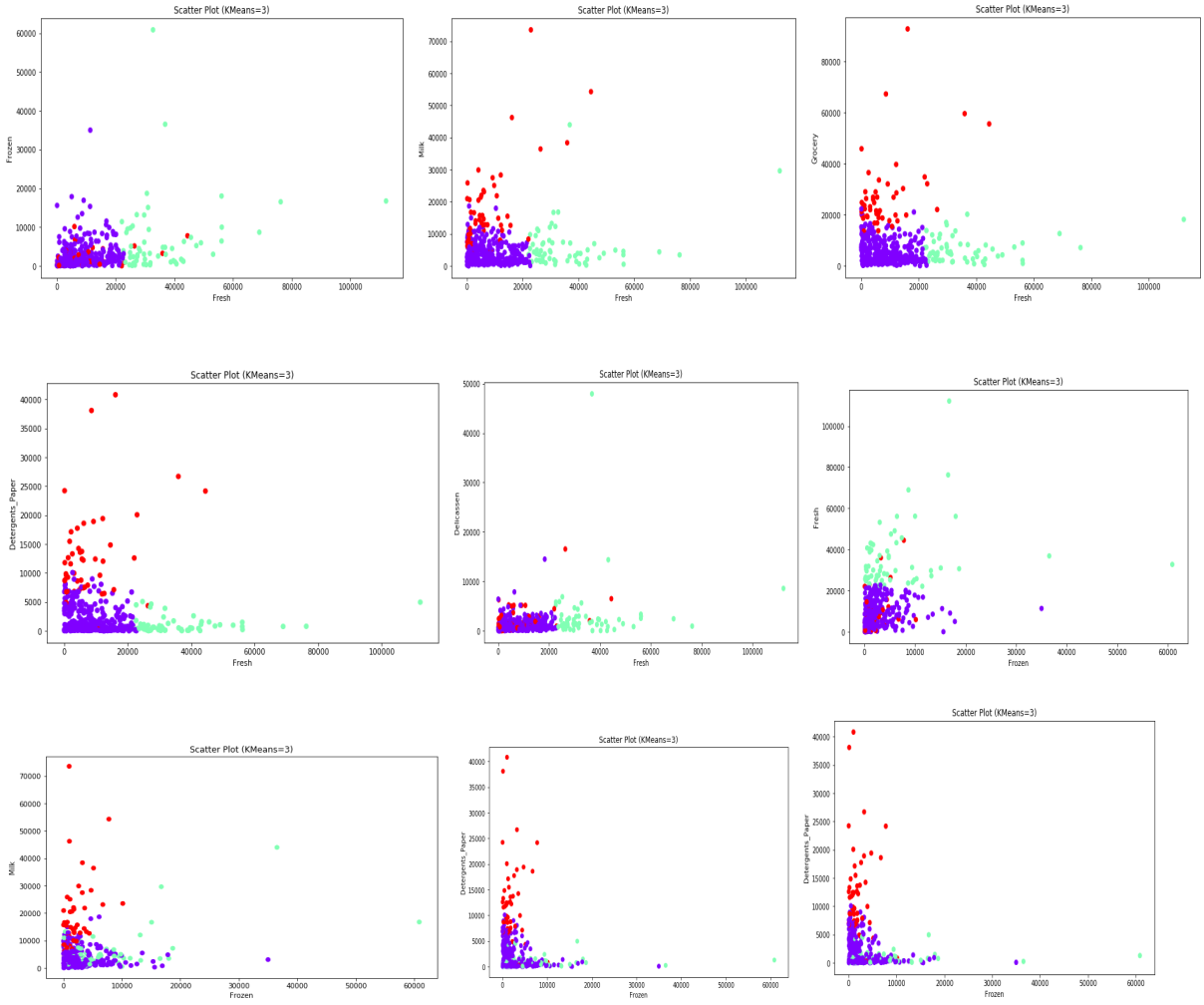
```
ad_data.describe().loc[['mean','min','max'],:]
```

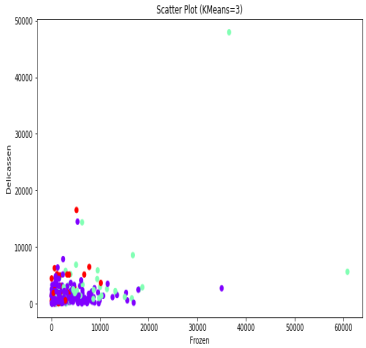
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

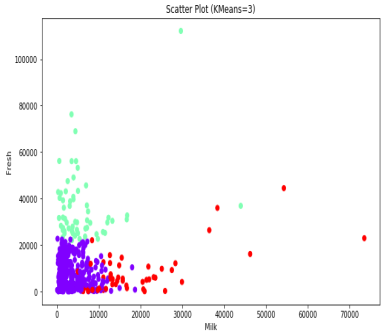
Scatter plot

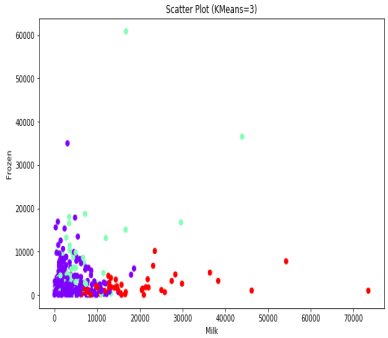
Question 2 ：

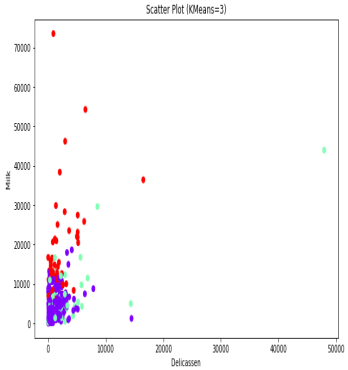
Looking at the scatter plot below the scatter plot between fresh vs milk is a better fit compared to other scatters plots below.

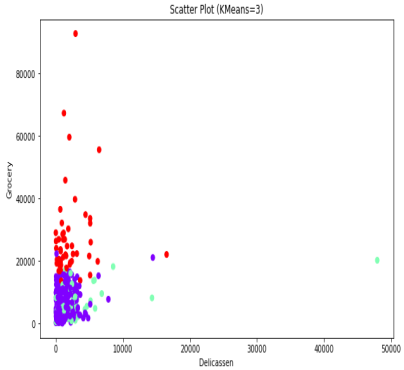


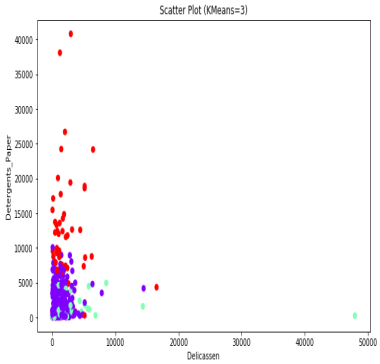

Scatter Plot (KMeans=3)
Delicassen
Frozen


Scatter Plot (KMeans=3)
Fresh
Milk


Scatter Plot (KMeans=3)
Frozen
Milk


Scatter Plot (KMeans=3)
Milk
Delicassen


Scatter Plot (KMeans=3)
Grocery
Delicassen


Scatter Plot (KMeans=3)
Detergents_Paper
Delicassen