DATA MINING COURSEWORK 1

BY

MEGHANA KUMAR

20093575

Regression

Question 1.1

Metric	Value
Number of Instances	48842
Number of Missing Values	6465
Fraction of Missing Values/All Attribute Values	0.010182
Number of Instances with Missing Values	3620
Fraction of Instances with Missing Values/All Instances	0.074117

Question 1.2

- All null values were padded with the string ‘NaN’.
- So it is seen as a different label called ‘NaN when the discrete values are printed.

Attribute	Values
'age'	array([0, 1, 2, 3, 4])
'workclass'	array(['Federal-gov', 'Local-gov', 'NaN', 'Never-worked', 'Private', 'Self-emp-inc', 'Self-emp-not-inc', 'State-gov', 'Without-pay'])
'education'	array(['10th', '11th', '12th', '1st-4th', '5th-6th', '7th-8th', '9th', 'Assoc-acdm', 'Assoc-voc', 'Bachelors', 'Doctorate', 'HS-grad', 'Masters', 'Preschool', 'Prof-school', 'Some-college'])
'education-num'	array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16])
'marital-status'	array(['Divorced', 'Married-AF-spouse', 'Married-civ-spouse',

	'Married-spouse-absent', 'Never-married', 'Separated', 'Widowed']
'occupation'	array(['Adm-clerical', 'Armed-Forces', 'Craft-repair', 'Exec-managerial', 'Farming-fishing', 'Handlers-cleaners', 'Machine-op-inspct', 'NaN', 'Other-service', 'Priv-house-serv', 'Prof-specialty', 'Protective-serv', 'Sales', 'Tech-support', 'Transport-moving']
'relationship'	array(['Husband', 'Not-in-family', 'Other-relative', 'Own-child', 'Unmarried', 'Wife']
'race'	array(['Amer-Indian-Eskimo', 'Asian-Pac-Islander', 'Black', 'Other', 'White']
'sex'	array(['Female', 'Male']
'capitalgain'	array([0, 1, 2, 3, 4])
'capitalloss'	array([0, 1, 2, 3, 4])
'hoursperweek'	array([0, 1, 2, 3, 4])
‘native-country'	array(['Cambodia', 'Canada', 'China', 'Columbia', 'Cuba', 'Dominican-Republic', 'Ecuador', 'El-Salvador', 'England', 'France', 'Germany', 'Greece', 'Guatemala', 'Haiti', 'Holand-Netherlands', 'Honduras', 'Hong', 'Hungary', 'India', 'Iran', 'Ireland', 'Italy', 'Jamaica', 'Japan', 'Laos', 'Mexico', 'NaN', 'Nicaragua', 'Outlying-US(Guam-USVI-etc)', 'Peru', 'Philippines', 'Poland', 'Portugal', 'Puerto-Rico', 'Scotland', 'South', 'Taiwan', 'Thailand', 'Trinadad&Tobago', 'United-States', 'Vietnam', 'Yugoslavia']

Question 1.3

Random State = 0 in all cases

- The evaluation is done for three different values of split: 10%, 25%, 33.33%
- A standard split and validation is done and the values are reported.
- An additional 10-Fold Cross Validation is also done and the error rate is reported.
- Error rate is the fraction in instances in the validation set that are misclassified.

With test split = 10%,

Evaluation Metric	Score
Training Error Rate	0.081230
Test Error Rate	0.175326
10 Fold Cross Validation Error Rate	0.176839

With test split = 25%,

Evaluation Metric	Score
Training Error Rate	0.079078
Test Error Rate	0.181054
10 Fold Cross Validation Error Rate	0.176706

With test split = 33.33%,

Evaluation Metric	Score
Training Error Rate	0.077449
Test Error Rate	0.176806
10 Fold Cross Validation Error Rate	0.177414

Question 1.4

Random State = 0 in all cases

D1’ = Missing values filled with the class label “missing”

D2’ = Missing values filled with the mode, i.e most popular value for each column

With test split = 10%
Number of instances in D1’/D2’ = 7240
Number of instances in Test picked from D - D1’/D2’ = int(7240/9) = 804

Evaluation Metric	Score
D1’ Training Error Rate	0.034684
D1’ Test Error Rate	**0.164365**
D1’ 10 Fold Cross Validation Error Rate	**0.168232**
D2’ Training Error Rate	0.041436
D2’ Test Error Rate	**0.185083**
D2’ 10 Fold Cross Validation Error Rate	**0.162293**

With test split = 25%
Number of instances in D1’/D2’ = 7240
Number of instances in Test picked from D - D1’/D2’ = int(7240/9) = 2413

Evaluation Metric	Score
D1’ Training Error Rate	0.032965
D1’ Test Error Rate	**0.177901**
D1’ 10 Fold Cross Validation Error Rate	**0.170718**
D2’ Training Error Rate	0.039963
D2’ Test Error Rate	**0.180110**
D2’ 10 Fold Cross Validation Error Rate	**0.170994**

With test split = 33.33%
Number of instances in D1’/D2’ = 7240
Number of instances in Test picked from D - D1’/D2’ = int(7240/2) ~ 3619

Evaluation Metric	Score
D1’ Training Error Rate	0.030460

D1’ Test Error Rate	**0.172328**
D1’ 10 Fold Cross Validation Error Rate	**0.171271**
D2’ Training Error Rate	0.037505
D2’ Test Error Rate	**0.181027**
D2’ 10 Fold Cross Validation Error Rate	**0.168785**

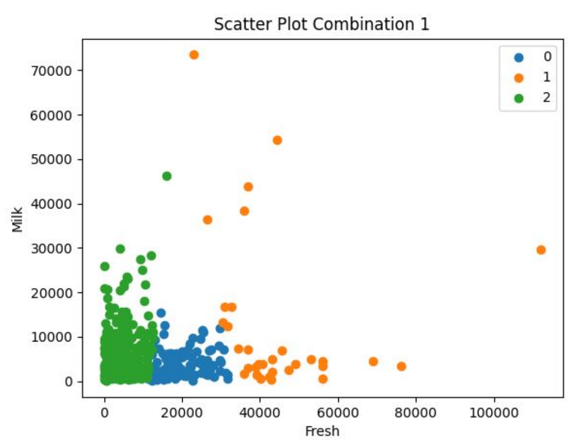
- All instances used in training are exclusively from D1’ and D2’. The test set is chosen separately from the remaining instances in D.
- The model is tested with three different validation splits and usually 0.25 to 0.33 is considered a reasonable split.
- In these two splits, for **dummy label approach D1’,** the training error is less but test error is more. Might indicate overfitting. For **popular value approach D2’,** the training error is more but test error is lesser, so it might indicate that the model D2’ generalizes better and might be a better approach to combat overfitting.
- The **popular value approach D2’** is likely to work better since it might be more representative of what the value could have been. It is a fair assumption to make that the missing value might have taken the most occurring popular value.
- An ever better approach would be to only consider the popular value if it is present for more than some x fraction of instances. (eg. at least 30% of instances). If the most popular value is only popular *slightly more* than other values, then it may not be representative of data.
- A **weighted approach of probabilities** of occurrence of a value to predict the missing value would be a more educated guess. Using random sampling with assigned weights of probability might give the best results as it exactly represents the data present.

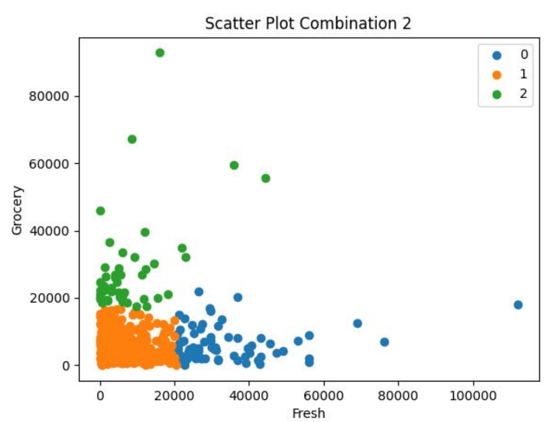
Clustering

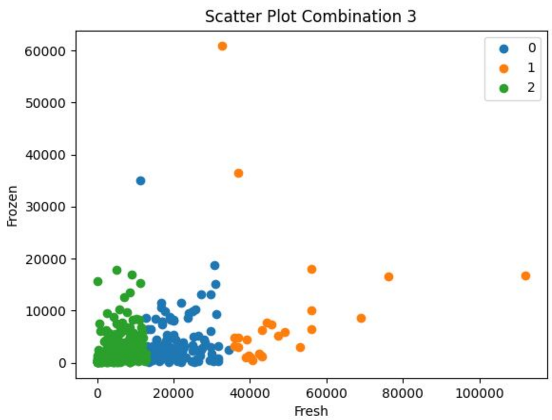
Question 2.1

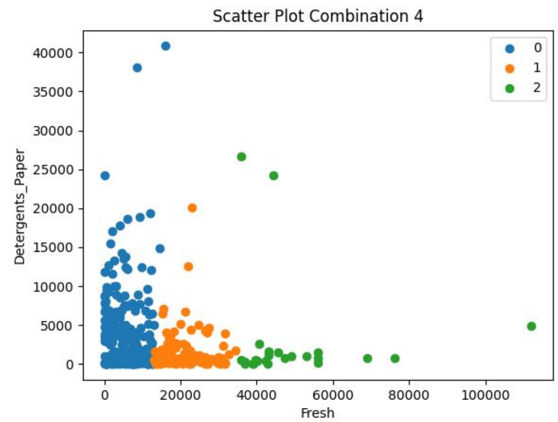
Attribute	Mean μ_j	Range $[x_{j,min}, x_{j,max}]$
FRESH	20725.890824	[3.0, 112151.0]
MILK	12439.986635	[55.0, 73498.0]
GROCERY	16030.211263	[3.0, 92780.0]
FROZEN	9385.388144	[25.0, 60869.0]
DETERGENTS_PAPER	6739.324704	[3.0, 40827.0]
DELICATESSEN	6990.622049	[3.0, 47943.0]

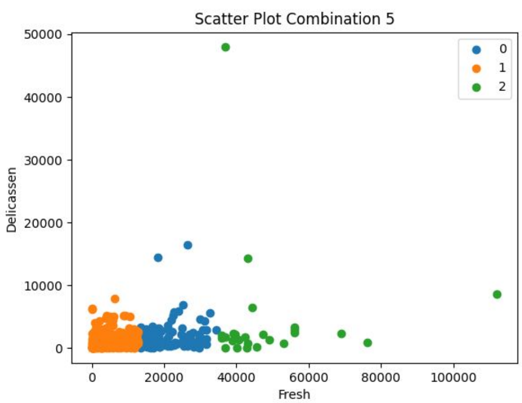
Question 2.2

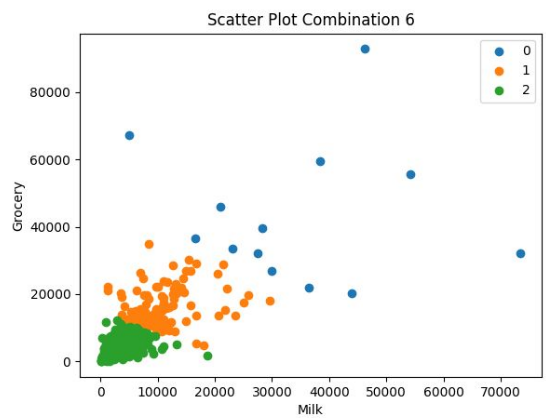


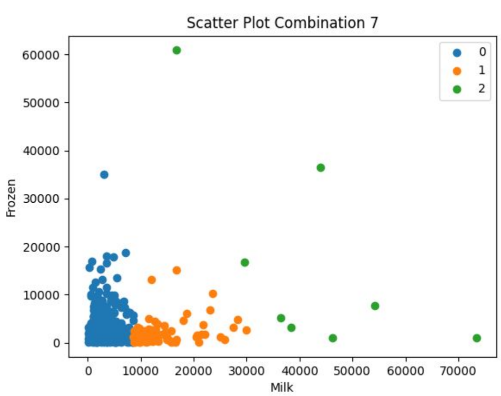


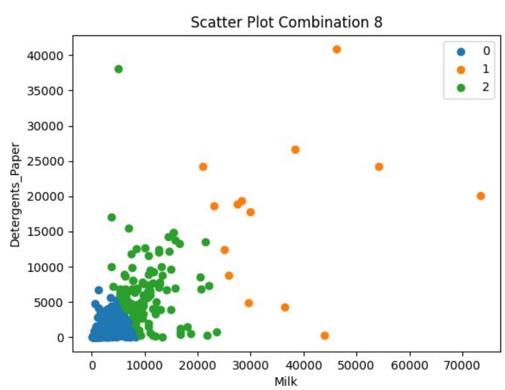

Scatter Plot Combination 3
Frozen
Fresh
0
1
2

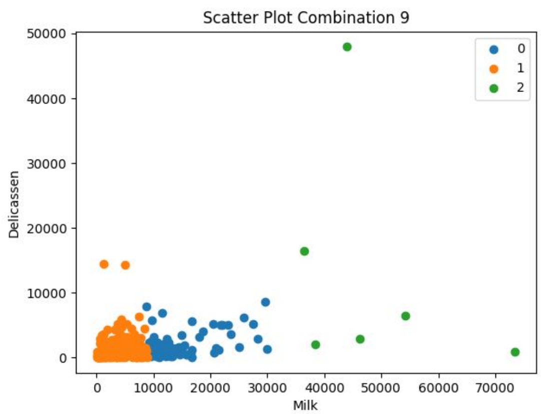

Scatter Plot Combination 4
Detergents_Paper
Fresh
0
1
2

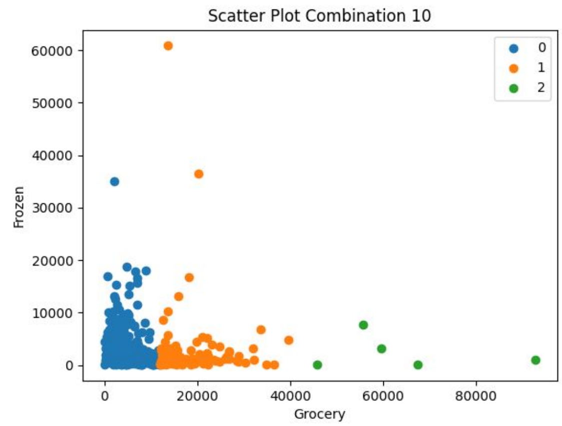

Scatter Plot Combination 5
Delicassen
Fresh
0
1
2

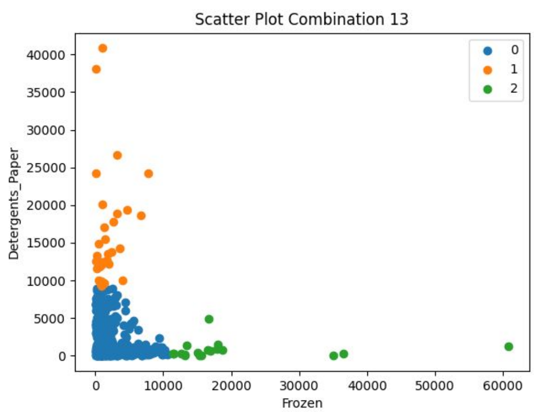

Scatter Plot Combination 6
Grocery
Milk
0
1
2

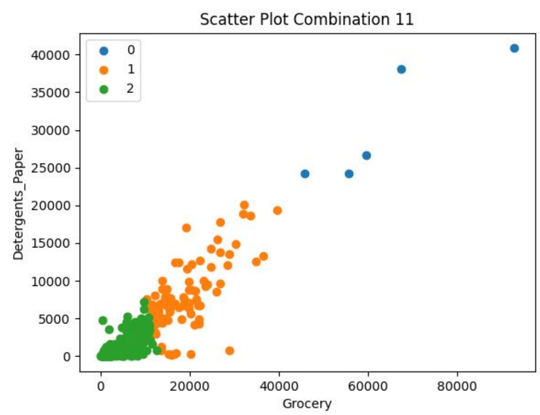

Scatter Plot Combination 7
Frozen
Milk
0
1
2

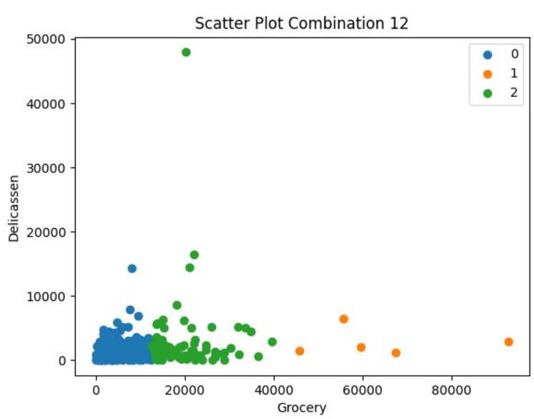

Scatter Plot Combination 8
Detergents_Paper
Milk
0
1
2

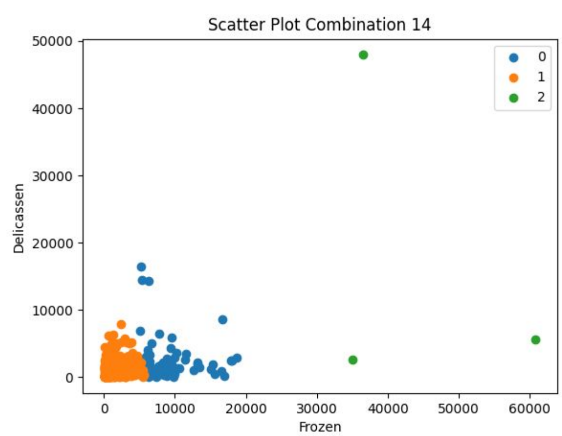

Scatter Plot Combination 9
Delicassen
Milk
0
1
2

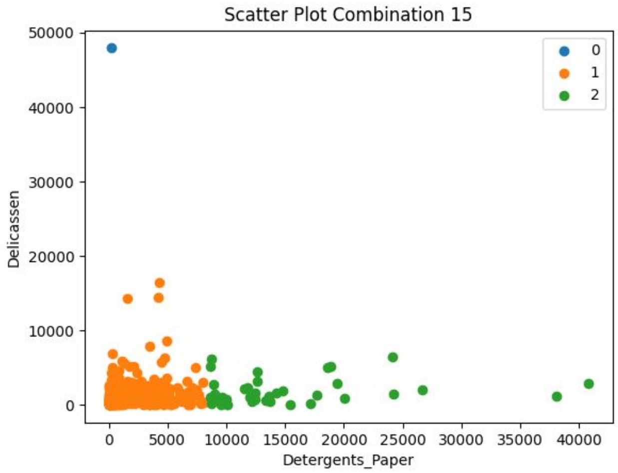

Scatter Plot Combination 10
Frozen
Grocery
0
1
2


Scatter Plot Combination 13
Detergents_Paper
Frozen
0
1
2


Scatter Plot Combination 11
Detergents_Paper
Grocery
0
1
2


Scatter Plot Combination 12
Delicassen
Grocery
0
1
2


Scatter Plot Combination 14
Delicassen
Frozen
0
1
2



Observations:

- Clusters are well separated for a lot of metrics, and can be clearly visually represented.
- Outliers are sometimes taken as a separate cluster for some attribute pairs Eg: Detergents_Paper vs Delicatessen (Plot 15, Plot 9, Plot 14) are other examples
- For some attribute pairs, a different K value might be better. A lower k value would make fewer clusters, and avoid classifying outliers as a different cluster and approximate them to an existing cluster.
- Optimal number of clusters can be found using Elbow Method or Silhoutte method
- The overall clustering space is done in a 6 dimensional space, and the 15 plots act as projections in 2D for every pair.

Question 2.3

	k = 3	**k = 5**	**k = 10**
BC	3162657727.586802	25621025526.678169	177230279225.94540
WC	80332413843.01633	52928148942.576141	29673646783.305515
BC/WC	0.039370	0.484072	5.972649

The WC (Within Cluster distance) can be calculated by obtaining the inertia.
A Custom Function is written to obtain the BC (Between Cluster Distance)

The BC/WC value denotes a metric of how good a k-means algorithm, such that it separates the data well. By the basic principle of the k-means algorithm, it separates the clusters based on decreasing within cluster distance to the centroid, by assigning to the nearest centroid, and increasing the distance between centroids of clusters to ensure they are well separated. Hence the BC/WC value is to be maximized to get a good clustering algorithm score.

But these values need to be normalized to get a proper inference from them, now the values are not converging, and are constantly increasing with the value of K.

We can normalize them and plot Calinski-Harabasz or the Silhouette score to determine the optimal clusters number. The below evaluation is made by using the Distortion Score: **It determines that the optimum value of k = 5.** This could not be solely determined by BC/WC scores since it needed to be normalized. In the BC/WC scores, the scores start becoming greater than 1 after k=5 and do not converge.