Data Mining Coursework 1

Rakan Zabian - 1741706 - MSc Data Science

18/02/2021

1 Classification

1.1 Question 1:

Number of instances	48842
Number of missing values	6465
Fraction of missing values over all attribute values	0.01
Number of instances with missing values	3620
Fraction of instances with missing values over all instances	0.07

1.2 Question 2:

age: [2 3 1 0 4]
workclass: [6 5 3 0 1 8 4 7 2]
education: [9 11 1 12 6 15 7 8 5 10 14 4 0 3 13 2]
education-num: [4 15 13 5 11 1 3 2 10 7 6 9 12 8 0 14]
marital-status: [4 2 0 3 5 1 6]
occupation: [0 3 5 9 7 11 2 13 4 6 12 14 10 1 8]
relationship: [1 0 5 3 4 2]
race: [4 2 1 0 3]
sex: [1 0]
capital-gain: [1 0 4 2 3]
capital-loss: [0 3 1 2 4]
hours-per-week: [2 0 3 4 1]
native-country: [38 4 22 18 41 25 34 32 15 8 1 10 19 29 21 30 3 0 36 6 24 35 13 31 5 7 9 12 2 23 40 28 27 33 37 11 26 39 16 20 17 14]

1.3 Question 3:

Error rate: 0.18

1.4 Question 4:

Error rate Dp1: 0.23502100375856727
Error rate Dp2: 0.2151227061684723

As expected, both error rates are consistently greater than that of Q3, because the data used in Q3 is complete since we only used instances without NaN values.
The error rates vary slightly with every program run since the training data varies due to the random selection of instances. Therefore, we cannot make a conclusion on which approach ('missing' or mode)

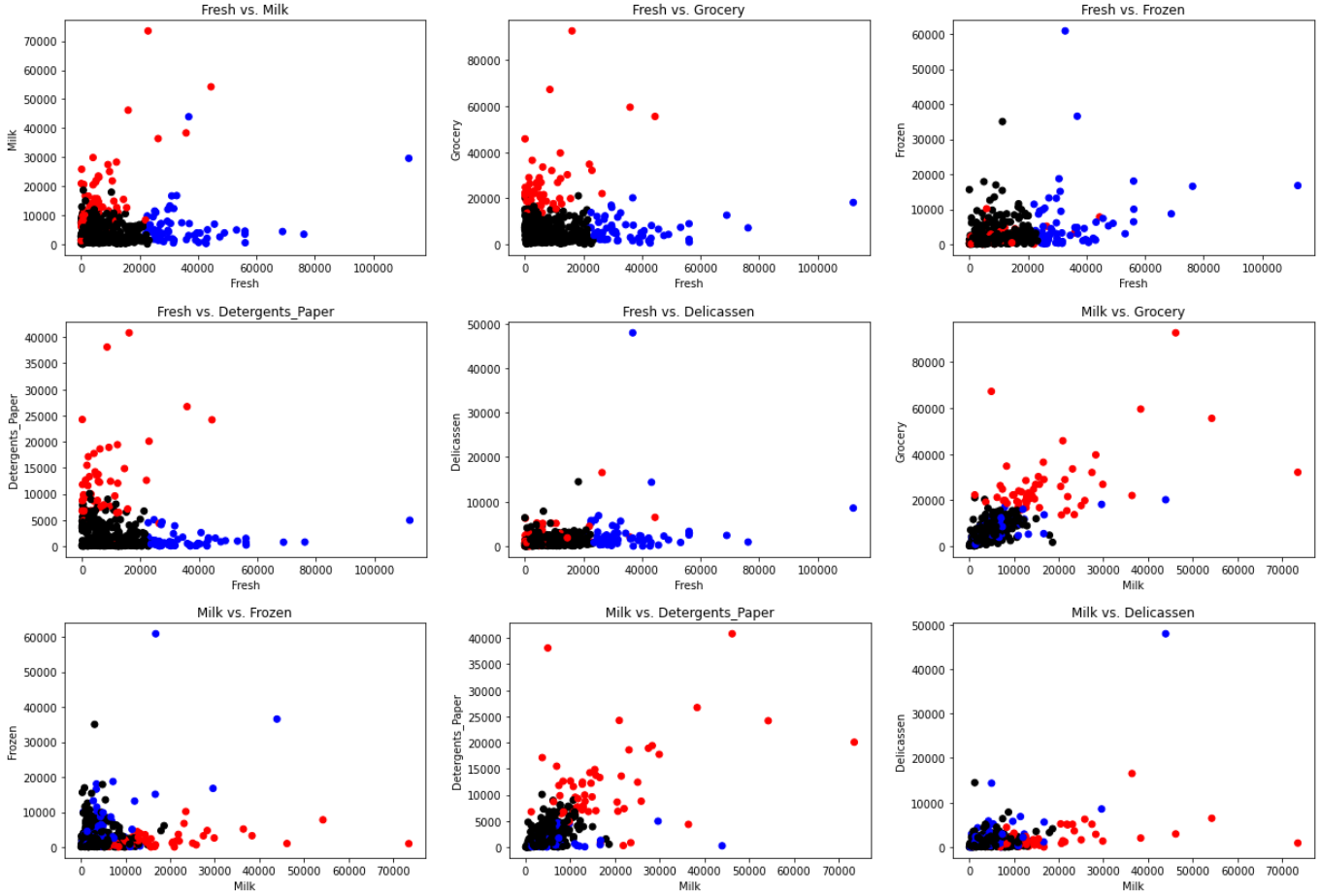
is constantly better. However, they are both almost always within a range less than 0.03, meaning that both approaches are suitable in this case.

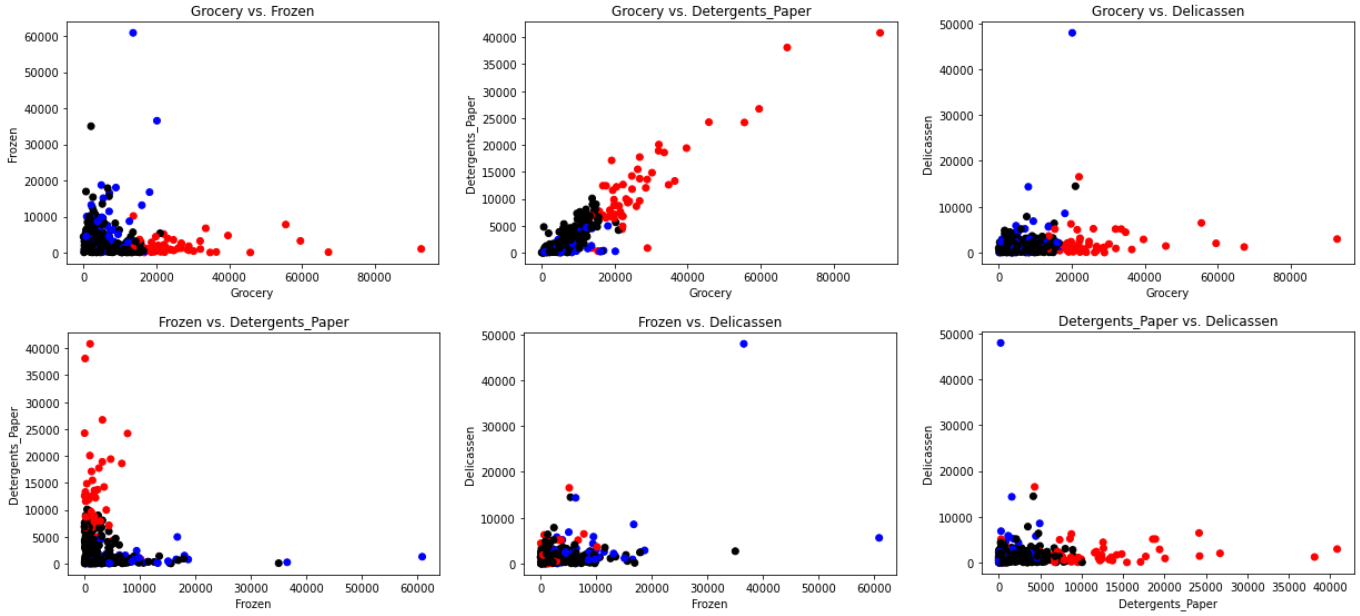
2 Clustering

2.1 Question 1:

	Mean	Range (max-min)
Fresh	12000.3	112148
Milk	5796.3	73443
Grocery	7951.3	92777
Frozen	3071.9	60844
Detergents_Paper	2881.5	40824
Delicassen	1524.9	47940

2.2 Question 2:





We observe firstly that Grocery, Milk, and Detergents_Paper are all related. They are directly proportional to each other, probably because customers buy them together.
When assessing the different clusters, the black cluster seems to group channels/regions where the total customer spend is below around 20,000. The channels/regions in this group spend mostly on Grocery.
The red and blue clusters seem to group channels/regions where the total customer spend is higher than that of the black cluster.
Customers in the channels/regions grouped by the blue cluster seem to spend mostly on Fresh products.
Customers in the channels/regions grouped by the red cluster seem to spend mostly on Milk and Grocery products.

2.3 Question 3:

	k = 3	k = 5	k = 10
BC	3110621948.5	25621025526.7	175002744077.8
WC	80342166920.9	52928148942.6	30488976230.3
BC/WC	0.04	0.48	5.7

We see that as k increases, BC/WC increases, meaning the clusters are becoming denser and more well separated. However, as k increases, bias increases, making the model too fitted to the test data. Hence, with further analysis, we can try finding the near-optimal k value so that the clustering model performs well without being overfitted.