Aman Rangapur
A20517739

CS484: Intro to Machine Learning- Assignment 2

Question-1

I invited six friends to watch a basketball game at home. They brought the following items along.

Friend	Items
Andrew	Cheese, Cracker, Soda, Wing
Betty	Cheese, Soda, Tortilla, Wing
Carl	Cheese, Ice Cream, Wing
Danny	Cheese, Ice Cream, Salsa, Soda, Tortilla
Emily	Salsa, Soda, Tortilla, Wing
Frank	Cheese, Cracker, Ice Cream, Wing

I noticed that my friends often brought Cheese, Soda, and Wing together. Since I prefer to spend on food instead of Soda, I study how likely my friends would bring Soda if they already bought Cheese and Wing. Therefore, please calculate the Lift of this association rule {Cheese, Wing} ==> {Soda} for me.

Friend	Items	{Cheese, Wings}	{Soda}	{Cheese, Wings, Soda}
Andrew	Cheese, Cracker, Soda, Wing	Yes	Yes	Yes
Betty	Cheese, Soda, Tortilla, Wing	Yes	Yes	Yes
Carl	Cheese, Ice Cream, Wing	Yes	No	No
Danny	Cheese, Ice Cream, Salsa, Soda, Tortilla	No	Yes	No
Emily	Salsa, Soda, Tortilla, Wing	No	Yes	No
Frank	Cheese, Cracker, Ice Cream, Wing	Yes	No	No

Support of {Cheese, Wings} = 4/6.
Support of {Soda} = 4/6.
Support of {Cheese, Wings, Soda} = 2/6.
Confidence of {Cheese, Wings}**->**{Soda} = Support of {Cheese, Wings, Soda} / Support of {Cheese, Wings} = (2/6) / (4/6) = 1/2.

Expected Confidence of the rule is Support of {Soda} = 4/6.
Therefore, the Lift of the rule is (1/2) / (4/6) = 3/4 = 0.75.

Question- 2

This question walks you through the typical process of discovering association rules. We will use the market basket data in the Groceries.csv file to discover association rules. Here are the data contents.

1. **Customer: Customer Identifier**
2. **Item: Name of Product Purchased**

For your information, we have sorted the observations in ascending order first by Customer and then by Item. Also, we have removed duplicated items for each customer.

a. What is the number of items in the Universal Set? What is the maximum number of item-sets that we can find in theory from the data? What is the maximum number of association rules that we can generate in theory from the data?

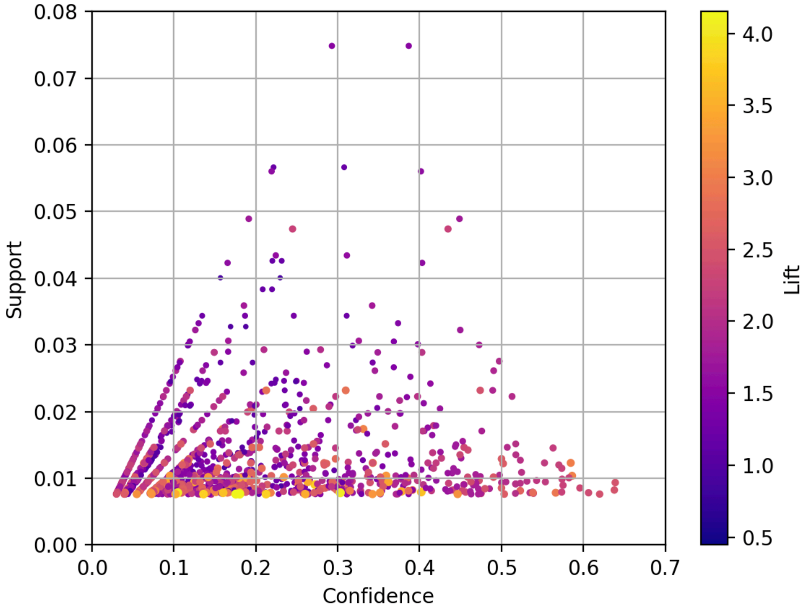
There are total of 169 items in the universal set. The maximum number of item sets we can find from the data are $2^{169} - 1$. The maximum number of association rules we can generate are $3^{169} - 2^{170} + 1$.

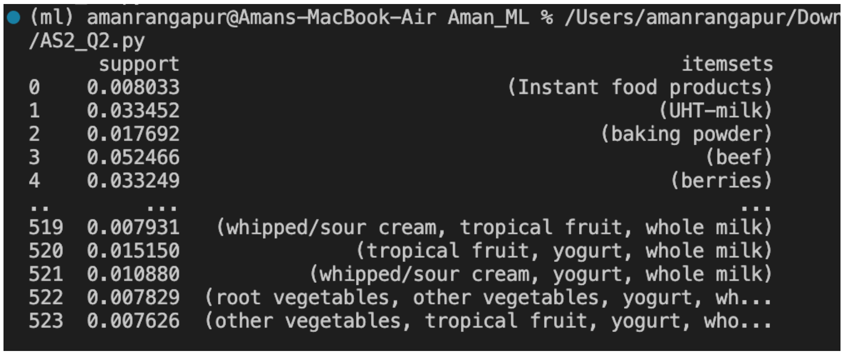
b. We are interested in the itemsets that can be found in the market baskets of at least seventy-five (75) customers. How many itemsets did we find? Also, what is the largest number of items, i.e., k, among these itemsets?

The minimal support should be set as MIN SUPPORT = 75/9835 = 0.0076258261311642. Since a consumer may only ever purchase 32 products at a time, I set the maximum length of an itemset to 32 in accordance. The criteria led to the discovery of 524 itemsets, with the greatest k value being 4.

c. We will use <u>up to</u> the largest k value we found in Part (b) and then generate the association rules whose Confidence metrics are greater than or equal to 1%. How many association rules can we find? Next, we plot the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for these association rules. We will use the Lift metrics to indicate the size of the marker. We will add a color gradient legend to the chart for the Lift metrics.

There are 1228 association rules for minimum threshold = 1/100.

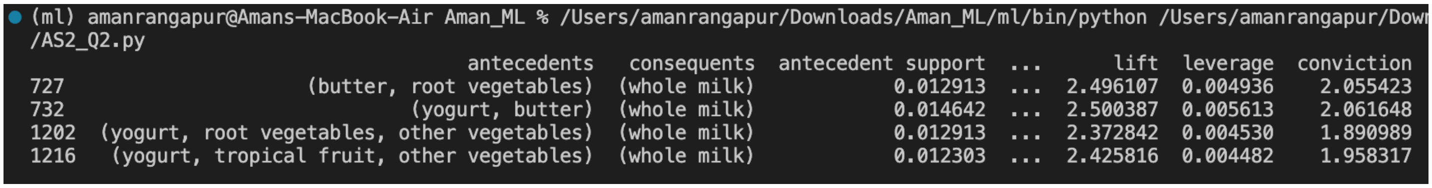




```
● (ml) amanrangapur@Amans-MacBook-Air Aman_ML % /Users/amanrangapur/Dowr
  /AS2_Q2.py
        support                                         itemsets
  0    0.008033                          (Instant food products)
  1    0.033452                                       (UHT-milk)
  2    0.017692                                  (baking powder)
  3    0.052466                                           (beef)
  4    0.033249                                        (berries)
  ..        ...                                              ...
  519  0.007931   (whipped/sour cream, tropical fruit, whole milk)
  520  0.015150               (tropical fruit, yogurt, whole milk)
  521  0.010880           (whipped/sour cream, yogurt, whole milk)
  522  0.007829  (root vegetables, other vegetables, yogurt, wh...
  523  0.007626  (other vegetables, tropical fruit, yogurt, who...
```

d. Among the rules that you found in Part (c), list the rules whose Confidence metrics are greater than or equal to 60%. Please show the rules in a table that shows the Antecedent, the Consequent, the Support, the Confidence, the Expected Confidence, and the Lift.

Antecedent	Consequent	Support	Confidence	Expected Confidence	Lift
{'butter', 'root vegetables'}	{'whole milk'}	0.00824	0.63780	0.25552	2.49611
{'yogurt', 'butter'}	{'whole milk'}	0.00935	0.63889	0.25552	2.50039
{'yogurt', 'root vegetables', 'other vegetables'}	{'whole milk'}	0.00783	0.60630	0.25552	2.37284
{'yogurt', 'tropical fruit', 'other vegetables'}	{'whole milk'}	0.00763	0.61983	0.25552	2.42582



```
● (ml) amanrangapur@Amans-MacBook-Air Aman_ML % /Users/amanrangapur/Downloads/Aman_ML/ml/bin/python /Users/amanrangapur/Dowr
  /AS2_Q2.py
                                   antecedents   consequents  antecedent support  ...      lift  leverage  conviction
  727                   (butter, root vegetables)  (whole milk)            0.012913  ...  2.496107  0.004936    2.055423
  732                            (yogurt, butter)  (whole milk)            0.014642  ...  2.500387  0.005613    2.061648
  1202  (yogurt, root vegetables, other vegetables)  (whole milk)            0.012913  ...  2.372842  0.004530    1.890989
  1216   (yogurt, tropical fruit, other vegetables)  (whole milk)            0.012303  ...  2.425816  0.004482    1.958317
```

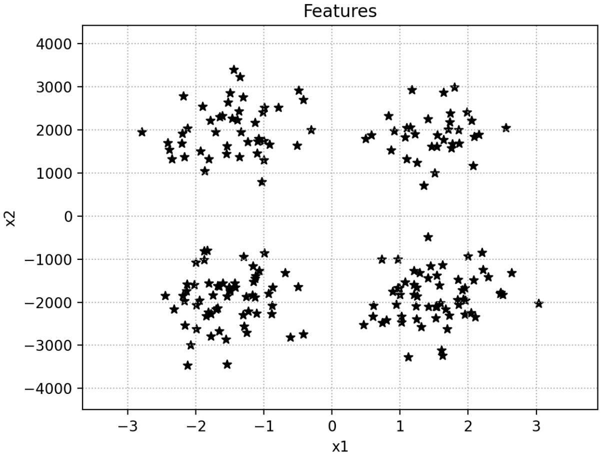
Question- 3

This question demonstrates the effect of rescaling input variables on the cluster results. We will discover clusters using all the observations in the TwoFeatures.csv file with the following specifications.

- **The input interval variables are x1 and x2**
- **The metric is the Manhattan distance**
- **The minimum number of clusters is 1**
- **The maximum number of clusters is 8**
- **Use the Elbow value for choosing the optimal number of clusters**

Since the sklearn.cluster.KMeans class works only with the Euclidean distance, you will need to develop custom Python codes to implement the K-Means algorithm with the Manhattan distance.

a) Plot x2 (vertical axis) versus x1 (horizontal axis). Add gridlines to both axes. Let the graph engine chooses the tick marks. How many clusters do you see in the graph?



There are 4 clusters.

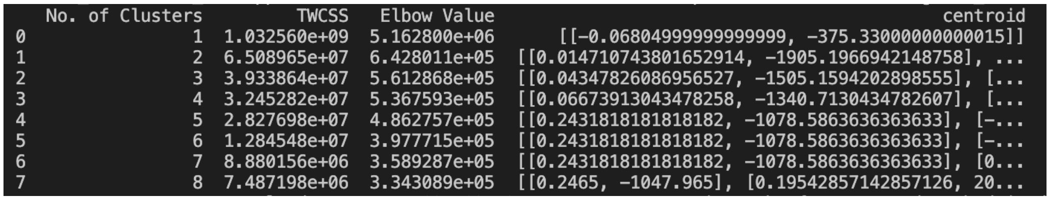
b) Discover the optimal number of clusters without any transformations. List the number of clusters, the Total Within-Cluster Sum of Squares (TWCSS), and the Elbow values in a table. Plot the Elbow Values versus the number of clusters. How many clusters do you find? What are the centroids of your optimal clusters?

The below table shows the Number of clusters, TWCSS, Elbow Value, Centroid value. I have found 8 clusters. The optimal cluster is 2 and it's centroids are [1.47107438e-02, -1.90519669e+03],

[-1.94810127e-01, 1.96788354e+03].

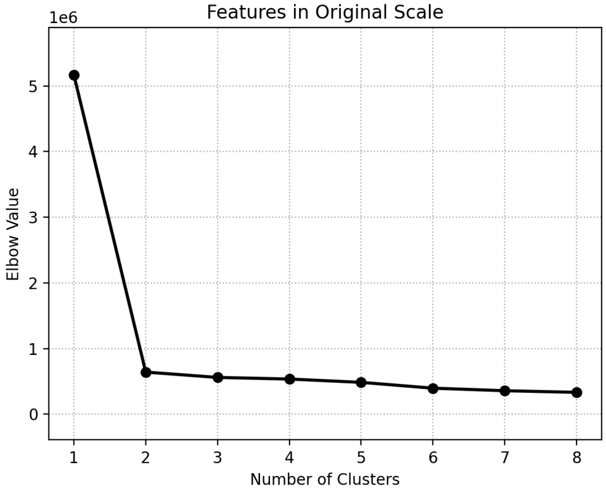
Number of Clusters	TWCSS	Elbow Value
1	1,032,560,056.57	5,162,800.28
2	65,089,654.39	642,801.09
3	39,338,640.78	561,286.75
4	32,452,819.79	536,759.32
5	28,276,984.15	486,275.71
6	12,845,481.32	397,771.48
7	8,880,156.13	358,928.70
8	7,487,198.05	334,308.89

Console screenshot:



```
   No. of Clusters         TWCSS   Elbow Value                                        centroid
0                1  1.032560e+09  5.162800e+06      [[-0.06804999999999999, -375.33000000000015]]
1                2  6.508965e+07  6.428011e+05  [[0.014710743801652914, -1905.1966942148758], ...
2                3  3.933864e+07  5.612868e+05  [[0.04347826086956527, -1505.1594202898555], [...
3                4  3.245282e+07  5.367593e+05  [[0.06673913043478258, -1340.7130434782607], [...
4                5  2.827698e+07  4.862757e+05  [[0.2431818181818182, -1078.5863636363633], [-...
5                6  1.284548e+07  3.977715e+05  [[0.2431818181818182, -1078.5863636363633], [-...
6                7  8.880156e+06  3.589287e+05  [[0.2431818181818182, -1078.5863636363633], [0...
7                8  7.487198e+06  3.343089e+05  [[0.2465, -1047.965], [0.19542857142857126, 20...
```

The below graphs illustrate Elbow value vs Number of Clusters graph.

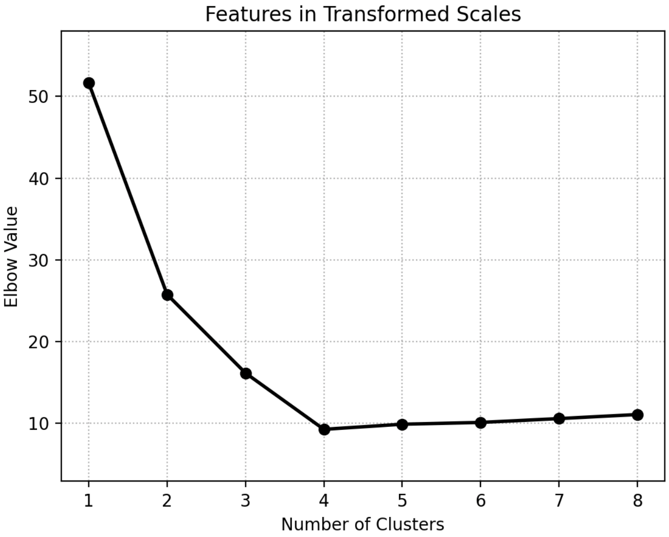


c) **Linearly rescale x1 such that the resulting variable has a minimum of zero and a maximum of ten. Likewise, rescale x2. Discover the optimal number of clusters from the transformed observations. List the number of clusters, the Total Within-Cluster Sum of Squares (TWCSS), and the Elbow values in a table. Plot the Elbow Values versus the number of clusters. How many clusters do you find? What are the centroids of your optimal clusters in the original scale of x1 and x2?**

The below table shows the no. of clusters, TWCSS, Elbow value.

Number of Clusters	TWCSS	Elbow Value
1	10,326.63	51.63
2	2,584.09	25.75
3	1,149.05	16.15
4	472.77	9.25
5	416.47	9.88
6	388.68	10.10
7	340.08	10.57
8	312.85	11.07

I have found 8 clusters. In the below graph, the elbow appears when the number of clusters is four. Therefore, the optimal number of clusters is four. The centroids for optimal clusters are [7.36, 7.82], [2.26, 7.99], [2.17, 2.23], [7.42, 2.31].



d) If you are doing everything correctly, you should discover two different optimal cluster solutions. In your words, how do you explain the difference?

The discovery of two different optimal cluster solutions for the original scale of the features and for the transformed data reveals the effect of the data transformation on the clustering results. The variation between the two solutions can be attributed to the influence that scaling of the features has on the similarity assessments between the data points. By transforming the features, patterns and connections that were not noticeable in the original data may be uncovered, leading to a differing optimal clustering solution. This underscores the importance of taking into account the scaling of the features when conducting clustering analysis and its potential impact on the results.

In short, when the features are on different scales, rescaling the features can assist in determining a more appropriate number of clusters.