Introduction to Machine Learning Assignment 2

Name: Vivekanand Reddy Malipatel

CWID: A20524971

**Question 1 Answer:**

The lift of an association rule {Cheese, Wing} => {Soda} can be calculated as follows:

Lift({Cheese, Wing} => {Soda}) = (P({Cheese, Wing} => {Soda}) / (P({Cheese, Wing}) \* P({Soda})))

P({Cheese, Wing}) = 4/6

P({Soda}) = 4/6

P({Cheese, Wing} => {Soda}) = 2/6

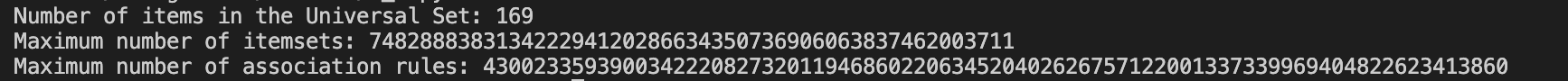
Lift({Cheese, Wing} => {Soda}) = (2/6) / (4/6 \* 4/6) = 0.75

So the lift of the association rule {Cheese, Wing} => {Soda} is 0.75, which means that the presence of Cheese and Wing in a transaction prohibits the likelihood of having Soda by 0.75 times compared to the likelihood of having Soda if Cheese and Wing were not present.

**Question 2 Answers:**

1. (The Python code for this is in the file: 2\_a.py)

Output Screenshot :



Answer:

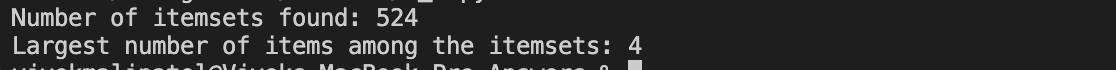
Number of items in the Universal Set: **169**

Maximum number of itemsets: **748288838313422294120286634350736906063837462003711**

Maximum number of association rules: **430023359390034222082732011946860220634520402626757122001337339969404822623413860**

1. (The Python code for this is in the file: 2\_b.py)

Output Screenshot :



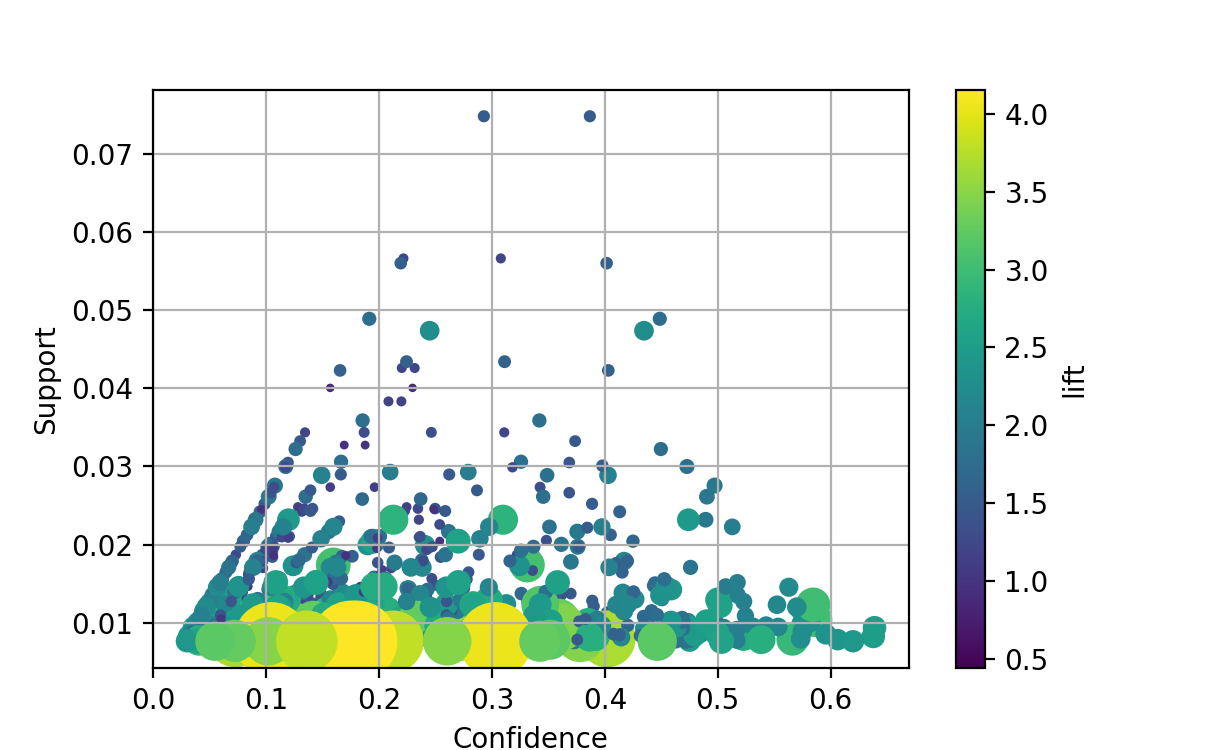
Answer:

Number of itemsets found: 524

Largest number of items among the itemsets: 4

1. (The Python code for this is in the file: 2\_c.py)

Plot :



Output Screenshot:

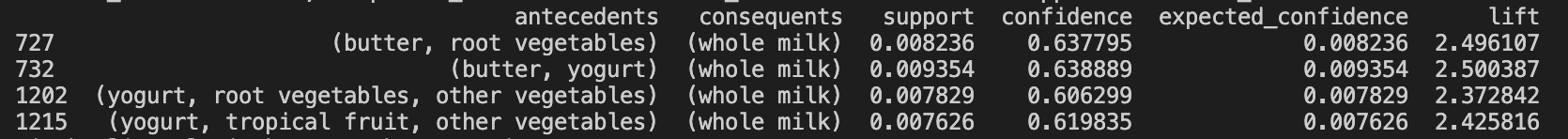


Answer:

Number of Association Rules : 1228

1. (The Python code for this is in the file: 2\_d.py)

Output Screenshot:



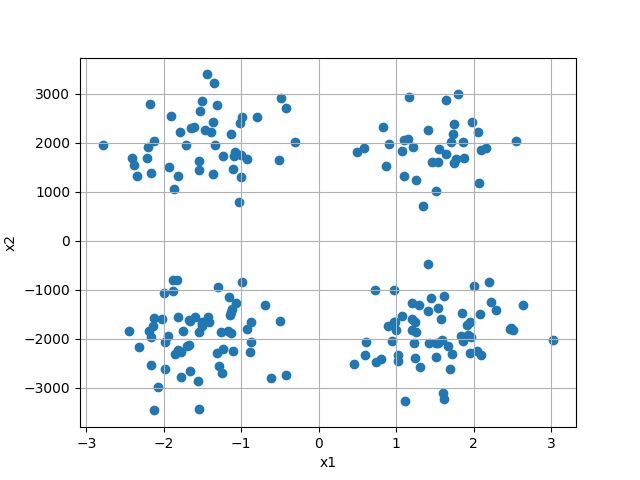
Answer:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| antecedents | consequents | support | confidence | expected\_confidence | lift |
| frozenset({'butter', 'root vegetables'}) | frozenset({'whole milk'}) | 0.008235892221657347 | 0.6377952755905512 | 0.008235892221657347 | 2.4961068585089814 |
| frozenset({'yogurt', 'butter'}) | frozenset({'whole milk'}) | 0.009354346720894764 | 0.6388888888888888 | 0.009354346720894764 | 2.500386877127824 |
| frozenset({'other vegetables', 'yogurt', 'root vegetables'}) | frozenset({'whole milk'}) | 0.007829181494661922 | 0.6062992125984252 | 0.007829181494661922 | 2.3728423222863158 |
| frozenset({'other vegetables', 'tropical fruit', 'yogurt'}) | frozenset({'whole milk'}) | 0.007625826131164209 | 0.6198347107438016 | 0.007625826131164209 | 2.4258155114068 |

**Question 3 Answers:**

1. The Python code for this is in the file: 3\_a.py)

Plot :



Answer :

After plotting the graph for the given csv file data. We can see **4** Clusters.

1. (The Python code for this is in the file: 3\_b.py)

Plot:

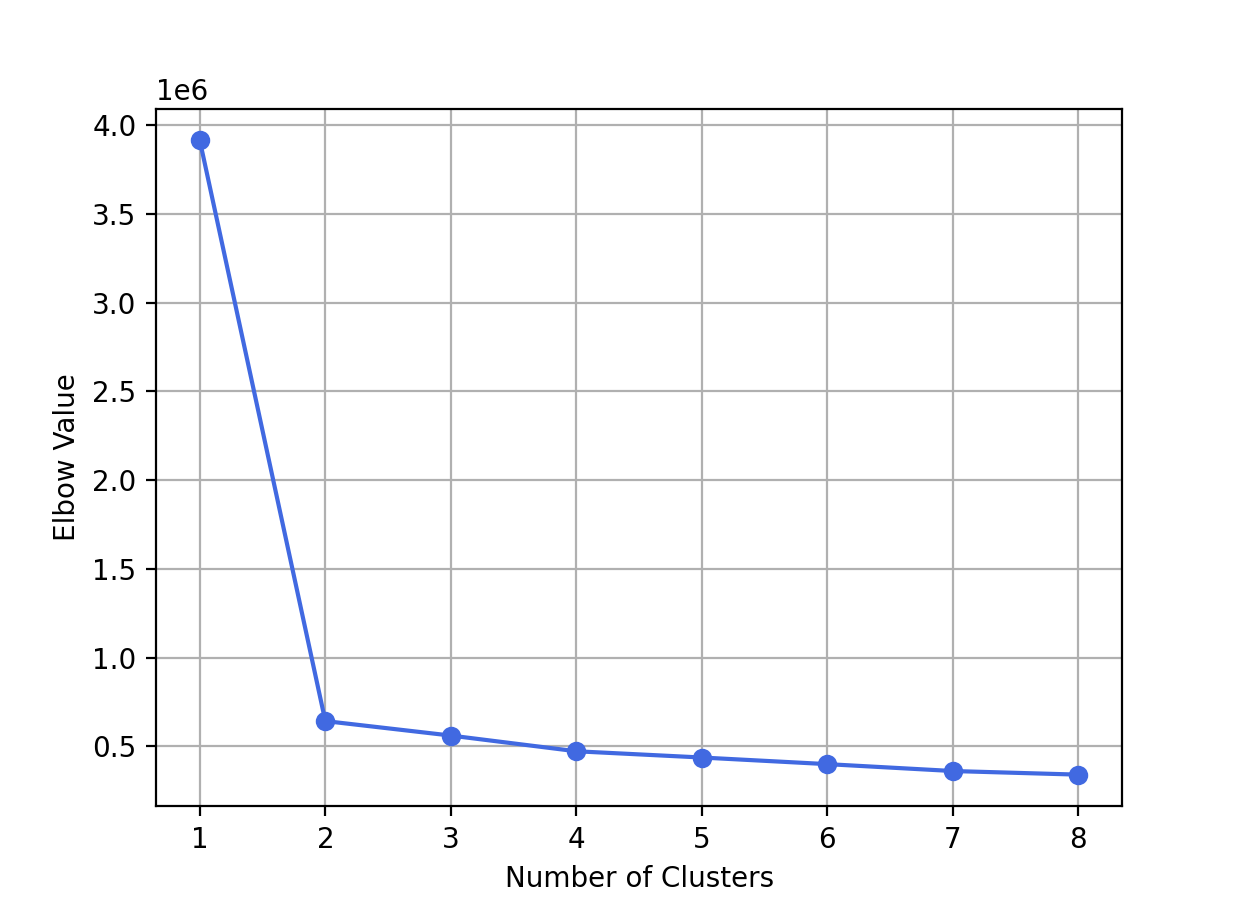


Table:

|  |  |  |
| --- | --- | --- |
| N Cluster | Total WCSS | Elbow |
| 1 | 782891013.0238838 | 3914455.0651194192 |
| 2 | 65089654.389735684 | 642801.0914895514 |
| 3 | 39336456.73787262 | 561190.5233420159 |
| 4 | 23904953.907457363 | 472686.29825282836 |
| 5 | 16753104.001151742 | 437484.9906273829 |
| 6 | 12920965.175807077 | 400209.84188469034 |
| 7 | 8955639.98858235 | 361367.06280921295 |
| 8 | 7101906.903299999 | 341090.52933626255 |

Answer:

The Optimal number of clusters with the given elbow values is: **2**

The Centroids of two cluster are: (x1,y1) (x2,y2) are the centroids

**X1 = -0.194810 Y1 = 1967.883544**

**X2 = 0.014711 Y2 = -1905.196694**

1. (The Python code for this is in the file: 3\_c.py)

Plot:

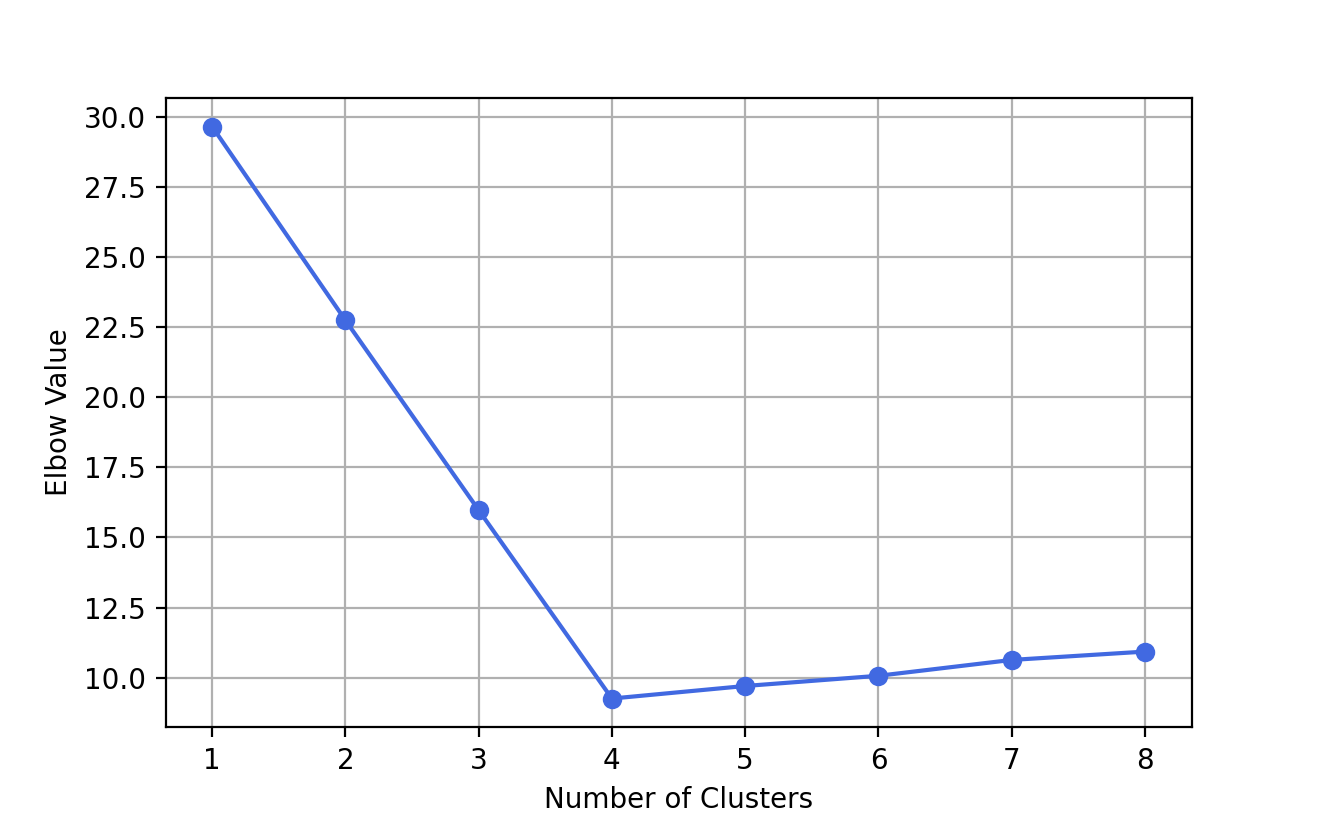


Table:

|  |  |  |
| --- | --- | --- |
| N Cluster | Total WCSS | Elbow |
| 1 | 5929.373746784754 | 29.64686873392377 |
| 2 | 2294.06353363172 | 22.74379295479966 |
| 3 | 1148.4589604208093 | 15.972892359357436 |
| 4 | 472.7690931119309 | 9.254667350391188 |
| 5 | 414.54248827566096 | 9.702143737199108 |
| 6 | 355.11013827238946 | 10.066034160723163 |
| 7 | 307.2879592054128 | 10.631243336329957 |
| 8 | 284.45993478786687 | 10.928476843346912 |

Answer:

The Optimal number of clusters with the given elbow values is: **4**

The Centroids of two cluster are:

**0 2.260023 7.994898**

**1 2.172680 2.237297**

**2 7.422117 2.313790**

**3 7.369618 7.826083**

**x1 x2**

1. Answer:

Prior to performing rescaling on the input variables, the optimal number of clusters was determined to be **2** based on the elbow value obtained through the analysis. However, upon applying the rescaling technique, the optimal number of clusters increased to **4**.

Prior to the rescaling, when visualizing the untouched TwoFeatures.csv data (3\_a.py), **4** clusters appeared to be an optimal solution. However, due to differences in the range and scale of x1 and x2, K-Means may have been giving more weight to one variable over the other, causing suboptimal clustering results. The rescaling of input variables allows for equal weighting of each variable, ultimately enabling K-Means to identify clusters based on both variables in an equitable manner.

Therefore, based on the above analysis, the optimal number of clusters for this particular dataset would be **4**.