Question 1)

1. Confidence: Confidence measures the probability of the consequent (Soda) occurring when the antecedent (Cheese and Wings) is true.
   - ➢ Confidence = P(Soda | Cheese and Wings)

2. Lift: Lift measures how much more likely the consequent is to occur when the antecedent is true compared to its overall occurrence.
   - ➢ Lift = Confidence / P(Soda)

3. Leverage: Leverage measures the difference between the observed frequency of Cheese and Wings occurring with Soda and what would be expected if they were independent.
   - ➢ Leverage = (Lift – 1)*(P(Cheese and Wings)*P(Soda))

4. Zhang's Metric: Zhang's metric is a modified version of the lift that accounts for the potential influence of other factors.
   - ➢ Zhang's Metric = (P(Soda | Cheese and Wings) - (P(Soda | Cheese and Wings) * P(Soda))) / 1-(P(Soda | Cheese and Wings) * P(Soda))

Metrics Calculation:

1. Confidence: Confidence = P(Soda | Cheese and Wings) = **0.5**

2. Lift: Lift = Confidence / P(Soda) = **0.5 / 0.667 ≈ 0.75**

3. Leverage: Leverage = (Lift – 1)*(P(Cheese and Wings)*P(Soda)) = (**0.75-1) - (4/6 * 4/6) ≈ - 0.111**

4. Zhang's Metric: Zhang's Metric = (P(Soda | Cheese and Wings) - (P(Soda | Cheese and Wings) * P(Soda))) / 1-(P(Soda | Cheese and Wings) * P(Soda)) Zhang's Metric = **(2/6 – 4/6*4/6) / 1 - (4/6 * 4/6) = -0.2**

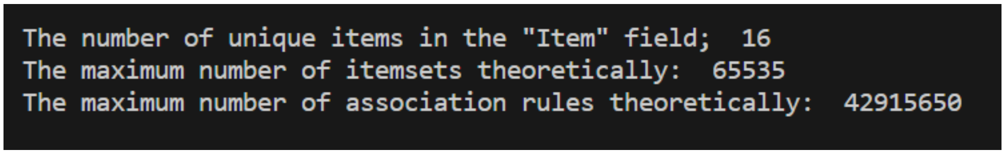
Question 2)

Dataset: Chinese_Bakery.csv

The dataset consists of two fields - 'Customer' and 'Item,' with observations sorted by Customer and duplicates removed.

Part 1: Universal Set and Theoretical Limits

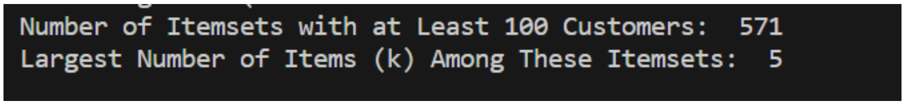
- Number of Items in the Universal Set: 16
- Maximum Number of Theoretical Itemsets: 65535
- Maximum Number of Theoretical Association Rules: 983040



```
The number of unique items in the "Item" field;  16
The maximum number of itemsets theoretically:  65535
The maximum number of association rules theoretically:  42915650
```

Part 2: Itemsets with Minimum Customer Threshold

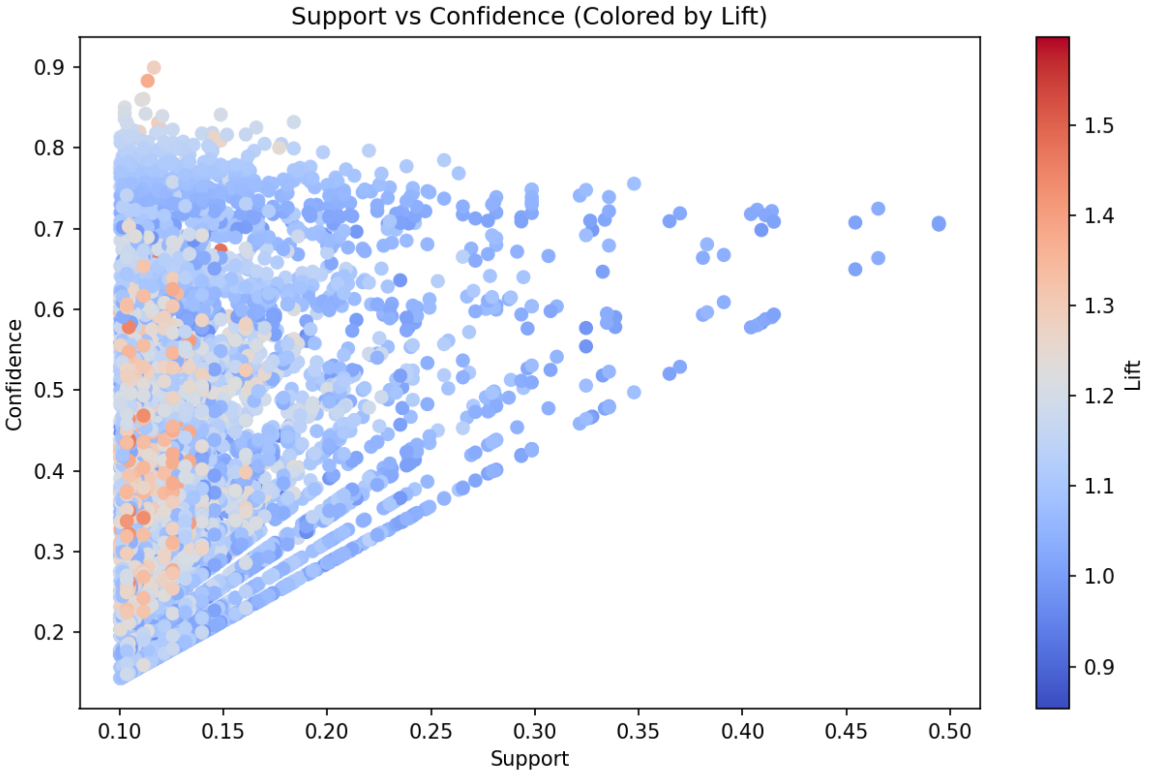
- Number of Itemsets with at Least 100 Customers: 571
- Largest Number of Items (k) Among These Itemsets: 5



```
Number of Itemsets with at Least 100 Customers:  571
Largest Number of Items (k) Among These Itemsets:  5
```

Part 3: Association Rules with Confidence Threshold

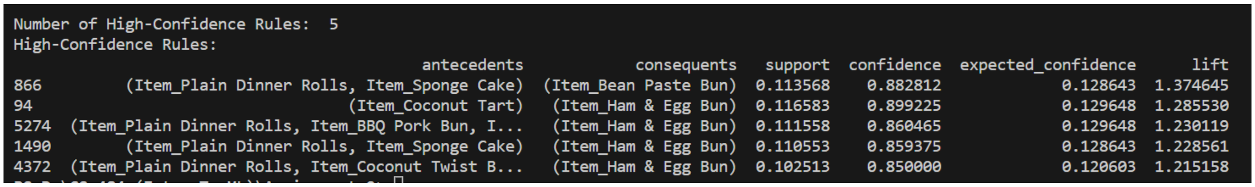
- Number of Association Rules with at Least 1% Confidence: 5424
- Support vs. Confidence Plot (Colored by Lift):



The vertical axis represents Support, the horizontal axis represents Confidence, and the color represents Lift.

Part 4: High-Confidence Rules (Confidence ≥ 85%)

- Number of High-Confidence Rules: 5
- Table of High-Confidence Rules (Sorted by Lift):



```
Number of High-Confidence Rules:  5
High-Confidence Rules:
                                        antecedents              consequents   support  confidence  expected_confidence      lift
866         (Item_Plain Dinner Rolls, Item_Sponge Cake)  (Item_Bean Paste Bun)  0.113568    0.882812             0.128643  1.374645
94                                (Item_Coconut Tart)     (Item_Ham & Egg Bun)  0.116583    0.899225             0.129648  1.285530
5274  (Item_Plain Dinner Rolls, Item_BBQ Pork Bun, I...    (Item_Ham & Egg Bun)  0.111558    0.860465             0.129648  1.230119
1490        (Item_Plain Dinner Rolls, Item_Sponge Cake)    (Item_Ham & Egg Bun)  0.110553    0.859375             0.128643  1.228561
4372  (Item_Plain Dinner Rolls, Item_Coconut Twist B...    (Item_Ham & Egg Bun)  0.102513    0.850000             0.120603  1.215158
```

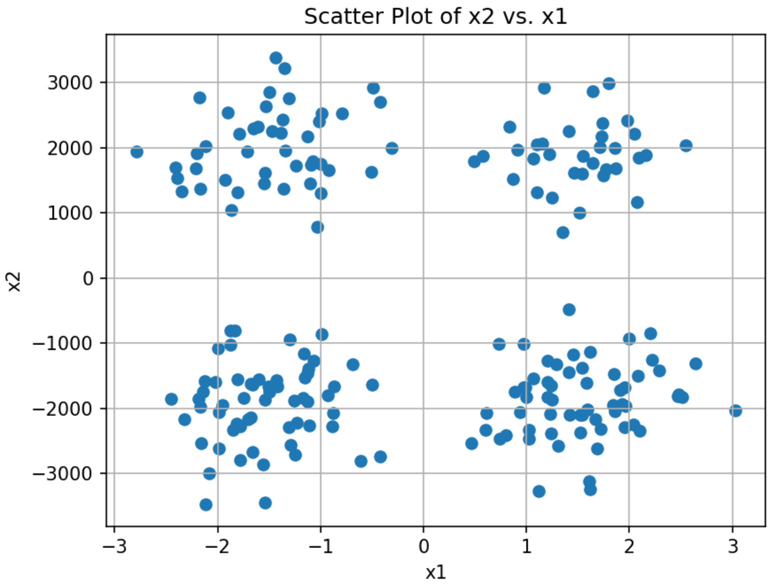
Columns: Antecedent, Consequent, Support, Confidence, Expected Confidence, Lift

Question 3)

Part a:

Plot x2 vs x1:

The assignment's first task involved generating a scatter plot using data from the 'TwoFeatures.csv' file. The horizontal axis represents variable x1, while the vertical axis represents variable x2. Gridlines were added to enhance the graph's readability. In the graph, the it was observed that there can be 4 clusters.



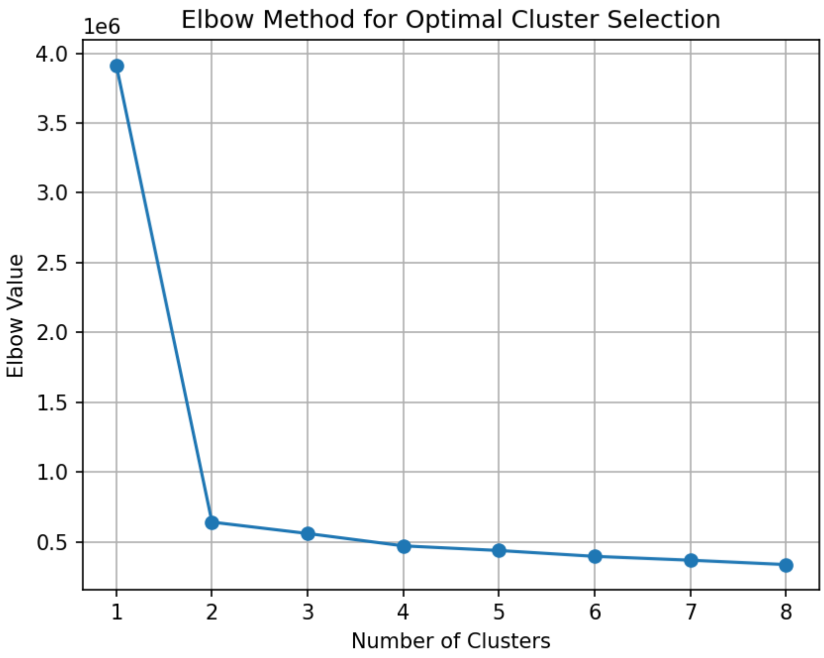
Part b):

- ➢ Below is the table computed for Dataset without any transformation:

NO OF CLUSTERS	TWCSS	ELBOW VALUE
1	7.83E+08	3.91E+06
2	6.51E+07	6.43E+05
3	3.93E+07	5.61E+05
4	2.39E+07	4.73E+05
5	1.68E+07	4.40E+05
6	1.28E+07	3.98E+05
7	8.96E+06	3.70E+05
8	7.03E+06	3.39E+05

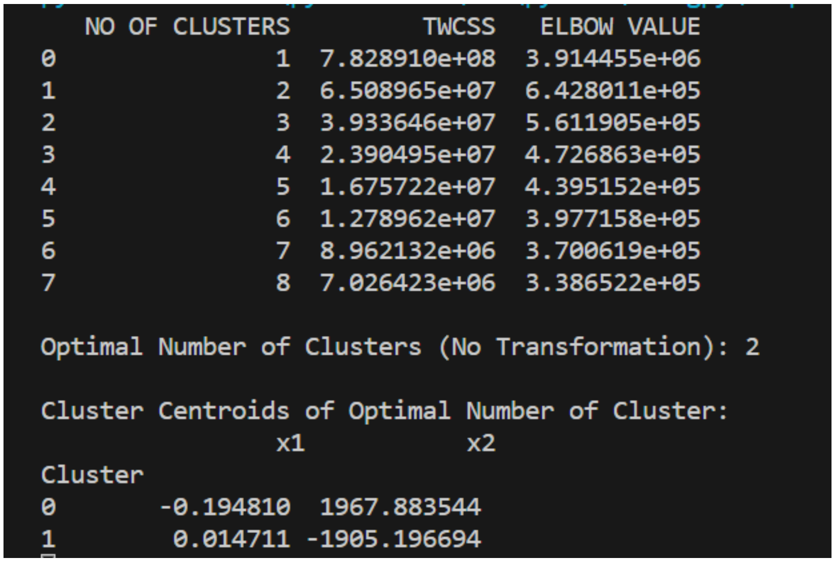
Table Columns: No of Cluster, TWCSS (Total Within-Cluster Squared Sum), Elbow Value

- Using the above Elbow values for each cluster below graph is plotted:



- The above values are calculated using the K-Means algorithm with the Manhattan distance metric, investigating a range of cluster numbers from 1 to 8.

- The optimal number of clusters found in this analysis is **2 Clusters**
- The centroids of these optimal clusters in the original scale of x1 and x2:



```
   NO OF CLUSTERS          TWCSS   ELBOW VALUE
0               1   7.828910e+08  3.914455e+06
1               2   6.508965e+07  6.428011e+05
2               3   3.933646e+07  5.611905e+05
3               4   2.390495e+07  4.726863e+05
4               5   1.675722e+07  4.395152e+05
5               6   1.278962e+07  3.977158e+05
6               7   8.962132e+06  3.700619e+05
7               8   7.026423e+06  3.386522e+05

Optimal Number of Clusters (No Transformation): 2

Cluster Centroids of Optimal Number of Cluster:
                x1           x2
Cluster
0        -0.194810  1967.883544
1         0.014711 -1905.196694
```

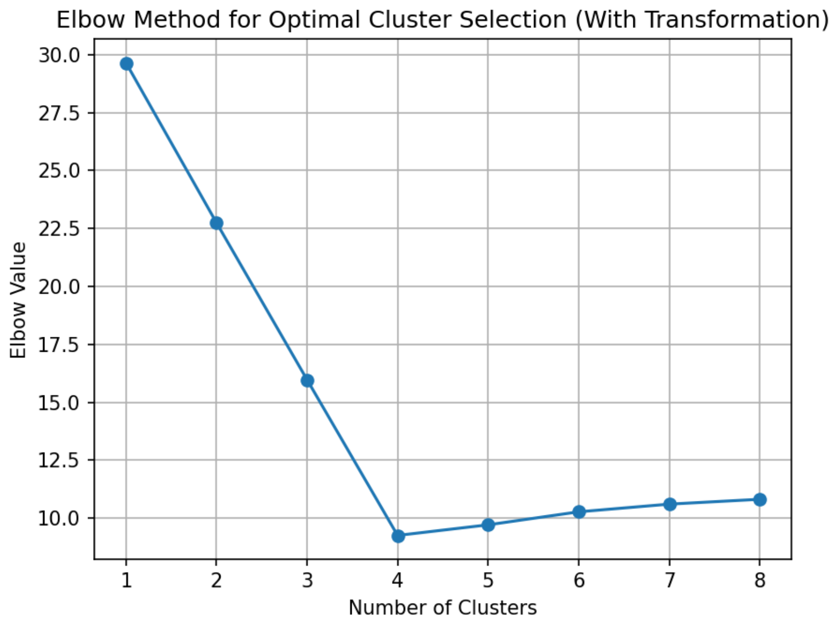
Part c):

- Below is the table computed for Dataset with transformation (within range of (0,10)):

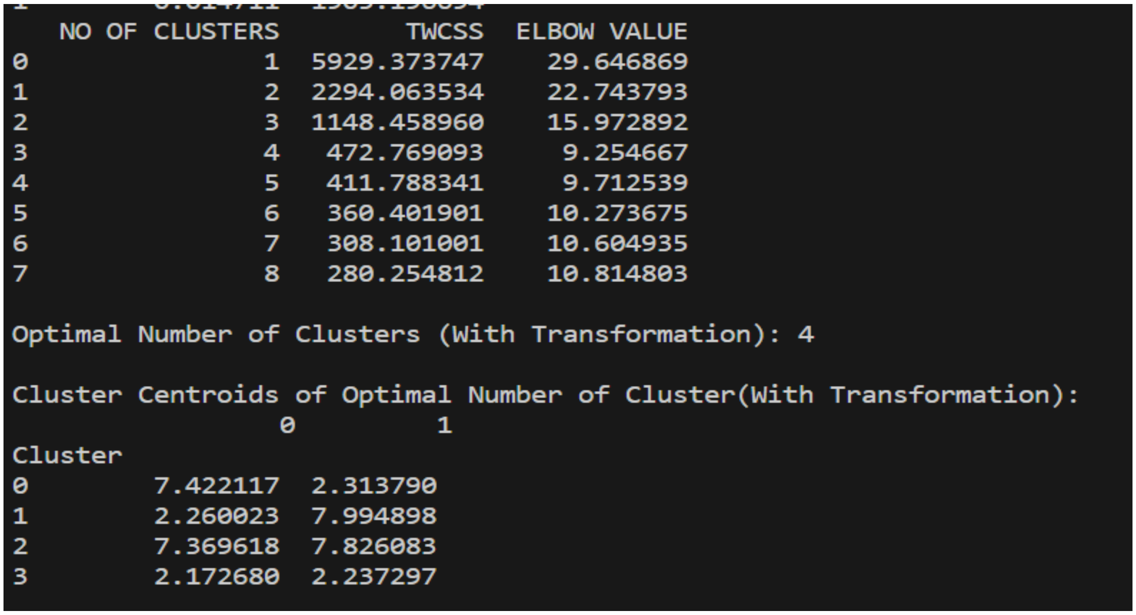
NO OF CLUSTERS	TWCSS	ELBOW VALUE
1	5929.373747	29.646869
2	2294.063534	22.743793
3	1148.45896	15.972892
4	472.769093	9.254667
5	411.788341	9.712539
6	360.401901	10.273675
7	308.101001	10.604935
8	280.254812	10.814803

Table Columns: No of Cluster, TWCSS (Total Within-Cluster Squared Sum), Elbow Value

- Using the above Elbow values for each cluster below graph is plotted:



- The optimal number of clusters found in this analysis is **4 Clusters**
- The centroids of these optimal clusters in the original scale of x1 and x2:



```
   NO OF CLUSTERS        TWCSS  ELBOW VALUE
0               1  5929.373747    29.646869
1               2  2294.063534    22.743793
2               3  1148.458960    15.972892
3               4   472.769093     9.254667
4               5   411.788341     9.712539
5               6   360.401901    10.273675
6               7   308.101001    10.604935
7               8   280.254812    10.814803

Optimal Number of Clusters (With Transformation): 4

Cluster Centroids of Optimal Number of Cluster(With Transformation):
                 0         1
Cluster
0         7.422117  2.313790
1         2.260023  7.994898
2         7.369618  7.826083
3         2.172680  2.237297
```

Part d)

The emergence of two distinct optimal cluster solutions stems from the application of data transformation to the input variables (x1 and x2) within the dataset:

- The presence of variables with higher variances can exert significant influence on the clustering process. When data is rescaled, it can shift the focus of the clustering algorithm toward variables with lower variances, potentially leading to the formation of distinct clusters due to the altered emphasis on variable relationships.
- K-means relies on a distance metric (typically Euclidean distance) to gauge the similarity between data points. Rescaling modifies the scale of these distance calculations. For instance, if one variable is originally measured in meters while another is in millimeters, the former will disproportionately affect distance calculations. Consequently, rescaling can result in clusters forming based on the variable with the larger scale.
- Rescaling also impacts how outliers are treated within the data. Outliers can wield disproportionate influence over clustering results. When data is rescaled, outliers that

were initially extreme may become less extreme, potentially leading to the assignment of data points to different clusters.

- The rescaling of data can alter the shape and compactness of clusters. In some instances, rescaling can make clusters more tightly packed, causing fewer, denser clusters to emerge. Conversely, it can also lead to less compact clusters, resulting in the formation of more clusters.
- K-means is sensitive to the initial placement of cluster centroids. Rescaling data can influence the initial positions of these centroids, consequently affecting the convergence of the clustering algorithm and the resultant cluster assignments.

In summary, rescaling data can indeed change the number of clusters in K-means clustering because it alters the distances between data points and the importance of individual variables in the clustering process. Standardizing or normalizing the data can help address these issues and lead to more reliable and interpretable clustering results.