CS 584-04: Machine Learning

NAME : SOURAV YADAV

AID: A20450418

Spring 2020 Assignment 2

# Question 1 (35 points)

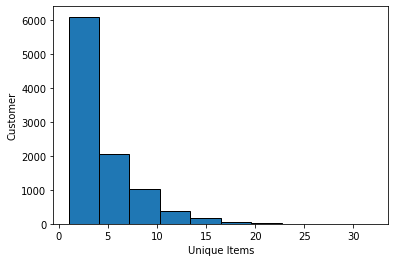
The file Groceries.csv contains market basket data. The variables are:

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

After you have imported the CSV file, please discover association rules using this dataset. For your information, the observations have been sorted in ascending order by Customer and then by Item. Also, duplicated items for each customer have been removed.

1. (5 points) Create a data frame that contains the number of unique items in each customer’s market basket. Draw a histogram of the number of unique items. What are the 25th, 50th, and the 75th percentiles of the histogram?

**Answer:**



**25th Percentaile:2.0**

**50th Percentiles:3.0**

**75th Percentiles:6.0**

1. (10 points) We are only interested in the *k*-itemsets that can be found in the market baskets of at least seventy five (75) customers. How many itemsets can we find? Also, what is the largest *k* value among our itemsets?

**Answer : Number of itemsets found:524**

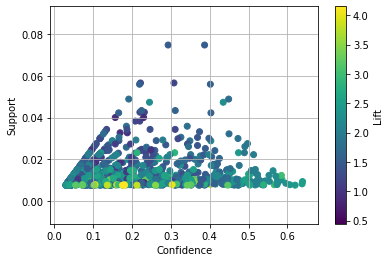
**The largest k value among our itemsets:4**

1. (10 points) Find out the association rules whose Confidence metrics are greater than or equal to 1%. How many association rules can we find? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent. Please **do not** display those rules in your answer.

**Answer : Number of Association Rules: 1228**

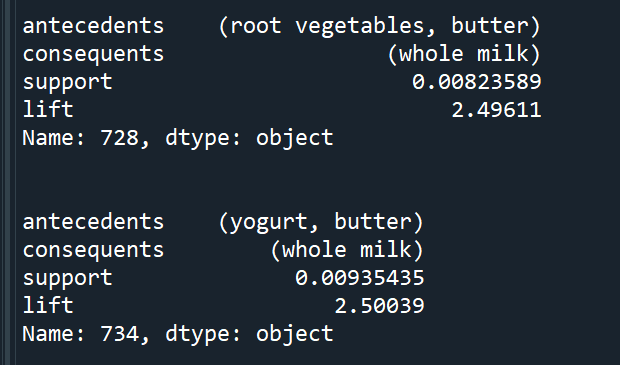
1. (5 points) Plot the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you have found in (c). Please use the Lift metrics to indicate the size of the marker.

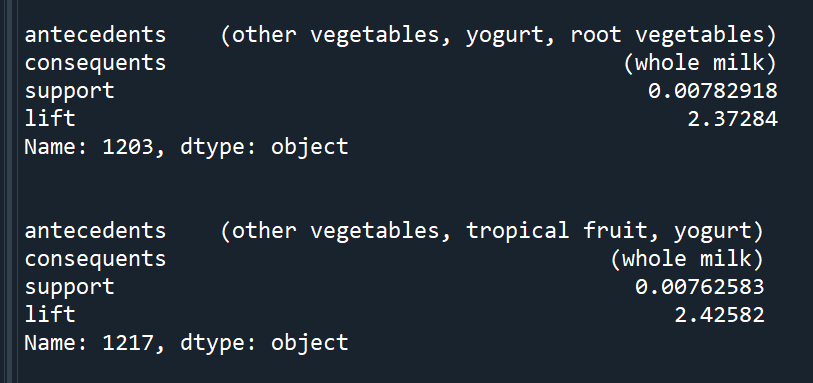
**Answer :**



1. (5 points) List the rules whose Confidence metrics are greater than or equal to 60%. Please include their Support and Lift metrics.

**Answer :**

****

****

# Question 2 (30 points)

The K-means algorithm works only with interval features. One way to apply the k-means algorithm to categorical features is to transform them into a new interval feature space. However, this approach can be very inefficient, and it does not produce good results.

For clustering categorical features, we should consider the K-modes clustering algorithm which extends the K-means algorithm by using different dissimilarity measures and a different method for computing cluster centers. See this article for more details. Huang, Z. (1997). “A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining.” In Proceedings of the *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1–8. New York: ACM Press.

Please implement the K-modes clustering method in Python and then apply the method to the cars.csv. Your input fields are these four categorical features: Type, Origin, DriveTrain, and Cylinders. **Please do not remove the missing or blank values in these four features**. Instead, consider these values as a separate category.

The cluster centroids are the modes of the input fields. In the case of tied modes, choose the lexically or numerically lowest one.

Suppose a categorical feature has observed values . Their global frequencies (i.e., number of observations) are . Please be noted that these global frequencies do not change with the cluster assignment. The distance metric between two values is if . Otherwise, . The distance between any two observations is the sum of the distance metric of the four categorical features.

1. (5 points) What are the frequencies of the categorical feature Type?

**Answer :**

**Frequencies of the categorical feature**

**[ 60 262 49 30 24 3]**

**'Type': Type Freq\_Type**

**0 SUV 60**

**1 Sedan 262**

**2 Sedan 262**

**3 Sedan 262**

**4 Sedan 262**

**5 Sedan 262**

**6 Sports 49**

**7 Sedan 262**

**8 Sedan 262**

**9 Sedan 262**

**10 Sedan 262**

**11 Sedan 262**

**12 Sedan 262**

**13 Sedan 262**

**14 Sedan 262**

**15 Sedan 262**

**16 Sedan 262**

**17 Sedan 262**

**18 Sedan 262**

**19 Sedan 262**

**20 Sports 49**

**21 Sports 49**

**22 Sports 49**

**23 Sports 49**

**24 Wagon 30**

**25 Wagon 30**

**26 SUV 60**

**27 SUV 60**

**28 Sedan 262**

**29 Sedan 262**

**.. ... ...**

**398 Truck 24**

**399 Truck 24**

**400 Wagon 30**

**401 SUV 60**

**402 Sedan 262**

**403 Sedan 262**

**404 Sedan 262**

**405 Sedan 262**

**406 Sedan 262**

**407 Sedan 262**

**408 Sedan 262**

**409 Sedan 262**

**410 Sedan 262**

**411 Sedan 262**

**412 Sedan 262**

**413 Wagon 30**

**414 Wagon 30**

**415 Wagon 30**

**416 SUV 60**

**417 Sedan 262**

**418 Sedan 262**

**419 Sedan 262**

**420 Sedan 262**

**421 Sedan 262**

**422 Sedan 262**

**423 Sedan 262**

**424 Sedan 262**

**425 Sedan 262**

**426 Wagon 30**

**427 Wagon 30**

1. (5 points) What are the frequencies of the categorical feature DriveTrain?

**Answer :**

**Frequencies of the categorical feature :** **[ 92 226 110]**

**'Type': DriveTrain Freq\_DriveTrain**

**0 AWD 92**

**1 FWD 226**

**2 FWD 226**

**3 FWD 226**

**4 FWD 226**

**5 FWD 226**

**6 RWD 110**

**7 FWD 226**

**8 FWD 226**

**9 FWD 226**

**10 AWD 92**

**11 AWD 92**

**12 FWD 226**

**13 AWD 92**

**14 FWD 226**

**15 AWD 92**

**16 AWD 92**

**17 AWD 92**

**18 AWD 92**

**19 AWD 92**

**20 FWD 226**

**21 FWD 226**

**22 AWD 92**

**23 AWD 92**

**24 AWD 92**

**25 AWD 92**

**26 AWD 92**

**27 AWD 92**

**28 RWD 110**

**29 RWD 110**

**.. ... ...**

**398 RWD 110**

**399 AWD 92**

**400 FWD 226**

**401 AWD 92**

**402 FWD 226**

**403 FWD 226**

**404 FWD 226**

**405 FWD 226**

**406 FWD 226**

**407 FWD 226**

**408 FWD 226**

**409 FWD 226**

**410 FWD 226**

**411 FWD 226**

**412 FWD 226**

**413 FWD 226**

**414 FWD 226**

**415 FWD 226**

**416 AWD 92**

**417 FWD 226**

**418 AWD 92**

**419 FWD 226**

**420 AWD 92**

**421 FWD 226**

**422 AWD 92**

**423 FWD 226**

**424 FWD 226**

**425 FWD 226**

**426 FWD 226**

**427 AWD 92**

1. (5 points) What is the distance metric between ‘Asia’ and ‘Europe’ for Origin?

**Answer : Distance between Asia and Europe is : 0.014459195224863643**

1. (5 points) What is the distance metric between Cylinders = 5 and Cylinders = Missing?

**Answer : Distance between Cylinder5 and Cylinder0 is : 0.6428571428571428**

1. (5 points) Apply the K-modes method with **three clusters**. How many observations in each of these three clusters? What are the centroids of these three clusters?

**Answer :**

**Number of Observations in Cluster1: 71**

**Number of Observations in Cluster2: 216**

**Number of Observations in Cluster3: 141**

**Centroid of CLuster 1: [60.0, 147.0, 92.0, 190.0] ['SUV', 'USA', 'AWD', 6.0]**

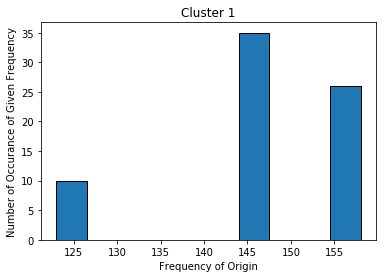
**Centroid of CLuster 2: [262.0, 158.0, 226.0, 136.0]** **['Sedan', 'Europe', 'FWD', 6.0]**

**Centroid of CLuster 3: [262.0, 123.0, 110.0, 190.0]** **['Sedan', 'Asia', 'FWD', 4.0]**

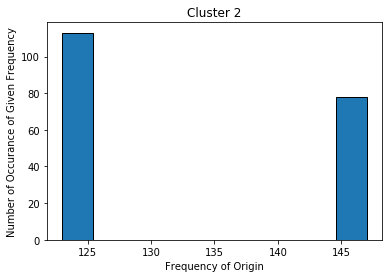
1. (5 points) Display the frequency distribution table of the Origin feature in each cluster.

**Answer:**

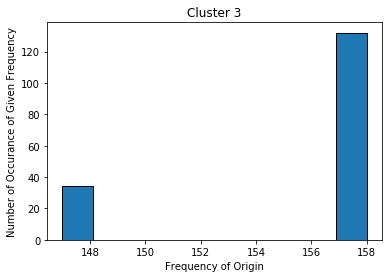
Frequency Distribution of Cluster 1: Asia :26, Europe:10, USA:35



**Frequency Distribution of Cluster 2: Asia :0, Europe:113, USA:78**



**Frequency Distribution of Cluster 3: Asia :132, Europe:0, USA:34**

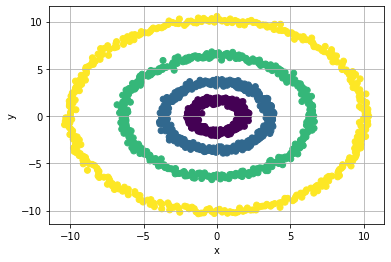


# Question 3 (35 points)

Apply the Spectral Clustering method to the FourCircle.csv. Your input fields are x and y. Wherever needed, specify random\_state = 60616 in calling the KMeans function.

1. (5 points) Plot y on the vertical axis versus x on the horizontal axis. How many clusters are there based on your visual inspection?

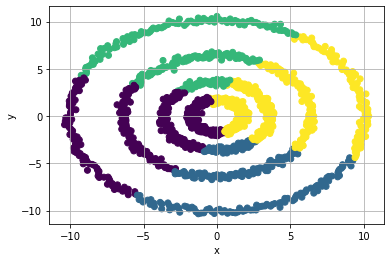
**Answer :** **Based on my observation , number of clusters are 4.**



1. (5 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifiers to control the color scheme. Please comment on this K-mean result.

**Answer :**

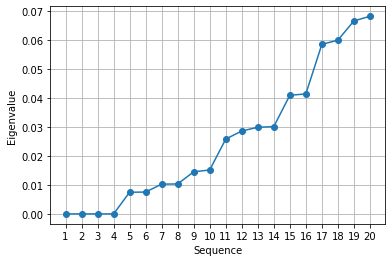
**The below graph shows that without transforming the data we get wrong results from K mean clustering algorithm. The points which belong to different clusters has been assigned to same clusters as shown below with different colors. Hence we can not use K mean clustering algorithm directly to categorical data without transforming it.**



1. (10 points) Apply the nearest neighbor algorithm using the Euclidean distance. We will consider the number of neighbors from 1 to 15. What is the smallest number of neighbors that we should use to discover the clusters correctly? Remember that we may need to try a couple of values first and use the eigenvalue plot to validate our choice.

**Answer :**

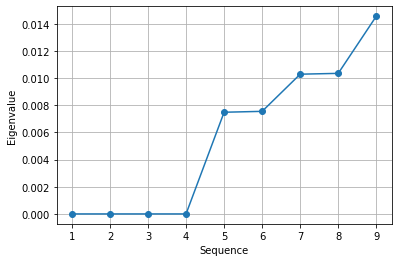
**The smallest number of neighbors that we should use to discover the clusters correctly: 10**



1. (5 points) Using your choice of the number of neighbors in (c), calculate the Adjacency matrix, the Degree matrix, and finally the Laplacian matrix. How many eigenvalues do you determine are practically zero? Please display values of the “zero” eigenvalues in scientific notation.

**Answer :**

**Values of the “zero” eigenvalues in scientific notation: [-7.33258899e-16 2.84331459e-16 1.28367511e-15 1.40853702e-15]**



**Adajacency Matrix**

**[[1. 0. 0. ... 0. 0. 0. ]**

**[0. 1. 0. ... 0.86087673 0. 0. ]**

**[0. 0. 1. ... 0. 0.96602229 0. ]**

**...**

**[0. 0.86087673 0. ... 1. 0. 0. ]**

**[0. 0. 0.96602229 ... 0. 1. 0. ]**

**[0. 0. 0. ... 0. 0. 1. ]]**

**Degree Matrix**

**[[8.37508495 0. 0. ... 0. 0. 0. ]**

**[0. 9.08018606 0. ... 0. 0. 0. ]**

**[0. 0. 8.67243781 ... 0. 0. 0. ]**

**...**

**[0. 0. 0. ... 9.22996601 0. 0. ]**

**[0. 0. 0. ... 0. 7.80266431 0. ]**

**[0. 0. 0. ... 0. 0. 7.17690085]]**

**Laplacian matrix**

**[[ 7.37508495 0. 0. ... 0. 0.**

**0. ]**

**[ 0. 8.08018606 0. ... -0.86087673 0.**

**0. ]**

**[ 0. 0. 7.67243781 ... 0. -0.96602229**

**0. ]**

**...**

**[ 0. -0.86087673 0. ... 8.22996601 0.**

**0. ]**

**[ 0. 0. -0.96602229 ... 0. 6.80266431**

**0. ]**

**[ 0. 0. 0. ... 0. 0.**

**6.17690085]]**

1. (10 points) Apply the K-mean algorithm on the eigenvectors that correspond to your “practically” zero eigenvalues. The number of clusters is the number of your “practically” zero eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme.

**Answer :**

