**Assignment – Unsupervised learning:**

1. Which of the following is a common use of unsupervised clustering?

a) Detect outliers

b) Determine a best set of projection for supervised learning

c) Evaluate the likely performance of a supervised learner model

d) Determine if meaningful relationships can be found in a dataset

e) All of the above

2. Which statement is true about the K-Means algorithm?

a) All attribute values must be categorical

b) The output attribute must be categorical

c) Attribute values may be either categorical or numeric

d) All attributes must be numeric

K-means algorithm defines an objective function which computes Euclidean distance between 2 points. It is not possible to find Euclidean distance for categorical values. So, all attributes must be numeric when we use K-Means algorithm.

3. Amongst below data transformation technique which works well when minimum and maximum values for a real-valued attribute are known.

a) min-max normalization

b) decimal scaling

c) z-score normalization

d) logarithmic normalization

4. This technique uses mean and standard deviation scores to transform real-valued attributes.

a) decimal scaling

b) min-max normalization

c) z-score normalization

d) logarithmic normalization

5. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.

a) agglomerative clustering

b) conceptual clustering

c) K-Means clustering

d) expectation maximization

6. What is the minimum no. of variables/features required to perform clustering?

a) 0

b) 1

c) 2

d) 3

7. Which of the following algorithm is most sensitive to outliers?

a) K-means clustering algorithm

b) K-medians clustering algorithm

c) K-modes clustering algorithm

d) K-medoids clustering algorithm

8. The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA).

1. PCA is an unsupervised method

2. It searches for the directions that data have the largest variance

3. Maximum number of principal components <= number of features

4. All principal components are orthogonal to each other

Which is above is true.

A. 1 and 2

B. 1 and 3

C. 2 and 3

D. 1, 2 and 3

E. 1,2 and 4

F. All of the above

9. Given historical weather records, can we predict if tomorrow's weather will be sunny or rainy using K-means. **FALSE**

10. Given a set of news articles from many different websites, using k-means can you find out what topics are the main topics covered. **TRUE**

11. Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model. **TRUE**

12. PCA can be used for projecting and visualizing data in lower dimensions. **TRUE**

**13. Pros and cons of :**

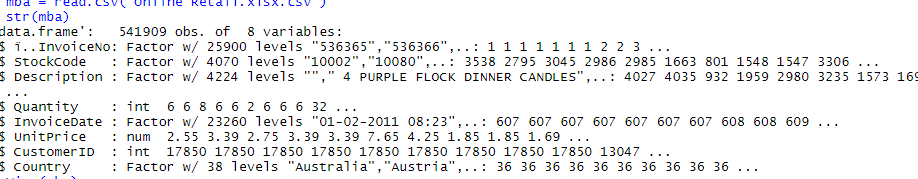
1. K-means clustering:
   1. Pros:
      1. Faster in computation.
      2. Easy to implement.
      3. Good at segmenting large dataset.
      4. Follows Expectation – Maximisation method which helps to adjust to changes quickly.
   2. Cons:
      1. Difficult to predict k value.
      2. Restricted to numeric data only.
      3. We have to specify the K-value at the beginning of algorithm which is difficult.
2. Scatter Plots:
   1. Pros:
      1. These are easy to draw.
      2. Can be easily understood and interpreted.
      3. Values of extreme variables do not affect this method as such points are isolated in the diagram.
      4. Can find outliers in a given dataset.
      5. Patterns between the dimensions can be shown with ease.
   2. Cons:
      1. Difficult to measure the exact amount of correlation.
      2. When decimal places are rounded off, measurements are not accurate enough.
3. Principal Component Analysis:
   1. Pros:
      1. Easy to find correlation of features for large datasets.
      2. As the number of features decrease, computational time reduces, in turn, performance of algorithm increases.
      3. It also helps in reducing overfitting problem.
   2. Cons:
      1. After principal components are formed, we cannot read the original data i.e., cannot interpret the original features.
      2. Sometimes, when principal components bias towards high variance, the results may be misinterpreted.
      3. Information can be lost sometimes if principal components are not chosen effectively.

**14. MARKET – BASKET ANALYSIS:**

From the given dataset, we have the details:

* InvoiceNo
* StockCode
* Description
* Quantity
* InvoiceDate
* UnitPrice
* CustomerID
* Country

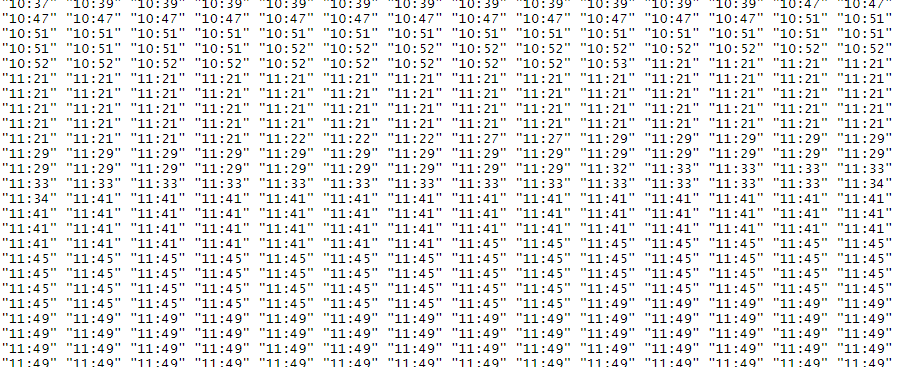
First, reading the given file and display the structure of the dataset. It is as follows:

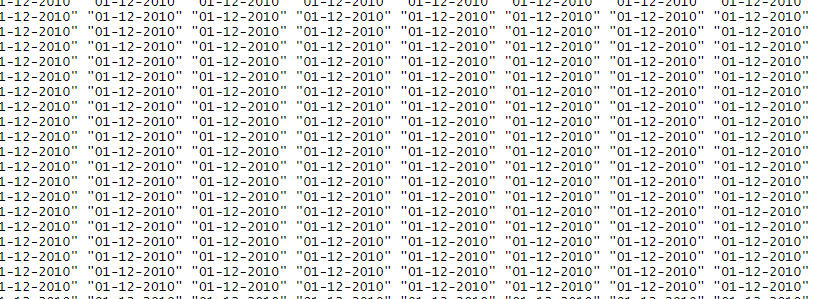


1. **What time do people often purchase online?**

To find this, we need the time of purchase of the customers. In the given dataset, we have “InvoiceDate” feature which contains date and time.

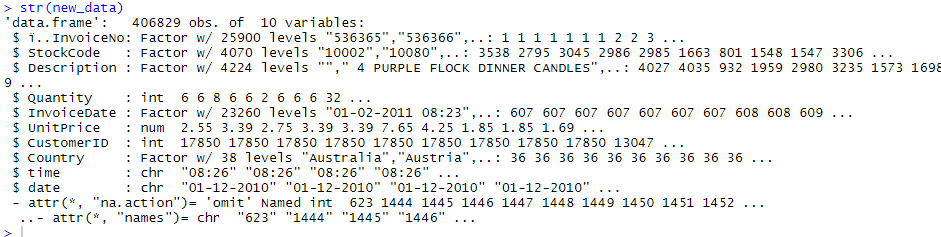
Now, extract the time and date from this feature.



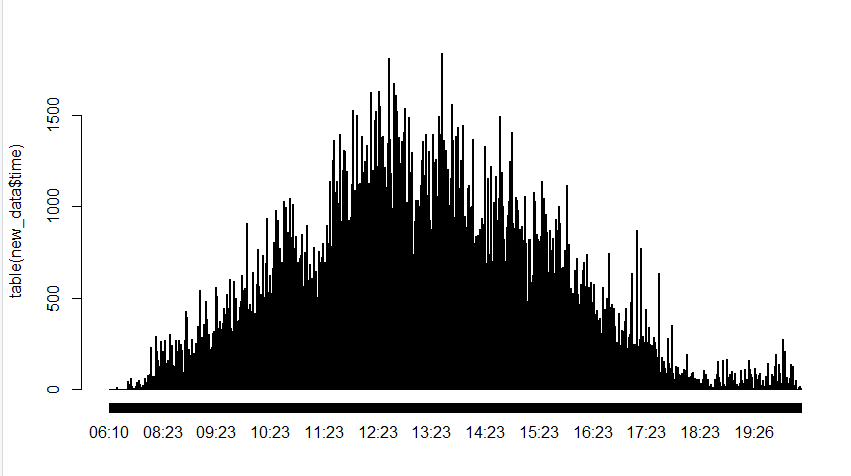


Remove the null values from the dataset.

The structure of new data is as follows:



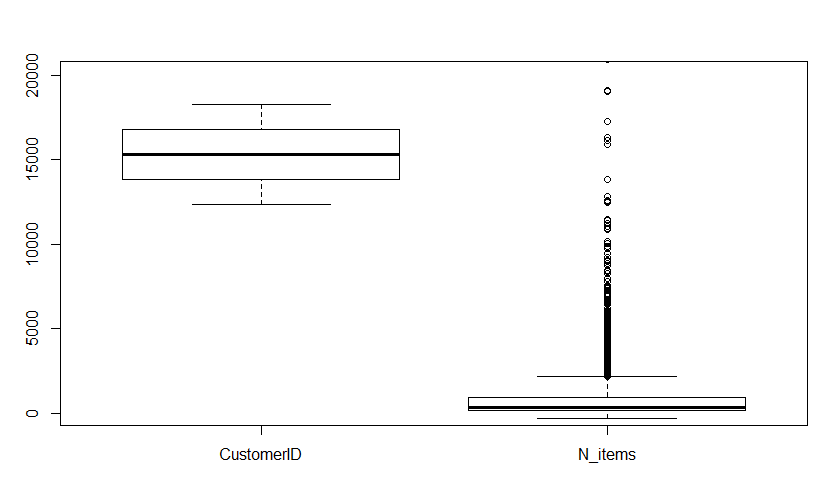
Now, form a table for the time in the dataset and find the time for which maximum number of customers are purchasing products online. The distribution of the above data is as follows:

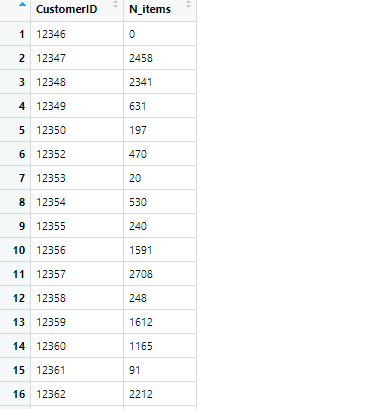


From the above plot, it is evident that most of the customers are buying products in the time range of 11:23 – 15:23. So, we can infer that from 11AM to 3PM people are often purchasing online.

**2.How many items each customer buy?**

To know how many items each customer bought, we have to know the details of the customer as well as the quantity of each customer.

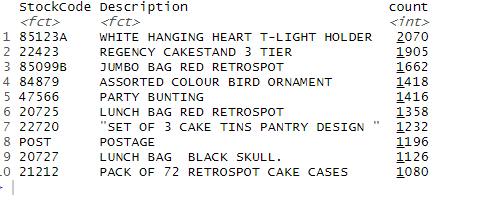




Totally, there are 4372 customers in this new data and each row represents the quantity of items bought by the customer. This is depicted in the box-plot.

**3.Top 10 best sellers:**

To find the top 10 best sellers, find the data based on stock code and description in the dataset and extract the top ten descriptions of sellers.



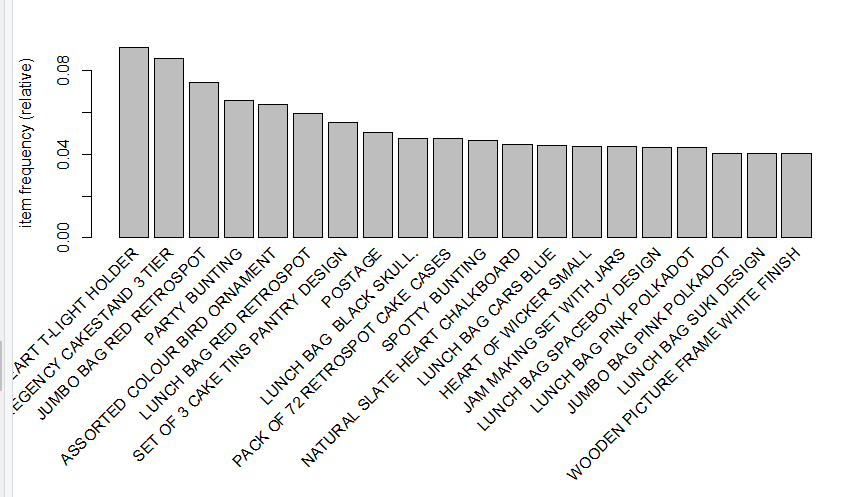
From the above table, we can find the top 10 best sellers based on the count value.

1. **To increase profits,**

Market-Basket analysis provides the results for the retailer to increase profits.

The results of the analysis is as follows:

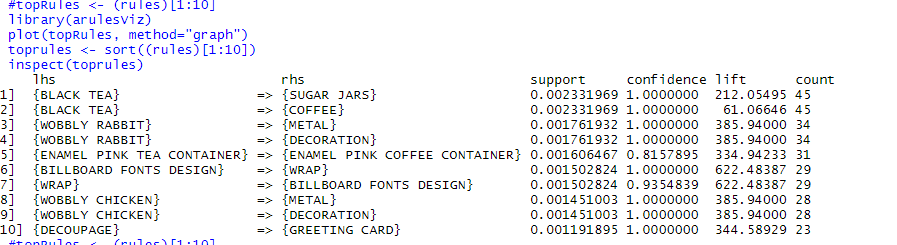
The most frequently bought items can be shown as:

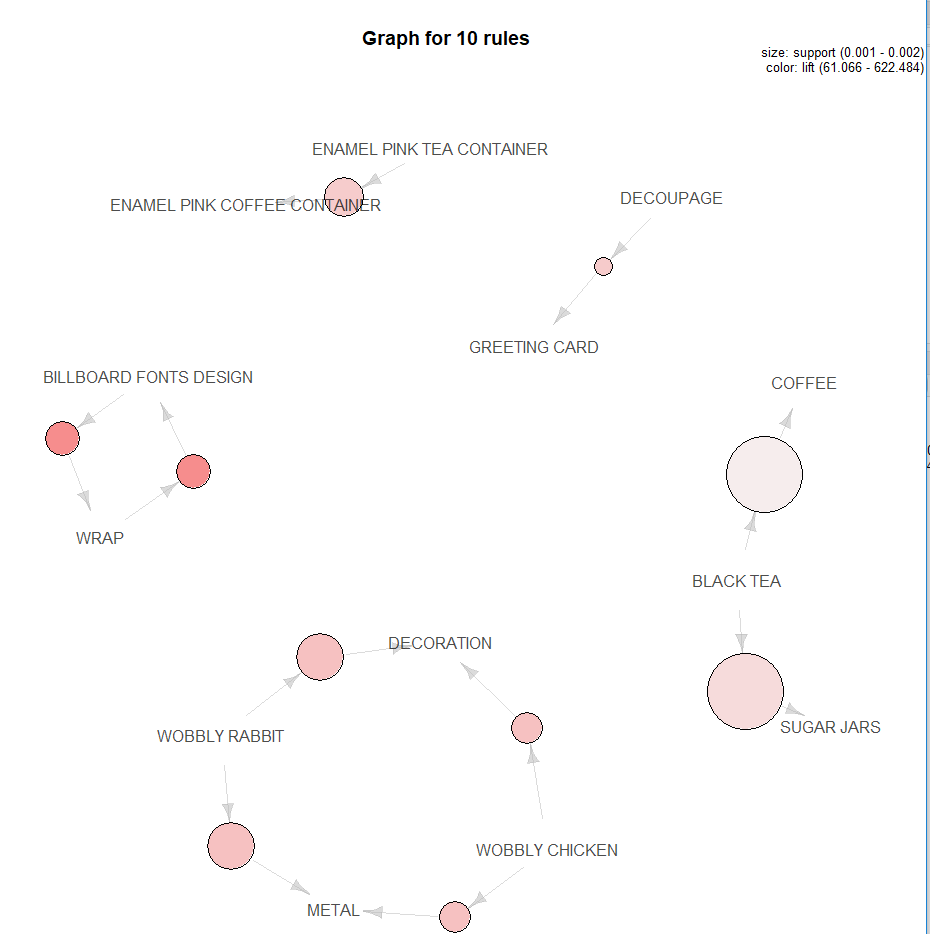


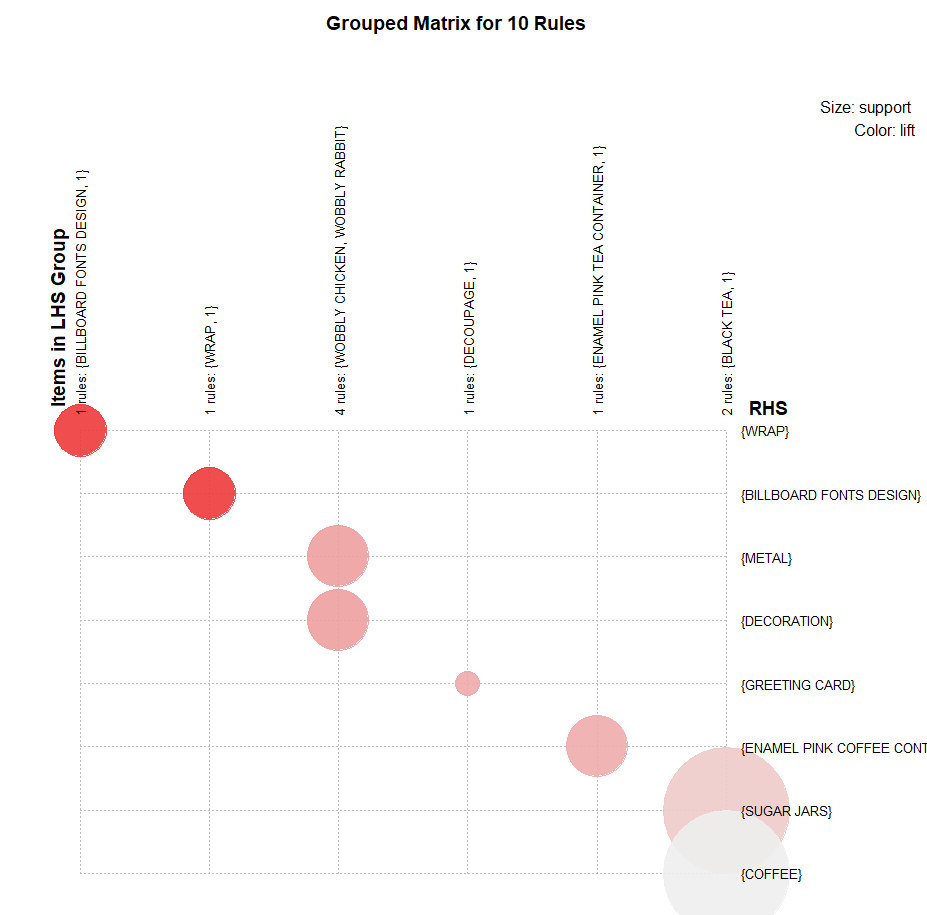
We can use apriori algorithm and search for frequent itemsets.

For the values ,support =0.001 and confidence =0.8, we get the association rules as follows:

When the data is sorted, top 10 association rules are :







So, for the retailer to increase the profits,

* Enamel pink tea and Enamel pink coffee containers
* Decoupage and greeting card
* Coffee, black tea and sugar jars
* Billboard fonts design and wrap
* Decoration, Wobbly rabbit, wobbly chicken and metal

These items must be kept together as the customers are frequently buying these items together. This arrangement gives retailer more profits.

Associate rules:

* { (Billboard fonts design), (wrap) } --🡪 2
* { (Wobbly rabbit) , (metal) } 🡪1
* { (Wobbly rabbit) , (decoration) } 🡪1
* { (black tea), (sugar jars) } 🡪 1
* { (black tea), (coffee) } 🡪 1
* { (Enamel pink tea container), (Enamel pink coffee container) } 🡪1
* { (Wobbly chicken) , (decoration) } 🡪1
* { (Wobbly rabbit) , (metal) } 🡪1
* { ( Decoupage), (greeting card) } 🡪1