**Data Mining Lab Report**

**(Problem 8 & 9)**

**Name: Md. Rahat Hasan**

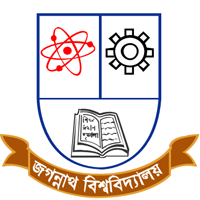
ID: B-160305019

Supervised by

**Dr. Md. Manowarul Islam**

Associate Professor

Department of Computer Science & Engineering



**Department of Computer Science and Engineering**

**Jagannath University, Dhaka-1100**

**Mall\_Customers**

# Introduction

Malls or shopping complexes are often indulged in the race to increase their customers and hence making huge profits. To achieve this task machine learning is being applied by many stores already. It is amazing to realize the fact that how machine learning can aid in such ambitions. The shopping complexes make use of their customers’ data and develop ML models to target the right ones. This not only increases sales but also makes the complexes efficient.

# Problem Statement

You are owing a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. Spending Score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data. The dataset “**Mall\_Customers.csv”** represents the information about the mall. So, we need to find out the following problem’s solution

1. Visualize male and female customer spending scores.
2. Identify similar group customers based on the attributes mentioned in the dataset. In other words, obtain insight into underlying patterns of different groups.

# The Dataset

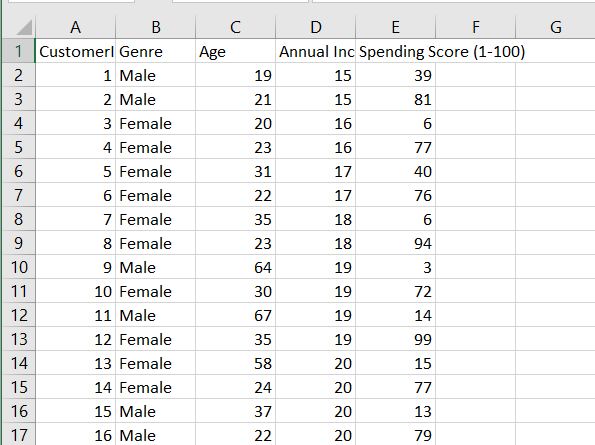


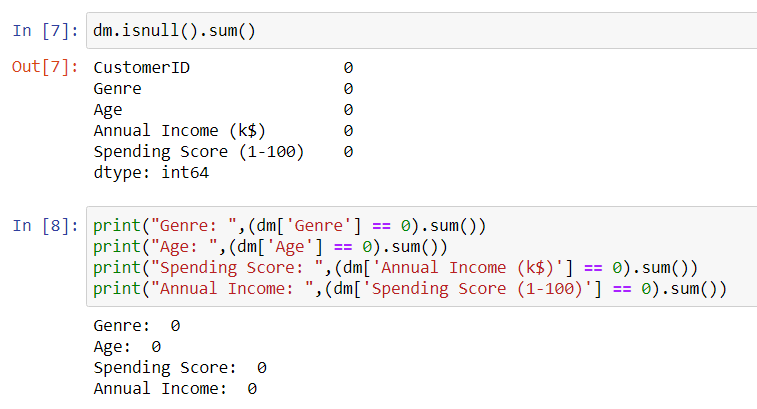
Fig :Dataset of the mall customers

* Here we have the following features :

1. CustomerID: It is the unique ID given to a customer
2. Gender: Gender of the customer
3. Age: The age of the customer
4. Annual Income(k$): It is the annual income of the customer
5. Spending Score: It is the score(out of 100) given to a customer by the mall authorities, based on the money spent and the behavior of the customer.

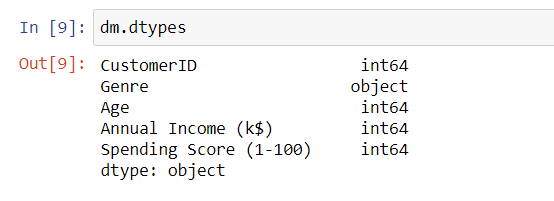
# Data preprocessing

**Checking Null values :**



* We have zero null values in any column. Sounds Good!

**Checking Categorical feature :**

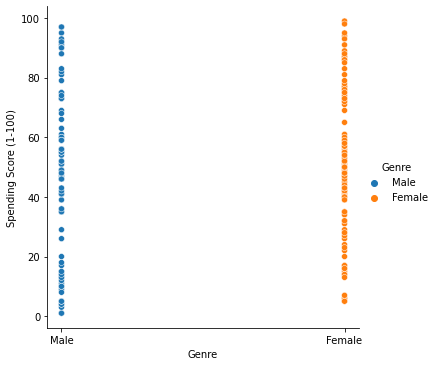


* We see that we have only one categorical feature: Gender.

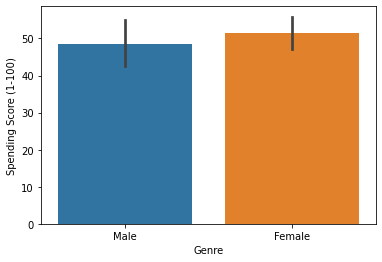
# Visualizing Male and Female customer spending scores

**Using Relational Plot :**

**Relational plots** are used for visualizing the statistical relationship between the data points. Visualization is necessary because it allows the human to see trends and patterns in the data. The process of understanding how the variables in the dataset relate each other and their relationships are termed as Statistical analysis.

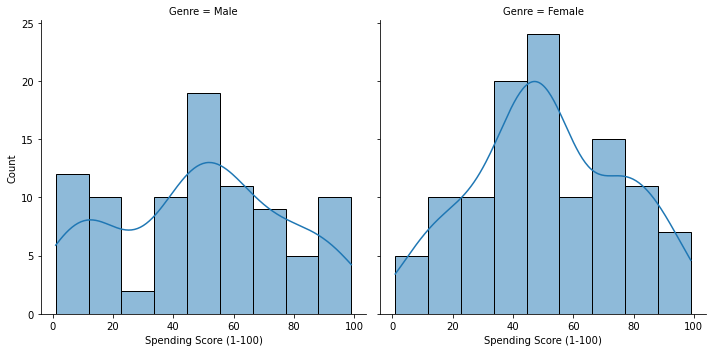


**Using Barplot :**

****

**Using Displot :**

Figure-level interface for drawing distribution plots onto a FacetGrid. This function provides access to several approaches for visualizing the univariate or bivariate distribution of data, including subsets of data defined by semantic mapping and faceting across multiple subplots.

****

After visualizing male and female customers spending scores using several plots, we can say that female customers spending score is slightly higher compared to male customers spending score.

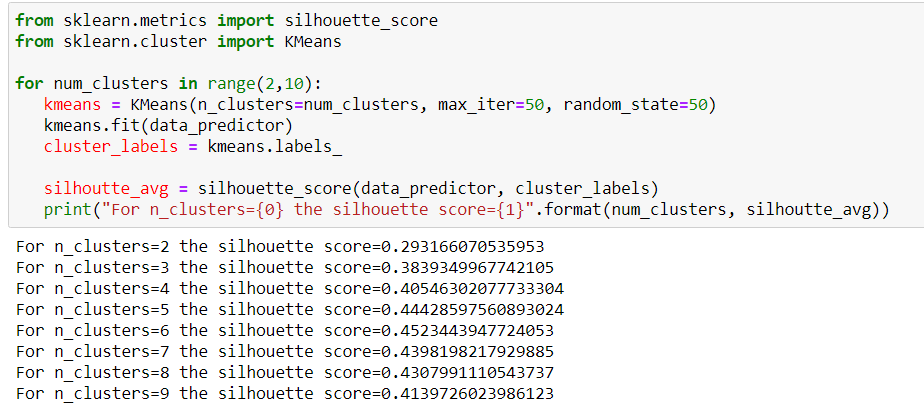
# Identifying similar group customers based on the attributes mentioned in the dataset

**Applying Algorithms & Methods**

Now the data preprocessing has been done and now let us move on to making the clustering model. I will use the K-Means Clustering algorithm to cluster the data. To implement K-Means clustering, we need to look at the Elbow Method.

**Selecting the number of clusters with silhouette analysis on KMeans clustering :**

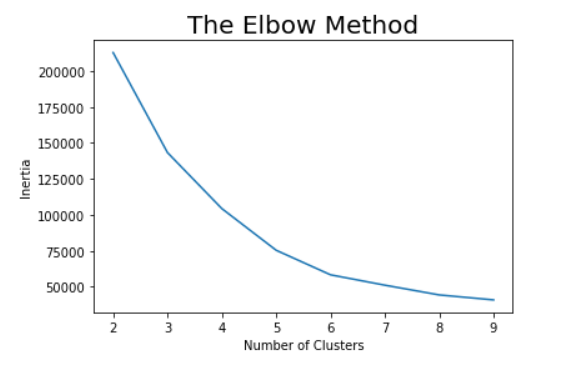
Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.



**Elbow method :**

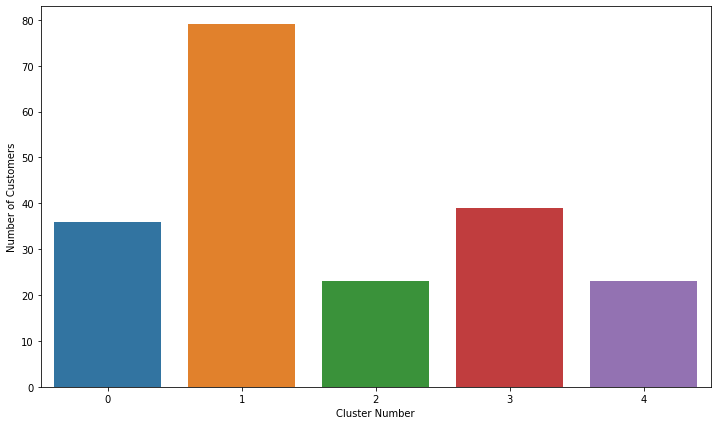
The Elbow method is a method of interpretation and validation of consistency within-cluster analysis designed to help to find the appropriate number of clusters in a dataset.

The following figure demonstrates the elbow method :

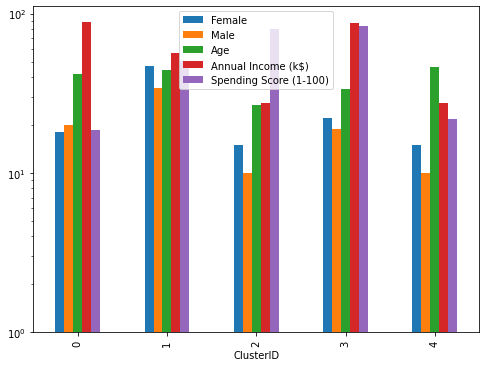


It is clear from the figure that we should take the number of clusters equal to 5, as the slope of the curve is not steep enough after it.

**Clusters IDs and Number of customers in each clusters:**

****

**K-means Clustering to identify similar group customers :**

****

# Analyzing the Results

We can see that the mall customers can be broadly grouped into 5 groups based on their purchases made in the mall.

In cluster 4, we can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less. The shops/mall will be least interested in people belonging to this cluster.

In cluster 2, we can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it’s because these people are more than satisfied with the mall services. The shops/malls might not target these people that effectively but still will not lose them.

In cluster 1,we see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall’s facilities.

In cluster 0, we see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall’s services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

Finally, based on our machine learning technique we may deduce that to increase the profits of the mall, the mall authorities should target people belonging to cluster 0 and cluster 4 and should also maintain its standards to keep the people belonging to cluster 1 and cluster 2 ,3 happy and satisfied.

To conclude, I would like to say that it is amazing to see how machine learning can be used in businesses to enhance profit.

**Grocery Store**

# Problem Statement

For the Data Mining lab where we had to execute algorithms like apriori, it was very difficult to get a small data set with only a few transactions. It was infeasible to run the algorithm with datasets containing over 10000 transactions. The dataset “**GroceryStoreDataSet.csv**” contains 11 items: JAM, MAGGI, SUGAR, COFFEE, CHEESE, TEA, BOURNVITA, CORNFLAKES, BREAD, BISCUIT and MILK consisting of 20 transactions.

1. Find most frequent itemsets that you will find in your experiment.
2. List all the association rules.

Assume minimum support = 0.01, confidence = 0.5; or you can set your own defined support and confidence threshold.

# Motivation

To find some more opportunities and more such products that can be tied together, the sales guy analyzed all sales records. What he found was intriguing. Many customers who purchased diapers also bought beers. The two products are obviously unrelated, so he decided to dig deeper. He found that raising kids is grueling. And to relieve stress, parents imprudently decided to buy beer. He paired diapers with beers and the sales escalated. This is a perfect example of Association Rules in data mining. This article takes you through a beginner’s level explanation of Apriori algorithm in data mining. We will also look at the definition of association rules.

# Dataset

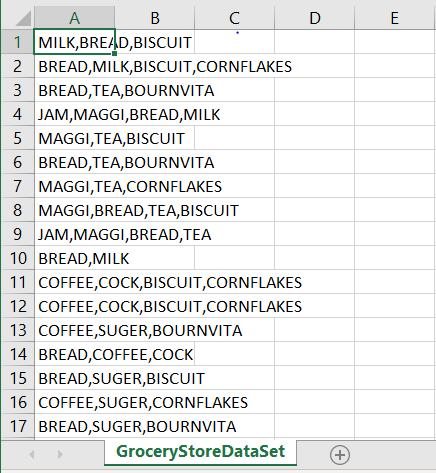
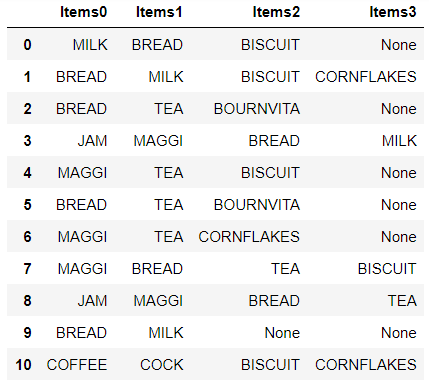


Fig : Dataset of the grocery store.

# Splitting the Dataset



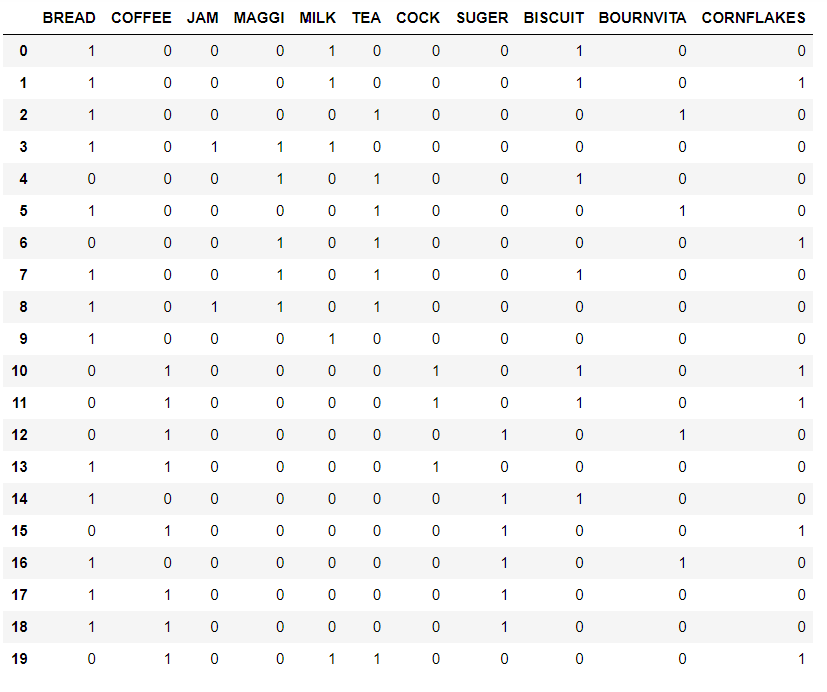
* As all the data was combined together , I have splitted the data into several attributes.

# One-Hot Encoding

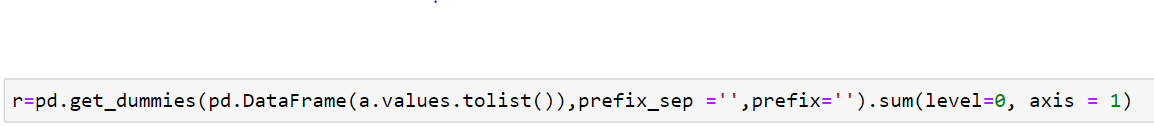
For categorical variables where no such ordinal relationship exists, the integer encoding is not enough.

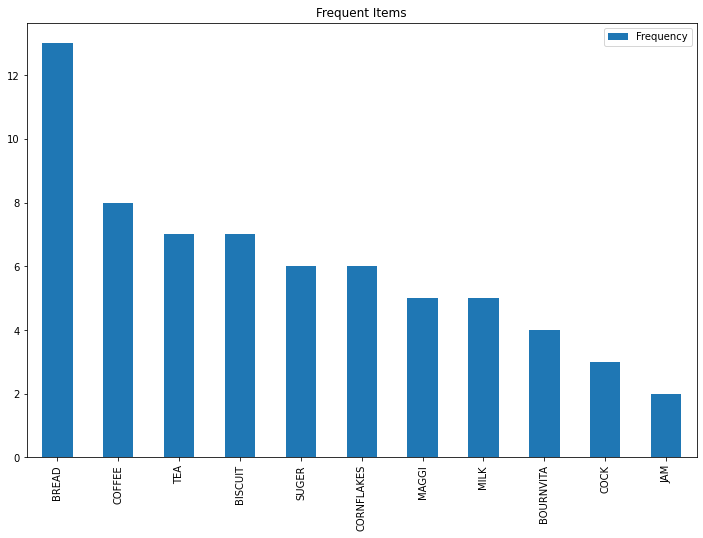
In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

In this case, a one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.



# Finding most frequent itemsets





# Applying Apriori Algorithm

With the quick growth in e-commerce applications, there is an accumulation vast quantity of data in months not in years. Data Mining, also known as Knowledge Discovery in Databases(KDD), to find anomalies, correlations, patterns, and trends to predict outcomes.

Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

It is very important for effective Market Basket Analysis and it helps the customers in purchasing their items with more ease which increases the sales of the markets. It has also been used in the field of healthcare for the detection of adverse drug reactions. It produces association rules that indicates what all combinations of medications and patient characteristics lead to ADRs.

# Association rules

Association rule learning is a prominent and a well-explored method for determining relations among variables in large databases.

Let I={i1,i2,i3,…,in} be a set of n attributes called items and D={t1,t2,…,tn} be the set of transactions. It is called database. Every transaction, ti in D has a unique transaction ID, and it consists of a subset of itemsets in I.  
A rule can be defined as an implication, X⟶Y where X and Y are subsets of I(X,Y⊆I), and they have no element in common, i.e., X∩Y. X and Y are the antecedent and the consequent of the rule, respectively. There are three common ways to measure association.

### **Support**

The support of an itemset X, supp(X) is the proportion of transaction in the database in which the item X appears. It signifies the popularity of an itemset.

supp(X)=Number of transaction in which Xappears / Total number of transactions.

In the example above, supp(Onion)=4/6=0.66667.

### **Confidence**

Confidence of a rule is defined as follows:

conf(X⟶Y)=supp(X∪Y)/supp(X)

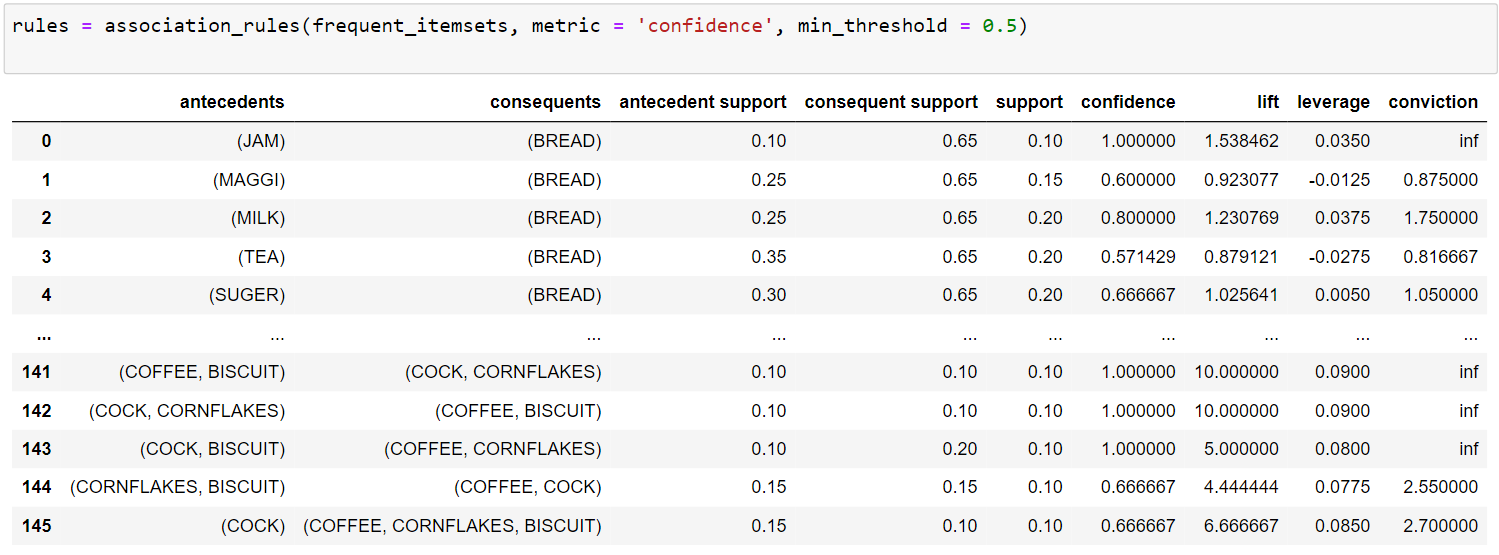
It signifies the likelihood of item Y being purchased when item X is purchased. So, for the rule {Onion, Potato} => {Burger},

### **Lift**

The lift of a rule is defined as:

lift(X⟶Y)=supp(X∪Y)/supp(X)∗supp(Y)

This signifies the likelihood of the itemset Ybeing purchased when item X is purchased while taking into account the popularity of Y.



# Analyzing results

If I consider number 2, The probability of Milk sales is 25% Milk and Bread are sold together in 20% of all purchases (support) 80% of customers who buy Milk will also buy Bread (confidence) Sales of Bread increased by 1.23 times in shopping with Milk (lift) Milk & Bread correlation with each other is seen as 1.75.