Data Warehousing

And

Business Intelligence

Kamaljeet Kaur Sidhu

Student ID 19489382

Module Code **7BUIS010W**

# Tasks

# Data Understanding:

* **Load the data to SQL:**

Text

Description automatically generated

**Output:** A sampleof five customer records is displayed below containing the Member number, Order date, Item name purchased that day as Item Description and Cost per item.

Graphical user interface

Description automatically generated with medium confidence

* **Distribution Analysis:**

Chart, line chart, histogram

Description automatically generated

* **Statistical Analysis:**

Statistical analysis of Member number and Cost is described below.

Graphical user interface, application

Description automatically generated

* **Correlation Analysis:**

Correlation among Member number and Cost variable is displayed below.

A picture containing application

Description automatically generated

* **Suitable transformation of variables:**

Table

Description automatically generated

It is clear from the Dataset info () that date is available as an Object datatype. But it needs to be in datetime format for having accurate analysis of data. So, the variable is transformed into required datatype using below code.



**Output:**

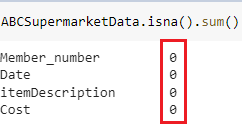


**Additionally:**

I have transformed the Date variable into Numeric values of the date. Because the **max (date)** function is not working on the normal date value. After using the numeric value of date column, I got the accurate results from the required query.

* **Management of Missing values:**

There are no missing values in the dataset as each customer has their unique Member number, shopping details. For your reference I have attached the output of missing value check code.



* **Data Cleaning:**

There are some duplicate records of customer’s visit to the store, so during cleaning all the 759 duplicate entries are removed from the dataset.

Graphical user interface, application

Description automatically generated

**Solution:**

A picture containing application

Description automatically generated

* **Data Visualisation:**

The charts describe the cost of 167 distinct items purchased by customers during the time period of two year (2014-2015).

# RFM Segmentation:

* **Definition of RFM Metrics:**

Diagram

Description automatically generated

RFM stands for Recency, Frequency and Monetary value.

**Recency** describes the recent date when the customer visited at the store. In the Business insights, it helps to describe about the freshness of the customer activity, be in purchases or visits.

**Frequency** refers to the total number of frequent visits of the customers at the store during a particular time period. The information got from the frequency metrics helps to answer that **how active a customer is?**

**Monetary value** defines the total amount a customer has spent in the store during a particular time period.

|  |  |  |
| --- | --- | --- |
| **Facts of RFM Metrics** | | **Information Usage** |
| **Recency** | **More recent visit** More likely to consider the business and brand in the future.  **Those have purchased months or longer period ago** the likelihood of future transactions is lower as compared to the recent customers. | * Target the recent customers to revisit the business and spend more. * Marketing efforts can help to convince users to revisit the store, while offering incentives or discounts to resume shopping. |
| **Frequency** | The factors such as, the type of an item, the cost of purchase and need for replacement may affect the frequency of a customer’s orders. | * The prediction of a customer’s purchase cycle can help to know when a customer needs to purchase more groceries. * Marketing strategies can be used to remind them to purchase their favourite items when its on discount or low price to increase the sales. |
| **Monetary Value** | Customers who spend more leads to increase the business profits. | * To produce a better Return on Investment, it is a good practice to focus on encouraging customers who spend more money to continue to do so in the future as well. |

* **Implementation in SQL:**

Text

Description automatically generated

**Output Analysis:**

The member number 1000 paid last visit at the store on 25th November 2015 (Recency). During the time period of two years (2014-2015), he visited the store for five times for grocery shopping and he spent 68.13 pounds in total.

**Find the attached dataset file.**

Table

Description automatically generated

The below query results a date to perform analysis. It filters the last date a customer has visited the store from the entire date column and add one day into it to get the difference of days. Based on the resulted date a difference can be calculate by subtracting the last visit date of a user from 31st December 2015. It will result, how many days ago a customer has visited at the store.

Text

Description automatically generated

Table

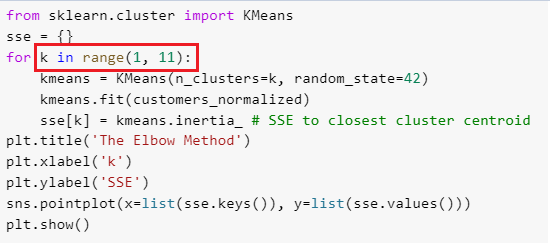
Description automatically generated

**Output:**

The abovementioned query results that Member number 1000 has purchased some items from the store 36 days ago.

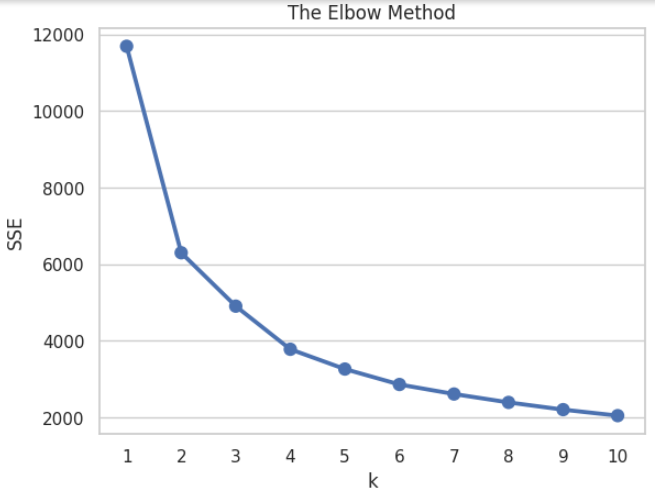
# Customer Segmentation with K-means:

* **Build K-means Model:**



A k-means model has been created by giving a range of values from (1-11) to K variable. The model will pass each number from this range to k continuously when the code is executed. The output of this piece of code will create an Elbowbased on the given values of K.

**Output:**

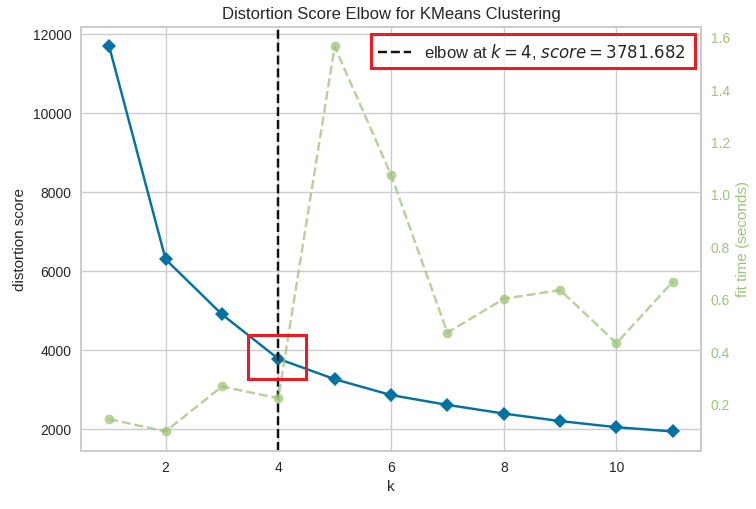


* **Correct justification of K value:**

In this elbow graph **SSE** stands for the sum of the squared Euclidean distance of each point to its closest centroid. However k means, the number of clusters. From the elbow joint point the optimal number of clusters has been taken.

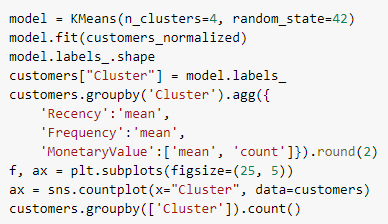
As this is the point where SSE value suddenly stops decreasing and helps to deliver best k value for clustering.

For further clarification KElbowVisualizer has been imported from Yellowbrick.cluster package to provide a visual representation of an optimal k value to form number of clusters.



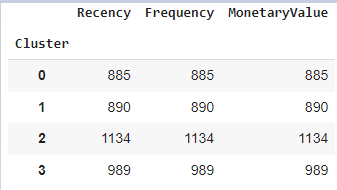
In this graph, a minimal drop has been noticed at 4. So, 4 is the number to be taken as number of clusters.

* **Testing of k means Model:**

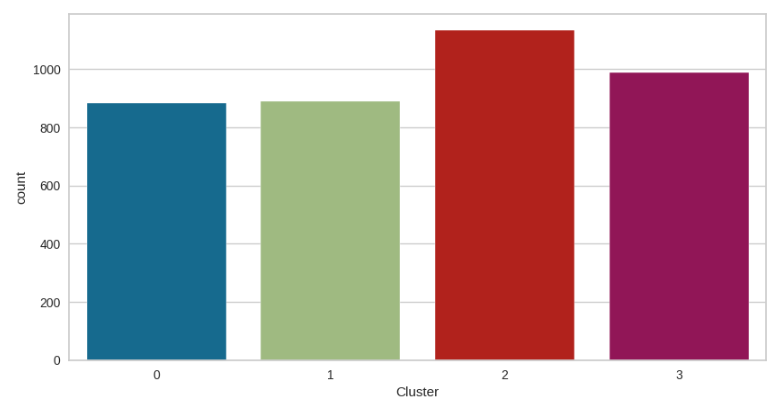


A k means model has been tested on cluster value of 4.

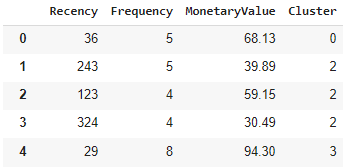
**Output:**



K means model has divided the total number of customers into 4 different clusters. The majority of customers lie in cluster ‘2’ followed by ‘3’,’1’ and ‘0’ in order.

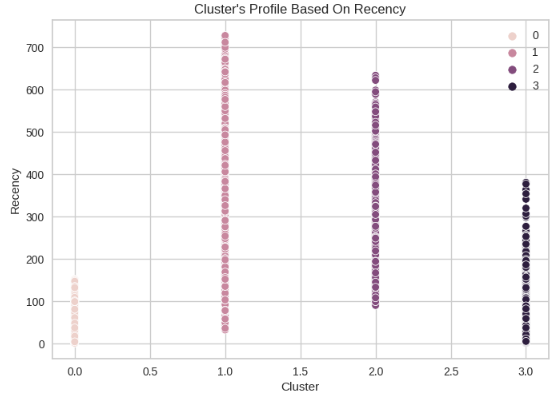


A sample of five customers has been attached below that displays the different customer number assigned to a particular customer.



**Visual representation:**

Further visual representation of each cluster has been displayed based on Recency, Frequency and MonetaryValue.



From the above illustration it is clear, the majority of customers who belong to cluster ‘1’ are the old customers. It means, it’s been a long time that they visited the store. However, cluster ‘0’ refers to the most recent customers.

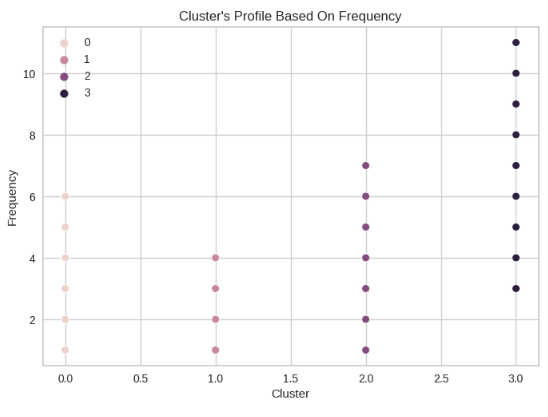
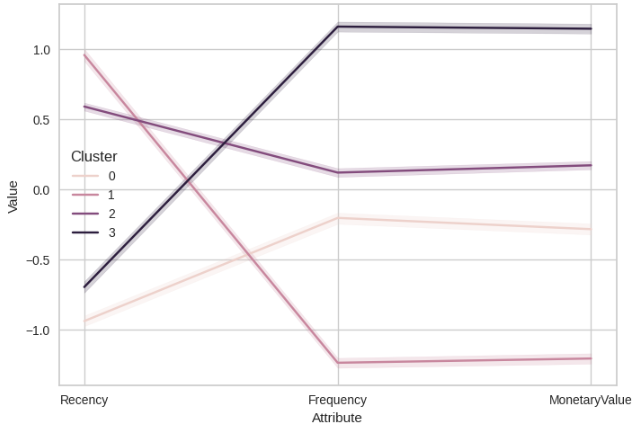


Illustration depicts the distribution of clusters based on Frequency. It is noticed that users categorised in cluster ‘3’ were more active. It means they visited the store frequently as compare to other users. While, those belong to cluster ‘1’ visited rarely.



This illustration depicts that the maximum monetary value that a store has earned is the result of purchases done by customers who belong to cluster ‘3’. On the other hand, users who lie in cluster ‘1’ had spent less money as compared to customers of other clusters.

**Final visual representation to discuss the profile and properties of each cluster**



|  |  |
| --- | --- |
| **Cluster Number** | **Properties of each Cluster** |
| 0 | Visited recently but do not visit often/frequently and spend less. |
| 1 | Visited long ago and do not visit often also spend very less. |
| 2 | Visited months ago but visit again after some time and do not spend considerable amount of money. |
| 3 | Visited few months ago and visit often as well as spend more. |

Based on the properties of each cluster they can be categorised into four different segments. It will help Business Analyst/Company Director/Marketers to make appropriate decision for growth of a business.

|  |  |
| --- | --- |
| **Cluster Number** | **Customer Segmentation based on Analysis** |
| 0 | Customer needing Attention |
| 1 | At Risk |
| 2 | Loyal |
| 3 | Best Customer |

# Review of Results:

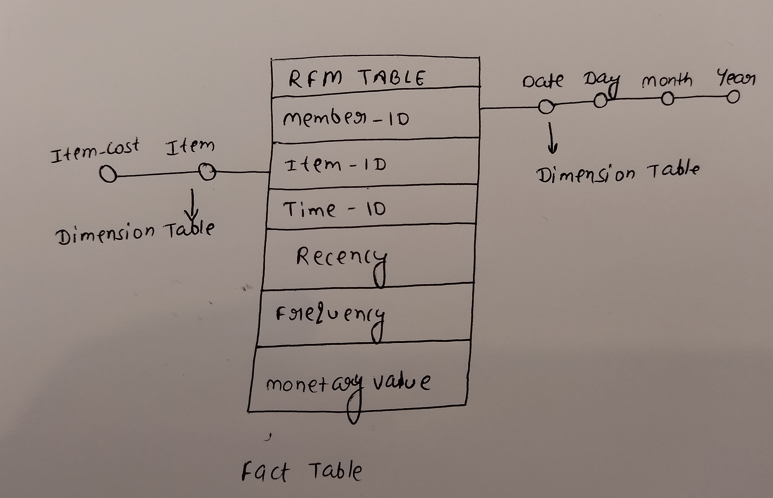
* **Identification of Business value Customer Segments:**

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **Customer Segments** | **Business value for Marketers** |
| 0 | Customer needing Attention | * Although, they have visited recently, but maybe it’s not their first choice. * Maybe, they have just tried the brand * They take some time to convince themselves to visit again. * They do not spend much maybe they think items do not worth how much they pay in the store. |
| 1 | At Risk | * They visit after more than a year of time or months. * They may do not like the brand or stock. That is why they appear after a long time. * They may not feel it’s worth spending in the store. Or they had bad experience with items or store services. |
| 2 | Loyal | * Although they had not appeared in the store for a year approximately. Still, they keep the brand in mind. * They do not visit frequently, either they shop excessively at one time or they visit somewhere else as well. * They do not spend much each time when they visit. |
| 3 | Best Customer | * Although they take months, but they visit the same store. * The comeback frequently because they are satisfied with the purchases. * They believe that the items available in the store are worth paying that is why they spend more. |

* **Correct Justification of their Business Value:**

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **Customer Segments** | **Correct Justification of Business Value** |
| 0 | Customer needing Attention | * **Maximize Customer Lifetime value** Create personalized marketing strategy to remind them about the business. It may convince them to purchase something. * **Increase Customer Loyalty** Get to know about their lifestyle, to think what affects them to take long time to visit again? Then, take actions accordingly. |
| 1 | At Risk | * **Maximize Customer Lifetime value** Arrange to conduct a survey to know what they think about price matrices of the items. Then, affordable products can be displayed to them to make more sales. * **Increase Customer Loyalty** Try to get insights about their store services and used products experience. Then, if there is any bad experience it can be converted into good one by fulfilling their needs. |
| 2 | Loyal | * **Increase Profit** Get to know about their interests and habits. Further focus on the target marketing to make more sales. * **Maximize Customer Lifetime value** Try to give offers on their favourite products so that they return back to same store frequently rather than visiting somewhere else. * **Increase Customer Loyalty** Focus on product quality so that they can select same store over others. |
| 3 | Best Customer | * **Increase Profit** Create personalized messages and Customized offers to make them feel high - value customers. Resultantly, they will purchase more because they are already satisfied with store services. * **Maximize Customer Lifetime value** Rewards and discounts for maximum purchase. * **Increase Customer Loyalty** Focus more on satisfying their needs. |

# **Data Mart Design**



* **Identification of Dimensions:**

There are two dimensioned mentioned in the Data Mart Design named as ’Item’ and ‘Date’.

* **Justification of selected Dimensions:**

**Item:** It refers to the name of different items that customers have purchased throughout the period of two years. Further the **‘Item-cost’** depicts the cost of each item present in the store. The item-cost is the dimensions that helps to calculate the total monetary value spend by a customer during two years.

**Date:** Date is the particular time period during the two years of time period when people visited the store. Each person has multiple entries in the store and it helps to analyse that how old a customer is. And, frequently he visits the store.

* **Identification of Measures:**

**‘Recency, Frequency and Monetary value’** are three measure in the fact table.

* **Justification of Measures:**

Recency refers to the last date when a customer visited the store. Frequency means that how many time they gave visited in the store in total till the end of 2015. Monetary Value refers to the total money that a customer has spent in the store.

**Find the attached GITHUB PYTHON file link.**

<https://github.com/kamalsidhu224/Datawarehouse-Coursework-files/blob/main/DWH_Coursework2.ipynb>