JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Customer Segmentation Using RFM Analysis: Realizing Through KMeans Clustering

[1]Kritika, [2]Dipak Kumar

[1]Student – M.S. in Business Analytics, [2]Student – M.S. in Business Analytics,
[1]RACE, REVA University, Bangalore India

Abstract : This study is an extension of paper "Customer Segmentation Using RFM Analysis: Realizing Through Python Implementation" by Dr. Ahmad M. A. Zamil and Dr. T. G. Vasista. Purpose: The purpose of this research paper is to segment the best contributed retail management customers of a super store and corresponding least contributed customers for a given transaction period based on the customer purchase behavior.

Design/Methodology/Approach: Theoretically Market segmentation is determined based on geographical segmentation and RFM analysis is adopted to be implemented for obtaining customer segmentation. Data mining technique called clustering technique is used to implement RFM analysis in order to identify the highly profitable, high-valued customers and low-risk customers. The secondary file containing secondary data is obtained from the source web site called Github.

Findings:(i) Segmenting customers based on their geographical residence is easy to work using SQL database and query (ii) segmenting customers based on their purchase behavior patterns of grocery items is easy to find using Orange Data Mining tool for visualization. Though there are prior researches conducted using Python libraries, however our approach is an effort of using SQL for generating RFM score and visually analyzing using KMeans clustering through Orange Data Mining Too.

Research Limitations/Implications: The development of modern patterns of grocery stores data analysis is limited to the case of Britain and has been described and analyzed through Python programming implementation. Though it is not possible to obtain patterns for all countries globally, however it is possible to obtain such patterns of some of the leading countries like UAE, Saudi Arabia and some countries from Europe etc.

Practical Implications: Customer Segmentation through RFM analysis technique implementation in Python provides an opportunity to analyze not only the Britain region but also for different countries is possible. RFM analysis technique that is mentioned theoretically can be well analyzed with little sql query and KMeans Clustering though GUI based Orange data mining tool. No familiarity with Python required to generate these graphs.

Social Implications: It helps grocery stores management in identifying profitable customers and thus can campaign various customer

***IndexTerms* - Keywords: Behavioral Segmentation, Clustering, Customer Segmentation, Electronic Customer Relationship Management, Geographical Customer Segmentation, Market Segmentation, RFM Analysis, SQL, Orange, KMeans Clustering.**

I. INTRODUCTION

RFM Analysis is a marketing technique used to quantitatively rank and segment customers based on their purchasing behavior. It stands for Recency, Frequency, and Monetary value:
Recency (R): How recently a customer made a purchase.
Frequency (F): How often a customer makes a purchase.
Monetary (M): How much money a customer spends on purchases.

RFM Analysis is significant in targeted marketing where it enables businesses to target specific customer segments with tailored marketing campaigns. For example, high RFM score customers can be targeted for loyalty programs.
Helps in Customer Retention as it Identifies at-risk customers (low Recency score) who haven't purchased recently, allowing businesses to re-engage them. It also helps in allocating marketing resources effectively by focusing on the most valuable customer

segments. Facilitates personalized communication strategies for different segments, enhancing customer experience. By focusing marketing efforts on the right segments, businesses can achieve higher returns on investment.

KMeans clustering is an effective mechanism to analyze RFM report and using Orange Data Mining tool, it enables easy visualization of different graphs.

.

Data and Sources of Data

Contemporarily technology based online retailing has reshaped the retail landscape. Shopping in Grocery stores is emerging as one of the fastest growing categories in the United Kingdom (Davies, Dolega & to more unknown similar item groups (Oliverira, 2012). The data for this research is leveraged from the base research paper "**Customer Segmentation Using RFM Analysis: Realizing Through Python Implementation**"

II. RESEARCH METHODOLOGY

The methodology section outline the plan and method that how the study is conducted. This includes creating SQLite Database from csv file, using SQL Query to generate RFM table. This is followed by KMeans Clustering Data analysis using Orange data mining tool.

Creating SQLite database

```
import pandas as pd
import numpy as np

# Convert 'InvoiceDate' to datetime, handling different formats
df_demo['InvoiceDate'] = pd.to_datetime(df_demo['InvoiceDate'], errors='coerce')

# Drop rows with NaT values
df_demo = df_demo.dropna(subset=['InvoiceDate'])

# Convert 'InvoiceDate' to Julian Day Number (JDN)
df_demo['JulianDayNumber'] = df_demo['InvoiceDate'].apply(lambda x: x.to_julian_date())

# Convert to integer
df_demo['JulianDayNumber'] = df_demo['JulianDayNumber'].astype(int)
```

```
## Creating a connection to SQLLIte database

db_path="C:\\Kritika\\BA14_Module03_Data Management and Visual Analytics(Dr. J B Simha)\\New Assignment_paper\\paper1.db"
#Establish a connection
conn = sqlite3.connect(db_path)
#Save the dataframe in the database
df_demo.to_sql("BASE",conn,if_exists = "replace",index = False)
```

Using SQL Query to generate RFM Score/Labels

```
select * from base;

select distinct CustomerID as CID from base group by CustomerID;

--delete the row where customerID is null
select CustomerID from base where CustomerID='NULL';

SELECT CustomerID
FROM base
WHERE CustomerID IS NULL
OR CustomerID = '';

DELETE FROM base
WHERE CustomerID IS NULL
OR CustomerID = '';

select * from base limit 50;

select distinct CustomerID as CID from base group by CustomerID;
```

---create a recency table

```
select CustomerID as CID,max(JulianDayNumber) as maxium_date from base group by CustomerID;

--create a table and then will manipulate the table
create table cp_r as select CustomerID as CID,max(JulianDayNumber) as maxium_date from base group by CustomerID;
--updating the table with the maxium transaction

select max(JulianDayNumber) from base ;

Alter table cp_r add column max_julian;
select * from cp_r;

UPDATE cp_r
SET max_julian = 2455904;

Alter table cp_r add column recency;

select * from cp_r;

UPDATE cp_r set recency=max_julian-maxium_date;

select * from cp_r group by CID;
```

--
---Frequency --

```
select CustomerID as CID,count(InvoiceNo) as frequency from base group by CustomerID;

create view cp_f as select CustomerID as CID,count(InvoiceNo) as frequency from base group by CustomerID;
```

----Monetization---

```
select distinct CustomerID as CID , UnitPrice*Quantity as Sales from base group by CustomerID;
--create table for monetization

create table cp_m as select distinct CustomerID as CID , UnitPrice*Quantity as Sales from base group by CustomerID;
--------------------------------------------------------------------
---- combine all the table in staging table

SELECT
t1.CID,
t1.recency,
t2.frequency,
```

```
t3.sales
FROM cp_r t1
JOIN cp_f t2 ON t1.CID = t2.CID
JOIN cp_m t3 ON t1.CID = t3.CID;


---------------------------------------------------------

--creating a RFM bin columns


create table RFM as SELECT
t1.CID,
t1.recency,
t2.frequency,
t3.sales
FROM cp_r t1
JOIN cp_f t2 ON t1.CID = t2.CID
JOIN cp_m t3 ON t1.CID = t3.CID;

-------------------------

select * from RFM;

----Get some metrics------------------

select max(recency) as max_recency,
min(recency) as min_recency,
avg(recency) as avg_recency,
max(frequency) as max_frequency,
min(frequency) as min_frequency,
avg(frequency) as avg_frequency,
max(sales) as max_m,
min(sales) as min_m,
avg(sales) as avg_m
from rfm;


-----------Alter the table to add to add column rbin,

alter table rfm add column rbin varchar(20);
alter table rfm add column fbin varchar(20);
alter table rfm add column mbin varchar(20);
--------------------------------Eyeball the data -----------------

select * from rfm limit 10;

-------------------Update the column----------

update rfm set rbin = 3 where recency between 248 and 373;
update rfm set rbin = 2 where recency between 124 and 248;
update rfm set rbin = 1 where recency <= 124;

select * from rfm limit 10;

----Update the fruency bin------
update rfm set fbin = 3 where frequency between 2297 and 3446;
update rfm set fbin = 2 where frequency between 1148 and 2297;
update rfm set fbin = 1 where frequency <= 1148;

select * from rfm limit 10;
---Update the monetary Bin-----------

update rfm set mbin = 3 where Sales between 111937.75 and 168469.6;
update rfm set mbin = 2 where Sales between 55405.88 and 111937.74;
update rfm set mbin = 1 where Sales <= 55405;

select * from rfm limit 10;
```

```
---create and update RFM label
alter table rfm add column label varchar(3);

update rfm set label = CAST(rbin AS TEXT) || CAST(fbin AS TEXT) || CAST(mbin AS TEXT);
select * from rfm limit 5;

----making label colmn as RFM---------------
SELECT
CAST(rbin AS TEXT) || CAST(fbin AS TEXT) || CAST(mbin AS TEXT) AS label
FROM
rfm;
---combining everthing on 1 column--------------------------
select CID,
CAST(rbin AS TEXT) || CAST(fbin AS TEXT) || CAST(mbin AS TEXT) as label,
recency, frequency, sales,
rbin, fbin, mbin
from rfm;
```

------------------**creating view** ------

```
create view view_rfm as
select CID,
CAST(rbin AS TEXT) || CAST(fbin AS TEXT) || CAST(mbin AS TEXT) as label,
recency, frequency, sales,
rbin, fbin, mbin
from rfm;
----------------------------------------------------------------------------
select * from view_rfm limit 5;
```

----**Generate RFM report for stats**---

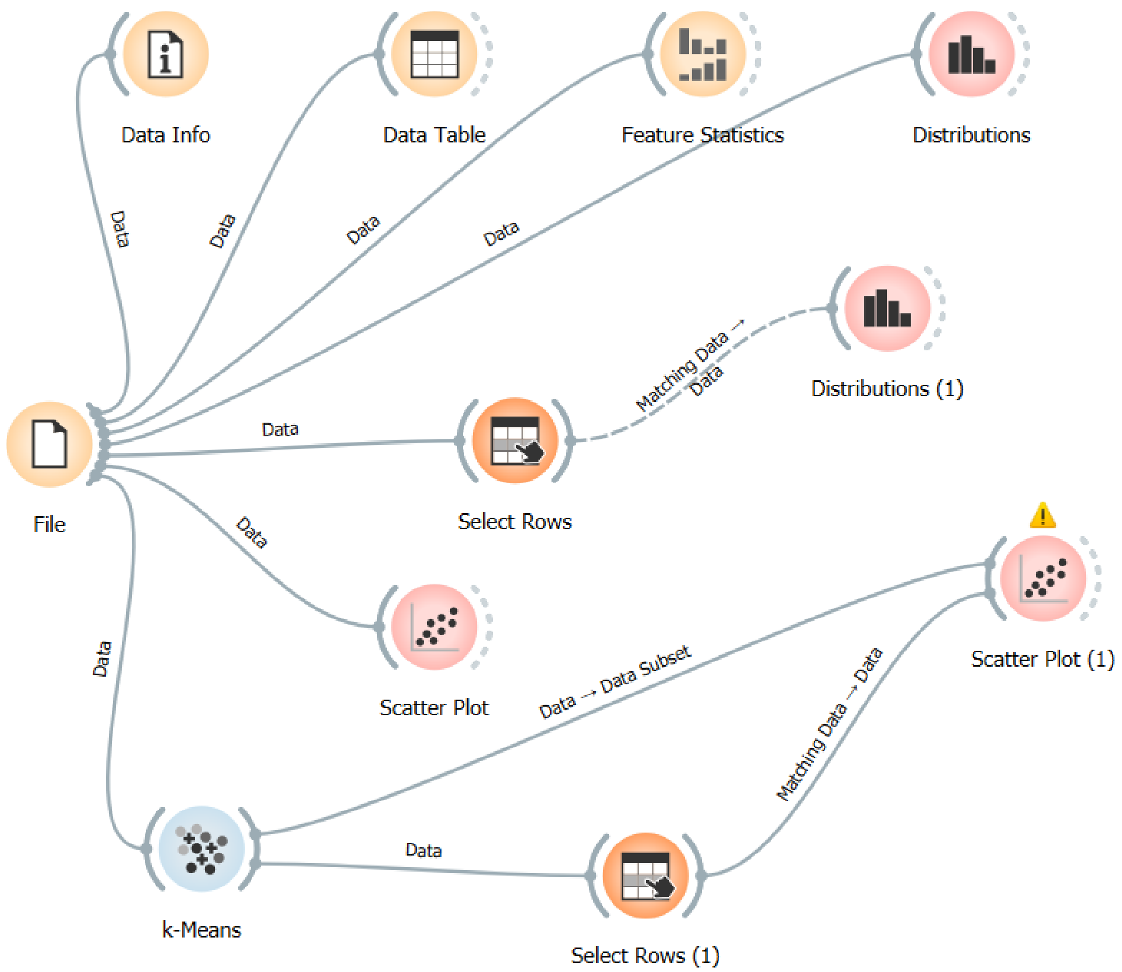
```
select label as "RFM Bin",
count(CID) as "Number of Customers",
sum(Sales) as Revenue,
avg(recency) as Avg_Recency,
avg(frequency) as Avg_Frequency,
avg(sales) as Avg_Revenue
from view_rfm
group by label;
```

------**create report view**--------------------

```
create view view_rfm_report as
select label as "RFM Bin",
count(CID) as "Number of Customers",
sum(Sales) as Revenue,
avg(recency) as Avg_Recency,
avg(frequency) as Avg_Frequency,
avg(sales) as Avg_Revenue
from view_rfm
group by label;

----top 3segment----------------

select * from view_rfm_report
order by Revenue desc limit 3;


-----Eyebal the data -----------
select * from view_rfm;


-----Alter the table with the country name ----
alter table RFM add column Country varchar(40);

---filling the country column----
```

```
UPDATE RFM
SET Country = BASE.Country
FROM BASE
WHERE RFM.CID = Base.CustomerID;
```

Means Custering using Orange



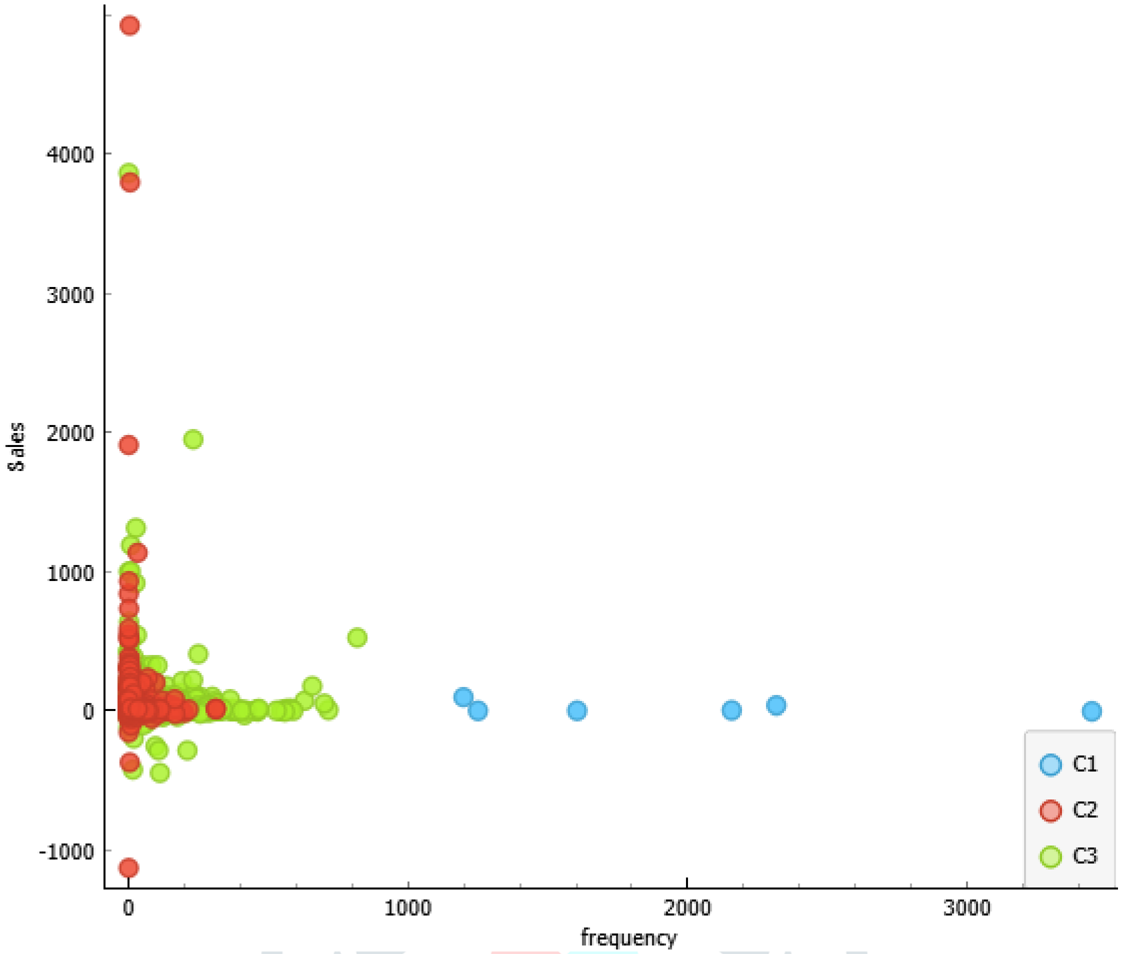


FIGURE 1

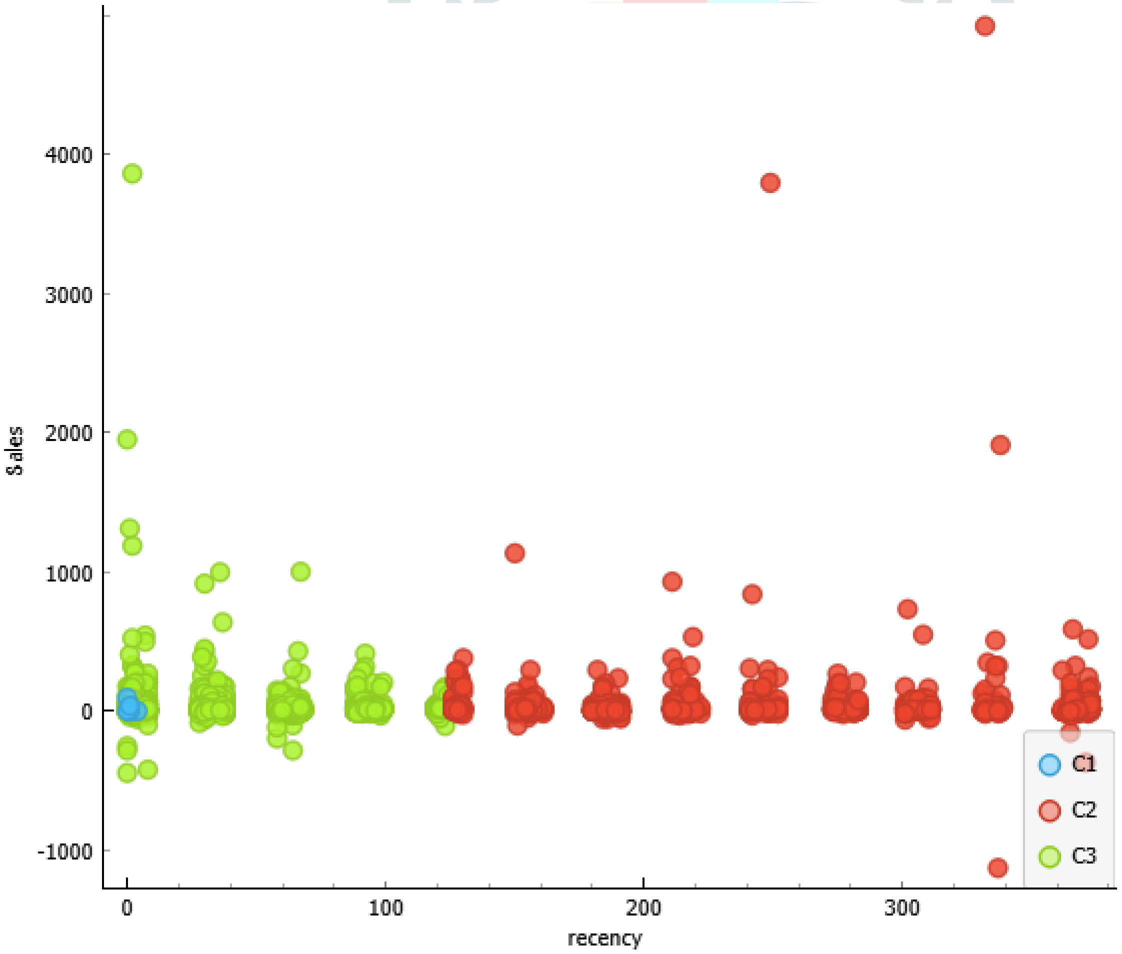


Figure 2

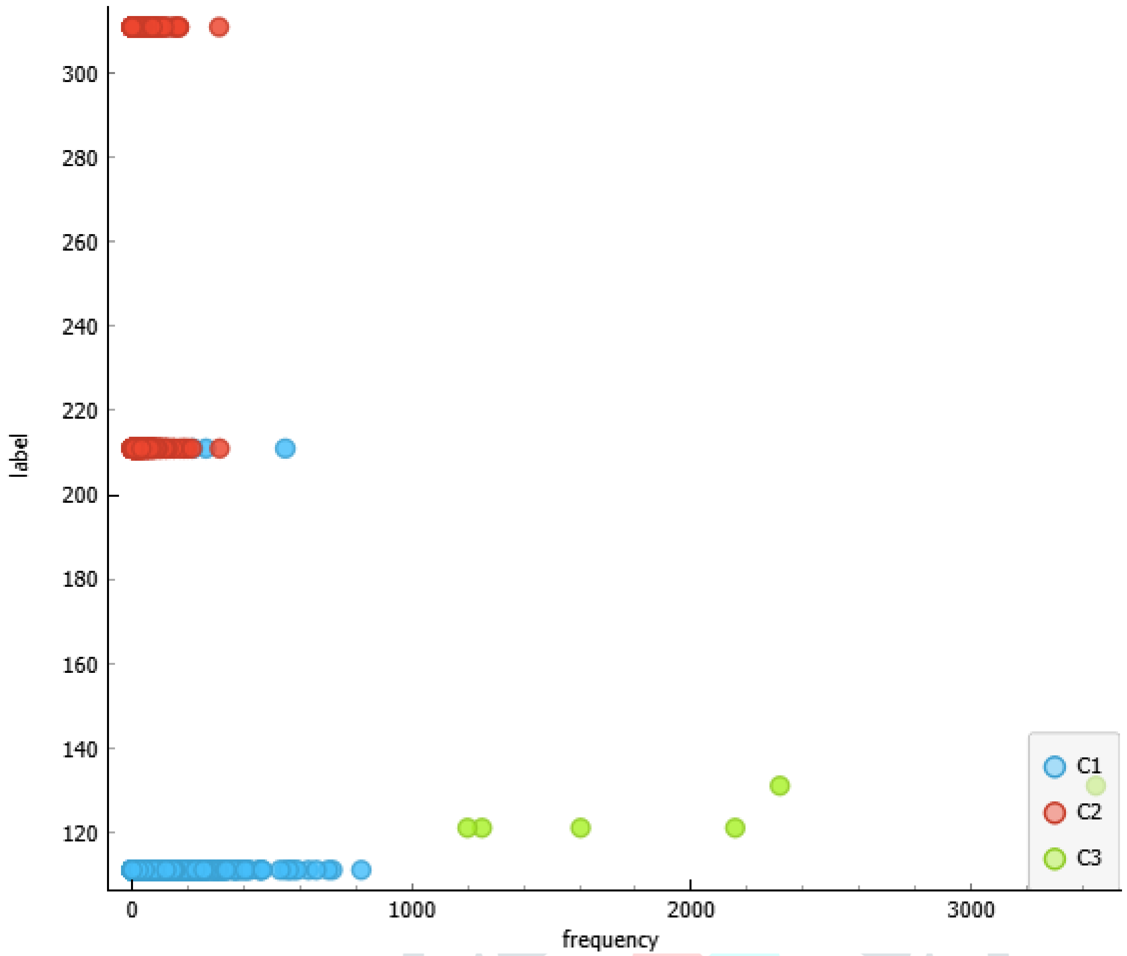
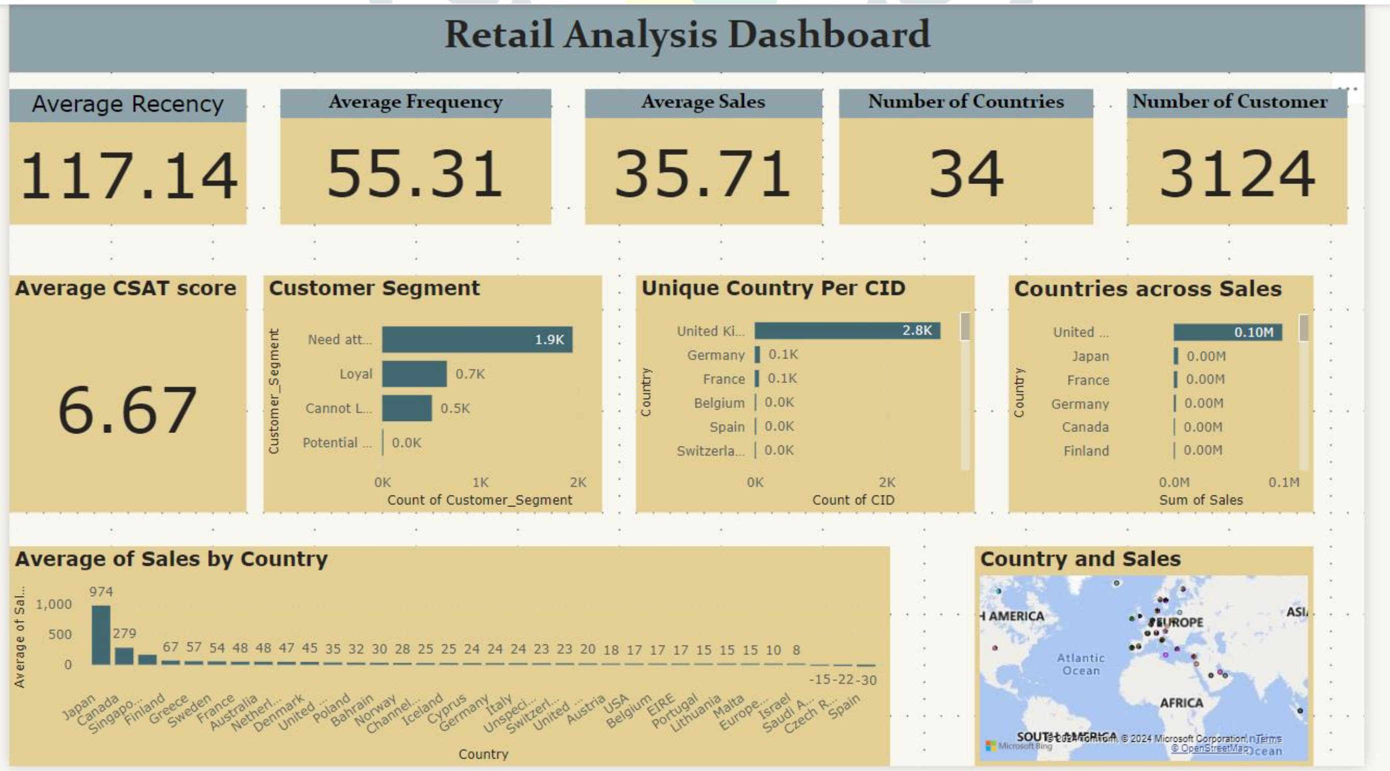
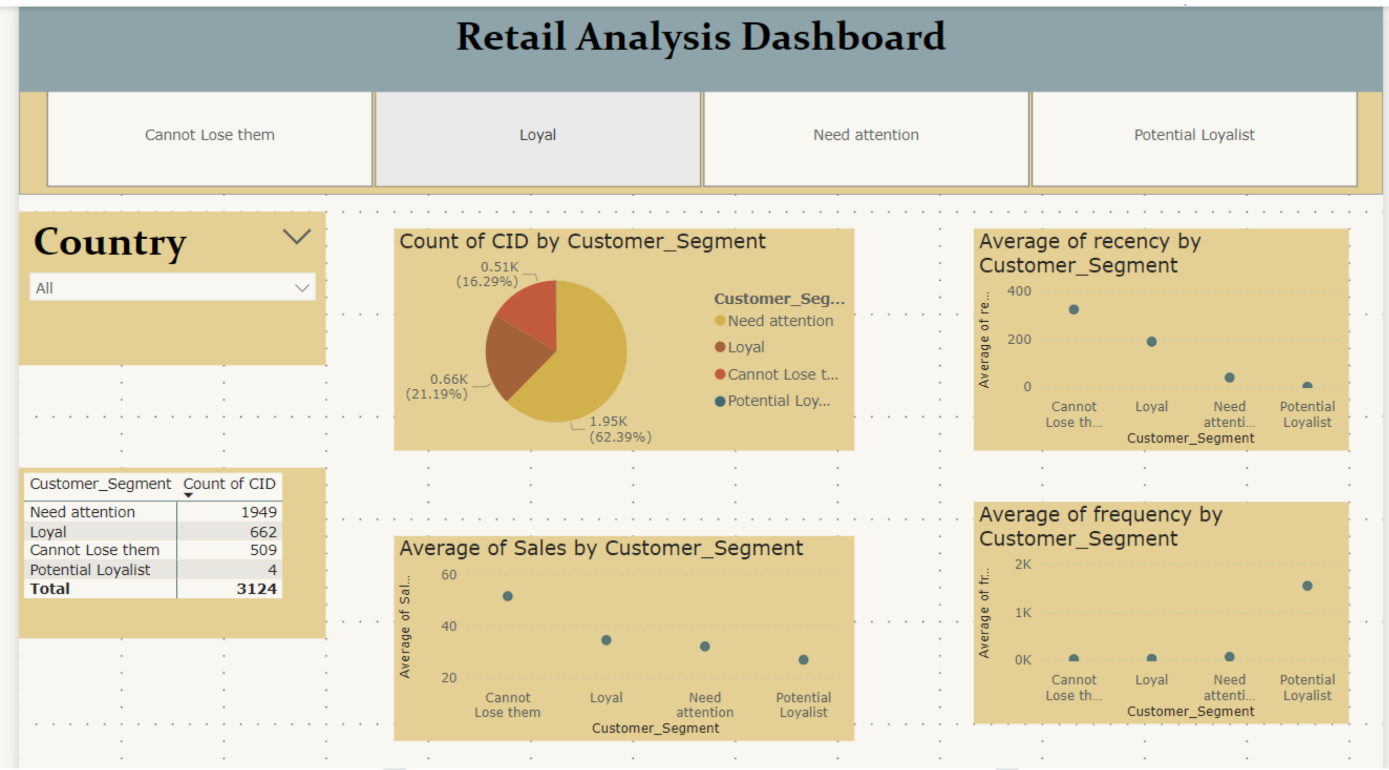


Figure 3

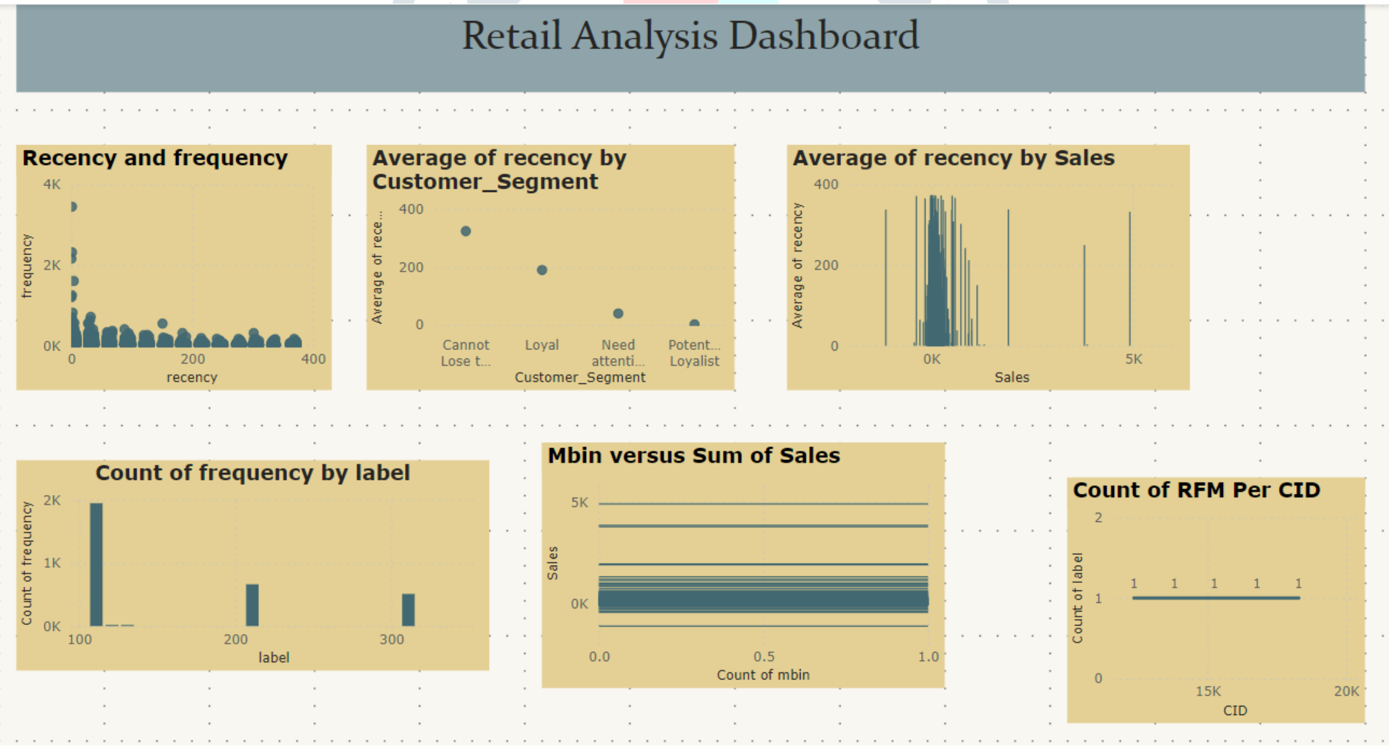
Dash board presentation using Power BI



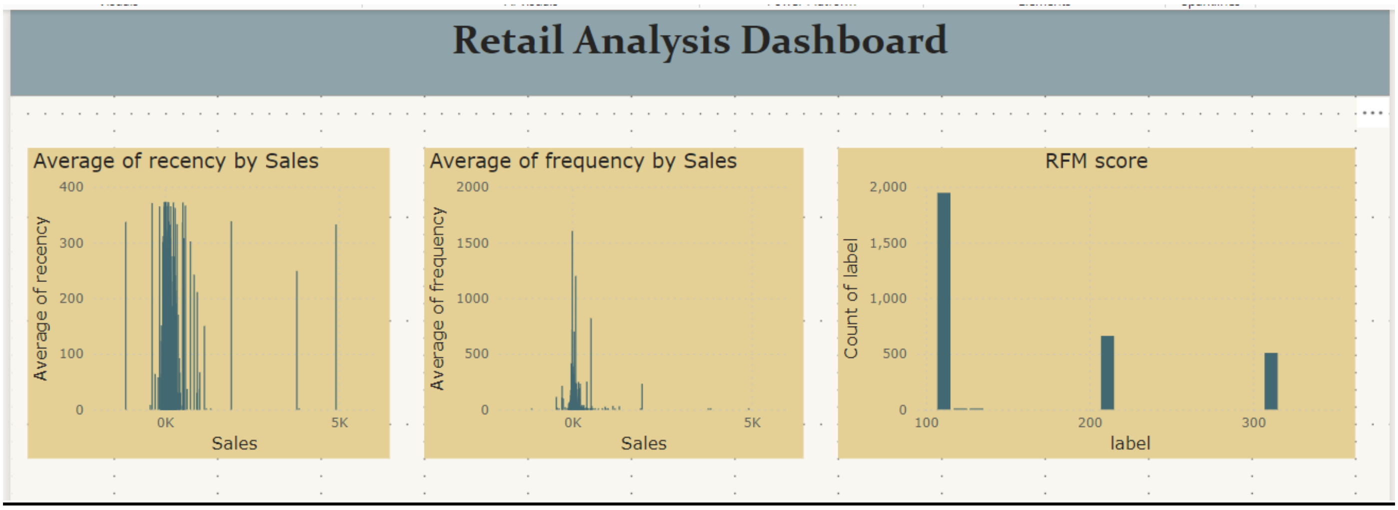
Customer Information



Segmentation of Customer's



RFM Analysis



Sales comparison using Recency , frequency and RFM score

III. CONCLUSION

From the above graph based on use case analysis, customers with frequency lower than a threshold (e.g. 1000 in the above graph) and with a higher label score can be a target of further marketing and boosting sales.

Sales is least impacted by frequency and recency. Across all 34 Country, Average of Sales ranged from -29.81 to 974.49.

The top 4 Sales all had Average of recency of 373.00.

The Average CSAT score for the above customer analysis is 6.67.

Consumer behavior understanding is crucial for developing loyalty programs that add significant value.

Analyzing consumer behavior allows businesses to create tailored loyalty programs that cater specifically to the needs and preferences of their customers. This can enhance customer satisfaction and retention.

By leveraging big data and analytics, businesses can gain insights into customer purchasing patterns, preferences, and behaviors. This information can be used to segment customers into different categories and create personalized offers that resonate with each segment, thereby enhancing customer loyalty.

In summary, understanding consumer behavior and using advanced data analytics for segmentation and personalization are key strategies for building effective loyalty programs.

REFERENCES

[1] Customer Segmentation Using RFM Analysis: Realizing Through Python Implementation" by Dr. Ahmad M. A. Zamil and Dr. T. G. Vasista