THE CURIOSITY CUP 2024

A Global SAS® Student Competition

Team Name: Segmint Masters

Unveiling Customer Behavior Insights: An RFM Analysis of Online Retail Data from archive.ics.uci.edu Using SAS

Youssef Azam, Amany Gaber, and Ali Rabie, Beni-suef & Alexandria University

ABSTRACT

This paper outlines a methodology for analyzing a dataset that has about 1,067,373 transactions from the Online Retail database from archive.ics.uci.edu. it represents transactions of a UK-based online retail company specializing in all-occasion gift-ware, primarily dealing with wholesalers. The main goal is to analyze customer behavior, and factors affecting sales performance to identify customer segments and propose strategies for increasing revenue. The insights from this analysis are good for improving marketing strategies, maximizing revenue, and enhancing overall performance in online retail.

INTRODUCTION

In the world of e-commerce, the RFM (Recency, Frequency, Monetary) model constitutes an effective strategy for understanding customer behavior. This model involves analyzing customer transactional data to segment them based on their behavior. By applying the RFM model and data analysis, companies can gain useful insights into customer behavior, which also leads to making strategic decisions to improve customer experience and increase revenue.

In this study, we'll analyze a dataset from a UK-based commercial company, encompassing purchase details such as invoices, product codes, quantities, dates, prices, customer IDs, and countries. Our analysis aims to understand customer behavior, identify sales drivers, and highlight top revenue contributors, given a total revenue of 17,000,000. We'll also employ data-driven insights, particularly from the RFM model, to tailor targeted marketing campaigns, thereby improving customer understanding and enhancing marketing effectiveness to meet business goals.

METHODOLOGY

The methodology employed in this study encompasses a five-step approach to analyze an Online Retail dataset. Subsequent sections will elucidate these stages in detail.

IMPORTING THE DATA

importing a CSV file by defining the file name and using PROC IMPORT to import it as CSV into "WORK" library named "IMPORT":

```
FILENAME REFFILE '/home/u63340110/sasuser.v94/OnlineRetailf.csv';
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.IMPORT REPLACE;
    GETNAMES=YES;
RUN;
```

DATA PREPROCESSING

Preprocessing very useful for data quality leads to better analysis. Starting by removing unnecessary columns named "StockCode" and "Description". Next, eliminating rows with missing values in the "CustomerID" column to ensure data completeness:

```
data df;
    set WORK.IMPORT;
    drop StockCode Description;
run;
data df_dropna;
    set df;
    where not missing(CustomerID);
run;
```

Next, extracting year, month, and day from "InvoiceDate" and ensures data integrity by converting negative values in "UnitPrice" and "Quantity" into positive ones:

```
   data df_dropna;
       set df_dropna;
       Year=year(InvoiceDate);
       Month=month(InvoiceDate);
       Day=day(InvoiceDate);
       if UnitPrice < 0 then UnitPrice = -1 * UnitPrice;
       if Quantity < 0 then Quantity = -1 * Quantity;
run;
```

CALCULATING RFM MATRIX

Calculating Recency (R) starting by retrieving the last transaction date for each customer and sets a reference date. Then, calculating Recency by getting the difference between each customer's last transaction date and the reference date:

```
proc sql;
    create table last_transaction_date as
        select CustomerID, max(InvoiceDate) as LastTransactionDate
format=date9.
        from df_dropna group by CustomerID; quit;
%let reference_date = '30DEC2011'd;
data RFM;
    set last_transaction_date;
    Recency=intck('day', LastTransactionDate, &reference_date.);
    format Recency Best.;
run;
```

Calculating Frequency (F) by using PROC SQL to create a table named "Frequency" from the cleaned dataset and grouping the data by Customer ID and counting the distinct number of invoices, it tells us how often customers interact with the company or make purchases, giving great insights into their behavior and preferences:

```
proc sql;
    create table Frequency as
        select CustomerID, count(InvoiceDate) as Frequency
        from df_dropna group by CustomerID; quit;
```

Calculating Monitory (M) by creating a table named "MonetaryValue" where the total monetary value of purchases, calculated as the sum of Quantity multiplied by Unit Price, this for each customer from the dataset. This helps assess the contribution of each customer to the business, to ease segmentation and targeted marketing strategies:

```
proc sql;
    create table MonetaryValue as
        select CustomerID, sum(Quantity * UnitPrice) as MonetaryValue
        from df_dropna group by CustomerID; quit;
```

SEGMENT CUSTOMERS

calculating Ranks of segments:

```
proc rank data=RFM out=RFM_ranked groups=6 ties=low;
    var Recency Frequency MonetaryValue;
    ranks R_Quartile F_Quartile M_Quartile;
run;
```

Segmentation step combines quartile ranks to segment customers based on RFM scores:

- By creating a dataset named "RFM_segments".
- Segment labels are assigned according to quartile ranks: "Champions" for high scores in all RFM categories, "Loyal Accounts" for high scores in MonetaryValue and Frequency, and "Lost" or "At Risk" for declining engagement:

```
data RFM_segments;
    set RFM_ranked;
    drop LastTransactionDate;
    length RFM_Segment $50.;
    if M_Quartile in (5, 4) and F_Quartile in (5, 4) and R_Quartile in (5, 4)
then RFM_Segment='Champions';
    else if M_Quartile in (5, 4) and F_Quartile in (5, 4) then
RFM_Segment='Loyal Accounts';
    else if R_Quartile in (5, 4) and M_Quartile in (3, 2, 1) then
RFM_Segment='Lost';
    else if R_Quartile in (5, 4) and M_Quartile in (4, 3, 2, 1) then
RFM_Segment='At Risk';
    else if R_Quartile in (5, 4) and M_Quartile=0 then RFM_Segment='About to
Sleep';
    else if R_Quartile in (5, 4) and F_Quartile=0 then RFM_Segment='Potential
Loyalist';
```

```
    else if M_Quartile=0 then RFM_Segment='Low Spenders';
    else if R_Quartile=0 then RFM_Segment='New Active Accounts';
    else if F_Quartile=0 then RFM_Segment='Promising';
    else RFM_Segment='Need Attention';
run;
```

Finally, merged all and this is the result:

RFM Segmentation Results

Obs	CustomerID	Recency	Frequency	MonetaryValue	R_Quartile	F_Quartile	M_Quartile	RFM_Segment
1	12346	346	48	155177.60	4	2	5	Need Attention
2	12347	23	253	5633.32	0	5	5	Loyal Accounts
3	12348	96	51	2019.40	2	2	4	Need Attention
4	12349	39	180	4452.84	1	4	5	Loyal Accounts

A1.The result of RFM Dataset

Insights:

1. During the end of the year (from October to December), noticed increasing in sales, frequency, and monetary, which we can use to improve the next campaigns.
2. The "Lost" segment represents about (19.39%) who have previously made purchases but are now inactive, requiring re-engagement.
3. "Loyal Accounts" represent about (24.69%) with high frequency and significance value of monetary, and need efforts to make them in targeted marketing.
4. The "Need Attention" segment represents about (26.91%) of trying to make them a great target for engagement-focused marketing efforts.
5. The "New Active Accounts" represent a small portion of customers (7.00%) who are new and have shown continuous activity.

Recommendations:

1. Prioritize targeting the "Champions" segment, comprising customers with the highest frequency and monetary, by making exclusive deals and rewards to encourage them to repeat purchases.
2. Focus on reactivating "Lost" customers, by personalized win-back campaigns and special promotions. Doing these campaigns the next month to re-engage and increase revenue
3. Schedule efforts of marketing during the end of year period (October to December), as we observed increase in sales, frequency, and monetary during this time.

CONCLUSION:

the RFM analysis gave us useful insights about customer behavior, making highlighting key segments like "Champions" and suggestions for improvement like re-engaging the "Lost" segment. at end of the year peaks and understanding correlations between customer metrics are useful for targeted marketing strategies. These insights and recommendations can lead the business in maximizing revenue and achieving customer satisfaction.

REFERENCES:

- SAS Viya for Learners Tutorials: https://www.sas.com/en_us/software/viya-for-learners.html (Accessed February 19, 2024)
- SAS Blogs: [https://blogs.sas.com/] (Accessed February 19, 2024)
- SAS Viya Help: https://documentation.sas.com/ (Accessed February 19, 2024)

ACKNOWLEDGMENTS

We would like to acknowledge Dr.Waleed Ead for his supervision and support throughout our work

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

yossefazammahfoze22_sd@fcis.bsu.edu.eg

cds.Amanygaber61708@alexu.edu.eg

aliatta205_sd@fcis.bsu.edu.eg

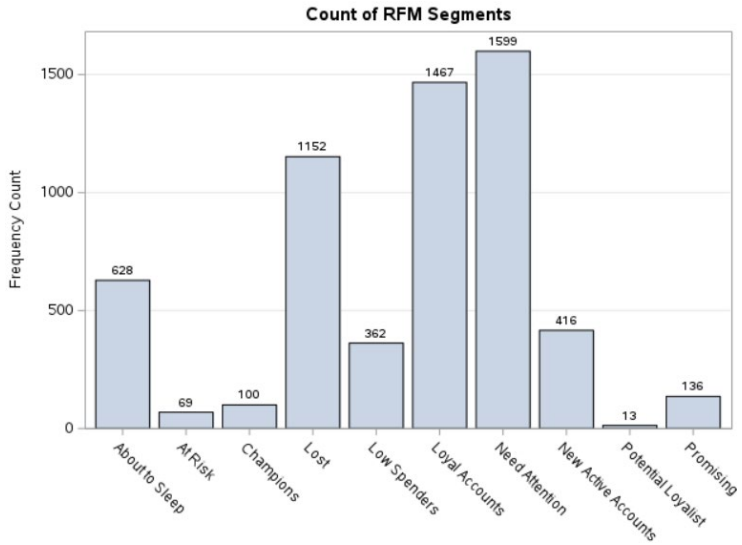
APPENDIX A

Visualizations (Using SAS Studio and Viya)

NumUniqueCustomers
5942

A1.Number of customers

A2.Total sales revenue

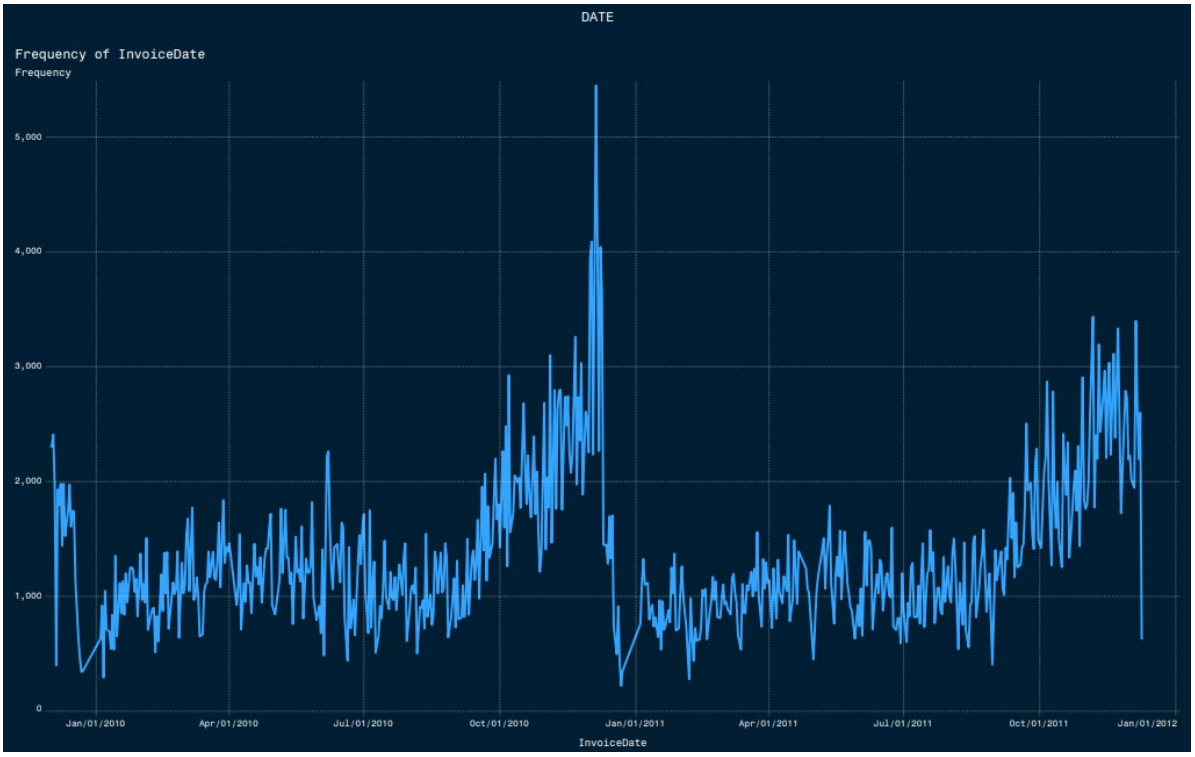


A3.Histogram about segments

All information for country

Country	Frequency	Quantity	Price
United Kingdom	741K	741K	2.6M
EIRE	16K	16K	110K
Germany	18K	18K	68K
France	14K	14K	67K
Norway	1.5K	1.5K	41K
Singapore	346	346	25K
Spain	3.8K	3.8K	21K
Portugal	2.5K	2.5K	17K
Netherlands	5.1K	5.1K	16K
Belgium	3.1K	3.1K	15K

A4.Top 10 countries per Price



A5.Frequency per Invoice date