Data Mining for Online Retail Industry

GROUP- 13

Simarpreet Kaur (0285168)

Krina Patel (0283599)

Tanvi Ajmera (0283944)

Kusam (0255724)

Prabhjot Singh (0278865)

Shannon School of Business, Cape Breton University

MGSC-5126-22: Data Mining

Dr. Jamileh Yousefi

Date: April 1st, 2024

ABSTRACT

The surge in online shopping over recent years has transformed consumer purchasing behaviors and financial services utilization. This shift is evident in the substantial growth of online sales, as highlighted by statistics from the Interactive Media in Retail Group (IMRG) and the Adobe Digital Economy Index. Online retail offers unique advantages such as real-time tracking of customer behavior, personalized customer interactions, and data-driven business intelligence. Data mining techniques, particularly the RFM (Recency, Frequency, Monetary Value) model, have become instrumental in understanding customer profitability and optimizing marketing strategies. This paper presents an overview of data preprocessing steps for an Online Retail dataset, exploring data structures, handling missing values, and creating new features for analysis and also a detailed exploration of the Online Retail dataset, which encompasses transactional data from December 2010 to December 2011, this study delves into preprocessing steps, data visualization, and the identification of key classes within the dataset. Market basket analysis, RFM model-based segmentation, and precision marketing strategies are discussed based on relevant literature, emphasizing their role in enhancing customer experiences, driving revenue, and fostering sustainable growth in the online retail sector. The findings underscore the significance of data analytics and customer insights in navigating the dynamic and competitive landscape of online retail.

INTRODUCTION

Over the last few years, we have seen steady and strong growth in online shopping. According to the Interactive Media in Retail Group (IMRG), online shoppers in the UK spent an estimated £50 billion in 2011, an increase of over 5,000% since 2000. According to the Adobe Digital Economy Index, UK shoppers spent around £110.6. billion online in 2022, nearly doubling in 11 years. This significant increase in online sales shows how consumers buy and use financial services has fundamentally changed.

Compared to shopping in traditional stores, online stores have some unique features: the buying process and performance of each customer can be tracked immediately and accurately, each customer's order usually has a delivery address and billing address, and each customer has an online store account with important contact and payment information. These desirable features of e-commerce have enabled e-commerce merchants to treat each customer as an individual with a personal understanding of each customer and to rely on customer-centric business intelligence.

To address these business concerns, data mining techniques have been widely adopted across the online retail sector, coupled with a set of well-known business metrics about customers' profitability and values, for instance, the recency, frequency, and monetary (RFM) model, and the customer life value model.

METHODOLOGY

Dataset Overview

Transactional data is contained in the Online Retail dataset from the UCI Machine Learning Repository for a UK-based online retail. Attributes like Invoice Number, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country are included in the dataset which span from December 2010 to December 2011. Description of synthetic dataset is shown in Table 1.

Sr. No.	Name of the Attribute	Type of the Attribute	Description of the Attribute
1	InvoiceNo	Character	Six-digit number uniquely assigned for each transaction.
2	StockCode	Character	Five-digit unique number assigned to each distinct product.
3	Description	Character	Name of the product
4	Quantity	Numeric	Quantities of each product per transaction
5	InvoiceDate	Numeric	Date and time of each transaction generated of x attribute
6	Unitprice	Numeric	Product price per unit
7	CustomerID	Numeric	Five-digit unique number assigned to a customer.
8	Country	Character	Name of the country

Table 1: Dataset overview

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047	United Kingdom
11	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-01 08:34:00	2.10	13047	United Kingdom
12	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-01 08:34:00	2.10	13047	United Kingdom
13	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	2010-12-01 08:34:00	3.75	13047	United Kingdom
14	536367	22310	IVORY KNITTED MUG COSY	6	2010-12-01 08:34:00	1.65	13047	United Kingdom
15	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	2010-12-01 08:34:00	4.25	13047	United Kingdom

Showing 1 to 15 of 541,909 entries, 8 total columns

Figure 1: Glimpse of dataset fields and values

Before preprocessing, the dataset includes 541,909 entries and 8 columns

Preprocessing Steps

Step 1: Load Necessary Libraries

```
library(readxl) # For reading Excel files

library(dplyr)  # For data manipulation
```

Step 2: Load the Dataset

```
Online_Retail <- read_excel("C:/Users/Prabhid/Desktop/Online Retail.xlsx")
```

#Step 3: Explore the Data

Explore the dataset to understand its structure, such as the columns available and the first few rows of data.

```
# Determine the number of instances (rows) and features (columns).

dim(retail_data)

# Display the structure of the dataset

str(Online_Retail)

# View the first few rows of the dataset

head(Online_Retail)
```

Step 4: Data Preprocessing

Common preprocessing steps include handling missing values, removing duplicate records, and possibly creating new features that could be useful for analysis.

Handling Missing Values

```
# Check for missing values
```

colSums(is.na(Online_Retail))

Depending on the analysis, you might decide to remove rows with missing values

For instance, if 'CustomerID' is crucial for your analysis

Online_Retail <- Online_Retail [!is.na(Online_Retail $CustomerID),]

Removing Duplicate Records

Online_Retail <- Online_Retail %>% distinct()

Creating New Features

For example, creating a TotalPrice feature might be useful.

Online_Retail <- Online_Retail %>%

  mutate(TotalPrice = Quantity * UnitPrice)

Step 5: Basic Exploration

Summary Statistics: Get a sense of the numeric columns.

summary(Online_Retail)

Step 6: Identification of Classes

unique(retail_data$ColumnName)

Step 7: Convert ‘InvoiceDate’ Using the Correct Format

Use the adjusted format string in as.POSIXct() to convert your InvoiceDate column:

Online_Retail$InvoiceDate <- as.POSIXct(Online_Retail$InvoiceDate, format = "%Y-%m-%d %H:%M:%S %p")

```
Online_Retail <- Online_Retail %>%
  mutate(
    Date = date(InvoiceDate),
    Time = format(InvoiceDate, "%H:%M:%S"),
    AMPM = format(InvoiceDate, "%p"))
```

LITERATURE REVIEW

Market basket analysis plays a pivotal role in understanding shopping behaviours at category and brand levels. İnanç Kabasakal's research in 2020 delves into the intricacies of market basket analysis, emphasizing the importance of identifying product affinities and cross-selling opportunities (Kabasakal, 2020). Furthermore, studies such as the one by Anitha and Patil in 2012 showcase the effectiveness of RFM model-based customer segmentation using the K-Means algorithm. This approach enables retailers to segment customers based on recency, frequency, and monetary value, allowing for targeted marketing strategies and personalized customer experiences (P. Anitha, 2019).

Additionally, Budilaksono et al.'s work in 2021 sheds light on precision marketing techniques, combining the RFM method, K-Means algorithm, and decision trees to create detailed customer profiles. These profiles aid in implementing precision marketing strategies, enhancing customer retention, and maximizing revenue generation (Sularso Budilaksono, 2021). Furthermore, Serwah et al.'s research underscores the significance of customer analytics and data mining techniques, such as weighted K-Means clustering and RFM analysis, in deriving actionable insights from large datasets. These insights empower online retailers to optimize marketing strategies, improve decision-making, and drive business growth (A. Serwah, 2023).

In conclusion, the literature reviewed underscores the critical role of data-driven approaches in understanding shopping behaviours and customer segmentation in the online retail industry. Leveraging market basket analysis, RFM model-based segmentation, data mining techniques, and precision marketing strategies enables retailers to gain a competitive edge, enhance customer experiences, and achieve sustainable growth. As online retail continues to evolve, leveraging data analytics and customer insights will remain integral to success in this dynamic and competitive landscape.

RESULTS AND DISCUSSION

Graph to show product popularity through description

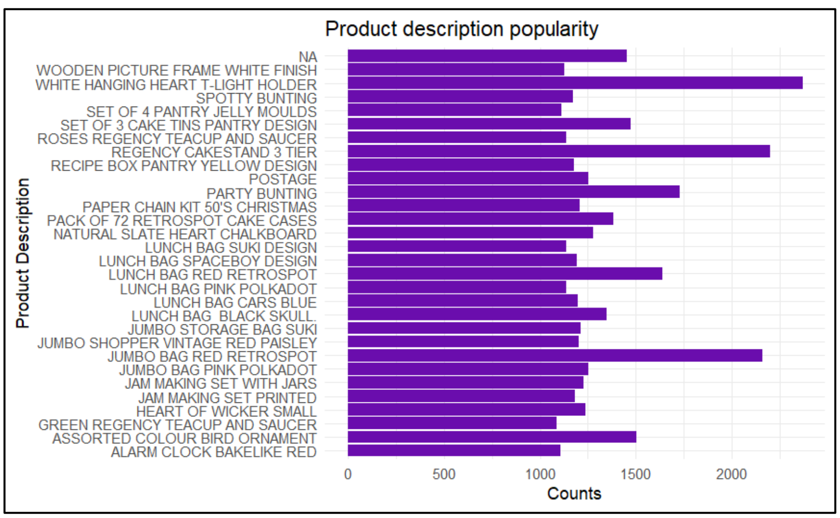

Figure 2: Product popularity graph

Visualizing product popularity through descriptions can provide valuable insights into consumer preferences and purchasing behaviors. As we can find out that white hanging heart – light holder is the most popular product.

Sales of Product in different countries

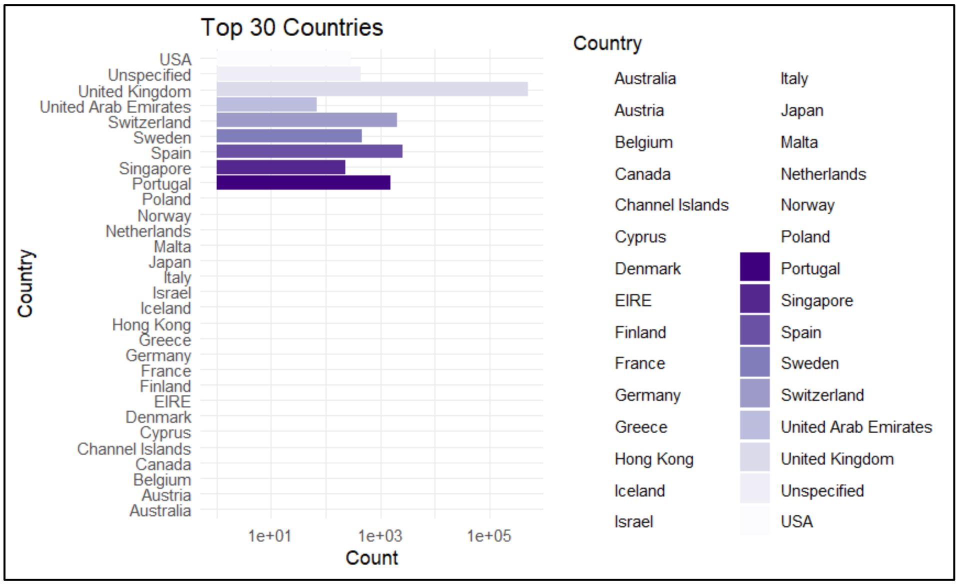

Figure 3: Country wise sales of products

Analyzing the sales of a product across different countries offers valuable insights into global market dynamics and consumer preferences. By examining sales data from diverse regions, businesses can identify lucrative markets, understand regional variations in demand, and tailor their marketing and distribution strategies accordingly. Visualizing sales figures through graphs can highlight geographical patterns, such as which countries contribute the most to overall sales or where there are opportunities for growth. Moreover, from this graph we can see that UK is the country that has highest sales.

Finding countries having highest product return rate

```
> print(Online_Retail_ex)
# A tibble: 541,909 × 8
   InvoiceNo StockCode Description    Quantity InvoiceDate         UnitPrice CustomerID Country
   <chr>     <chr>     <chr>             <dbl> <dttm>                  <dbl>      <dbl> <chr>
 1 536365    85123A    WHITE HANGING…        6 2010-12-01 08:26:00      2.55      17850 United…
 2 536365    71053     WHITE METAL L…        6 2010-12-01 08:26:00      3.39      17850 United…
 3 536365    84406B    CREAM CUPID H…        8 2010-12-01 08:26:00      2.75      17850 United…
 4 536365    84029G    KNITTED UNION…        6 2010-12-01 08:26:00      3.39      17850 United…
 5 536365    84029E    RED WOOLLY HO…        6 2010-12-01 08:26:00      3.39      17850 United…
 6 536365    22752     SET 7 BABUSHK…        2 2010-12-01 08:26:00      7.65      17850 United…
 7 536365    21730     GLASS STAR FR…        6 2010-12-01 08:26:00      4.25      17850 United…
 8 536366    22633     HAND WARMER U…        6 2010-12-01 08:28:00      1.85      17850 United…
 9 536366    22632     HAND WARMER R…        6 2010-12-01 08:28:00      1.85      17850 United…
10 536367    84879     ASSORTED COLO…       32 2010-12-01 08:34:00      1.69      13047 United…
# ℹ 541,899 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figure 4: Glimpse of output showing countries having highest product return rate

The dataset has around 5,41,909 values and 8 columns.

After omitting NA values the data set has around 4,06,829 values and 8 columns.

```
# A tibble: 406,829 × 8
   InvoiceNo StockCode Description     Quantity InvoiceDate         UnitPrice CustomerID Country
   <chr>     <chr>     <chr>              <dbl> <dttm>                  <dbl>      <dbl> <chr>
 1 536365    85123A    WHITE HANGING…         6 2010-12-01 08:26:00      2.55      17850 United…
 2 536365    71053     WHITE METAL L…         6 2010-12-01 08:26:00      3.39      17850 United…
 3 536365    84406B    CREAM CUPID H…         8 2010-12-01 08:26:00      2.75      17850 United…
 4 536365    84029G    KNITTED UNION…         6 2010-12-01 08:26:00      3.39      17850 United…
 5 536365    84029E    RED WOOLLY HO…         6 2010-12-01 08:26:00      3.39      17850 United…
 6 536365    22752     SET 7 BABUSHK…         2 2010-12-01 08:26:00      7.65      17850 United…
 7 536365    21730     GLASS STAR FR…         6 2010-12-01 08:26:00      4.25      17850 United…
 8 536366    22633     HAND WARMER U…         6 2010-12-01 08:28:00      1.85      17850 United…
 9 536366    22632     HAND WARMER R…         6 2010-12-01 08:28:00      1.85      17850 United…
10 536367    84879     ASSORTED COLO…        32 2010-12-01 08:34:00      1.69      13047 United…
# ℹ 406,819 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figure 5: Glimpse of output after data cleaning

#return_data containing only the rows where the Quantity column is less than zero.

```
# A tibble: 8,905 × 8
   InvoiceNo StockCode Description     Quantity InvoiceDate         UnitPrice CustomerID Country
   <chr>     <chr>     <chr>              <dbl> <dttm>                  <dbl>      <dbl> <chr>
 1 C536379   D         Discount              -1 2010-12-01 09:41:00     27.5       14527 United…
 2 C536383   35004C    SET OF 3 COLO…        -1 2010-12-01 09:49:00      4.65      15311 United…
 3 C536391   22556     PLASTERS IN T…       -12 2010-12-01 10:24:00      1.65      17548 United…
 4 C536391   21984     PACK OF 12 PI…       -24 2010-12-01 10:24:00      0.29      17548 United…
 5 C536391   21983     PACK OF 12 BL…       -24 2010-12-01 10:24:00      0.29      17548 United…
 6 C536391   21980     PACK OF 12 RE…       -24 2010-12-01 10:24:00      0.29      17548 United…
 7 C536391   21484     CHICK GREY HO…       -12 2010-12-01 10:24:00      3.45      17548 United…
 8 C536391   22557     PLASTERS IN T…       -12 2010-12-01 10:24:00      1.65      17548 United…
 9 C536391   22553     PLASTERS IN T…       -24 2010-12-01 10:24:00      1.65      17548 United…
10 C536506   22960     JAM MAKING SE…        -6 2010-12-01 12:38:00      4.25      17897 United…
# ℹ 8,895 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figure 6: Glimpse of dataset having values less than zero in Quantity column

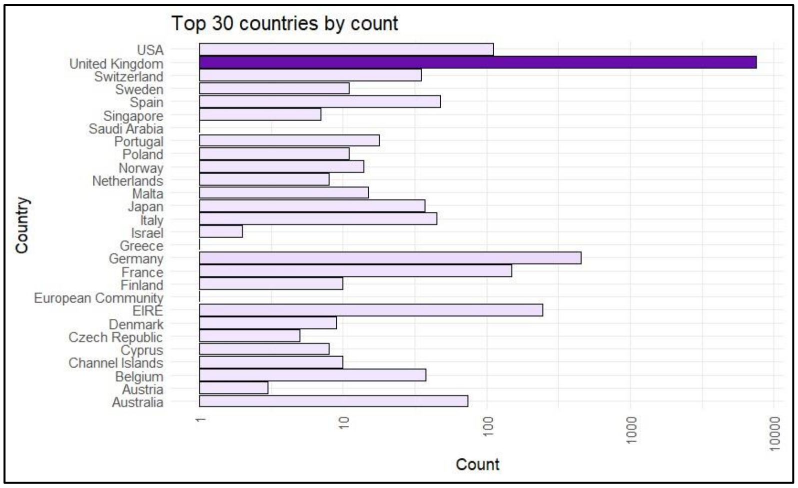

Figure 7: Countries having the maximum product returns

Analyzing product returns data with respect to different countries provides critical insights into customer satisfaction, product quality, and logistical challenges across global markets. By examining which country has the highest rate of product returns, businesses can pinpoint potential issues such as customer dissatisfaction, shipping complications, or discrepancies between product descriptions and actual offerings. From the above graph UK has most of the return this is obvious because this country is the one with the highest sales as well.

Matrix Correlation between revenue and most sale month

Revenue, Year, Quarter, Month, Week, Weekday and Day columns are created from InvoiceDate columns to work further to find the daily sales.

```
# A tibble: 5 × 16
  InvoiceNo StockCode Description     Quantity InvoiceDate         UnitPrice CustomerID Country
  <chr>     <chr>     <chr>              <dbl> <dttm>                  <dbl>      <dbl> <chr>
1 550159    21175     GIN + TONIC DI…       96 2011-04-14 15:46:00      2.08      14062 United…
2 560592    21175     GIN + TONIC DI…       96 2011-07-19 16:36:00      2.08      13225 United…
3 561645    23318     BOX OF 6 MINI …       96 2011-07-28 15:16:00      2.08      14911 EIRE
4 562374    23318     BOX OF 6 MINI …       96 2011-08-04 14:40:00      2.08      14911 EIRE
5 566089    21175     GIN + TONIC DI…       96 2011-09-09 10:31:00      2.08      14062 United…
# ℹ 8 more variables: Revenue <dbl>, Year <dbl>, Quarter <int>, Month <dbl>, Week <dbl>,
#   Weekday <dbl>, Day <int>, DescriptionLength <int>
```

Figure 8: Glimpse of new columns created from existing columns

Finding popular days to shop

#Sort the data by 'Revenue' in descending order and count occurrences of each day and show them in graph

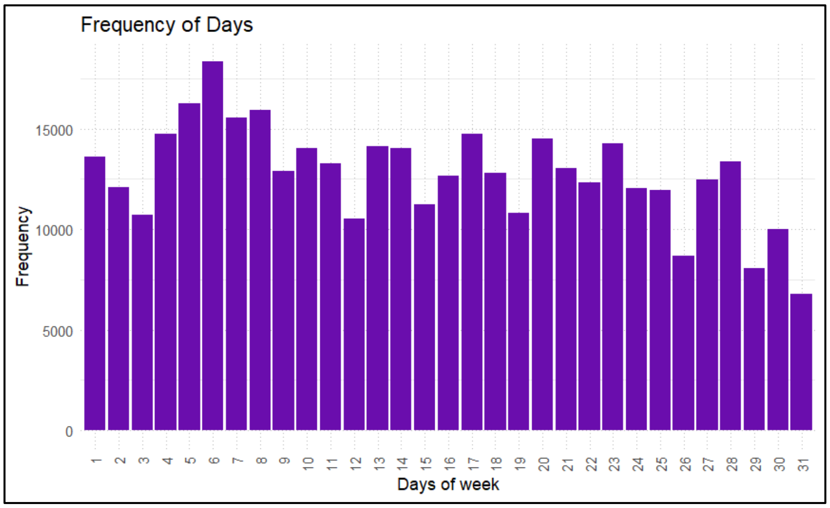

Figure 9: Frequency of days based on revenue

From the graph, it can be seen that day 6th of a week is popular in online retail.

Finding popular month to shop

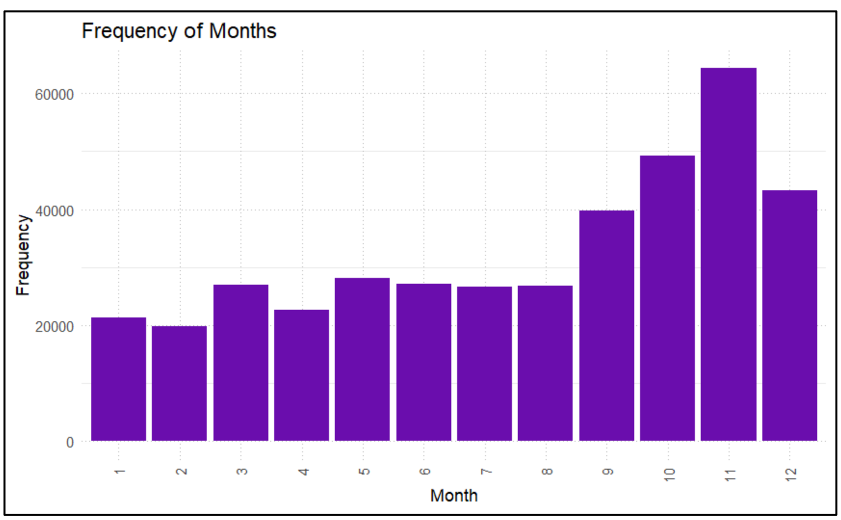

Figure 10: Best month to shop online

The November month has the highest sales throughout the year.

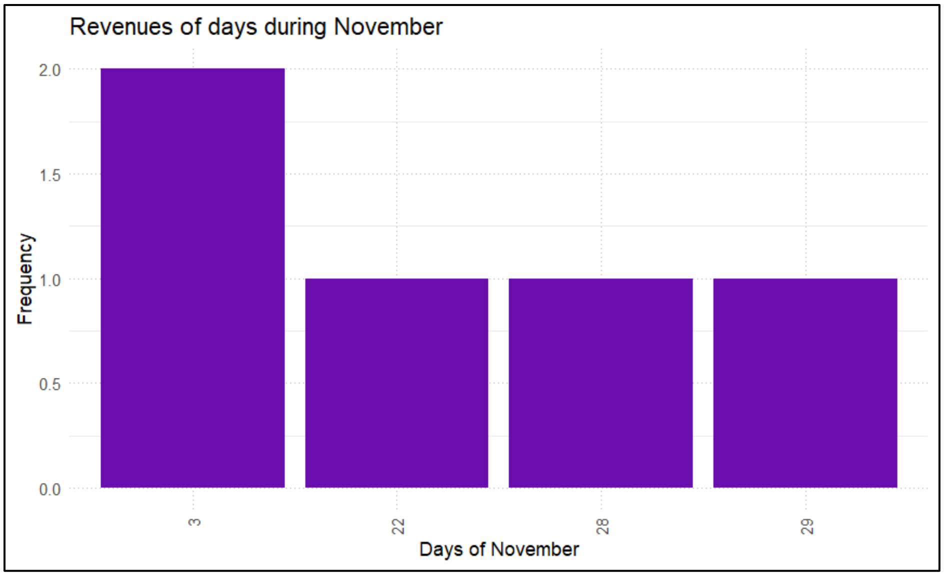

Figure 11: Revenue in November

From the graph above, it is visible that November 3rd is popular.

Graph showing the maximum occurrence of a stock code

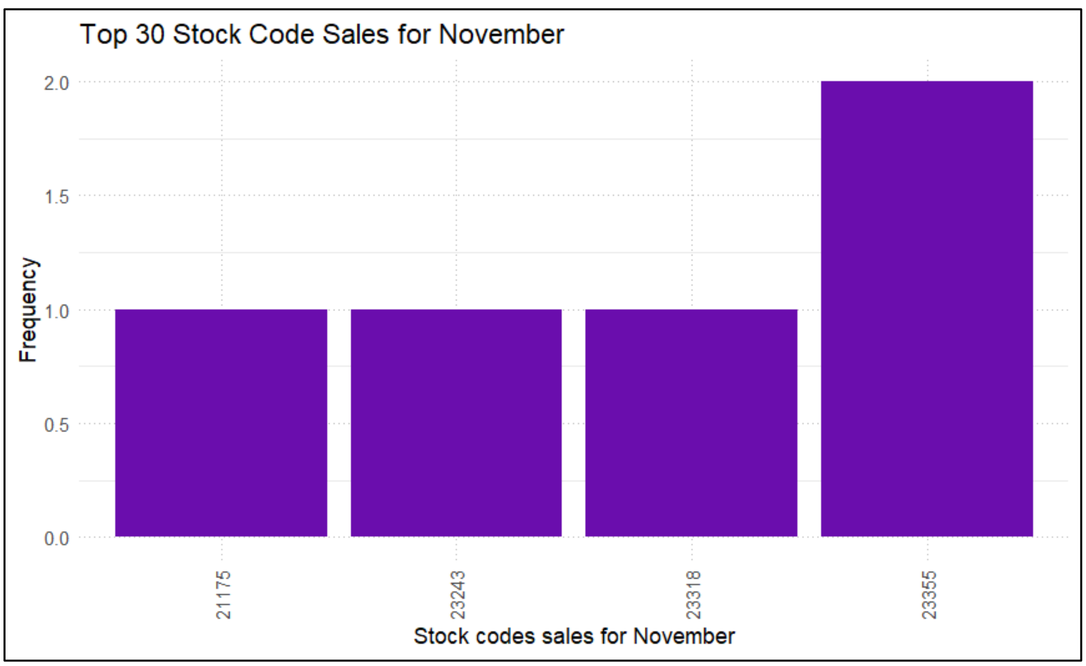

Figure 12: Top stock code sales for November

Looks like product code 23355 is very popular during month of November.

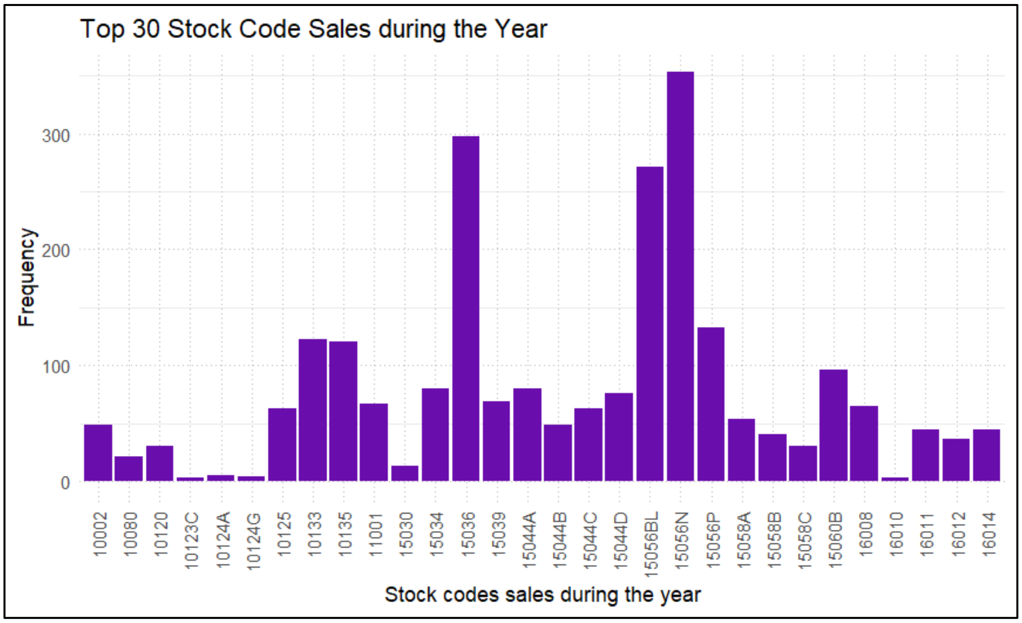

Figure 13: Stock code sales during the year

Looks like stock code 15056 is popular throughout the year opposed to code 23355 during the month of November.

Calculate the 1st and 99th percentiles for 'Quantity' and 'Revenue' then remove infinite values. Here is the summarized data.

```
  InvoiceNo           StockCode          Description           Quantity
 Length:396018       Length:396018      Length:396018      Min.    : -2.000
 Class :character    Class :character   Class :character   1st Qu.:  2.000
 Mode  :character    Mode  :character   Mode  :character   Median :  5.000
                                                           Mean   :  9.468
                                                           3rd Qu.: 12.000
                                                           Max.   :120.000
  InvoiceDate                        UnitPrice        CustomerID       Country
 Min.   :2010-12-01 08:26:00.00   Min.   :  0.000   Min.   :12347   Length:396018
 1st Qu.:2011-04-07 10:24:00.00   1st Qu.:  1.250   1st Qu.:13969   Class :character
 Median :2011-07-31 13:33:00.00   Median :  1.950   Median :15159   Mode  :character
 Mean   :2011-07-10 22:28:24.99   Mean   :  2.918   Mean   :15293
 3rd Qu.:2011-10-20 15:09:15.00   3rd Qu.:  3.750   3rd Qu.:16794
 Max.   :2011-12-09 12:50:00.00   Max.   :195.000   Max.   :18287
    Revenue             Year          Quarter           Month             Week
 Min.   : -9.90   Min.   :2010   Min.   :1.000   Min.   : 1.000   Min.   : 1.00
 1st Qu.:  4.25   1st Qu.:2011   1st Qu.:2.000   1st Qu.: 5.000   1st Qu.:19.00
 Median : 10.82   Median :2011   Median :3.000   Median : 8.000   Median :34.00
 Mean   : 16.46   Mean   :2011   Mean   :2.856   Mean   : 7.612   Mean   :30.96
 3rd Qu.: 18.00   3rd Qu.:2011   3rd Qu.:4.000   3rd Qu.:11.000   3rd Qu.:44.00
 Max.   :199.68   Max.   :2011   Max.   :4.000   Max.   :12.000   Max.   :51.00
    Weekday            Day
 Min.   :1.000   Min.   : 1.00
 1st Qu.:2.000   1st Qu.: 7.00
 Median :4.000   Median :15.00
 Mean   :3.508   Mean   :15.04
 3rd Qu.:5.000   3rd Qu.:22.00
 Max.   :6.000   Max.   :31.00
```

Figure 14: Summary of percentiles of Quantity and Revenue

Here is the histogram of daily product sales with respect to frequency.

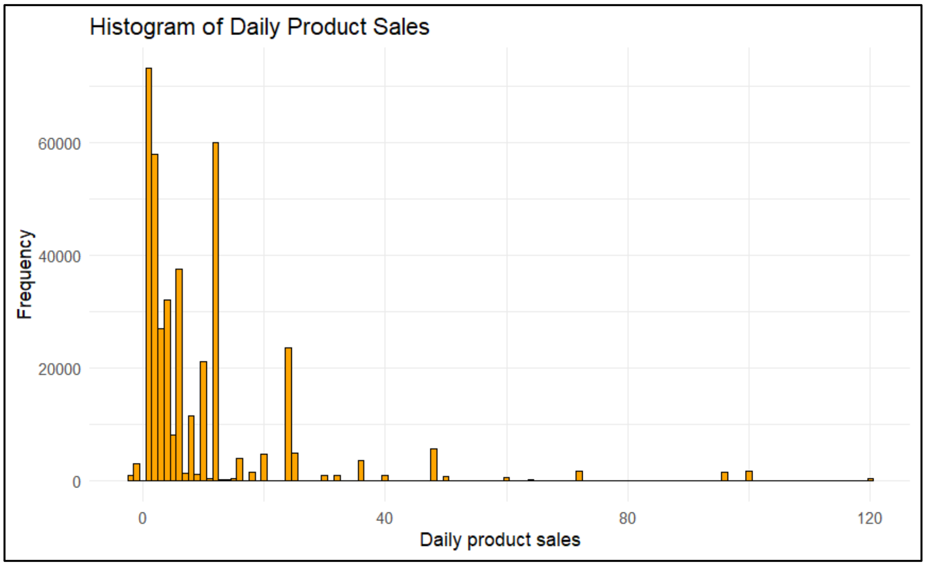

Figure 15: Daily product sales

We can see the lower quantities are popular during purchase

Correlation Matrix

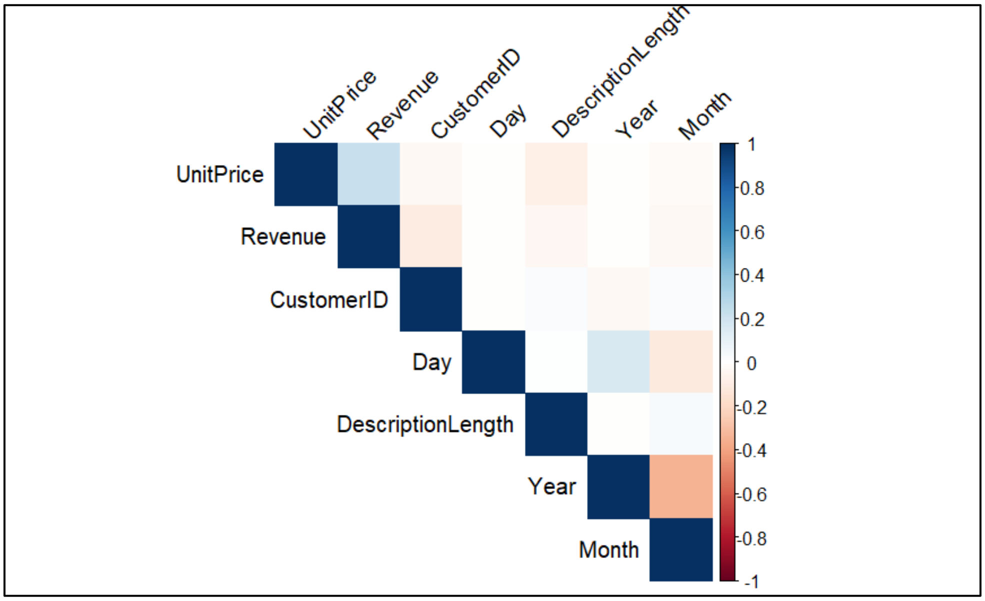

Figure 16: Correlation Matrix

#We can see by correlation matrix how month is related to revenue of the product and its unit price and revenue.

CONCLUSION

In conclusion, the literature reviewed underscores the critical role of data-driven approaches in understanding shopping behaviors and customer segmentation in the online retail industry. Leveraging market basket analysis, RFM model-based segmentation, data mining techniques, and precision marketing strategies enables retailers to gain a competitive edge, enhance customer experiences, and achieve sustainable growth.

The insights derived from this review highlight the importance of leveraging data analytics and customer insights to drive informed decision-making and optimize marketing strategies. As online retail continues to evolve, businesses must remain agile and adaptable, leveraging emerging technologies and innovative methodologies to meet evolving consumer demands and stay ahead of the competition.

By embracing data-driven strategies, online retailers can unlock new opportunities for personalized customer engagement, targeted marketing campaigns, and revenue optimization. Ultimately, success in the dynamic and competitive online retail landscape hinges on the strategic integration of data analytics, customer-centric approaches, and continuous innovation to deliver exceptional value and drive long-term growth.

REFERENCES

[PDF] customer Profilling for precision marketing using RFM method, K-MEANS algorithm and decision tree. (n.d.). Semantic Scholar | AI-Powered Research Tool. https://www.semanticscholar.org/reader/7e42d434b009bf564a1b5688258241acf38d9b45

Applying data mining for online CRM marketing strategy: An empirical case of the coffee shop industry in Taiwan. (2018, March 5). Discover Journals, Books & Case Studies | Emerald Insight. https://www.emerald.com/insight/content/doi/10.1108/BFJ-02-2017-0075/full/html

(n.d.). CEUR-WS.org - CEUR Workshop Proceedings (free, open-access publishing, computer science/information systems). https://ceur-ws.org/Vol-2649/paper2.pdf

Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. (2012, August 27). SpringerLink. https://link.springer.com/article/10.1057/dbm.2012.17

Understanding shopping behaviors with category- and brand-level market basket analysis. (0001, January 1). IGI Global: International Academic Publisher. https://www.igi-global.com/gateway/chapter/235905

RFM model for customer purchase behaviour. (n.d.). using the K-means algorithm. https://www.sciencedirect.com/science/article/pii/S1319157819309802?via%3Dihub

Customer analytics for online retailers using weighted k-means and RFM analytics. (n.d.). Semantic

Scholar. https://pdfs.semanticscholar.org/2094/679b670242e4177670b839fd7736fc4 54afa.pdf?_gl=1*t39621*_ga*MzIzOTc0MzQuMTcwODI5NTI1NA..*_ga_H7P4ZT52 H5*MTcxMDg1ODA0NS4yLjEuMTcxMDg1ODk0NC41OS4wLjA