MKTG 2505: Marketing Analytics

Online Retail Market Analysis Final Project

Dr. Sujata Ramnarayan

Santa Clara University, Leavey School of Business

Group 3

Tommy Cleland, Justin Hurcombe, Shilipi Kumari,

Komal Srivastava, Ameha Teshome

March 15th, 2023

Table of Contents

Data description

This dataset contains transactional data from an online gift retailer, including information on the products sold, the quantity, the price, customer ID, and country of origin. By analyzing this data, we can gain valuable insights into which products are popular, customers' purchasing behavior, and how much customers are willing to pay for different products. This information can help retailers decide on product offerings and marketing campaigns to drive growth and expansion.

The dataset also includes information on whether a transaction was a purchase or a return, providing additional insight into customers' behavior. This information can help the retailer identify and target customers more likely to purchase and improve their overall customer experience. By analyzing this rich dataset, online gift retailers can gain valuable insights into their business and make data-driven decisions to drive growth and success.

Situation Analysis

❖ The first plot tried to answer the question "How do sales vary by month?" It shows that sales are the highest from September to November/December, especially in November. Holiday shopping starts in these months, and that caused the sales to be higher than any of the months in a year.

❖ Using the n_distinct() function, we learned that 4290 unique customers are involved in online retail purchases.

❖ The second plot answered, "What is the most frequently purchased product, and how much revenue does it generate?" In this dataset, the most frequently purchased product

and revenue it generates is "regency cakestand 3 tier"; its total sales or revenues are 128794.15.

❖ The third plot answered, "What is the total sales trend over time?" The procedures we used for this plot are to divide the total sales into four quarters and compute the total sales for each quarter to compare the trends. The fourth quarter had the highest total sale and was $2.5 million+.

❖ The following plot shows which customers are the top consumers and their contribution to total sales or revenue. This plot helps identify the top customers of the online retail business and understand their impact on the overall sales revenue.

❖ The pie chart shows the total sale percentage by season. Fall has the highest sales, with ~37% of all total sales, and summer has the lowest.

❖ In Fig 1.6 shows the top products by total sales across 2010 and 2011. The resulting data frame, "top_products", shows the top 10 products by total sales revenue across 2010 and 2011, along with their total revenue and the year they were sold. This information can be used to identify which products were the most popular and profitable over the two years and to make informed decisions about inventory management and marketing strategies.

❖ To better understand the daily sales trend, we created a plot to track every sale daily using a line plot. The line plot can help identify any patterns or trends in the sales data over time, such as seasonal fluctuations or changes in sales trends. It can also help identify any outliers or unusual spikes in sales that may require further investigation.

❖ The following bar plot visualizes the top 5 countries with the most purchases. Based on the results, most of the sales were from the UK. This plot is important because it visually represents the top 5 countries with the highest purchases. It allows the online retail

business to identify essential markets and make informed decisions about targeting these countries for growth.

❖ The last bar plot visualizes the UK's top 5 most purchased items. The insights that can be derived from this analysis include the suggestion to offer promotions or bundles for the top-selling products to incentivize customers to purchase them. This analysis is specific to the UK market, but other countries can use these insights to improve their strategies by determining if similar products would succeed.

❖ Market Basket Analysis: The rule suggests that customers who purchase both the "GREEN REGENCY TEACUP AND SAUCER" and the "ROSES REGENCY TEACUP AND SAUCER" are likely to also purchase the "PINK REGENCY TEACUP AND SAUCER" with a high confidence of 0.720 and a lift of 23.363. The support of this rule is 0.022, meaning that this combination of purchases occurs in only 2.2% of all transactions. The lift of 23.363 suggests that this combination of purchases is much more likely to occur together than expected if the purchases were independent. Overall, this rule can be used to improve sales by placing these three items close to each other, offering bundle deals or discounts on the combination of these three items, or promoting the purchase of "PINK REGENCY TEACUP AND SAUCER" to customers who have already bought "GREEN REGENCY TEACUP AND SAUCER" and "ROSES REGENCY TEACUP AND SAUCER."

Identification of Problem

This analysis aims to identify items with higher return rates in the online retail dataset. Understanding the critical factor influencing a higher return on purchased items is essential.

Approach for Data Analysis

- The first step is to filter the data for only negative values of the "Quantity" variable, which would indicate a return.

- Then, calculates the return rate by dividing the number of returns by the total number of transactions in the dataset and multiplying by 100 to get a percentage.

- Next, group the returned items by their stock code and description and calculate the return rate for each item by dividing the number of returns for that item by the absolute value of the sum of its quantities.

- Finally, arranges the items in descending order of return rate, select the top 10 items, remove any rows with missing values, and sort the remaining rows in descending order of return rate.

Techniques for Data Analysis (EDA)

- Initial cleaning involved removing rows with negative values in Quantity or UnitPrice and missing values, transactions with no customer ID, and duplicates.

- Outliers were handled by removing rows with extreme values in Quantity or UnitPrice.

- Inconsistent data in the Description column was standardized in terms of formatting.

- Skewness was addressed through log transformation of the Quantity and UnitPrice columns.

- For clustering, another data frame was created that removed text columns (StockCode, Description, InvoiceDate, Month, and Country), decreasing columns from 60 to 56.

- Both data frames had the same number of rows as the original dataset.

Cluster Analysis

We segmented customers based on their orders using cluster analysis. Customer data was prepared by grouping and summarizing variables such as total amount spent, purchase frequency, quantity ordered, average unit price, and days since the last purchase. We determined the best number of clusters using an elbow chart and silhouette scores and decided on k-means clustering with three groups. Analyzing the summary characteristics of each cluster, we found three distinct customer segments to address (Fig. 1.11).

The customers can be divided into wholesalers, loyal customers, and opportunistic customers (Fig. 1.13 - 1.15). Wholesalers represent the smallest group but have the highest revenue potential for the company. They have a high purchase frequency and order volume (Fig. 1.13), and maintaining positive relationships with them is essential for the company's success. Loyal customers are the largest group and the most significant source of revenue. They make frequent purchases and spend more than opportunistic customers (Fig. 1.14). The company needs to ensure they maintain their purchase frequency and do not become opportunistic customers. In contrast, opportunistic customers are the smaller and contribute the least to the company's revenue (Fig. 1.15). They only purchase when they have a need, and little effort should be spent on marketing to this segment.

Understanding these customer segments' characteristics allows the company to tailor its marketing strategies and prioritize efforts to maximize revenue opportunities. The company can ensure a steady revenue stream by prioritizing efforts to maintain positive relationships with wholesalers and loyal customers. Meanwhile, while not neglecting the opportunistic customers, the company can focus more of its marketing efforts on attracting and retaining loyal customers.

This customer segmentation approach can help the company optimize its marketing strategies and resources to better target and serve its customers.

Conclusions

Based on the first insight, the company should increase inventory to meet the higher demand, since sales are the highest during the holiday season. This will ensure that they have enough stock to fulfill orders and reduce the chances of stockout.

The company should take cues from the second insights and promote related products. Since the "regency cakestand 3 tier" is a popular product, the company can promote related products, such as other cake stands or baking accessories, to encourage customers to make additional purchases.

From the "top consumers" insights, the company should provide personalized services to them, Since the top customers contribute significantly to the company's revenue. This can include offering exclusive discounts, early access to new products, and dedicated customer support. Additionally, The company can build customer loyalty among its top customers by offering loyalty programs that reward repeat purchases or referrals. This can help to increase customer retention and reduce customer churn.

Since the UK is the country with the highest purchases, the company should focus on this market to increase sales. This can include targeted marketing campaigns, promotions, and discounts to attract more customers in the UK. For the top products that are purchased in the UK, the company should analyze why the "White Hanging Heart tlight holder" is the most frequently purchased product. Is it because of its quality, design, or price? This information can be used to inform future product development and marketing strategies.
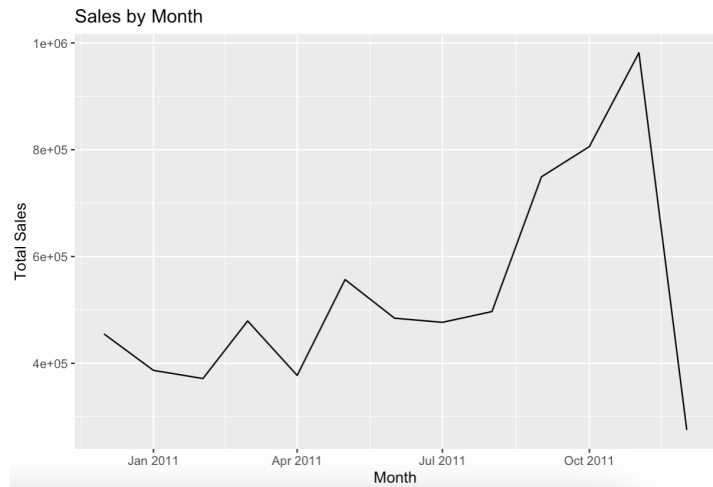
Appendix

Sales by Month

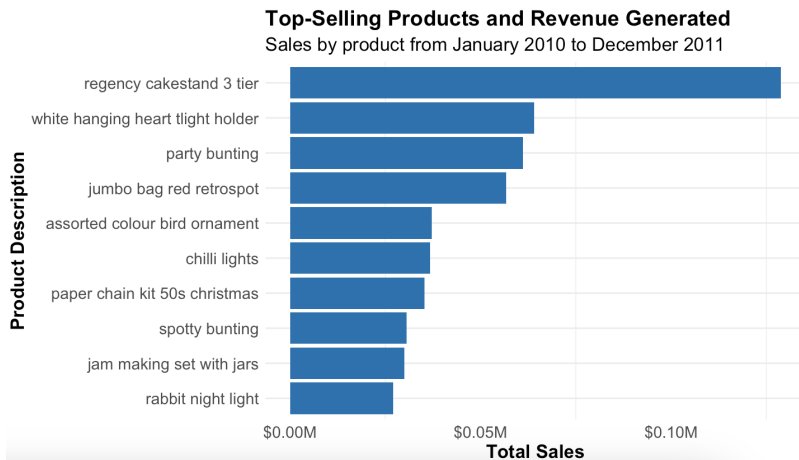**Fig 1.1: Sales by months**

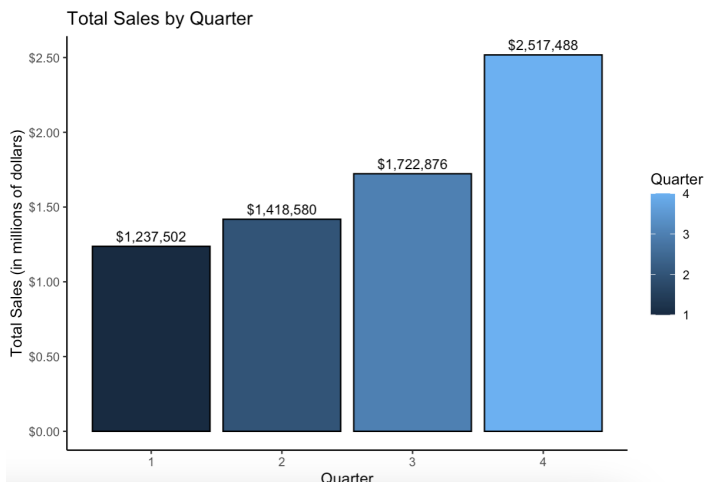**Fig 1.2: Top selling products and revenue generated**
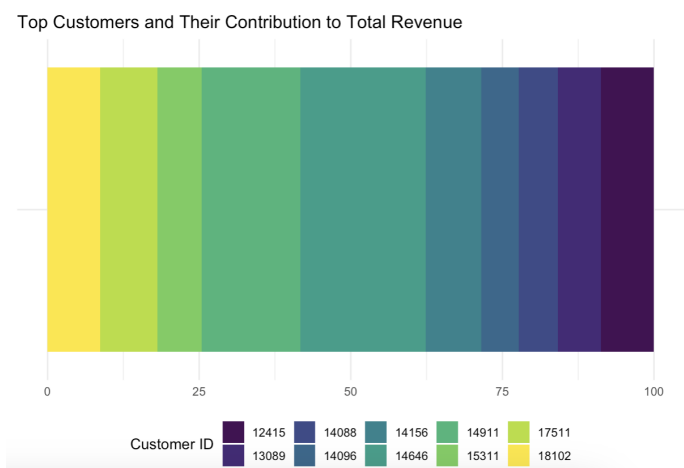

Fig 1.3: Trend of total sales

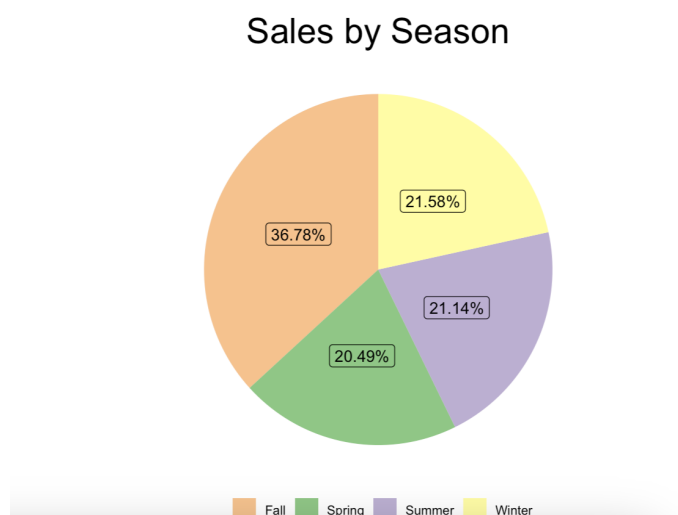**Fig 1.4: Top Customers and Their Contribution to Total Revenue**

Fig 1.5: Sales by season

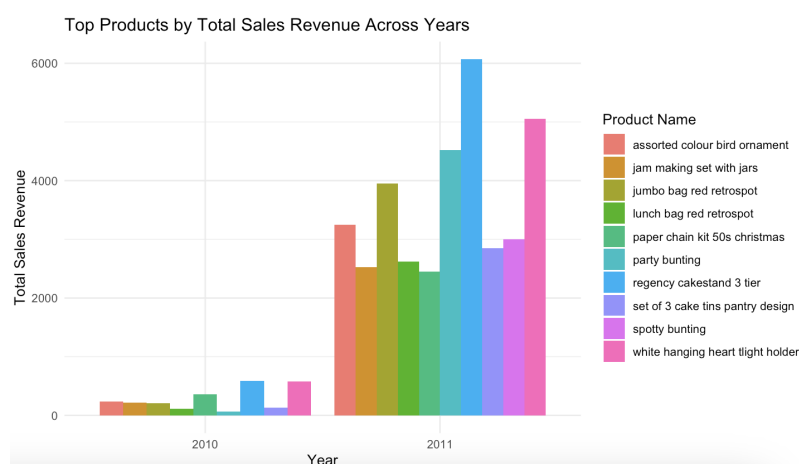**Fig 1.6: Top products by total sales revenue across years**

Fig 1.7: Daily Sales Trend

Fig 1.8: Popular words in the product description

Fig 1.9: Top 5 countries with the highest number of purchases

roses regency teacup and saucer   green regency teacup and saucer

support
→ 0.022
→ 0.023
→ 0.024
→ 0.025

lift
23
22
21
20

pink regency teacup and saucer
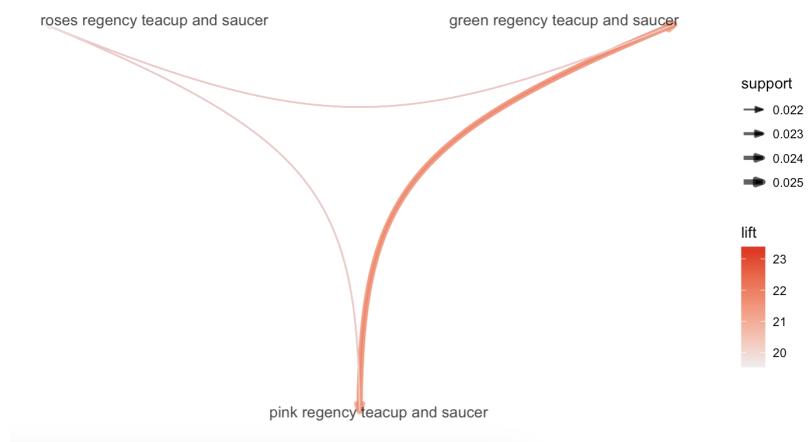
Fig 1.10: Market Basket Analysis

Cluster plot

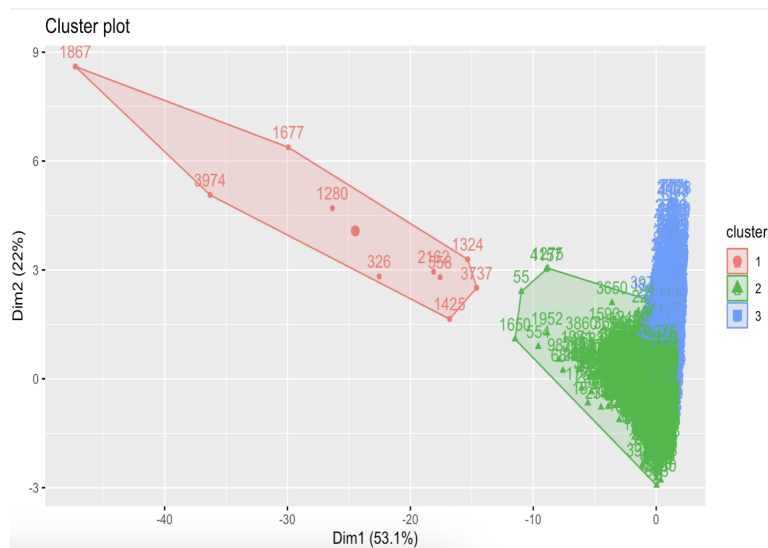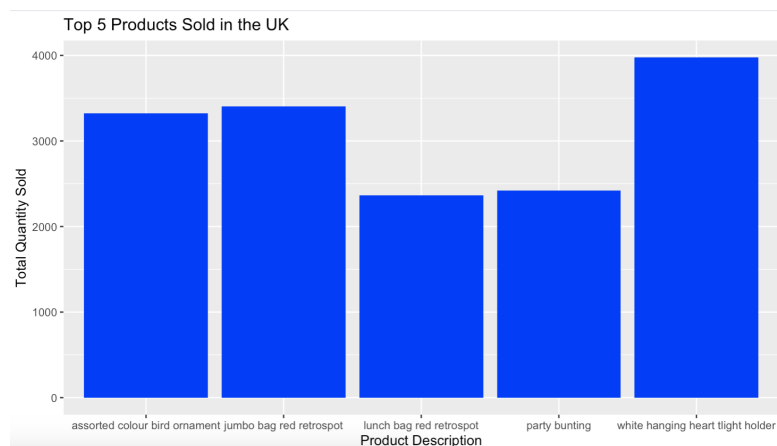**Fig 1.11: K-means Cluster Plot**

cluster
● 1
▲ 2
■ 3

Top 5 Products Sold in the UK

Fig 1.12: Top 5 products sold in the UK

Fig. 1.13: Characteristics of Wholesalers (Cluster 1)

```
> summary(online_retail_cust[online_retail_cust$cluster==1,])
   CustomerID  amount_spent     purchase_frequency quantity_puchased avg_unit_price
 12748  :1    Min.   : 28761   Min.   : 849       Min.   : 3089     Min.   :0.7974
 13089  :1    1st Qu.: 47439   1st Qu.:1574       1st Qu.: 4235     1st Qu.:1.0620
 14096  :1    Median : 55903   Median :2070       Median : 5298     Median :1.1064
 14156  :1    Mean   : 70204   Mean   :3177       Mean   : 5816     Mean   :1.1204
 14298  :1    3rd Qu.: 72393   3rd Qu.:4767       3rd Qu.: 5801     3rd Qu.:1.1938
 14646  :1    Max.   :158968   Max.   :7642       Max.   :12703     Max.   :1.3927
 (Other):4
 days_last_purchase    cluster
 Min.   : 1.00       Min.   :1
 1st Qu.: 2.25       1st Qu.:1
 Median : 3.50       Median :1
 Mean   : 5.60       Mean   :1
 3rd Qu.: 9.50       3rd Qu.:1
 Max.   :12.00       Max.   :1
```

Fig 1.14: Characteristics of Loyal Customers (Cluster 2)

```
> summary(online_retail_cust[online_retail_cust$cluster==2,])
   CustomerID     amount_spent      purchase_frequency quantity_puchased  avg_unit_price
 12347  :   1   Min.   :    6.2   Min.   :   1.0      Min.   :   1.386   Min.   :0.2546
 12348  :   1   1st Qu.:  381.1   1st Qu.:  25.0      1st Qu.:  51.153   1st Qu.:0.9947
 12349  :   1   Median :  854.7   Median :  57.0      Median : 111.736   Median :1.1293
 12352  :   1   Mean   : 1788.1   Mean   : 101.8      Mean   : 195.391   Mean   :1.1376
 12356  :   1   3rd Qu.: 1965.0   3rd Qu.: 122.0      3rd Qu.: 240.679   3rd Qu.:1.2748
 12357  :   1   Max.   :67004.4   Max.   :2657.0      Max.   :2750.613   Max.   :2.1248
 (Other):3189
 days_last_purchase    cluster
 Min.   : 1.00       Min.   :2
 1st Qu.: 12.00      1st Qu.:2
 Median : 30.00      Median :2
 Mean   : 42.24      Mean   :2
 3rd Qu.: 63.00      3rd Qu.:2
 Max.   :189.00      Max.   :2
```

Fig 1.15: Characteristics of Opportunistic Customers (Cluster 3)

```
> summary(online_retail_cust[online_retail_cust$cluster==3,])
   CustomerID    amount_spent     purchase_frequency quantity_puchased  avg_unit_price
 12350  :   1  Min.   :   2.9   Min.   :  1.00      Min.   :  0.6931   Min.   :0.3477
 12353  :   1  1st Qu.: 157.8   1st Qu.:  8.00      1st Qu.: 17.2312   1st Qu.:1.0907
 12354  :   1  Median : 300.9   Median : 18.00      Median : 33.3613   Median :1.2825
 12355  :   1  Mean   : 443.7   Mean   : 25.94      Mean   : 46.5615   Mean   :1.3269
 12361  :   1  3rd Qu.: 500.3   3rd Qu.: 31.00      3rd Qu.: 57.0549   3rd Qu.:1.5065
 12365  :   1  Max.   :8951.3   Max.   :289.00      Max.   :549.5834   Max.   :2.7695
 (Other):1079
 days_last_purchase    cluster
 Min.   : 10.0       Min.   :3
 1st Qu.:186.0       1st Qu.:3
 Median :241.0       Median :3
 Mean   :241.4       Mean   :3
 3rd Qu.:297.0       3rd Qu.:3
 Max.   :374.0       Max.   :3
```