MONASH University

ETW2001 - Foundations of data analysis

Assignment 3: Group Assignment

Title:

A Comprehensive Data Analytics to analyze E-Commerce Profitability

Team Members:

Lam Man Hyin (32023367)

Lee Zhen Yan (32023634)

Chai Jia Jing (31920772)

Kong Pui Fun (32023170)

Sia Yi Bin (33363129)

Word Count: 2939 words

Problem Statement 1

Profitability Analysis by Product Category

Problem: The profitability of different product categories varies, resulting in an inefficient allocation of resources for inventory and marketing.

Solution: Determine the *product categories that generate the highest profit margin*. Investigate the factors contributing to these margins, such as consumer demand (Quantity Sold) and production costs (Total Sales).

Result: The company can enhance its ability to organise inventory management and marketing campaigns to optimise resource allocation and profit.

Task 2 for Problem Statement 1

Step 1: Data Manipulation and Cleaning Techniques

```
# Task 2
# Load necessary libraries
library(dplyr)

# Read Amazon Sale Report from csv file into amazon_data data frame
amazon_data <- read.csv("C:/Users/Sean/Documents/ETW2001 Group Assignment/Amazon Sale Report.csv")
# Return the column names of the amazon_data data frame
colnames(amazon_data)
```

```
# Drop NA
# Removing columns "promotion.ids" and "Unnamed..22", then filtering out rows with
# missing values in columns "Order.ID", "Date", "Category", and "Amount"
amazon_data_cleaned <- amazon_data %>%
  select(-c(promotion.ids, Unnamed..22)) %>%
  filter(!is.na(Order.ID) & !is.na(Date) & !is.na(Category) & !is.na(Amount))

# Converts the "Date" column in the amazon_data_cleaned data frame to the Date data type
# using the specified format ("%m-%d-%y")
amazon_data_cleaned$Date <- as.Date(amazon_data_cleaned$Date, format = "%m-%d-%y")
# Display the content of amazon_data_cleaned
amazon_data_cleaned
```

```
# Analysis profit by Category column
category_analysis <- amazon_data_cleaned %>%
  group_by(Category) %>%
  summarise(Total_Sales = sum(Amount, na.rm = TRUE),
            Average_Amount = mean(Amount, na.rm = TRUE),
            Total_Quantity_Sold = sum(Qty, na.rm = TRUE)) %>%
  mutate(Price_Per_Unit = Total_Sales / Total_Quantity_Sold)
# Display the content of category_analysis
category_analysis

# Arranges the data frame "category_analysis" in descending order based on the "Price_Per_Unit"
# column to identify the categories with the highest profit per unit.
highest_profit_categories <- category_analysis %>%
  arrange(desc(Price_Per_Unit))
# Display the content of highest_profit_categories
highest_profit_categories
```

To clean and manipulate data in R, we import the Dplyr library. Then, we import the data by utilising the read.csv() function. We use the colnames (amazon_data) function to examine the column names and gain a clear understanding of the dataset's structure. We perform data cleansing by eliminating redundant columns (promotion. ids and Unnamed..22) and excluding rows with missing values in significant columns (Order. ID, Date, Category, Amount) through the select and filter capabilities.

Next, we convert the 'Date' column to the Date data type using the as.Date() function on the amazon_data_cleaned dataset. The date should be formatted as "mm-dd-yy". A descriptive analysis is conducted by grouping the data based on the 'Category' variable and calculating the total sales, average amount, and total quantity sold. Additionally, the price per unit is determined using the mutate function. Finally, we sort the categories in descending order based on the cost per unit to choose the most profitable categories using the arrange(desc(Price_Per_Unit)) function.

Output:

highest_profit_categories:

	Category	Total_Sales	Average_Amount	Total_Quantity_Sold	Price_Per_Unit
1	Set	39204124.0	833.3856	45225	866.8684
2	Saree	123933.8	799.5726	152	815.3537
3	Western Dress	11216072.7	762.7906	13939	804.6540
4	Ethnic Dress	791217.7	723.8954	1053	751.3938
5	Blouse	458408.2	520.3271	844	543.1377
6	Top	5347792.3	526.0986	9899	540.2356
7	kurta	21299546.7	455.9271	44970	473.6390
8	Bottom	150668.0	358.7333	397	379.5163
9	Dupatta	915.0	305.0000	3	305.0000

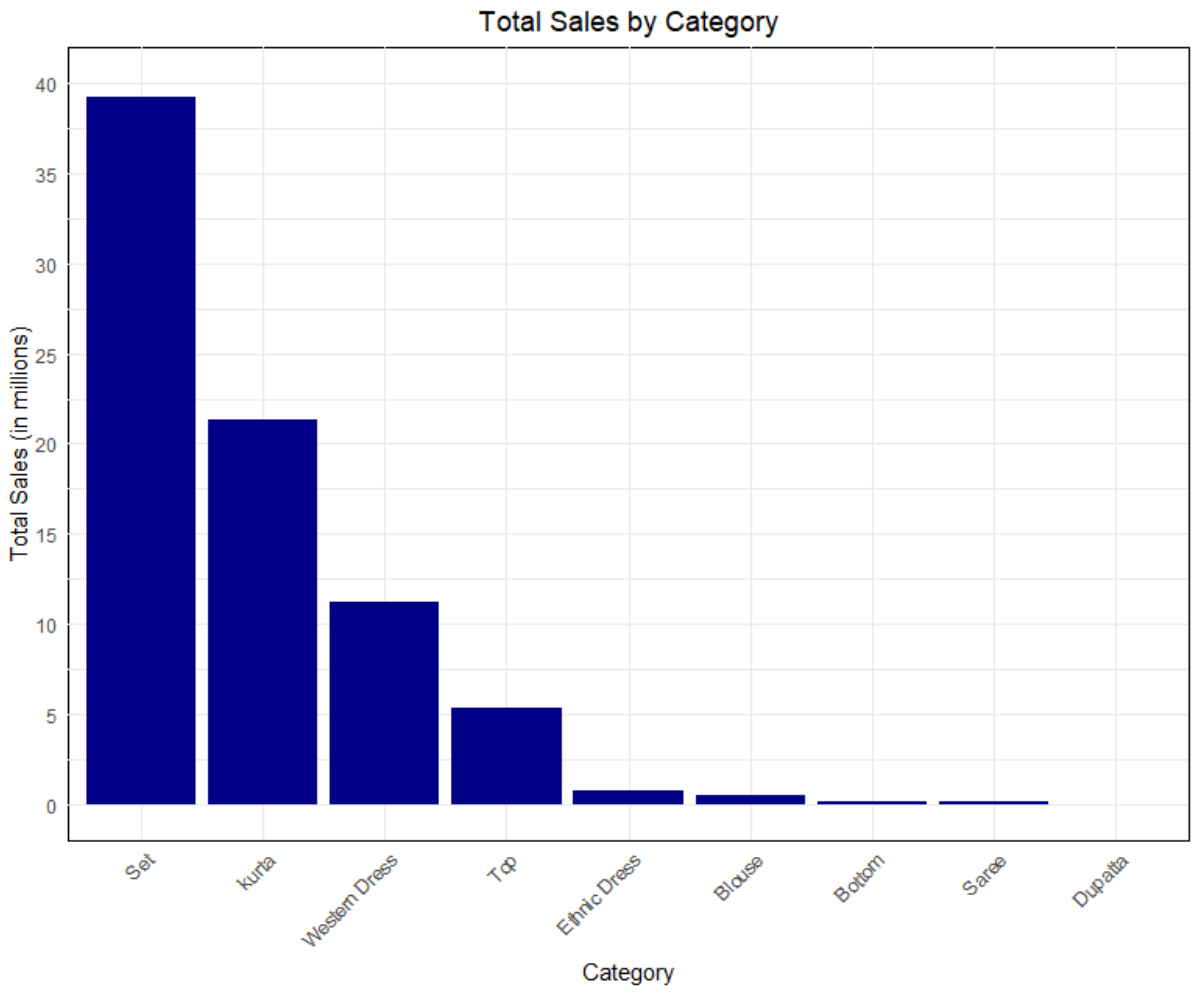
The highest_profit_categories dataset provides a comprehensive overview of sales performance across various product categories. It includes key metrics such as total sales, average sales amount, total quantity sold, and price per unit for each category. The dataset is structured as follows:

1. **Category:** The product category, such as Set, Saree, Western Dress, etc.

2. **Total_Sales:** The aggregated total sales amount for each category, indicating the overall revenue generated.
3. **Average_Amount:** The average sales amount per transaction for each category, providing insights into the typical transaction size.
4. **Total_Quantity_Sold:** The total quantity of products sold in each category, reflecting the volume of sales.
5. **Price_Per_Unit:** The average price per unit for each category, indicating the pricing strategy.

By focusing on these metrics, businesses can optimize their resource allocation, marketing efforts, and pricing strategies to enhance overall profitability and market presence.

Task 3 for Problem Statement 1



The bar plot displays the total sales for various product categories, including Blouse, Bottom, Dupatta, Ethnic Dress, Kurta, Saree, Set, Top, and Western Dress. The

y-axis represents the total sales in millions, and the x-axis specifies different categories of products. The height of each bar visually measures the sales data, comparing total sales across different categories. This visual representation facilitates rapid evaluation of the relative performance of different categories.

The bar plot illustrates substantial differences in total sales across many categories. When combined with unit pricing and cost of goods data for each category, this variation in sales can serve as an indicator for understanding the potential profitability and popularity of each category.

> **Key Insights**: Highest and Lowest Sales: The "Set" category shows the highest total sales, reaching approximately 40 million, indicating it is a top-performing and potentially high-profit category. On the other hand, categories like Blouse, Bottom, Saree, and Dupatta have relatively low sales, suggesting they might be less popular or less profitable.
>
> **Resource Allocation Implications:** To maximize profitability, it would be strategic to allocate more resources to the top four categories: "Set," "Kurta," "Western Dress," and "Top." This could include actions such as increasing advertisements for these products. Featuring these categories prominently on the first page of the selling platform. Implementing promotional activities to boost visibility and attract more customers.

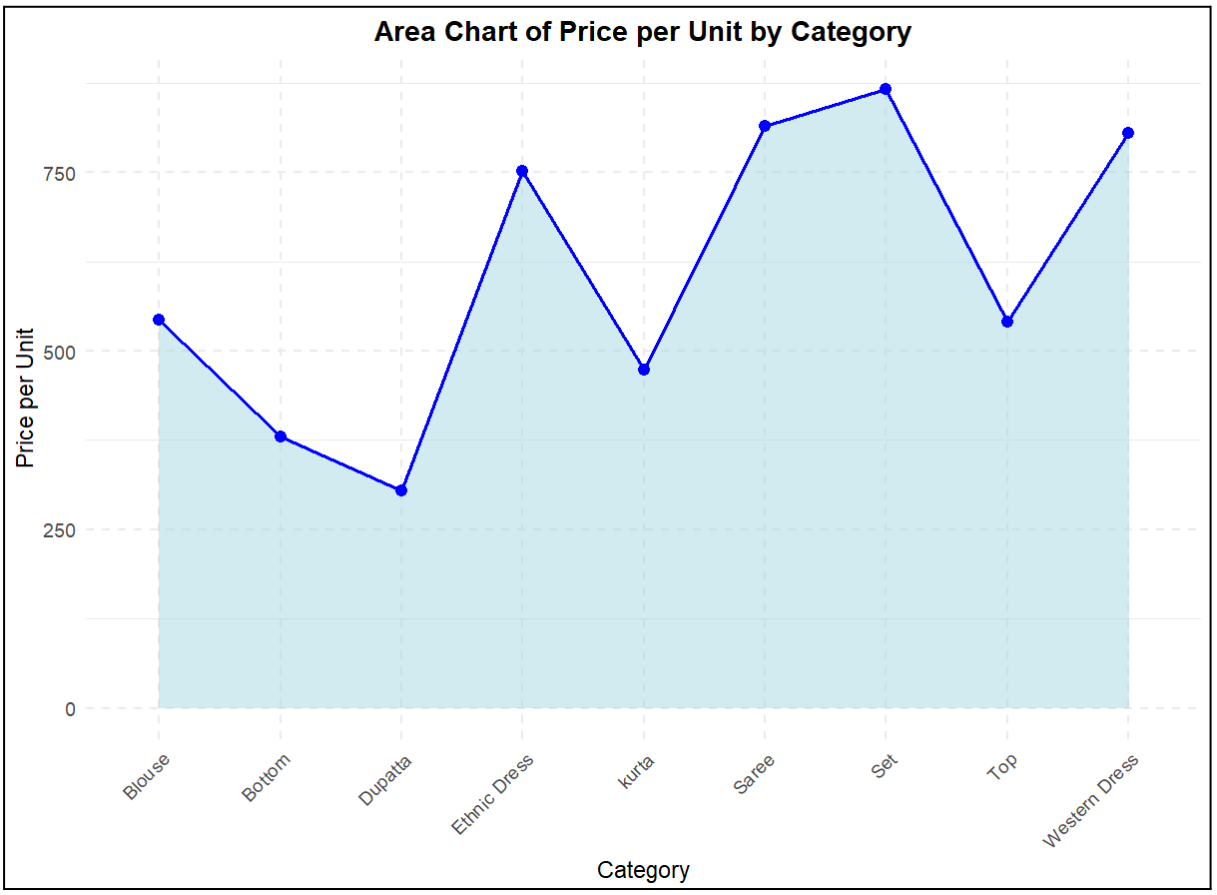
For instance, if the "Set" category not only has high total sales but also exhibits a high price per unit, it would be advantageous to invest more in its production and advertising. Conversely, if a category like "Dupatta" is not generating significant profits despite its lower sales, it may be necessary to reconsider the strategy, such as decreasing inventory levels or improving marketing efforts.

By analyzing the total sales data for different product categories, businesses can make informed decisions to maximize profitability and enhance market presence. The "Set" category's leading sales position highlights its strong market demand and potential for higher profit margins. Allocating more resources towards promoting and marketing top-performing categories like "Set," "Kurta," "Western Dress," and "Top" can drive further growth and capitalize on their popularity.

Conversely, the lower sales figures for categories such as Blouse, Bottom, Saree, and Dupatta indicate areas that may require strategic reassessment. Businesses

might consider optimizing their inventory, refining their marketing strategies, or even innovating product designs to boost appeal and sales in these less popular categories. By balancing investment and promotional efforts across both high and low-performing categories, businesses can achieve a more robust and profitable product portfolio.

In summary, focusing on top-performing categories while strategically managing lower-performing ones will enable businesses to enhance their overall profitability and market competitiveness.



The area chart displays the unit pricing for many product categories, including Blouse, Bottom, Dupatta, Ethnic Dress, Kurta, Saree, Set, Top, and Western Dress. The y-axis represents the unit price, and the X-axis specifies various types of products. The shaded area below the line visually measures the data, comparing prices across different categories. The shaded region facilitates rapid evaluation of the relative rankings of different categories.

The area in the chart above illustrates substantial fluctuations in the price per unit across many categories. When combined with the cost of goods and total sales data for each category, this variation in price can serve as an indicator for understanding the potential profitability per unit.

Higher and lower price thresholds: The pricing of Saree and Set has significant peaks, indicating that they may be positioned as premium products or have more considerable production costs. On the other hand, categories such as Bottom and Top are positioned towards the lower range of prices, suggesting that they consist of typical or cheaper items.

Resource allocation implications: When the company combines price information with profitability data to optimise its inventory and advertising campaigns, it can be advantageous to spend more money on production and advertising. For instance, if Set exhibits both a high price and high profitability, spending more money on its production and advertising would be advantageous. On the other hand, if Dupatta is not generating significant profits despite its low prices, it may be necessary to reconsider the strategy, such as decreasing inventory levels.

The area chart is helpful for visually representing pricing strategies in several categories. The corporation can conduct extensive studies to investigate the factors behind the increased pricing of certain categories and their correlation with consumer demand and manufacturing costs, which are crucial for maximising profit margins. The data shown above may additionally support strategic decisions in inventory management and targeted marketing, leading to more efficient resource allocation and potentially increased overall profitability.

Problem Statement 2

Comparative Analysis of Domestic and International Sales Performance

Problem: Sales and Profitability differ between domestic and international markets, creating difficulties in understanding market-specific demand and optimizing inventory and marketing strategies effectively.

Solution: Comparatively examine sales information from the same time period from both domestic (Amazon) and foreign markets. Determine the patterns and trends in each market's revenue, profit margins, and sales volume. Examine monthly and seasonal fluctuations to learn more about the performance and demand of a certain market.

Result: By comparing sales patterns between domestic and international markets, the company can enhance demand forecasting, optimize inventory levels, and plan targeted marketing campaigns. This strategy will boost profitability in a variety of markets and improve resource allocation.

Task 2 for Problem Statement 2

Step 1: Data Manipulation and Cleaning Techniques

```
> # Load necessary libraries
> library(dplyr)
> library(lubridate)
> library(ggplot2)
>
> # Load datasets
> amazon_sales <- read.csv("Amazon Sale Report.csv")
> international_sales <- read.csv("International sale Report.csv")
>
> # Convert date columns to Date type
> amazon_sales$Date <- as.Date(amazon_sales$Date, format = "%m-%d-%y")
> international_sales$DATE <- as.Date(international_sales$DATE, format = "%m-%d-%y")
>
> # Filter out cancelled orders from the Amazon sales data
> amazon_sales <- amazon_sales %>%
+   filter(Status != "Cancelled")
>
> # Omit rows with any NA values
> amazon_sales <- na.omit(amazon_sales)
> international_sales <- na.omit(international_sales)
```

```
> # Rename columns to standardize across datasets
> # Select required variables
> # Convert relevant columns to numeric
> # Filter datasets for April 2022
> # Aggregate data by date
> amazon_sales_april <- amazon_sales %>%
+   rename(Quantity = Qty) %>%
+   select(Date, Quantity, Amount) %>%
+   mutate(Quantity = as.numeric(Quantity), Amount = as.numeric(Amount)) %>%
+   filter(Date >= as.Date("2022-04-01") & Date <= as.Date("2022-04-30")) %>%
+   group_by(Date) %>%
+   summarize(Domestic_Quantity = sum(Quantity), Domestic_Amount = sum(Amount))
>
> international_sales_april <- international_sales %>%
+   rename(Quantity = PCS, Amount = GROSS.AMT, Date = DATE) %>%
+   select(Date, Quantity, Amount) %>%
+   mutate(Quantity = as.numeric(Quantity), Amount = as.numeric(Amount)) %>%
+   filter(Date >= as.Date("2022-04-01") & Date <= as.Date("2022-04-30")) %>%
+   group_by(Date) %>%
+   summarize(International_Quantity = sum(Quantity), International_Amount = sum(Amoun
t))
```

First, we loaded the necessary libraries and the datasets containing domestic (Amazon Sale Report) and international sales data. The date columns in both datasets were converted to the Date type to ensure consistency and enable time-based filtering and analysis. Next, we filtered out transactions marked as

"Cancelled" in the Amazon sales dataset to ensure that only completed transactions were included in the analysis. We then omitted rows with any missing values (NAs) from both datasets to clean the data and ensure completeness. To standardize the column names across the datasets, we renamed the columns and selected only the required variables (Date, Quantity, Amount). Additionally, we converted the relevant columns to numeric types to facilitate accurate calculations and analysis. We then filtered both datasets to include only transactions within the date range of April 2022. This step ensured that we focused our analysis on the specified month.

Step 2: Data Joining Techniques

```
> # Combine datasets with full join to include all dates from both datasets
> combined_data_april <- full_join(amazon_sales_april, international_sales_april, by
="Date")
>
> # Handle potential NAs introduced by the join
> combined_data_april <- combined_data_april %>%
+   mutate(
+     Domestic_Quantity = ifelse(is.na(Domestic_Quantity), 0, Domestic_Quantity),
+     Domestic_Amount = ifelse(is.na(Domestic_Amount), 0, Domestic_Amount),
+     International_Quantity = ifelse(is.na(International_Quantity), 0, International_
Quantity),
+     International_Amount = ifelse(is.na(International_Amount), 0, International_Amou
nt)
+   )
```

After filtering the datasets to include only April 2022 transactions, we aggregated the sales data by date for both the Amazon and international datasets. This involved summing the quantities and amounts for each date, reducing the number of rows and simplifying the analysis. We then performed a full join on the Date column to combine the aggregated data from both datasets. This ensured that all dates from both datasets were included, even if there were no matching transactions on a given date. After joining, we handled any missing values (NA) introduced by the join operation by replacing them with zeros in the Quantity and Amount columns, accurately representing days with no sales.

Description of the ‘combined_data_april’ Dataset

ETW2001_Assessment3.R × combined_data_april ×

Filter

	Date	Domestic_Quantity	Domestic_Amount	International_Quantity	International_Amount
1	2022-04-01	1237	779737	30	33575.00
2	2022-04-02	1302	811110	0	0.00
3	2022-04-03	1453	925049	0	0.00
4	2022-04-04	1270	813708	16	10240.00
5	2022-04-05	1416	879694	307	380453.50
6	2022-04-06	1361	815553	102	103515.00
7	2022-04-07	1296	814016	60	53260.00
8	2022-04-08	1435	922614	74	67425.00
9	2022-04-09	1377	873568	0	0.00
10	2022-04-10	1557	987630	0	0.00

Showing 1 to 10 of 30 entries, 5 total columns

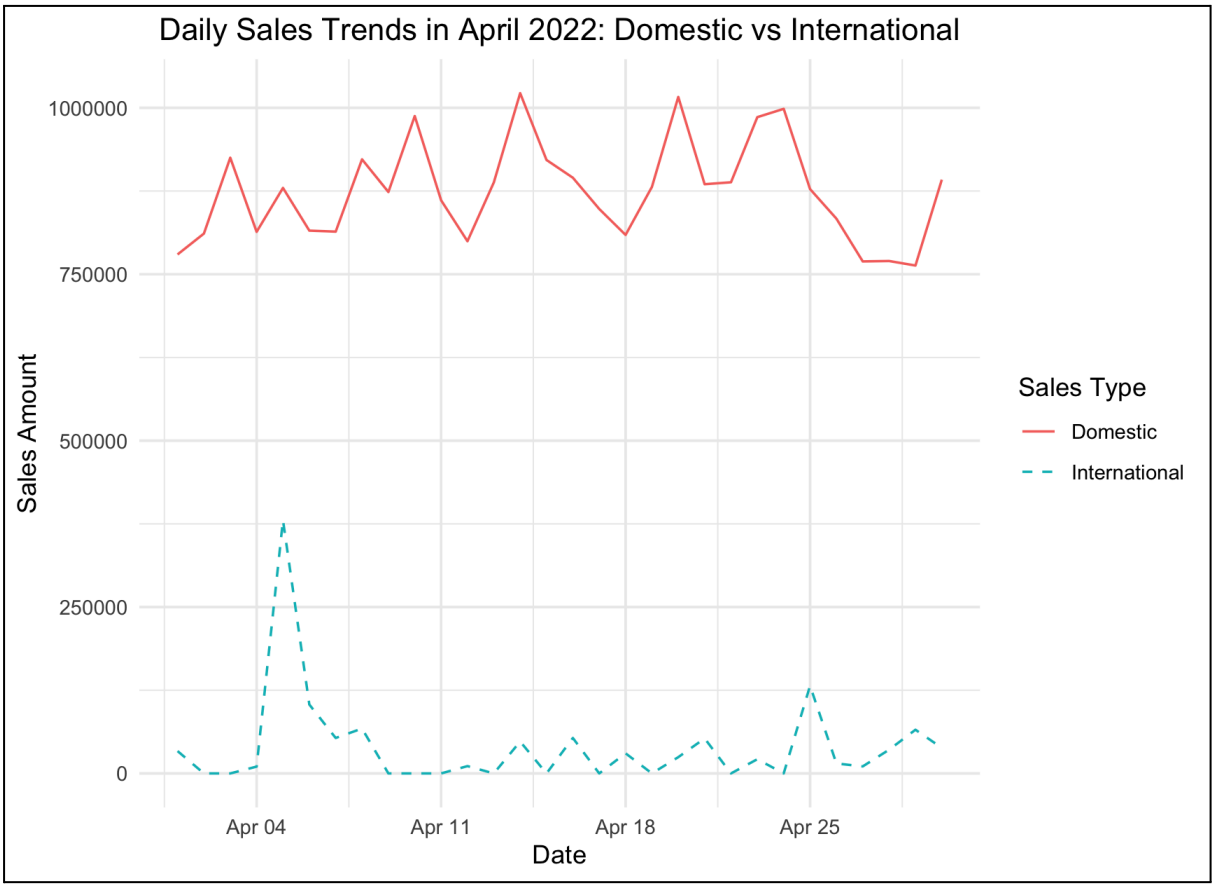
The combined_data_april dataset includes a comprehensive overview of aggregated sales data for the month of April 2022 from both domestic (Amazon) and international markets. It consists of the following columns:

1. **Date**: This column includes the dates of each transaction, enabling a day-by-day analysis of sales trends over April 2022.
2. **Domestic_Quantity**: This column records the aggregated quantity of products sold domestically through Amazon for each day in April 2022. It captures the sales volume for domestic transactions.
3. **Domestic_Amount**: This column represents the aggregated total sales amount from domestic transactions through Amazon for each day in April 2022. It reflects the revenue generated from domestic sales.
4. **International_Quantity**: This column captures the aggregated quantity of products sold internationally for each day in April 2022. It provides insights into the sales volume in international markets.
5. **International_Amount**: This column shows the aggregated total sales amount from international transactions for each day in April 2022, indicating the revenue generated from international sales.

By focusing on the month of April 2022 and aggregating the data by date, this combined_data_april dataset ensures that comparisons between domestic and international sales are made for the same month. This dataset allows for a detailed

analysis of sales performance in April, providing valuable insights into market-specific demand and revenue generation during this period.

Task 3 for Problem Statement 2



The graph comparing daily sales trends for April 2022 between domestic and international markets reveals several key insights that can inform demand forecasting, inventory management, and marketing strategies.

The domestic market, represented by the red solid line, shows higher and more consistent sales amounts throughout the month. This indicates a steady demand, allowing for predictable revenue generation and reliable inventory management. The consistent sales suggest a stable customer base and regular purchasing patterns in the domestic market.

In contrast, the international market, shown by the blue dashed line, exhibits notable volatility with significant sales spikes around April 4 and April 25. These spikes could result from specific promotions or product launches. However, apart from these

spikes, the baseline sales in the international market are relatively low, indicating fragmented and less predictable demand.

This volatility in the international market presents both challenges and opportunities. The unpredictable spikes require flexible inventory management, potentially including safety stock to handle sudden increases in demand. Additionally, understanding the drivers behind these spikes can help plan targeted marketing campaigns to boost international sales consistently.

The higher and consistent revenue from the domestic market suggests prioritizing resources like inventory, marketing, and customer service to maintain and enhance performance. For the international market, resources might be better allocated to understanding market dynamics, optimizing promotional timing, and improving supply chain responsiveness.

In summary, the insights highlight the steady performance of the domestic market, enabling predictable inventory management and consistent revenue. In contrast, the international market's volatility requires a dynamic approach but also offers opportunities for targeted marketing. Leveraging these insights can enhance demand forecasting, optimize inventory levels, and plan effective marketing campaigns, leading to better resource allocation and increased profitability across both markets.

Problem Statement 3

Profitability Analysis by Product Color and Size

Problem: Lack of inventory turnover of colour and size option for the following products might result in decreased sales volumes and affected overall revenue.

Solution: Analyze the relationship between product color variations and their corresponding sales amount. Analyze the relationship between product size variations and their corresponding sales amount. Quantify the influence of color and size on sales revenue and profitability metrics.

Result: Gain actionable insights into how different product colors and sizes affect financial outcomes. This analysis benefits to optimize product assortments and marketing strategies based on color preferences and sizes preferences to enhance profitability and customer satisfaction.

Task 2 for Problem Statement 3

Step 1: Data Manipulation

```
> library(dplyr)
> library(ggplot2)
> # Problem Statement 3
> Amazon = read.csv("Amazon Sale Report.csv")
> Report = read.csv("Sale Report.csv")
> # Task 2
> Amazon_Sale_Report = inner_join(Amazon, Report, by = c('SKU' = 'SKU.Code'), relationship = "many-to-many")
>
```

Read the Amazon Sale Report.csv file and Sale Report.csv file to investigate the problem statement 3 stated above. Next, using inner_join() function to find all the observations that match both dataset.

Step 2: Data Structure

```
> str(Amazon_Sale_Report) # data structure
'data.frame':   121269 obs. of  30 variables:
 $ index.x           : int  0 1 2 3 4 5 6 7 9 11 ...
 $ Order.ID          : chr  "405-8078784-5731545" "171-9198151-1101146" "404-0687676-7273146" "403-9615377-8133951" ...
 $ Date              : chr  "04-30-22" "04-30-22" "04-30-22" "04-30-22" ...
 $ Status            : chr  "Cancelled" "Shipped - Delivered to Buyer" "Shipped" "Cancelled" ...
 $ Fulfilment        : chr  "Merchant" "Merchant" "Amazon" "Merchant" ...
 $ Sales.Channel     : chr  "Amazon.in" "Amazon.in" "Amazon.in" "Amazon.in" ...
 $ ship.service.level: chr  "Standard" "Standard" "Expedited" "Standard" ...
 $ Style             : chr  "SET389" "JNE3781" "JNE3371" "JO341" ...
 $ SKU               : chr  "SET389-KR-NP-S" "JNE3781-KR-XXXL" "JNE3371-KR-XL" "JO341-DR-L" ...
 $ Category.x        : chr  "Set" "kurta" "kurta" "Western Dress" ...
 $ Size.x            : chr  "S" "3XL" "XL" "L" ...
 $ ASIN              : chr  "B09KXVBD7Z" "B09K3WFS32" "B07WV4JV4D" "B099NRCT7B" ...
 $ Courier.Status    : chr  "" "Shipped" "Shipped" "" ...
 $ Qty               : int  0 1 1 0 1 1 1 1 1 1 ...
 $ currency          : chr  "INR" "INR" "INR" "INR" ...
 $ Amount            : num  648 406 329 753 574 ...
 $ ship.city         : chr  "MUMBAI" "BENGALURU" "NAVI MUMBAI" "PUDUCHERRY" ...
 $ ship.state        : chr  "MAHARASHTRA" "KARNATAKA" "MAHARASHTRA" "PUDUCHERRY" ...
 $ ship.postal.code  : num  400081 560085 410210 605008 600073 ...
 $ ship.country      : chr  "IN" "IN" "IN" "IN" ...
 $ promotion.ids     : chr  "" "Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27EJC,Amazon PLCC Free-Financing Un
iversal Merchant A"| __truncated__ "IN Core Free Shipping 2015/04/08 23-48-5-108" "" ...
 $ B2B               : chr  "False" "False" "True" "False" ...
 $ fulfilled.by      : chr  "Easy Ship" "" "Easy Ship" ...
 $ Unnamed..22       : chr  "" "" "" "" ...
 $ index.y           : int  8978 5971 3805 2530 5336 8242 1096 3945 4261 4463 ...
 $ Design.No.        : chr  "SET389" "JNE3781" "JNE3371" "JO341" ...
 $ Stock             : num  32 96 4 193 6 81 1 423 2 16 ...
 $ Category.y        : chr  "SET" "KURTA" "KURTA" "DRESS" ...
 $ Size.y            : chr  "S" "XXXL" "XL" "L" ...
 $ Color             : chr  "White" "Green" "Light Green" "Blue" ...
```

Figure 1: Data Structure of Amazon_Sale_Report

According to Figure 1, the total observations is 121269 and total number of variables is 30. The Amazon_Sale_Report dataset contains 24 variables in character format, 3 integer variables and 3 numerical variables based on the data types.

Step 3: Data Missing

```
> colSums(is.na(Amazon_Sale_Report)) # number of column contain missing data
          index.x           Order.ID               Date             Status        Fulfilment      Sales.Channel
                0                  0                  0                  0                 0                  0
ship.service.level              Style                SKU        Category.x             Size.x               ASIN
                0                  0                  0                  0                 0                  0
    Courier.Status                Qty           currency             Amount          ship.city         ship.state
                0                  0                  0               7339                 0                  0
  ship.postal.code       ship.country      promotion.ids                B2B       fulfilled.by       Unnamed..22
               32                  0                  0                  0                 0                  0
          index.y        Design.No.              Stock        Category.y             Size.y              Color
                0                  0                  0                  0                 0                  0
```

Figure 2: Missing value variables in Amazon_Sale_Report dataset

There are 2 variables contain missing values (NAs), which are the Amount and ship.postal.code. According to Figure 2, Amount has 7339 missing values and ship.postal.code has 32 missing values.

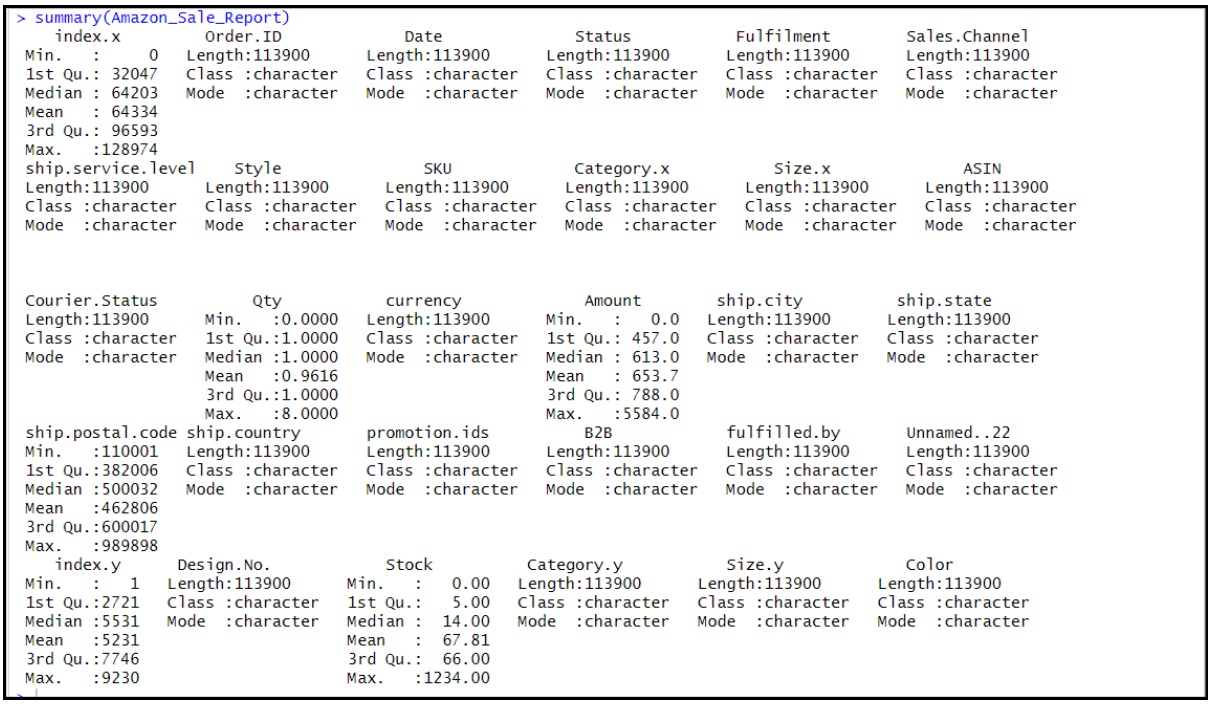
Step 4: Data Cleaning

```
> Amazon_Sale_Report = na.omit(Amazon_Sale_Report) # remove missing value
> nrow(Amazon_Sale_Report) # number of rows
[1] 113900
> ncol(Amazon_Sale_Report) # number of columns
[1] 30
```

Figure 3: Number of rows and columns in cleaned Amazon_Sale_Report dataset

Using na.omit() function to remove all the missing values in the dataset. After removing the missing vlaues in the dataset, the total number of observations is 113900 records and total number of variables remains 30.

Step 5: Data Description

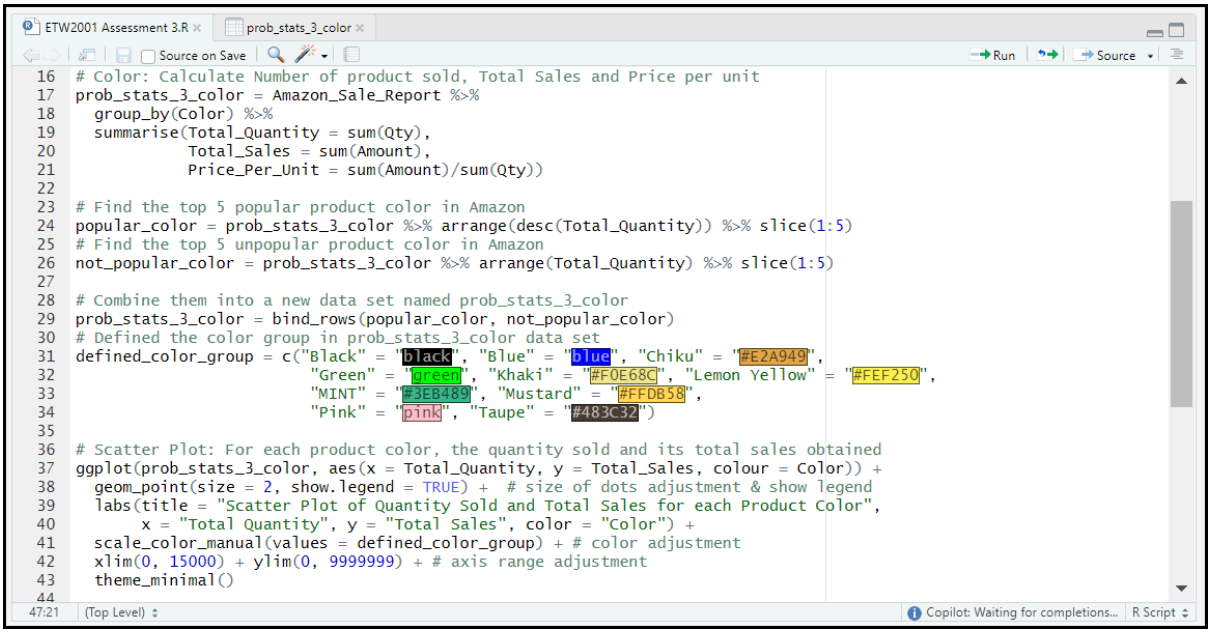


```
> summary(Amazon_Sale_Report)
    index.x        Order.ID            Date              Status          Fulfilment        Sales.Channel
 Min.   :     0   Length:113900      Length:113900      Length:113900      Length:113900      Length:113900
 1st Qu.: 32047   Class :character   Class :character   Class :character   Class :character   Class :character
 Median : 64203   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   : 64334
 3rd Qu.: 96593
 Max.   :128974
 ship.service.level    Style              SKU            Category.x          Size.x             ASIN
 Length:113900      Length:113900      Length:113900      Length:113900      Length:113900      Length:113900
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



 Courier.Status          Qty           currency            Amount        ship.city          ship.state
 Length:113900      Min.   :0.0000   Length:113900      Min.   :   0.0   Length:113900      Length:113900
 Class :character   1st Qu.:1.0000   Class :character   1st Qu.: 457.0   Class :character   Class :character
 Mode  :character   Median :1.0000   Mode  :character   Median : 613.0   Mode  :character   Mode  :character
                    Mean   :0.9616                      Mean   : 653.7
                    3rd Qu.:1.0000                      3rd Qu.: 788.0
                    Max.   :8.0000                      Max.   :5584.0
 ship.postal.code ship.country       promotion.ids          B2B            fulfilled.by       Unnamed..22
 Min.   :110001   Length:113900      Length:113900      Length:113900      Length:113900      Length:113900
 1st Qu.:382006   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :500032   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :462806
 3rd Qu.:600017
 Max.   :989898
    index.y      Design.No.            Stock          Category.y          Size.y             Color
 Min.   :   1   Length:113900      Min.   :   0.00   Length:113900      Length:113900      Length:113900
 1st Qu.:2721   Class :character   1st Qu.:   5.00   Class :character   Class :character   Class :character
 Median :5531   Mode  :character   Median :  14.00   Mode  :character   Mode  :character   Mode  :character
 Mean   :5231                      Mean   :  67.81
 3rd Qu.:7746                      3rd Qu.:  66.00
 Max.   :9230                      Max.   :1234.00
```

Figure 4: Summary of cleaned Amazon_Sale_Report dataset

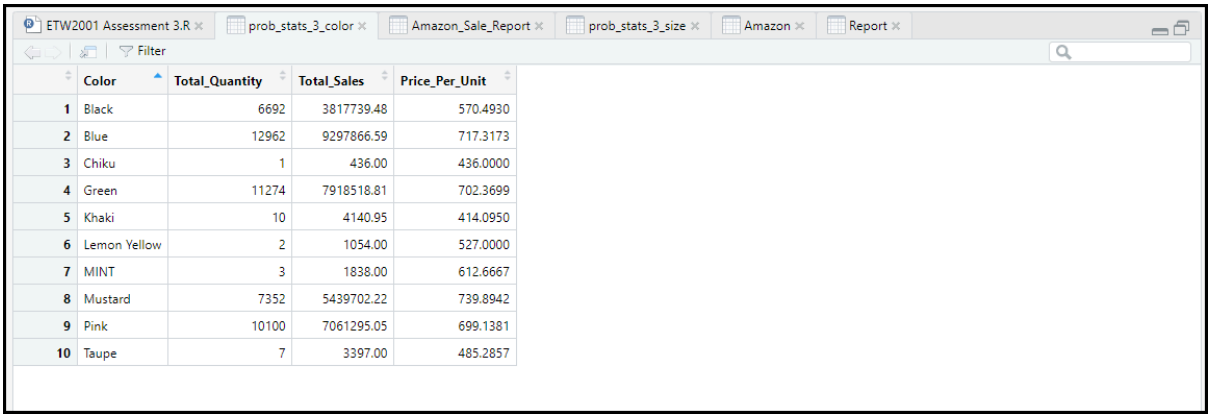
According to the Figure 4, maximum number of orders is 8 orders and maximum amount among all the orders is 5584. Next, the maximum number of stocks among all the SKU is 1234 items.

Step 6: prob_stats_3_color dataset



```r
# Color: Calculate Number of product sold, Total Sales and Price per unit
prob_stats_3_color = Amazon_Sale_Report %>%
  group_by(Color) %>%
  summarise(Total_Quantity = sum(Qty),
            Total_Sales = sum(Amount),
            Price_Per_Unit = sum(Amount)/sum(Qty))

# Find the top 5 popular product color in Amazon
popular_color = prob_stats_3_color %>% arrange(desc(Total_Quantity)) %>% slice(1:5)
# Find the top 5 unpopular product color in Amazon
not_popular_color = prob_stats_3_color %>% arrange(Total_Quantity) %>% slice(1:5)

# Combine them into a new data set named prob_stats_3_color
prob_stats_3_color = bind_rows(popular_color, not_popular_color)
# Defined the color group in prob_stats_3_color data set
defined_color_group = c("Black" = "black", "Blue" = "blue", "Chiku" = "#E2A949",
                        "Green" = "green", "Khaki" = "#F0E68C", "Lemon Yellow" = "#FEF250",
                        "MINT" = "#3EB489", "Mustard" = "#FFDB58",
                        "Pink" = "pink", "Taupe" = "#483C32")

# Scatter Plot: For each product color, the quantity sold and its total sales obtained
ggplot(prob_stats_3_color, aes(x = Total_Quantity, y = Total_Sales, colour = Color)) +
  geom_point(size = 2, show.legend = TRUE) +  # size of dots adjustment & show legend
  labs(title = "Scatter Plot of Quantity Sold and Total Sales for each Product Color",
       x = "Total Quantity", y = "Total Sales", color = "Color") +
  scale_color_manual(values = defined_color_group) + # color adjustment
  xlim(0, 15000) + ylim(0, 9999999) + # axis range adjustment
  theme_minimal()
```

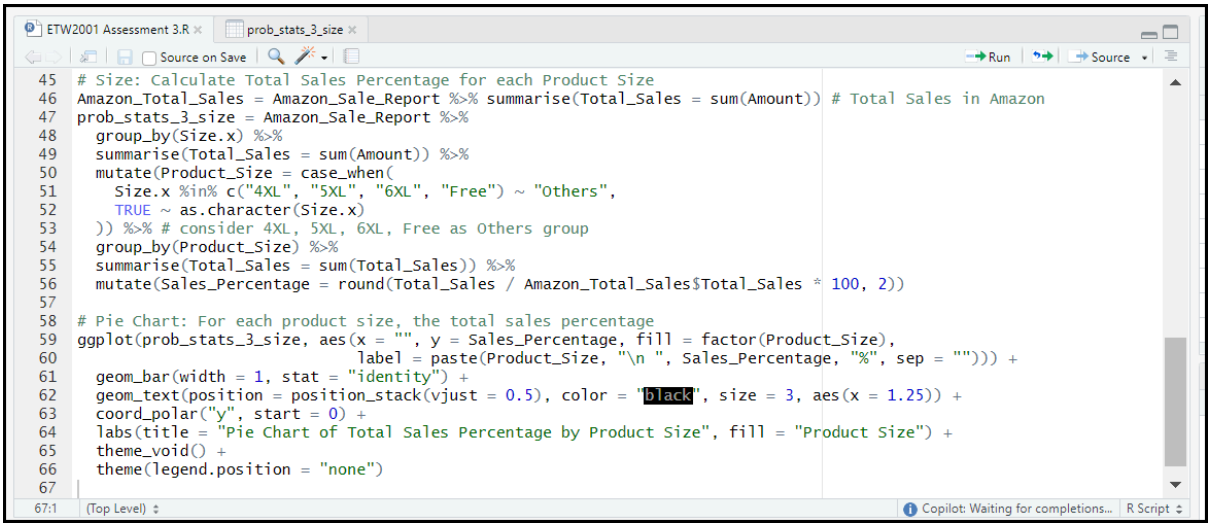
First, calculate the total number of product sold (Total_Quantity), total sales of product (Total_Sales), and price per product unit (Price_Per_Unit) for each product colors in the cleaned Amazon_Sale_Report dataset. Besides that, find the top 5 popular product colors and top 5 unpopular product colors by filtering their total number of product sold and total sales of product sold. Combine the results as new dataset and named in as prob_stats_3_color. Defined the color group in the prob_stats_3_color dataset and named it as defined_color_group for further plotting purpose. Furthermore, Plot a scatter plot to show the quantity sold and its total sales obtained for each product color. Next, dot size is adjusted by using geom_point(size = 2), x-axis range is adjusted from 0 to 15000, y-axis range is adjusted from 0 to 9999999, and using the defined_color_group as our dot colors.



	Color	Total_Quantity	Total_Sales	Price_Per_Unit
1	Black	6692	3817739.48	570.4930
2	Blue	12962	9297866.59	717.3173
3	Chiku	1	436.00	436.0000
4	Green	11274	7918518.81	702.3699
5	Khaki	10	4140.95	414.0950
6	Lemon Yellow	2	1054.00	527.0000
7	MINT	3	1838.00	612.6667
8	Mustard	7352	5439702.22	739.8942
9	Pink	10100	7061295.05	699.1381
10	Taupe	7	3397.00	485.2857

Figure 5: The prob_stats_3_color dataset

Step 7: prob_stats_3_size dataset



```
# Size: Calculate Total Sales Percentage for each Product Size
Amazon_Total_Sales = Amazon_Sale_Report %>% summarise(Total_Sales = sum(Amount)) # Total Sales in Amazon
prob_stats_3_size = Amazon_Sale_Report %>%
  group_by(Size.x) %>%
  summarise(Total_Sales = sum(Amount)) %>%
  mutate(Product_Size = case_when(
    Size.x %in% c("4XL", "5XL", "6XL", "Free") ~ "Others",
    TRUE ~ as.character(Size.x)
  )) %>% # consider 4XL, 5XL, 6XL, Free as Others group
  group_by(Product_Size) %>%
  summarise(Total_Sales = sum(Total_Sales)) %>%
  mutate(Sales_Percentage = round(Total_Sales / Amazon_Total_Sales$Total_Sales * 100, 2))

# Pie Chart: For each product size, the total sales percentage
ggplot(prob_stats_3_size, aes(x = "", y = Sales_Percentage, fill = factor(Product_Size),
                              label = paste(Product_Size, "\n ", Sales_Percentage, "%", sep = ""))) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(position = position_stack(vjust = 0.5), color = "black", size = 3, aes(x = 1.25)) +
  coord_polar("y", start = 0) +
  labs(title = "Pie Chart of Total Sales Percentage by Product Size", fill = "Product Size") +
  theme_void() +
  theme(legend.position = "none")
```

	Total_Sales
1	74451603

First, calculate the total sales in Amazon_Sale_Report which is 74451603 and named it as Amazon_Total_Sales. Next, calculate the total sales percentage for each product size by grouping 4XL, 5XL, 6XL and Free size as Others since these sizes are minority and round the Sales_Percentage into two decimal places. Besides that, plot a pie chart that shows the total sales percentage for each product size. Adjust the data label in form of (‘Product Size’ nextline ‘Sales Percentage’ %)

	Product_Size	Total_Sales	Sales_Percentage
1	M	13095147.7	17.59
2	L	12678303.4	17.03
3	XL	11972814.8	16.08
4	S	10225595.2	13.73
5	XXL	10192005.3	13.69
6	3XL	8878817.1	11.93
7	XS	6748660.4	9.06
8	Others	660258.8	0.89

Figure 6: The prob_stats_3_size dataset

Task 3 for Problem Statement 3:

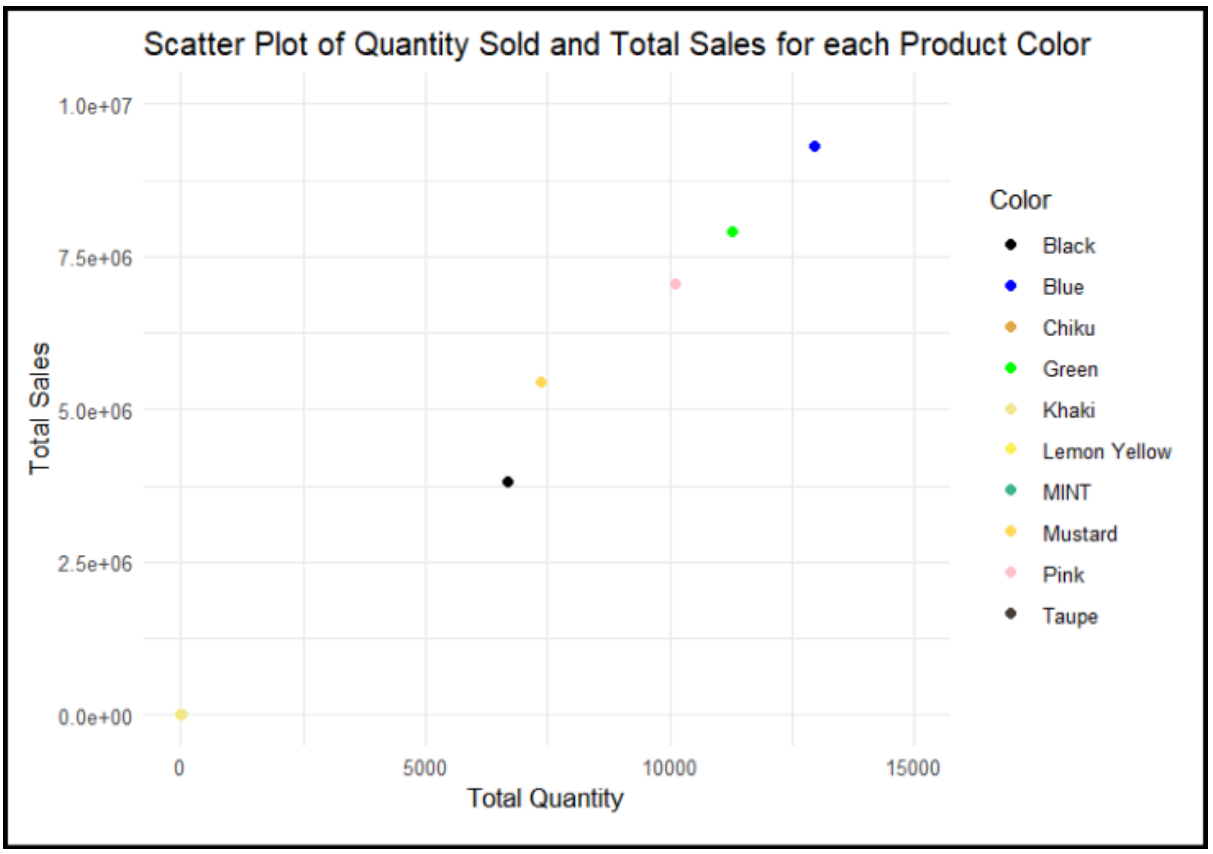


Figure 7: Scatter plot that showing Total Quantity and Total Sales for each Product Colors sold in Amazon

The availability of product color options can significantly influence consumer purchasing decisions, as certain colors may be more popular than others. According to the Figure 5 and Figure 7 above, we observed that blue color (12962 products sold), green color (11274 products sold) and pink color (10100 products sold) are the top three famous among all the product. It means that many consumers prefer to buy the product that in blue, green and pink color. Hence, sellers may focus more about the stock availability of these popular product color to ensure getting higher sales. It is because the demands of these product colors are slightly higher than other product colors. Conversely, products that in chiku color (1 product sold), lemon yellow color (2 products sold), mint color (3 products sold) are unlikable in the market. Thus, sellers may not restock the products in these color options or thinking on promotion and marketing strategies to clear the stocks.

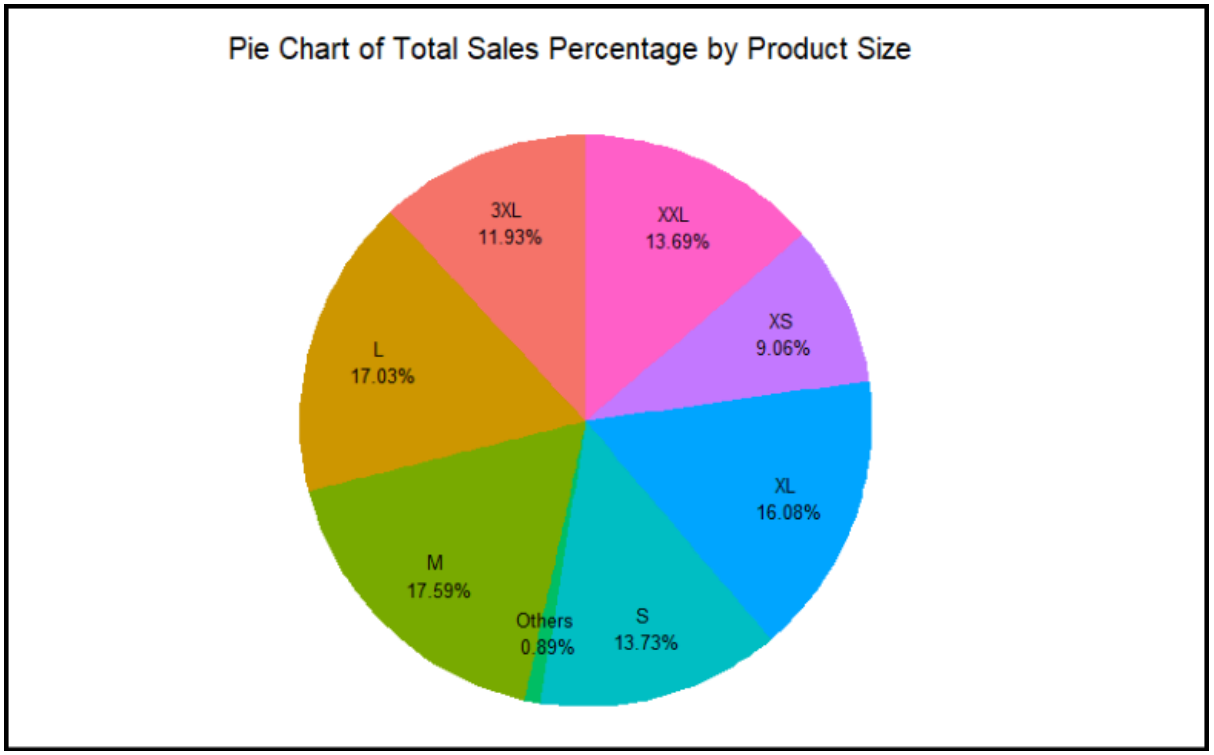


Figure 8: Pie Chart of Total Sales Percentage by Product Size

According to Figure 6 and Figure 8, the consumers normally purchase the product in M size, L size and XL size. It is because their sales percentages are quite similar which are 17.59%, 17.03% and 16.08% respectively. Hence, sellers in Amazon may need to prepare the stock availability for these sizes to fulfill the demand of consumer. However, the M size has the highest sales percentage which is 17.59% among these three product sizes. It means that sellers should prepare more quantity of product in M size compared with other size. In Others product size category (4XL, 5XL, 6XL and Free size), sellers may allocate less budget and cost on it because there are minor consumers will purchase these sizes (0.89% in total).