Sales Data Analysis

By: Badrinath Sanagavaram
Kamani Madasu
Shuchi Shah

INTRODUCTION

```
# A tibble: 6 × 14
  Region         Country `Item Type` `Sales Channel` `Order Priority` `Order Date`
  <chr>          <chr>   <chr>       <chr>           <chr>            <chr>
1 Sub-Saharan … Namibia Household   Offline         M                8/31/2015
2 Europe         Iceland Baby Food  Online          H                11/20/2010
3 Europe         Russia  Meat       Online          L                6/22/2017
4 Europe         Moldova Meat       Online          L                2/28/2012
5 Europe         Malta   Cereal     Online          M                8/12/2010
6 Asia           Indone… Meat       Online          H                8/20/2010
# ℹ 8 more variables: `Order ID` <dbl>, `Ship Date` <chr>, `Units Sold` <dbl>,
#   `Unit Price` <dbl>, `Unit Cost` <dbl>, `Total Revenue` <dbl>,
#   `Total Cost` <dbl>, `Total Profit` <dbl>
```

This expansive sales dataset, spanning 2009 to 2017, encapsulates over 50,000 records across diverse regions and product categories. It details sales dynamics, order priorities, financial metrics, and temporal aspects, offering insights crucial for understanding consumer behavior and strategic business optimization.

Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
Sub-Saharan	Namibia	Household	Offline	M	8/31/15	897751939	10/12/15	3604	668.27	502.54	2408445.08	1811154.16	597290.92
Europe	Iceland	Baby Food	Online	H	11/20/10	599480426	1/9/11	8435	255.28	159.42	2153286.8	1344707.7	808579.1
Europe	Russia	Meat	Online	L	6/22/17	538911855	6/25/17	4848	421.89	364.69	2045322.72	1768017.12	277305.6
Europe	Moldova	Meat	Online	L	2/28/12	459845054	3/20/12	7225	421.89	364.69	3048155.25	2634885.25	413270
Europe	Malta	Cereal	Online	M	8/12/10	626391351	9/13/10	1975	205.7	117.11	406257.5	231292.25	174965.25
Asia	Indonesia	Meat	Online	H	8/20/10	472974574	8/27/10	2542	421.89	364.69	1072444.38	927041.98	145402.4
Sub-Saharan	Djibouti	Household	Online	M	2/3/11	854331052	3/3/11	4398	668.27	502.54	2939051.46	2210170.92	728880.54
Europe	Greece	Household	Online	L	9/11/15	895509612	9/26/15	49	668.27	502.54	32745.23	24624.46	8120.77
Sub-Saharan	Cameroon	Cosmetics	Offline	M	1/31/14	241871583	2/4/14	4031	437.2	263.33	1762353.2	1061483.23	700869.97
Sub-Saharan	Nigeria	Cosmetics	Online	C	11/21/15	409090793	12/7/15	7911	437.2	263.33	3458689.2	2083203.63	1375485.57
Sub-Saharan	Senegal	Fruits	Offline	M	8/29/16	733153569	10/5/16	5288	9.33	6.92	49337.04	36592.96	12744.08
Middle East	Afghanistan	Cosmetics	Offline	L	10/21/16	620358741	12/1/16	6792	437.2	263.33	2969462.4	1788537.36	1180925.04
Asia	India	Vegetables	Online	C	3/21/10	897317636	4/5/10	5084	154.06	90.93	783241.04	462288.12	320952.92
Middle East	Lebanon	Vegetables	Online	L	10/15/10	660954082	11/19/10	9855	154.06	90.93	1518261.3	896115.15	622146.15
Middle East	Turkey	Office Supplie	Online	L	10/4/10	428504407	11/13/10	2831	651.21	524.96	1843575.51	1486161.76	357413.75
Middle East	Iraq	Cosmetics	Offline	M	10/14/14	787517440	10/19/14	2766	437.2	263.33	1209295.2	728370.78	480924.42
Sub-Saharan	Rwanda	Personal Care	Offline	M	6/15/13	145854508	7/8/13	445	81.73	56.67	36369.85	25218.15	11151.7
Europe	Ukraine	Baby Food	Offline	M	5/7/17	581689441	5/29/17	3687	255.28	159.42	941217.36	587781.54	353435.82
Europe	Finland	Office Supplie	Online	H	5/21/15	193508565	7/3/15	2339	651.21	524.96	1523180.19	1227881.44	295298.75
Sub-Saharan	South Sudan	Beverages	Offline	H	6/28/16	750110709	7/14/16	3283	47.45	31.79	155778.35	104366.57	51411.78
Central Ame	Antigua and	Fruits	Offline	M	7/6/15	940607202	8/12/15	5428	9.33	6.92	50643.24	37561.76	13081.48
Middle East	Kuwait	Personal Care	Online	C	2/18/12	424421870	3/23/12	4718	81.73	56.67	385602.14	267369.06	118233.08
Europe	United Kingd	Office Supplie	Offline	C	9/9/14	281291043	9/14/14	9125	651.21	524.96	5942291.25	4790260	1152031.25
Central Ame	Saint Kitts ar	Personal Care	Online	C	8/24/13	761263549	9/25/13	3656	81.73	56.67	298804.88	207185.52	91619.36
Central Ame	Antigua and	Personal Care	Offline	M	9/4/15	834700715	9/16/15	5345	81.73	56.67	436846.85	302901.15	133945.7
Central Ame	Saint Lucia	Cosmetics	Offline	L	2/26/17	442276370	3/28/17	8261	437.2	263.33	3611709.2	2175369.13	1436340.07
Middle East	Kuwait	Personal Care	Offline	H	7/15/15	944976842	8/18/15	8502	81.73	56.67	694868.46	481808.34	213060.12
Sub-Saharan	South Sudan	Office Supplie	Online	L	2/9/13	174100959	2/18/13	9197	651.21	524.96	5989178.37	4828057.12	1161121.25
Middle East	Tunisia	Snacks	Online	M	3/27/13	981260049	4/4/13	5509	152.58	97.44	840563.22	536796.96	303766.26

DATASET SUMMARY

- Region: Categorical data denoting geographical regions.
- Country: Character data representing sales countries.
- Item type: Categorical variable specifying product categories.
- Sales Channel: Character data indicating the sales medium (online/offline).
- Order Priority: Character data depicting the priority of orders (High/Medium/Low).
- Order date & Ship date: Date variables in MM/DD/YYYY format.
- Order ID: Integer data representing unique order identifiers.
- Units sold: Integer variable denoting the quantity sold.
- Unit Price, Unit Cost, Total Revenue, Total Cost, Total Profit: Decimal variables portraying financial aspects of sales.

```
> colnames(your_data)
 [1] "Region"         "Country"        "Item Type"      "Sales Channel"
 [5] "Order Priority" "Order Date"     "Order ID"       "Ship Date"
 [9] "Units Sold"     "Unit Price"     "Unit Cost"      "Total Revenue"
[13] "Total Cost"     "Total Profit"
```

```
    Region           Country           Item Type         Sales Channel
 Length:37554       Length:37554       Length:37554       Length:37554
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



 Order Priority      Order Date          Order ID          Ship Date
 Length:37554       Length:37554       Min.   :100013196   Length:37554
 Class :character   Class :character   1st Qu.:323151636   Class :character
 Mode  :character   Mode  :character   Median :549518812   Mode  :character
                                       Mean   :549364475
                                       3rd Qu.:776320504
                                       Max.   :999999463
   Units Sold      Unit Price       Unit Cost       Total Revenue
 Min.   :   1   Min.   :  9.33   Min.   :  6.92   Min.   :     28
 1st Qu.:2502   1st Qu.: 81.73   1st Qu.: 35.84   1st Qu.: 275987
 Median :5024   Median :154.06   Median : 97.44   Median : 779015
 Mean   :5007   Mean   :265.40   Mean   :187.14   Mean   :1323884
 3rd Qu.:7493   3rd Qu.:421.89   3rd Qu.:263.33   3rd Qu.:1811068
 Max.   :9999   Max.   :668.27   Max.   :524.96   Max.   :6682032
   Total Cost       Total Profit         Profit_Z          Revenue_Z
 Min.   :     21   Min.   :     7.2   Min.   :-1.033865   Min.   :-0.904226
 1st Qu.: 159424   1st Qu.:  93888.2   1st Qu.:-0.785344   1st Qu.:-0.715715
 Median : 466536   Median : 279144.9   Median :-0.294934   Median :-0.372092
 Mean   : 933396   Mean   : 390487.4   Mean   :-0.000189   Mean   : 0.000114
 3rd Qu.:1188631   3rd Qu.: 564063.0   3rd Qu.: 0.459299   3rd Qu.: 0.332916
 Max.   :5243300   Max.   :1738178.4   Max.   : 3.567408   Max.   : 3.660324
```

DATA CLEANING

Data cleaning is the process of identifying and rectifying errors, inconsistencies, and inaccuracies in a dataset to enhance its quality and reliability for analysis.

Data Cleaning involves the following processes:

1.Identify and Handle Missing Data

2.Detect and Handle Outliers

3.Address Inconsistent Data

4.Validate and Cleanse Categorical Data

IDENTIFYING AND HANDLING MISSING DATA

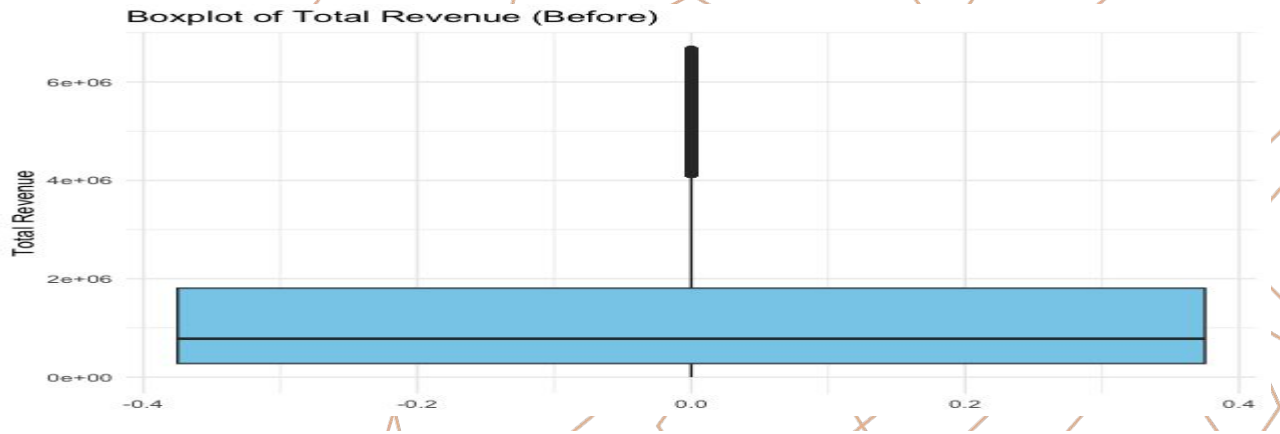
Upon reviewing the dataset, no missing values were identified; hence, there is no need for further action, such as imputation or deletion, to address missing data. This conclusion is drawn as the dataset is complete with no null values.

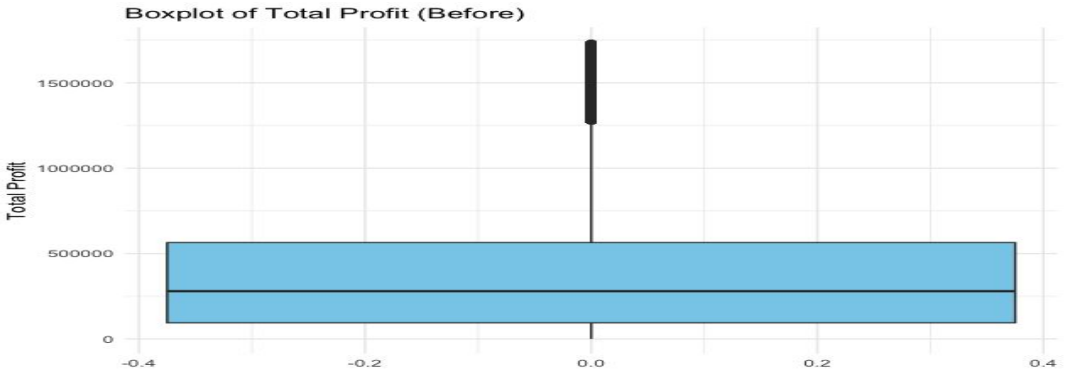
```
null_count <- colSums(is.na(your_data))

# Display columns with null values and their counts
print(null_count)
```

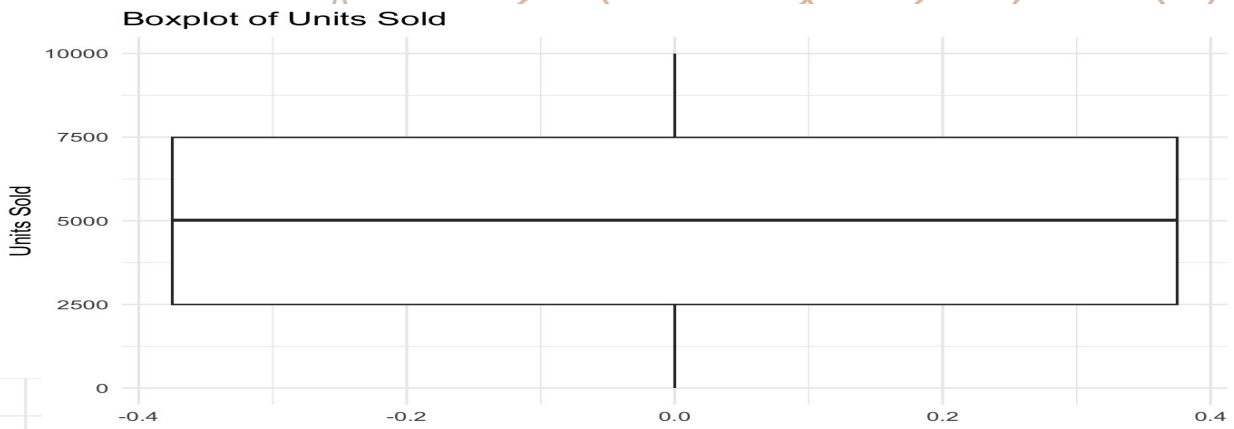
```
> print(null_count)
      Region        Country      Item Type  Sales Channel Order Priority
           0              0              0              0              0
  Order Date       Order ID      Ship Date     Units Sold     Unit Price
           0              0              0              0              0
   Unit Cost  Total Revenue     Total Cost   Total Profit
           0              0              0              0
```

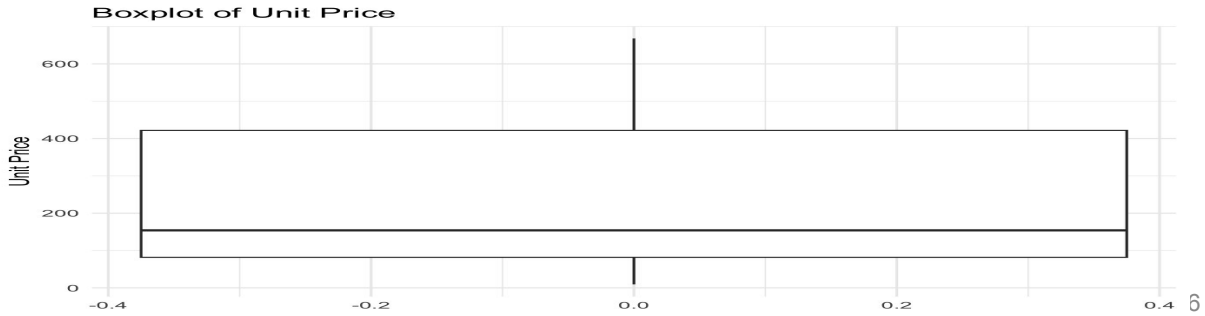
DETECTING OUTLIERS

I utilized visualizations such as box plots to pinpoint outliers and subsequently made decisions on whether to eliminate them or apply transformations to enhance data integrity.



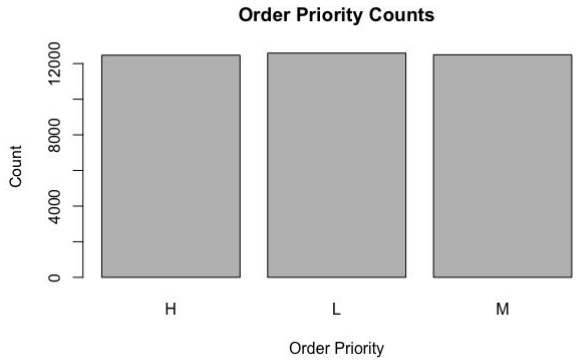


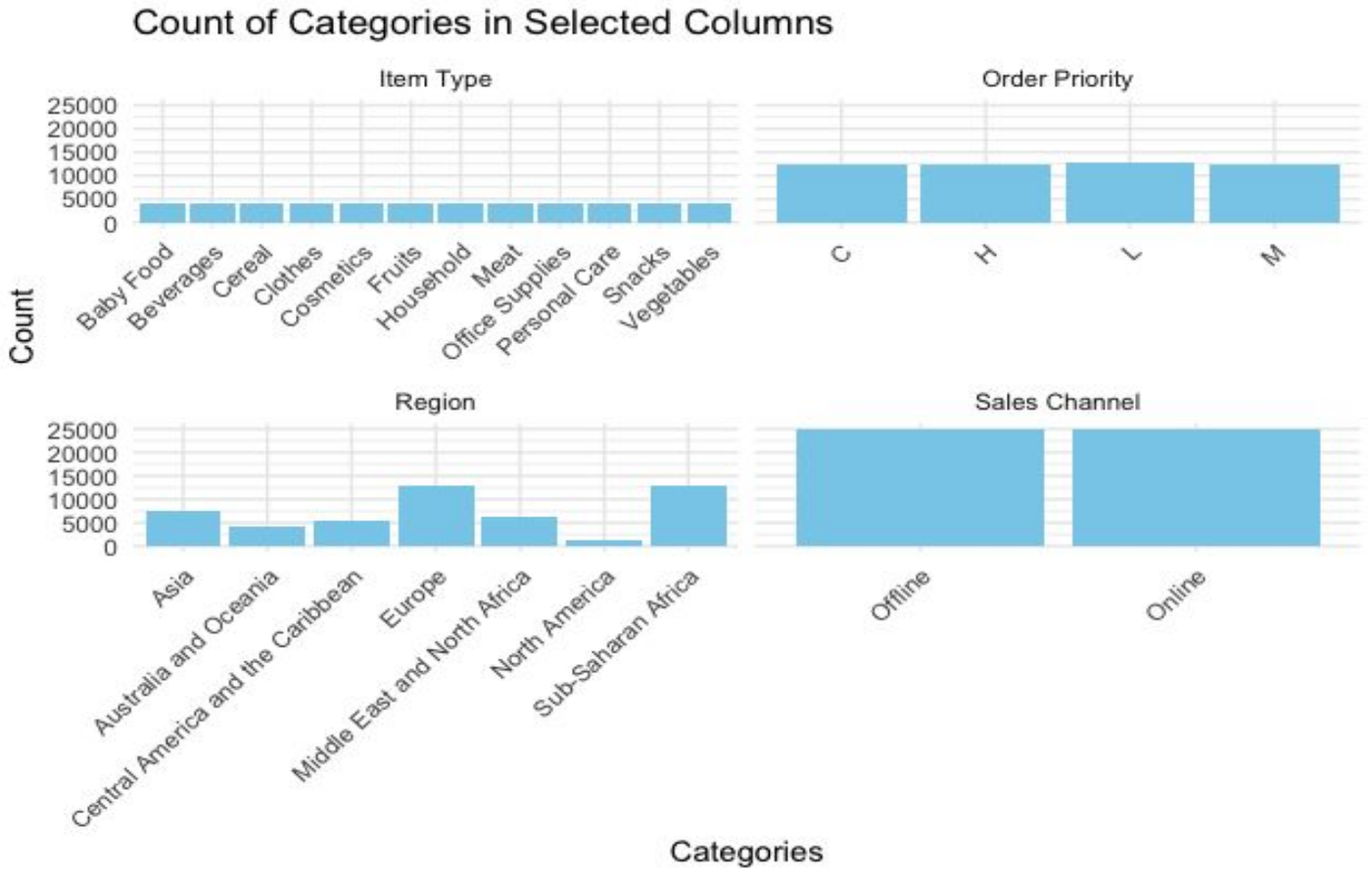




IDENTIFICATION OF DATA INCONSISTENCIES

Examine the dataset to identify any entries that exhibit inconsistencies or errors, and subsequently undertook the necessary measures to standardize or rectify such inconsistent data, ensuring the overall maintenance of data consistency and integrity.





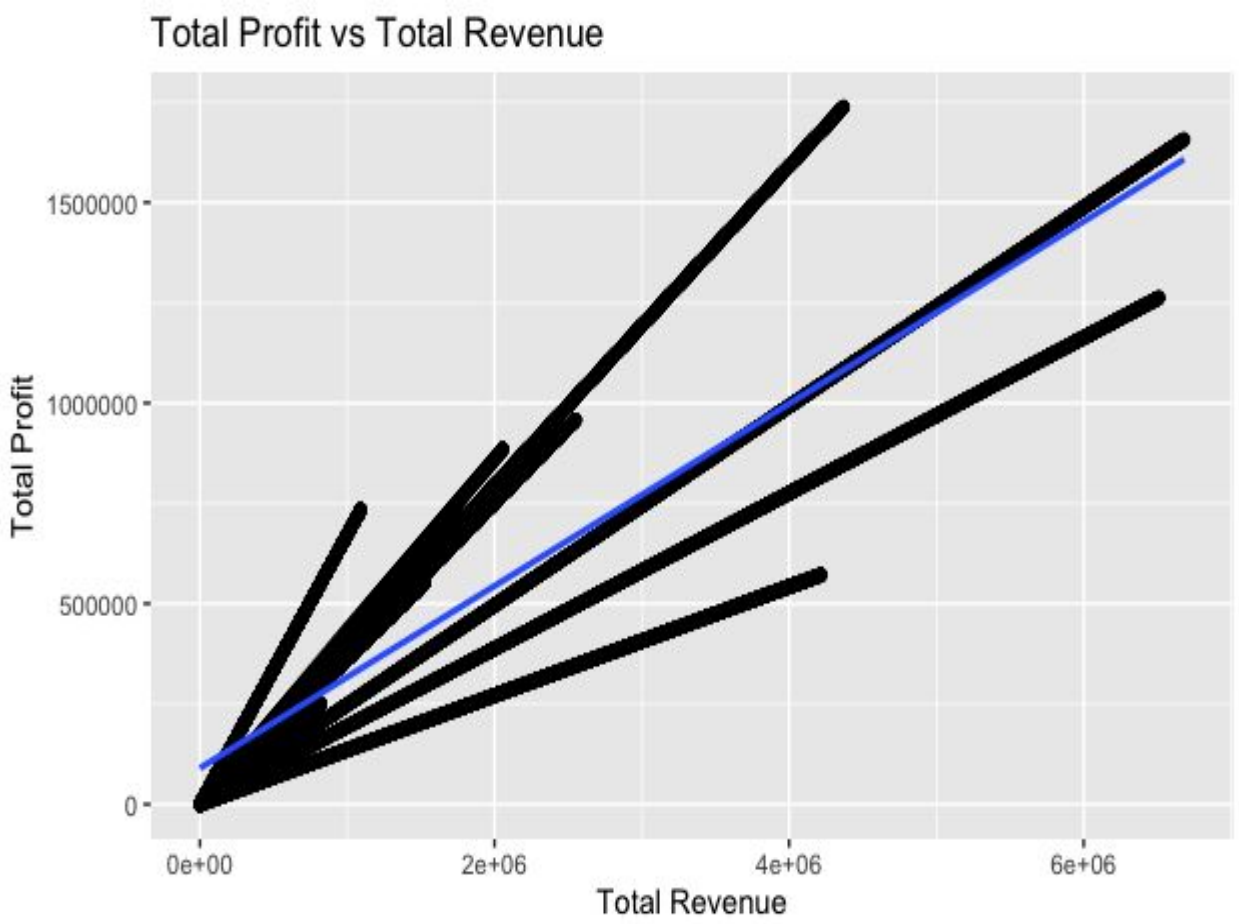
RESEARCH QUESTIONS

QUESTION 1:

1. Correlation between Total Profit earned by a product and Total Revenue (Selling Price) of the product?

I performed regression analysis to ascertain the relationship between a product's total profit and the revenue generated. The correlation between total profit and revenue, which stands at 0.8801112, suggests a strong positive linear relationship. The correlation graph and coefficient affirm that total revenue significantly impacts a product's profitability, implying that variations in revenue directly affect profit margins.

```
> correlation <- cor(your_data$`Total Profit`, your_data$`Total Revenue`)
> print(correlation)
[1] 0.8801112
```

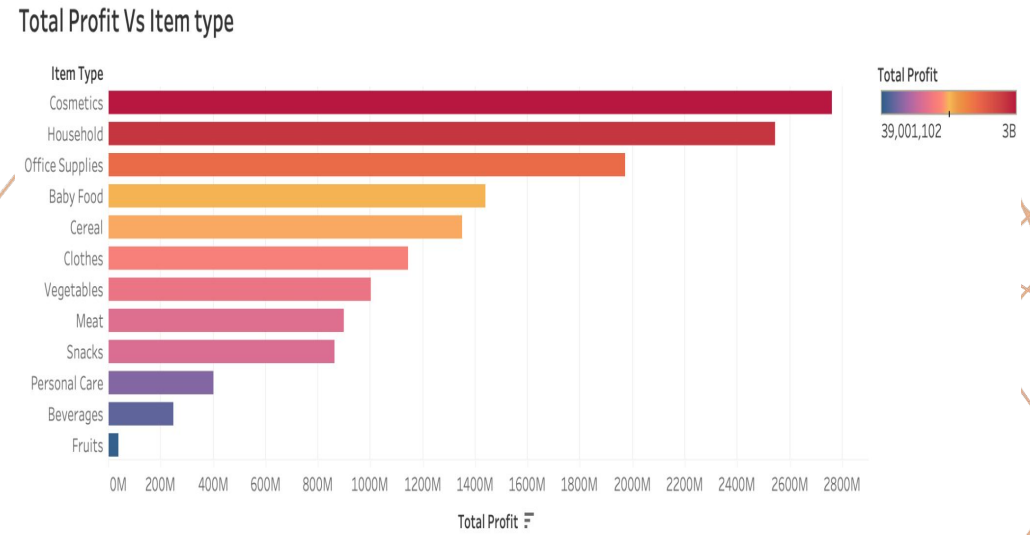


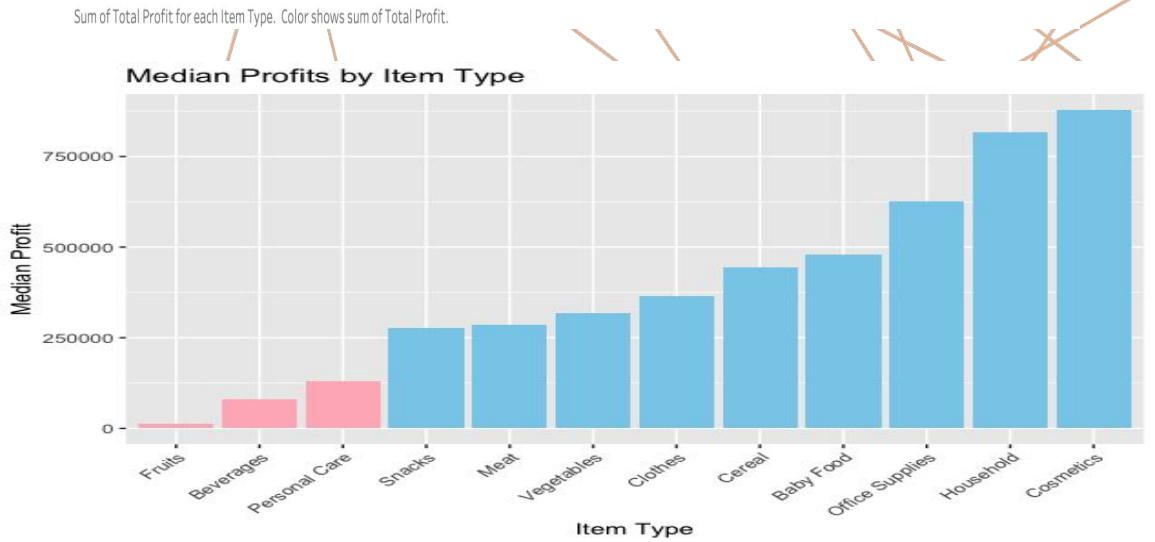
RESEARCH QUESTIONS

Question 2 :

Identification of products which are creating losses?

While there are no products operating at a loss (profits less than zero, as evident in the adjacent graph), certain products show notably lower median profits. In the R-generated graph, these lower-profit products are depicted in a light pink shade. In the Tableau-generated graph, these products are represented by varying shades of purple.



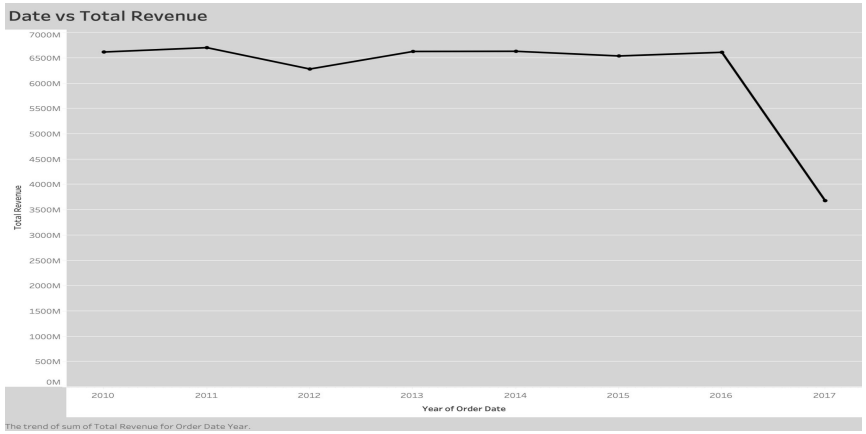


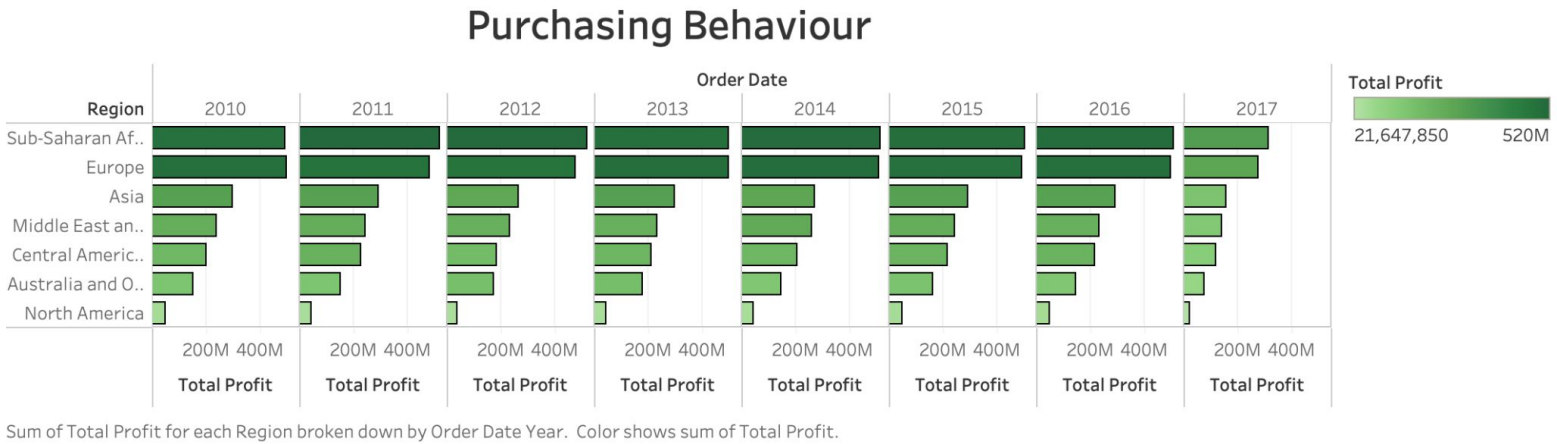
RESEARCH QUESTIONS

Question 3:

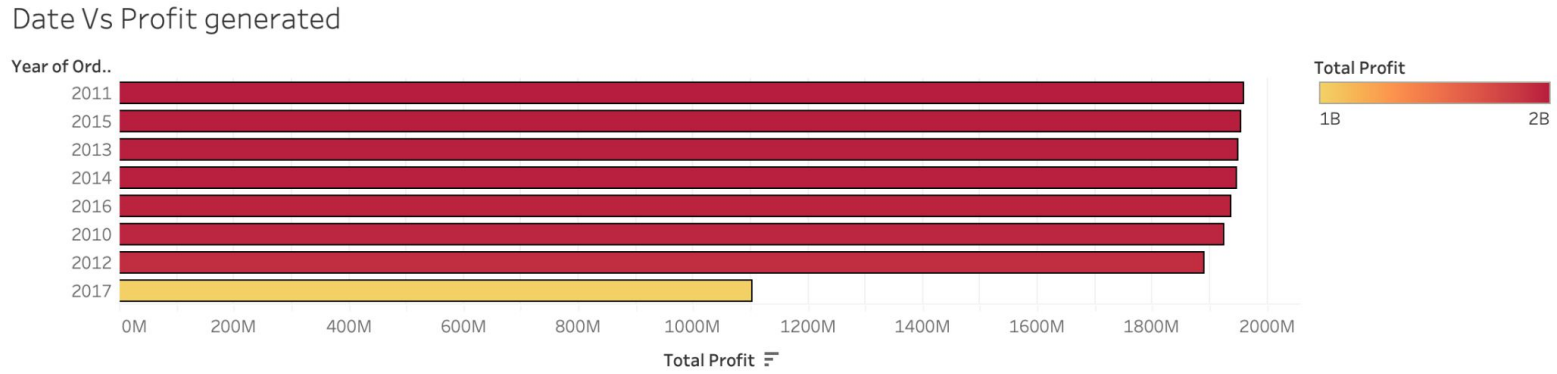
Identification of Seasonality Patterns in sales data and purchasing behaviour

The purchase behavior of consumers shows consistency across regions, with Sub-Saharan Africa and Europe maintaining dominance. However, by 2017, profits from both these regions and well as other regions also decline, consequently impacting the company's overall profits. reducing the profits below 1.5 Billion





Sum of Total Profit for each Region broken down by Order Date Year. Color shows sum of Total Profit.



Sum of Total Profit for each Order Date Year. Color shows sum of Total Profit.

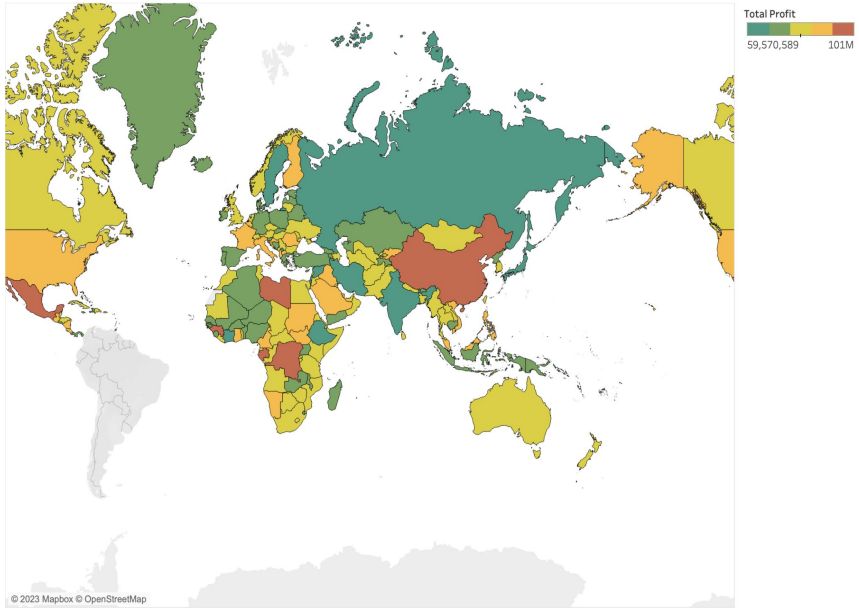
RESEARCH QUESTIONS

Question 4:

What are the top item types, top regions and top countries with higher revenue and highest profit?

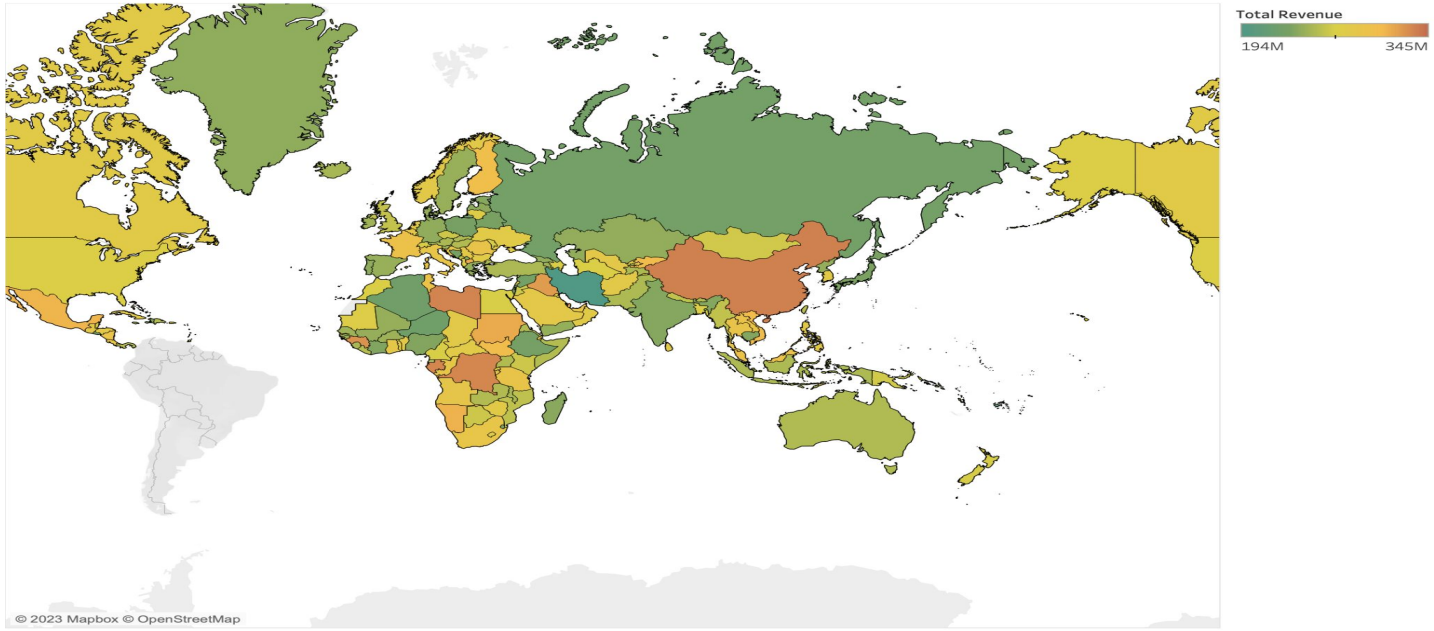
China, Libya, Congo, Gabon, Mexico, and Guinea are the top revenue-generating countries, with Household Items and Office Supplies being the leading contributors. These countries, including China, Libya, Congo, Gabon, Mexico, and Guinea, are also driving the highest profits for the company.

Total Profit Vs Country



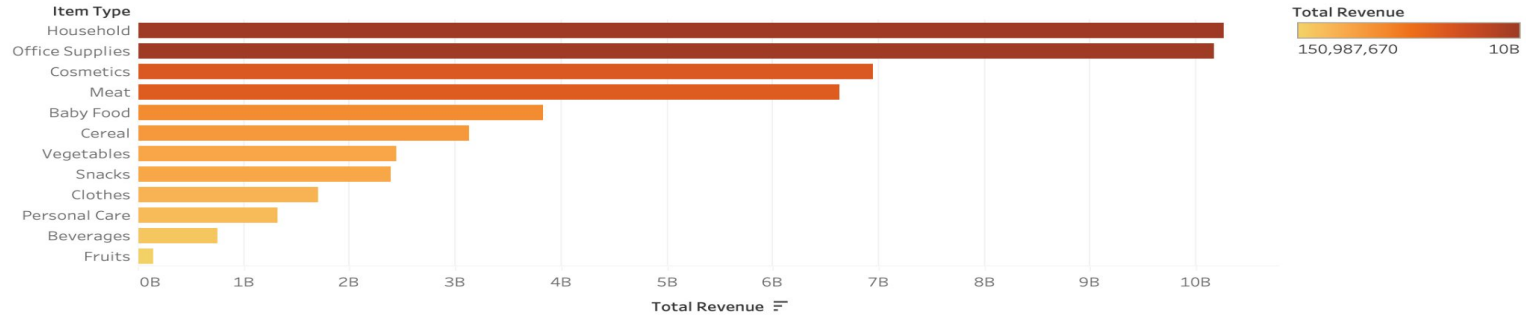
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Total Profit. Details are shown for Country.

Total Revenue Vs Country



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Total Revenue. Details are shown for Country.

Item Type Vs Total Revenue



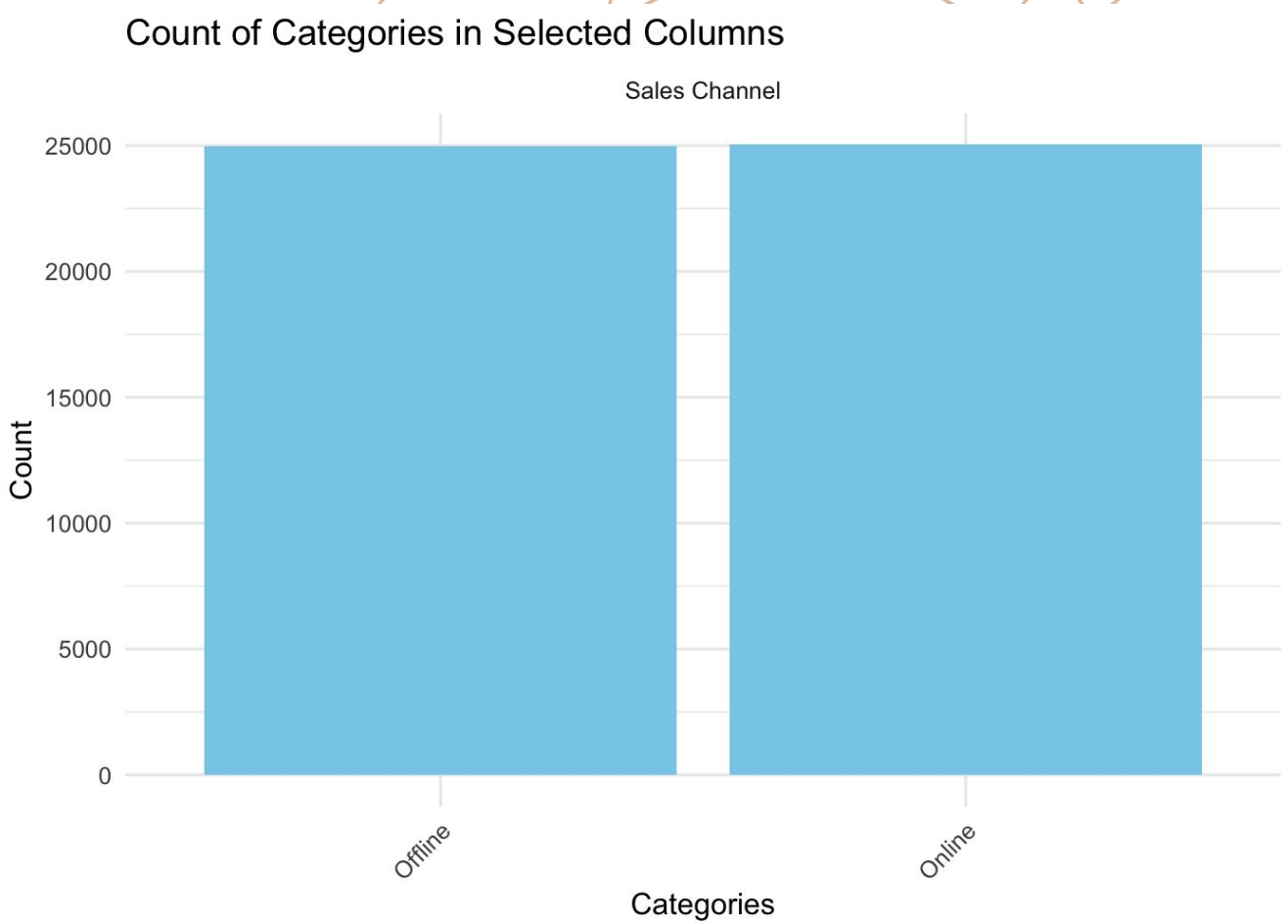
Sum of Total Revenue for each Item Type. Color shows sum of Total Revenue.

RESEARCH QUESTIONS

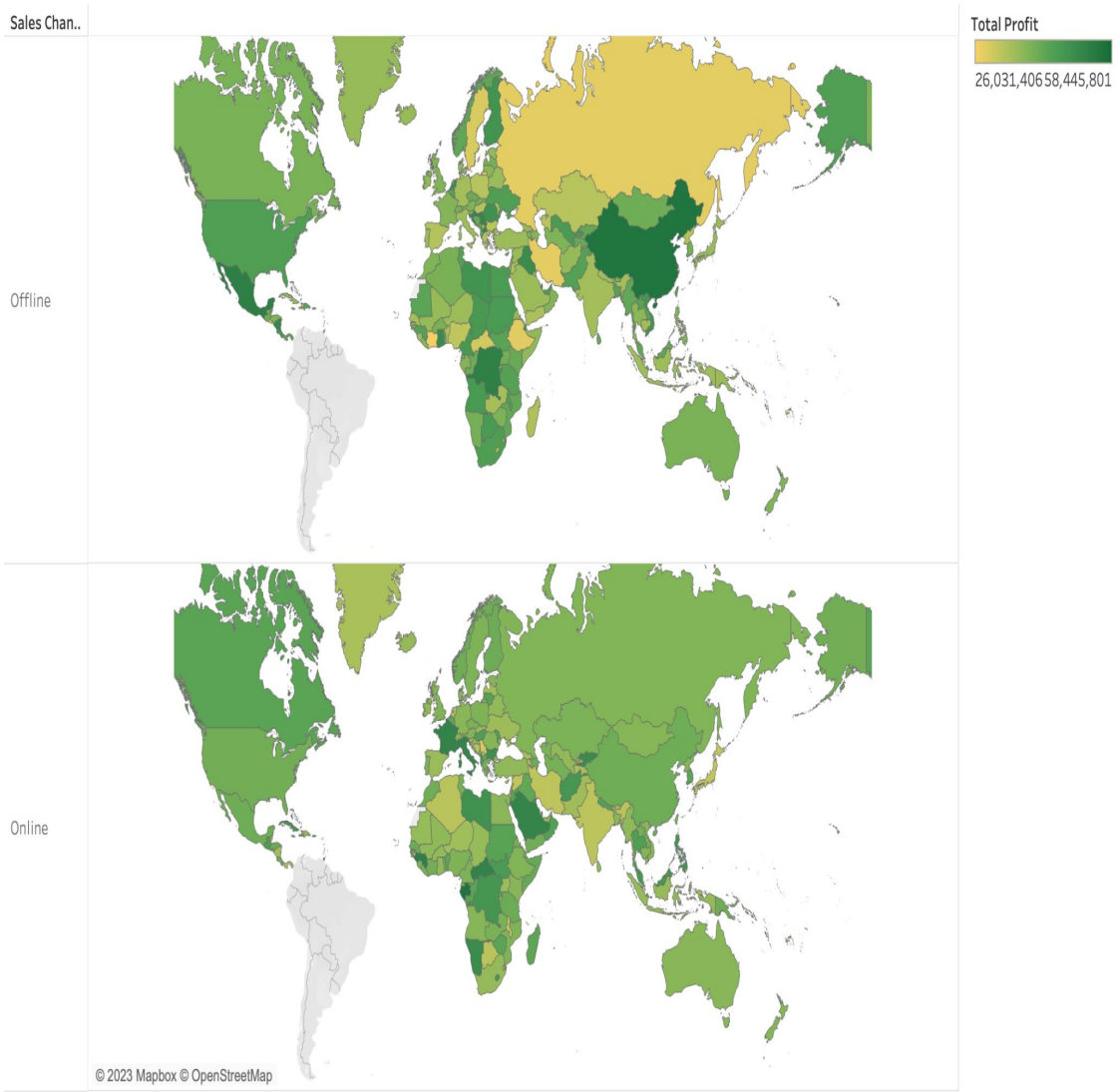
Question 4:

What is the distribution of sales between online and offline channels, and how can I devise distinct strategies to enhance sales for each order type?

Despite variations in revenue across countries, both online and offline sales make equal contributions. To boost sales in both channels, consider implementing targeted marketing campaigns, optimizing the online user experience, expanding physical store presence, and offering exclusive promotions for each platform.

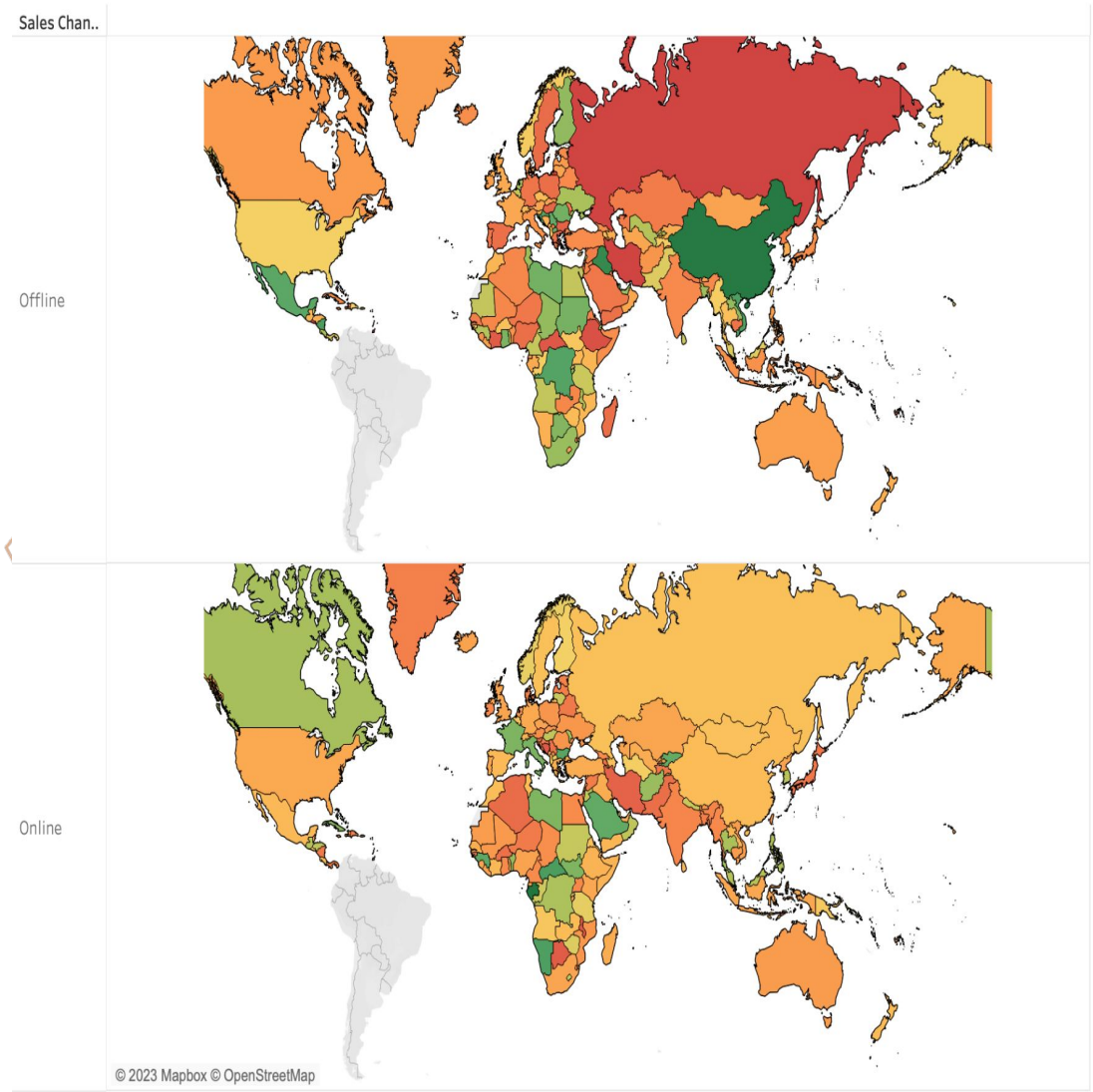


Sales Channel vs Total Profit



Map based on Longitude (generated) and Latitude (generated) broken down by Sales Channel. Color shows sum of Total Profit. Details are shown for Country.

Sales Channel Vs Total Revenue



Map based on Longitude (generated) and Latitude (generated) broken down by Sales Channel. Color shows sum of Total Revenue. Details are shown for Country.

RESEARCH QUESTIONS

Question 5:

Identify the regions with growth potential in terms of revenue and profit.

Considering the responses to the preceding research questions, it can be inferred that Sub-Saharan African regions and Europe have significantly contributed to the company's profits. However, there is untapped potential for growth in other regions, particularly Asia. Given the high internet penetration in Asian and Middle Eastern regions, it is recommended that the company expands its presence in these areas, reducing dependence on Europe and Sub-Saharan Africa.

RESEARCH QUESTIONS

Identifying Important Features for 'Total Revenue' or 'Total Profit' using Random Forest: