STAT 515 Final Project

Name: Shuchi Nirav Shah

Group Members: Shuchi Nirav Shah, Badrinath Sanagavaram, Kamani Madasu

Index

Description of the dataset

- Source of Data: https://www.kaggle.com/datasets/annakhew/sample-country-sales-dataset
- Dataset owner: Annakhew
- Description: It is the dataset based on country sale which includes various aspects of the item type like, Order method, Order priority, Order date. And this dataset gives insights of the profit and revenue after sale of the product.
- Number of Rows: 50000
- Number of Columns: 14 columns

About data

```
# View the first few rows of the dataset
head(data)
```

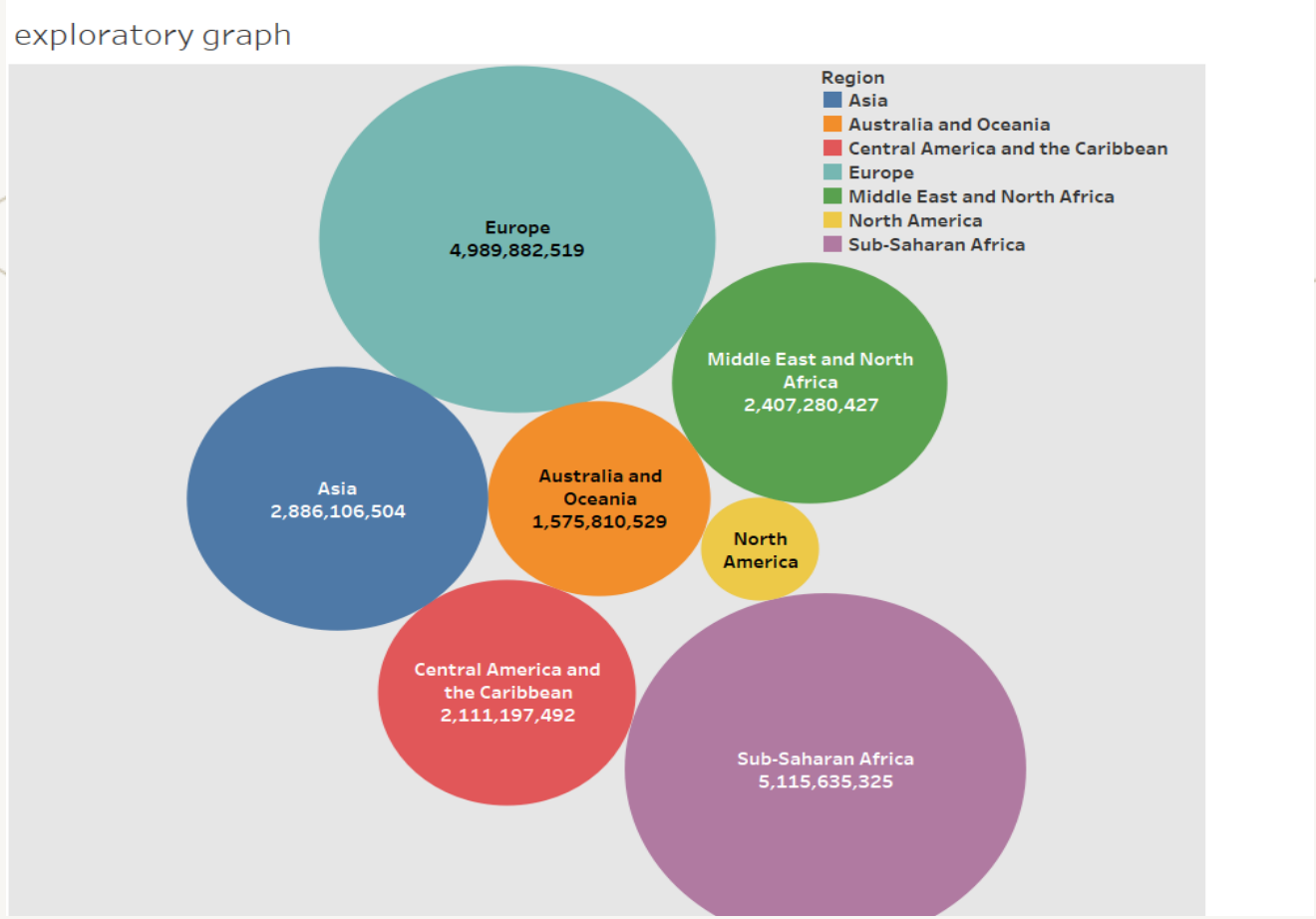
	Region	Country	Item.Type	Sales.Channel	Order.Priority	Order.Date	Order.ID	Ship.Date	Units.Sold
Sub-Saharan	Africa	Namibia	Household	Offline	M	2015-08-31	897751939	2015-10-12	3604
	Europe	Iceland	Baby Food	Online	H	2010-11-20	599480426	2011-01-09	8435
	Europe	Russia	Meat	Online	L	2017-06-22	538911855	2017-06-25	4848
	Europe	Moldova	Meat	Online	L	2012-02-28	459845054	2012-03-20	7225
	Europe	Malta	Cereal	Online	M	2010-08-12	626391351	2010-09-13	1975
	Asia	Indonesia	Meat	Online	H	2010-08-20	472974574	2010-08-27	2542

Unit.Price	Unit.Cost	Total.Revenue	Total.Cost	Total.Profit
668.27	502.54	2408445.1	1811154.2	597290.9
255.28	159.42	2153286.8	1344707.7	808579.1
421.89	364.69	2045322.7	1768017.1	277305.6
421.89	364.69	3048155.2	2634885.2	413270.0
205.70	117.11	406257.5	231292.2	174965.2
421.89	364.69	1072444.4	927042.0	145402.4

```
> #summary of the data
> summary_data <- summary(data)
> print(summary_data)
                                Region                                      Country                  Item.Type
Asia                                : 7348   Trinidad and Tobago                :  321   Fruits        : 4221
Australia and Oceania               : 4017   Guinea                             :  318   Meat          : 4221
Central America and the Caribbean: 5451   Cape Verde                         :  315   Cosmetics     : 4193
Europe                              :12841   Maldives                           :  311   Vegetables    : 4191
Middle East and North Africa        : 6128   Finland                            :  310   Personal Care: 4186
North America                       : 1099   Democratic Republic of the Congo:  308   Beverages     : 4173
Sub-Saharan Africa                  :13116   (Other)                            :48117   (Other)       :24815
 Sales.Channel   Order.Priority      Order.Date       Order.ID            Ship.Date       Units.Sold
 Offline:24966   C:12446        1/21/2017 :   34   Min.   :100013196   7/16/2014 :   35   Min.   :    1
 Online :25034   H:12471        4/14/2013 :   32   1st Qu.:324007046   12/28/2012:   34   1st Qu.: 2498
                 L:12588        12/29/2014:   31   Median :550422394   12/8/2014 :   33   Median : 5018
                 M:12495        2/24/2010 :   31   Mean   :549733027   10/10/2010:   32   Mean   : 5000
                                5/28/2017 :   31   3rd Qu.:776782381   10/6/2011 :   32   3rd Qu.: 7493
                                5/3/2011  :   31   Max.   :999999463   11/17/2013:   32   Max.   :10000
                                (Other)   :49810                       (Other)   :49802
  Unit.Price       Unit.Cost      Total.Revenue      Total.Cost       Total.Profit
 Min.   :  9.33   Min.   :  6.92   Min.   :     28   Min.   :     21   Min.   :      7.2
 1st Qu.: 81.73   1st Qu.: 35.84   1st Qu.: 276487   1st Qu.: 160637   1st Qu.:  94150.9
 Median :154.06   Median : 97.44   Median : 781325   Median : 467104   Median : 279536.4
 Mean   :265.65   Mean   :187.32   Mean   :1323716   Mean   : 933157   Mean   : 390558.7
 3rd Qu.:421.89   3rd Qu.:263.33   3rd Qu.:1808642   3rd Qu.:1190390   3rd Qu.: 564286.7
 Max.   :668.27   Max.   :524.96   Max.   :6682032   Max.   :5249075   Max.   :1738178.4
```

```
# Check for NULL values in each column
null_counts <- colSums(is.na(data))

# Display columns with NULL values and their counts
print(null_counts)
      Region        Country      Item.Type  Sales.Channel Order.Priority     Order.Date       Order.ID      Ship.Date
           0              0              0              0              0              0              0              0
  Units.Sold     Unit.Price      Unit.Cost  Total.Revenue     Total.Cost   Total.Profit
           0              0              0              0              0              0
```

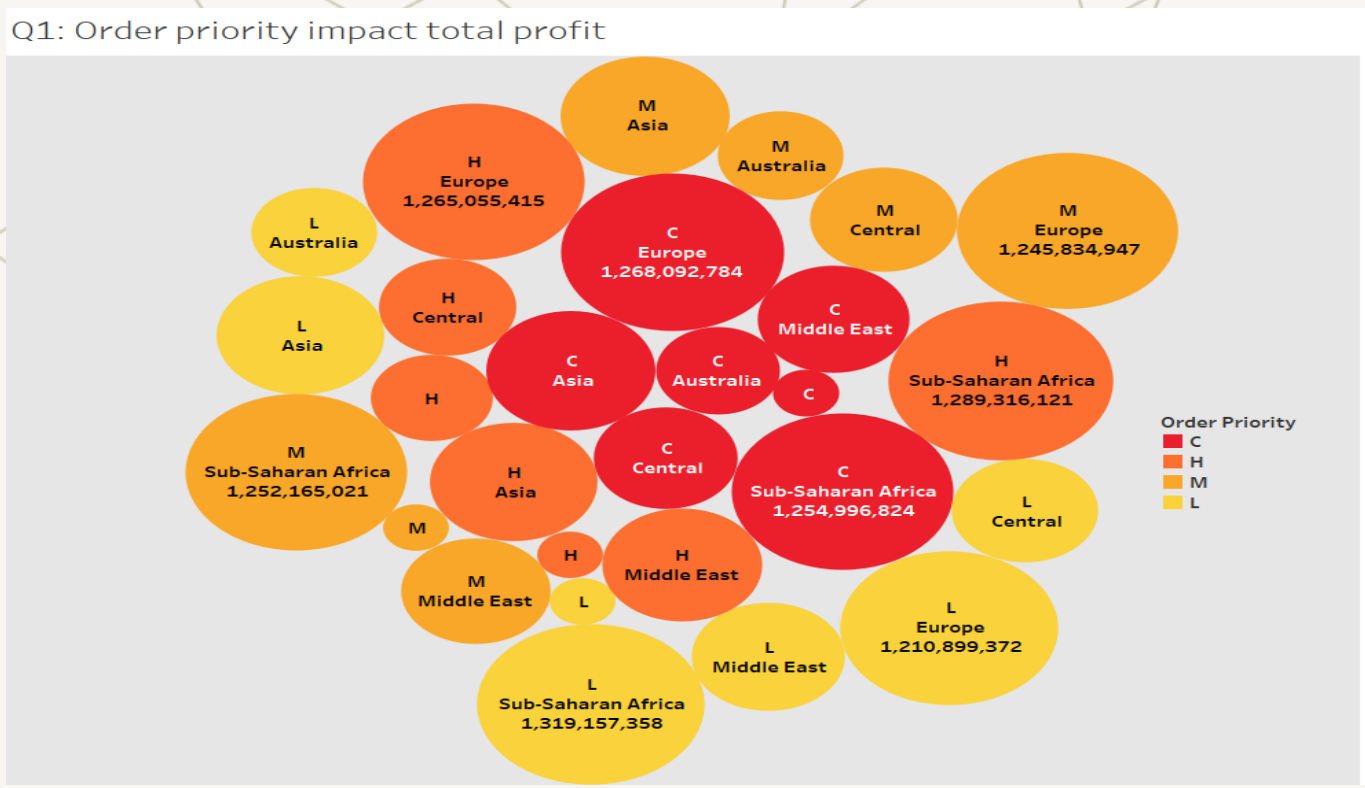
Basic graph



- Here the basic graph is represented which shows the relationship between Region and its Total Profit.

Research Q1

- How does order priority impact total profit, and are there regional variations in this correlation?
- Used linear regression

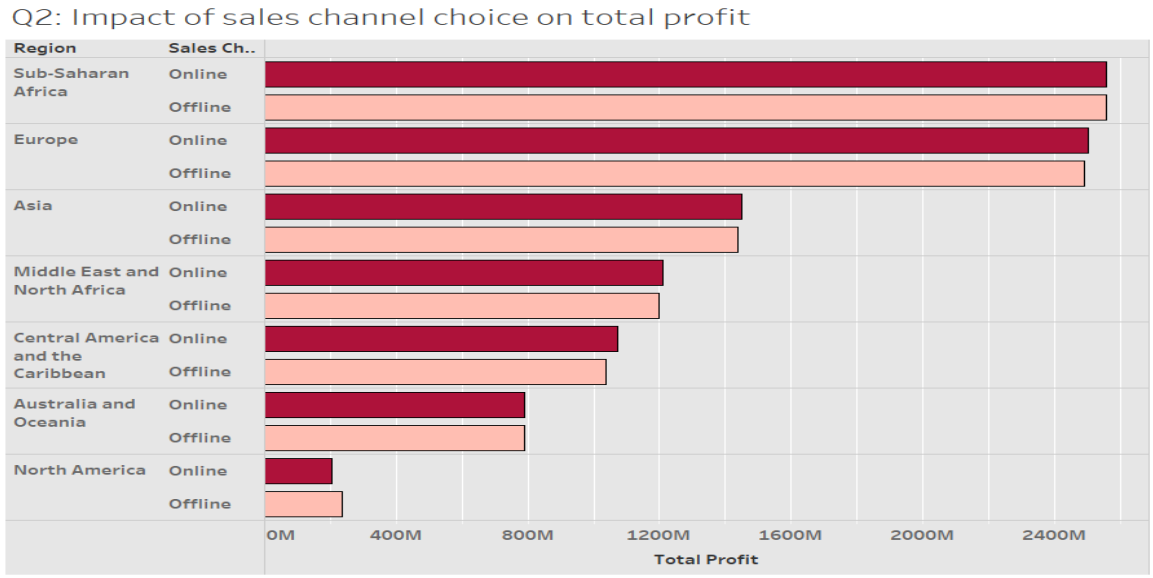


As per the graph there is no direct impact of order priority on Total profit. Because the values are almost similar in each region for order priority.

```
Summary for 1 :
# A tibble: 4 × 5
  term             estimate std.error statistic p.value
  <chr>               <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)      391446.      8802.    44.5     0
2 Order.PriorityH    -354.     12461.    -0.0284  0.977
3 Order.PriorityL    7222.     12530.     0.576   0.564
4 Order.PriorityM   -1365.     12401.    -0.110   0.912

Summary for 2 :
# A tibble: 4 × 5
  term             estimate std.error statistic   p.value
  <chr>               <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)      397490.     12122.     32.8    3.62e-209
2 Order.PriorityH  -24002.     17007.     -1.41   1.58e-  1
3 Order.PriorityL   -7590.     16934.     -0.448  6.54e-  1
4 Order.PriorityM   11525.     17147.      0.672  5.02e-  1

Summary for 3 :
# A tibble: 4 × 5
  term             estimate std.error statistic   p.value
  <chr>               <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)      377856.      9983.     37.9    1.53e-278
2 Order.PriorityH    5456.     14515.      0.376  7.07e-  1
3 Order.PriorityL   16514.     14146.      1.17   2.43e-  1
4 Order.PriorityM   15400.     14075.      1.09   2.74e-  1

Summary for 4 :
# A tibble: 4 × 5
  term             estimate std.error statistic p.value
  <chr>               <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)      393451.      6693.     58.8    0
2 Order.PriorityH   -2519.      9455.     -0.266  0.790
3 Order.PriorityL  -20178.      9450.     -2.14   0.0328
4 Order.PriorityM    3565.      9529.      0.374  0.708
Summary for 5 :
# A tibble: 4 × 5
  term             estimate std.error statistic   p.value
  <chr>               <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)      396400.      9887.     40.1    2.54e-312
2 Order.PriorityH    1436.     13646.      0.105  9.16e-  1
3 Order.PriorityL   -1398.     13913.     -0.100  9.20e-  1
4 Order.PriorityM  -14802.     13959.     -1.06   2.89e-  1

Summary for 6 :
# A tibble: 4 × 5
  term             estimate std.error statistic  p.value
  <chr>               <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      417168.     23135.     18.0    7.53e-64
2 Order.PriorityH    8679.     33001.      0.263  7.93e- 1
3 Order.PriorityL  -31997.     32199.     -0.994  3.21e- 1
4 Order.PriorityM  -33357.     32199.     -1.04   3.00e- 1

Summary for 7 :
# A tibble: 4 × 5
  term             estimate std.error statistic p.value
  <chr>               <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)      386510.      6583.     58.7    0
2 Order.PriorityH   12536.      9322.      1.34   0.179
3 Order.PriorityL    9753.      9253.      1.05   0.292
4 Order.PriorityM   -8098.      9267.     -0.874  0.382
```

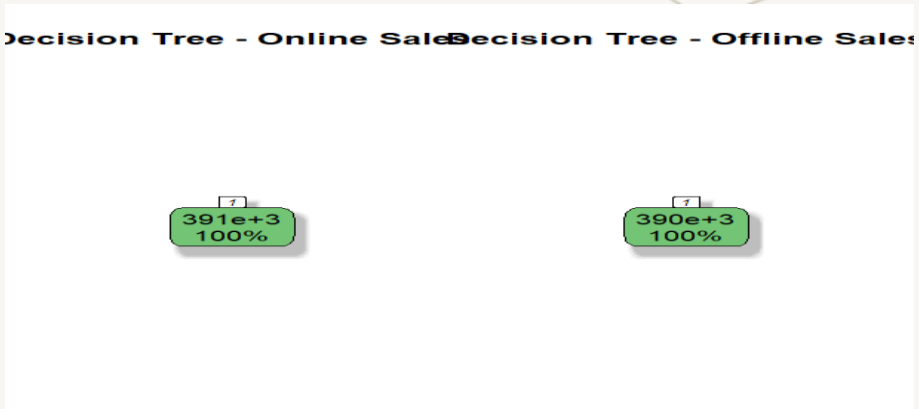
Research Q2:

- What is the impact of sales channel choice on total profit across different regions?
- Used Regression Tree



Groups: Sales.Channel [2]

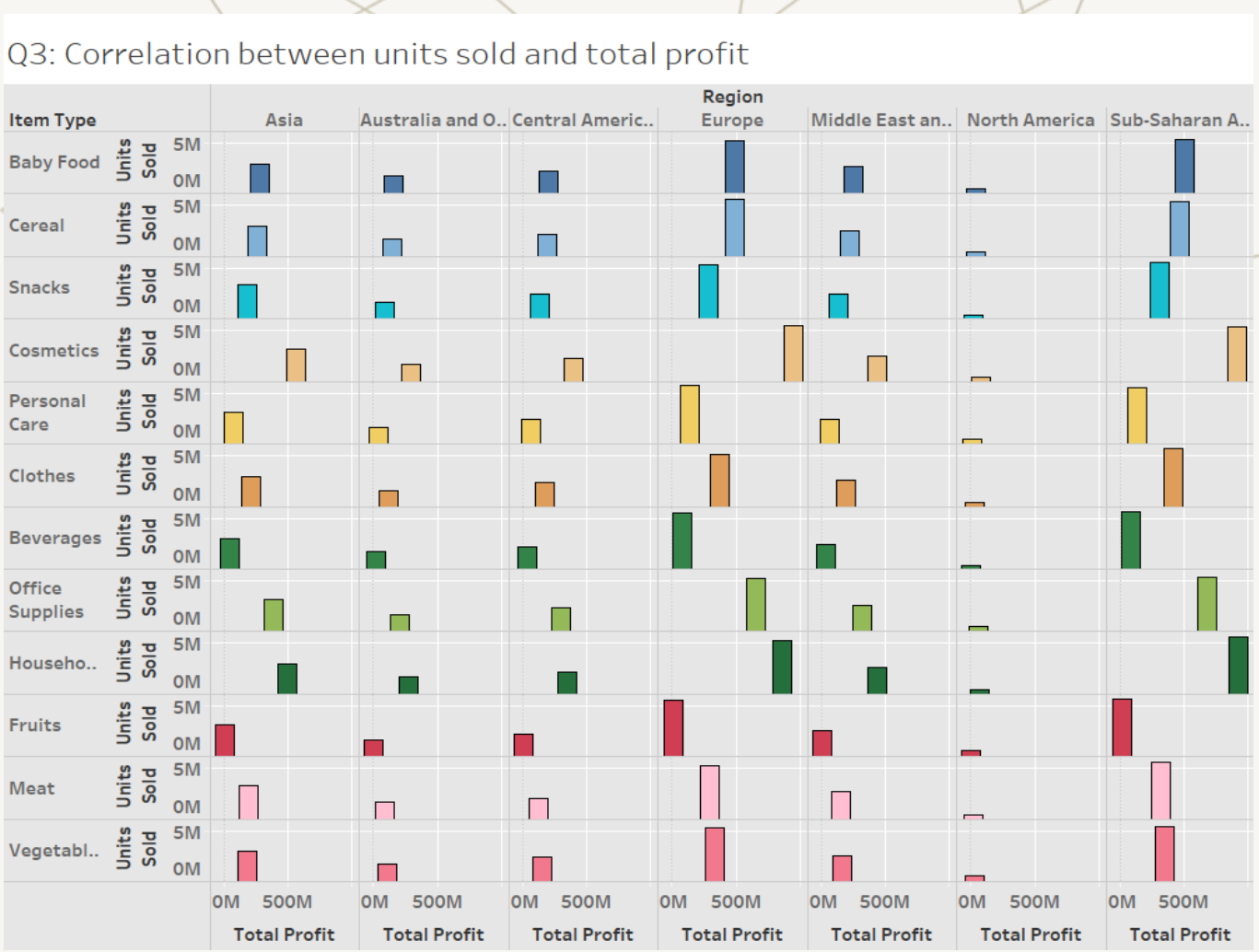
Sales.Channel	Region	Avg_Profit
<fct>	<fct>	<dbl>
Offline	Asia	392559.
Offline	Australia and Oceania	395115.
Offline	Central America and the Caribbean	380528.
Offline	Europe	389374.
Offline	Middle East and North Africa	394297.
Offline	North America	414472.
Offline	Sub-Saharan Africa	388597.
Online	Asia	392988.
Online	Australia and Oceania	389496.
Online	Central America and the Caribbean	394104.
Online	Europe	387813.
Online	Middle East and North Africa	391396.
Online	North America	389037.
Online	Sub-Saharan Africa	391475.



- Here the Decision Tree is presented.
- The highest online and offline sales in sub-Saharan Africa as per the graph.
- The Avg. profit of online and offline are presented.

Research Q3

- Is there a consistent correlation between units sold and total profit, and how does this vary by product type and region?
- Used Linear Regression

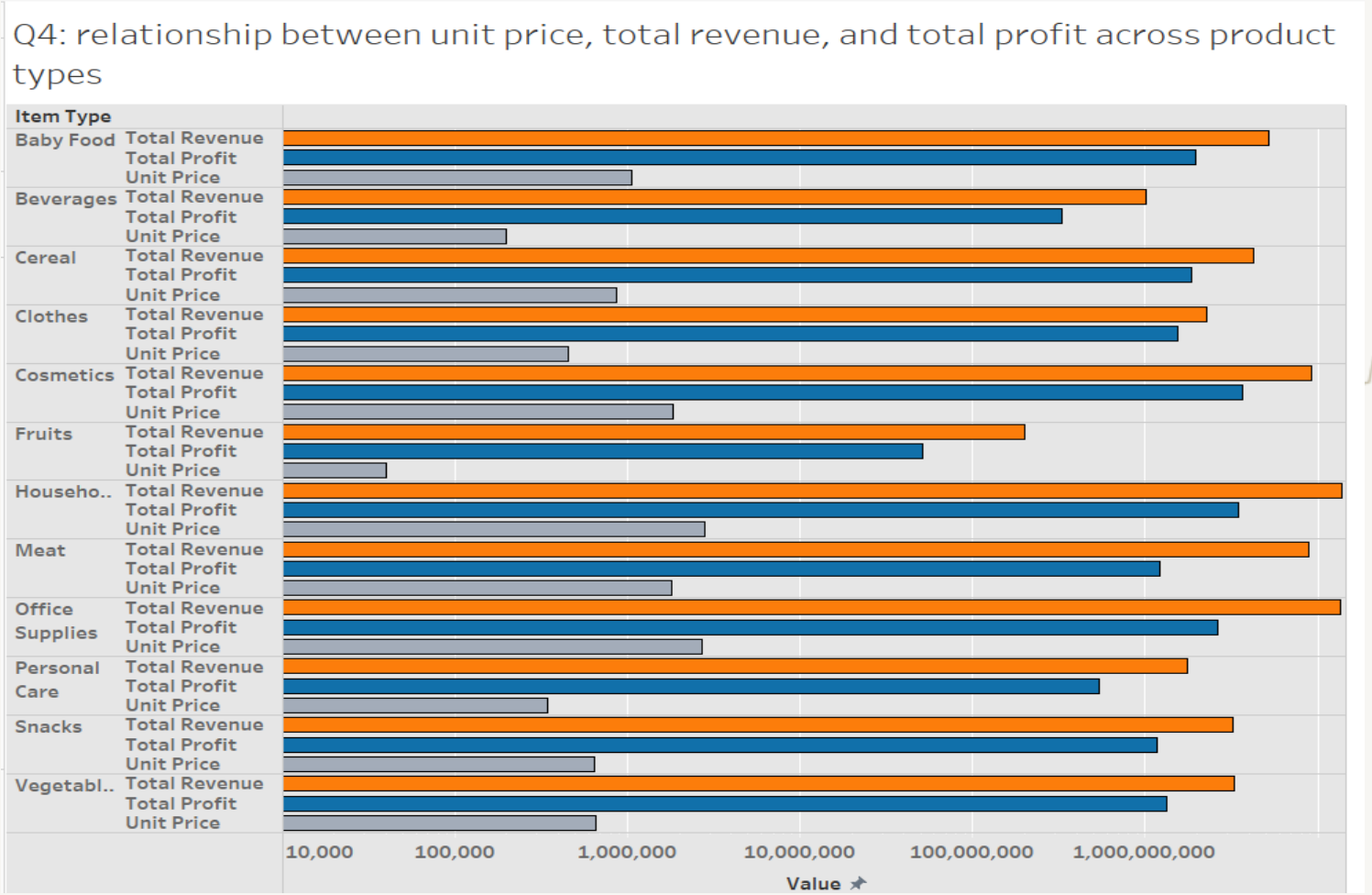



```
> 
> # View the ordered data showing Regions where an item was sold the most and their profits
> print(ordered_data)
```

	Region	Item.Type	Units.Sold	Total.Profit
67	Europe	Personal Care	5818499	145811585
28	Sub-Saharan Africa	Clothes	5679314	417088820
42	Sub-Saharan Africa	Fruits	5649292	13614794
14	Sub-Saharan Africa	Beverages	5606808	87802613
56	Sub-Saharan Africa	Meat	5583837	319395476
49	Sub-Saharan Africa	Household	5581711	925056964
11	Europe	Beverages	5555760	87003202
18	Europe	Cereal	5552719	491915376
77	Sub-Saharan Africa	Snacks	5540079	305479956
39	Europe	Fruits	5534534	13338227
70	Sub-Saharan Africa	Personal Care	5486705	137496827
84	Sub-Saharan Africa	Vegetables	5471542	345418446

Here the graph between Unit sold and total profit base on region and Item type is presented. As the Europe has highest Unit sold on personal care where profit is also higher.

Research Q4



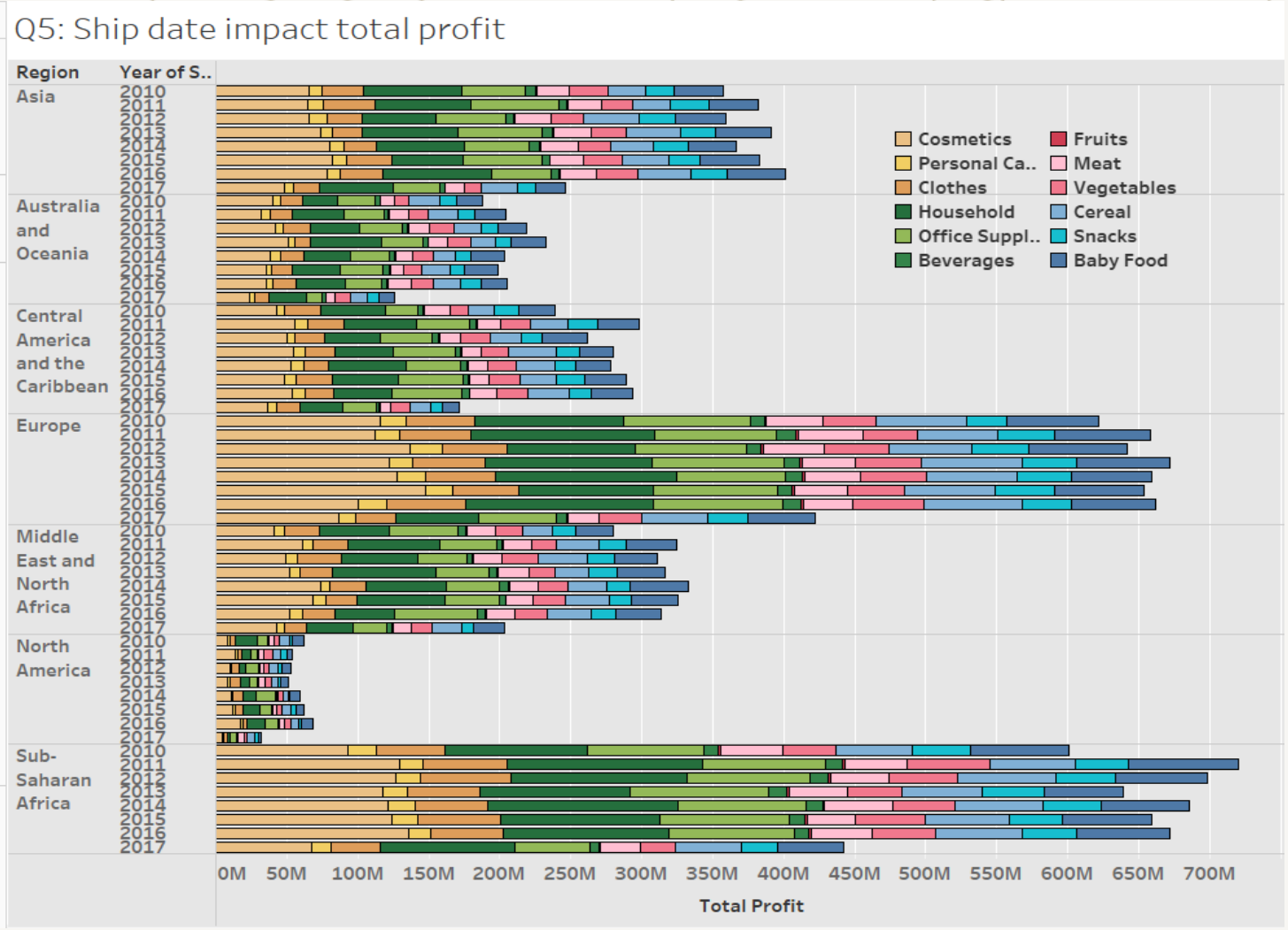
- How does the relationship between unit price, total revenue, and total profit vary across product types?
- Used Linear regression

```
Call:
lm(formula = Total.Profit ~ Item.Type, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-868061 -139443     331  140596  869770

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             476946       4261 111.943  < 2e-16 ***
Item.TypeBeverages     -398405       5991 -66.500  < 2e-16 ***
Item.TypeCereal         -36297       6002  -6.047 1.49e-09 ***
Item.TypeClothes       -110703       5997 -18.458  < 2e-16 ***
Item.TypeCosmetics      391462       5984  65.418  < 2e-16 ***
Item.TypeFruits        -464846       5974 -77.809  < 2e-16 ***
Item.TypeHousehold      345330       6003  57.524  < 2e-16 ***
Item.TypeMeat          -193289       5974 -32.354  < 2e-16 ***
Item.TypeOffice Supplies  152788       6003  25.451  < 2e-16 ***
Item.TypePersonal Care -349000       5986 -58.298  < 2e-16 ***
Item.TypeSnacks        -200635       5995 -33.469  < 2e-16 ***
Item.TypeVegetables    -161180       5985 -26.932  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 272100 on 49988 degrees of freedom
Multiple R-squared:  0.4814,    Adjusted R-squared:  0.4812
F-statistic:  4218 on 11 and 49988 DF,  p-value: < 2.2e-16
```

- Here the relationship between unit price, total revenue , and total profit across product type is presented. Where there are major difference in unit cost and total profit as per the Item type.

Research Q5

How does the ship date impact total profit, and does this influence vary by region and product category?
Used Linear regression




```
> print(max_profit_fast_delivery)
# A tibble: 84 × 3
# Groups:   Region [7]
   Region                          Item.Type  Max_Total_Profit
   <fct>                           <fct>                 <dbl>
 1 Sub-Saharan Africa              Cosmetics          1737483.
 2 Central America and the Caribbean Cosmetics        1730528.
 3 Europe                          Cosmetics          1728963.
 4 Middle East and North Africa    Cosmetics          1720096.
 5 North America                   Cosmetics          1673325.
 6 Asia                            Cosmetics          1670543.
 7 Sub-Saharan Africa              Household          1654814.
 8 Australia and Oceania           Cosmetics          1638725.
 9 Middle East and North Africa    Household          1637412.
10 Europe                          Household          1632938.
# i 74 more rows
```

Here the ship date impact the total profit. And in sub-Saharan Africa total profit is maximum where the Item type is cosmetic.

challenges

- As there are hardly differences in the given values of the columns so it is hard to find differences.
- Not all the data are directly related to Total profit, so it is hard to find it.
- Dataset is large as there are lot of countries, so I had to take Regions

Further Analysis

- Further Future predictions can be done to manage total profit.
- Reduce the total cost per region by selling Item which are available there.

Conclusion

- In conclusion there are 50000 sales data where, through many aspects are directly contribute for higher Total profit. Some regions have higher value of total profit due to ship date, sales channel, unit cost, Item type etc.