Online Retail Sales

Venu Gopalan Krishnagiri Tuppal | 11613143

INFO 5709 | Data Visualization and Communication

Instructor: Dr. Dr. Gahangir Hossain

Term Project

May 10, 2023

Introduction

Online Purchase has become one the most efficient and common way of purchases in recent years. Groceries, Fashion and Clothing, Electronics, Baby Care, Footwear, Daily needs, Kitchen ware, Toys, Medicine, almost everything is available, and people can get anything they desire to their doorstep with few clicks on their smartphones, laptops, tablets, and other devices. Many companies are changing their strategies and business models to meet the requirements of the customers and see that their standards are met by adding more services on online purchases.

Dramatic changes have occurred over the past 20 years because of the quick spread of computer and information technology among business and consumer communities. A significant improvement in the way buyers and sellers communicate is the Internet's application to purchasing behavior. The Pew Internet and American Life Project (2014) reports that as of March 2014, 87% of American adults (18 and older) used the Internet, up from 73% in 2006, with usage nearing saturation among those who live in households earning $75,000 or more annually (99%), young adults (18 to 29), and those with college degrees (97%).

Dataset

The dataset is obtained from Kaggle website. This is a transactional data set that contains the record of the transactions that occurred from December 1, 2010, to December 9, 2011. This dataset belongs to a Non-Store Online Retail which is UK- based and Registered. This Retail company usually deals with customers of various companies that are wholesalers. The products sold by these company is unique to all-occasion gifts. This dataset contains 581,587 entries with 8 attributes which gives us idea on the Customer, Invoice and Country of the transaction. We will not focus on whole data and consider few countries for reference in our work. Following is the Source URL for the dataset.

URL: https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset

Attribute List:

We have 8 attributes in Online Retail Dataset that is used to describe the transaction. Those attributes are:

1. **InvoiceNo**: Invoice number of the Transaction - a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
2. **StockCode**: Product/Item code, a 5-digit integral number uniquely assigned to each distinct product.
3. **Description**: Product /Item Name that describes the product.
4. **Quantity**: The quantities of each product/item per transaction.

5. **InvoiceDate**: Invoice Date and time, the day and time when each transaction was generated.
6. **UnitPrice**: Product/Item Price per 1 Unit. Amount mentioned is in Sterling.
7. **CustomerID**: Customer number - a 5-digit integral number uniquely assigned to each customer.
8. **Country**: Country Name, the name of the country where each customer resides.

Tools / Software Used:

- R (4.3.0) and R Studio
- Tableau

Data Processing:

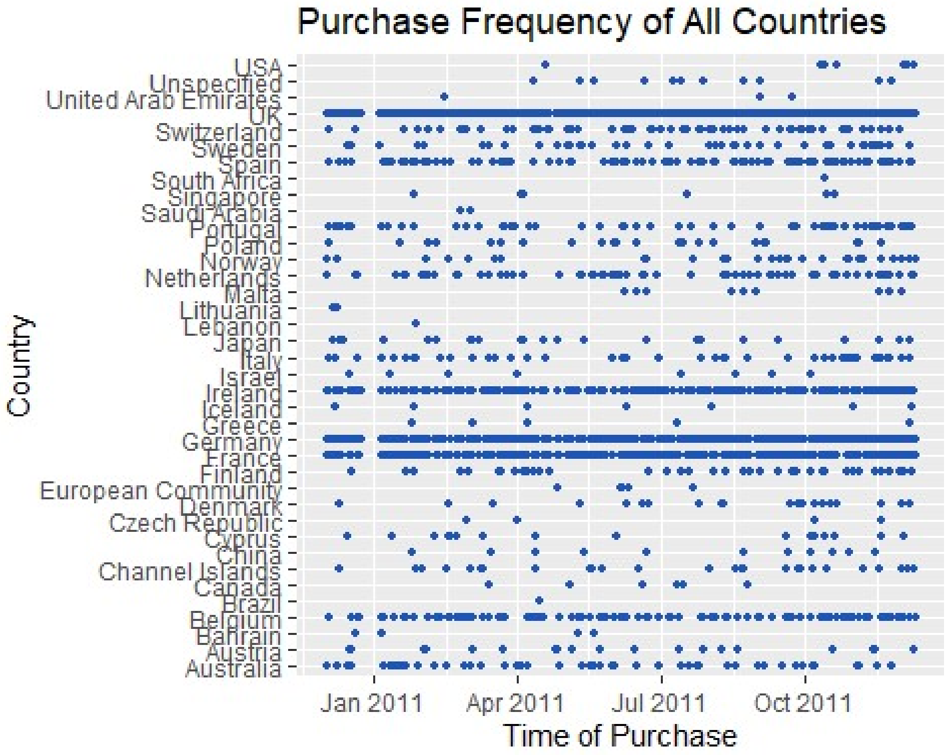
- The data types of the attributes Country, CustomerID, Description, InvoiceID, StockCode are changed from numeric and character/string type to factor type (in R). This makes the mapping and handling the data easy.
- Country Names in the dataset are also changed (as part of cleaning) to as follows:
  - United Kingdom – UK
  - EIRE – Ireland
  - RSA – South Africa
  - Hong Kong – China

Data Analysis:

By taking a few countries as examples, sample graphs are plotted. A reference graph to give the overall idea of the dataset is also plotted which is as follows:

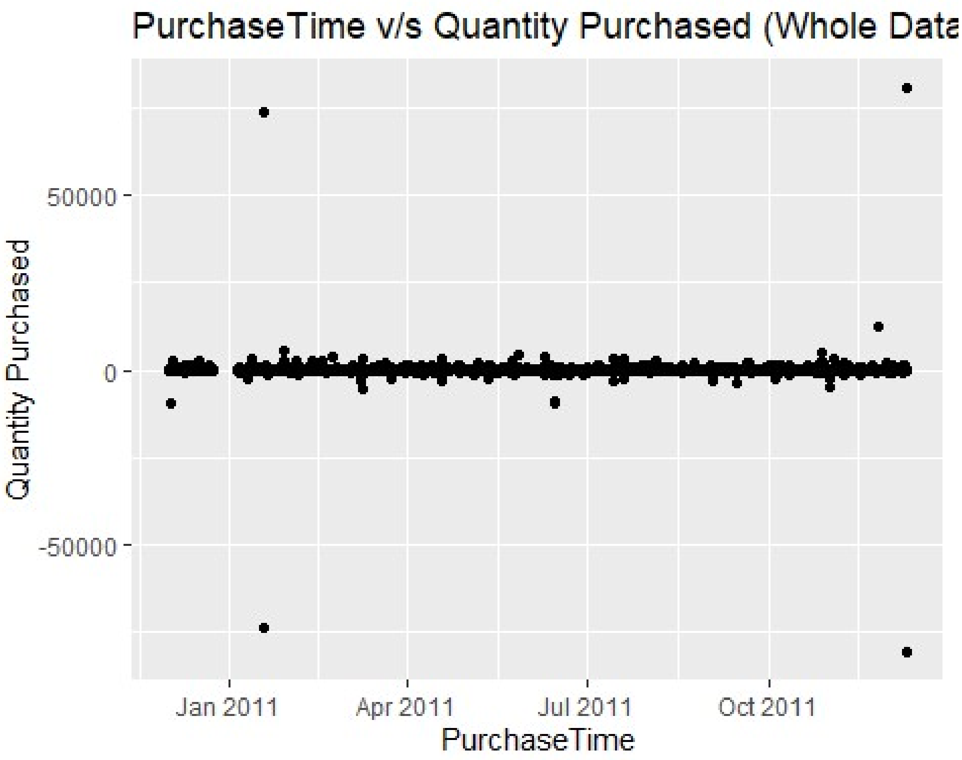
```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)

# Purchase Time v/s Country
RetailData%>%
  ggplot(aes(InvoiceDate, Country))+
  geom_point(color = "#1D55B6", size = 1) +
  labs(x = "Time of Purchase", y = "Country",
       title = "Purchase Frequency of All Countries")
```



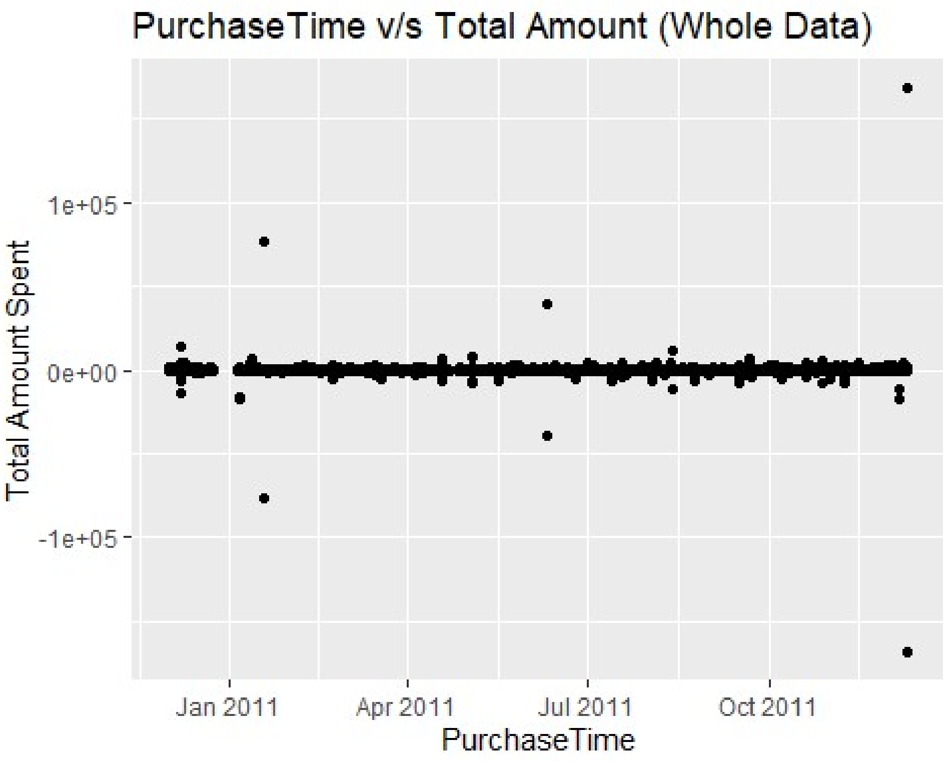
The above plotted graph gives Country wise purchases for the whole time (December 1, 2010, to December 9, 2011)

```
# Purchase Time v/s Quantity for whole Data:
qplot(x = InvoiceDate, y = Quantity,
      data = RetailData,
      xlab = "PurchaseTime",
      ylab = "Quantity Purchased",
      main = "PurchaseTime v/s Quantity Purchased (Whole Data)")

## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



The above graph shows the Total Quantity Purchased time (December 1, 2010, to December 9, 2011)

```
# Purchase Time v/s Price Spent for whole Data:
qplot(x = InvoiceDate, y = Quantity*UnitPrice,
      data = RetailData,
      xlab = "PurchaseTime",
      ylab = "Total Amount Spent",
      main = "PurchaseTime v/s Total Amount (Whole Data)")
```



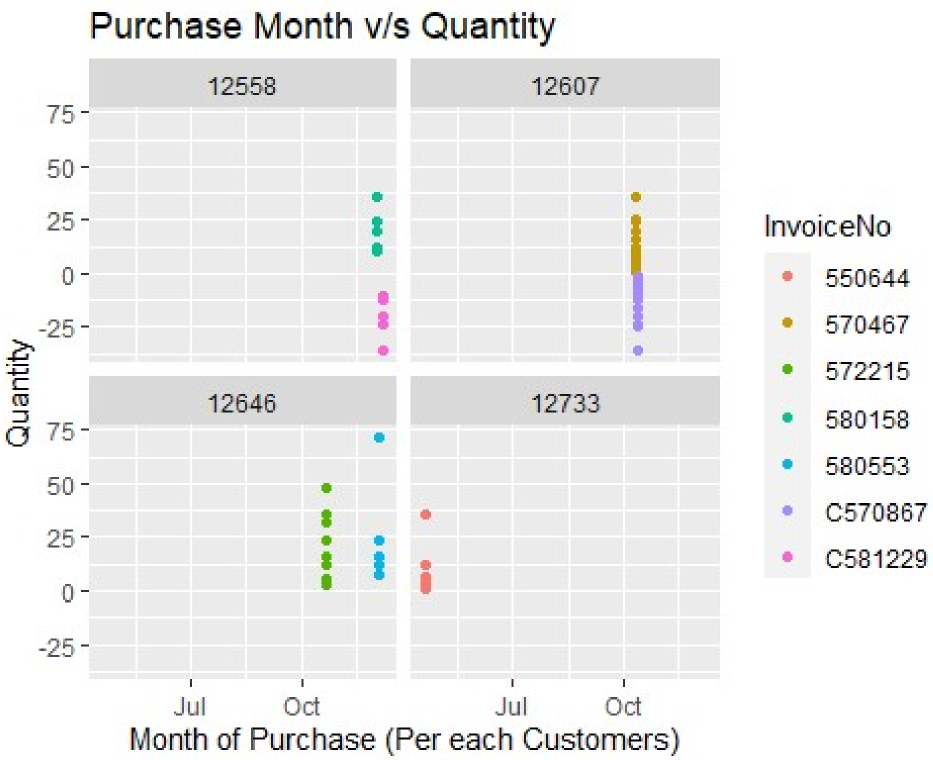
The above graph shows the Total Amount Spent (in Sterling) time (December 1, 2010, to December 9, 2011)

Country-wise Analysis:

By taking 5 Countries (US, Spain, Japan, Italy, Canada) and Field where Country is Unspecified as sample, graphs are plotted to find the pattern in purchase

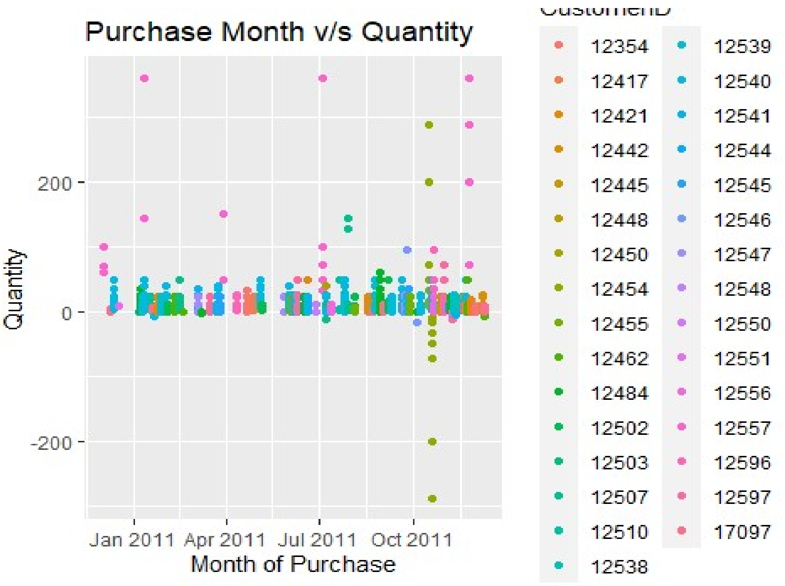
1) Country - **US**

```
# Tests based on Countries
# 1) USA
filter(RetailData, Country == "USA") -> us_data

# qplot(x = InvoiceDate, y = Quantity, data = us_data, color = CustomerID)
# qplot(x = InvoiceDate, y = Quantity, data = us_data, color = InvoiceNo)
qplot(x = InvoiceDate, y = Quantity,
      data = us_data,
      xlab = "Month of Purchase (Per each Customers)",
      ylab = "Quantity",
      main = "Purchase Month v/s Quantity",
      color = InvoiceNo, # Coloring Points based on 'InvoiceNo'
      facets = ~ CustomerID) # Separated Graphs for CustomerID
```



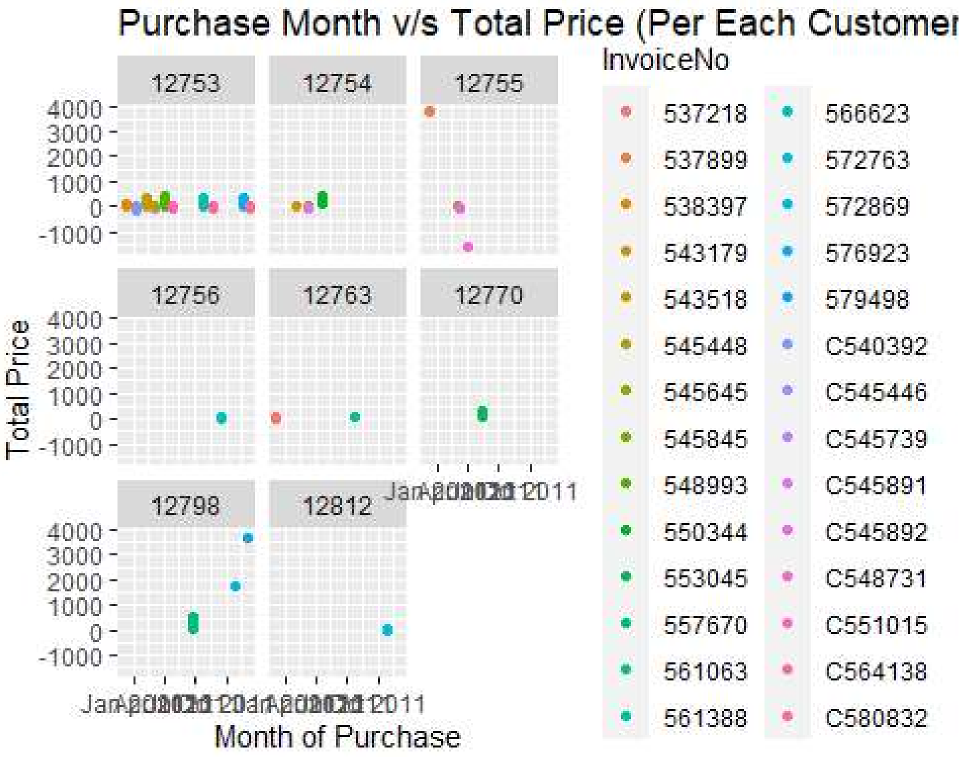
2) Country - **Spain**

```
# 2) Spain
filter(RetailData, Country == "Spain") -> spn_data

qplot(x = InvoiceDate, y = Quantity,
      data = spn_data,
      xlab = "Month of Purchase",
      ylab = "Quantity",
      main = "Purchase Month v/s Quantity",
      color = CustomerID) #color = CustomerID - MANAGABLE - 31 Customers
```



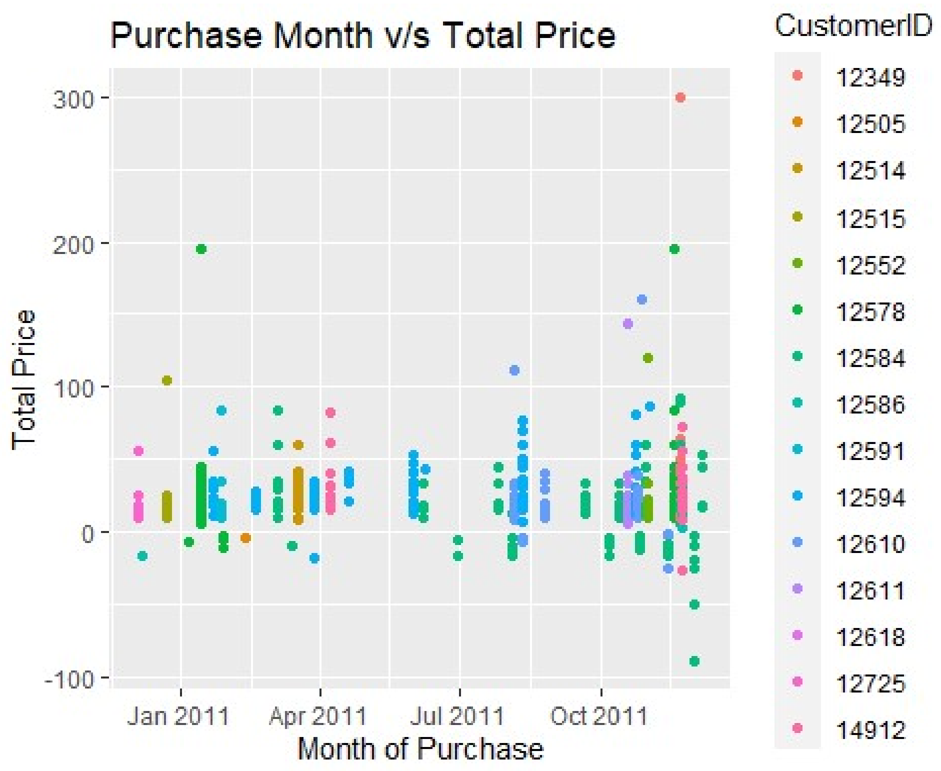
3) Country – **Japan**

```
# 3) Japan
filter(RetailData, Country == "Japan") -> jpn_data
# qplot(x = InvoiceDate, y = Quantity*UnitPrice, data = jpn_data, color = Cus
tomerID)
# 8 Customers
qplot(x = InvoiceDate, y = Quantity*UnitPrice,
      data = jpn_data,
      xlab = "Month of Purchase",
      ylab = "Total Price",
      main = "Purchase Month v/s Total Price (Per Each Customer)",
      color = InvoiceNo,
      facets = ~ CustomerID)
```



4) Country – **Italy**

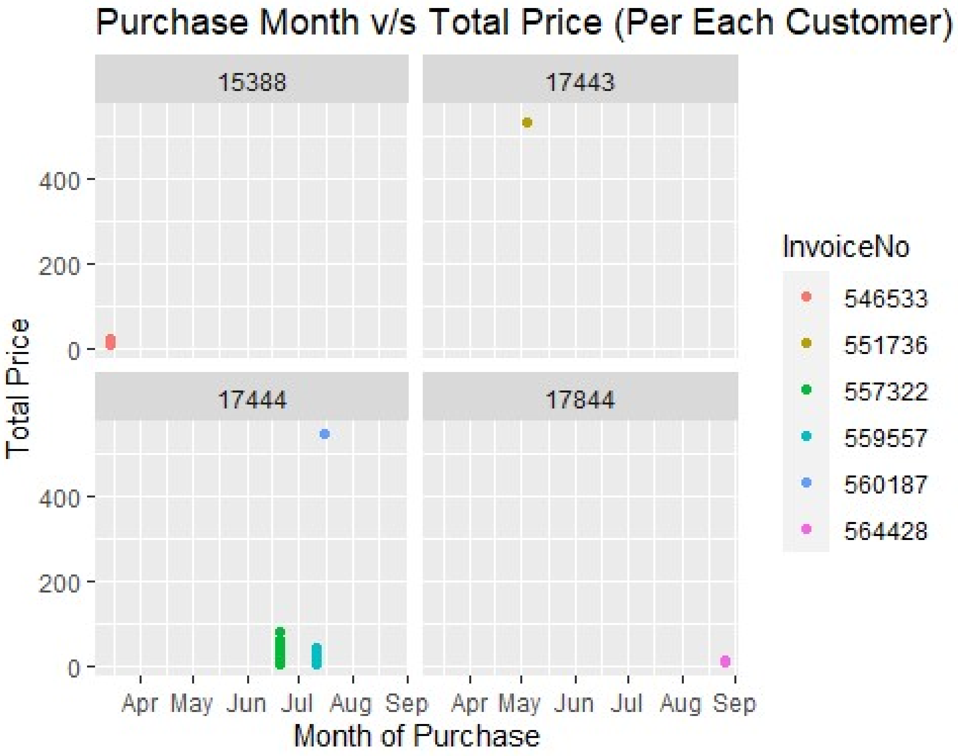
```
# 4) Italy
filter(RetailData, Country == "Italy") -> ity_data
qplot(x = InvoiceDate, y = Quantity*UnitPrice,
      data = ity_data,
      xlab = "Month of Purchase",
      ylab = "Total Price",
      main = "Purchase Month v/s Total Price",
      color = CustomerID)
```



5) Country – **Canada**

```
# 5) Canada
filter(RetailData, Country == "Canada") -> cnd_data
summary(cnd_data)

##    InvoiceNo    StockCode                             Description
##  559557 :77   10133  :  2   COLOURING PENCILS BROWN TUBE       :  3
##  557322 :57   23190  :  2   BUNDLE OF 3 ALPHABET EXERCISE BOOKS:  2
##  546533 :10   23192  :  2   BUNDLE OF 3 SCHOOL EXERCISE BOOKS  :  2
##  564428 : 5   79030D :  2   10 COLOUR SPACEBOY PEN             :  1
##  551736 : 1   10135  :  1   12 PENCILS TALL TUBE POSY          :  1
##  560187 : 1   15044A :  1   4 TRADITIONAL SPINNING TOPS        :  1
##  (Other): 0   (Other):141   (Other)                            :141
##     Quantity       InvoiceDate                      UnitPrice
##  Min.   :  1.0   Min.   :2011-03-14 13:53:00.00   Min.   :  0.10
##  1st Qu.:  6.0   1st Qu.:2011-06-20 09:04:00.00   1st Qu.:  0.83
##  Median : 12.0   Median :2011-07-11 10:33:00.00   Median :  1.65
##  Mean   : 18.3   Mean   :2011-06-26 16:27:03.57   Mean   :  6.03
##  3rd Qu.: 20.0   3rd Qu.:2011-07-11 10:33:00.00   3rd Qu.:  2.95
##  Max.   :504.0   Max.   :2011-08-25 11:27:00.00   Max.   :550.94
##
##    CustomerID       Country
##  17444  :135   Canada   :151
##  15388  : 10   Australia:  0
##  17844  :  5   Austria  :  0
##  17443  :  1   Bahrain  :  0
##  12346  :  0   Belgium  :  0
##  12347  :  0   Brazil   :  0
##  (Other):  0   (Other)  :  0
```

```
# qplot(x = InvoiceDate, y = Quantity*UnitPrice, data = cnd_data, color = Cus
tomerID)
qplot(x = InvoiceDate, y = Quantity*UnitPrice,
      data = cnd_data,
      xlab = "Month of Purchase",
      ylab = "Total Price",
      main = "Purchase Month v/s Total Price (Per Each Customer)",
      color = InvoiceNo,
      facets = ~ CustomerID)
```

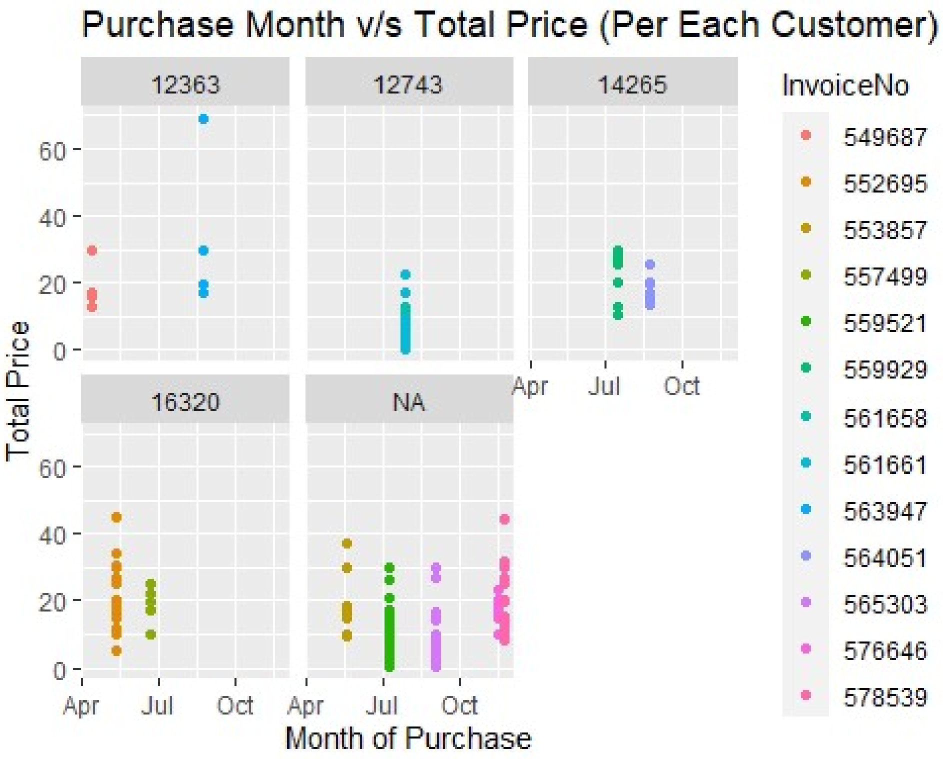


6) Field where Country name is **Unspecified.**

```
# 6) Unspecified
filter(RetailData, Country == "Unspecified") -> unsp_data
summary(unsp_data)
```

```
##    InvoiceNo     StockCode                                Description
##  561658 :83   22150   :  4   3 STRIPEY MICE FELTCRAFT           :  4
##  559521 :72   20983   :  3   12 PENCILS TALL TUBE RED RETROSPOT:  3
##  565303 :66   21124   :  3   4 TRADITIONAL SPINNING TOPS        :  3
##  561661 :51   21591   :  3   ASSORTED COLOUR BIRD ORNAMENT      :  3
##  552695 :47   21888   :  3   BINGO SET                          :  3
##  578539 :34   21889   :  3   CHILDRENS CUTLERY DOLLY GIRL       :  3
##  (Other):93   (Other):427   (Other)                           :427
##     Quantity       InvoiceDate                       UnitPrice
##  Min.   : 1.000   Min.   :2011-04-11 13:29:00.00   Min.   : 0.19
##  1st Qu.: 1.000   1st Qu.:2011-07-08 16:26:00.00   1st Qu.: 0.85
##  Median : 3.000   Median :2011-07-28 16:06:00.00   Median : 1.65
##  Mean   : 7.399   Mean   :2011-07-30 15:13:21.66   Mean   : 2.70
##  3rd Qu.:12.000   3rd Qu.:2011-09-02 12:17:00.00   3rd Qu.: 3.35
##  Max.   :48.000   Max.   :2011-11-24 14:55:00.00   Max.   :16.95
##
##    CustomerID         Country
##  12743  :134   Unspecified:446
##  16320  : 56   Australia  :  0
##  14265  : 31   Austria    :  0
##  12363  : 23   Bahrain    :  0
##  12346  :  0   Belgium    :  0
##  (Other):  0   Brazil     :  0
##  NA's   :202   (Other)    :  0
```

```
# qplot(x = InvoiceDate, y = Quantity*UnitPrice, data = unsp_data, color = Cu
stomerID)
qplot(x = InvoiceDate, y = Quantity*UnitPrice,
      data = unsp_data,
      xlab = "Month of Purchase",
      ylab = "Total Price",
      main = "Purchase Month v/s Total Price (Per Each Customer)",
      color = InvoiceNo,
      facets = ~ CustomerID)
```

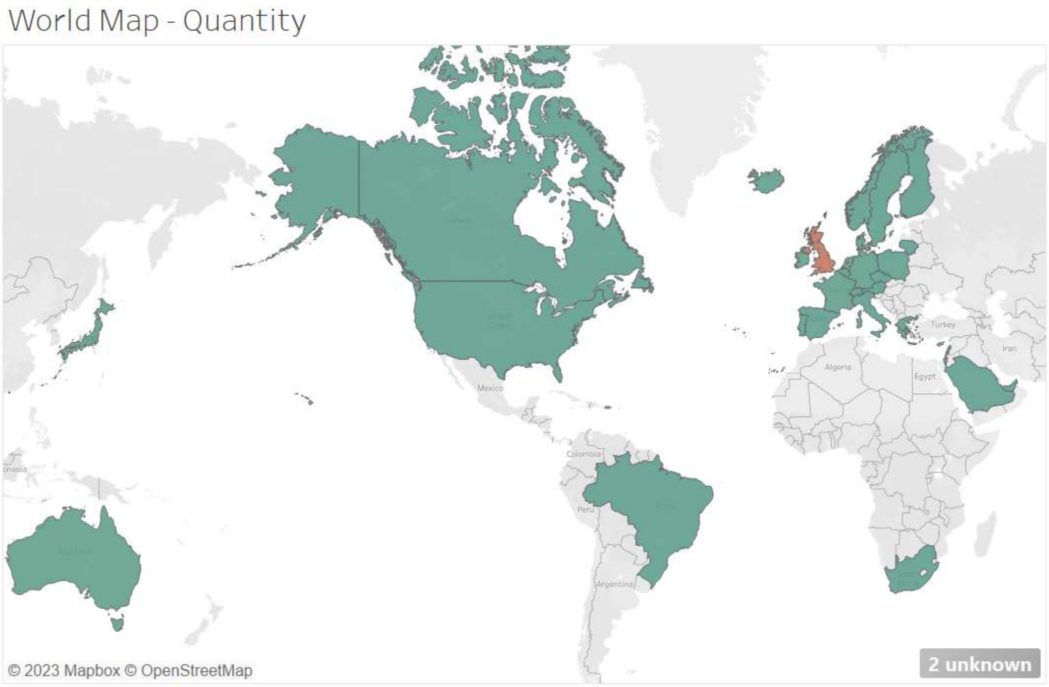


Hypothesis

1. Which region does the Business mostly deal with for their Online Purchases? Find the Country with highest Purchases? Who were the Top Customers?
2. Which Customer does the Least Purchase? Which Country does the Customer belong to? What Products did the Customer Purchase?
3. Does any customer who purchased the products belong to Austria? If so, How many invoices were generated by the customers? What are the prices of the Top 5 Most Purchased product by the Customers?

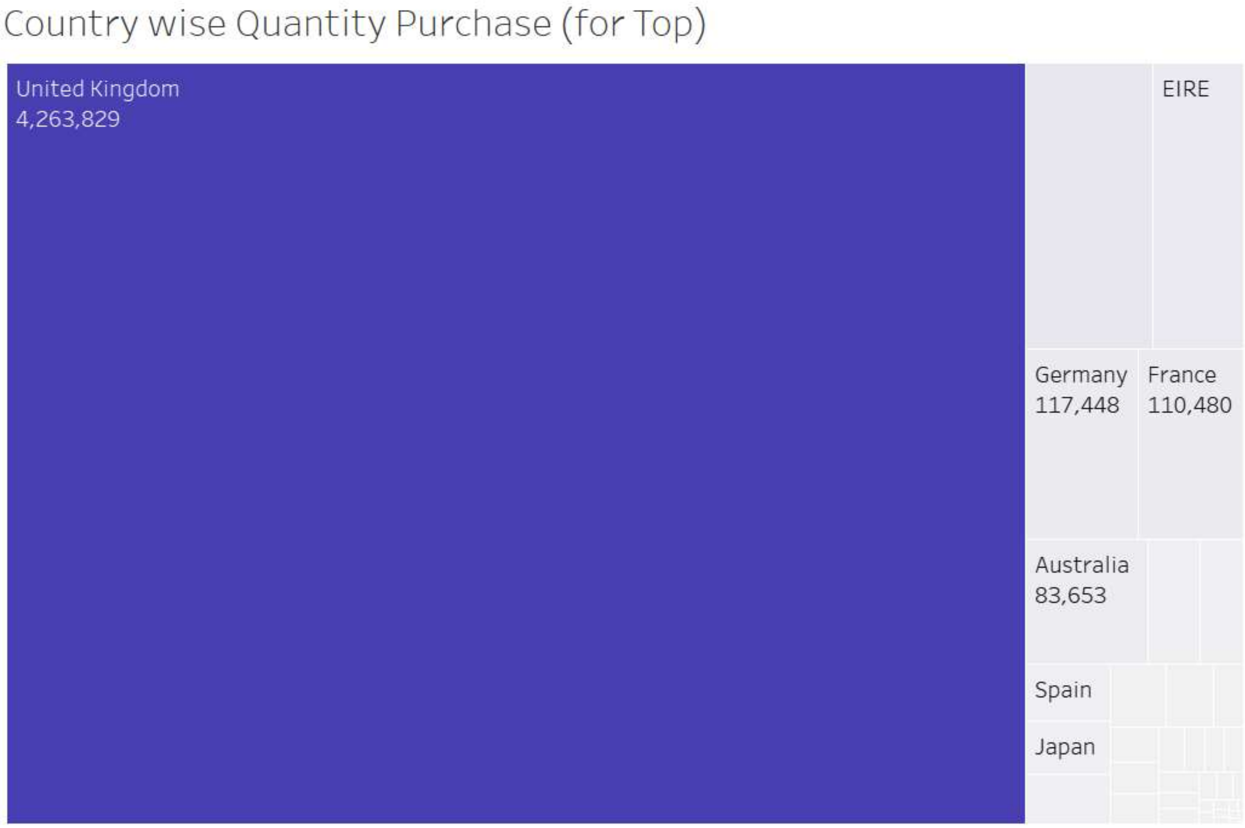
Hypothesis Results

1) The Region that the Online Retail Company mostly deals with is **Australia**, **Europe,** and **North America**. Even though it is not mentioned in the dataset, when plotted with the World Map, we can get an idea of the Countries and Regions where the purchase is done. Also, few countries such as Brazil, Japan and South Africa and Saudi Arabia.

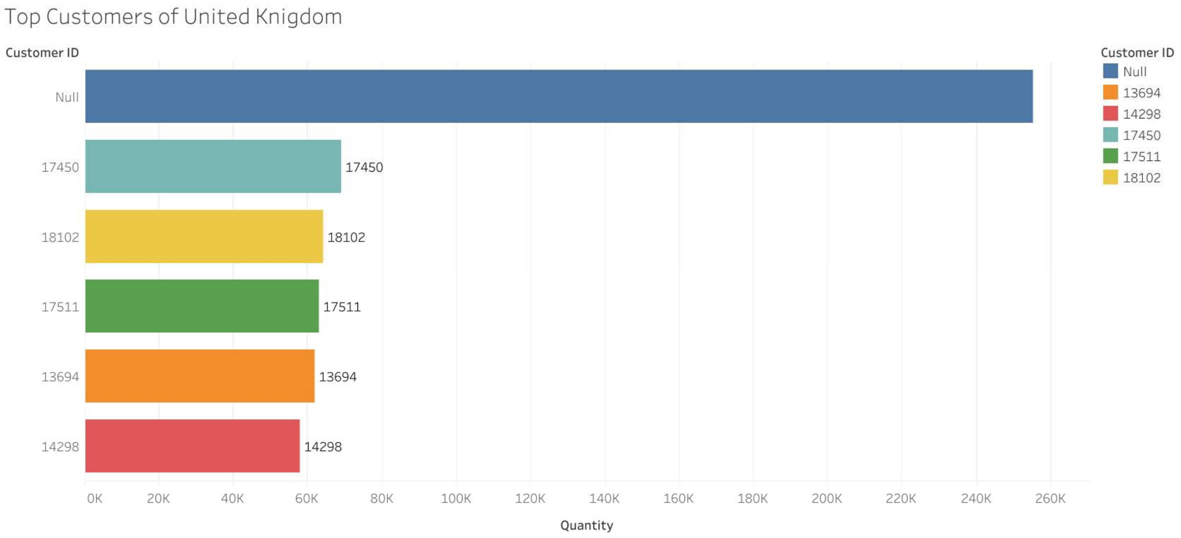


Country with Highest Purchase is **United Kingdom** with the Total Quantity of **4,263,829.**

Using Packed Bubbles (Square Shape) makes the graph and values easy to read and grasp the data quickly.

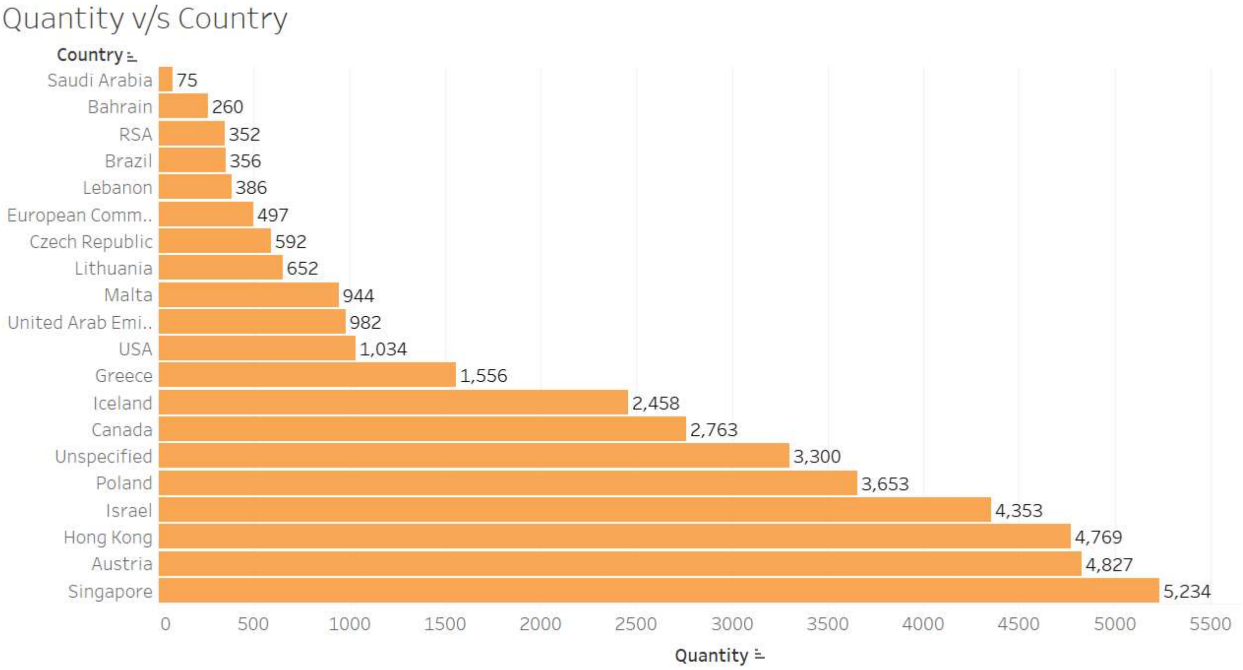


Top customers of United Kingdom are as follows:



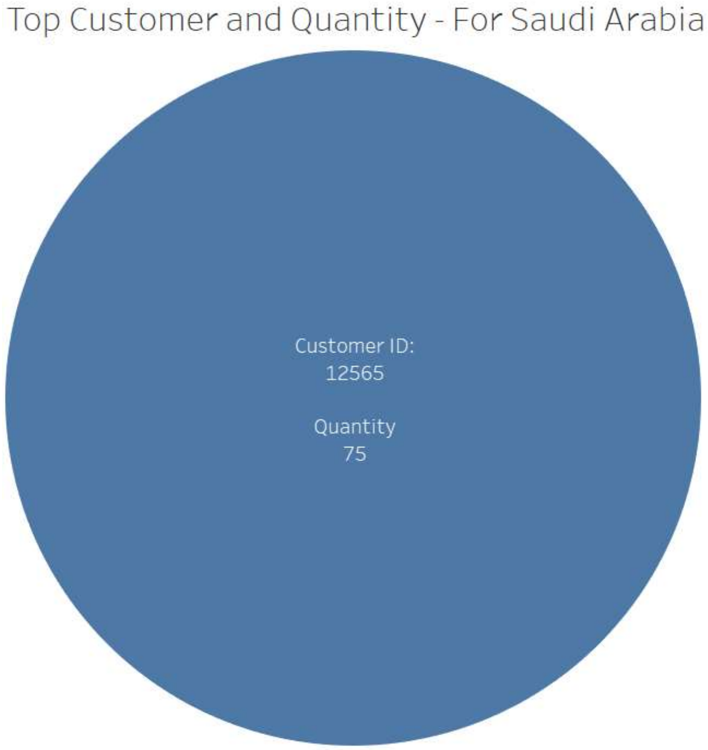
Top Customers of UK are shown above (Customer IDs). Null Value for Customer ID cannot be replaced with Mean or other values. It cannot be ignored also. So Null value is left untouched.

2) Country with Lowest Purchase is **Saudi Arabia** with total Quantity of **75** Total Purchase



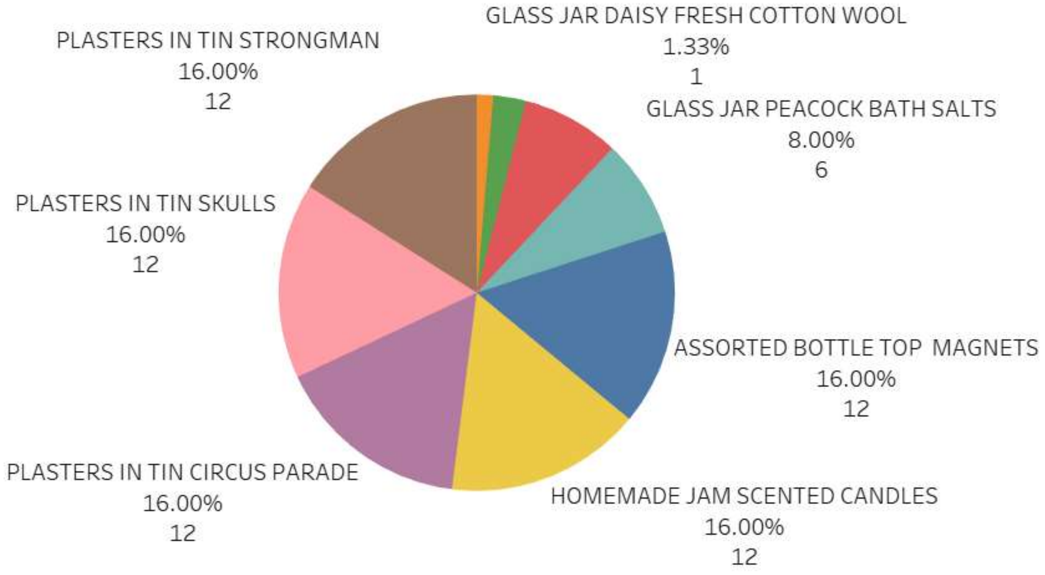
Bar Graph makes it easy to understand the levels and difference clearly.

There is only 1 Customer for Saudi Arabia that Purchased with Online Retail and the Customer ID is 12565. Use of Pie Chart makes it easy to compare on a whole (We have only 1 entry so Pie Chart or any other graph couldn't be used efficiently)



The Products purchased by Customer 12565 are as follows:

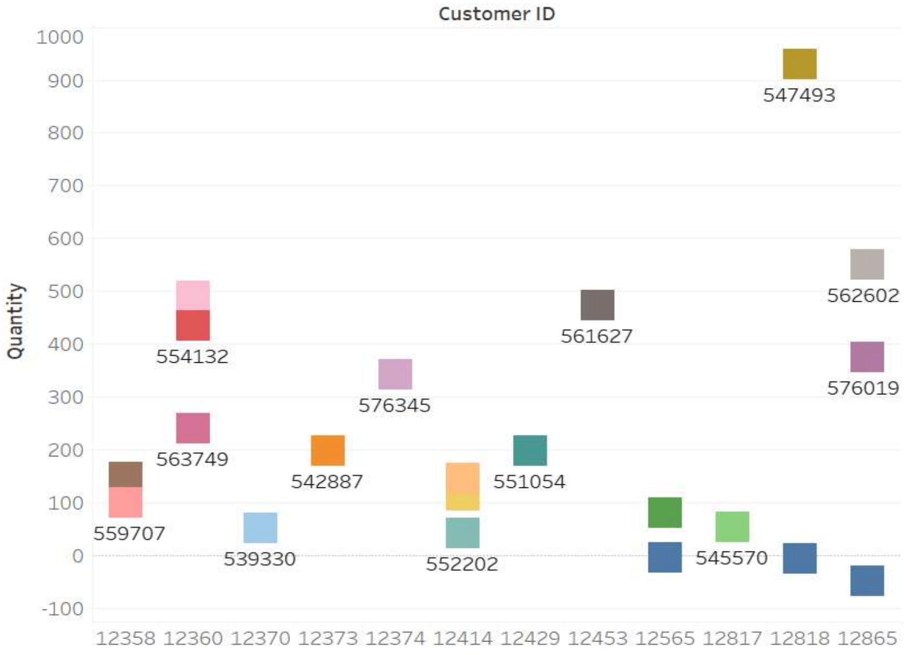
Product Description for Saudi Arabia’s Purchases



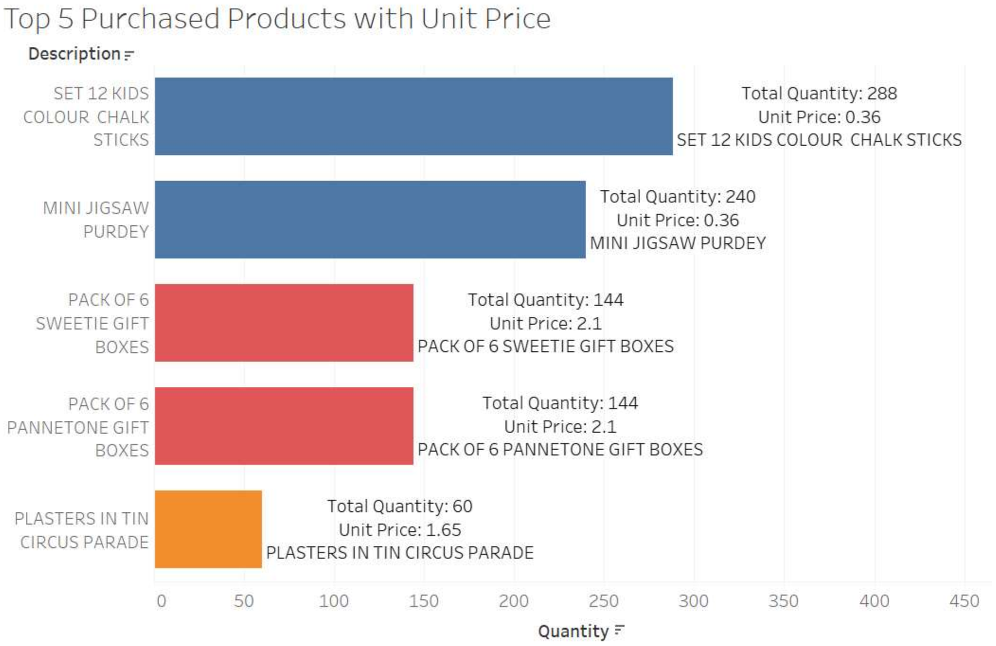
Total of 9 different Products were purchased by Customer 12565 from Saudi Arabia. Above Pie Chart describes the % of each product from the Total Purchase.

3) Yes, Customers purchase from Austria. Following are the List of Customers from Austria and Invoice No. for each Customer’s Purchase

Austria’s Customer ID with Invoice No.



Following are the products with their Unit Price that are mostly purchased by the Customers in Austria:



Conclusion

While the business of Online Retail Vendor mainly focuses on Europe, it has its sales covers in different regions. Yet the products that are being sold are in less quantity in other regions when compared to European Countries. There are cases where Customers without Customer ID are making purchases.

This dataset has instances of negative Quantity. It means that there were cases where Products were returned, or the entry of the data was not correct. This may lead to scenarios where obtained results after processing and visualizing data may be incorrect and accurate and effective business decisions cannot be taken without them.

References

Cai, Y., & Cude, B. J. (2016). Online shopping. *Handbook of consumer finance research*, 339-355.

Online Retail Dataset. (n.d.). Kaggle.
https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset