INFOST 687

Final Project

Justin Sanders

May 4, 2024

Cleaning the Data

```
InvoiceNo       StockCode       Description

Length:541909     Length:541909     Length:541909

Class :character   Class :character   Class :character

Mode  :character   Mode  :character   Mode  :character

Quantity        InvoiceDate         UnitPrice

Min.   :-80995.00   Length:541909     Min.   :-11062.06

1st Qu.:     1.00   Class :character   1st Qu.:     1.25

Median :     3.00   Mode  :character   Median :     2.08

Mean   :     9.55                      Mean   :     4.61

3rd Qu.:    10.00                      3rd Qu.:     4.13

Max.   : 80995.00                      Max.   : 38970.00

CustomerID      Country

Min.   :12346    Length:541909

1st Qu.:13953    Class :character

Median :15152    Mode  :character

Mean   :15288

3rd Qu.:16791

Max.   :18287

NA's   :135080
```

Something is not right with Quantity and Unit Price. Checking for NA values. CustomerID variable doesn’t seem to be necessary and can be removed. all the unwanted values in description will be gone once we values of UnitPrice= 0. As they will not contribute to calculating sales.

After removing free orders (Unit price = 0), returned orders by eliminating negative UnitPrice, and Customer ID variable the summary of the data looks like:

```
InvoiceNo        StockCode        Description

Length:539394     Length:539394     Length:539394

Class :character   Class :character   Class :character

Mode  :character   Mode  :character   Mode  :character

Quantity        InvoiceDate        UnitPrice

Min.   :-80995.00   Length:539394     Min.   :-11062.06

1st Qu.:     1.00   Class :character   1st Qu.:     1.25

Median :     3.00   Mode  :character   Median :     2.08

Mean   :     9.85                     Mean   :     4.63

3rd Qu.:    10.00                     3rd Qu.:     4.13

Max.   : 80995.00                     Max.   : 38970.00

Country

Length:539394

Class :character

Mode  :character
```

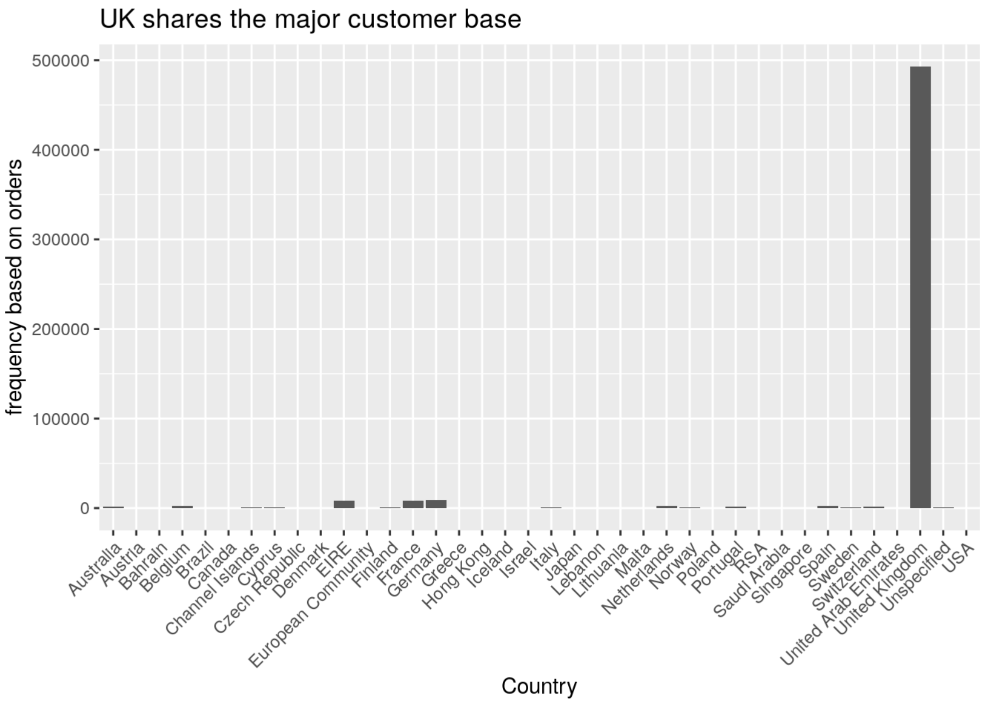
Data pre-processing 1 for single variable plot

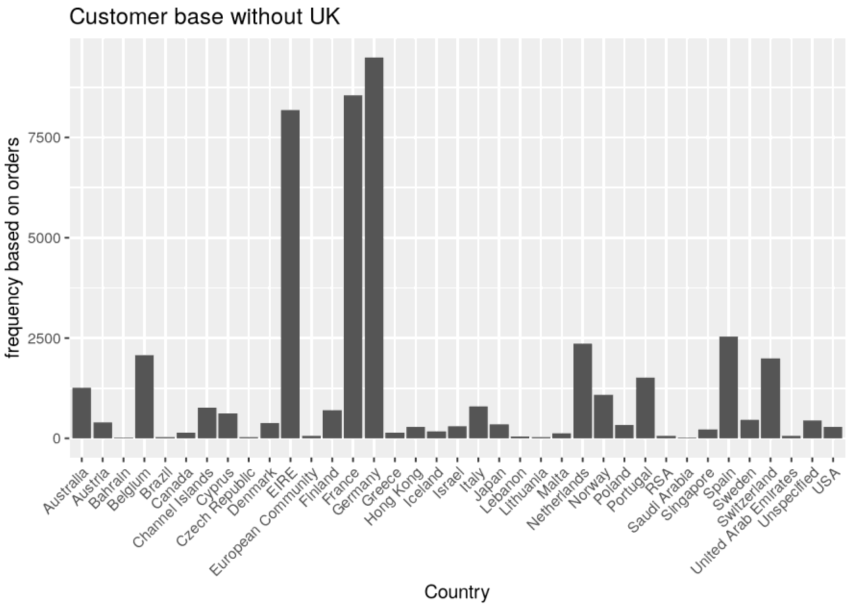
Hourly, daily and monthly split of date time for Invoice date will be used in single and multivariable plot.

```
InvoiceNo StockCode Description Quantity InvoiceDate        UnitPrice

<chr>    <chr>    <fct>        <dbl> <dttm>              <dbl>

1 536365    85123A    WHITE HANG…       6 2010-12-01 08:26:00      2.55

2 536365    71053    WHITE META…       6 2010-12-01 08:26:00      3.39
```

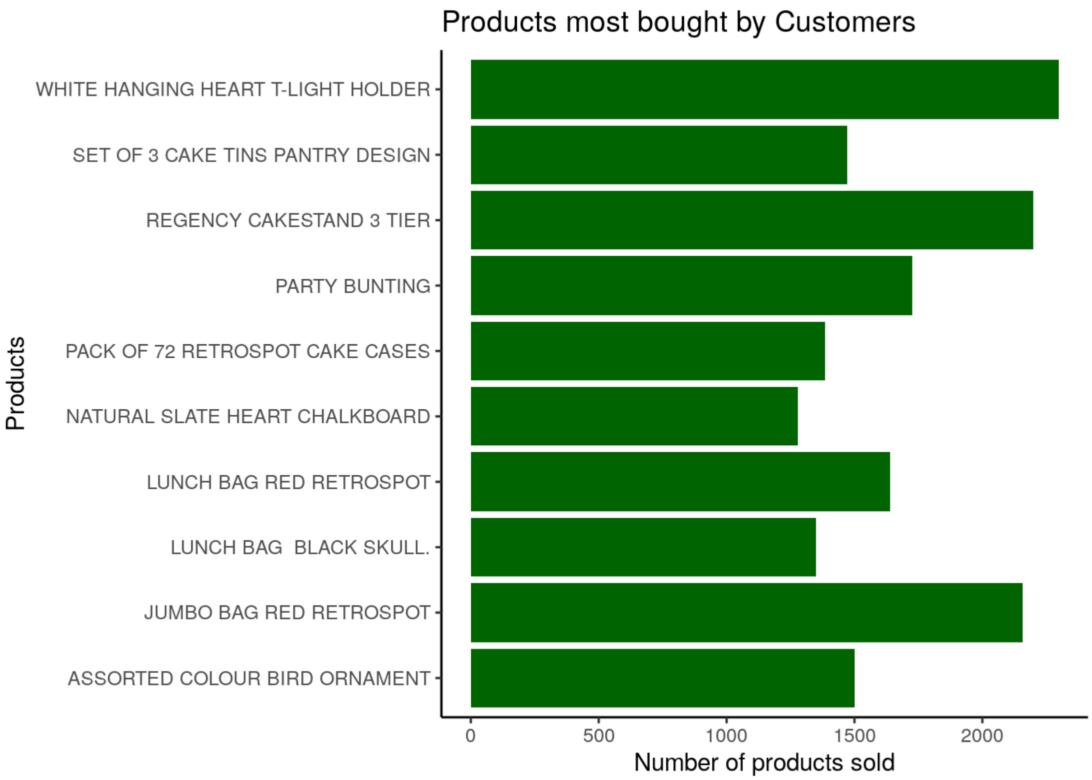
```
3 536365    84406B    CREAM CUPI...       8 2010-12-01 08:26:00      2.75
4 536365    84029G    KNITTED UN...       6 2010-12-01 08:26:00      3.39
5 536365    84029E    RED WOOLLY...       6 2010-12-01 08:26:00      3.39
6 536365    22752     SET 7 BABU...       2 2010-12-01 08:26:00      7.65
```

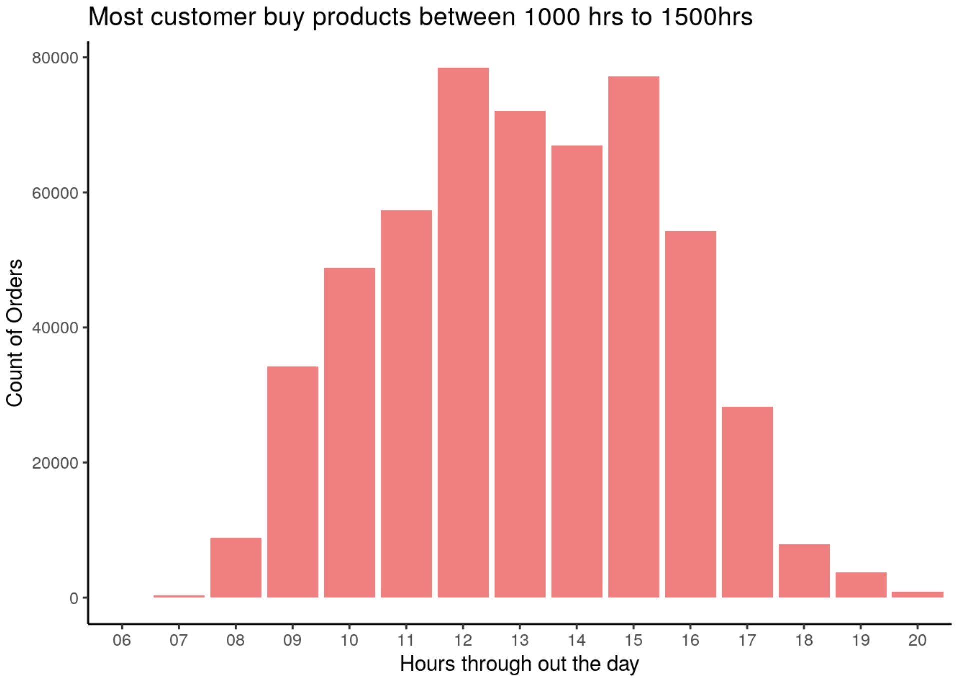
The graph displays that UK has the major portion of customers compared to other countries. Germany, France and Ireland are top 3 countries where online retail is working but it’s very low in comparison to UK. We will remove all other countries to focus only on UK.


UK shares the major customer base


Customer base without UK

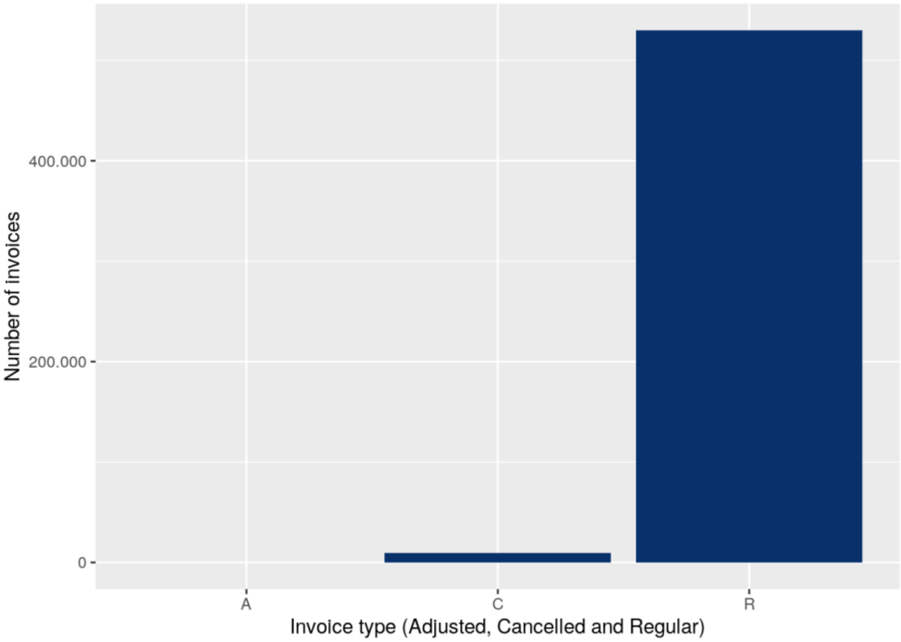
This graph represents the top 10 products which are mostly ordered by users. The next graph explains that between 10 am and 3pm most of the orders are placed on the online portal from every country.





Canceled orders

From the graph it can be inferred that there are a significant number of cancelled orders. We can check to see how this is impacting the Total Sales.



To check, we can check the relation of cancelled orders with top grossing sales by comparing the exact total sales amount.

```
InvoiceNo StockCode Description       Quantity Total_sales InvoicePrefix

  <chr>     <chr>     <fct>                <dbl>       <dbl> <chr>

       1 581483    23843     PAPER CRAFT , LI…    80995     168470. R

       2 C581484   23843     PAPER CRAFT , LI…   -80995    -168470. C

       3 541431    23166     MEDIUM CERAMIC T…    74215      77184. R

       4 C541433   23166     MEDIUM CERAMIC T…   -74215     -77184. C

       5 556444    22502     PICNIC BASKET WI…       60      38970  R

       6 C556445   M         Manual                  -1     -38970  C

       7 C537630   AMAZONFEE AMAZON FEE              -1     -13541. C

       8 537632    AMAZONFEE AMAZON FEE               1      13541. R

       9 C537651   AMAZONFEE AMAZON FEE              -1     -13541. C

       10 A563185   B         Adjust bad debt          1      11062. A
```

We can see from this data that the top 3 earning the highest total sales are cancelled orders and the rest of them are fine.

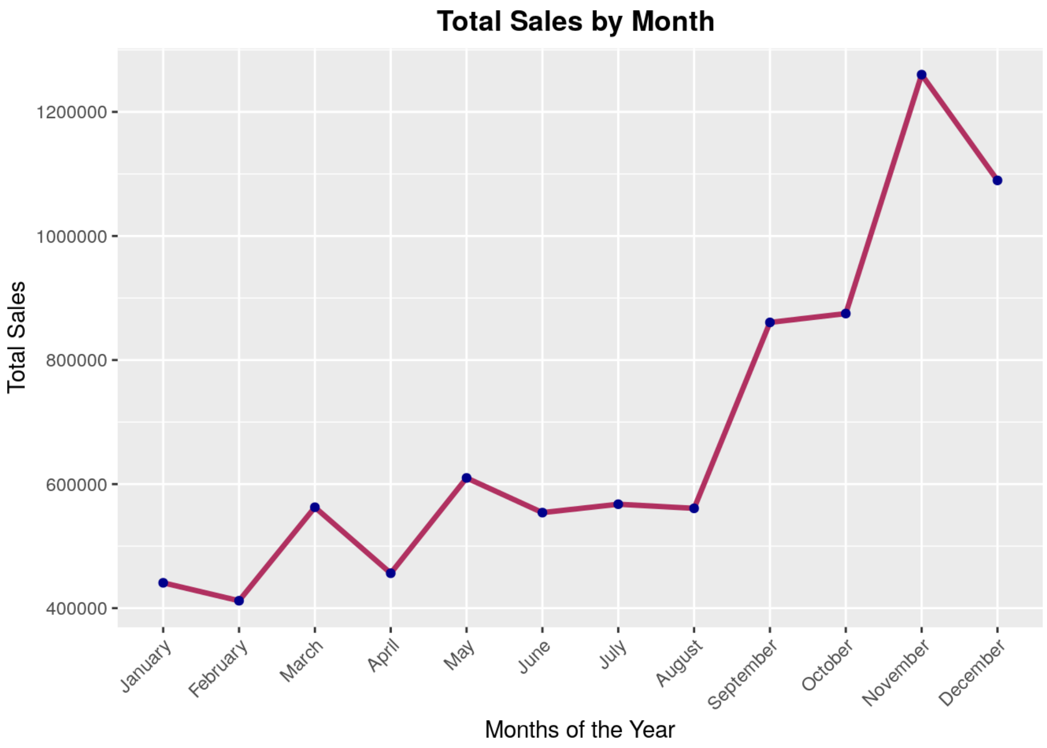
Cleaning unwanted orders and top cancelled orders

```
Groups:   StockCode, Description, Total_sales [20]

StockCode Description                          Total_sales

<chr>     <fct>                                      <dbl>

1 23243     SET OF TEA COFFEE SUGAR TINS PANTRY       7145.

2 21108     FAIRY CAKE FLANNEL ASSORTED COLOUR        6539.

3 23084     RABBIT NIGHT LIGHT                        4992

4 22086     PAPER CHAIN KIT 50'S CHRISTMAS            4782.

5 85123A    WHITE HANGING HEART T-LIGHT HOLDER        4632

6 48185     DOORMAT FAIRY CAKE                        4522.

7 23173     REGENCY TEAPOT ROSES                      4401

8 48185     DOORMAT FAIRY CAKE                        4254.

9 84879     ASSORTED COLOUR BIRD ORNAMENT             4176

10 22470     HEART OF WICKER LARGE                     4122.

11 22413     METAL SIGN TAKE IT OR LEAVE IT            3861

12 21623     VINTAGE UNION JACK MEMOBOARD              3828

13 23113     PANTRY CHOPPING BOARD                     3825.

14 22328     ROUND SNACK BOXES SET OF 4 FRUITS         3794.

15 23084     RABBIT NIGHT LIGHT                        3652.

16 22722     SET OF 6 SPICE TINS PANTRY DESIGN         3621

17 22197     POPCORN HOLDER                            3549
```

```
18 21175     GIN + TONIC DIET METAL SIGN                 3380

19 22086     PAPER CHAIN KIT 50'S CHRISTMAS              3322.

20 47556B    TEA TIME TEA TOWELS                         3315.
```

```
This data is a more accurate representation of legitimate total sales from
which multivariate analysis will be done monthly, daily and hourly. Removi
ng variables country, description, stockcode, invoicedate, invoiceprefix w
hile considering UK data for further analysis.
```

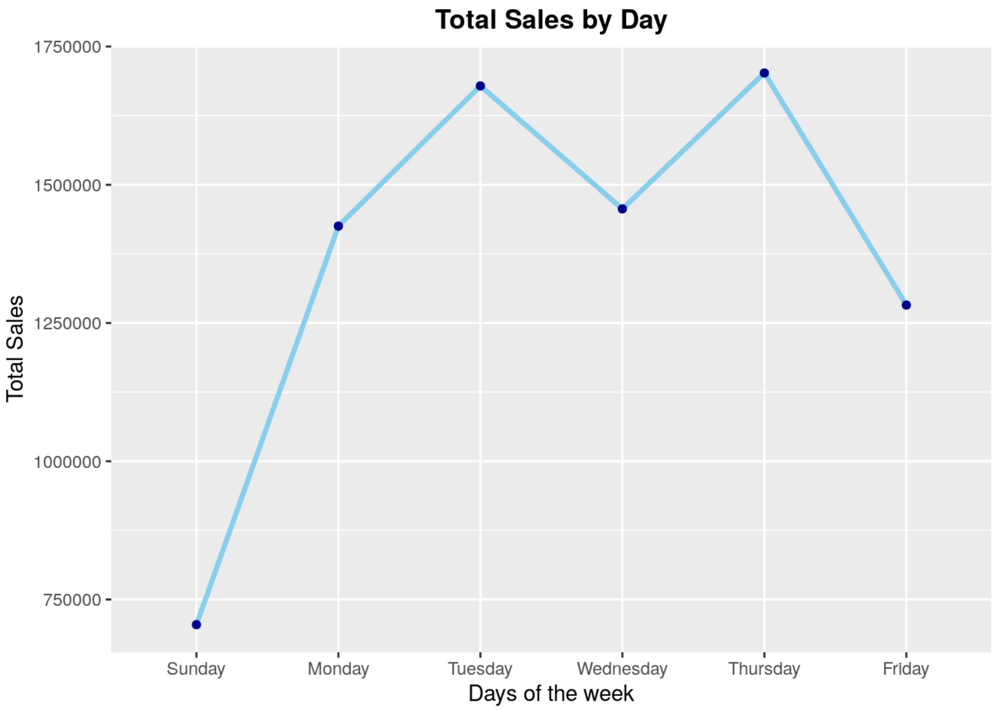
Total Sales (monthly)



This graph shows from September to December the sales are high in comparison to other months of the year. November is the peak season. This could be because of the number of holidays around that time.

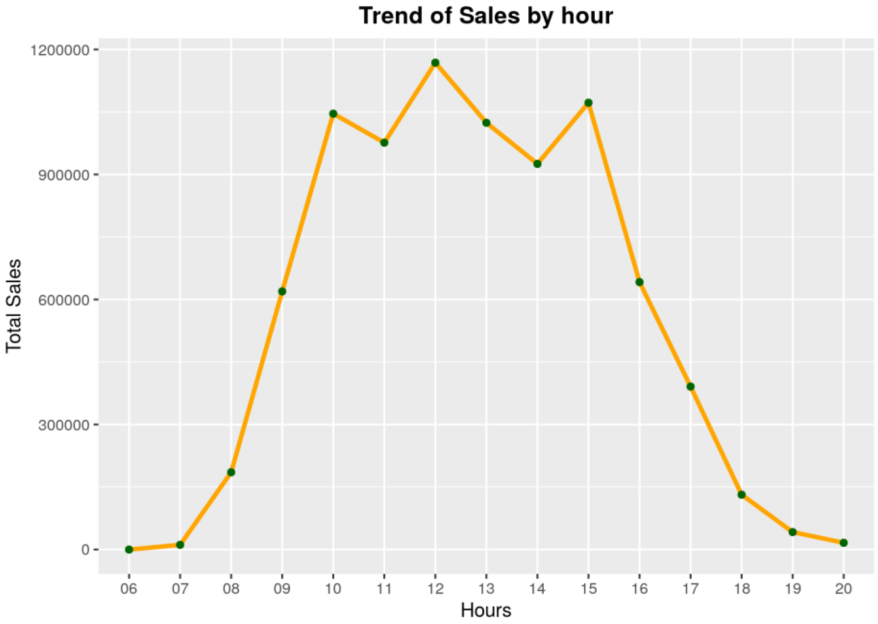
Total Sales (daily)

The daily graph shows that Tuesday's and Thursday's are the days where more sales are happening in comparison to other weekdays.



Total Sales (hourly)

This graph shows that the hours between 10am and 3pm generate the most sales during the day. This would probably be because this is the time when most customers are at work, and thus, putting in orders.



From the above analysis, we can identify that the busiest time orders are put in, are in November to December for most of the holidays, on a Tuesday or Thursday, and between the hours of 10am and 3pm.