Introduction:

The provided data looks like a sales data of a wholesale/retail company. It has product, customer and sales related data. In analyzing the data, as a first step, I have used R to clean the data. I learnt R during this exercise to perform the initial steps as I have no previous experience in R. As next step, I used Excel to extract month, date, year data from the datetime field. Then, I have used Tableau to perform the analysis and create the visualization. I have included the R code and visualizations in this document.

Data Cleanup:

Importing a CSV file as a dataframe in R. It has 541909 rows and 8 columns: Invoice no, Stock code, Description quantity, Invoice date, Unit price, Customer ID and Country.

```
> data <- read.csv(file.choose(), header = T)
> nrow(data)
[1] 541909
> head(data)
  InvoiceNo StockCode                  Description Quantity  InvoiceDate UnitPrice CustomerID        Country
1  C581569    20979   36 PENCILS TUBE RED RETROSPOT      -5 12/9/11 11:58      1.25      17315 United Kingdom
2  C581569    84978 HANGING HEART JAR T-LIGHT HOLDER      -1 12/9/11 11:58      1.25      17315 United Kingdom
3  C581568    21258      VICTORIAN SEWING BOX LARGE      -5 12/9/11 11:57     10.95      15311 United Kingdom
4  C581499        M                          Manual      -1 12/9/11 10:28    224.69      15498 United Kingdom
5  C581490    22178  VICTORIAN GLASS HANGING T-LIGHT     -12  12/9/11 9:57      1.95      14397 United Kingdom
6  C581490    23144  ZINC T-LIGHT HOLDER STARS SMALL     -11  12/9/11 9:57      0.83      14397 United Kingdom
```

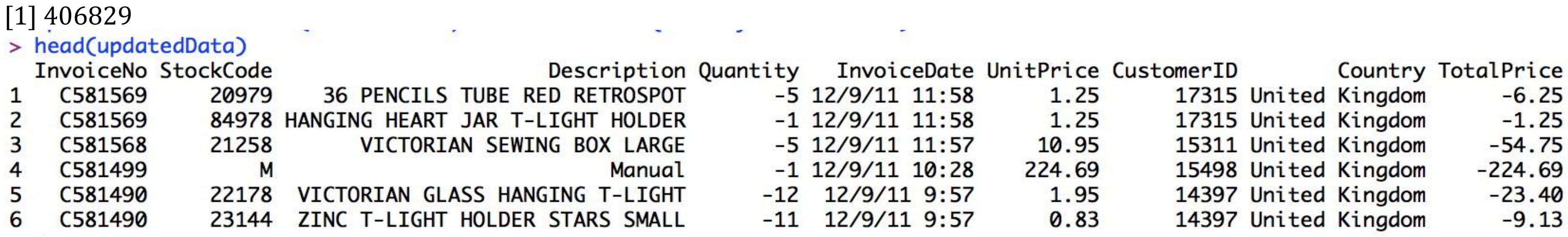
As a first step in cleaning data, Nulls should be handled. The nulls in the description field are not ignored. Because if there are nulls in the description field, it can be checked for duplicate stock codes (Two stock codes that have the same description), if not, those nulls can be replaced with the description that matches the stock code; On the contrary, if there are duplicates, the rows with null in description is removed. Nulls in every other column(except description) will require that row to be removed. Country column that has the term "Unspecified" is also removed for our analysis.

```
> cleanedData <- data[complete.cases(data$InvoiceNo), ]
> nrow(cleanedData)
[1] 541909
> cleanedData <- data[complete.cases(data$StockCode), ]
> nrow(cleanedData)
[1] 541909
> cleanedData <- data[complete.cases(data$CustomerID), ]
> nrow(cleanedData)
[1] 406829
> cleanedData <- data[complete.cases(cleanedData$UnitPrice), ]
> nrow(cleanedData)
[1] 406829
> cleanedData <- data[complete.cases(data$CustomerID), ]
> cleanedData <- cleanedData[cleanedData$Country != "Unspecified", ]
> nrow(cleanedData)
[1] 406829
> cleanedData <- cleanedData[complete.cases(cleanedData$UnitPrice), ]
> nrow(cleanedData)
[1] 406829
> cleanedData <- cleanedData[complete.cases(cleanedData$InvoiceDate), ]
> nrow(cleanedData)
[1] 406829
> cleanedData <- cleanedData[complete.cases(cleanedData$Quantity), ]
> nrow(cleanedData)
[1] 406829
> cleanedData <- cleanedData[complete.cases(cleanedData$InvoiceDate), ]
> nrow(cleanedData)
[1] 406829
```

Overall 135,080 lines are omitted as part of the cleanup.

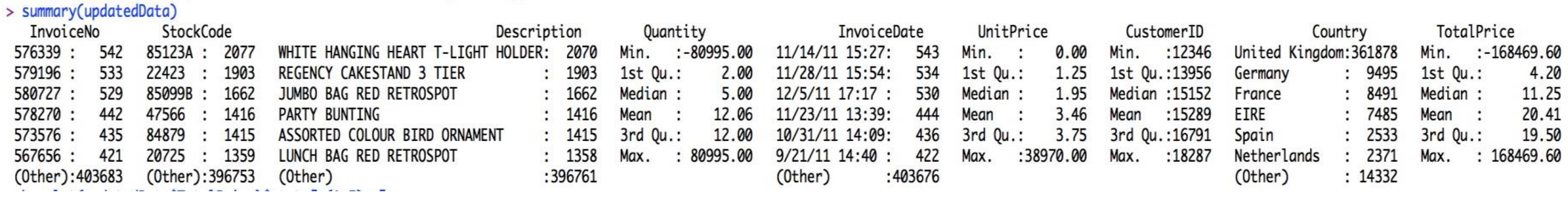
Added another column called total price that is the product of quantity and unit price.

```
> updatedData <- within(cleanedData, TotalPrice <- Quantity * UnitPrice)
> nrow(updatedData)
[1] 406829
```



```
> head(updatedData)
  InvoiceNo StockCode                      Description Quantity  InvoiceDate UnitPrice CustomerID        Country TotalPrice
1   C581569     20979    36 PENCILS TUBE RED RETROSPOT       -5 12/9/11 11:58      1.25      17315 United Kingdom      -6.25
2   C581569     84978 HANGING HEART JAR T-LIGHT HOLDER       -1 12/9/11 11:58      1.25      17315 United Kingdom      -1.25
3   C581568     21258       VICTORIAN SEWING BOX LARGE       -5 12/9/11 11:57     10.95      15311 United Kingdom     -54.75
4   C581499         M                           Manual       -1 12/9/11 10:28    224.69      15498 United Kingdom    -224.69
5   C581490     22178  VICTORIAN GLASS HANGING T-LIGHT      -12  12/9/11 9:57      1.95      14397 United Kingdom     -23.40
6   C581490     23144  ZINC T-LIGHT HOLDER STARS SMALL      -11  12/9/11 9:57      0.83      14397 United Kingdom      -9.13
```

```
> summary(updatedData)
```



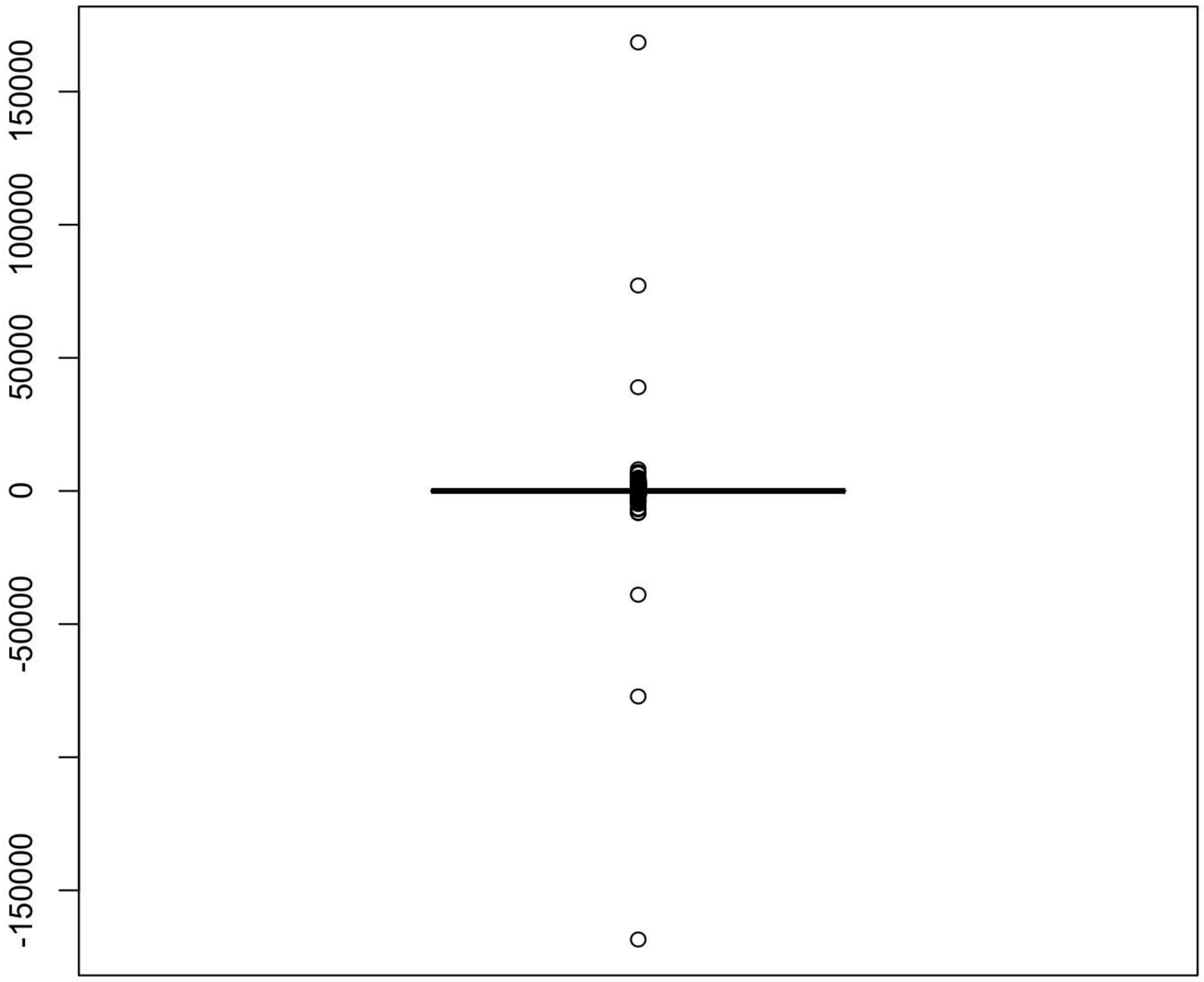
```
> summary(updatedData)
   InvoiceNo        StockCode                                  Description        Quantity                InvoiceDate        UnitPrice           CustomerID             Country          TotalPrice
 576339 :    542   85123A :  2077   WHITE HANGING HEART T-LIGHT HOLDER:  2070   Min.    :-80995.00   11/14/11 15:27:    543   Min.    :    0.00   Min.    :12346   United Kingdom:361878   Min.    :-168469.60
 579196 :    533   22423  :  1903   REGENCY CAKESTAND 3 TIER          :  1903   1st Qu.:     2.00   11/28/11 15:54:    534   1st Qu.:    1.25   1st Qu.:13956   Germany       :  9495   1st Qu.:      4.20
 580727 :    529   85099B :  1662   JUMBO BAG RED RETROSPOT           :  1662   Median :     5.00   12/5/11 17:17 :    530   Median :    1.95   Median :15152   France        :  8491   Median :     11.25
 578270 :    442   47566  :  1416   PARTY BUNTING                     :  1416   Mean   :    12.06   11/23/11 13:39:    444   Mean   :    3.46   Mean   :15289   EIRE          :  7485   Mean   :     20.41
 573576 :    435   84879  :  1415   ASSORTED COLOUR BIRD ORNAMENT     :  1415   3rd Qu.:    12.00   10/31/11 14:09:    436   3rd Qu.:    3.75   3rd Qu.:16791   Spain         :  2533   3rd Qu.:     19.50
 567656 :    421   20725  :  1359   LUNCH BAG RED RETROSPOT           :  1358   Max.   : 80995.00   9/21/11 14:40 :    422   Max.   :38970.00   Max.   :18287   Netherlands   :  2371   Max.   : 168469.60
 (Other):403683   (Other):396753   (Other)                           :396761                        (Other)       :403676                                        (Other)       : 14332
```

Let us get the final dataframe as csv for further analysis in Tableau.

```
> write.csv(updatedData, "/Users/NagaSoundari/Documents/UpdatedDataSet.csv")
```

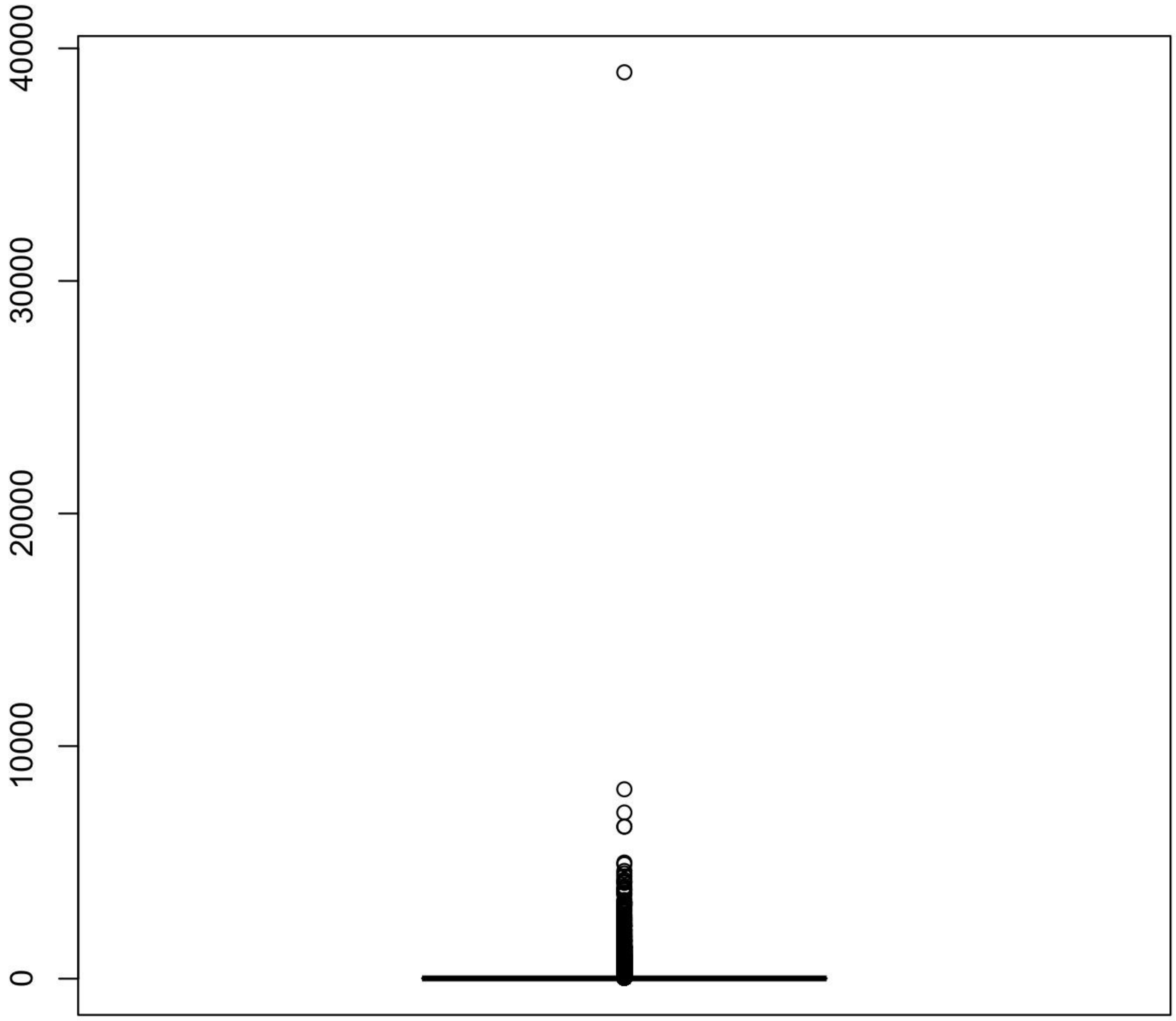
Now let us look at the data spread of Total price using a Box and Whisker plot. This is done to check if the data is skewed by any outliers.

```
> boxplot(updatedData$TotalPrice)$stats[c(1,5), ]
[1] -18.75  42.45
```



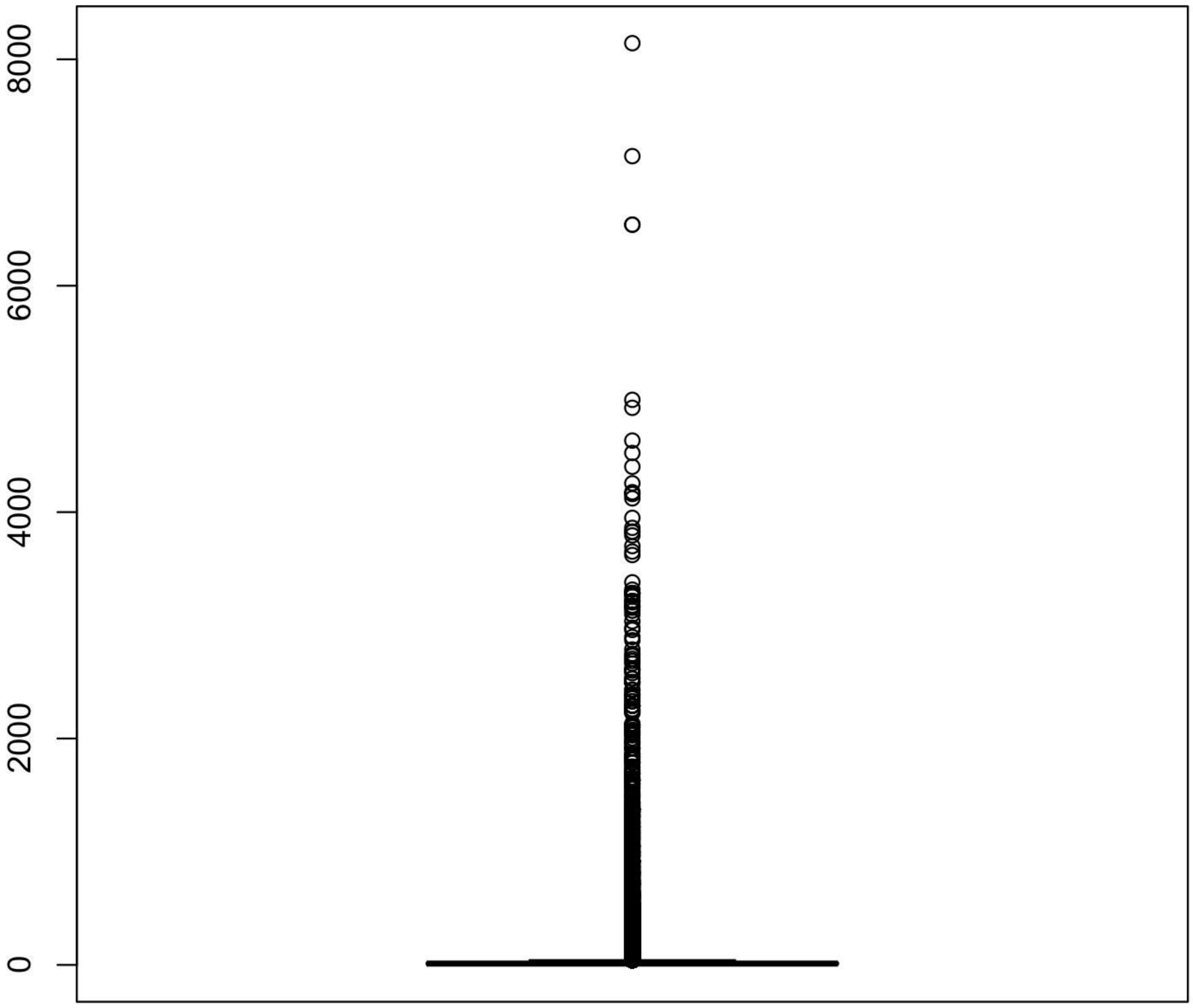
Let’s zoom in to see around the median, starting from 0 and 40,000.

```
> positiveData = subset(updatedData,  updatedData$TotalPrice >= 0 & updatedData$TotalPrice <= 40000)
> boxplot(positiveData$TotalPrice)$stats[c(1,5), ]
[1]  0.00 42.45
```



Zooming in further to see data between 0 and 10,000.

```
> positiveSubsetData = subset(updatedData,  updatedData$TotalPrice >= 0 & updatedData$TotalPrice <= 10000)
> boxplot(positiveSubsetData$TotalPrice)$stats[c(1,5), ]
[1]  0.00 42.45
```



Insights from the data Cleaning:
From the data, we can guess that this a sales data of a global sales company that has customers across different countries. Most of the customers seems to be wholesale customers. The data is available for the year 2011 and December of 2010. As the data is not available for entire 2010, sales data cannot be compared between years. There are 8 customer ID's which corresponds to two countries. This might be either duplicate customer ID or the customer is a multinational company. As we don't have sufficient information on this, the duplicate customer ID is not removed or altered. The Country data for some of the sales row were mentioned as "Unspecified". However, their contribution to the overall purchase is higher than few other countries. Hence this data cannot be ignored.

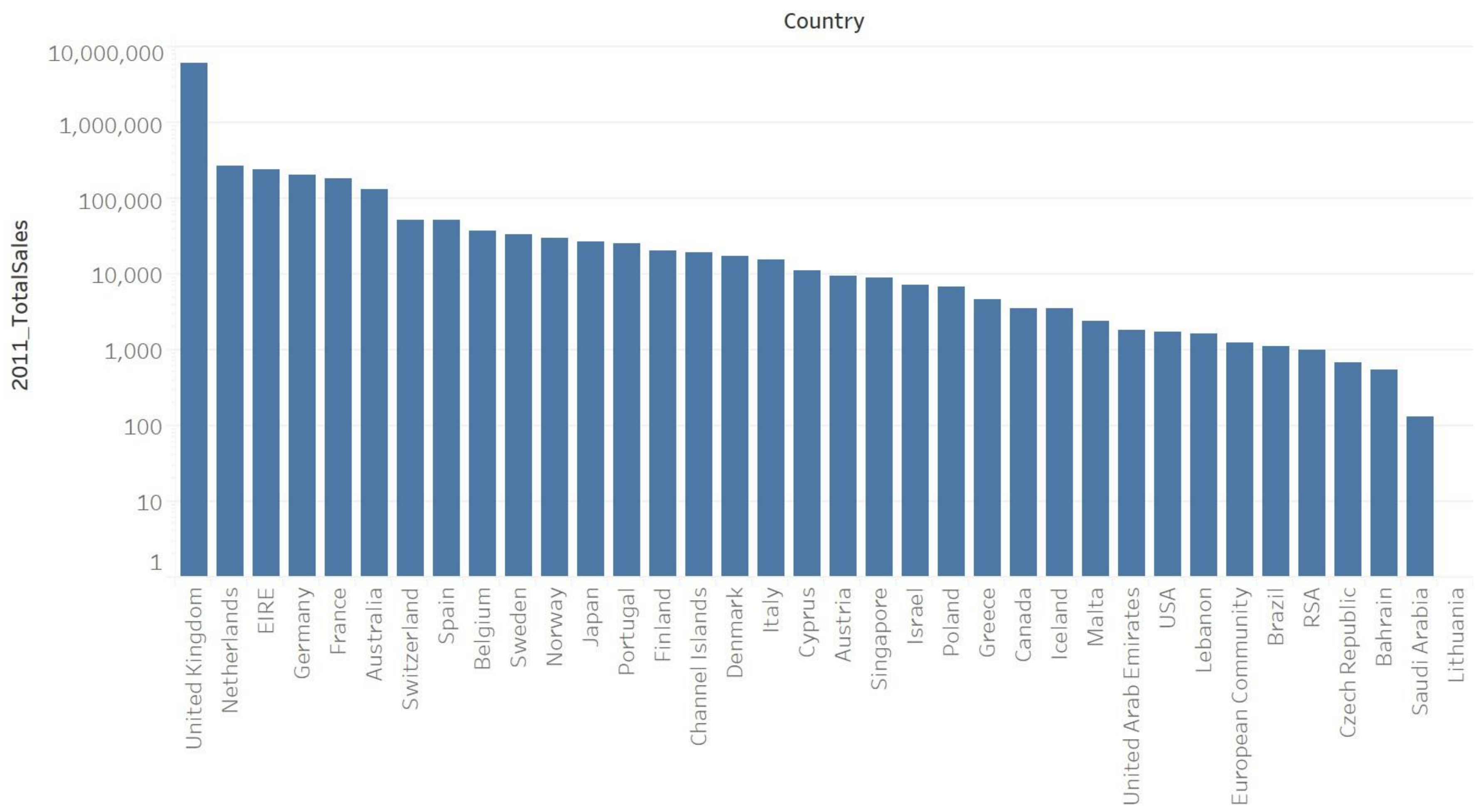
Now we can proceed to the visualization part where interesting insights can be identified. Some of the insights that can be drawn are:

1. Compare sales across different countries
2. Compare the sales in different months of a year
3. Compare December of 2010 and 2011.
4. Compare the sales in different days of week
5. Find the top 5 customers making most purchase
6. Compare sales of different product across countries
7. Which is the most popular product sold in each country
8. Is there any specific product that is popular among all the countries
9. What type of products are cancelled
10. Is there any specific customer cancelling the order
11. Is there any specific product that is returned by most of the customers

Let us analyze them one by one:
The analysis is done for the year 2011 (omitted December 2010 data).

Sales across countries

Chart: 1

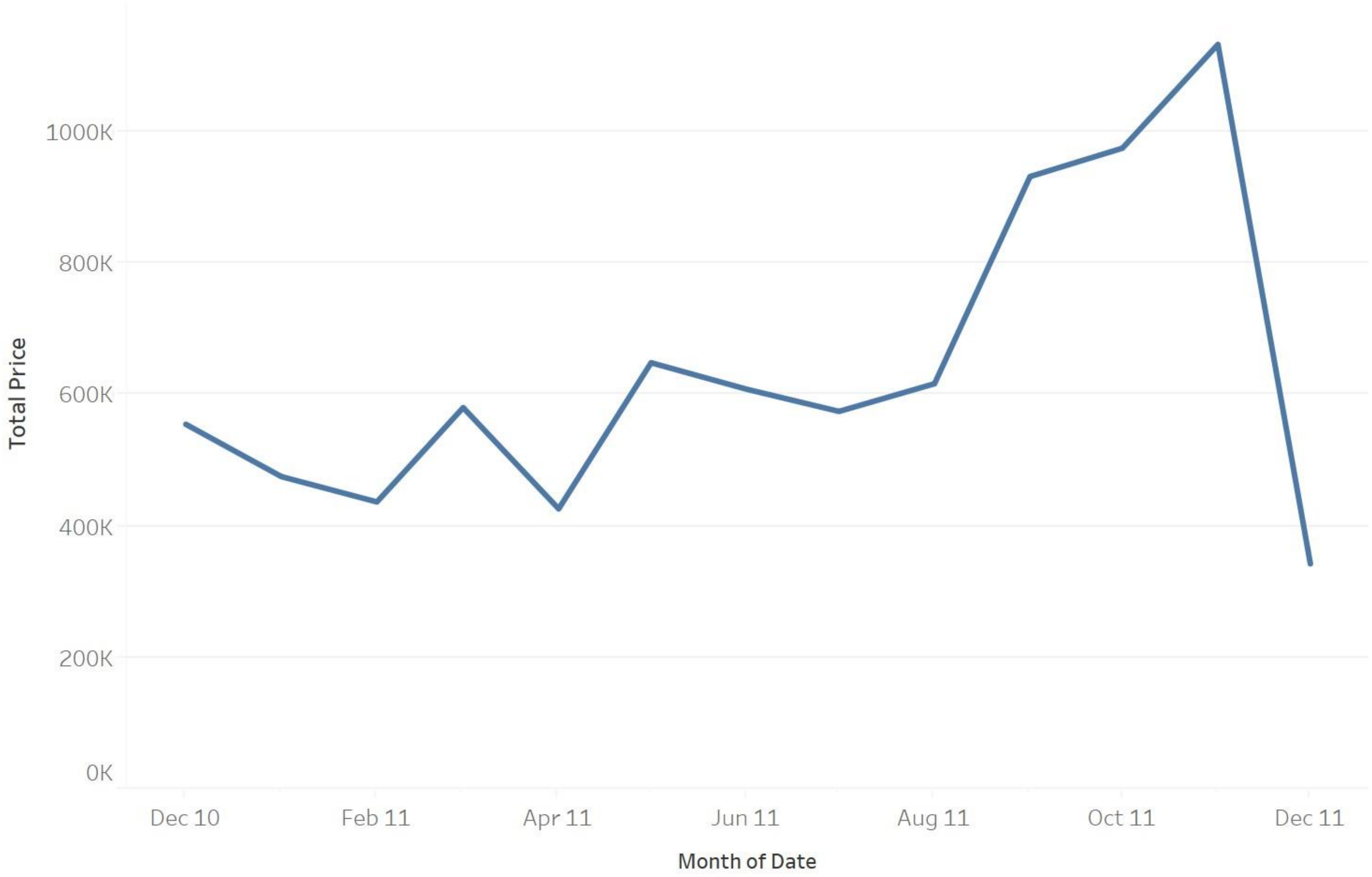


Sum of 2011_TotalSales for each Country.

United Kingdom has the most sales (6,284,074) for the year 2011. Lithuania is the country that has shown negative sales. But comparing only the sales data of different countries might not be an optimal solution. These countries differ in their geographical area and population size. Hence computing a ratio of population data vs sales data for each country and comparing that ratio will show the exact sales comparison. This will also be useful to identify the potential for growth in each country. But considering the available data, UK is the biggest market. Now let us drill down to look at the sales in different months of year.

Rolling 13 months Sales

Chart: 2

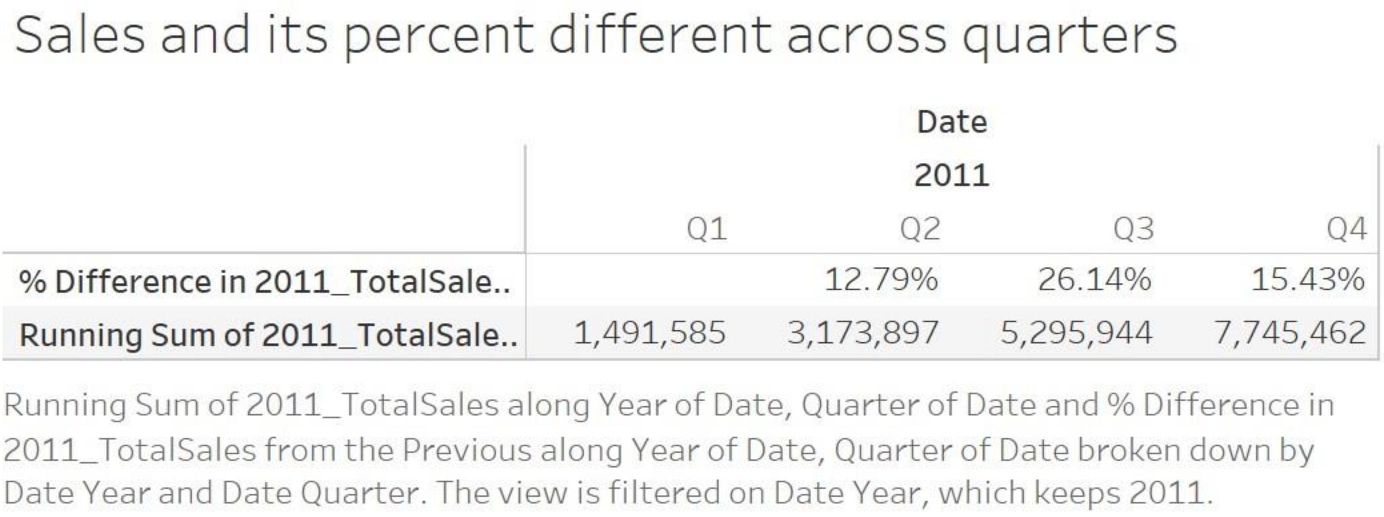


The trend of sum of Total Price for Date Month.

The maximum sales was in November (1,132,408) and minimum in December (342,506). If provided with data over years, pattern availing among different months would have been identified. But comparing December 2010 sales with December 2011 sales shows that the sales has dropped in December 2011 by approximately 38%. From the 2011 sales, it is clear that November sales was at peak. The sales percentage difference between November and December 2011 was 70% . If the same scenario prevailed in 2010, then the November 2010 sales would have been much higher that November 2011. Hence there could be a decrease in sales in 2011 which can be confirmed if the 2010 sales data is available.

The quarter sales data is compared in the below chart.

Chart: 3



Sales and its percent different across quarters

	Date			
	2011			
	Q1	Q2	Q3	Q4
% Difference in 2011_TotalSale..		12.79%	26.14%	15.43%
Running Sum of 2011_TotalSale..	1,491,585	3,173,897	5,295,944	7,745,462

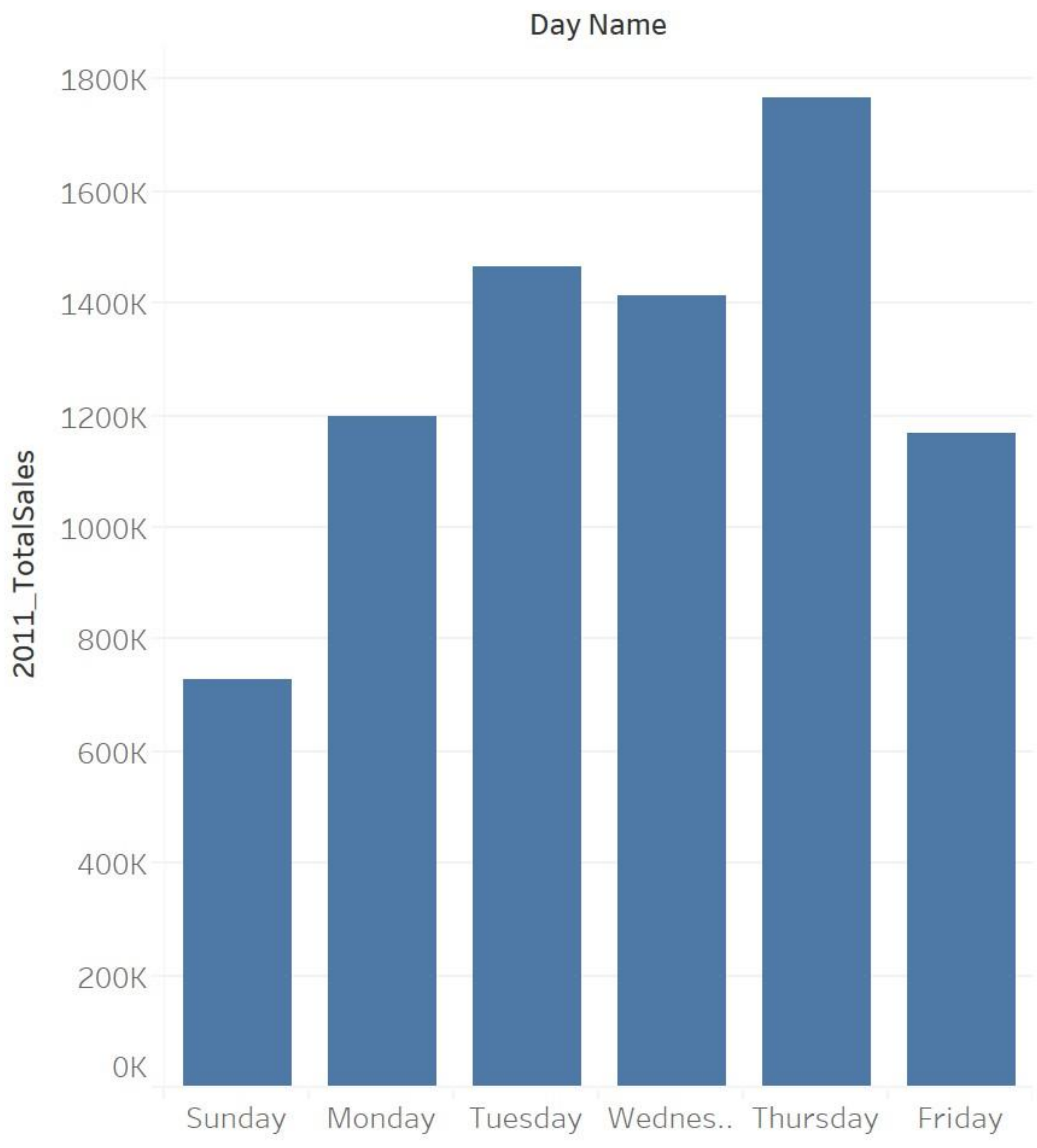
Running Sum of 2011_TotalSales along Year of Date, Quarter of Date and % Difference in 2011_TotalSales from the Previous along Year of Date, Quarter of Date broken down by Date Year and Date Quarter. The view is filtered on Date Year, which keeps 2011.

There is a tremendous rise and fall in the sales percentage over different quarters. Though the sales were at peak during the months of September, October and November, quarter 4 has shown a striking fall in sales.

Looking at the sales across different days of a week reveals an interesting information that none of the sales were done on Saturdays. So it can be either assumed as a non-working day or the reason behind it should be identified. Also the sales on Thursdays are high and Sundays are low showing a staunch difference. Based on this sales staff can be allocated and deallocated respectively for cost benefits.

Sales across weekdays

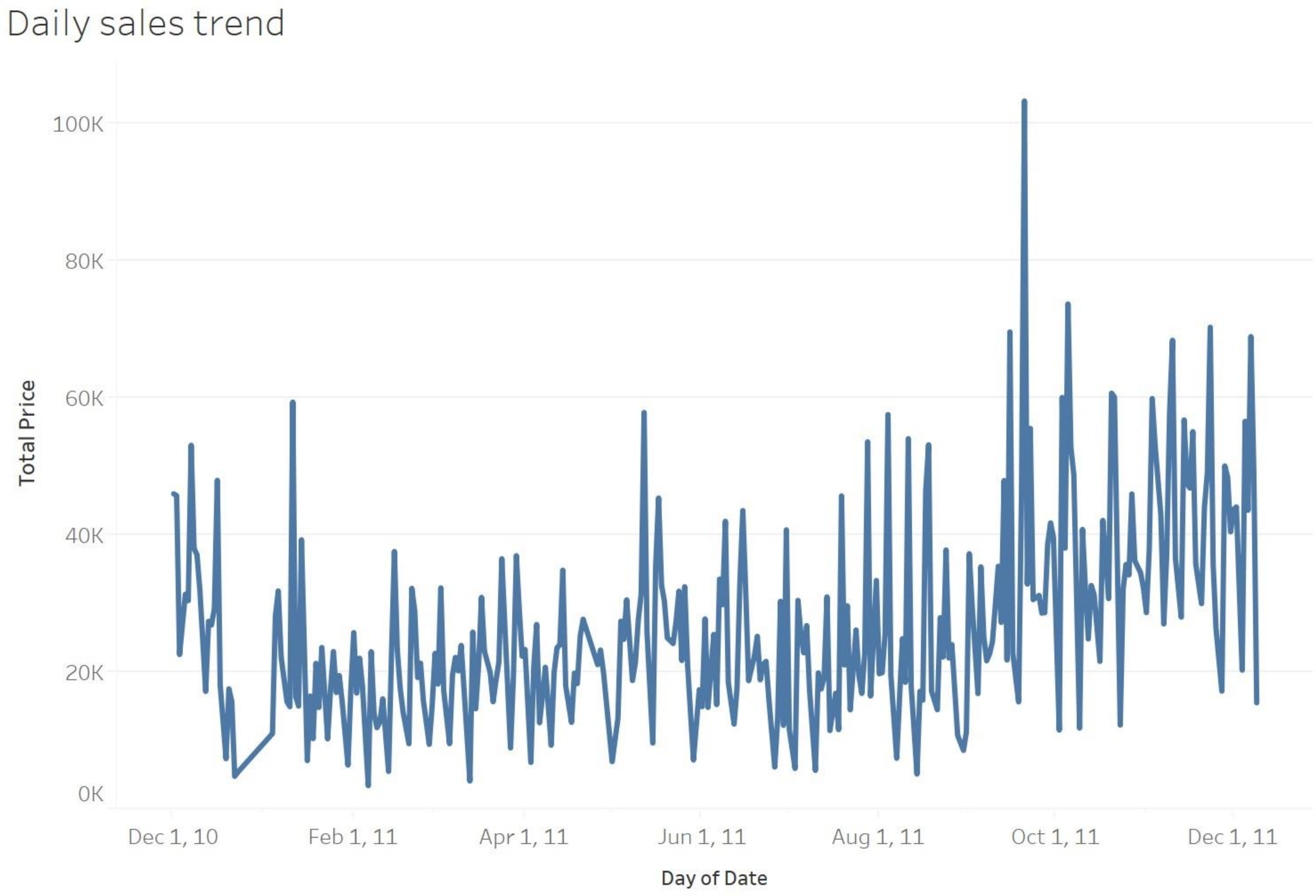
Chart: 4



Sum of 2011_TotalSales for each Day Name.

As the day of week showed an engrossing information, the trend across the year were also analyzed to check if the sales show any pattern around same week of different months (i.e., any specific week of all months has a significant sales). Though any of such pattern could not be seen, the sales was sky-high (103,385) on September 20 that is 30% of the whole December month sales.

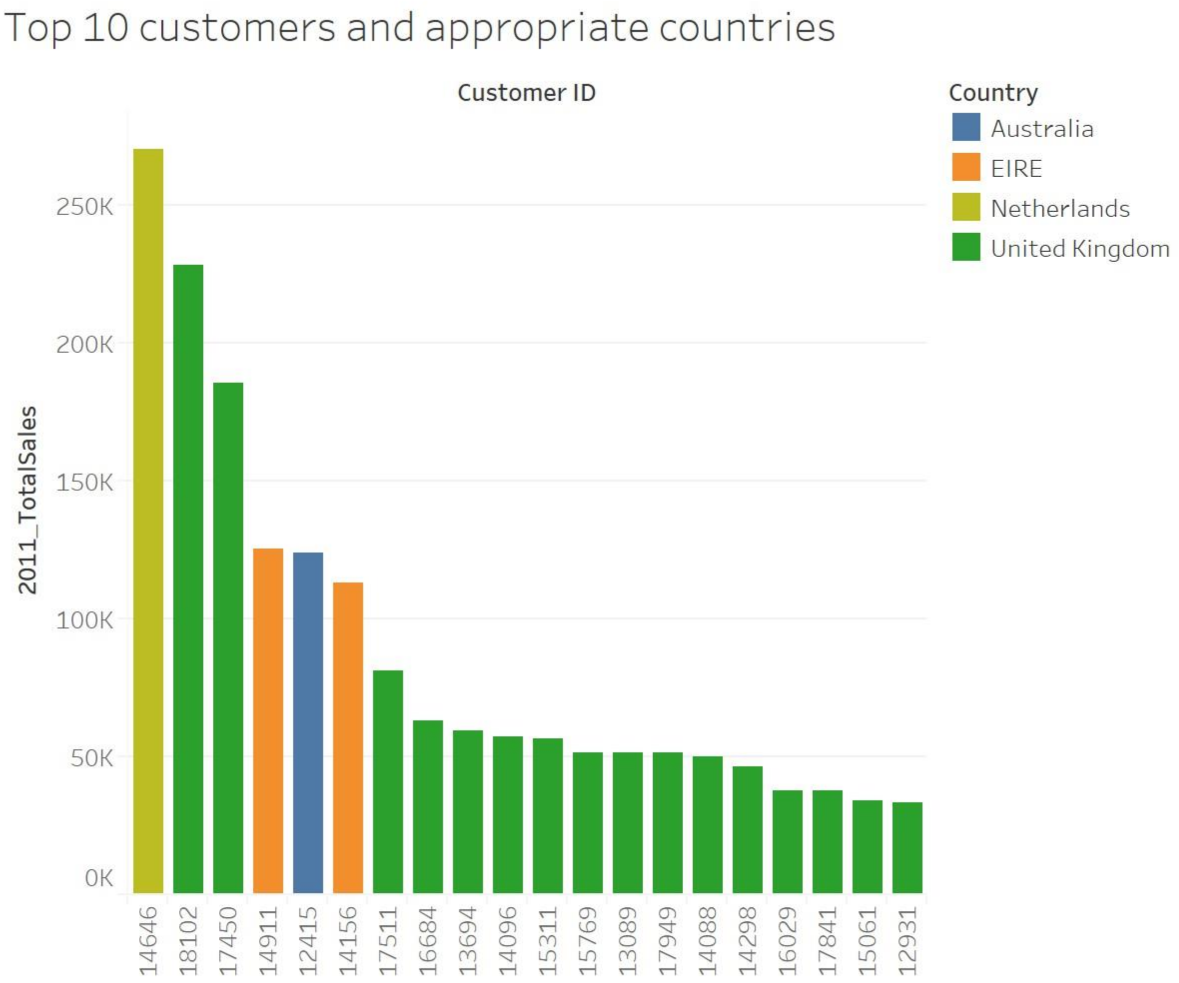
Chart: 5



The trend of sum of Total Price for Date Day.

The top 10 customers of the company who bring in most sales is identified. This is done also to identify the countries to which they belong.

Chart: 6



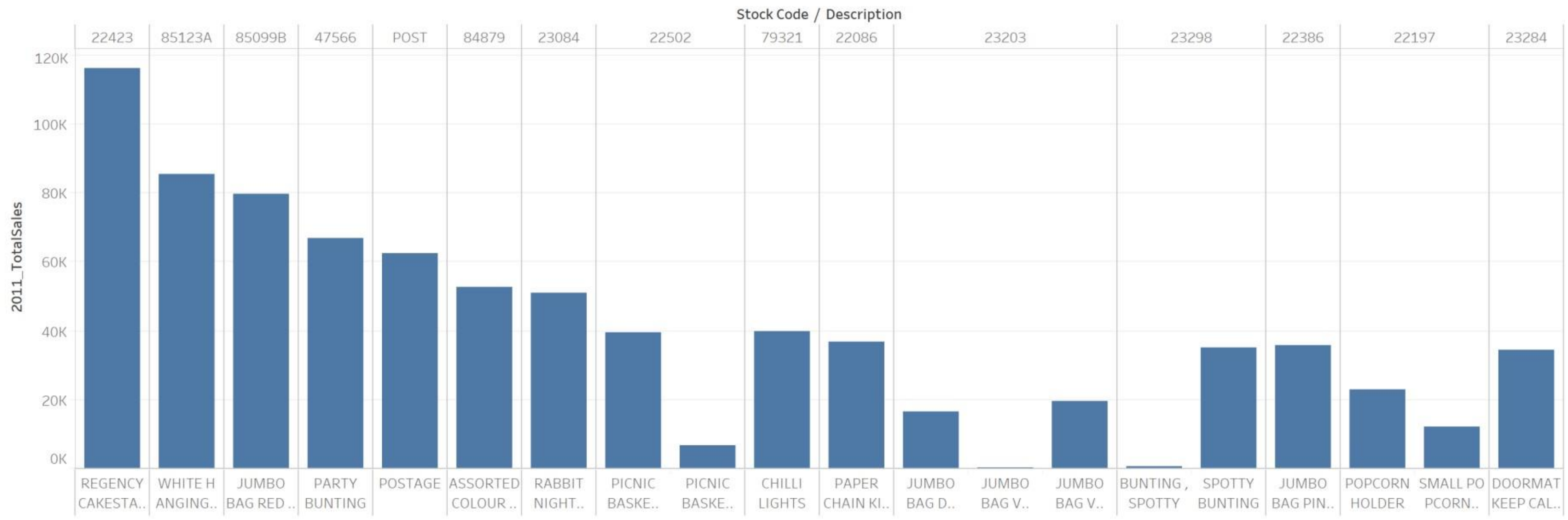
Sum of 2011_TotalSales for each Customer ID. Color shows details about Country. The view is filtered on Customer ID, which keeps 20 of 4,372 members.

Moving on with the analysis based on Products,

The sales are dominated completely by United Kingdom. Thus, the most sold products in UK and other countries except UK were found and compared against each other. This can tell us which products are popular and are to be pitched in different countries.

Chart: 7

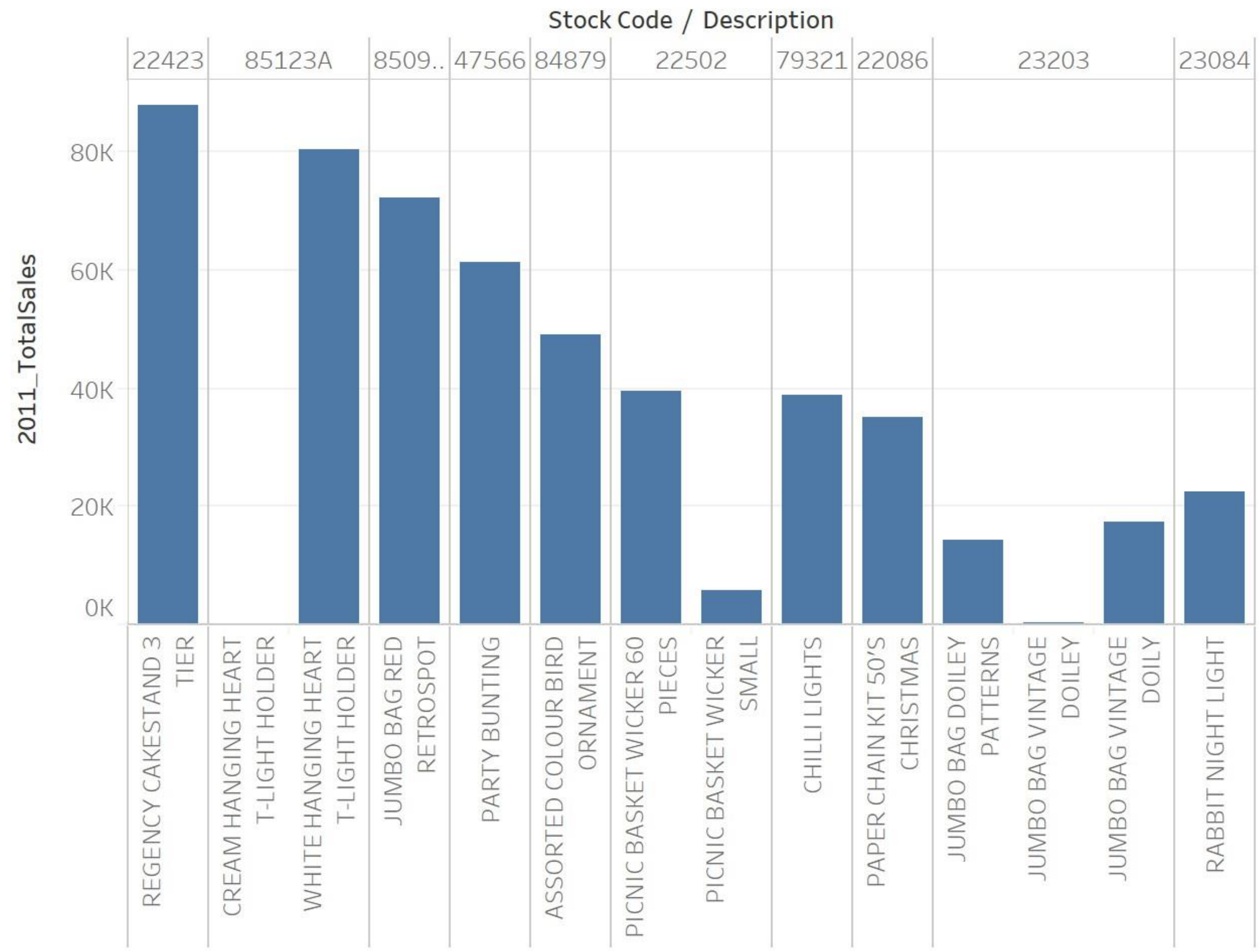
Top selling products (Overall)



Sum of 2011_TotalSales for each Description broken down by Stock Code. The view is filtered on Stock Code and Description. The Stock Code filter keeps 15 of 3,684 members. The Description filter keeps 3,884 of 3,885 members.

Chart: 8

Top 10 products in UK



Sum of 2011_TotalSales for each Description broken down by Stock Code. The data is filtered on Country, which keeps United Kingdom. The view is filtered on Stock Code, which keeps 10 of 3,684 members.

Most of the top selling products are similar in both cases (UK and Other countries). Only exception is the Product POSTAGE which is the top most selling product in other countries while in UK it shows a negative sale or returns.

The most selling products should be given more importance and can be customized with different color, pattern, variety to attract customers which in turn will reflect in the sales. On the other hand, it is equally important to identify the products that are sold the least. These products can either be discontinued or analyzed further to improve their sales.

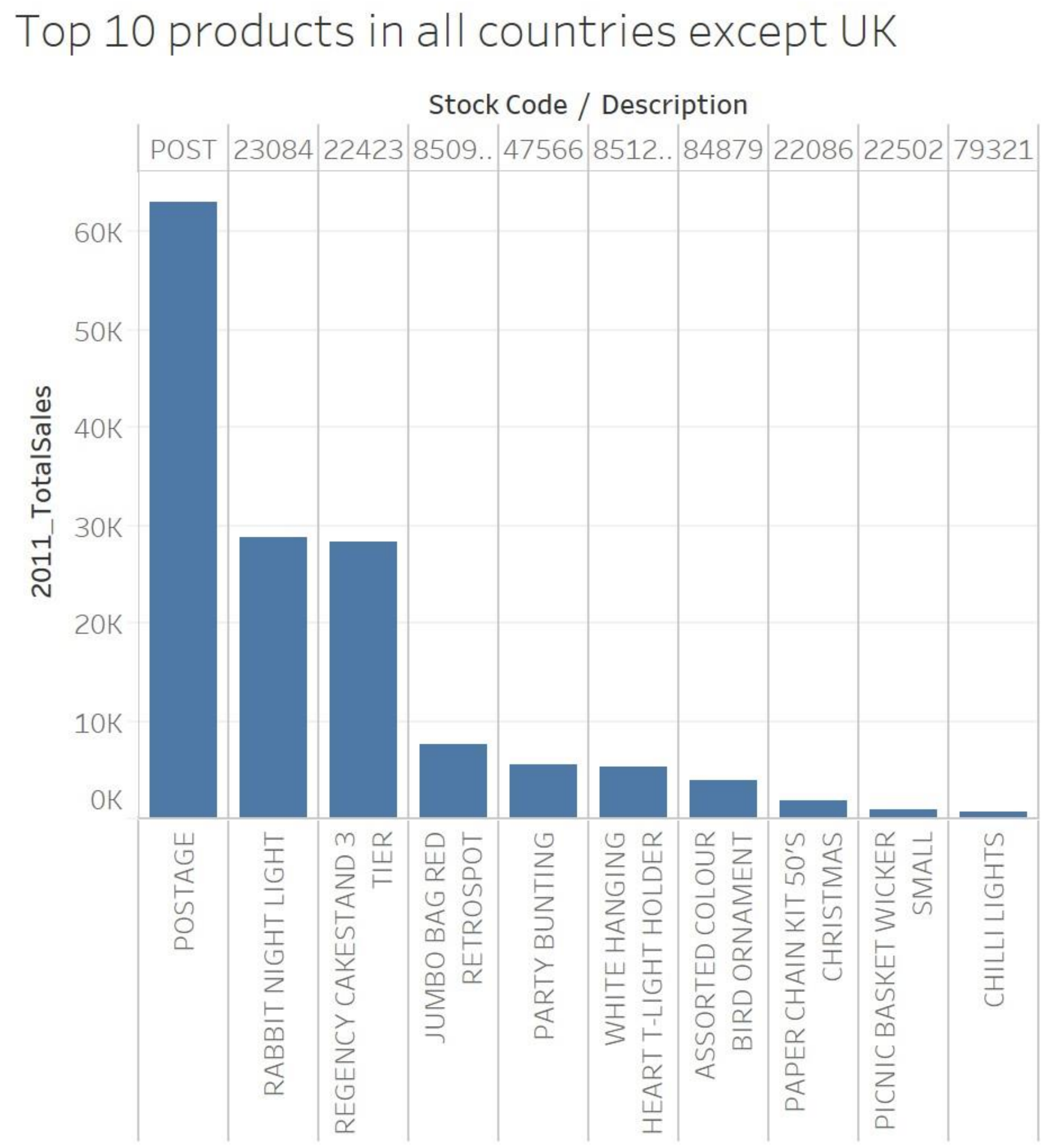
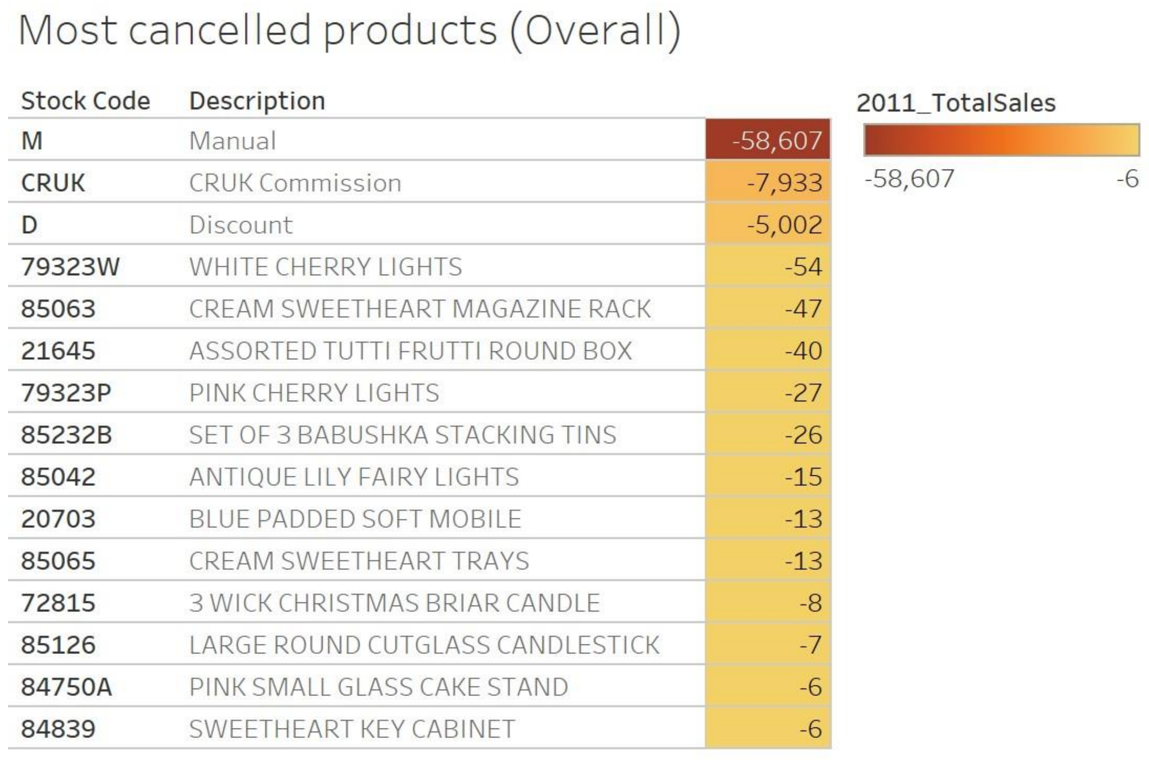


Chart: 9

Sum of 2011_TotalSales for each Description broken down by Stock Code. The data is filtered on Country, which excludes United Kingdom and Unspecified. The view is filtered on Stock Code, which keeps 10 of 3,684 members.

Another criteria that needs to be analyzed are the products that are cancelled the most. This is done to identify if there is a specific product that is returned by most customers.



Most cancelled products (Overall)

Stock Code	Description	
M	Manual	-58,607
CRUK	CRUK Commission	-7,933
D	Discount	-5,002
79323W	WHITE CHERRY LIGHTS	-54
85063	CREAM SWEETHEART MAGAZINE RACK	-47
21645	ASSORTED TUTTI FRUTTI ROUND BOX	-40
79323P	PINK CHERRY LIGHTS	-27
85232B	SET OF 3 BABUSHKA STACKING TINS	-26
85042	ANTIQUE LILY FAIRY LIGHTS	-15
20703	BLUE PADDED SOFT MOBILE	-13
85065	CREAM SWEETHEART TRAYS	-13
72815	3 WICK CHRISTMAS BRIAR CANDLE	-8
85126	LARGE ROUND CUTGLASS CANDLESTICK	-7
84750A	PINK SMALL GLASS CAKE STAND	-6
84839	SWEETHEART KEY CABINET	-6

Chart: 10

Sum of 2011_TotalSales broken down by Stock Code and Description. Color shows sum of 2011_TotalSales. The marks are labeled by sum of 2011_TotalSales. The view is filtered on Stock Code, which keeps 15 of 3,684 members.

From the above chart, it can be seen that the most returned (or negative valued) product's descriptions are Manual, Discount and CRUK Commission. So it can be assumed that the cancellation of products are done manually. Looking further, the discount amount is high, the sales incurred as a consequence of this discount should be higher. Also the CRUK commission amount is high; this clearly shows that the sales person is bringing in more business. Hence these can be promoted to increase the sales.

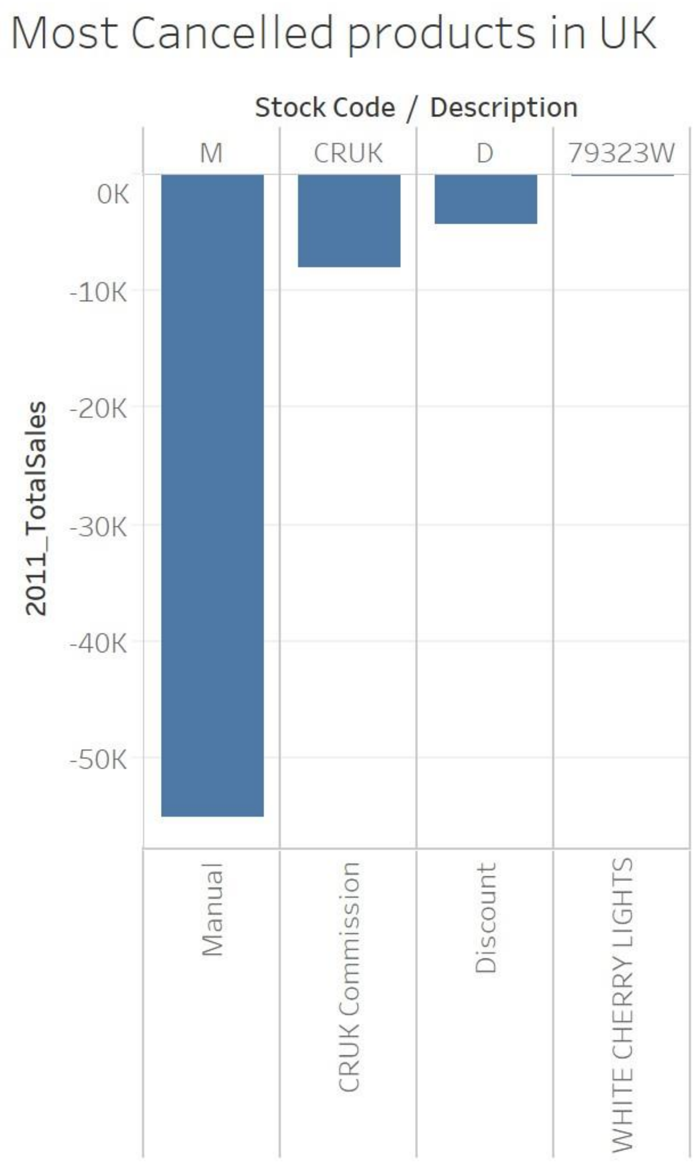
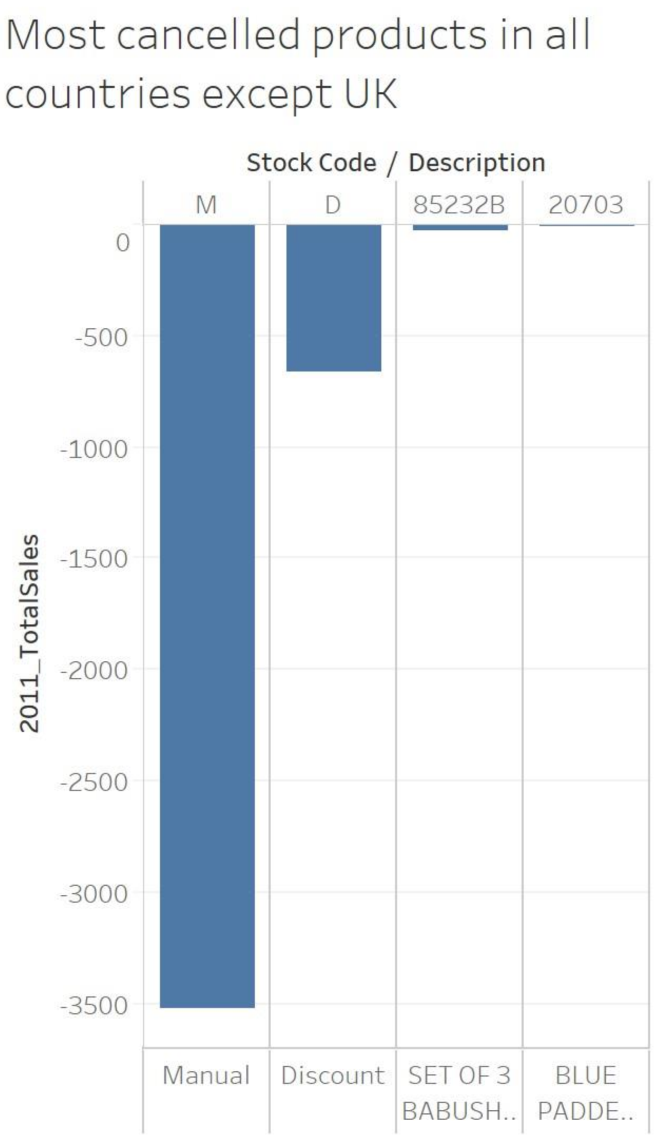


Chart: 11

Sum of 2011_TotalSales for each Description broken down by Stock Code. The data is filtered on Country, which keeps United Kingdom. The view is filtered on Stock Code, which keeps 79323W, CRUK, D and M.



Sum of 2011_TotalSales for each Description broken down by Stock Code. The data is filtered on Country, which excludes United Kingdom and Unspecified. The view is filtered on Stock Code, which keeps 10 of 3,684 members.

Conclusion

The provided dataset contained only data for December month of 2010. The rest of the data doesn't contain sales information on Saturdays. The data has "Unspecified" in the country column and it is not clear how to clean/classify unspecified.

The analysis done above gives us few insights from the data like,

- Days of week and months of year to be concentrated
- Countries in which sales can be projected
- Customers to be valued
- Products to be given attention
- Discount/Commission to be used as incentive

The analysis can still be drilled down to make detailed analysis with much more data that can provide us with business insights.