Online Retail

Shinjini,Nitish,Sambit

25 December 2017

Data Set Information:

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Import the Dataset

A look at the dataset we're working on

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047	United Kingdom
##Structure	of Data and	NA's					

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    541909 obs. of  8 variables:
##  $ InvoiceNo  : chr  "536365" "536365" "536365" "536365" ...
##  $ StockCode  : chr  "85123A" "71053" "84406B" "84029G" ...
##  $ Description: chr  "WHITE HANGING HEART T-LIGHT HOLDER" "WHITE METAL LANTERN" "CREAM CUPID HEARTS
COAT HANGER" "KNITTED UNION FLAG HOT WATER BOTTLE" ...
##  $ Quantity   : num  6 6 8 6 6 2 6 6 6 32 ...
##  $ InvoiceDate: POSIXct, format: "2010-12-01 08:26:00" "2010-12-01 08:26:00" ...
##  $ UnitPrice  : num  2.55 3.39 2.75 3.39 3.39 7.65 4.25 1.85 1.85 1.69 ...
##  $ CustomerID : num  17850 17850 17850 17850 17850 ...
##  $ Country    : chr  "United Kingdom" "United Kingdom" "United Kingdom" "United Kingdom" ...
##   InvoiceNo   StockCode Description    Quantity InvoiceDate   UnitPrice
##   0.0000000   0.0000000   0.2683107   0.0000000   0.0000000   0.0000000
##  CustomerID     Country
##  24.9266943   0.0000000
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

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

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

Date and Time

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Sales	InvoiceTime
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010/12/01	2.55	17850	United Kingdom	15.30	08:26:00
536365	71053	WHITE METAL LANTERN	6	2010/12/01	3.39	17850	United Kingdom	20.34	08:26:00
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010/12/01	2.75	17850	United Kingdom	22.00	08:26:00
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010/12/01	3.39	17850	United Kingdom	20.34	08:26:00
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010/12/01	3.39	17850	United Kingdom	20.34	08:26:00
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010/12/01	7.65	17850	United Kingdom	15.30	08:26:00

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Sales	InvoiceTime
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010/12/01	4.25	17850	United Kingdom	25.50	08:26:00
536366	22633	HAND WARMER UNION JACK	6	2010/12/01	1.85	17850	United Kingdom	11.10	08:28:00
536366	22632	HAND WARMER RED POLKA DOT	6	2010/12/01	1.85	17850	United Kingdom	11.10	08:28:00
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010/12/01	1.69	13047	United Kingdom	54.08	08:34:00

Descriptive Analysis

1. Summary

```
##   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            Sales           
##  Min.   :12346    United Kingdom:495478   Min.   :-168469.60  
##  1st Qu.:13953    Germany       :  9495   1st Qu.:      3.40  
##  Median :15152    France        :  8557   Median :      9.75  
##  Mean   :15288    EIRE          :  8196   Mean   :     17.99  
##  3rd Qu.:16791    Spain         :  2533   3rd Qu.:     17.40  
##  Max.   :18287    Netherlands   :  2371   Max.   : 168469.60  
##  NA's   :135080   (Other)       : 15279                       
##  InvoiceTime       
##  Length:541909     
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
## 
```

```
## [1] 378.8108
```

OUTLIERS TREATMENT ALONG WITH SKEWNESS & KURTOSIS

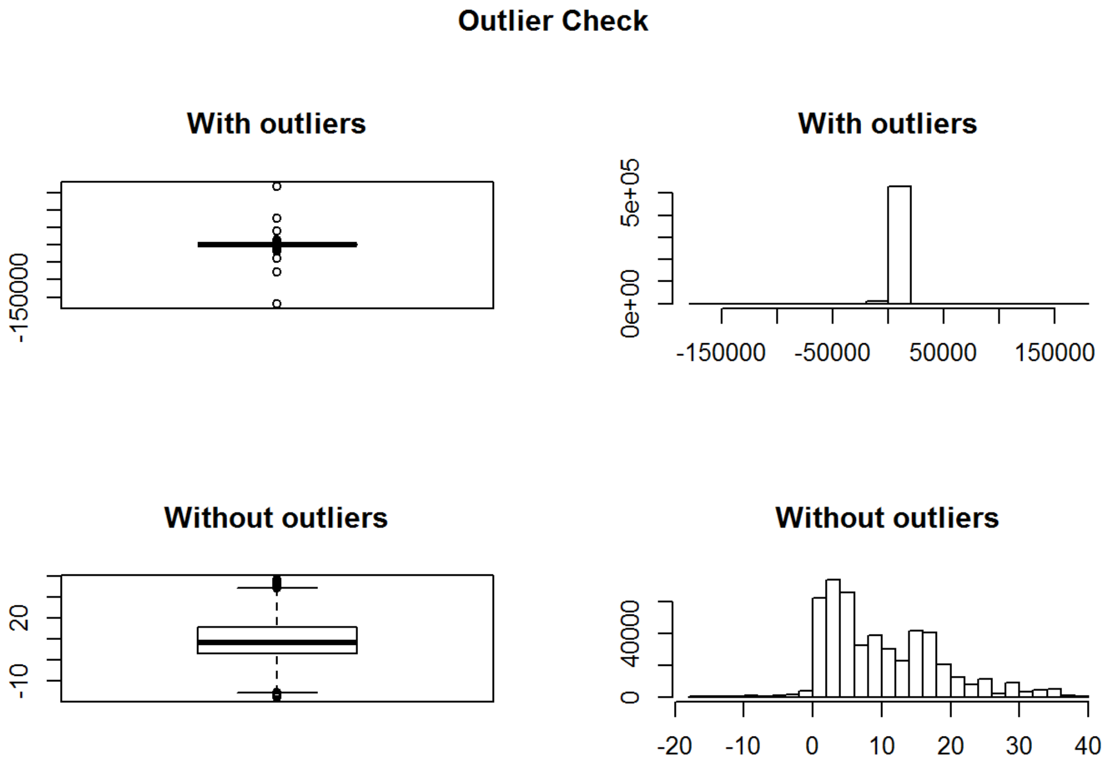
- In statistics, an outlier is defined as an observation which stands far away from the most of other observations. Often an outlier is present due to the measurements error. Therefore, one of the most important tasks in data analysis is to identify and only if it is necessary to remove the outlier.
- Skewness is the measurement of how the data is distributed. To check the symmetry of the data distribution
- Intuitively, the kurtosis describes the tail shape of the data distribution. The normal distribution has zero kurtosis and thus the standard tail shape. It is said to be mesokurtic. Negative kurtosis would indicate a thin-tailed data distribution, and is said to be platykurtic. Positive kurtosis would indicate a fat-tailed distribution, and is said to be leptokurtic.

Function to identify Outliers

```
##
## Attaching package: 'e1071'
```

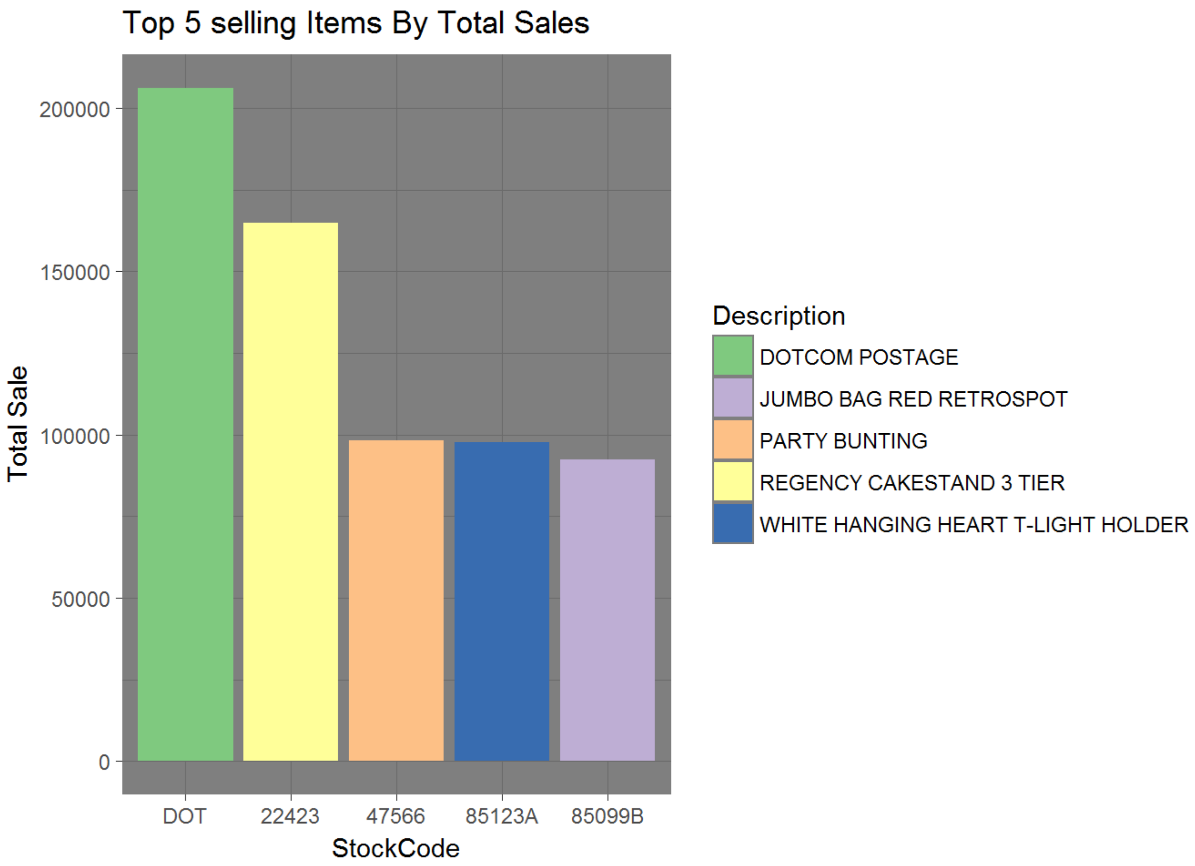
```
## The following objects are masked from 'package:moments':
##
##     kurtosis, moment, skewness
```

```
outlierfn(retail1,Sales) #Calling the function
```

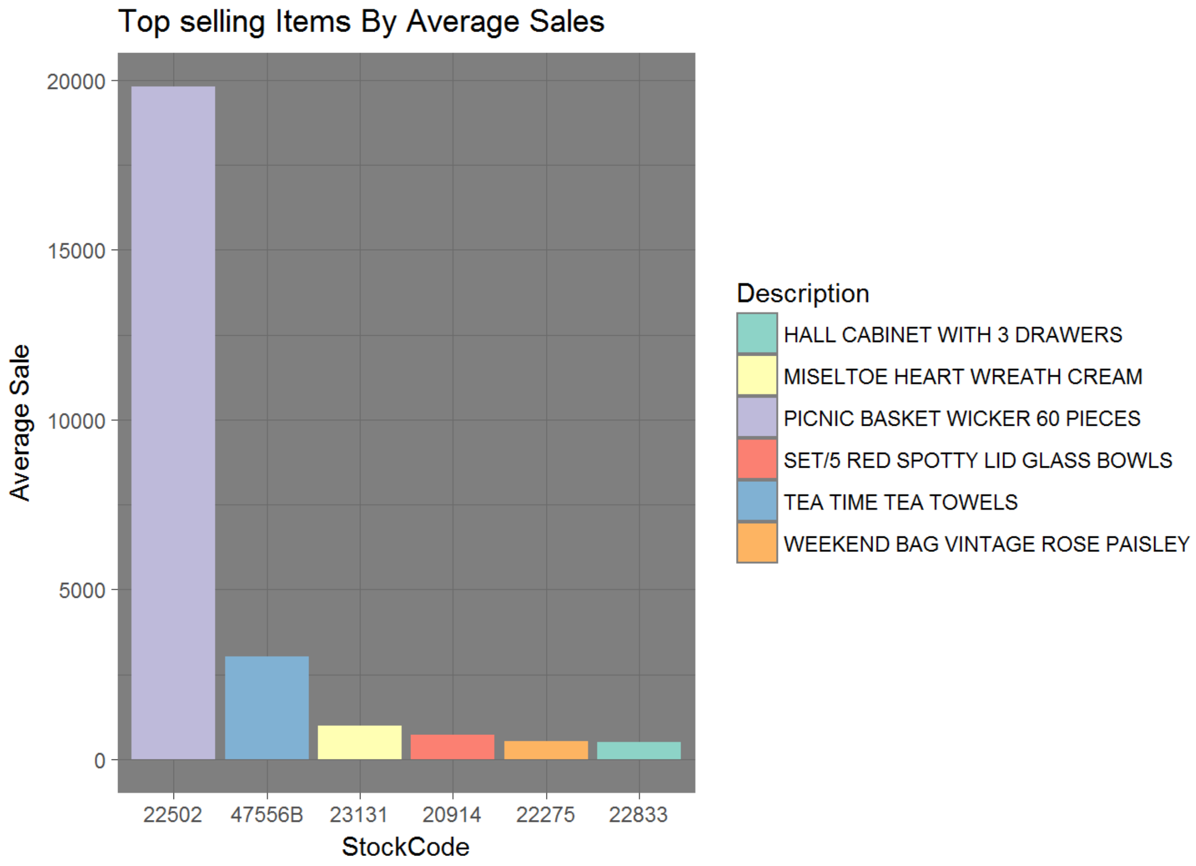


1.Top 5 selling items by overall sales

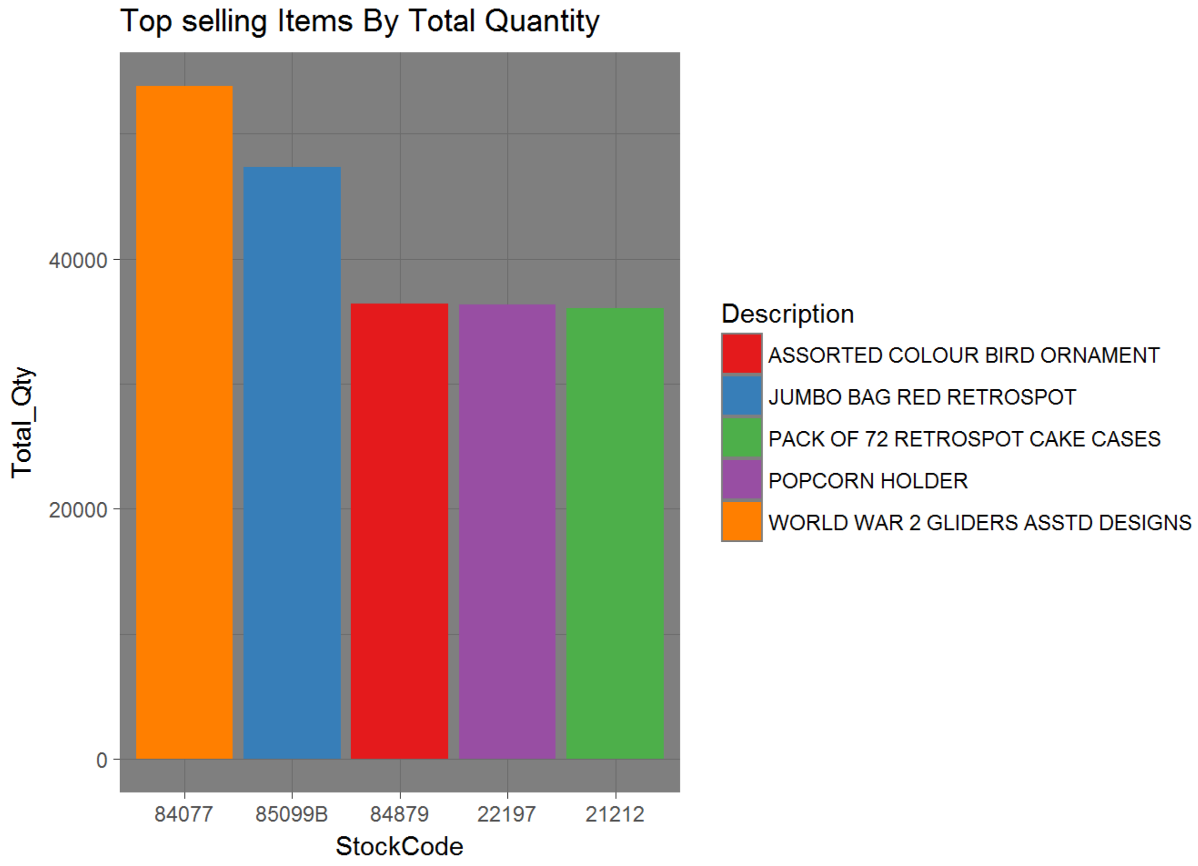
- Sales column created above using Quantity and Unit Price columns



2.Average Sales



3.Top 5 selling items by quantities sold

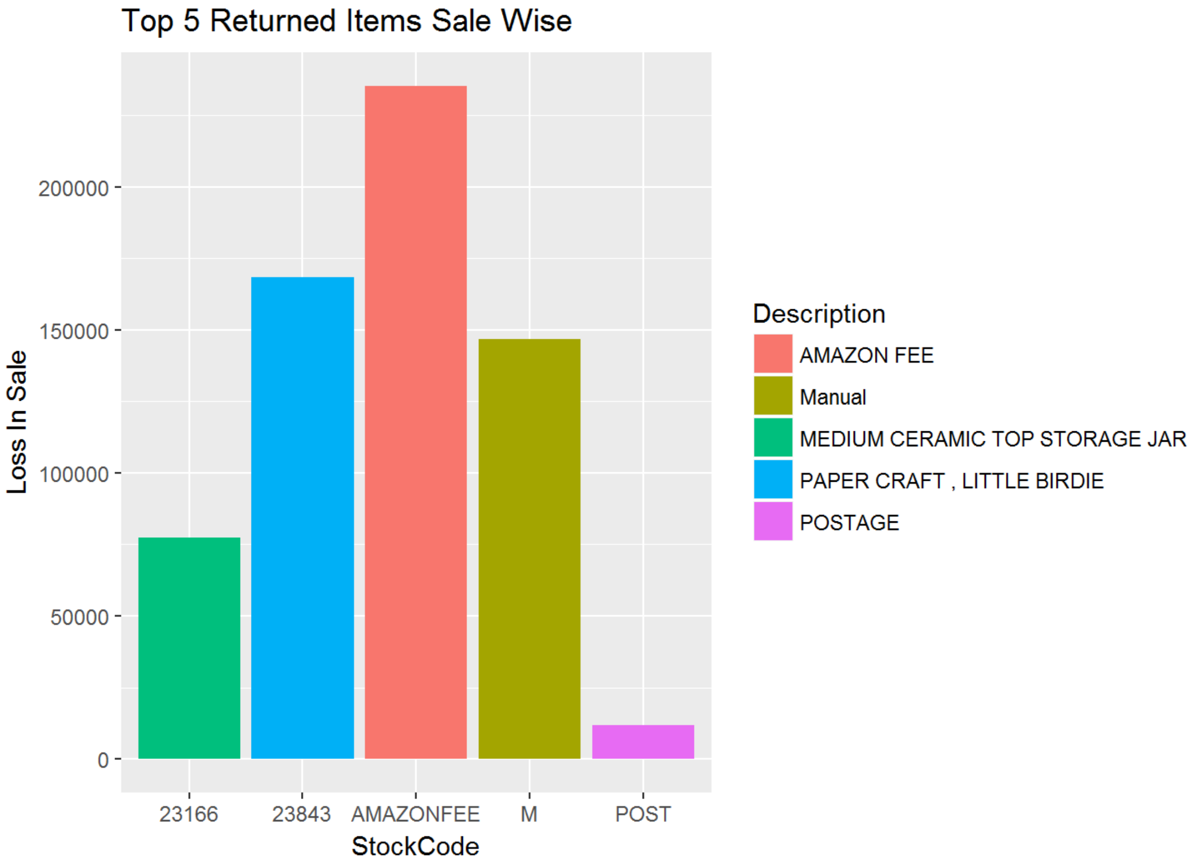


Top selling Items By Total Quantity

subset for cancelled products

- Here we filtered those products where the Invoice number started with ‘C’(c - denotes cancelled items)

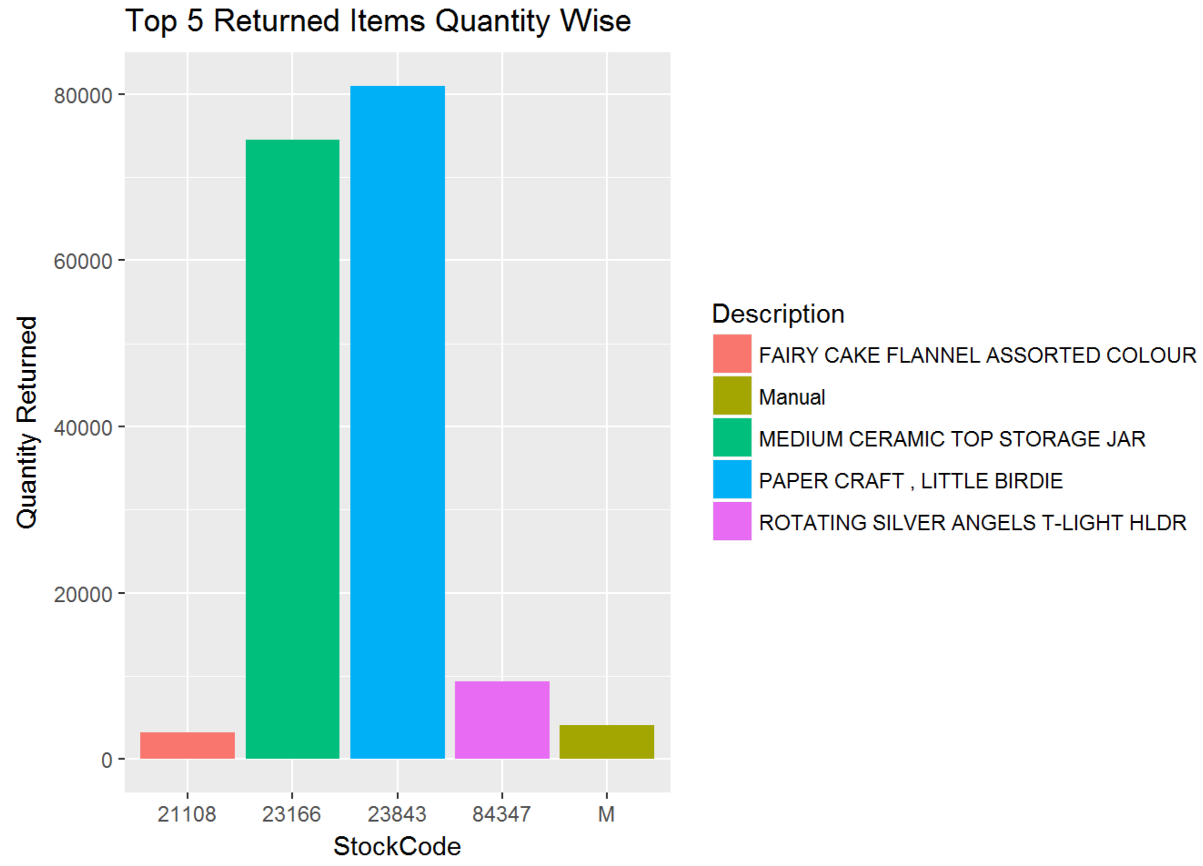
```
## # A tibble: 5 x 3
## # Groups:   StockCode [5]
##   StockCode                  Description totalsales
##       <chr>                        <chr>      <dbl>
## 1 AMAZONFEE                   AMAZON FEE -235281.59
## 2     23843    PAPER CRAFT , LITTLE BIRDIE -168469.60
## 3         M                       Manual -146784.46
## 4     23166 MEDIUM CERAMIC TOP STORAGE JAR  -77479.64
## 5      POST                      POSTAGE  -11871.24
```



# products

returned quantity wise

```
## # A tibble: 5 x 3
## # Groups:   StockCode [5]
##   StockCode                         Description totalquantity
##       <chr>                               <chr>         <dbl>
## 1     23843         PAPER CRAFT , LITTLE BIRDIE        -80995
## 2     23166      MEDIUM CERAMIC TOP STORAGE JAR        -74494
## 3     84347 ROTATING SILVER ANGELS T-LIGHT HLDR         -9376
## 4         M                              Manual         -4066
## 5     21108  FAIRY CAKE FLANNEL ASSORTED COLOUR         -3150
```



COUNTRY WISE SALES

```
country_sales = retail1 %>% group_by(Country) %>% summarise(totalsales=sum(Sales)) %>% arrange(-totalsa
les)
#country_sales
# country wise return
country_return = cancelled_products %>% group_by(Country) %>% summarise(totalreturn=sum(Sales)) %>% arr
ange(totalreturn)
#country_return
country_total = merge(country_sales,country_return)
#country_total

country_total$buisness = country_total$totalsales + country_total$totalreturn
country_total = country_total %>% arrange(-buisness)
library(reshape2)
```

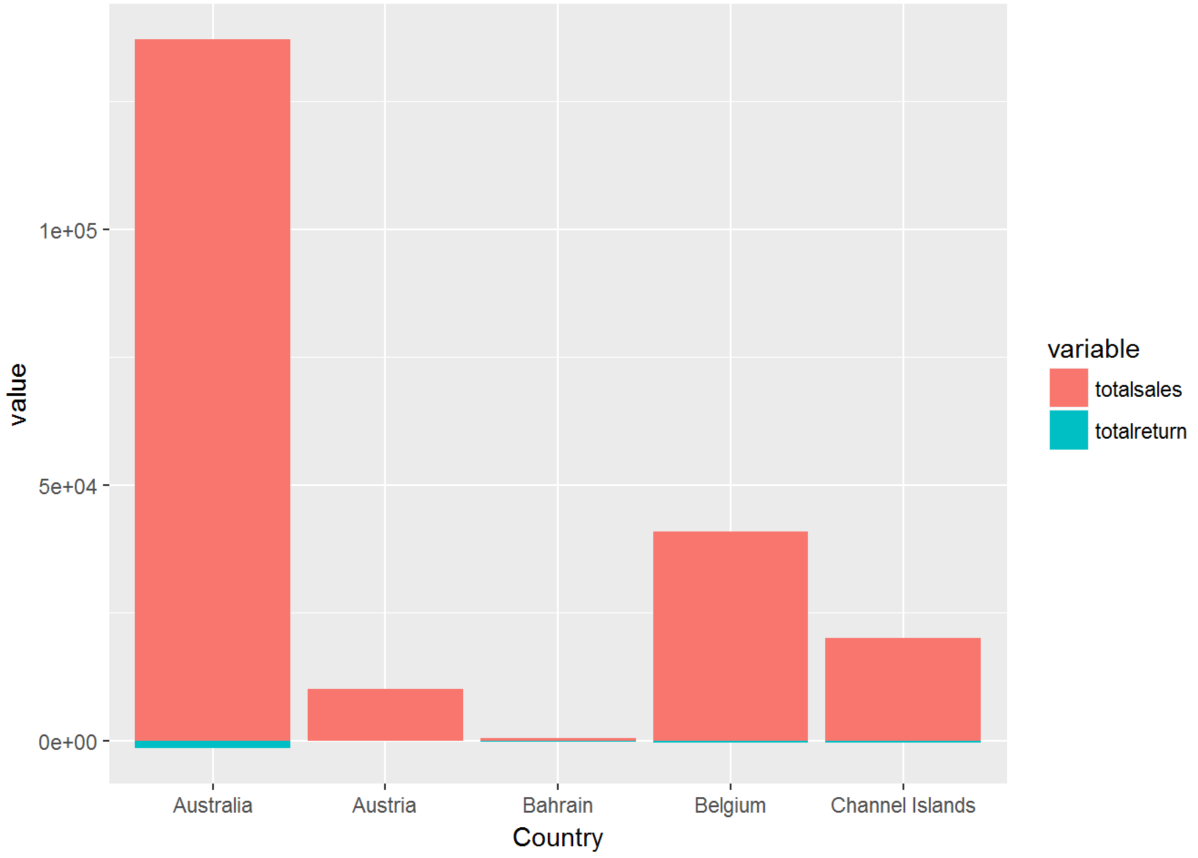
```
## Warning: package 'reshape2' was built under R version 3.4.3
```

```
knitr::kable(x = country_total)
```

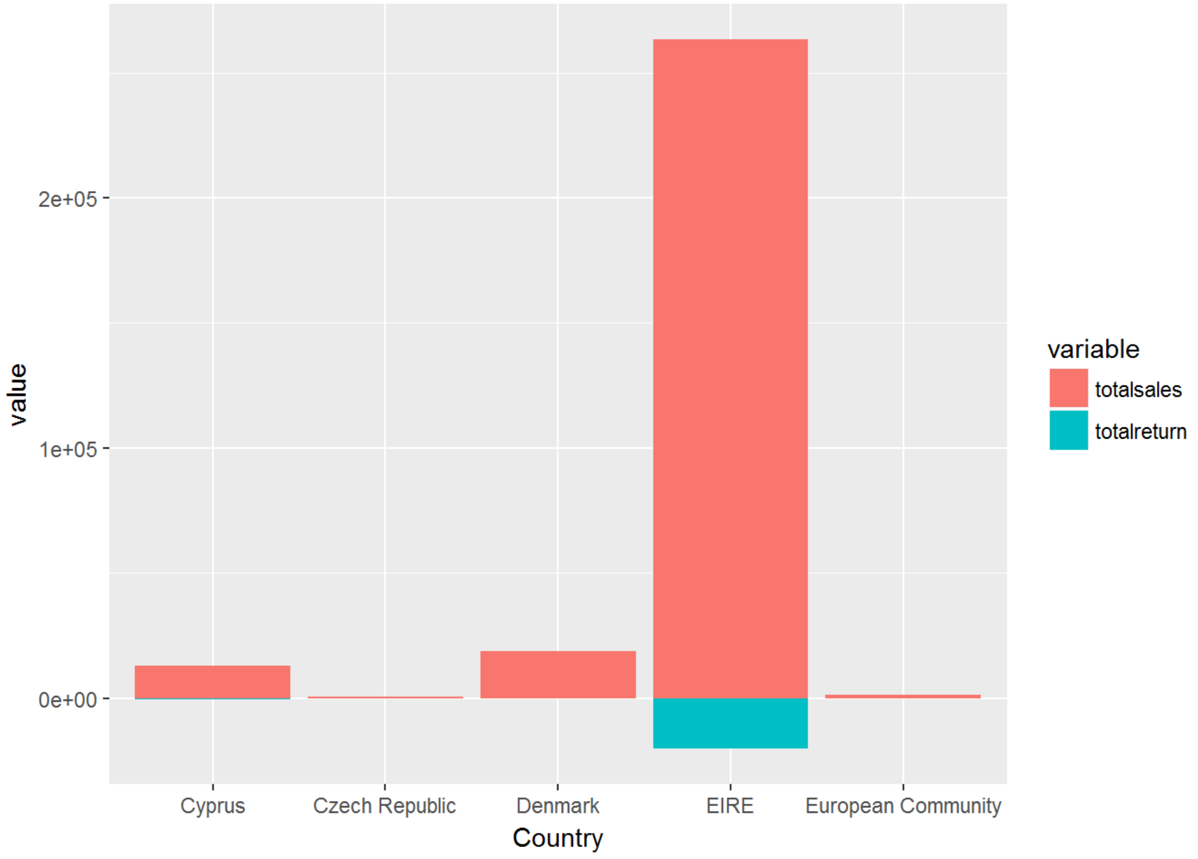
Country	totalsales	totalreturn	buisness
United Kingdom	8187806.36	-815291.60	7372514.76
Netherlands	284661.54	-784.80	283876.74
EIRE	263276.82	-20177.14	243099.68
Germany	221698.21	-7168.93	214529.28
France	197403.90	-12311.21	185092.69
Australia	137077.27	-1444.04	135633.23

Country	totalsales	totalreturn	buisness
Switzerland	56385.35	-704.55	55680.80
Spain	54774.58	-6802.53	47972.05
Belgium	40910.96	-285.38	40625.58
Sweden	36595.91	-1782.42	34813.49
Norway	35163.46	-1001.98	34161.48
Japan	35340.62	-2075.75	33264.87
Portugal	29367.02	-4380.08	24986.94
Finland	22326.74	-219.34	22107.40
Channel Islands	20086.29	-364.15	19722.14
Denmark	18768.14	-187.20	18580.94
Italy	16890.51	-592.73	16297.78
Cyprus	12946.29	-644.09	12302.20
Austria	10154.32	-44.36	10109.96
Israel	7907.82	-227.44	7680.38
Poland	7213.14	-121.51	7091.63
Greece	4710.52	-50.00	4660.52
Hong Kong	10117.04	-5574.76	4542.28
Malta	2505.47	-220.12	2285.35
European Community	1291.75	-8.50	1283.25
Czech Republic	707.72	-119.02	588.70
Bahrain	548.40	-205.74	342.66
Saudi Arabia	131.17	-14.75	116.42
USA	1730.92	-1849.47	-118.55
Singapore	9120.39	-12158.90	-3038.51

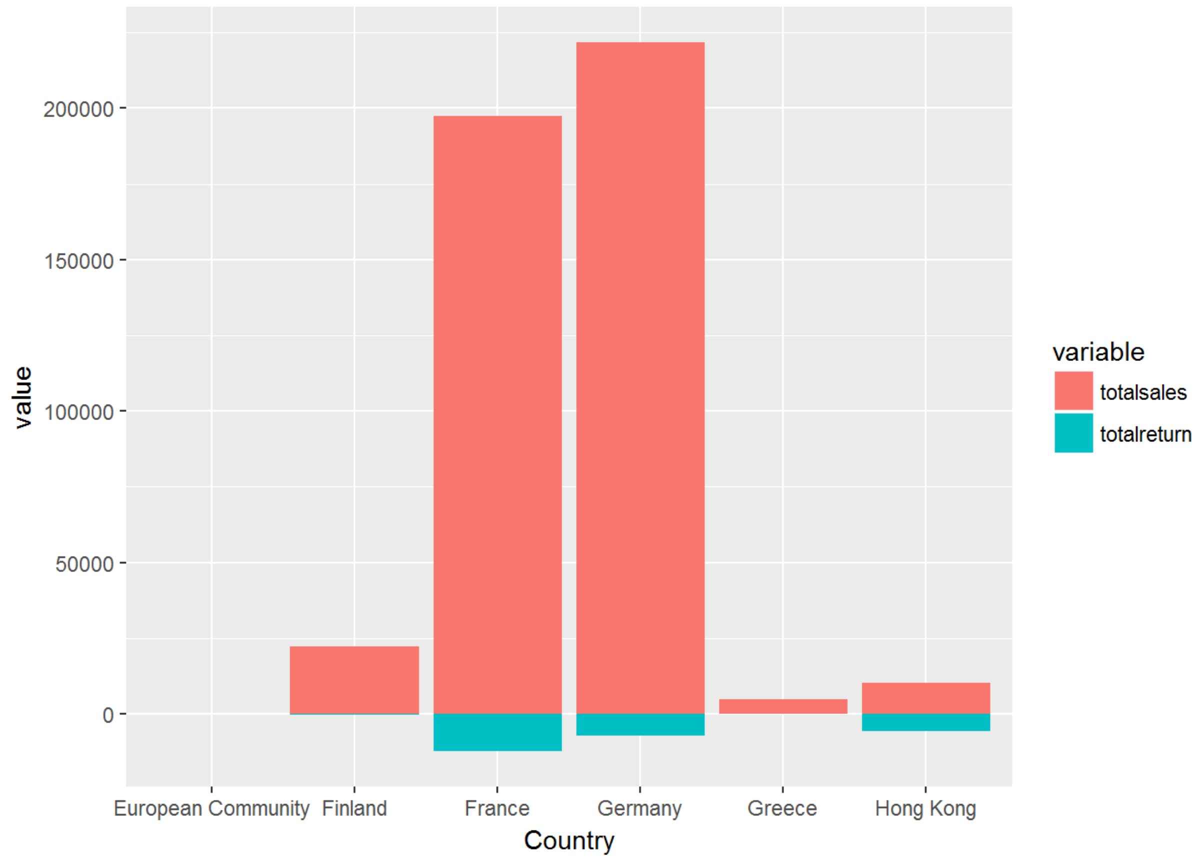
```
c1=reshape2::melt(country_total[,1:3],id="Country")%>%arrange(Country)


g1=ggplot(c1[1:10,],aes(x=Country,y=value))+geom_bar(stat="identity",aes(fill=variable))

g2=ggplot(c1[11:20,],aes(x=Country,y=value))+geom_bar(stat="identity",aes(fill=variable))

g3=ggplot(c1[20:30,],aes(x=Country,y=value))+geom_bar(stat="identity",aes(fill=variable))

g4=ggplot(c1[40:50,],aes(x=Country,y=value))+geom_bar(stat="identity",aes(fill=variable))

g5=ggplot(c1[50:60,],aes(x=Country,y=value))+geom_bar(stat="identity",aes(fill=variable))

g1
```


1e+05
5e+04
0e+00
value
Australia
Austria
Bahrain
Belgium
Channel Islands
Country
variable
totalsales
totalreturn

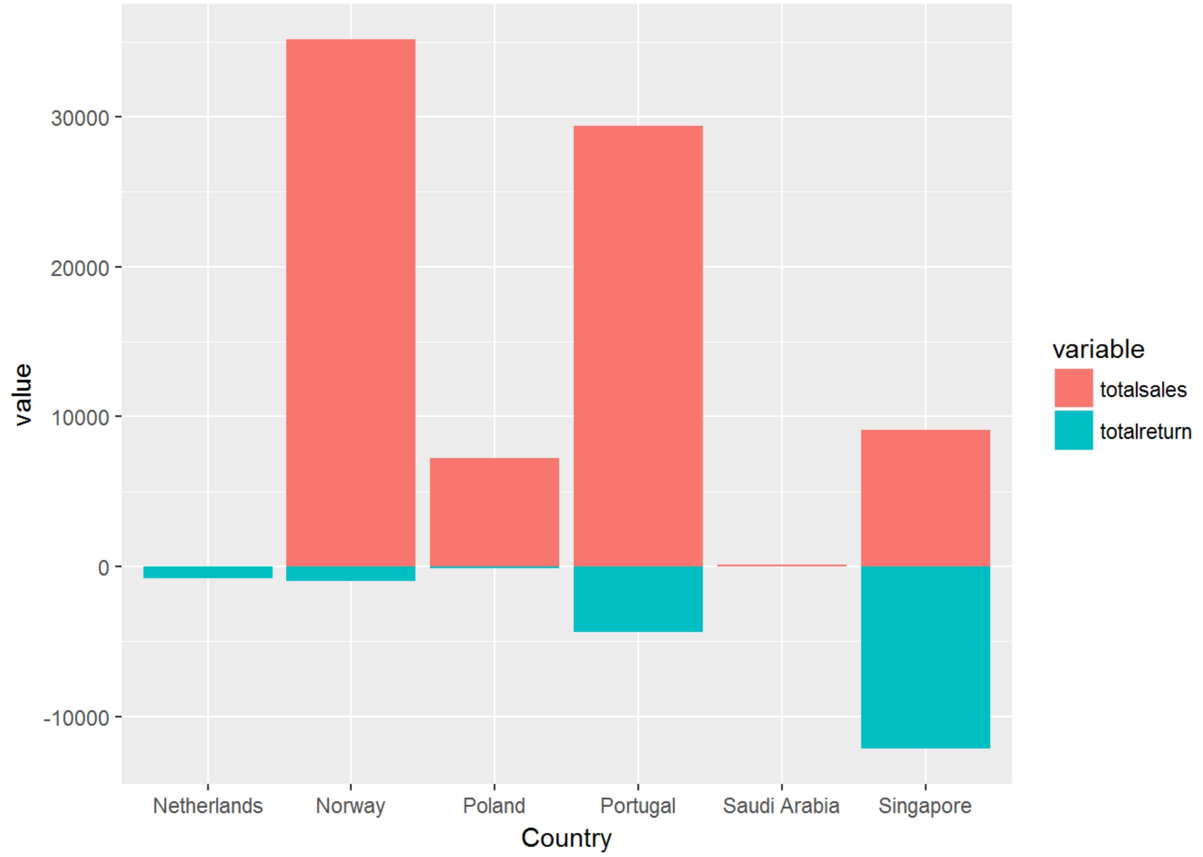
```
g2
```


2e+05
1e+05
0e+00
value
Cyprus
Czech Republic
Denmark
EIRE
European Community
Country
variable
totalsales
totalreturn

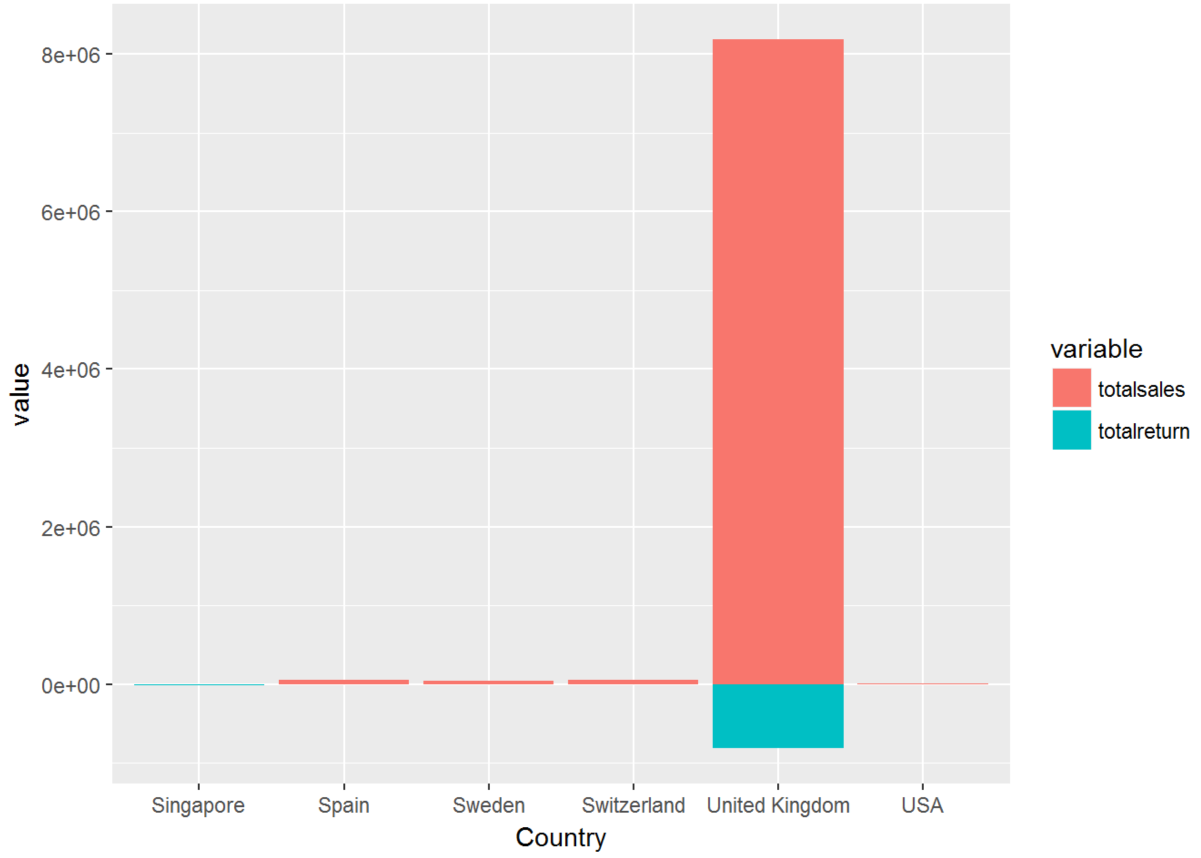
```
g3
```


value
200000
150000
100000
50000
0
European Community
Finland
France
Germany
Greece
Hong Kong
Country
variable
totalsales
totalreturn

```
g4
```


value
30000
20000
10000
0
-10000
Netherlands
Norway
Poland
Portugal
Saudi Arabia
Singapore
Country
variable
totalsales
totalreturn

```
g5
```



3.Maximum Sale Day Wise

```
## Loading required package: quantmod
```

```
## Warning: package 'quantmod' was built under R version 3.4.3
```

```
## Loading required package: xts
```

```
## Warning: package 'xts' was built under R version 3.4.3
```

```
## Loading required package: zoo
```

```
## 
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric
```

```
## 
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:dplyr':
## 
##     first, last
```

```
## Loading required package: TTR
## Warning: package 'TTR' was built under R version 3.4.3
```

```
## Version 0.4-0 included new data defaults. See ?getSymbols.
```

```
## Loading required package: scales
```

```
## 
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:plyr':
## 
##     here
```

```
## The following object is masked from 'package:base':
## 
##     date
```

```
#Converting InvoiceTime Column from Character to Time

library(chron)
```

```
## Warning: package 'chron' was built under R version 3.4.3
```

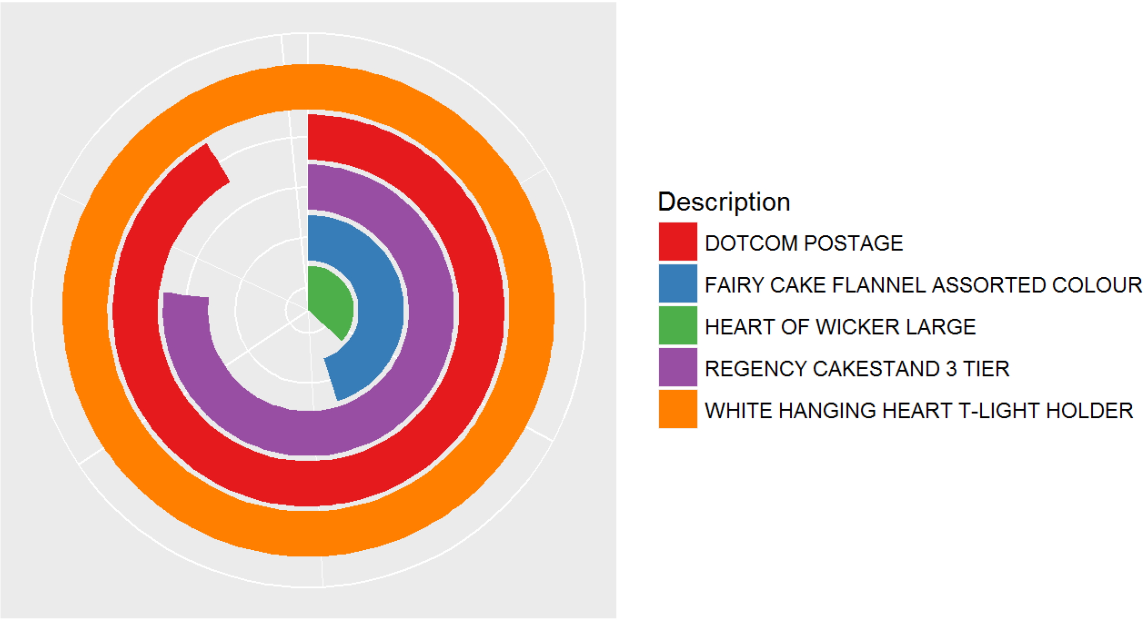
```
## 
## Attaching package: 'chron'
```

```
## The following objects are masked from 'package:lubridate':
## 
##     days, hours, minutes, seconds, years
```

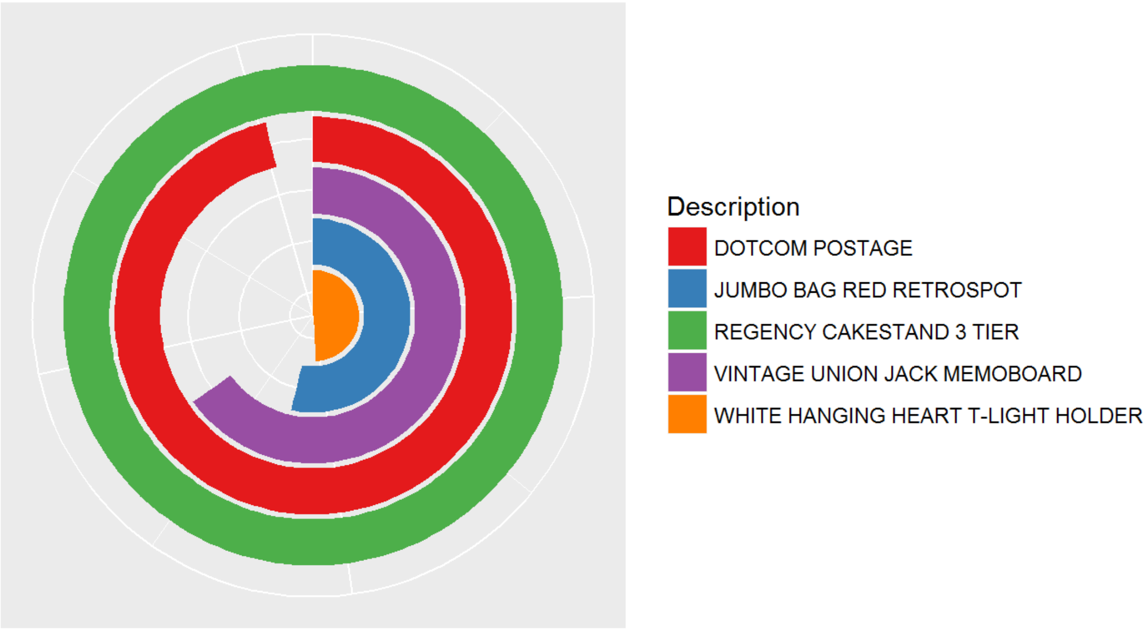
```
data$InvoiceTime=chron(times. = data$InvoiceTime)
```

Top Selling Item in a Month

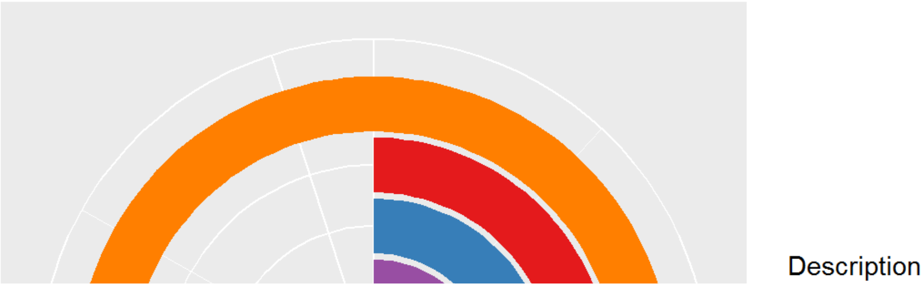
Top Selling Products in January

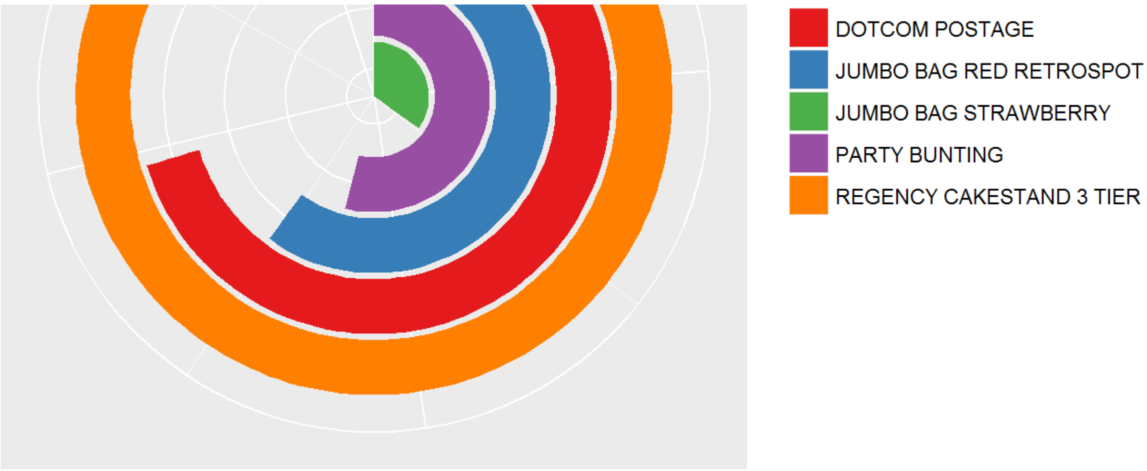


Top Selling Products in February

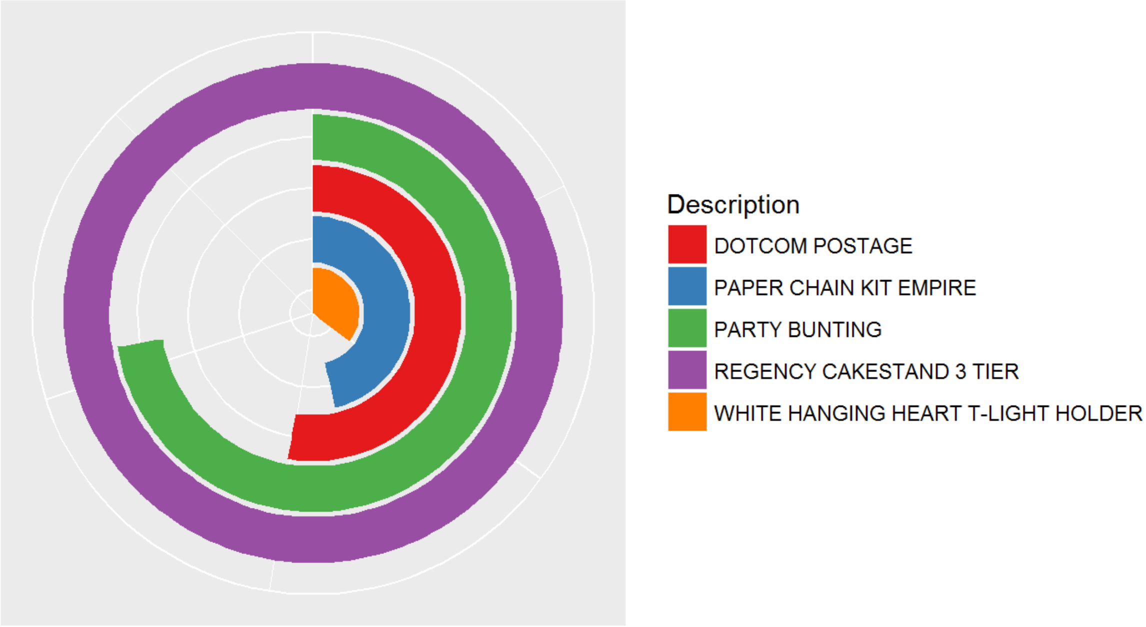


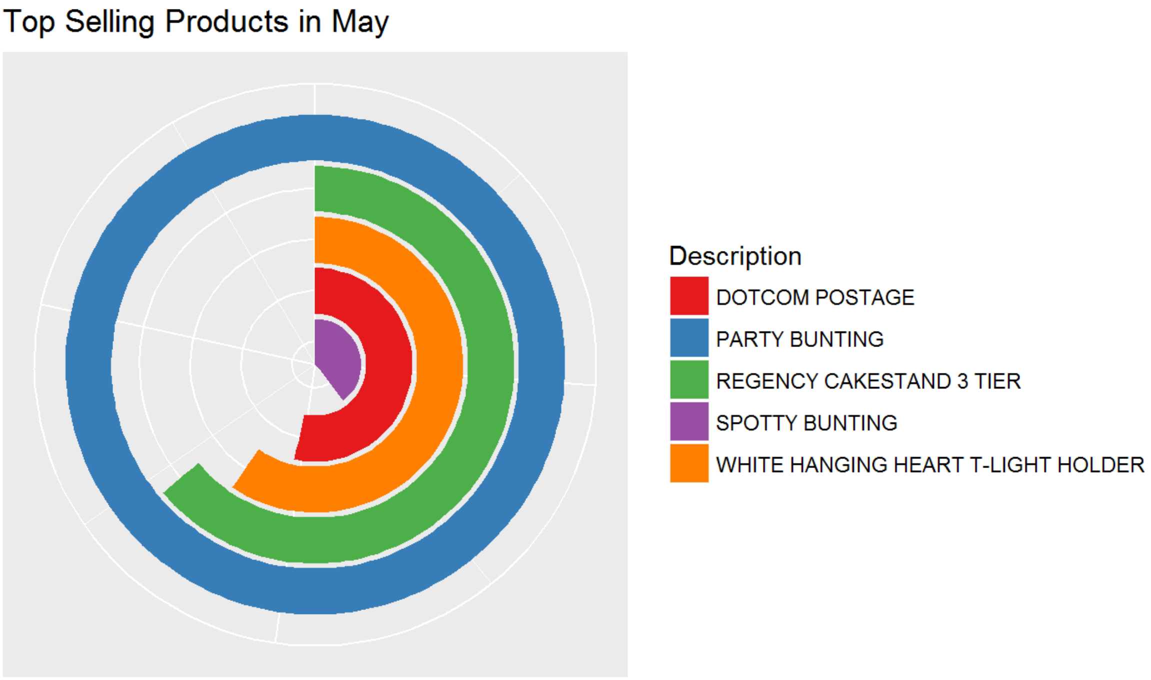
Top Selling Products in March

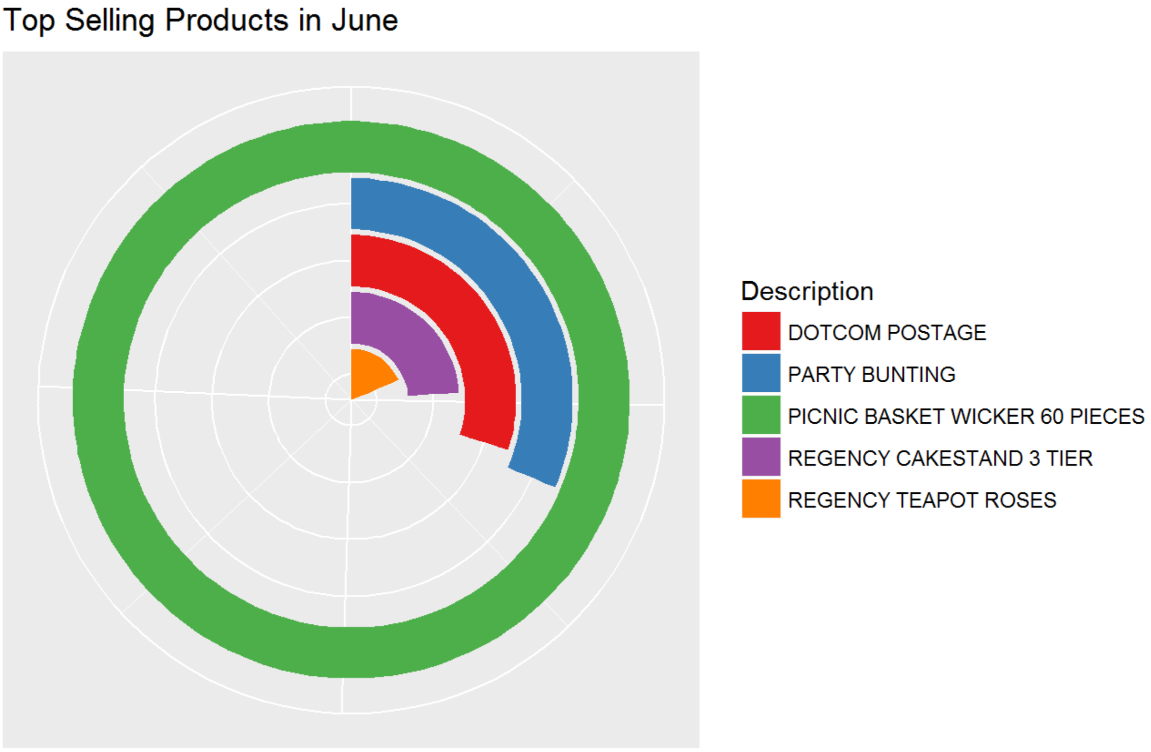




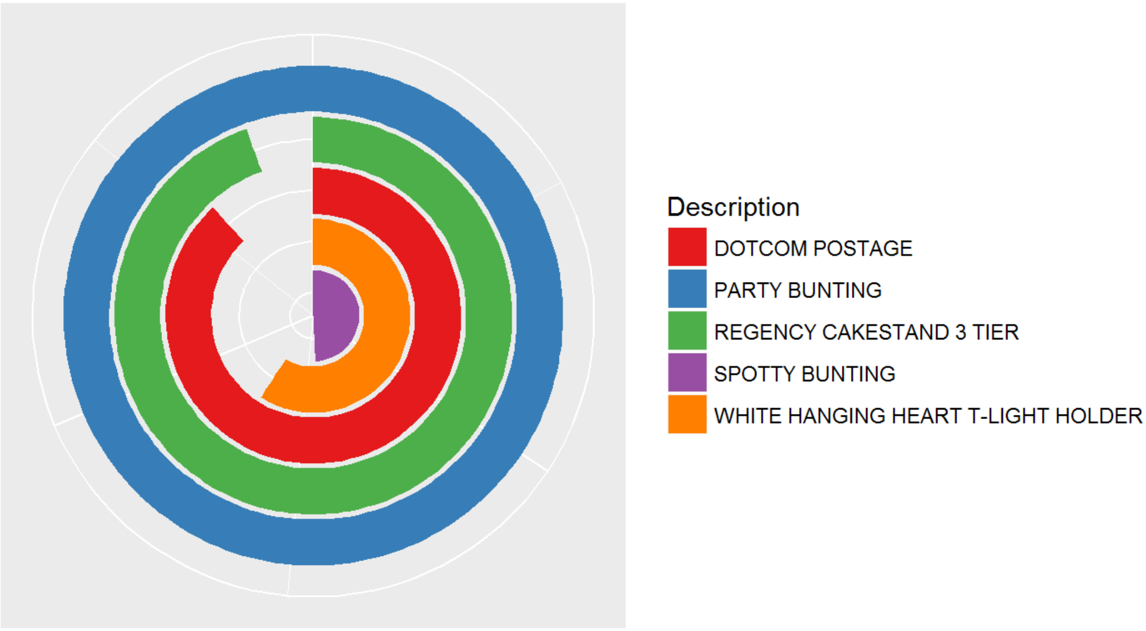
Top Selling Products in April



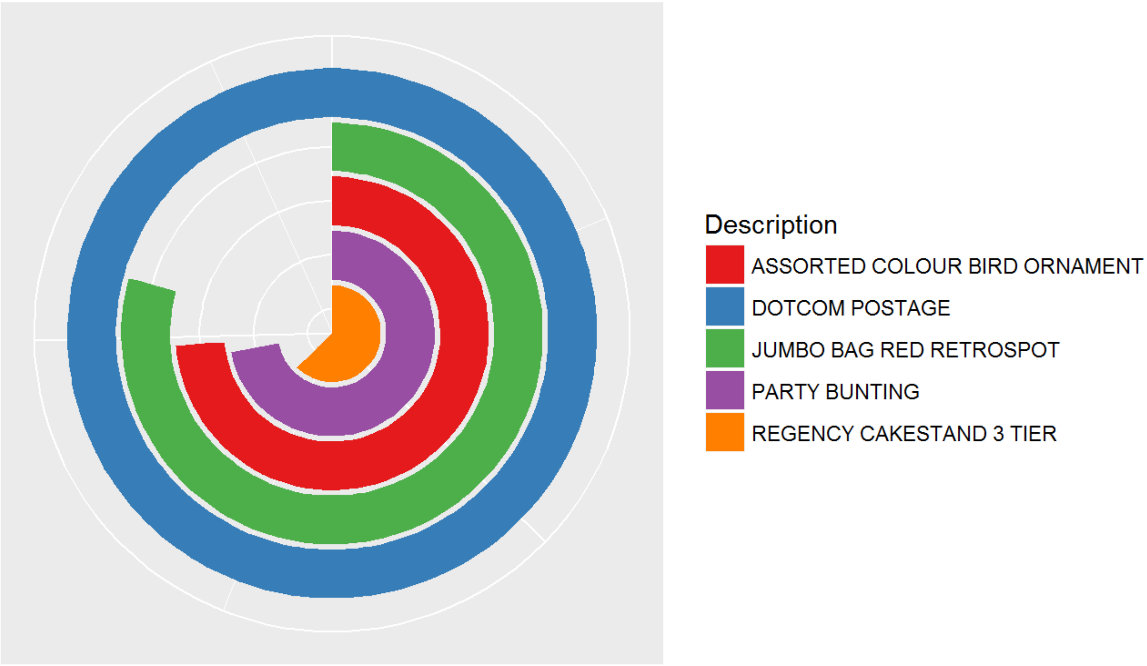

Top Selling Products in May
Description
DOTCOM POSTAGE
PARTY BUNTING
REGENCY CAKESTAND 3 TIER
SPOTTY BUNTING
WHITE HANGING HEART T-LIGHT HOLDER


Top Selling Products in June
Description
DOTCOM POSTAGE
PARTY BUNTING
PICNIC BASKET WICKER 60 PIECES
REGENCY CAKESTAND 3 TIER
REGENCY TEAPOT ROSES

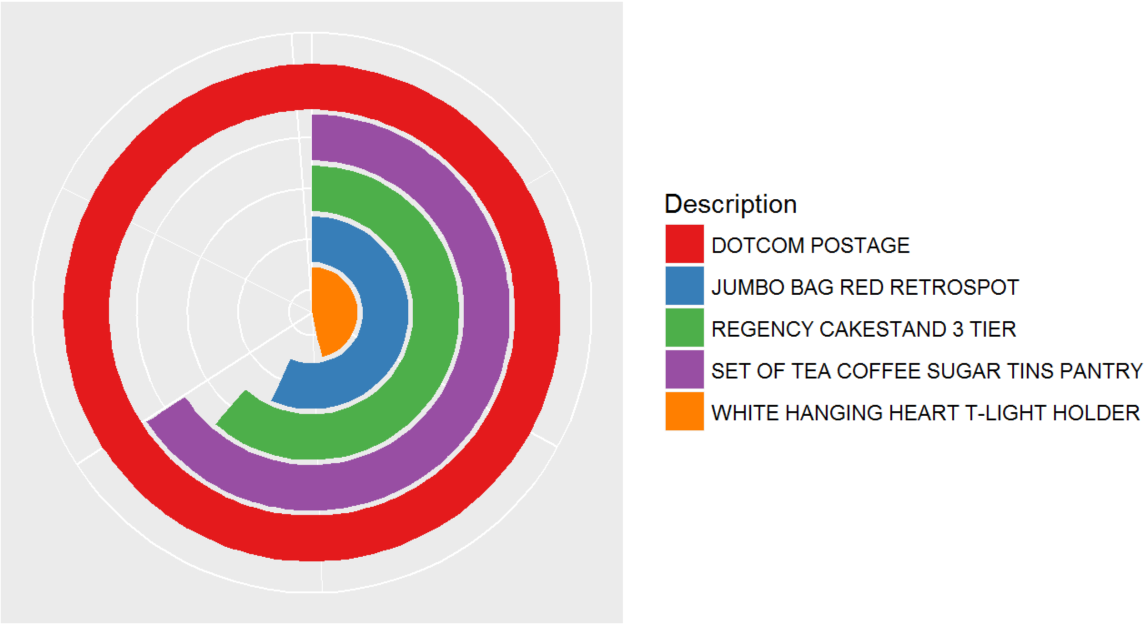
Top Selling Products in July



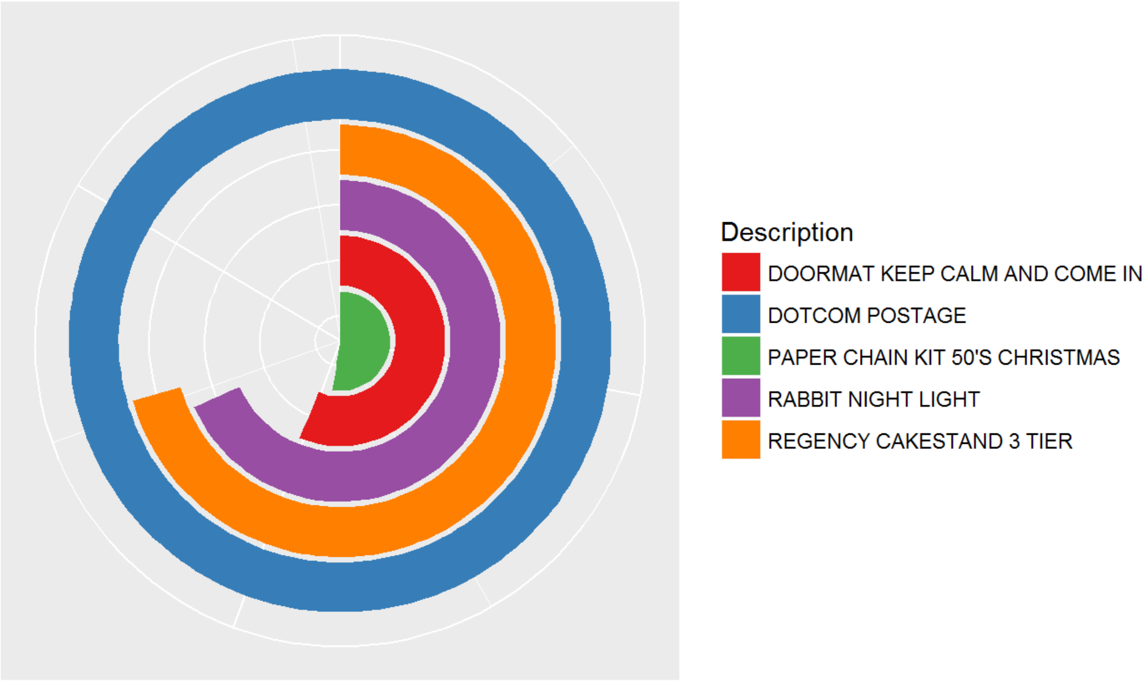
Top Selling Products in August



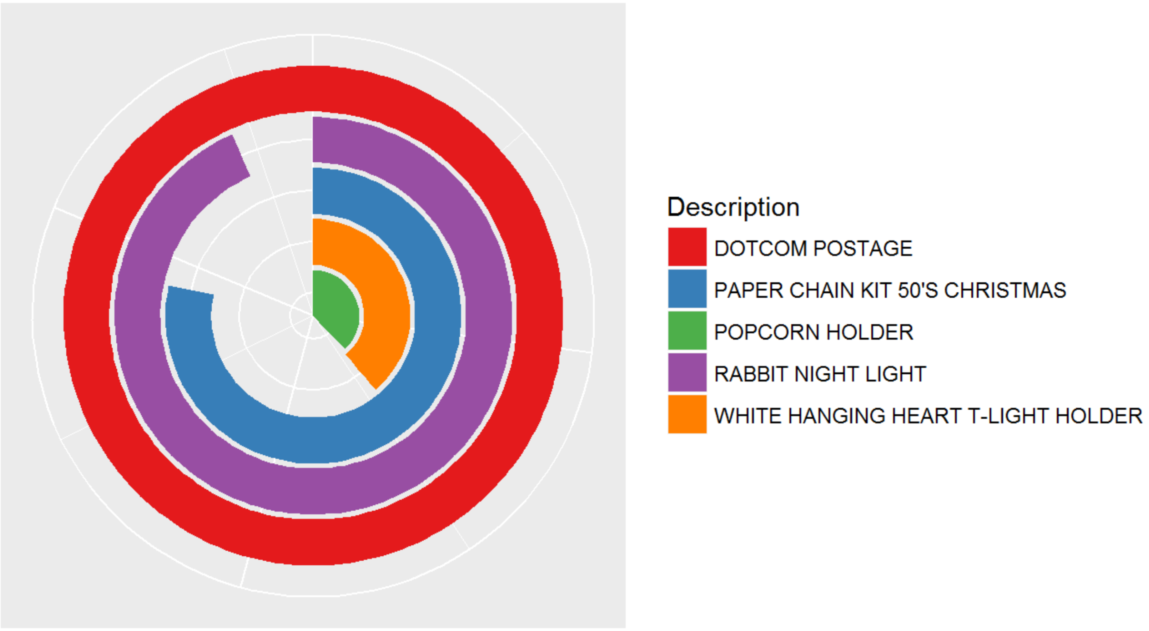
Top Selling Products in September



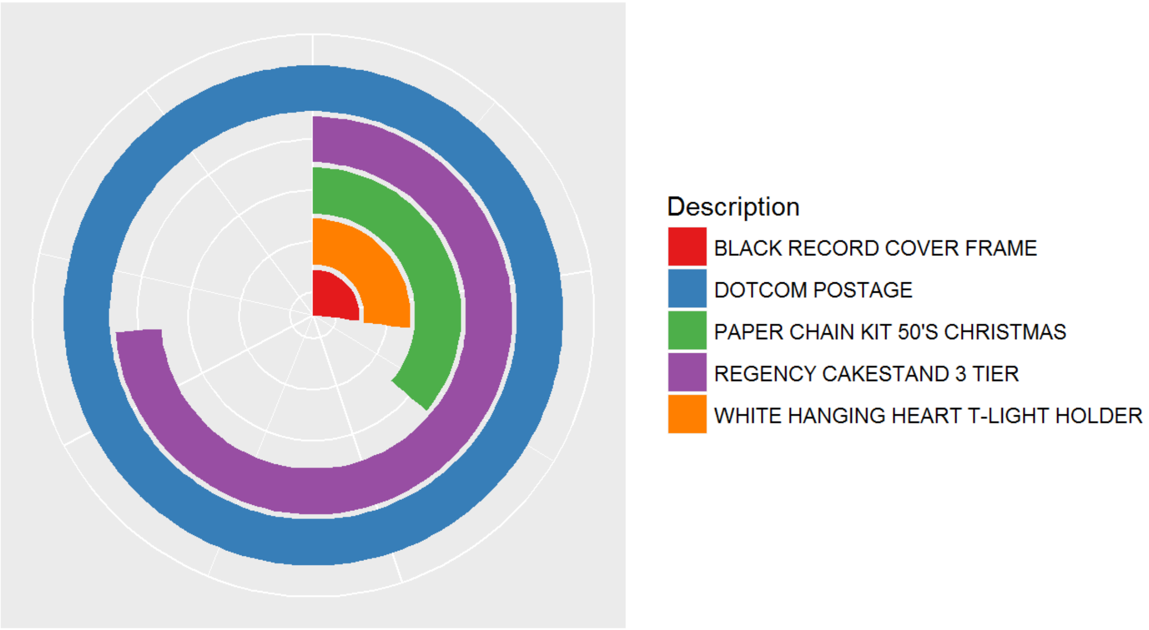
Top Selling Products in October



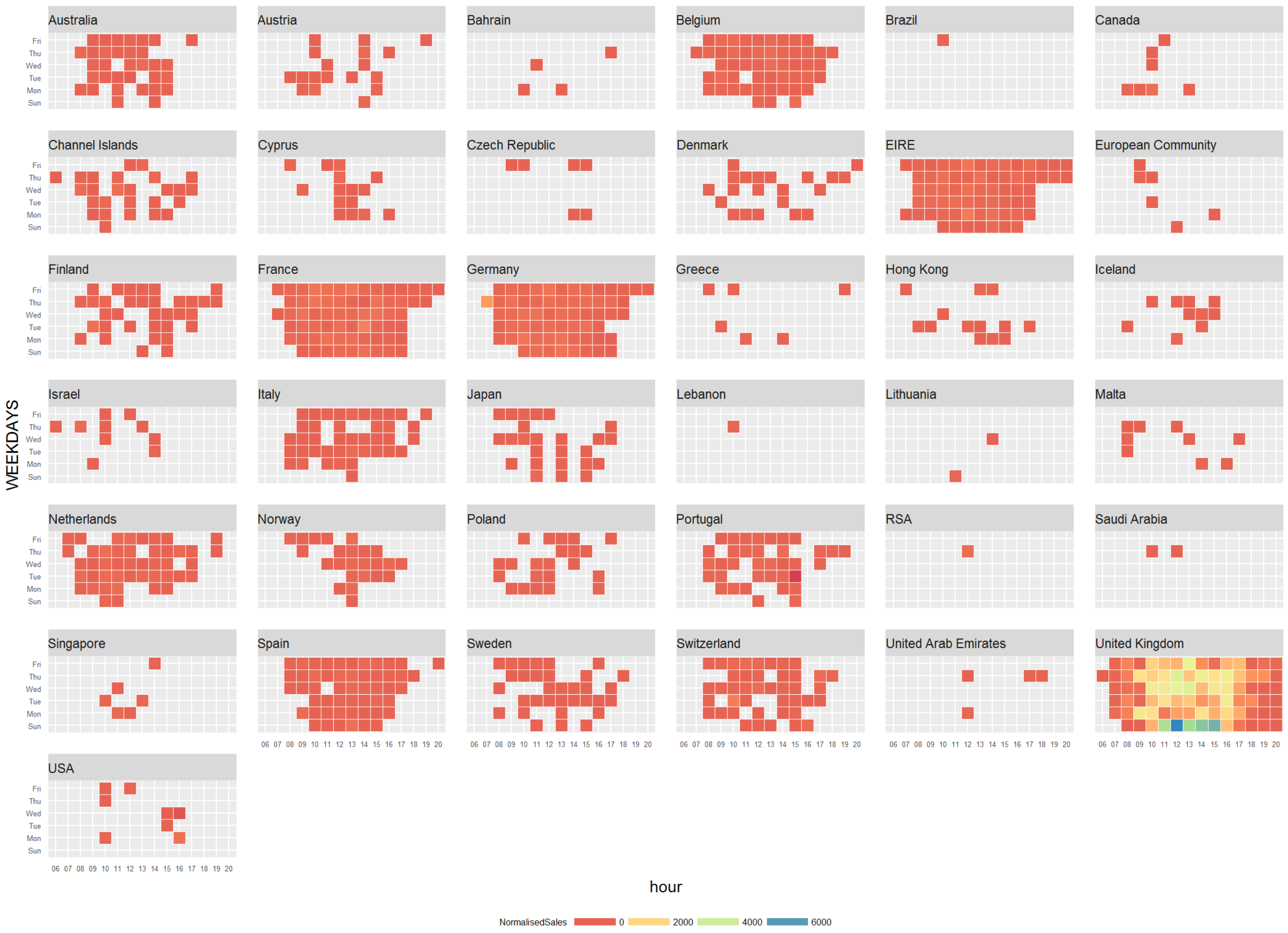
Top Selling Products in November



Top Selling Products in December

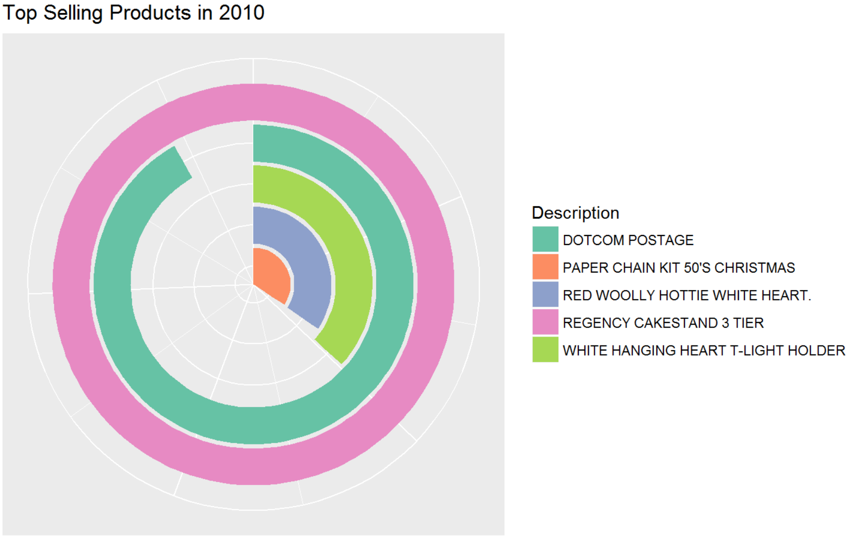


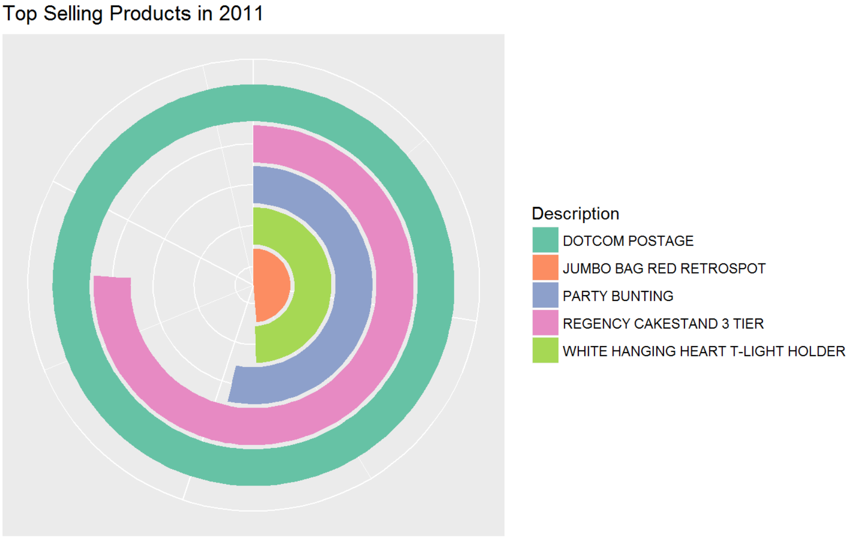
CountryWise Weekday wise Analysis



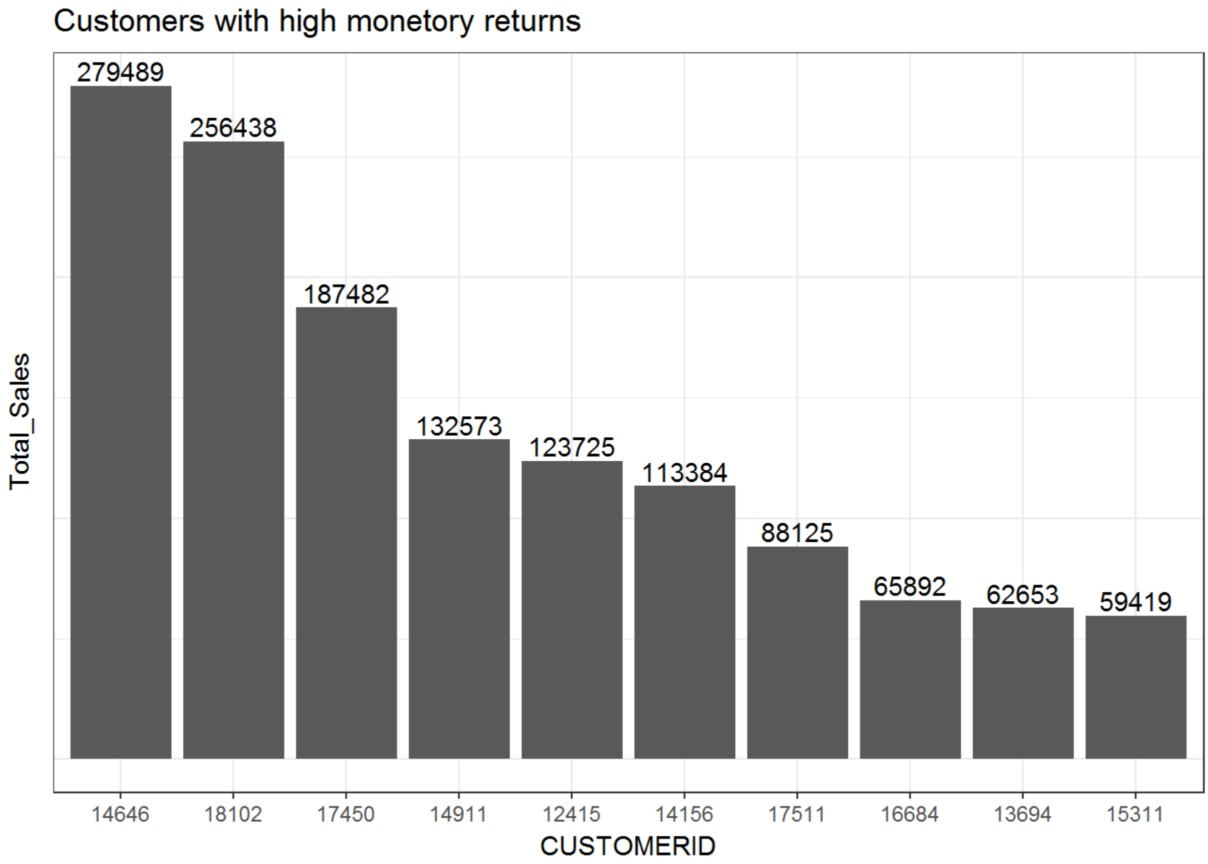
###Top revenue item in a year

```
q1=data%>%dplyr::group_by(InvoiceYear,Description)%>%dplyr::summarise(TotalSales=sum(Sales))%>%dplyr::a
rrange(InvoiceYear,-TotalSales)
q2= head(filter(q1,InvoiceYear==2010)%>%arrange(-TotalSales),5)%>%bind_rows(head(filter(q1,InvoiceYear=
=2011)%>%arrange(-TotalSales),5))

gg1=ggplot(q2[1:5,],aes(x=reorder(Description,TotalSales),y=TotalSales))+geom_bar(stat="identity",aes(f
ill=Description))+theme(axis.text.x = element_blank(),axis.ticks.x = element_blank(),axis.ticks.y = ele
ment_blank(),axis.text.y = element_blank())+coord_polar(theta = "y") +labs(title="Top Selling Products
 in 2010")+xlab("") + ylab("")+
    scale_fill_brewer(palette="Set2")


gg2=ggplot(q2[6:10,],aes(x=reorder(Description,TotalSales),y=TotalSales))+geom_bar(stat="identity",aes
(fill=Description))+theme(axis.text.x = element_blank(),axis.ticks.x = element_blank(),axis.ticks.y = e
lement_blank(),axis.text.y = element_blank())+coord_polar(theta = "y") +labs(title="Top Selling Product
s in 2011")+xlab("") + ylab("")+
    scale_fill_brewer(palette="Set2")

gridExtra::grid.arrange(gg1,gg2,nrow=2)
```





Total Sales Customer wise(Customers with high monetry value)


Customers with high monetory returns
Total_Sales
279489
256438
187482
132573
123725
113384
88125
65892
62653
59419
14646
18102
17450
14911
12415
14156
17511
16684
13694
15311
CUSTOMERID