Online Retail Kaggle

Tri Juhari

10/30/2019

Analisis Deskriptif

Data Retail Online didapatkan dari website kaggle dengan nama Retail Online atau dapat diunduh di https://www.kaggle.com/sanjeet41/online-retail. Analisis akan dilakukan secara mendetail bagi perusahaan yang mempunyai data online retail tersebut untuk memajukan perusahan tersebut dengan melakukan kebijakan berdasarkan rekomendasi dari hasil analisis yang dilakukan. Sebelum melakukan analisis, run package- package yang dibutuhkan seperti ggplot2 untuk proses visualisasi data, readr untuk proses import data, dplyr untuk manipulasi data.

```
data <- read.csv("C:/Users/tijeee/Downloads/DATASCIENCE/CHALLENGE/KAGGLE/online_retail.csv")
costumer <- data[!(data$StockCode==" "),]
```

Checking Dimension and Variable Data

```
dim(costumer)
```

```
## [1] 240007      8
```

```
summary(costumer)
```

```
##   ï..InvoiceNo      StockCode
##  537434 :   675   85123A :  1294
##  538071 :   652   22423  :  1254
##  538349 :   620   85099B :  1023
##  537638 :   601   47566  :   985
##  537237 :   597   20725  :   808
##  536876 :   593   21212  :   754
##  (Other):236269   (Other):233889
##                              Description        Quantity
##  WHITE HANGING HEART T-LIGHT HOLDER:  1319   Min.   :-74215.00
##  REGENCY CAKESTAND 3 TIER          :  1251   1st Qu.:     1.00
##  JUMBO BAG RED RETROSPOT           :  1023   Median :     3.00
##  PARTY BUNTING                     :   985   Mean   :     9.28
##                                    :   901   3rd Qu.:    10.00
##  LUNCH BAG RED RETROSPOT           :   807   Max.   : 74215.00
##  (Other)                           :233721
##          InvoiceDate       UnitPrice          CustomerID
##  12/6/10 16:57 :   675   Min.   :    0.00   Min.   :12346
##  12/9/10 14:09 :   652   1st Qu.:    1.25   1st Qu.:13842
##  12/10/10 14:59:   621   Median :    2.10   Median :15132
##  12/7/10 15:28 :   601   Mean   :    5.12   Mean   :15275
##  12/6/10 9:58  :   597   3rd Qu.:    4.21   3rd Qu.:16814
##  12/3/10 11:36 :   593   Max.   :38970.00   Max.   :18287
##  (Other)       :236268                      NA's   :67225
##           Country
##  United Kingdom:220279
##  Germany       :  4208
##  France        :  3642
```

```
##  EIRE          :  3034
##  Netherlands   :  1142
##  Spain         :  1142
##  (Other)       :  6560
```

```
str(costumer)
```

```
## 'data.frame':    240007 obs. of  8 variables:
##  $ ï..InvoiceNo: Factor w/ 12468 levels "536365","536366",..: 1 1 1 1 1 1 1 2 2 3 ...
##  $ StockCode   : Factor w/ 3645 levels "10002","10080",..: 3148 2447 2672 2620 2619 1652 793 1537 153
##  $ Description : Factor w/ 3618 levels "","? 4 PURPLE FLOCK DINNER CANDLES",..: 3455 3463 841 1729 263
##  $ Quantity    : int  6 6 8 6 6 2 6 6 6 32 ...
##  $ InvoiceDate : Factor w/ 11240 levels "1/10/11 10:04",..: 1452 1452 1452 1452 1452 1452 1452 1453 1
##  $ UnitPrice   : num  2.55 3.39 2.75 3.39 3.39 7.65 4.25 1.85 1.85 1.69 ...
##  $ CustomerID  : int  17850 17850 17850 17850 17850 17850 17850 17850 17850 13047 ...
##  $ Country     : Factor w/ 38 levels "Australia","Austria",..: 36 36 36 36 36 36 36 36 36 36 ...
```

Berdasarkan pengecekan dimensi dan variabel data dalam dataset ini terdapat 240007 baris dan 8 kolom, Variabel variabel tersebut dapat dilihat seperti diatas. InvoiceNo sebagai kode pembelian costumer dengan tipe data faktor. StockCode sebagai kode barang dengan tipe data Faktor, Description sebagai deskripsi barang atau nama barangnya dengan tipe data factor, Quantity sebagai jumlah dari barang yang dibeli oleh costumer dengan tipe data integer. InvoiceDate sebagai waktu atau tanggal pembelian dengan tipe data factor.UnitPrice sebagai harga produk perunitnya dengan tipe data numbner. CustomerID sebagai kode unik dari si customer atau bisa dibilang id si pembeli yang bersifat unik.Country adalah negara asal costumer dengan tipe data factor.

Data Cleansing

```
sapply(costumer, function(x) sum(is.na(x)))
```

```
## ï..InvoiceNo    StockCode  Description     Quantity  InvoiceDate
##            0            0            0            0            0
##    UnitPrice   CustomerID      Country
##            0        67225            0
```

```
sum(is.na(costumer$CustomerID))
```

```
## [1] 67225
```

```
costumer = dplyr::filter(costumer, !is.na(CustomerID))
```

Frekunsi based Costumer Country

```
table(costumer$Country)
```

```
##
##           Australia             Austria             Bahrain
##                 642                 127                  17
##             Belgium              Brazil              Canada
##                 933                  32                  68
##      Channel Islands              Cyprus      Czech Republic
##                 368                 353                  17
##             Denmark                EIRE  European Community
##                 184                2718                  32
##             Finland              France             Germany
##                 312                3625                4208
```

```
##               Greece            Hong Kong              Iceland
##                   85                    0                  102
##               Israel                Italy              Japan
##                   18                  309                  251
##              Lebanon            Lithuania              Malta
##                   45                   35                   47
##          Netherlands               Norway             Poland
##                 1142                  378                  187
##             Portugal         Saudi Arabia          Singapore
##                  624                   10                  118
##                Spain               Sweden        Switzerland
##                 1142                  200                  708
##                 Unit United Arab Emirates       United Kingdom
##                    1                   30               153620
##          Unspecified                  USA
##                   72                   22
```

```
sort(table(costumer$Country))
```

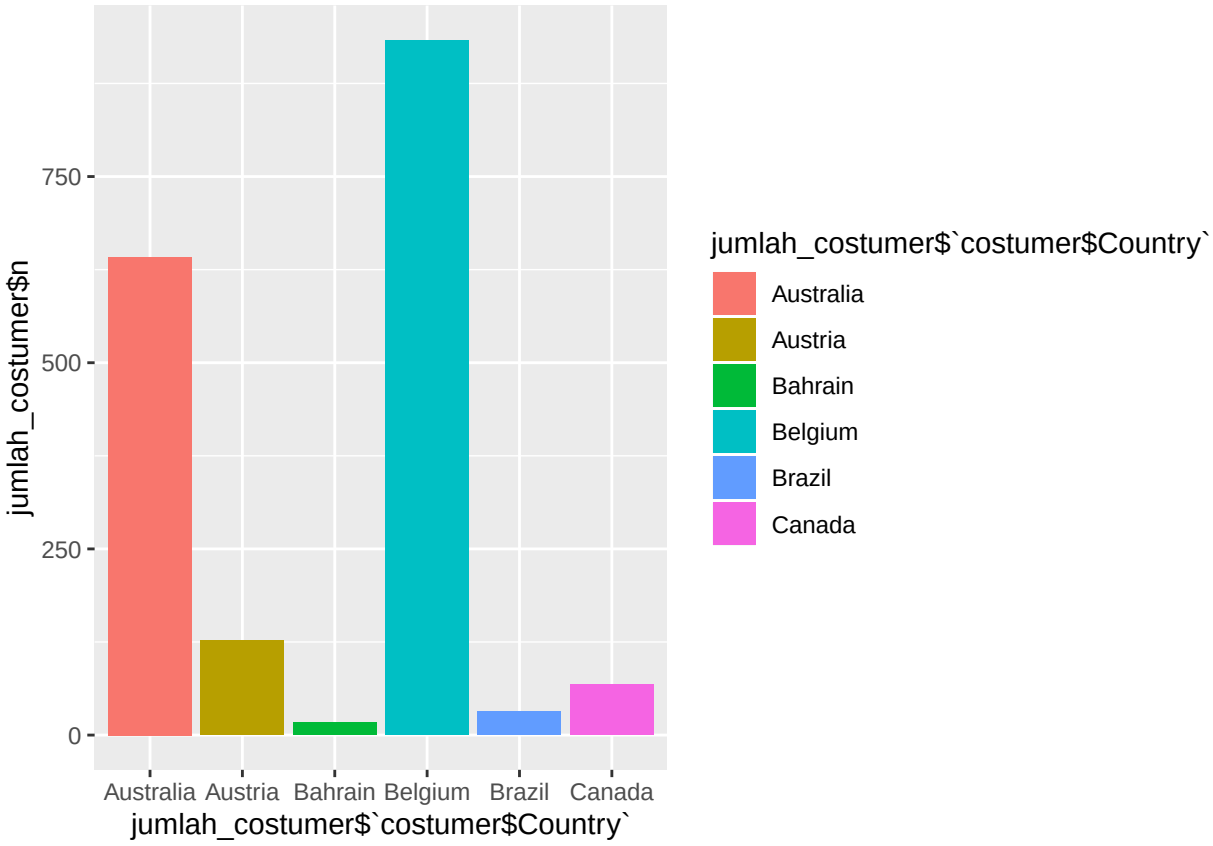
```
##
##            Hong Kong                 Unit         Saudi Arabia
##                    0                    1                   10
##              Bahrain       Czech Republic              Israel
##                   17                   17                   18
##                  USA United Arab Emirates               Brazil
##                   22                   30                   32
##   European Community            Lithuania              Lebanon
##                   32                   35                   45
##                Malta               Canada          Unspecified
##                   47                   68                   72
##               Greece              Iceland            Singapore
##                   85                  102                  118
##              Austria              Denmark              Poland
##                  127                  184                  187
##               Sweden              Japan                Italy
##                  200                  251                  309
##              Finland              Cyprus      Channel Islands
##                  312                  353                  368
##               Norway             Portugal            Australia
##                  378                  624                  642
##          Switzerland             Belgium          Netherlands
##                  708                  933                 1142
##                Spain                EIRE               France
##                 1142                 2718                 3625
##              Germany       United Kingdom
##                 4208               153620
```

```
jumlah_costumer= costumer %>% group_by(costumer$Country) %>% summarise(n=n()) %>% ungroup() %>% arrange
jumlah_costumer = head(jumlah_costumer)
jumlah_costumer
```

```
## # A tibble: 6 x 2
##   `costumer$Country`     n
##   <fct>              <int>
## 1 Australia            642
```

```
## 2 Austria              127
## 3 Bahrain               17
## 4 Belgium              933
## 5 Brazil                32
## 6 Canada                68
```

```
ggplot(data = jumlah_costumer, aes(x= jumlah_costumer$`costumer$Country`, y=jumlah_costumer$n, fill=juml
```



Berdasarkan dataset diatas, costumer tersebar dari 38 negara dimana negara dengan costumer yang paling banyak berasal dari UK berjumlah 220279 dan yang ter sediikit dari UEA berjumlah 1.

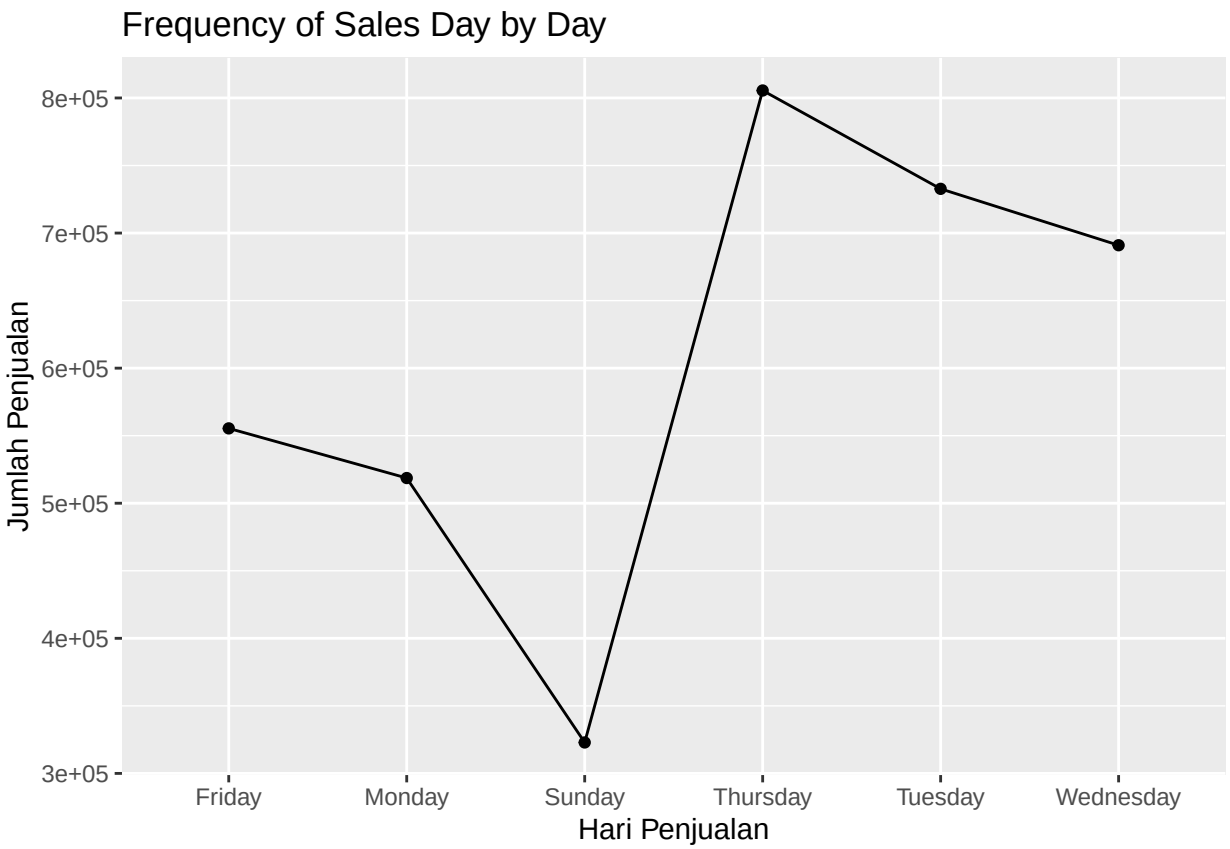
Frequency of sales day by day

```
max_week_sale <- filter(costumer, !is.na(CustomerID), !is.na(StockCode))
max_week_sale$InvoiceDate=mdy_hm(max_week_sale$InvoiceDate)
max_week_sale$weekdays <- weekdays(max_week_sale$InvoiceDate)
max_week_sale$Sales <- max_week_sale$Quantity * max_week_sale$UnitPrice
highsales = max_week_sale %>% group_by(max_week_sale$weekdays) %>% summarize(Salesamount = sum(Sales)) %
highsales= head(highsales)
highsales
```

```
## # A tibble: 6 x 2
##   `max_week_sale$weekdays` Salesamount
##   <chr>                          <dbl>
## 1 Thursday                     805537.
## 2 Tuesday                      732736.
## 3 Wednesday                    690984.
```

```
## 4 Friday                        555412.
## 5 Monday                        518657.
## 6 Sunday                        322900.
```

```
ggplot(highsales, aes(x= highsales$`max_week_sale$weekdays`, y= highsales$Salesamount, group=1))+ geom_l
```



Berdasarkaan frekuensi penjualan perhari bahwa dalam satu minggu, diketahui jumlah penjualan perhari dalam 1 minggu. Penjualan tertinggi terjadi pada hari Kamis dengan penjualan 805537 produk, kemudian hari Selasa dengan 732736 produk, hari Rabu dengan 690984 produk, hari Jum'at dengan 555412 produk, hari Senin dengan 518657 produk, dan penjualan terendah terjadi pada hari Minggu dengan penjualan 322900 produk.

Costumer yang Datang Beli lagi

```
repeatcost= costumer %>% group_by((costumer$CustomerID), n_distinct(InvoiceDate))%>%summarise(Count=n())
repeatcost
```

```
## # A tibble: 2,974 x 3
##    `(costumer$CustomerID)` `n_distinct(InvoiceDate)` Count
##                      <int>                     <int> <int>
##  1                   12346                      9735     2
##  2                   12347                      9735   102
##  3                   12348                      9735    28
##  4                   12350                      9735    17
##  5                   12352                      9735    48
##  6                   12353                      9735     4
##  7                   12354                      9735    58
```

```
##  8                   12355                       9735    13
##  9                   12356                       9735    57
## 10                   12359                       9735   145
## # ... with 2,964 more rows
```

Which Product Bring Most Revenue

```
revenue=costumer %>% group_by(costumer$StockCode) %>% summarise(sales=sum(Quantity*UnitPrice)) %>% ungr
revenue
```

```
## # A tibble: 3,282 x 2
##    `costumer$StockCode`  sales
##    <fct>                 <dbl>
##  1 22423                79452.
##  2 85123A               52422.
##  3 22502                46783.
##  4 47566                39990.
##  5 85099B               37543.
##  6 POST                 30577.
##  7 84879                23981.
##  8 79321                20446.
##  9 82484                20214.
## 10 21623                19326.
## # ... with 3,272 more rows
```

```
Sales_Detail<-costumer %>% mutate(Sales_Amount = Quantity*UnitPrice)

sales<-Sales_Detail%>% filter(!is.na(Sales_Amount))

sales %>%group_by(Country)%>% summarise(SalesAmount =sum(Sales_Amount= Quantity*UnitPrice)) %>%arrange(
```

```
## # A tibble: 37 x 2
##    Country        SalesAmount
##    <fct>                <dbl>
##  1 United Kingdom    2936812.
##  2 Netherlands        125721.
##  3 Germany            103526.
##  4 EIRE                94106.
##  5 France              87173.
##  6 Australia           79071.
##  7 Spain               24723.
##  8 Switzerland         22654.
##  9 Japan               21133.
## 10 Belgium             17251.
## # ... with 27 more rows
```

Gambar diatas menjelaskan tentang 10 produk yang paling banyak diminati oleh customer. Urutan pertama yang paling diminati adalah produk dengan kode 22423, dimana produk ini telah terjual sebanyak 101062 produk dan kode DOT sebanyak 87936 produk, dan begitu seterusnya dimana semakin bawah urutan kode produk maka semakin sedikit pembeli dari produk tersebut.

Gambar diatas menunjukkan 10 peringkat tertinggi dengan jumlah produk yang terjual terhadap negara tertentu. Penjualan tertinggi berada pada negara United Kingdom dengan jumlah penjualan 3572911 produk. Kemudian dilanjutkan dengan Netherlands sebanyak 125721 produk, Germany sebanyak 103526 produk, EIRE sebanyak 99384 produk, France sebanyak 87443 produk, Australia sebanyak 79071 produk, Spain sebanyak 24723 produk, Switzerland sebanyak 22654 produk, Japan sebanyak 21133 produk, dan urutan ke

10 adalah Belgium sebanyak 17251 produk, dan begitu seterusnya dimana semakin bawah urutan negara maka semakin sedikit produk yang terjual pada negara tersebut. ## #Can you find out which hours are most crowded and, therefore, need more staff?

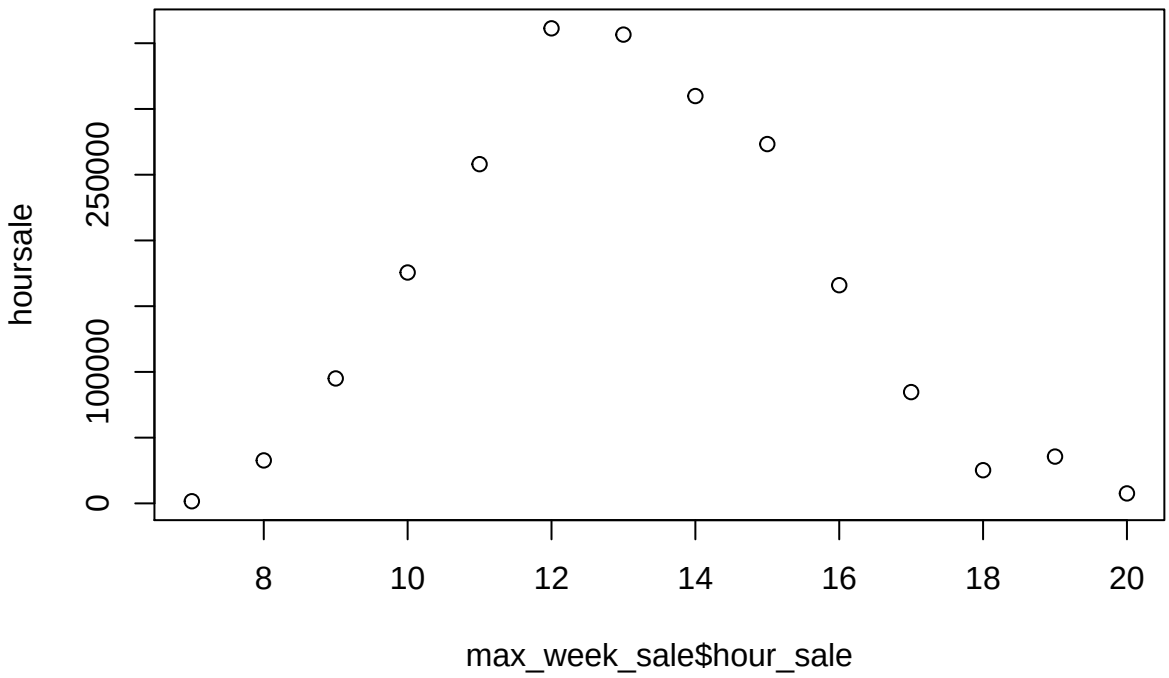
```
max_week_sale$hour_sale = hour(max_week_sale$InvoiceDate)
peakhour = max_week_sale %>% group_by(max_week_sale$hour_sale)%>% summarise(hoursale = sum(hour_sale))%
head(peakhour)
```

```
## # A tibble: 6 x 2
##   `max_week_sale$hour_sale` hoursale
##                       <int>    <int>
## 1                        12   361320
## 2                        13   356564
## 3                        14   309820
## 4                        15   273300
## 5                        11   258016
## 6                        10   175610
```

```
peakhour
```

```
## # A tibble: 14 x 2
##    `max_week_sale$hour_sale` hoursale
##                        <int>    <int>
##  1                        12   361320
##  2                        13   356564
##  3                        14   309820
##  4                        15   273300
##  5                        11   258016
##  6                        10   175610
##  7                        16   165952
##  8                         9    95022
##  9                        17    84643
## 10                        19    35587
## 11                         8    32664
## 12                        18    25218
## 13                        20     7580
## 14                         7     1624
```

```
plot(peakhour)
```



Penjualan dengan jumlah yang tinggi juga memerlukan staff tambahan untuk agar proses penjualan tetap stabil. Berasarkan plot diatas merupakan jumlah produk yang terjual dalam jam tertentu, sehingga pemilik perusahaan dapat mempertimbangkan pada waktu yang mana ia akan membutuhkan staff tambahan. Gambar diatas menunjukkan bahwa penjualan dimulai pagi hari dengan jumlah produk yang terjual terus meningkat hingga mencapai titik puncak dimana penjualan tertinggi akan terjadi pada pukul 12:00 sebanyak 361320 penjualan dan menurun dengan signifikan hingga malam hari. Penjualan dengan 5 peringkat teratas berada pada pukul 12:00, dan menurun pada pukul 13:00 dengan jumlah 356564 penjualan, pukul 14:00 sebanyak 309820 penjualan, pukul 15:00 sebanyak 273300 penjualan, pukul 11:00 sebanyak 258016 , pukul 10:00 sebanyak 175610 penjualan dan seterusnya akan menurun.

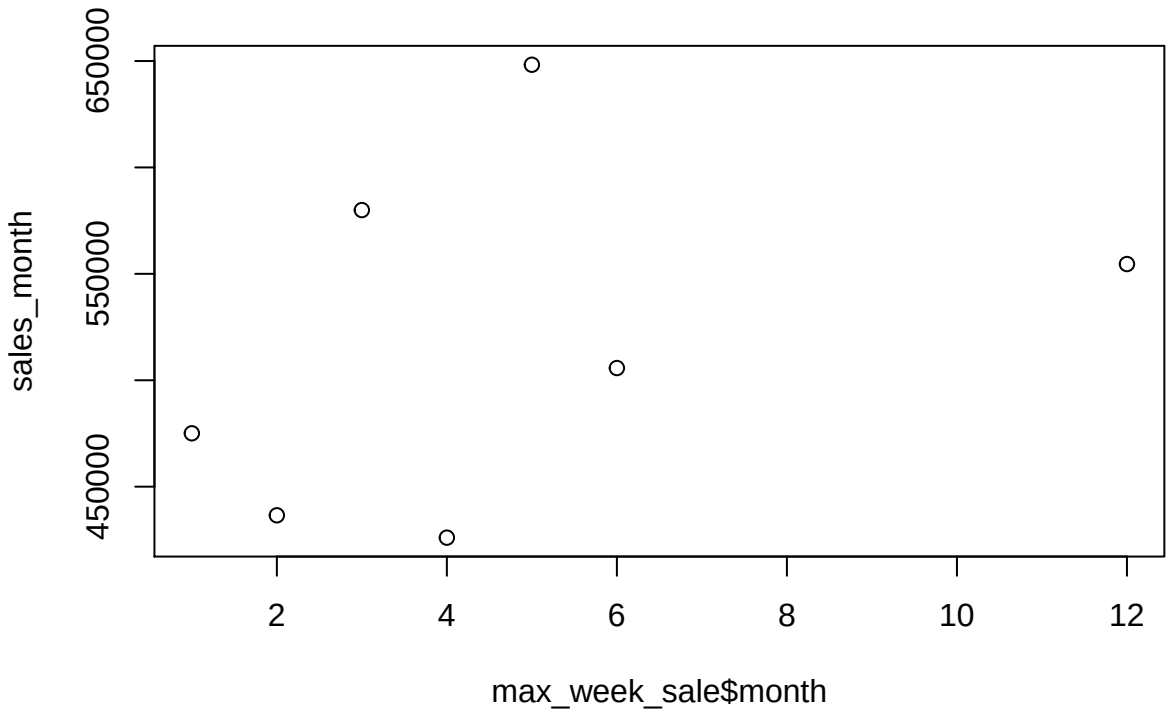
#Can you find out which mont are most crowded and, therefore, need more staff?

```
max_week_sale$month = month(max_week_sale$InvoiceDate)
peakmonth =  max_week_sale%>% group_by(max_week_sale$month) %>% summarise(sales_month= sum(Sales))
peakmonth
```

```
## # A tibble: 7 x 2
##   `max_week_sale$month` sales_month
##                   <dbl>       <dbl>
## 1                     1     475074.
## 2                     2     436546.
## 3                     3     579965.
## 4                     4     426048.
## 5                     5     648251.
## 6                     6     505736.
```

```
## 7                    12     554604.
```

```
plot(peakmonth)
```



Berikut merupakan plot dari penjualan produk perbulan, sehingga pemilik perusahaan dapat mengetahui puncak penjualan tertinggi yang dapat dipertimbangkan sebagai kebijakan dalam menarik perhatian konsumen. Berdasarkan plot yang ada diketahui bahwa penjualan tertinggi terjadi pada bulan ke-5 atau Mei dengan jumlah 648251 produk dan penjualan terendah terjadi pada bulan April dengan jumlah 426048 produk. ## TOP 10 COSTUMER

```
topten = max_week_sale %>% group_by(max_week_sale$CustomerID)%>% summarise(spend = sum(Sales))%>% arran
topten
```

```
## # A tibble: 10 x 2
##    `max_week_sale$CustomerID`   spend
##                         <int>   <dbl>
##  1                      14646 121929.
##  2                      18102 106443.
##  3                      12415  73717.
##  4                      17450  59462.
##  5                      14156  48281.
##  6                      14911  42460.
##  7                      17511  39868.
##  8                      15311  35582.
##  9                      15061  31974.
## 10                      15769  30766.
```

Diketahui bahwa customer dengan ID 14646 merupakan customer paling banyak yang melakukan retail online dengan jumlah produk terbeli sebanyak 121929 produk. Kemudian diurutan kedua diduduki oleh customer

dengan ID 18102 dengan jumlah pembelian sebanyak 106443 produk, urutan ketiga customer dengan ID 12415 dengan pembelian 73717 produk, dan seterusnya dimana semakin bawah keduduk ID customer maka semakin sedikit produk yang dibeli.

FINAL

Setelah melakukan analisis deskriptif maka dapat direkomendasikan beberapa kebijakan yang dapat digunakan untuk memajukan perusahaan tersebut, diantara lain :

1. Gratis Ongkos kirim atau discount ongkos kirim untuk produk yang dibeli bagi negara berdasarkan analisis frekuensi transaksi atas negara dengan customer tertinggi.
2. Menambah jumlah staff pada hari kamis berdasarkan analisis frekurnsi tertinggi penjualan perhari.

3.Memberi poin setiap kali pembelian kepada customer yang dapat ditukarkan dengan produk tertentu berdasarkan analisis frekuensi pembelian produk atas ID costumer.

4.Memberi diskon atau potongan harga pada bulan mei berdasarkan analasis penjualan tertinggi perbulan.

5.Menjalankan flashsale pada jam 12:00 berdasarkan analisis penjualan tertinggi atas jam.

6. Mengadakan kuis berhadiah voucher belanja pada hari kamis berdasarkan analisis frekuensi tertinggi atas penjualan perhari

7.Memberikan cashback 10% atas pembelian produk dengan code 22423 berdasarkan analisis produk yang paling sering terjual.