Onlin Retail

James Fang

04/09/2020

Data proprocessing

```
library(tidyverse)
retail.df <- read.csv('online_retail_II.csv')
Date <- as.Date(sapply(strsplit(retail.df$InvoiceDate,' '),function(x) x[1]))
retail.df <- data.frame(retail.df,Date)
```

Value = Quantity * Unit Price

```
retail.df$Value <- retail.df$Quantity* retail.df$Price
```

Volume can be negative due to sales returns, but price normally should be positive. Let's have a look.

```
head(retail.df[retail.df$Price <=0 ,])
```

```
##      Invoice StockCode  Description Quantity         InvoiceDate Price
## 264   489464     21733 85123a mixed      -96 2009-12-01 10:52:00     0
## 284   489463     71477        short     -240 2009-12-01 10:52:00     0
## 285   489467    85123A  21733 mixed     -192 2009-12-01 10:53:00     0
## 471   489521     21646                   -50 2009-12-01 11:44:00     0
## 3115  489655     20683                   -44 2009-12-01 17:26:00     0
## 3162  489659     21350                   230 2009-12-01 17:39:00     0
##      Customer.ID        Country       Date Value
## 264           NA United Kingdom 2009-12-01     0
## 284           NA United Kingdom 2009-12-01     0
## 285           NA United Kingdom 2009-12-01     0
## 471           NA United Kingdom 2009-12-01     0
## 3115          NA United Kingdom 2009-12-01     0
## 3162          NA United Kingdom 2009-12-01     0
```

It makes sense that these data should be removed no matter for calculating revenue or sales revenue.

```
retail.df <- retail.df[retail.df$Price >0 ,]
```

Badwords like ?????, damaged,lost, etc. should be removed. By observation, we get the following badwords list.

```
badwords <- c('?',
'?????',
'back charges',
'bad quality',
'Came as green?',
'Came as green?',
'cant find',
'cant find',
'check',
'checked',
'checked',
'code mix up 72597',
'code mix up 72597',
'coding mix up',
'crushed',
'crushed',
'damaged',
'damaged/dirty',
'damaged?',
'damages',
'damages etc',
'damages, lost bits etc',
'damages?',
'damges',
'Damp and rusty',
'dirty',
'dirty, torn, thrown away.',
'display',
'entry error',
'faulty',
'for show',
'given away',
'gone',
'Gone',
'incorrect credit',
'lost',
'lost in space',
'lost?',
'missing',
'Missing',
'missing (wrongly coded?)',
'missing?',
'missings',
'reverse mistake',
'Rusty ',
'Rusty connections',
'show',
'show display',
'smashed',
'sold in wrong qnty',
'This is a test product.',
'used for show display',
'wet',
'wet & rotting',
'wet and rotting',
'wet cartons',
'wet ctn',
'wet damages',
'Wet, rusty-thrown away',
'wet/smashed/unsellable',
'wrong code',
'wrong ctn size',
'Zebra invcing error')
retail.df$Description[retail.df$Description %in% badwords]
```

```
##  [1] "This is a test product." "This is a test product."
##  [3] "This is a test product." "This is a test product."
##  [5] "This is a test product." "This is a test product."
##  [7] "This is a test product." "This is a test product."
##  [9] "This is a test product." "This is a test product."
## [11] "This is a test product." "This is a test product."
## [13] "This is a test product." "This is a test product."
```

Remove these recoreds that contain bad words.

```
retail.df <- retail.df[!retail.df$Description %in% badwords,]
```

There are 236871 NA value in Customer.ID, we replace NA to 99999

```
summary(retail.df)
```

```
##    Invoice           StockCode         Description           Quantity
##  Length:1061150     Length:1061150     Length:1061150     Min.   :-80995.0
##  Class  :character  Class  :character  Class  :character  1st Qu.:      1.0
##  Mode   :character  Mode   :character  Mode   :character  Median :      3.0
##                                                           Mean   :     10.3
##                                                           3rd Qu.:     10.0
##                                                           Max.   : 80995.0
##
##  InvoiceDate            Price             Customer.ID       Country
##  Length:1061150     Min.   :    0.00   Min.   :12346    Length:1061150
##  Class  :character  1st Qu.:    1.25   1st Qu.:13975    Class  :character
##  Mode   :character  Median :    2.10   Median :15257    Mode   :character
##                     Mean   :    4.83   Mean   :15325
##                     3rd Qu.:    4.15   3rd Qu.:16797
##                     Max.   :38970.00   Max.   :18287
##                                        NA's   :236871
##       Date                 Value
##  Min.   :2009-12-01   Min.   :-168469.60
##  1st Qu.:2010-07-09   1st Qu.:      3.75
##  Median :2010-12-07   Median :      9.90
##  Mean   :2011-01-02   Mean   :     18.33
##  3rd Qu.:2011-07-22   3rd Qu.:     17.70
##  Max.   :2011-12-09   Max.   : 168469.60
##
```

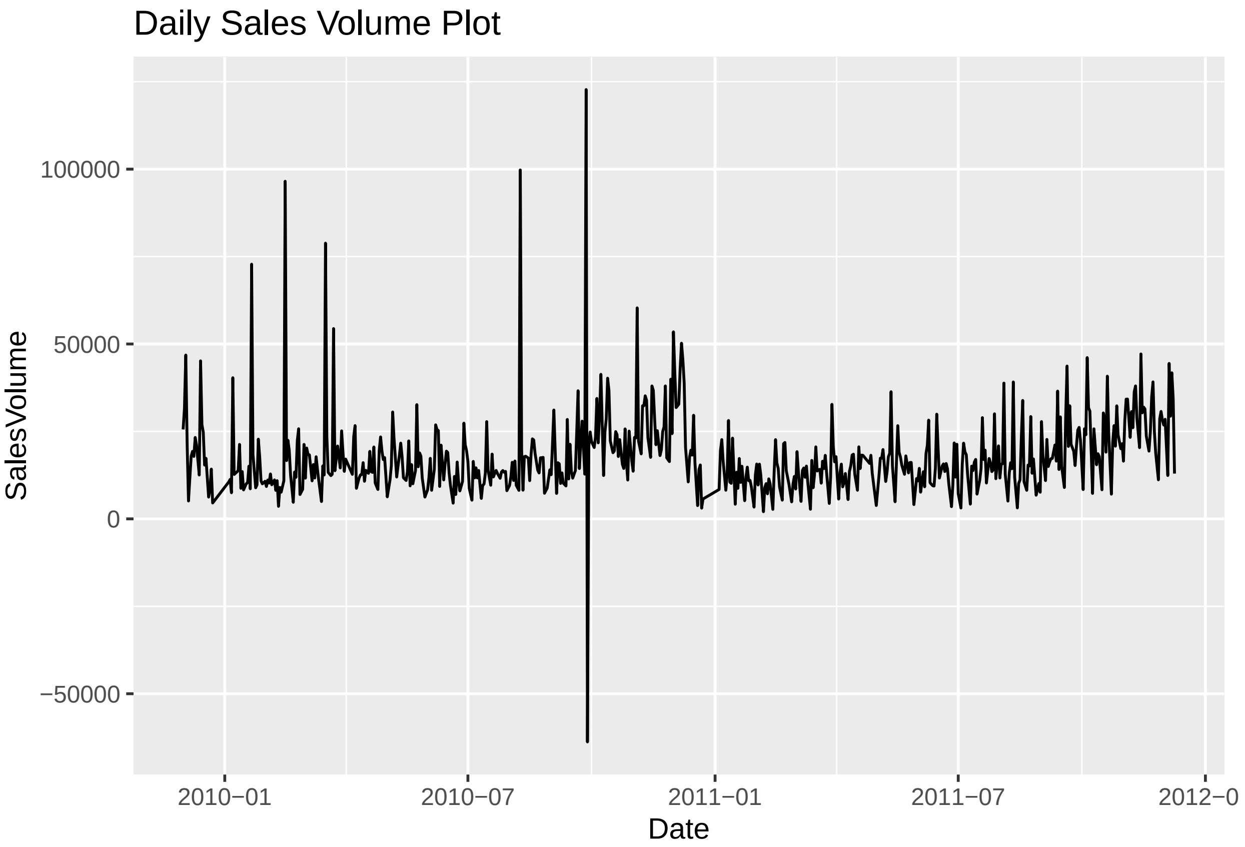
```
retail.df$Customer.ID[is.na(retail.df$Customer.ID)] <- 99999
```

Task I Visualization

(a)

Daily Sales Volume Plot

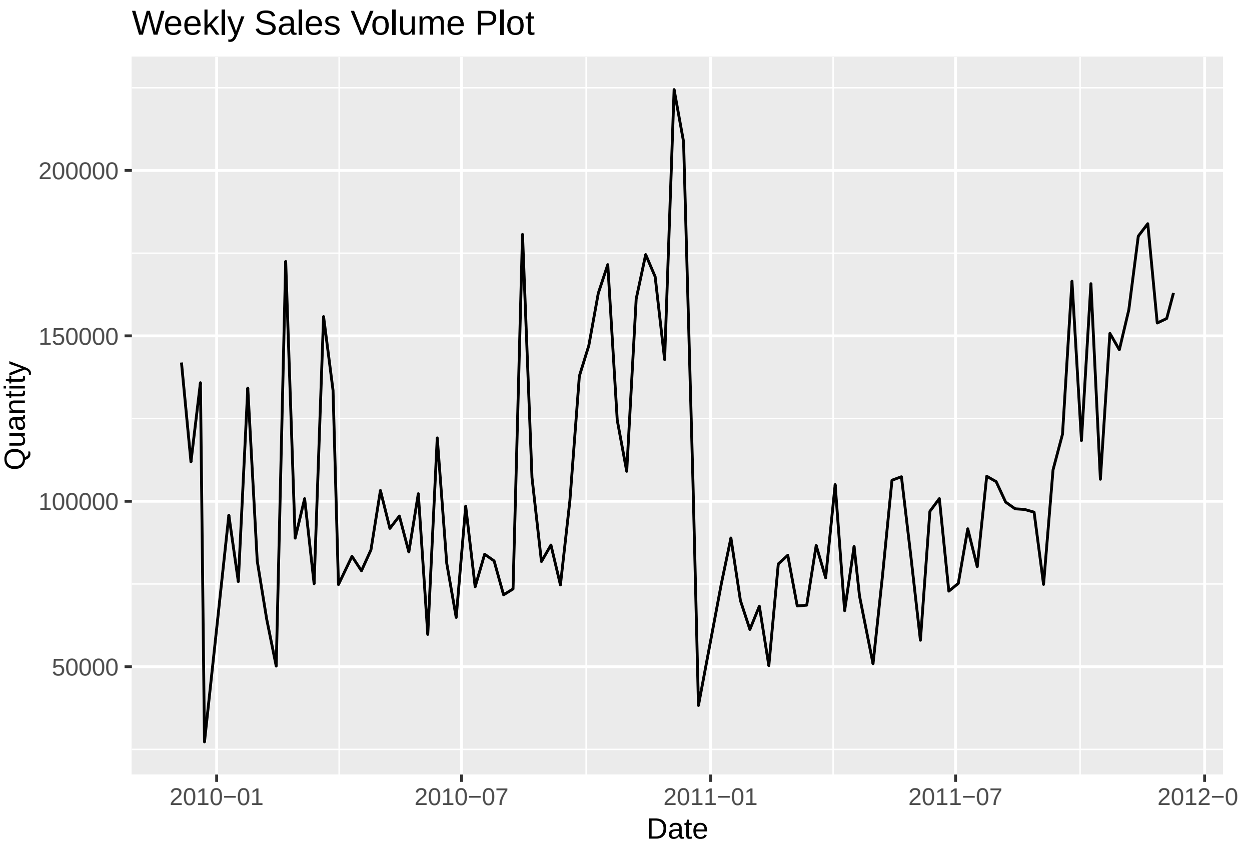
```
SalesVolume_byDay <- retail.df %>% group_by(Date) %>% summarize(SalesVolume = sum(Quantity))
ggplot(SalesVolume_byDay,aes(x=Date,y=SalesVolume)) + geom_line() + ggtitle("Daily Sales Volume Plot")
```



There is a flip around 2010-10. Sale Volume went up to a abnormal peak, then down to a negative value next day.

Weekly Sales Volume Plot

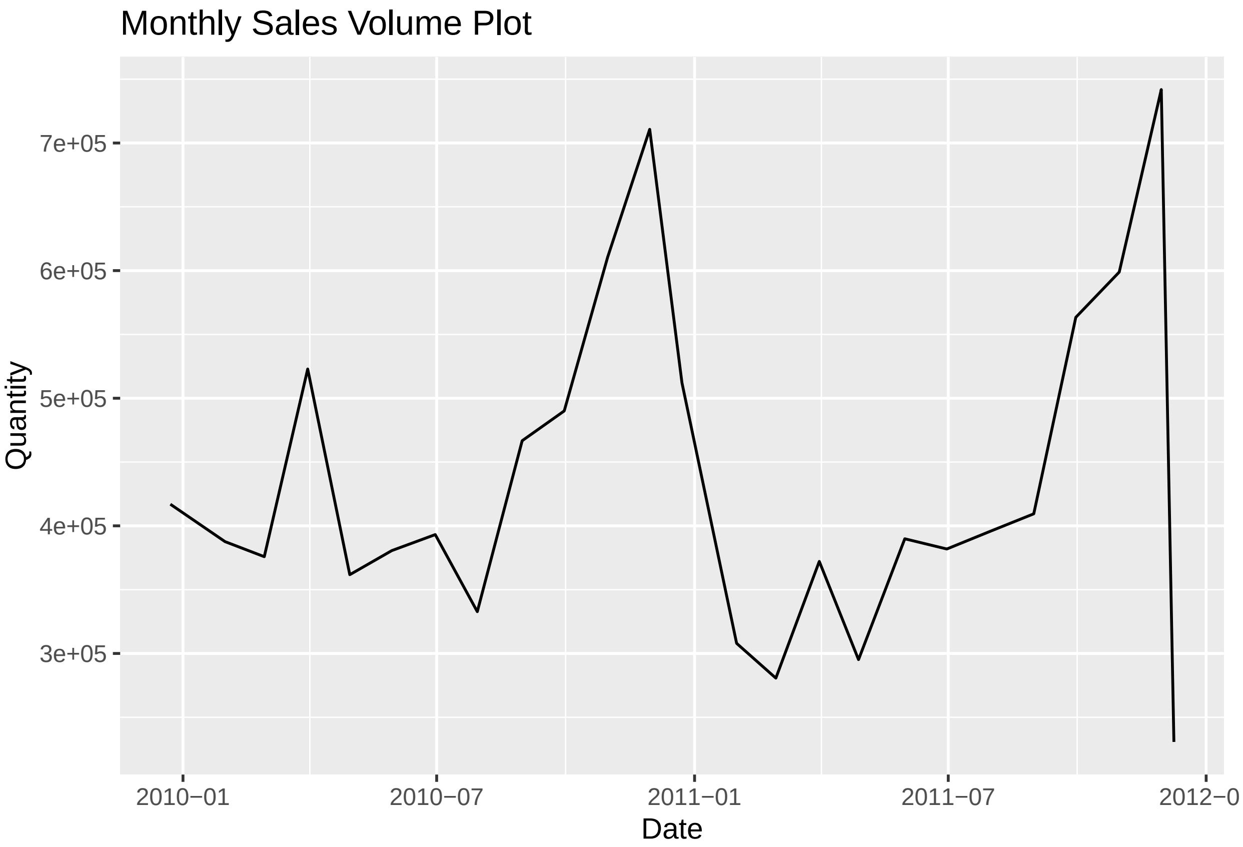
```
library(tidyquant)
SalesVolume_byWeek <- retail.df %>% tq_transmute(select = Quantity,
                           mutate_fun = apply.weekly,
                           FUN = sum)
ggplot(SalesVolume_byWeek,aes(x=Date, y=Quantity)) + geom_line() + ggtitle("Weekly Sales Volume Plot")
```



It can be seen that there exits a peak around December. Is there anything to do with Black Friday? Of course.

Monthly Sales Volume Plot

```
SalesVolume_byMonth <- retail.df %>% tq_transmute(select = Quantity,
                                                 mutate_fun = apply.monthly,
                                                 FUN = sum)
ggplot(SalesVolume_byMonth,aes(x=Date, y=Quantity)) + geom_line() + ggtitle("Monthly Sales Volume Plot")
```

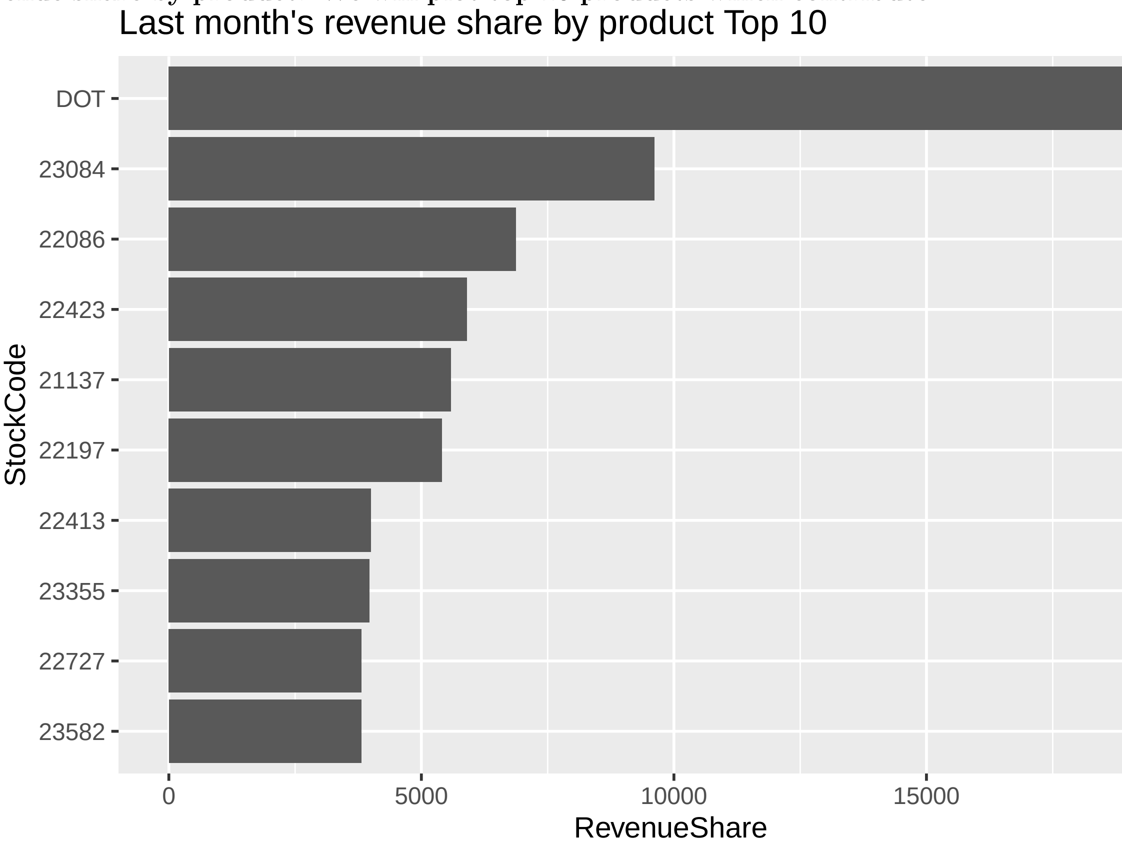


Yikes, Black Fridays again.

(b)

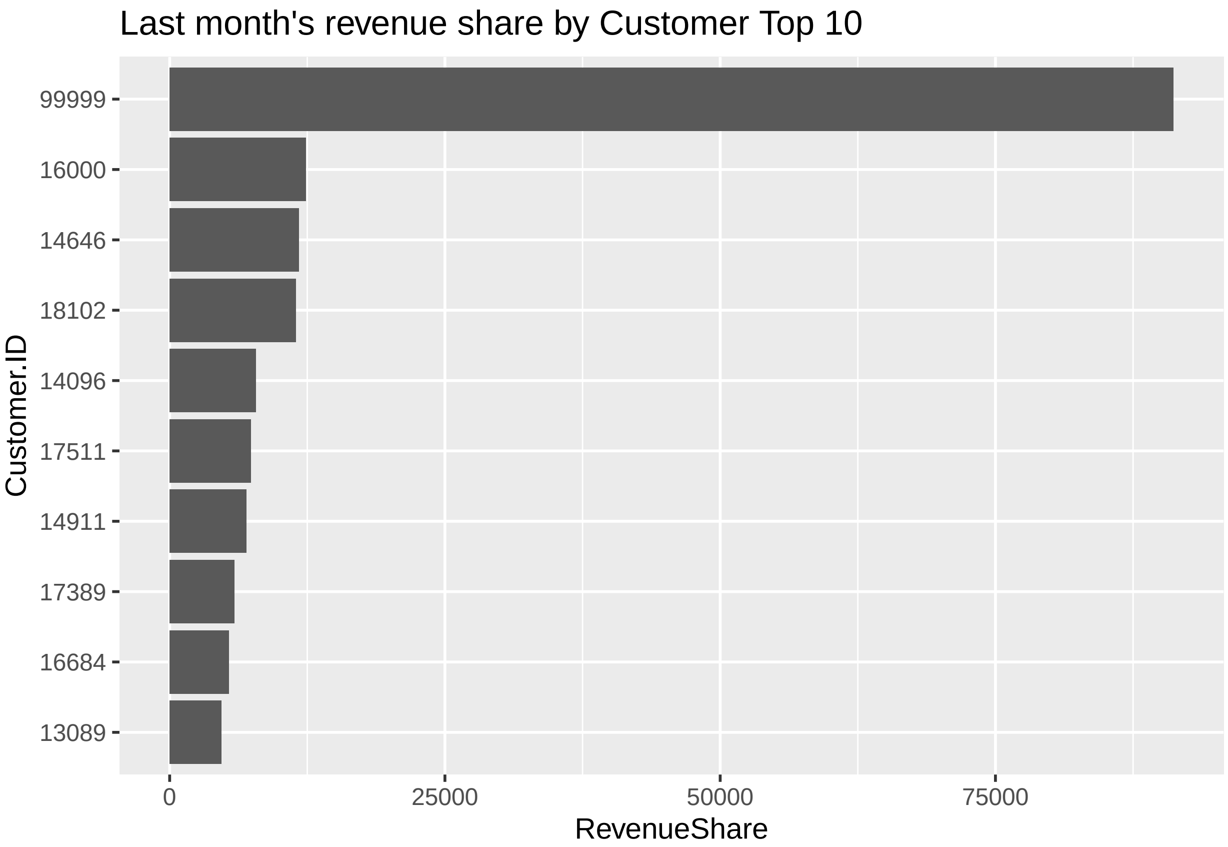
```
RS_byProduct <- retail.df %>% filter(Date >= "2011-12-01") %>%
          group_by(StockCode) %>%
          summarise(RevenueShare = sum(Value)) %>%
          arrange(desc(RevenueShare))
ggplot(RS_byProduct[1:10,], aes(x = reorder(StockCode,RevenueShare),y = RevenueShare)) +
  geom_col() + coord_flip() + xlab("StockCode") + ggtitle("Last month's revenue share by product Top 10
```

Plot of Last month’s revenue share by product. We will plot top 10 products which contribute most to revenue share.



Plot of Last month’s revenue share by customer. We will plot top 10 customers which contributes most to revenue share.

```
RS_byCustomer <- retail.df %>% filter(Date >= "2011-12-01") %>%
  group_by(Customer.ID) %>%
  summarise(RevenueShare = sum(Value))%>%
  arrange(desc(RevenueShare))
ggplot(RS_byCustomer[1:10,], aes(x = reorder(Customer.ID,RevenueShare),y = RevenueShare)) +
    geom_col() + coord_flip() + xlab("Customer.ID") + ggtitle("Last month's revenue share by Customer T
```

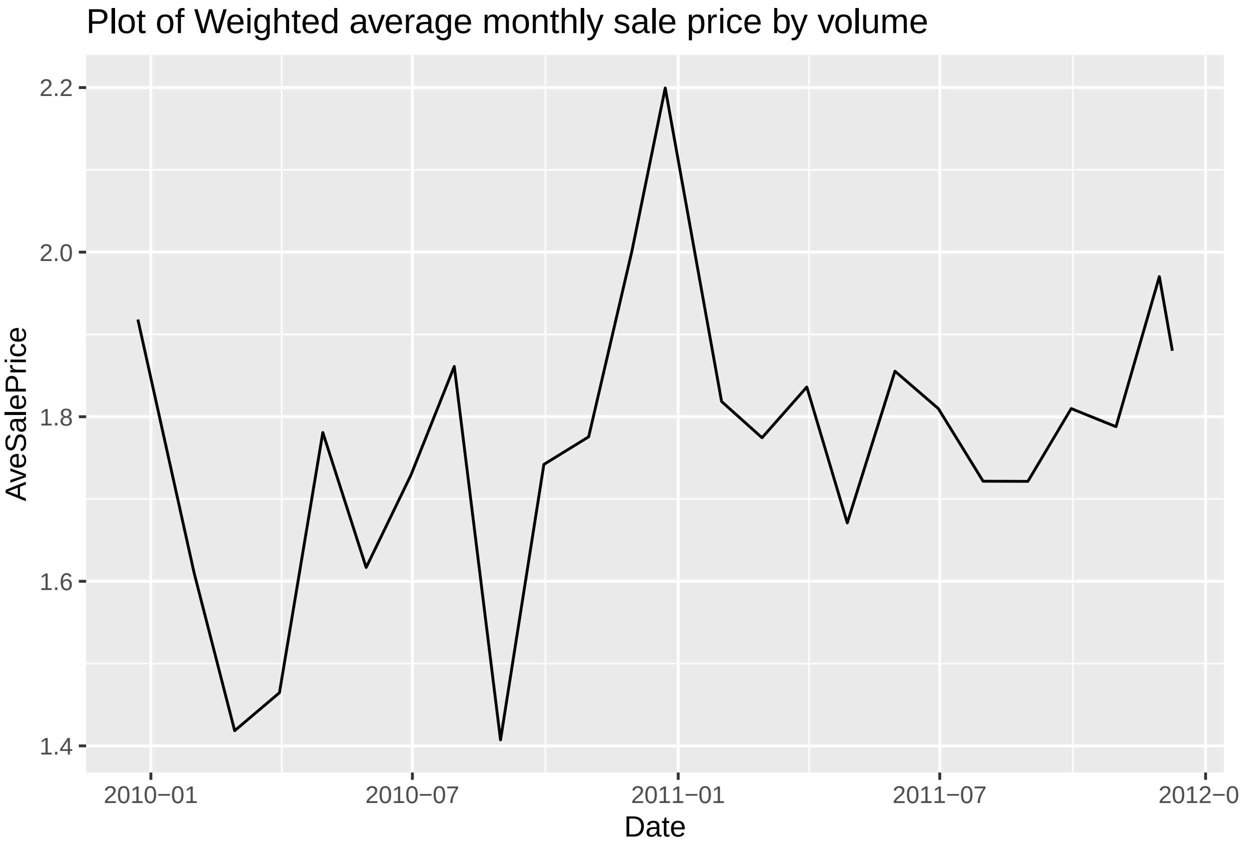


NA customers contribute a lot to revenue. The shop owner should trace these customers to get better classify his customer group.

### (c) #### Plot of Weighted average monthly sale price by volume

```
a <- retail.df %>% tq_transmute(select = Value,
                                                  mutate_fun = apply.monthly,
                                                  FUN = sum)
b <- retail.df %>% tq_transmute(select =Quantity ,
                           mutate_fun = apply.monthly,
                           FUN = sum)

SalesVolume_byMonth <- data.frame(Date = a$Date,AveSalePrice = a$Value/b$Quantity)
ggplot(SalesVolume_byMonth,aes(x=Date, y=AveSalePrice)) + geom_line() + ggtitle("Plot of Weighted averag
```



According to the plot, we can say that people would like to buy more expensive stuff around December. Maybe because of Christmas Day's Gift.

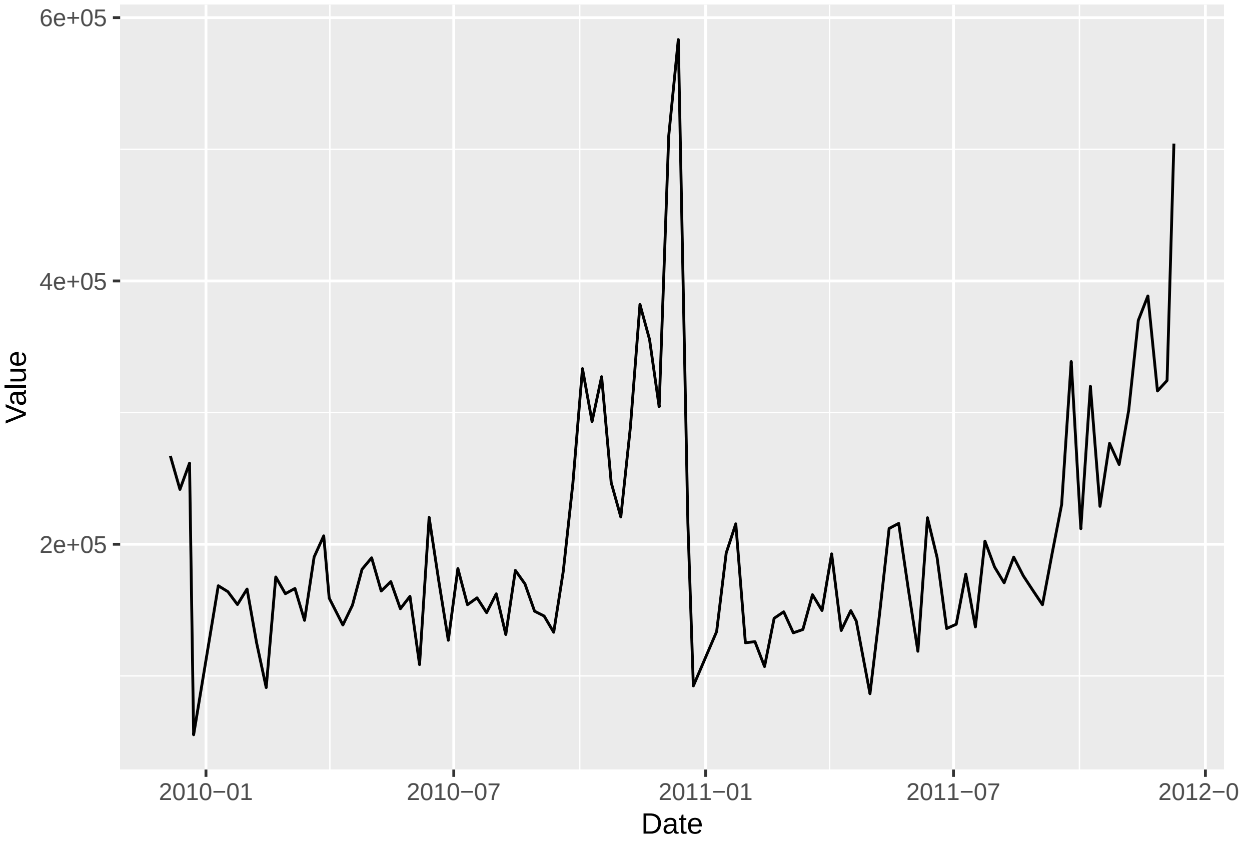
Task II Modelling

As being told, there are negative numbers in Quantity. We should drop these values.

```
retail.df <- retail.df %>% filter(Quantity>0)
```

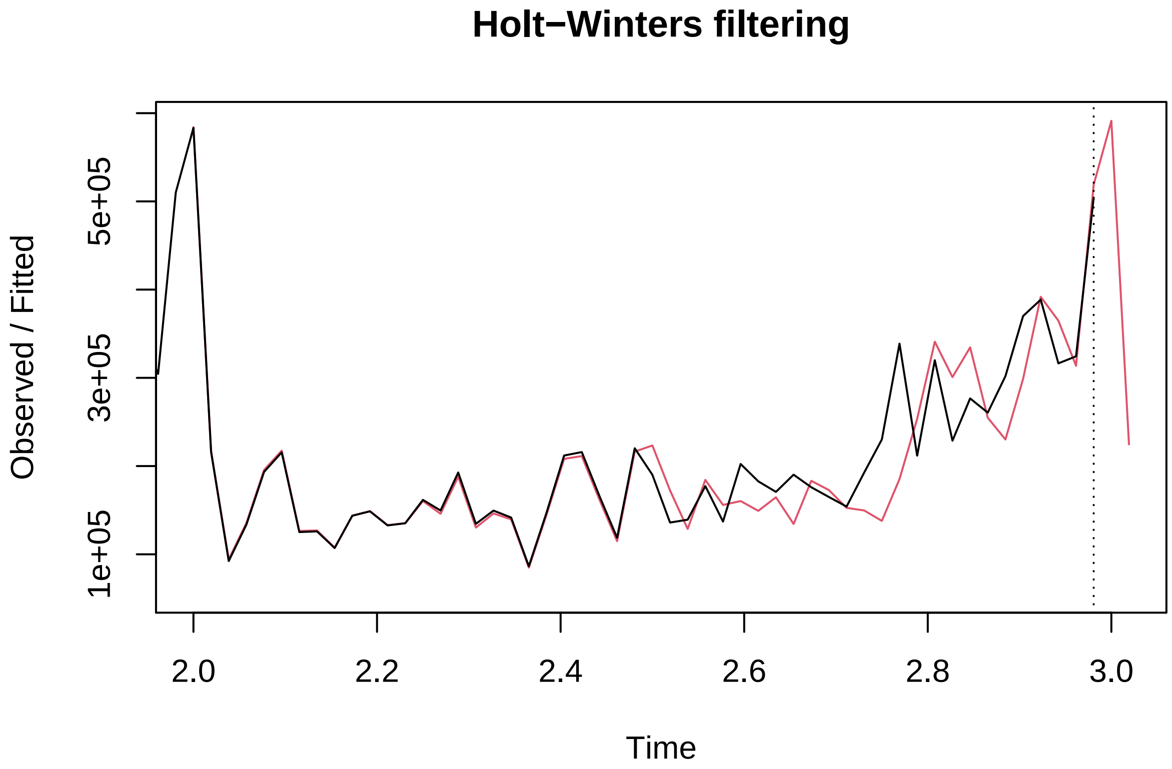
By observation, there are 3 business weeks in December, and we already have the revenue of the first week of 12/2011. To get total revenue of this months, we need predict the revenue of the second and the third week. Let's draw a revenue time series curve by week unit.

```
Value_byWeek <- retail.df %>% tq_transmute(select = Value,
                                                 mutate_fun = apply.weekly,
                                                 FUN = sum)
ggplot(Value_byWeek,aes(x=Date, y=Value)) + geom_line()
```



This time series curve looks good to me. We can build a time series model to predict the revenue. We will use modified Holt-Winters' method to forcast.

```
#By observation, there are 52  weeks in a year.
Value_byWeek.ts = ts(Value_byWeek$Value,frequency = 52)
rv.fit  <- HoltWinters(Value_byWeek.ts)
pred <- predict(rv.fit, n.ahead =2 )
plot(rv.fit,pred)
```

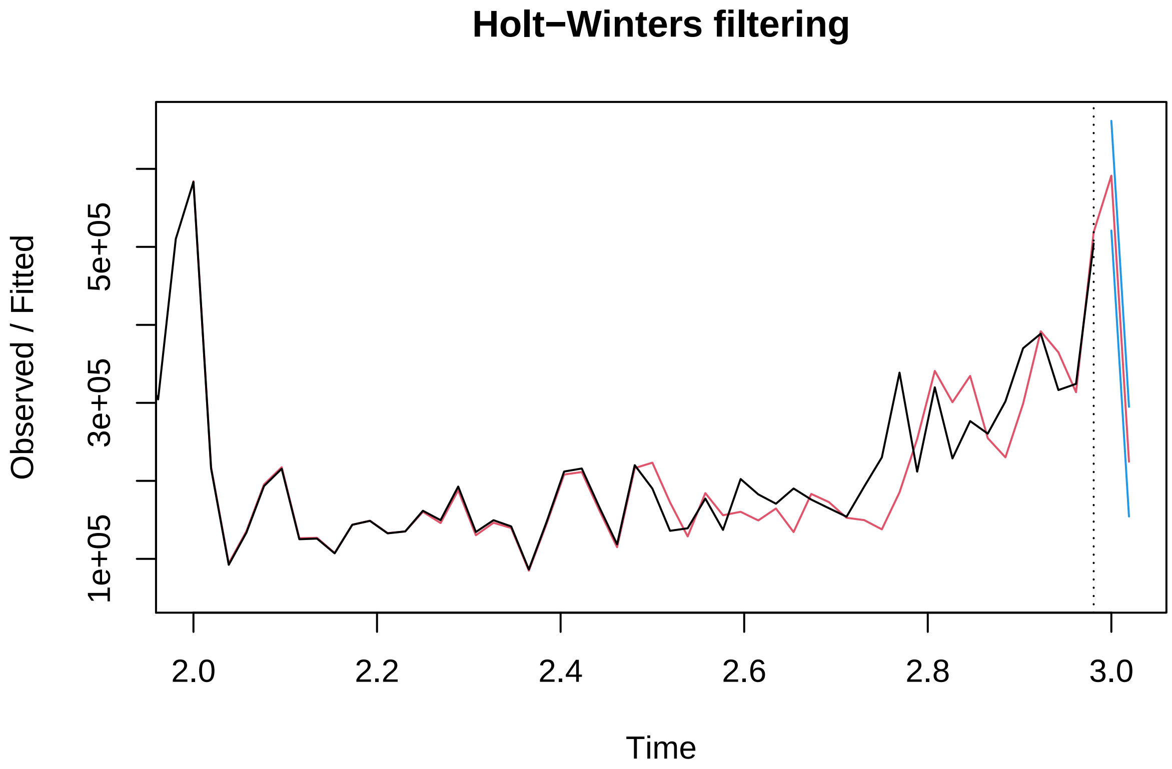


```
pred
```

```
## Time Series:
## Start = c(3, 1)
## End = c(3, 2)
## Frequency = 52
##           fit
## [1,] 591272.6
## [2,] 224595.1
```

Accoording to the prediction plot, the prediction lines are pretty close to the actual line. And We got 2 week's prediction. We can also put a prediction interval on our plot

```
pred.interval <- predict(rv.fit, n.ahead =2 ,prediction.interval = TRUE)
plot(rv.fit,pred.interval)
```



```
pred.interval
```

```
## Time Series:
## Start = c(3, 1)
## End = c(3, 2)
## Frequency = 52
##                fit      upr      lwr
## 3.000000 591272.6 661561.0 520984.2
## 3.019231 224595.1 294885.2 154305.0
```

Now, we will give the answer of predction of last month's revenue

```
Value_byWeek.ts[length(Value_byWeek.ts)] + sum(pred)
```

```
## [1] 1320196
```

Last, Let's talk about the uncertainties we still need to explore. One thing is that Our model didn't take full use of user's country information. With the company's business expanding, its oversea market will also expand. If this assumption holds, we can build our model according to different countries, because the consumption habits vary between different countries. For example, will Australia customers spend as much as Uk's customers on Black Friday, with their December on Summer? the answer is not necessary positive.