Assignment-Online Retail Analytics

library(dplyr)  
Online\_Retail<-read.csv("./Online\_Retail.csv")

**1. Breakdown of the number of transactions by countries in both percentage and count, and showing the countries accounting for more than 1% of the total transactions:**

trans\_countries<-Online\_Retail %>% group\_by(Country) %>% summarise(cnt = n()) %>% mutate(perc =round((cnt/sum(cnt))\*100,5)) %>% filter(perc>1)  
head(trans\_countries)

## # A tibble: 4 × 3  
## Country cnt perc  
## <chr> <int> <dbl>  
## 1 EIRE 8196 1.51  
## 2 France 8557 1.58  
## 3 Germany 9495 1.75  
## 4 United Kingdom 495478 91.4

**2. Creating a new variable TransactionValue and adding it to the dataframe:**

TransactionValue<-Online\_Retail$Quantity\*Online\_Retail$UnitPrice  
  
#creating a dataframe and adding TransactionValue to it  
  
Online\_Retail\_new<-data.frame(InvoiceNo=Online\_Retail$InvoiceNo,StockCode= Online\_Retail$StockCode,Description=Online\_Retail$Description,Quantity=Online\_Retail$Quantity, InvoiceDate=Online\_Retail$InvoiceDate,UnitPrice=Online\_Retail$UnitPrice, CustomerID=Online\_Retail$CustomerID,Country=Online\_Retail$Country, TransactionValue=TransactionValue)

**3. Showing the breakdown of transaction values by countries in total sum of transaction value. Displaying countries with total transaction exceeding 13000**

Trans\_value\_countries<- Online\_Retail\_new %>% group\_by(Country) %>% summarise(sum\_TransactionValue = sum(TransactionValue)) %>% filter(sum\_TransactionValue>130000)  
  
head(Trans\_value\_countries)

## # A tibble: 6 × 2  
## Country sum\_TransactionValue  
## <chr> <dbl>  
## 1 Australia 137077.  
## 2 EIRE 263277.  
## 3 France 197404.  
## 4 Germany 221698.  
## 5 Netherlands 284662.  
## 6 United Kingdom 8187806.

**4. Optional question**

Temp=strptime(Online\_Retail\_new$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')  
head(Temp)

## [1] "2010-12-01 08:26:00 GMT" "2010-12-01 08:26:00 GMT"  
## [3] "2010-12-01 08:26:00 GMT" "2010-12-01 08:26:00 GMT"  
## [5] "2010-12-01 08:26:00 GMT" "2010-12-01 08:26:00 GMT"

#New\_Invoice\_Date  
Online\_Retail\_new$New\_Invoice\_Date<- as.Date(Temp)  
  
Online\_Retail\_new$New\_Invoice\_Date[20000]- Online\_Retail\_new$New\_Invoice\_Date[10]

## Time difference of 8 days

#Invoice\_Week  
  
Online\_Retail\_new$Invoice\_Day\_Week= weekdays(Online\_Retail\_new$New\_Invoice\_Date)  
  
#Invoice\_Hour  
  
Online\_Retail\_new$New\_Invoice\_Hour = as.numeric(format(Temp, "%H"))  
  
#Invoice\_month  
  
Online\_Retail\_new$New\_Invoice\_Month = as.numeric(format(Temp, "%m"))

**4(a). Percentage of transactions (by numbers) by days of the week**

perc\_transc<- Online\_Retail\_new %>% group\_by(Invoice\_Day\_Week) %>% summarise(count=n()) %>% mutate(perc =round((count/sum(count))\*100,5))  
  
head(perc\_transc)

## # A tibble: 6 × 3  
## Invoice\_Day\_Week count perc  
## <chr> <int> <dbl>  
## 1 Friday 82193 15.2  
## 2 Monday 95111 17.6  
## 3 Sunday 64375 11.9  
## 4 Thursday 103857 19.2  
## 5 Tuesday 101808 18.8  
## 6 Wednesday 94565 17.5

**4(b). Percentage of transactions (by transaction volume) by days of the week**

perc\_trans\_week<- Online\_Retail\_new %>% group\_by(Invoice\_Day\_Week) %>% summarise(Total=sum(TransactionValue)) %>% mutate(percentage = (Total/sum(Total))\*100)  
head(perc\_trans\_week)

## # A tibble: 6 × 3  
## Invoice\_Day\_Week Total percentage  
## <chr> <dbl> <dbl>  
## 1 Friday 1540611. 15.8   
## 2 Monday 1588609. 16.3   
## 3 Sunday 805679. 8.27  
## 4 Thursday 2112519 21.7   
## 5 Tuesday 1966183. 20.2   
## 6 Wednesday 1734147. 17.8

**4(c). Percentage of transactions (by transaction volume) by month of the year**

perc\_trans\_month<- Online\_Retail\_new %>% group\_by(New\_Invoice\_Month) %>% summarise(Total=sum(TransactionValue)) %>% mutate(percentage = (Total/sum(Total))\*100)  
head(perc\_trans\_month)

## # A tibble: 6 × 3  
## New\_Invoice\_Month Total percentage  
## <dbl> <dbl> <dbl>  
## 1 1 560000. 5.74  
## 2 2 498063. 5.11  
## 3 3 683267. 7.01  
## 4 4 493207. 5.06  
## 5 5 723334. 7.42  
## 6 6 691123. 7.09

**4(d). The date with the highest number of transactions from Australia**

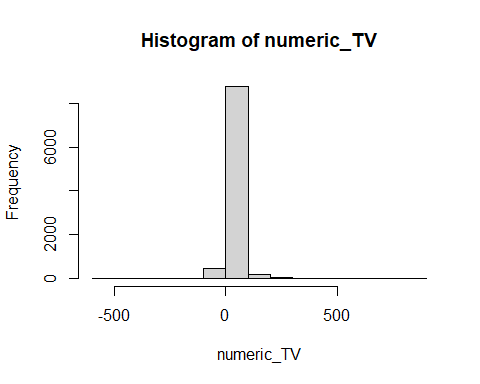
date\_trans<- Online\_Retail\_new %>% filter(Country == 'Australia') %>% group\_by(New\_Invoice\_Date) %>% summarise(count=n()) %>% arrange(desc(count))  
  
head(date\_trans)

## # A tibble: 6 × 2  
## New\_Invoice\_Date count  
## <date> <int>  
## 1 2011-06-15 139  
## 2 2011-07-19 137  
## 3 2011-08-18 97  
## 4 2011-03-03 84  
## 5 2011-10-05 82  
## 6 2011-05-17 73

#Australia recorded highest number of transactions on 2011-06-15

**5. Histogram of transaction Values from Germany**

histogram<- Online\_Retail\_new %>% filter(Country == 'Germany')   
numeric\_TV<- as.integer(histogram$TransactionValue)  
hist(numeric\_TV)



**6. Identifying the customer with highest number of transactions and finding the most valuable customer**

cust\_count<-Online\_Retail\_new %>% group\_by(CustomerID) %>% summarise(cntt = n()) %>% arrange(desc(cntt))  
head(cust\_count)

## # A tibble: 6 × 2  
## CustomerID cntt  
## <int> <int>  
## 1 NA 135080  
## 2 17841 7983  
## 3 14911 5903  
## 4 14096 5128  
## 5 12748 4642  
## 6 14606 2782

#Customer 17841 has the highest number of transactions.

cust\_sum<-Online\_Retail\_new %>% group\_by(CustomerID) %>% summarise(sum\_cnt =sum(TransactionValue)) %>% arrange(desc(sum\_cnt))  
  
head(cust\_sum)

## # A tibble: 6 × 2  
## CustomerID sum\_cnt  
## <int> <dbl>  
## 1 NA 1447682.  
## 2 14646 279489.  
## 3 18102 256438.  
## 4 17450 187482.  
## 5 14911 132573.  
## 6 12415 123725.

#Customer 14646 is the most valuable

**7. Percentage of missing values for each variable in the dataset**

missing\_values<- (colMeans(is.na(Online\_Retail\_new))\*100)  
  
missing\_values

## InvoiceNo StockCode Description Quantity   
## 0.00000 0.00000 0.00000 0.00000   
## InvoiceDate UnitPrice CustomerID Country   
## 0.00000 0.00000 24.92669 0.00000   
## TransactionValue New\_Invoice\_Date Invoice\_Day\_Week New\_Invoice\_Hour   
## 0.00000 0.00000 0.00000 0.00000   
## New\_Invoice\_Month   
## 0.00000

**8. Number of transactions with missing CustomerID records by countries**

missing<-Online\_Retail\_new %>% filter(is.na(CustomerID)) %>% group\_by(Country) %>% summarise(count\_by\_countries = n())  
  
View(missing)

**9. On average, how often the costumers comeback to the website for their next shopping?**

difference\_days<-Online\_Retail\_new %>% select(CustomerID,New\_Invoice\_Date) %>% group\_by(CustomerID) %>% distinct(New\_Invoice\_Date) %>% arrange(desc(CustomerID)) %>% mutate(days = New\_Invoice\_Date-lag(New\_Invoice\_Date)) %>% filter(!is.na(days))   
View(difference\_days)  
mean(difference\_days$days)

## Time difference of 38.4875 days

# On an average,customers come back after 38 days to the website for their next shopping.

**10. Return rate for the French customers**

cancelled\_customers <- Online\_Retail\_new %>% filter(Country=='France',Quantity<0) %>% summarise(count = n())  
Total\_customers<- Online\_Retail\_new %>% filter(Country=='France') %>% count()  
  
return\_rate\_french\_cust=((cancelled\_customers/Total\_customers)\*100)  
  
head(return\_rate\_french\_cust)

## count  
## 1 1.741264

**11. Product that has generated the highest Revenue for the retailer**

item\_sum<-Online\_Retail\_new %>% group\_by(Description) %>% summarise(sum\_cnt = sum(TransactionValue)) %>% arrange(desc(sum\_cnt))  
head(item\_sum)

## # A tibble: 6 × 2  
## Description sum\_cnt  
## <chr> <dbl>  
## 1 DOTCOM POSTAGE 206245.  
## 2 REGENCY CAKESTAND 3 TIER 164762.  
## 3 WHITE HANGING HEART T-LIGHT HOLDER 99668.  
## 4 PARTY BUNTING 98303.  
## 5 JUMBO BAG RED RETROSPOT 92356.  
## 6 RABBIT NIGHT LIGHT 66757.

#DOTCOM POSTAGE generates highest revenue for the retailer

**12. Unique customers in the dataset**

unique\_cust<- Online\_Retail\_new %>% distinct(CustomerID) %>% summarise(ncount = n())  
head(unique\_cust)

## ncount  
## 1 4373

#There are 4373 unique customers in the dataset