Assignment 2

```
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((c
head(trans_countries)
```

```
## # A tibble: 4 x 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,Desc

head(Online_Retail_new)
```

```
##   InvoiceNo StockCode                         Description Quantity
## 1    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6
## 2    536365     71053                 WHITE METAL LANTERN        6
## 3    536365    84406B      CREAM CUPID HEARTS COAT HANGER        8
## 4    536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6
## 5    536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6
## 6    536365     22752        SET 7 BABUSHKA NESTING BOXES        2
##      InvoiceDate UnitPrice CustomerID        Country TransactionValue
## 1 12/1/2010 8:26      2.55      17850 United Kingdom            15.30
## 2 12/1/2010 8:26      3.39      17850 United Kingdom            20.34
## 3 12/1/2010 8:26      2.75      17850 United Kingdom            22.00
## 4 12/1/2010 8:26      3.39      17850 United Kingdom            20.34
## 5 12/1/2010 8:26      3.39      17850 United Kingdom            20.34
## 6 12/1/2010 8:26      7.65      17850 United Kingdom            15.30
```

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 = sur

head(Trans_value_countries)
```

```
## # A tibble: 6 x 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.
```

```
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

head(perc_transc)
```

```
## # A tibble: 6 x 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(TransactionVa
head(perc_trans_week)
```

```
## # A tibble: 6 x 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(Transactio
head(perc_trans_month)
```

```
## # A tibble: 6 x 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

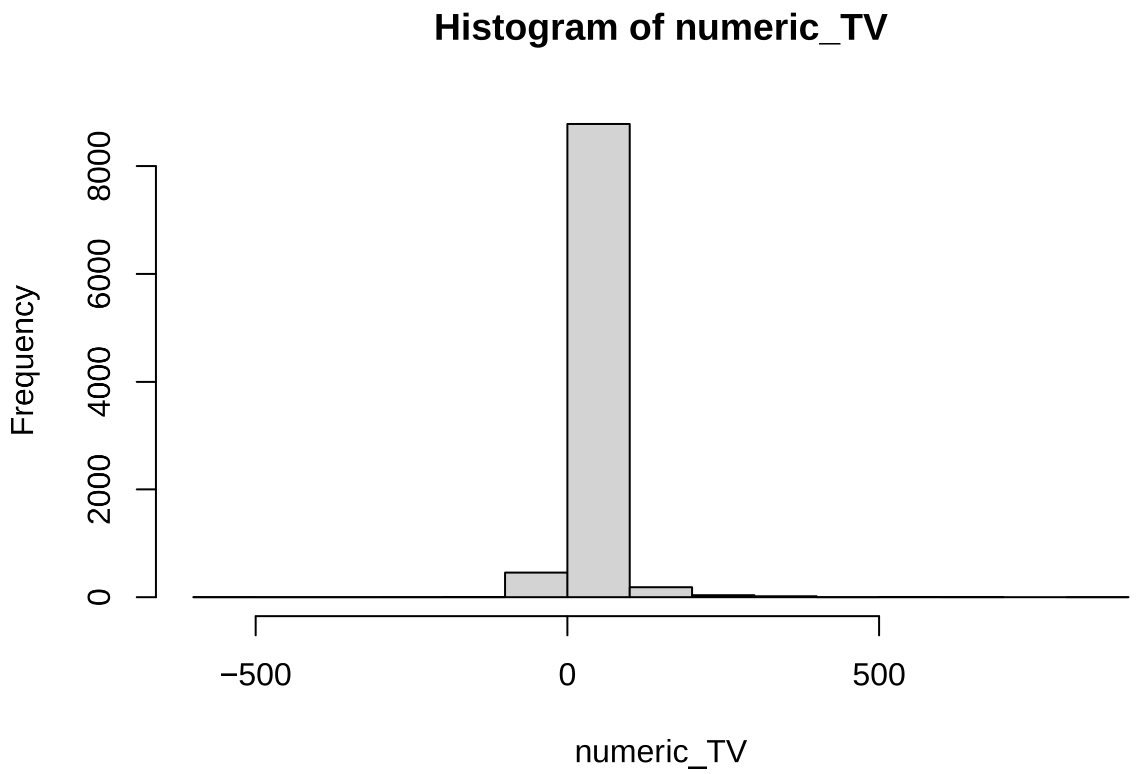
```
date_trans<- Online_Retail_new %>% filter(Country == 'Australia') %>% group_by(New_Invoice_Date) %>% su

head(date_trans)
```

```
## # A tibble: 6 x 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 x 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)) %>% a

head(cust_sum)
```

```
## # A tibble: 6 x 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_c

head(missing)
```

```
## # A tibble: 6 x 2
##   Country   count_by_countries
##   <chr>                  <int>
## 1 Bahrain                    2
## 2 EIRE                     711
## 3 France                    66
## 4 Hong Kong                288
## 5 Israel                    47
## 6 Portugal                  39
```

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) %>%
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

```
return_rate <- Online_Retail_new %>% filter(Country=='France')
cancelled_customers <- nrow(subset(return_rate,TransactionValue<0))
View(cancelled_customers)

return_rate_french_cust=(cancelled_customers/8556)

head(return_rate_french_cust)
```

```
## [1] 0.01741468
```

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

```
item_sum<-Online_Retail_new %>% group_by(Description) %>% summarise(sum_cnt = sum(TransactionValue)) %>
head(item_sum)
```

```
## # A tibble: 6 x 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
```