BA_Assignment 2

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()

getwd()

## [1] "/Users/thupiliabhinav/Desktop/BA/BA_Assignement 2"

setwd("/Users/thupiliabhinav/Desktop/BA/BA_Assignement 2")
assign_1 <- read.csv("Online_Retail.csv")
```

#1.Breakdown of the number of transactions by countries. Transactions in percentages. Only 1% of transactions.

```
ans1<- group_by(assign_1, Country)%>% count(Country)
ans1

## # A tibble: 38 × 2
## # Groups:   Country [38]
##    Country            n
##    <chr>          <int>
##  1 Australia       1259
##  2 Austria          401
##  3 Bahrain           19
##  4 Belgium         2069
##  5 Brazil            32
##  6 Canada           151
##  7 Channel Islands  758
##  8 Cyprus           622
##  9 Czech Republic    30
## 10 Denmark          389
## # … with 28 more rows

ans12<- ans1$n*100/sum(ans1$n)
ans12
```

```
##  [1]  0.232326830  0.073997664  0.003506124  0.381798420  0.005905050
##  [6]  0.027864457  0.139875883  0.114779419  0.005535985  0.071783270
## [11]  1.512431054  0.011256502  0.128250315  1.579047405  1.752139197
## [16]  0.026941793  0.053145454  0.033584975  0.054806250  0.148179860
## [21]  0.066062752  0.008303977  0.006458649  0.023435669  0.437527334
## [26]  0.200402651  0.062925694  0.280305365  0.010702904  0.001845328
## [31]  0.042258017  0.467421652  0.085254166  0.369434721  0.012548232
## [36] 91.431956288  0.082301641  0.053699053

ans123<-subset(ans12, ans12>1)
ans123

## [1]  1.512431  1.579047  1.752139 91.431956
```

#2.New variable “TransactionValue” and binding to the original dataframe.

```
TransactionValue<- assign_1$Quantity*assign_1$UnitPrice
b_ans1<-cbind(assign_1,TransactionValue)
head(b_ans1)

##   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.Breakdown of transaction values by countries. Total transaction exceeding 130,000 British Pound.

```
c_ans1<- summarise(group_by(b_ans1,Country), total.value=
sum(TransactionValue))
c_ans12 <- filter(c_ans1, total.value>130000)
c_ans12

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

#4.Converting ‘InvoiceDate’ into a POSIXlt object.

```
Temp=strptime(b_ans1$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"
```

#4.i. Day of the week and hour components dataframe with names as New_Invoice_Date, Invoice_Day_Week and New_Invoice_Hour:

```
b_ans1$New_Invoice_Date <- as.Date(Temp)
```

$4.ii.Date objects

```
b_ans1$New_Invoice_Date[20000]- b_ans1$New_Invoice_Date[10]

## Time difference of 8 days
```

#4.iii.Convert dates to days of the week

```
b_ans1$Invoice_Day_Week= weekdays(b_ans1$New_Invoice_Date)
```

#4.iv.Convert into a normal numerical value

```
b_ans1$New_Invoice_Hour = as.numeric(format(Temp, "%H"))
```

#4.v.Month as a separate numeric variable

```
b_ans1$New_Invoice_Month = as.numeric(format(Temp, "%m"))
```

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

```
n_transactions<- group_by(b_ans1, Invoice_Day_Week) %>% summarise(value=n())
%>% mutate(percentage=value/nrow(b_ans1)*100)
n_transactions

## # A tibble: 6 × 3
##   Invoice_Day_Week  value percentage
##   <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

```
n_transactions1 <- group_by(b_ans1, Invoice_Day_Week) %>% summarise(value=
sum(TransactionValue)) %>% mutate(total= value/sum(value)*100)
n_transactions1
```

```
## # A tibble: 6 × 3
##   Invoice_Day_Week    value total
##   <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

```
n_transactions2 <- group_by(b_ans1, New_Invoice_Month) %>% summarise(value=
sum(TransactionValue)) %>% mutate(total= value/sum(value)*100)
n_transactions2
```

```
## # A tibble: 12 × 3
##    New_Invoice_Month    value total
##                <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
##  7                 7  681300.  6.99
##  8                 8  682681.  7.00
##  9                 9 1019688. 10.5
## 10                10 1070705. 11.0
## 11                11 1461756. 15.0
## 12                12 1182625. 12.1
```

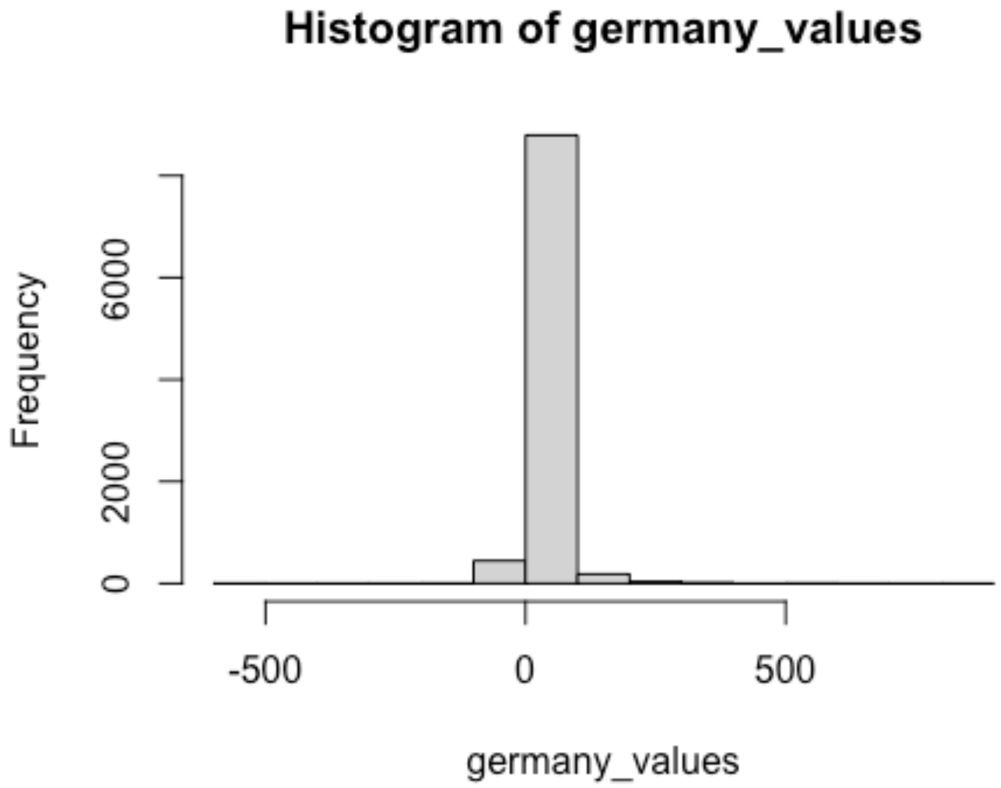
#4.d.The date with the highest number of transactions from Australia?

```
n_transactions3<- group_by(b_ans1, Country) %>% filter(Country=="Australia")
%>% group_by(New_Invoice_Date) %>% summarise(value= n()) %>%
arrange(desc(value))
n_transactions3
```

```
## # A tibble: 49 × 2
##    New_Invoice_Date value
##    <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
##  7 2011-02-15         69
##  8 2011-01-06         48
##  9 2011-07-14         35
## 10 2011-09-16         34
## # … with 39 more rows
```

#5

```
germany_values<- subset(b_ans1$TransactionValue, b_ans1$Country == 'Germany')
hist(germany_values)
```



#6.Customer had the highest number of transactions. Most valuable customer.

```
f_1 <-group_by(b_ans1,CustomerID) %>% select('CustomerID') %>%
na.omit(b_ans1) %>% summarise(value = n()) %>% arrange(desc(value))
f_1[which.max(f_1$value),]

## # A tibble: 1 × 2
##   CustomerID value
##        <int> <int>
## 1      17841  7983
```

```r
#Customer-ID 17841 has the highest number of transactions

f_ans<- summarise(group_by(b_ans1,CustomerID), Value= sum(TransactionValue))
%>% na.omit(b_ans1)
f_ans[which.max(f_ans$Value),]
```

```
## # A tibble: 1 × 2
##   CustomerID   Value
##        <int>   <dbl>
## 1      14646 279489.
```

```r
#The most valuable customer is Customer-ID-14646.
```

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

```r
missing_val<- colMeans(is.na(b_ans1)*100)
missing_val
```

```
##         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?

```r
missing_transaction <- b_ans1 %>% filter(is.na(CustomerID)) %>%
group_by(Country)
summary(missing_transaction$Country)
```

```
##    Length     Class      Mode
##    135080 character character
```

#10.What is the return rate for the French customers?

```r
returns <- filter(b_ans1,Country=="France", Quantity<0) %>% count()
total_value<- filter(b_ans1, Country=="France") %>% count()

percentage_returns<- returns/total_value*100
percentage_returns
```

```
##          n
## 1 1.741264
```

#11.Product that has generated the highest revenue for the retailer

```r
revenue<-b_ans1 %>% select(StockCode,TransactionValue) %>%
group_by(StockCode) %>% summarise(sum= sum(TransactionValue)) %>%
arrange(desc(sum))
revenue
```

```
## # A tibble: 4,070 × 2
##    StockCode     sum
##    <chr>       <dbl>
##  1 DOT       206245.
##  2 22423     164762.
##  3 47566      98303.
##  4 85123A     97894.
##  5 85099B     92356.
##  6 23084      66757.
##  7 POST       66231.
##  8 22086      63792.
##  9 84879      58960.
## 10 79321      53768.
## # … with 4,060 more rows
```

```
#DOT has the highest revenue generated with sum of 206245.48
```

#12.unique customers are represented in the dataset

```
unique_customer<- b_ans1%>% select(CustomerID) %>% unique() %>% count()
unique_customer
```

```
##      n
## 1 4373
```