R Notebook

Read the Data First and then set as working directory.

```
library(readr)
df <- read.csv("Online_Retail.csv")
```

Question 1

** Show the breakdown of the number of transactions by countries i.e., how many transactions are in the dataset for each country (consider all records including cancelled transactions). Show this in total number and also in percentage. Show only countries accounting for more than 1% of the total transactions. **

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df %>% group_by(Country) %>% summarise(TransCount=n(), Percetage_Trans=n()*100/nrow(df)) %>%
filter(Percetage_Trans>1)%>% as.data.frame()
```

```
##          Country TransCount Percetage_Trans
## 1           EIRE       8196        1.512431
## 2         France       8557        1.579047
## 3        Germany       9495        1.752139
## 4 United Kingdom     495478       91.431956
```

Question 2

** Create a new variable 'TransactionValue' that is the product of the exising 'Quantity' and 'UnitPrice' variables. Add this variable to the dataframe.**

```
df <- mutate(df,Transactionvalue = Quantity*UnitPrice)
```

Question 3

** Using the newly created variable, TransactionValue, show the breakdown of transaction values by countries i.e. how much money in total has been spent each country. Show this in total sum of transaction values. Show only countries with total transaction exceeding 130,000 British Pound. **

```
summary(df$Transactionvalue)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -168469.60       3.40       9.75      17.99      17.40  168469.60
```

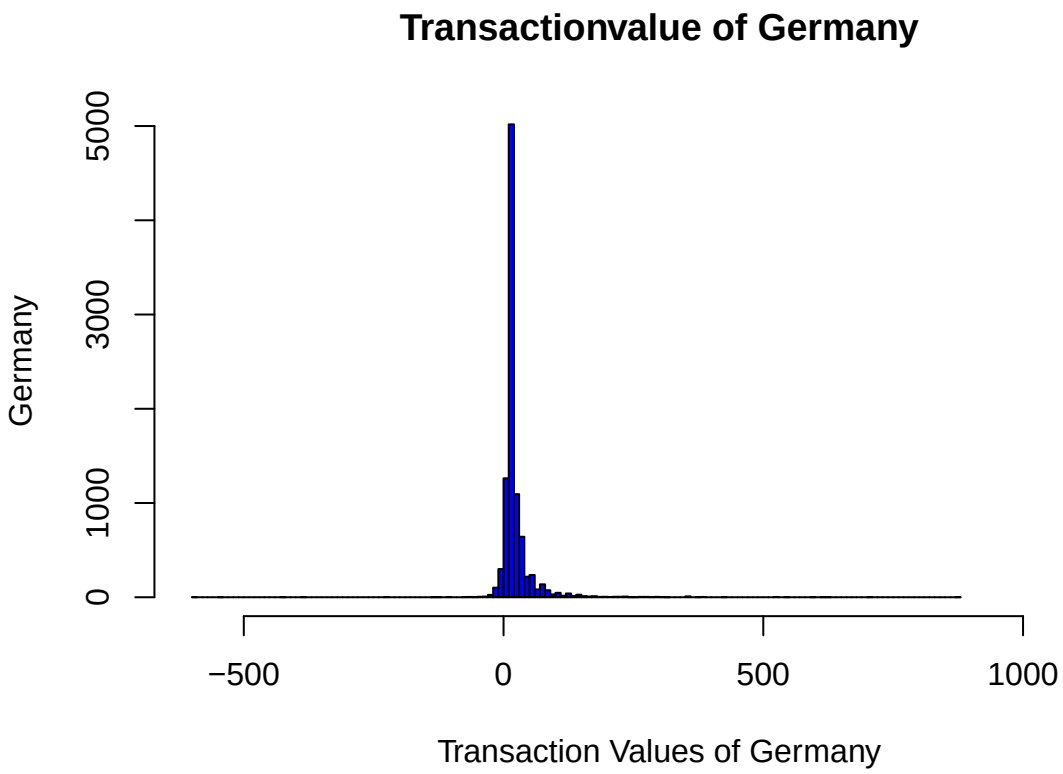
```
df %>% group_by(Country) %>% summarise(sum_Trans=sum(Transactionvalue)) %>% filter(sum_Trans>130000)
```

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

Question 5

** Plot the histogram of transaction values from Germany. **

```
Filtered_data = subset(df$Transactionvalue, df$Country == "Germany")
hist(Filtered_data, xlim = c (-600, 1200), breaks = 120,
main="Transactionvalue of Germany",xlab="Transaction Values of Germany",
     ylab = "Germany", col="blue", pch=16)
```



Question 6

** Which customer had the highest number of transactions? Which customer is most valuable? **

```
Temp_1=group_by(df, CustomerID)
Temp_2=summarise(Temp_1, count=n())
Temp_3=arrange(Temp_2,desc(count))
head(as.data.frame(Temp_3))
```

```
##   CustomerID  count
## 1         NA 135080
## 2      17841   7983
## 3      14911   5903
## 4      14096   5128
## 5      12748   4642
## 6      14606   2782
```

```
Temp_1=group_by(df, CustomerID)
Temp_2=summarise(Temp_1, Sum_value=sum(Transactionvalue))
Temp_3=arrange(Temp_2,desc(Sum_value))
head(as.data.frame(Temp_3))
```

```
##   CustomerID Sum_value
```

```
## 1         NA 1447682.1
## 2      14646  279489.0
## 3      18102  256438.5
## 4      17450  187482.2
## 5      14911  132572.6
## 6      12415  123725.4
```

Question 7

** Calculate the percentage of missing values for each variable in the dataset.**

```
missing_values <- colMeans(is.na(df)*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
##          0.00000
```

Question 8

** What are the number of transactions with missing CustomerID records by countries?**

```
df %>% group_by(Country) %>% filter(is.na(CustomerID))
```

```
## # A tibble: 135,080 x 9
## # Groups:   Country [9]
##    InvoiceNo StockCode Description     Quantity InvoiceDate UnitPrice CustomerID
##    <chr>     <chr>     <chr>              <int> <chr>           <dbl>      <int>
##  1 536414    22139     ""                    56 12/1/2010 ~      0            NA
##  2 536544    21773     "DECORATIVE RO~        1 12/1/2010 ~      2.51         NA
##  3 536544    21774     "DECORATIVE CA~        2 12/1/2010 ~      2.51         NA
##  4 536544    21786     "POLKADOT RAIN~        4 12/1/2010 ~      0.85         NA
##  5 536544    21787     "RAIN PONCHO R~        2 12/1/2010 ~      1.66         NA
##  6 536544    21790     "VINTAGE SNAP ~        9 12/1/2010 ~      1.66         NA
##  7 536544    21791     "VINTAGE HEADS~        2 12/1/2010 ~      2.51         NA
##  8 536544    21801     "CHRISTMAS TRE~       10 12/1/2010 ~      0.43         NA
##  9 536544    21802     "CHRISTMAS TRE~        9 12/1/2010 ~      0.43         NA
## 10 536544    21803     "CHRISTMAS TRE~       11 12/1/2010 ~      0.43         NA
## # ... with 135,070 more rows, and 2 more variables: Country <chr>,
## #   Transactionvalue <dbl>
```

```
summary(df$Country)
```

```
##    Length     Class      Mode
##    541909 character character
```

Question 10

** In the retail sector, it is very important to understand the return rate of the goods purchased by customers. In this example, we can define this quantity, simply, as the ratio of the number of transactions cancelled (regardless of the transaction value) over the total number of transactions. With this definition, what is the return rate for the French customers? Consider the cancelled transactions as those where the 'Quantity' variable has a negative value. **

```
French_orders <- filter(df,Country=="France")
French_cancelled_orders <- filter(French_orders, Quantity < 0)
nrow(French_cancelled_orders)*100/nrow(French_orders)
```

```
## [1] 1.741264
```

Question 11

** What is the product that has generated the highest revenue for the retailer? **

```
Temp_1= group_by(df, StockCode)
Temp_2=summarise(Temp_1, Sum_Trans=sum(Transactionvalue))
Temp_3=arrange(Temp_2, desc(Sum_Trans))
head(as.data.frame(Temp_3))
```

```
##   StockCode Sum_Trans
## 1       DOT 206245.48
## 2     22423 164762.19
## 3     47566  98302.98
## 4    85123A  97894.50
## 5    85099B  92356.03
## 6     23084  66756.59
```

Question 12

** How many unique customers are represented in the dataset? You can use unique() and length()functions. **

```
unique_customers <- unique(df$CustomerID)
length(unique_customers)
```

```
## [1] 4373
```