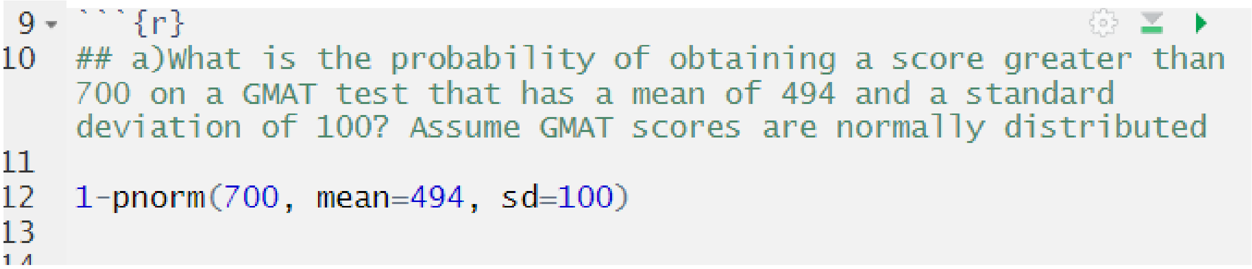
# MIS-64036: Business Analytics

## Assignment I

## Part A) Descriptive Statistics & Normal Distributions

1. **a) What is the probability of obtaining a score greater than 700 on a GMAT test that has a mean of 494 and a standard deviation of 100? Assume GMAT scores are normally distributed (5 marks).**

**Response:** [1] 0.01969927



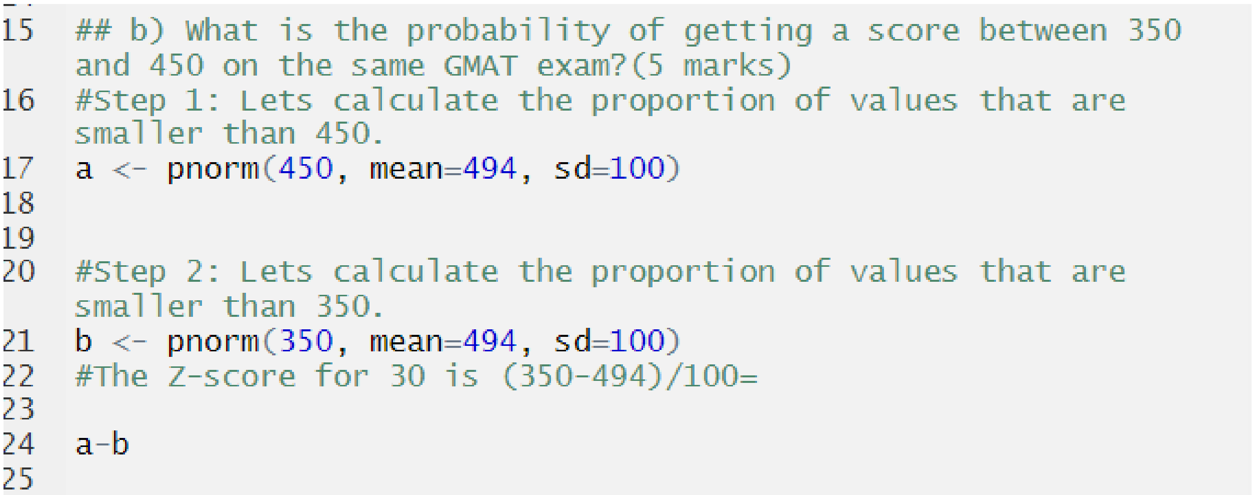
~~~
```{r}
## a)What is the probability of obtaining a score greater than 700 on a GMAT test that has a mean of 494 and a standard deviation of 100? Assume GMAT scores are normally distributed

1-pnorm(700, mean=494, sd=100)

~~~

**b) What is the probability of getting a score between 350 and 450 on the same GMAT exam?(5 marks)**

**Response:**



```
## b) What is the probability of getting a score between 350 and 450 on the same GMAT exam?(5 marks)
#Step 1: Lets calculate the proportion of values that are smaller than 450.
a <- pnorm(450, mean=494, sd=100)


#Step 2: Lets calculate the proportion of values that are smaller than 350.
b <- pnorm(350, mean=494, sd=100)
#The Z-score for 30 is (350-494)/100=

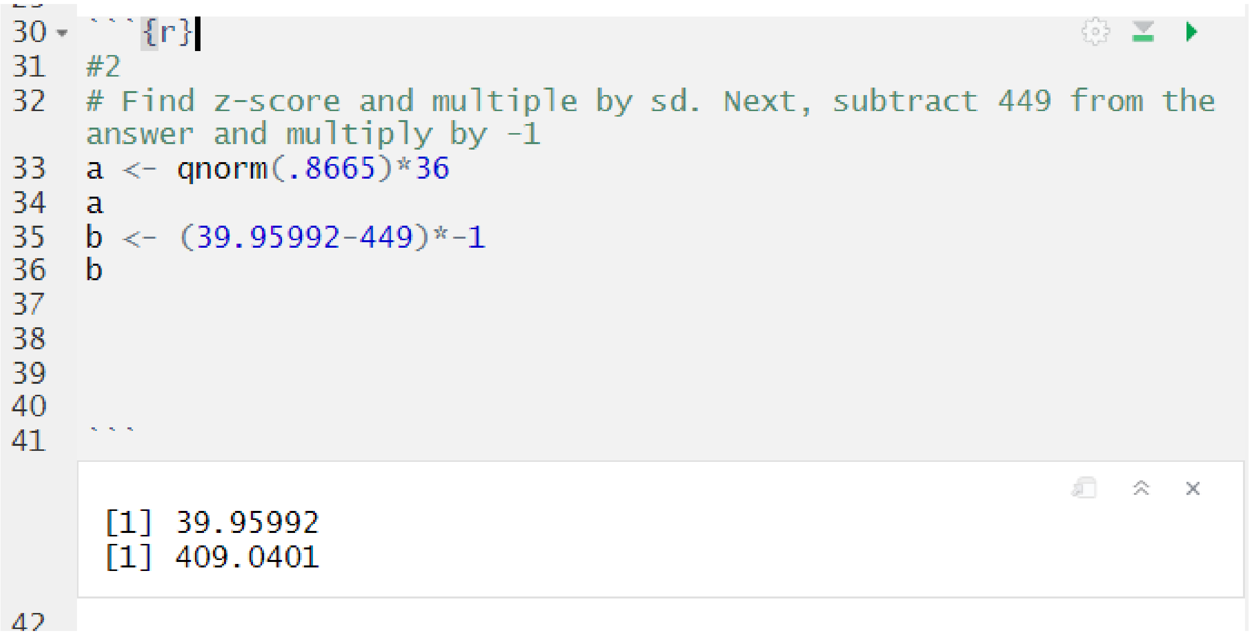
a-b

```

```
[1] 0.2550349
```

2. **Runzheimer International publishes business travel costs for various cities throughout the world. In particular, they publish per diem totals, which represent the average costs for the typical business traveler including three meals a day in business-class restaurants and single-rate lodging in business-class hotels and motels. If 86.65% of the per diem costs in Buenos Aires, Argentina, are less than $449 and if the standard deviation of per diem costs is $36, what is the average per diem cost in Buenos Aires? Assume that per diem costs are normally distributed (10 marks)**

**Response: [1] 409.0401**



~~~
```{r}
#2
# Find z-score and multiple by sd. Next, subtract 449 from the answer and multiply by -1
a <- qnorm(.8665)*36
a
b <- (39.95992-449)*-1
b




```
~~~

```
[1] 39.95992
[1] 409.0401
```

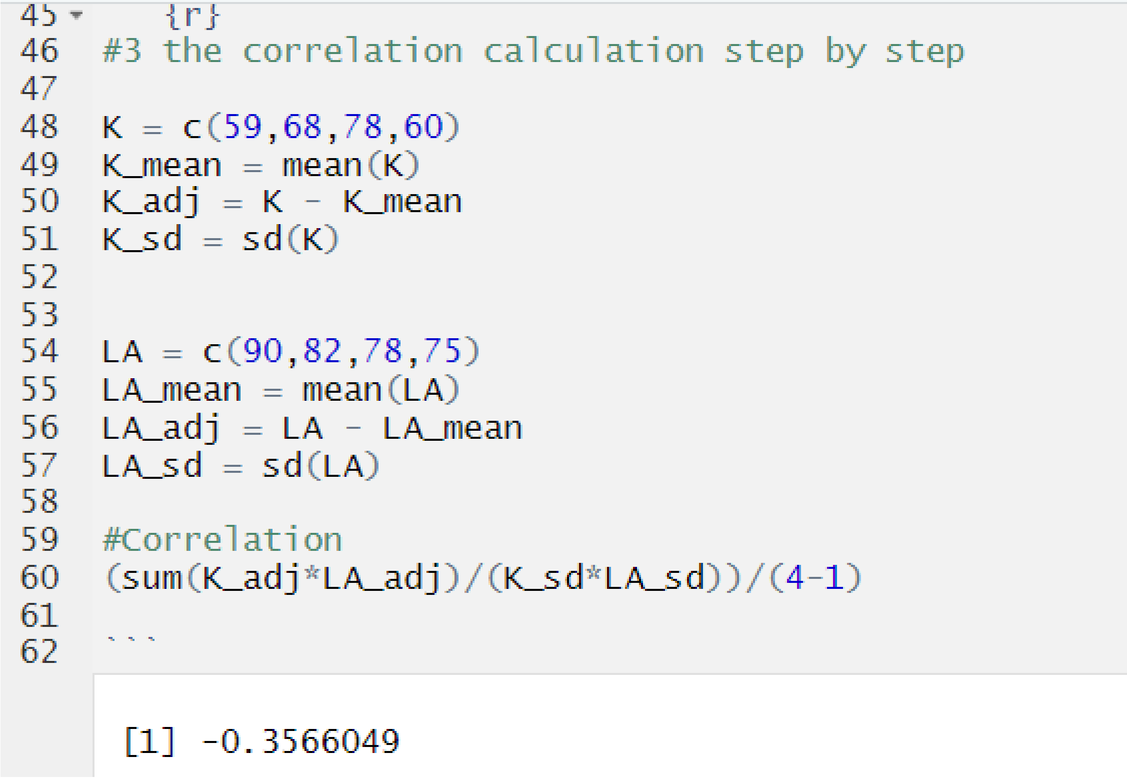
3. **Chris is interested in understanding the correlation between temperature in Kent, OH and Los Angeles, CA. He has got the following data for September 2017 from Alpha Knowledgebase. (5 marks)**

   **He has sampled the mid-day temperature for days from Sep 2 to Sep 6 as follows:**

```
Kent=c(59, 68, 78, 60)
Los_Angeles=c(90, 82, 78, 75)
```

**Calculate the correlation (Pearson Correlation Coefficient) between the temperatures of the two cities without using any R commands i.e. calculate step by step.**

**Response:** [1] -0.3566049



```
   {r}
#3 the correlation calculation step by step

K = c(59,68,78,60)
K_mean = mean(K)
K_adj = K - K_mean
K_sd = sd(K)


LA = c(90,82,78,75)
LA_mean = mean(LA)
LA_adj = LA - LA_mean
LA_sd = sd(LA)

#Correlation
(sum(K_adj*LA_adj)/(K_sd*LA_sd))/(4-1)

```
```

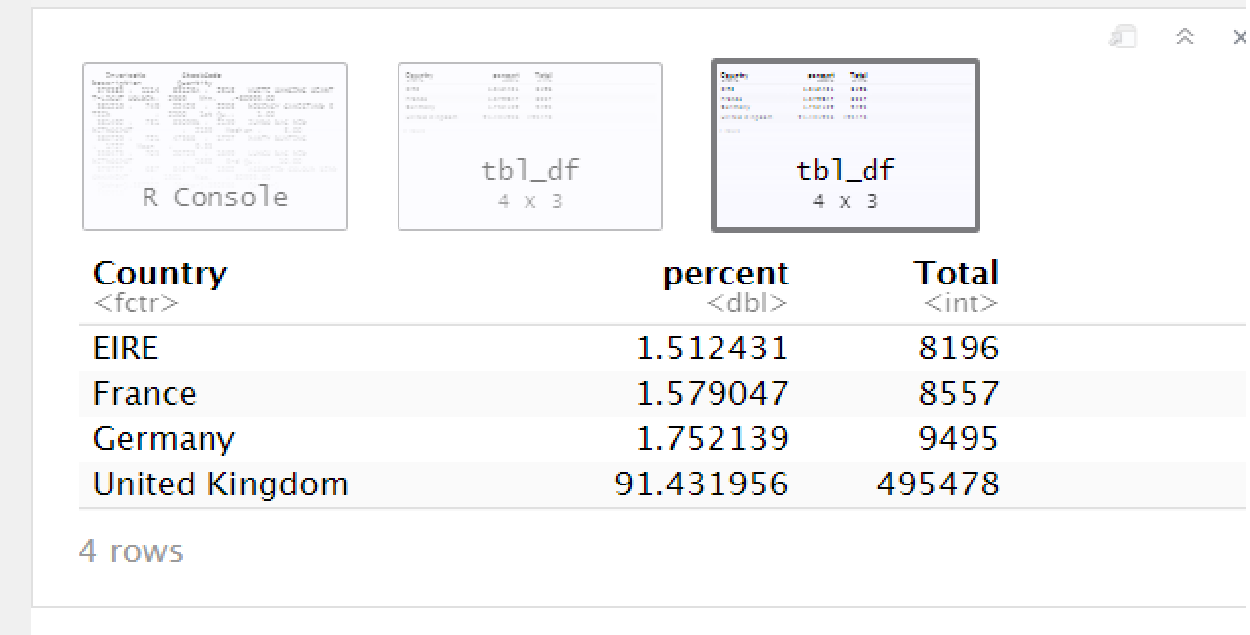
```
[1] -0.3566049
```
```

## Part B) Data Wrangling

For the questions in this part, you need to use the 'Online Retail' dataset which can be downloaded in CSV format from the course portal under the assignment folder. This is a transnational data set which contains all the transactions occurring between 01 Dec 2010 and 09 Dec 2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

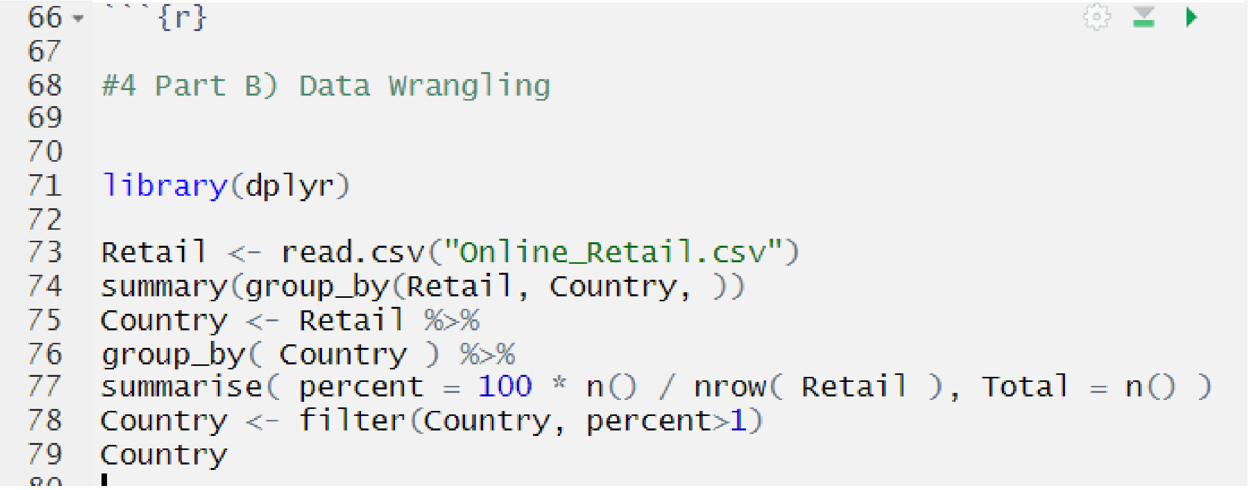
4. Show the breakdown of the number of transactions by countries i.e. how many transactions are in the dataset for each country (consider all records including cancelled transactions). Show this in total number and also in percentage. Show only countries accounting for more than 1% of the total transactions. (5 marks)

**Response:**



| Country <br>&lt;fctr&gt; | percent <br>&lt;dbl&gt; | Total <br>&lt;int&gt; |
|---|---|---|
| EIRE | 1.512431 | 8196 |
| France | 1.579047 | 8557 |
| Germany | 1.752139 | 9495 |
| United Kingdom | 91.431956 | 495478 |

4 rows



~~~
```{r}

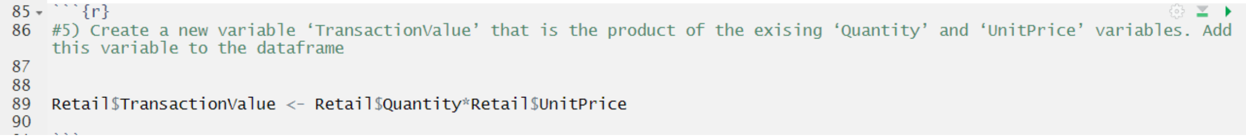
#4 Part B) Data Wrangling


library(dplyr)

Retail <- read.csv("Online_Retail.csv")
summary(group_by(Retail, Country, ))
Country <- Retail %>%
group_by( Country ) %>%
summarise( percent = 100 * n() / nrow( Retail ), Total = n() )
Country <- filter(Country, percent>1)
Country
~~~

5. **Create a new variable ‘TransactionValue’ that is the product of the exising ‘Quantity’ and ‘UnitPrice’ variables. Add this variable to the dataframe. (5 marks)**

**Response:**



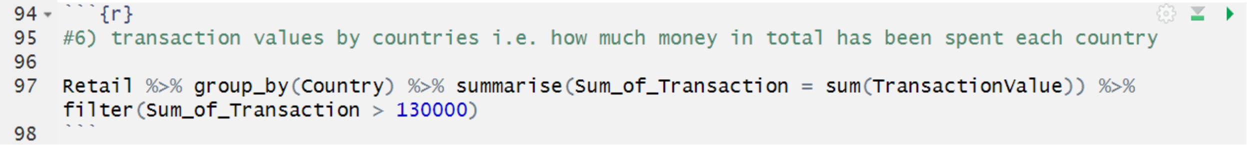
~~~
```{r}
#5) Create a new variable ‘TransactionValue’ that is the product of the exising ‘Quantity’ and ‘UnitPrice’ variables. Add this variable to the dataframe


Retail$TransactionValue <- Retail$Quantity*Retail$UnitPrice

```
~~~

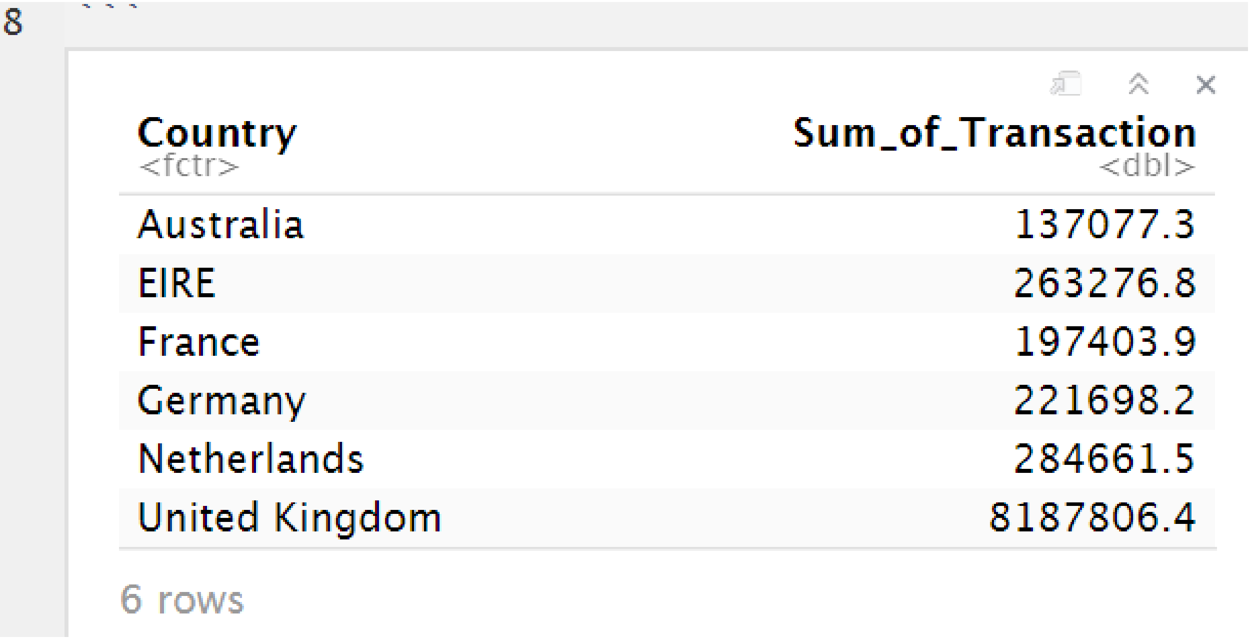
6. **Using the newly created variable, TransactionValue, show the breakdown of transaction values by countries i.e. how much money in total has been spent each country. Show this in total sum of transaction values. Show only countries with total transaction exceeding 130,000 British Pound. (10 marks)**

**Response:**



~~~
```{r}
#6) transaction values by countries i.e. how much money in total has been spent each country

Retail %>% group_by(Country) %>% summarise(Sum_of_Transaction = sum(TransactionValue)) %>% filter(Sum_of_Transaction > 130000)
```
~~~



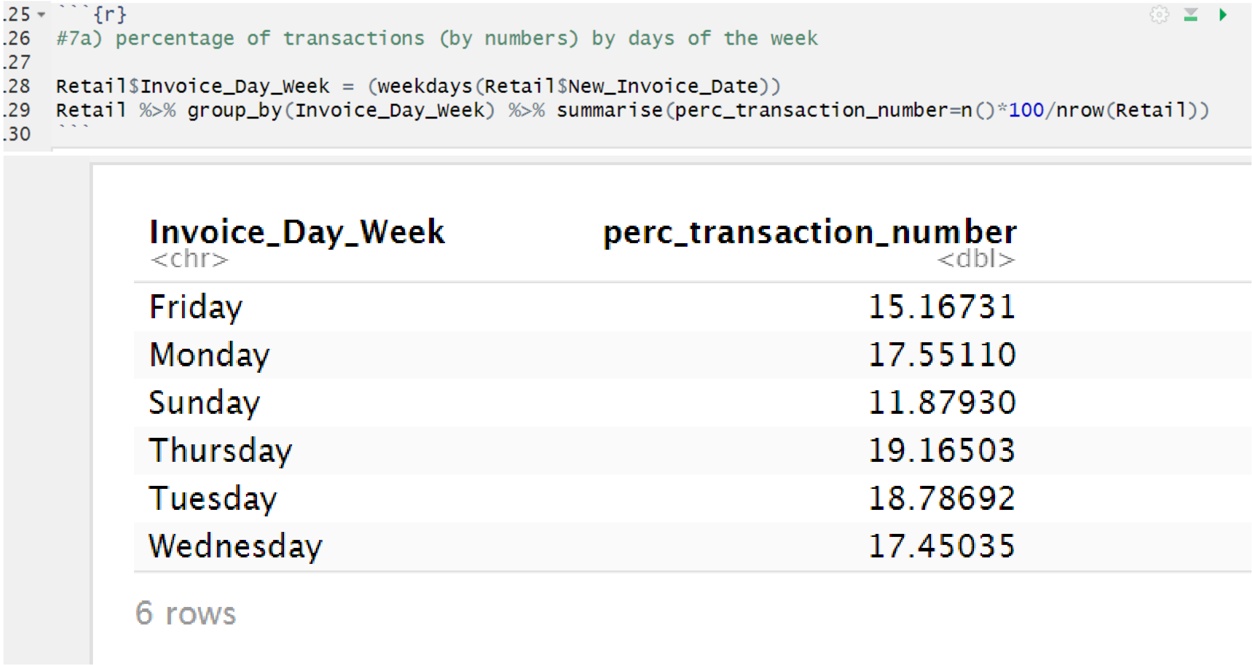
~~~
```
~~~

| Country<br><fctr> | Sum_of_Transaction<br><dbl> |
|---|---:|
| Australia | 137077.3 |
| EIRE | 263276.8 |
| France | 197403.9 |
| Germany | 221698.2 |
| Netherlands | 284661.5 |
| United Kingdom | 8187806.4 |

6 rows

7. This is an optional question which carries additional marks (golden questions). In this question, we are dealing with the InvoiceDate variable. The variable is read as a categorical when you read data from the file. Now we need to explicitly instruct R to interpret this as a Date variable. "POSIXlt" and "POSIXct" are two powerful object classes in R to deal with date and time. Show the percentage of transactions (by numbers) by days of the week (extra 2 marks)

**Response:**



~~~
```{r}
#7a) percentage of transactions (by numbers) by days of the week

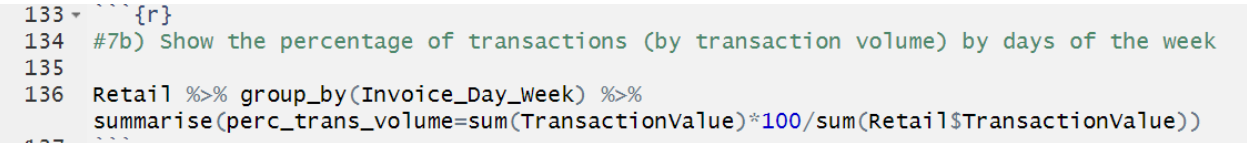
Retail$Invoice_Day_Week = (weekdays(Retail$New_Invoice_Date))
Retail %>% group_by(Invoice_Day_Week) %>% summarise(perc_transaction_number=n()*100/nrow(Retail))
```
~~~

| Invoice_Day_Week<br><chr> | perc_transaction_number<br><dbl> |
|---|---:|
| Friday | 15.16731 |
| Monday | 17.55110 |
| Sunday | 11.87930 |
| Thursday | 19.16503 |
| Tuesday | 18.78692 |
| Wednesday | 17.45035 |

6 rows

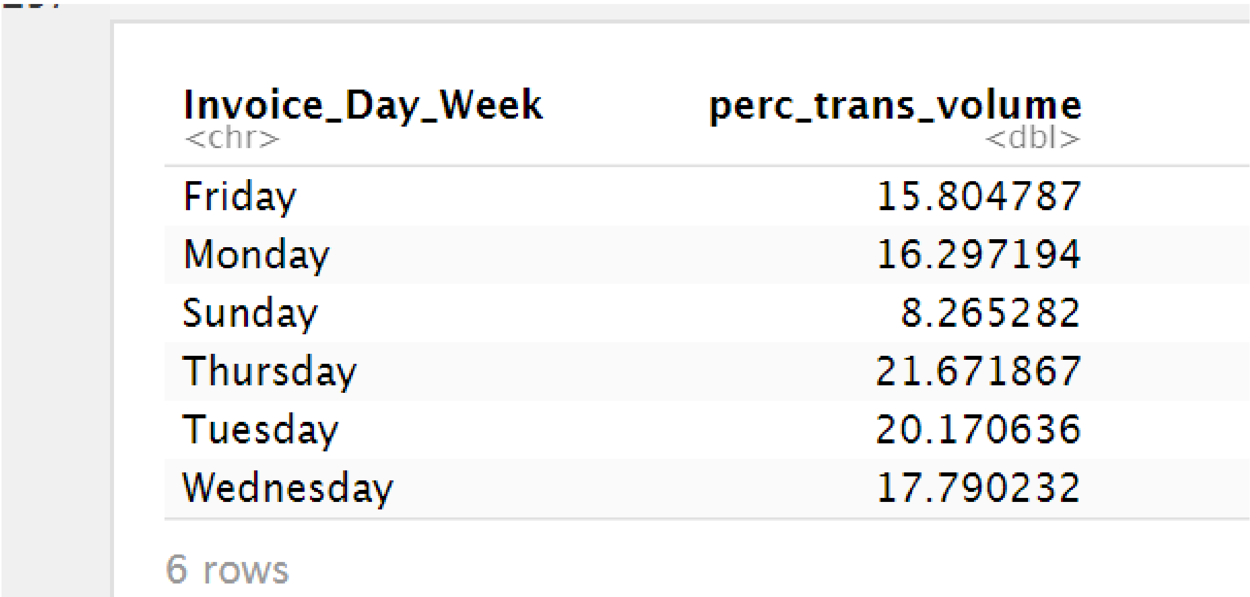
a) **Show the percentage of transactions (by transaction volume) by days of the week (extra 1 marks)**

**Response:**



~~~
```{r}
#7b) Show the percentage of transactions (by transaction volume) by days of the week

Retail %>% group_by(Invoice_Day_Week) %>%
summarise(perc_trans_volume=sum(TransactionValue)*100/sum(Retail$TransactionValue))
```
~~~

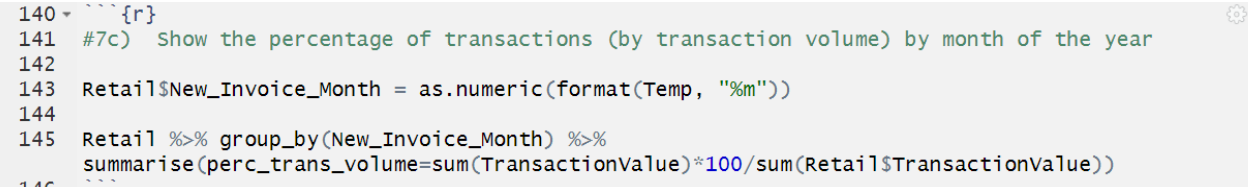


| Invoice_Day_Week<br>&lt;chr&gt; | perc_trans_volume<br>&lt;dbl&gt; |
|---|---|
| Friday | 15.804787 |
| Monday | 16.297194 |
| Sunday | 8.265282 |
| Thursday | 21.671867 |
| Tuesday | 20.170636 |
| Wednesday | 17.790232 |

6 rows

b) **Show the percentage of transactions (by transaction volume) by month of the year (extra 1 marks)**

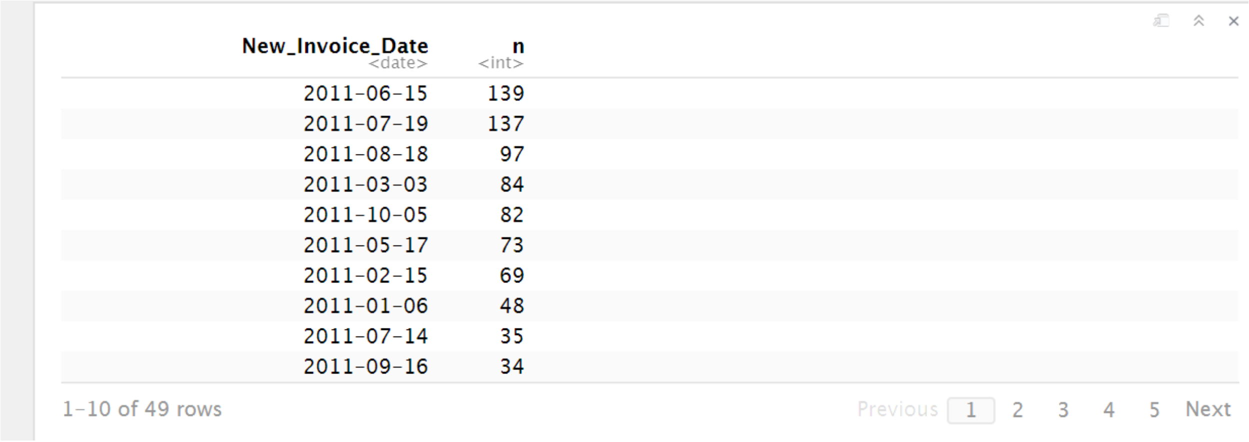
**Response: This is a 49 rows output so please refer to my RMD file in github for complete output table.**



~~~
```{r}
#7c)  Show the percentage of transactions (by transaction volume) by month of the year

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

Retail %>% group_by(New_Invoice_Month) %>%
summarise(perc_trans_volume=sum(TransactionValue)*100/sum(Retail$TransactionValue))
```
~~~

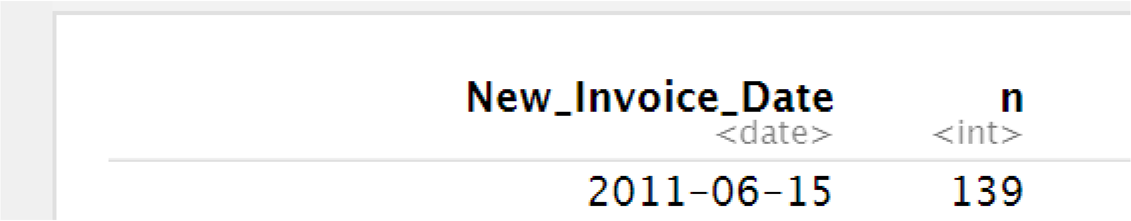


| New_Invoice_Date<br>&lt;date&gt; | n<br>&lt;int&gt; |
|---|---|
| 2011-06-15 | 139 |
| 2011-07-19 | 137 |
| 2011-08-18 | 97 |
| 2011-03-03 | 84 |
| 2011-10-05 | 82 |
| 2011-05-17 | 73 |
| 2011-02-15 | 69 |
| 2011-01-06 | 48 |
| 2011-07-14 | 35 |
| 2011-09-16 | 34 |

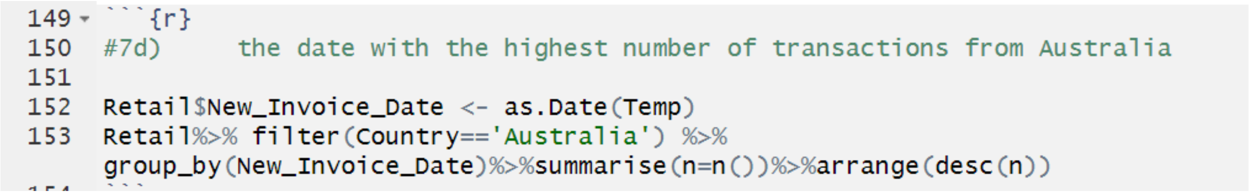
1-10 of 49 rows    Previous 1 2 3 4 5 Next

c) What was the date with the highest number of transactions from Australia? (3 marks)

**Response: the date with the highest number of transactions from Australia is**

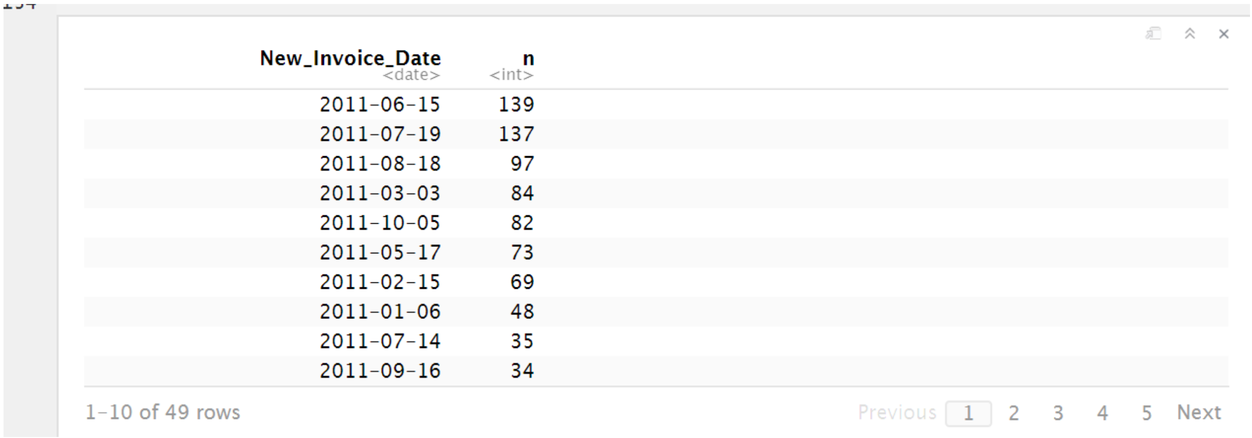


| New_Invoice_Date<br>&lt;date&gt; | n<br>&lt;int&gt; |
|---|---|
| 2011-06-15 | 139 |



~~~
```{r}
#7d)     the date with the highest number of transactions from Australia

Retail$New_Invoice_Date <- as.Date(Temp)
Retail%>% filter(Country=='Australia') %>%
group_by(New_Invoice_Date)%>%summarise(n=n())%>%arrange(desc(n))
```
~~~

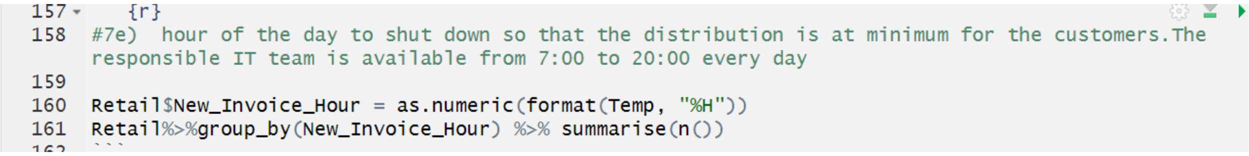


| New_Invoice_Date<br>&lt;date&gt; | n<br>&lt;int&gt; |
|---|---|
| 2011-06-15 | 139 |
| 2011-07-19 | 137 |
| 2011-08-18 | 97 |
| 2011-03-03 | 84 |
| 2011-10-05 | 82 |
| 2011-05-17 | 73 |
| 2011-02-15 | 69 |
| 2011-01-06 | 48 |
| 2011-07-14 | 35 |
| 2011-09-16 | 34 |

1-10 of 49 rows    Previous 1 2 3 4 5 Next

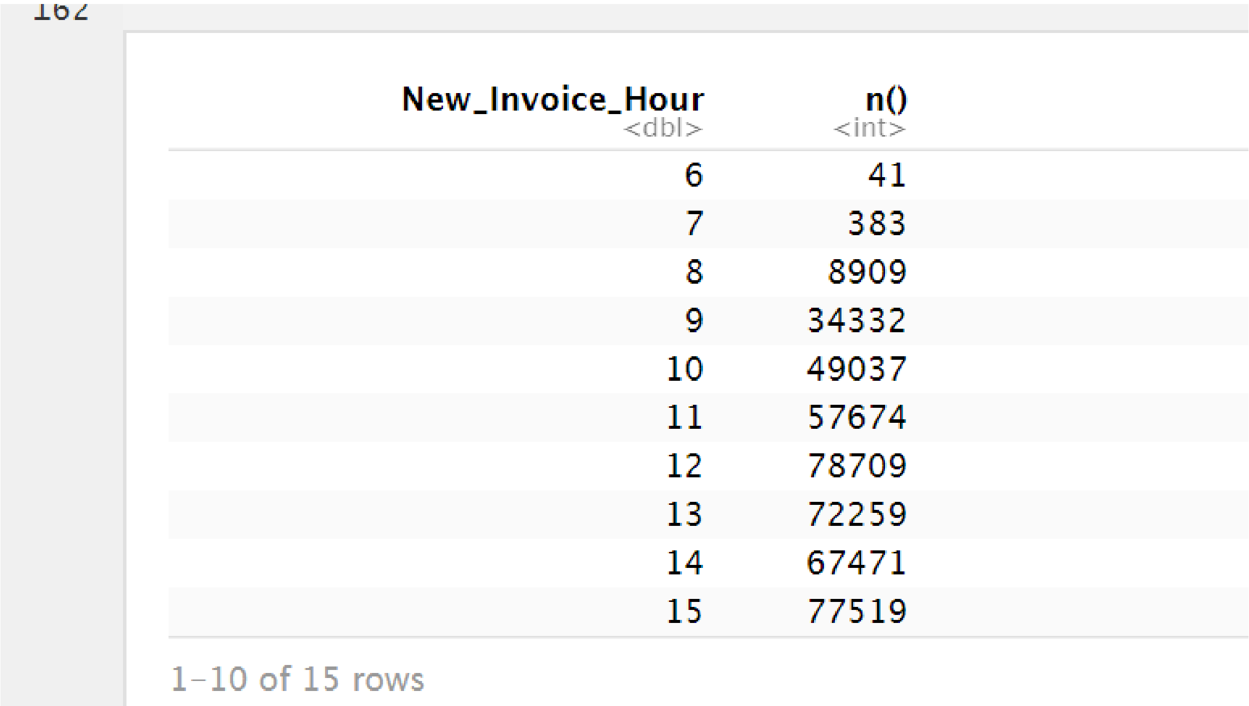
d) The company needs to shut down the website for two consecutive hours for maintenance. What would be the hour of the day to start this so that the distribution is at minimum for the customers? The responsible IT team is available from 7:00 to 20:00 every day(3 marks)

**Response:**



~~~~
   {r}
#7e)  hour of the day to shut down so that the distribution is at minimum for the customers.The responsible IT team is available from 7:00 to 20:00 every day

Retail$New_Invoice_Hour = as.numeric(format(Temp, "%H"))
Retail%>%group_by(New_Invoice_Hour) %>% summarise(n())
```
~~~~

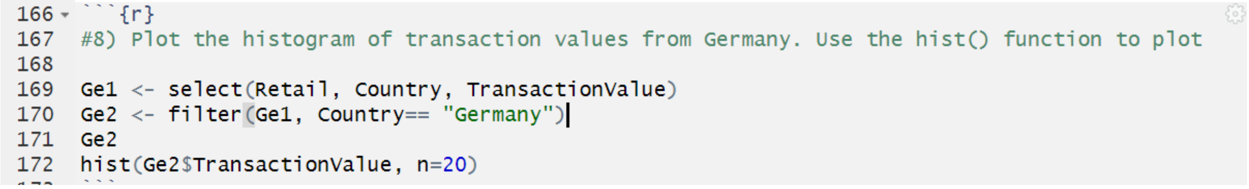


| New_Invoice_Hour<br><dbl> | n()<br><int> |
|---:|---:|
| 6 | 41 |
| 7 | 383 |
| 8 | 8909 |
| 9 | 34332 |
| 10 | 49037 |
| 11 | 57674 |
| 12 | 78709 |
| 13 | 72259 |
| 14 | 67471 |
| 15 | 77519 |

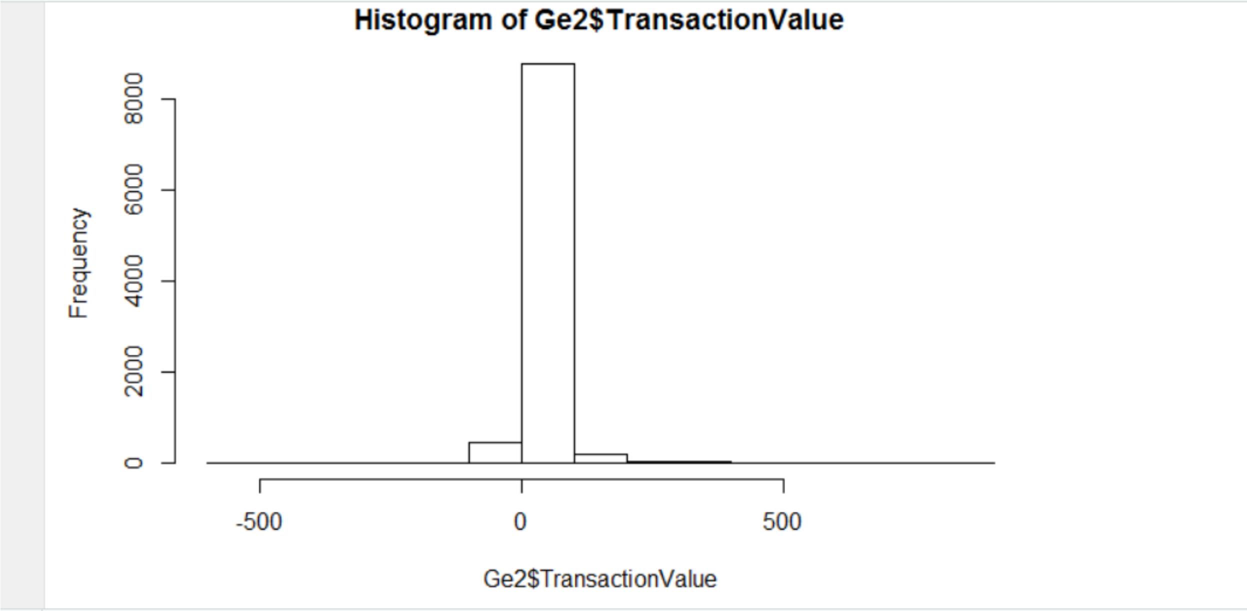
1–10 of 15 rows

8. **Plot the histogram of transaction values from Germany. Use the hist() function to plot. (5 marks)**

**Response:**

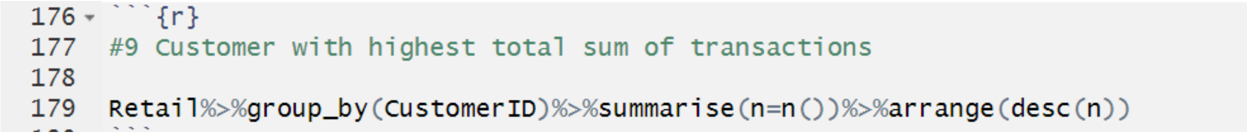


~~~~
```{r}
#8) Plot the histogram of transaction values from Germany. Use the hist() function to plot

Ge1 <- select(Retail, Country, TransactionValue)
Ge2 <- filter(Ge1, Country== "Germany")
Ge2
hist(Ge2$TransactionValue, n=20)
```
~~~~

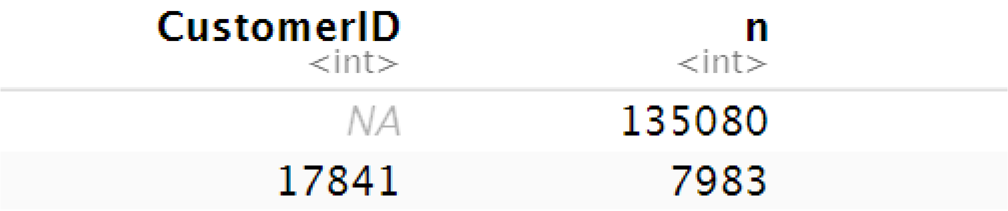


9. **Which customer had the highest number of transactions? Which customer is most valuable (i.e. highest total sum of transactions)? (10 marks)**

**Response:**



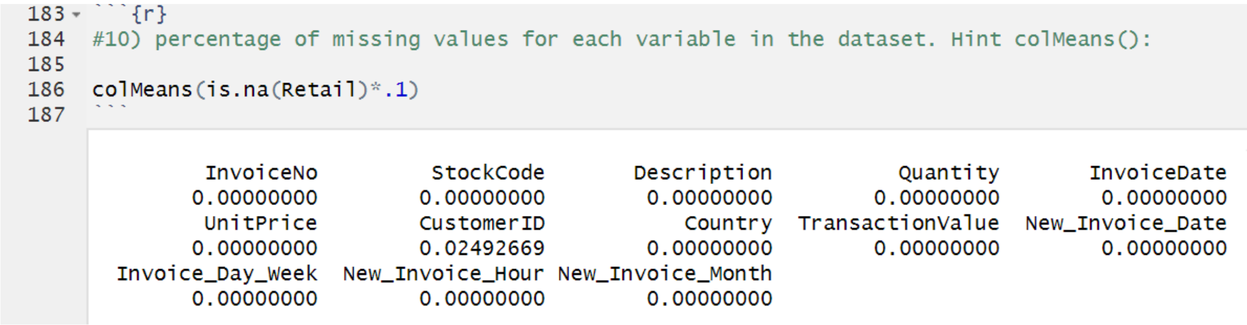
~~~
```{r}
#9 Customer with highest total sum of transactions

Retail%>%group_by(CustomerID)%>%summarise(n=n())%>%arrange(desc(n))
~~~



```
CustomerID        n
     <int>    <int>
        NA   135080
     17841     7983
```

10. **Calculate the percentage of missing values for each variable in the dataset (5 marks). Hint colMeans():**

**Response:**

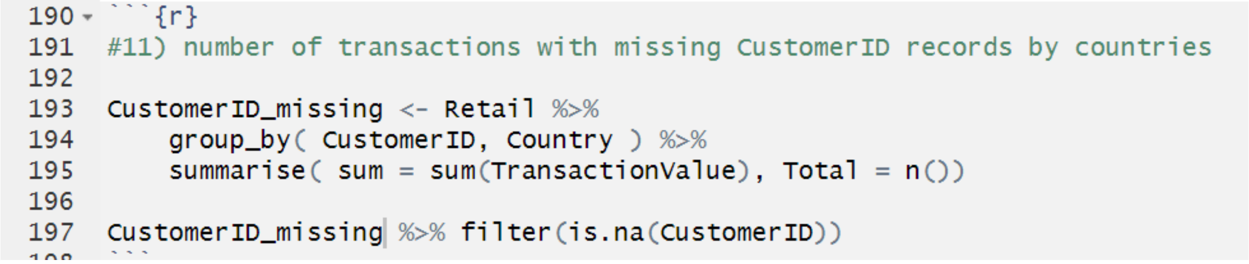


~~~
```{r}
#10) percentage of missing values for each variable in the dataset. Hint colMeans():

colMeans(is.na(Retail)*.1)
```
~~~

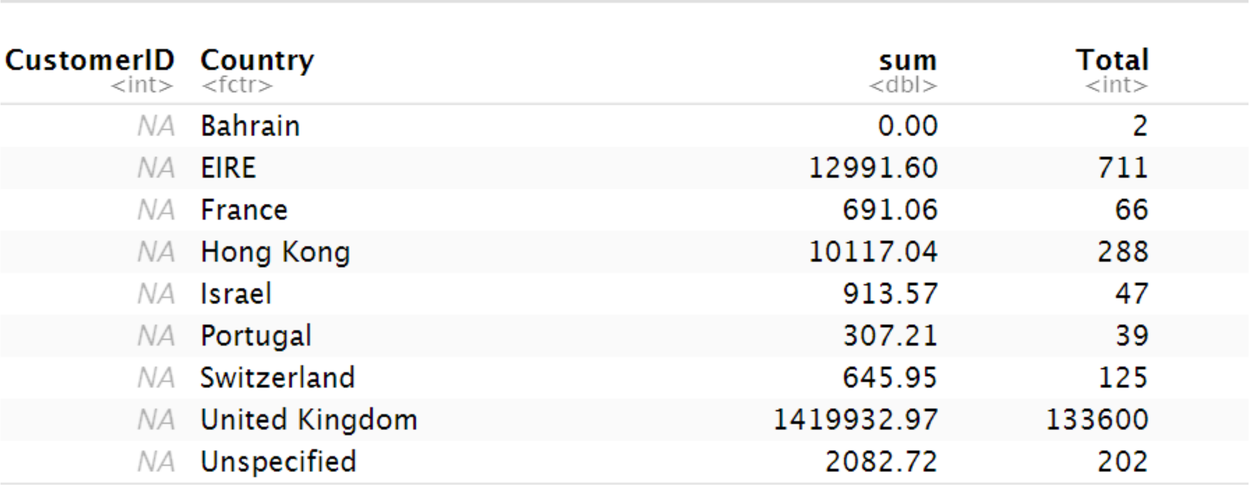
```
         InvoiceNo          StockCode        Description           Quantity        InvoiceDate 
        0.00000000         0.00000000         0.00000000         0.00000000         0.00000000 
         UnitPrice         CustomerID            Country   TransactionValue   New_Invoice_Date 
        0.00000000         0.02492669         0.00000000         0.00000000         0.00000000 
  Invoice_Day_Week   New_Invoice_Hour  New_Invoice_Month 
        0.00000000         0.00000000         0.00000000 
```

11. **What are the number of transactions with missing CustomerID records by countries? (10 marks)**

**Response:**



~~~
```{r}
#11) number of transactions with missing CustomerID records by countries

CustomerID_missing <- Retail %>%
    group_by( CustomerID, Country ) %>%
    summarise( sum = sum(TransactionValue), Total = n())

CustomerID_missing %>% filter(is.na(CustomerID))
```
~~~



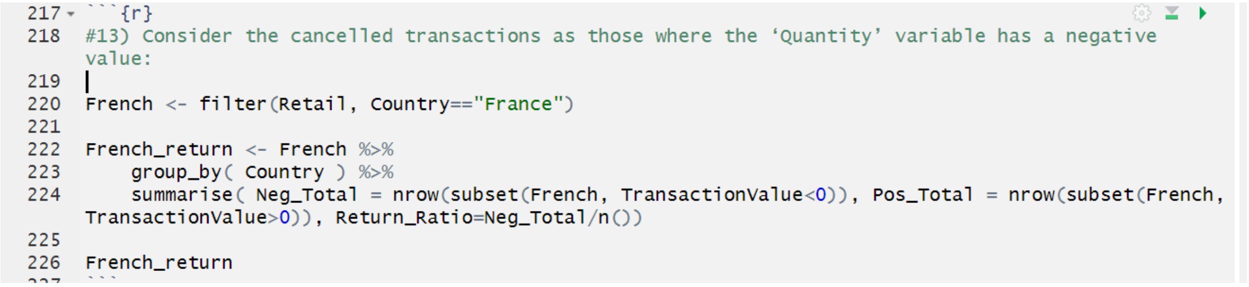
| CustomerID<br>&lt;int&gt; | Country<br>&lt;fctr&gt; | sum<br>&lt;dbl&gt; | Total<br>&lt;int&gt; |
|---|---|---|---|
| NA | Bahrain | 0.00 | 2 |
| NA | EIRE | 12991.60 | 711 |
| NA | France | 691.06 | 66 |
| NA | Hong Kong | 10117.04 | 288 |
| NA | Israel | 913.57 | 47 |
| NA | Portugal | 307.21 | 39 |
| NA | Switzerland | 645.95 | 125 |
| NA | United Kingdom | 1419932.97 | 133600 |
| NA | Unspecified | 2082.72 | 202 |

12. **On average, how often the costumers comeback to the website for their next shopping? (i.e. what is the average number of days between consecutive shopping) (Optional/Golden question: 18 additional marks!) Hint: 1. A close approximation is also acceptable and you may find diff() function useful.**

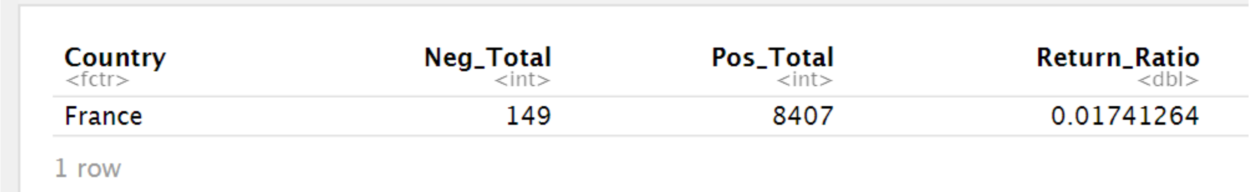
**Response: I couldn't fix this.**

13. 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? (10 marks). Consider the cancelled transactions as those where the 'Quantity' variable has a negative value.

**Response:**



~~~
```{r}
#13) Consider the cancelled transactions as those where the ‘Quantity’ variable has a negative
value:

French <- filter(Retail, Country=="France")

French_return <- French %>%
    group_by( Country ) %>%
    summarise( Neg_Total = nrow(subset(French, TransactionValue<0)), Pos_Total = nrow(subset(French,
TransactionValue>0)), Return_Ratio=Neg_Total/n())

French_return
```
~~~

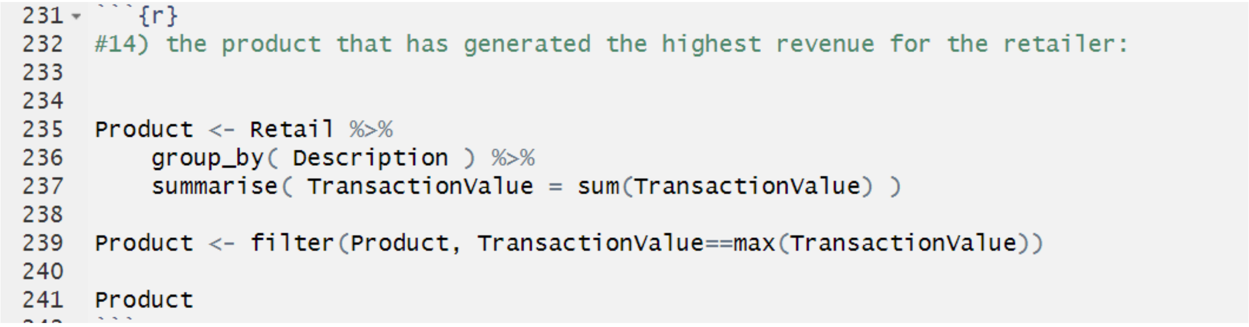


| Country<br>&lt;fctr&gt; | Neg_Total<br>&lt;int&gt; | Pos_Total<br>&lt;int&gt; | Return_Ratio<br>&lt;dbl&gt; |
|---|---|---|---|
| France | 149 | 8407 | 0.01741264 |

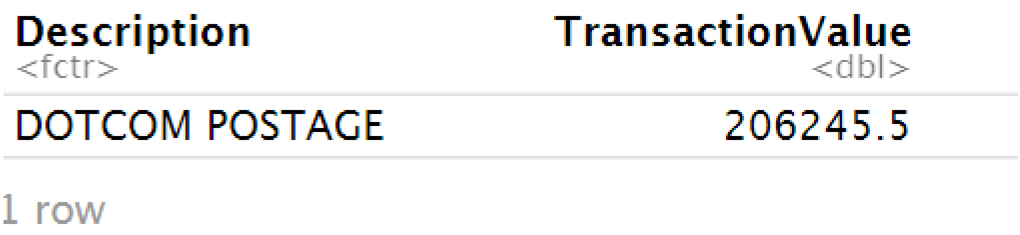
1 row

14. What is the product that has generated the highest revenue for the retailer? (i.e. item with the highest total sum of 'TransactionValue')(10 marks)

**Response:**



~~~
```{r}
#14) the product that has generated the highest revenue for the retailer:


Product <- Retail %>%
    group_by( Description ) %>%
    summarise( TransactionValue = sum(TransactionValue) )

Product <- filter(Product, TransactionValue==max(TransactionValue))

Product
```
~~~

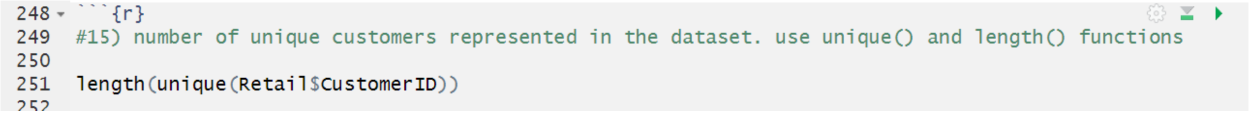


| Description<br>&lt;fctr&gt; | TransactionValue<br>&lt;dbl&gt; |
|---|---|
| DOTCOM POSTAGE | 206245.5 |

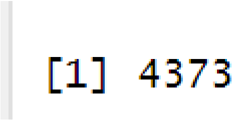
1 row

15. **How many unique customers are represented in the dataset? You can use unique() and length() functions. (5 marks)**

**Response:**



~~~
```{r}
#15) number of unique customers represented in the dataset. use unique() and length() functions

length(unique(Retail$CustomerID))
~~~



```
[1] 4373
```