Online Retail Analytics

Lets see few values of the Online Retail data:

```
##   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
## 1 12/1/2010 8:26      2.55      17850 United Kingdom
## 2 12/1/2010 8:26      3.39      17850 United Kingdom
## 3 12/1/2010 8:26      2.75      17850 United Kingdom
## 4 12/1/2010 8:26      3.39      17850 United Kingdom
## 5 12/1/2010 8:26      3.39      17850 United Kingdom
## 6 12/1/2010 8:26      7.65      17850 United Kingdom
```

DATA EXPLORATION:

```
#Descrptive statistics
summary(Online_data)
```

```
##    InvoiceNo          StockCode         Description          Quantity
##  Length:541909      Length:541909      Length:541909      Min.   :-80995.00
##  Class :character   Class :character   Class :character   1st Qu.:     1.00
##  Mode  :character   Mode  :character   Mode  :character   Median :     3.00
##                                                           Mean   :     9.55
##                                                           3rd Qu.:    10.00
##                                                           Max.   : 80995.00
##
##  InvoiceDate          UnitPrice           CustomerID       Country
##  Length:541909      Min.   :-11062.06   Min.   :12346    Length:541909
##  Class :character   1st Qu.:     1.25   1st Qu.:13953    Class :character
##  Mode  :character   Median :     2.08   Median :15152    Mode  :character
##                     Mean   :     4.61   Mean   :15288
##                     3rd Qu.:     4.13   3rd Qu.:16791
##                     Max.   : 38970.00   Max.   :18287
##                                         NA's   :135080
```

ASSIGNMENT QUESTIONS

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.**

#Total number of transactions by each country accounting more than 1% of total transactions

```
##            Country n_Transactions percentage
## 1   United Kingdom         495478 91.4319563
## 2          Germany           9495  1.7521392
## 3           France           8557  1.5790474
## 4             EIRE           8196  1.5124311
## 5            Spain           2533  0.4674217
## 6      Netherlands           2371  0.4375273
## 7          Belgium           2069  0.3817984
## 8      Switzerland           2002  0.3694347
## 9         Portugal           1519  0.2803054
## 10       Australia           1259  0.2323268
## 11          Norway           1086  0.2004027
## 12           Italy            803  0.1481799
## 13 Channel Islands            758  0.1398759
## 14         Finland            695  0.1282503
## 15          Cyprus            622  0.1147794
```

2. Create a new variable 'Transaction Value' that is the product of the existing 'Quantity' and 'Unit Price' variables. Add this variable to the data frame

 # Added New variable 'TransactionValue' to the end of the dataset (Highlighted by a red color line)

 # let's see the top rows of the dataset after the variable is added

```
##   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. Using the newly created variable, Transaction Value, 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.

 # List of countries with total transaction exceeding 130,000 British Pounds
 # See the Top values of the Output

```
##          Country Total_Sum_of_Transactions
## 1      Australia                  137077.3
## 2           EIRE                  263276.8
## 3         France                  197403.9
## 4        Germany                  221698.2
## 5    Netherlands                  284661.5
## 6 United Kingdom                 8187806.4
```

4. Conversion of categorical variable 'Invoice Date' into POSIXIt Object

See the Data after Conversion with new columns (Highlighted by a red color line)

#Lets see the few values of the dataset with new columns (New_Invoice_Date, Invoice_Day_Week, New_Invoice_Hour, New_Invoice_Month)

```
##   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
##   New_Invoice_Date Invoice_Day_Week New_Invoice_Hour New_Invoice_Month
## 1       2010-12-01        Wednesday                8                12
## 2       2010-12-01        Wednesday                8                12
## 3       2010-12-01        Wednesday                8                12
## 4       2010-12-01        Wednesday                8                12
## 5       2010-12-01        Wednesday                8                12
## 6       2010-12-01        Wednesday                8                12
```

A) Show the percentage of transactions (by numbers) by days of the week

```
##   Invoice_Day_Week Percent_of_Trans_Num_by_week
## 1           Friday                      15.16731
## 2           Monday                      17.55110
## 3           Sunday                      11.87930
## 4         Thursday                      19.16503
## 5          Tuesday                      18.78692
## 6        Wednesday                      17.45035
```

B) Show the percentage of transactions (by transaction volume) by days of the week

```
##   Invoice_Day_Week Percent_of_Trans_Vol_by_week
## 1           Friday                   15.804787
## 2           Monday                   16.297194
## 3           Sunday                    8.265282
## 4         Thursday                   21.671867
## 5          Tuesday                   20.170636
## 6        Wednesday                   17.790232
```

C) Show the percentage of transactions (by transaction volume) by month of the year

	New_Invoice_Month	Percent_of_Trans_by_month
1	1	5.744919
2	2	5.109515
3	3	7.009487
4	4	5.059703
5	5	7.420519
6	6	7.090080
7	7	6.989308
8	8	7.003469
9	9	10.460751
10	10	10.984123
11	11	14.995836
12	12	12.132290

D) What was the date with the highest number of transactions from Australia?

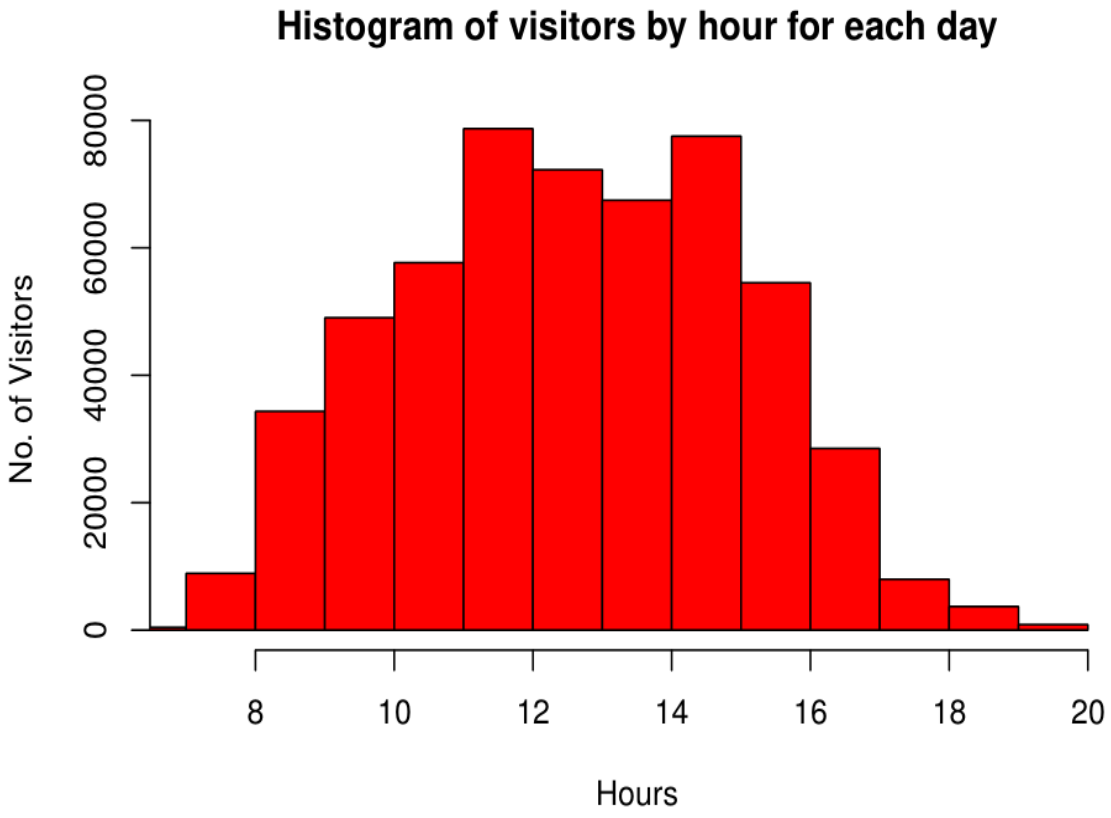
	InvoiceDate	Australia_highest_no_transactions
1	6/15/2011 13:37	139

E) 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.

 # Dataframe with Hour and its corresponding no of transactions per day

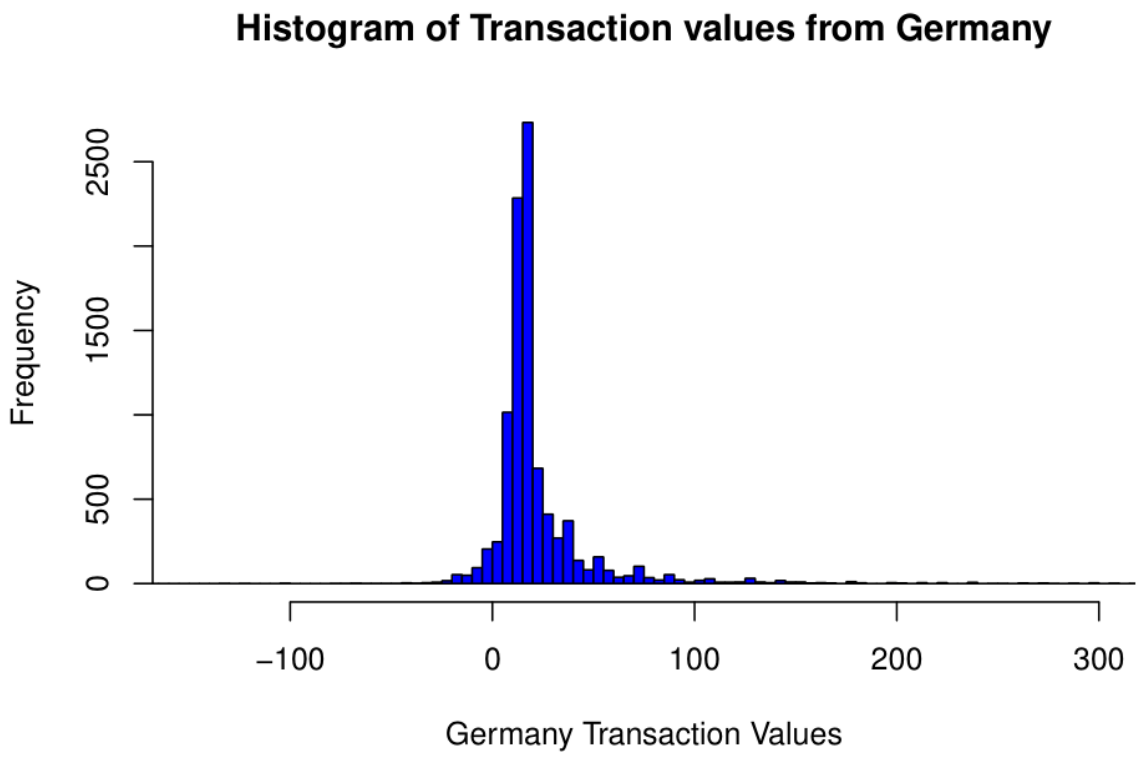
```
##    New_Invoice_Hour No_Of_Transactions  Percentage
## 1                 7                383  0.07067607
## 2                 8               8909  1.64400296
## 3                 9              34332  6.33538103
## 4                10              49037  9.04893626
## 5                11              57674 10.64274629
## 6                12              78709 14.52439432
## 7                13              72259 13.33415758
## 8                14              67471 12.45061440
## 9                15              77519 14.30480025
## 10               16              54516 10.05999162
## 11               17              28509  5.26084638
## 12               18               7974  1.47146477
## 13               19               3705  0.68369413
## 14               20                871  0.16072809
```

 # Plotting a graph to show the website visitors for transactions per hour



It is clearly evident that the good time for maintenance shutdown would be 6.00 am and 20:00 pm. As it is mentioned in the question that responsible IT team would be available from 7.00am to 20:00 pm, the best time would be 7.00 am and 20:00 pm as the distribution would be minimum at these hours.

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



6. Which customer had the highest number of transactions? Which customer is most valuable (i.e.highest total sum of transactions)?

Assumption 1:

a) Considering the Number of Transactions to find most Valuable Customer (Without NA Values)

```
##   CustomerID Highest_no_of_Trans
## 1      17841                7983
```

Assumption 2:

b) Considering the Number of Transactions to find most Valuable Customer (With NA Values)

```
##   CustomerID Highest_no_of_Trans_with_NAValues
## 1         NA                             135080
## 2      17841                               7983
## 3      14911                               5903
```

Assumption 3:

c) Considering the Total sum of Transactions to find most Valuable Customer (Without NA Values)

```
##   CustomerID Highest_Trans_Volume
## 1      14646               279489
```

Assumption 4:

d) Considering the Total sum of Transactions to find most Valuable Customer (With NA Values)

```
##   CustomerID Highest_Trans_Volume_with_NAValues
## 1         NA                          1447682.1
## 2      14646                           279489.0
## 3      18102                           256438.5
```

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

 # Percentage of missing values in the dataset

	Missing_Values_Percent <dbl>
InvoiceNo	0.0000000
StockCode	0.0000000
Description	0.0000000
Quantity	0.0000000
InvoiceDate	0.0000000
UnitPrice	0.0000000
CustomerID	0.2492669
Country	0.0000000
TransactionValue	0.0000000
New_Invoice_Date	0.0000000

Only CustomerID variable has the missing value with 24.92 %

8. What are the number of transactions with missing Customer ID records by countries?

 # No. of transactions with missing CustomerID records by countries

```
## # A tibble: 9 x 2
##   Country        No_of_missing_ID
##   <chr>                     <int>
## 1 United Kingdom           133600
## 2 EIRE                        711
## 3 Hong Kong                   288
## 4 Unspecified                 202
## 5 Switzerland                 125
## 6 France                       66
## 7 Israel                       47
## 8 Portugal                     39
## 9 Bahrain                       2
```

9. On average, how often the customers come back to the website for their next shopping? (i.e. what is the average number of days between consecutive shopping) . Hint: 1. A close approximation is also acceptable and you may find diff() function useful

Assumption 1:

Let's see the top rows of the average days of consecutive shopping Per customer (Including Cancelled transactions)

	CustomerID <int>	avg <time>
1	12347	60.83333 days
2	12348	94.33333 days
3	12352	43.33333 days
4	12356	151.50000 days
5	12358	149.00000 days
6	12359	64.80000 days

6 rows

Assumption 2:

Lets see the top rows of the average days of consecutive shopping per customer (Without Cancelled Transactions)

CustomerID <int>	avg <time>
12347	60.833333 days
12348	94.333333 days
12352	43.333333 days
12356	151.500000 days
12358	149.000000 days
12359	91.333333 days
12360	74.000000 days
12362	32.444444 days
12363	133.000000 days
12364	35.000000 days

The average days of consecutive shopping for all the customers (Without Cancelled Transactions)

avg_days_between_shopping <time>
78.42025 days

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.

```
# [1] "The return rate for the french customers is : 1.7412644618441"
```

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

```
##       Description High_Revenue
## 1 DOTCOM POSTAGE     206245.5
```

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

```
## [1] "The number of Unique Customers in the dataset are: 4373"
```