Assignment 2, Online Retail Analytics

By Jeremy Glasgow

jglasgo2@kent.edu


```
# read in csv into dataframe
data <- read.csv("Online_Retail.csv")
```

```
# 1
countryCount <- as.data.frame(table(data$Country))
colnames(countryCount) <- c("Country", "TotalTransactions")
totalCount <- nrow(data)
countryCount$PercentageOfTotal <- (countryCount$TotalTransactions / totalCount) * 100
print(subset(countryCount, PercentageOfTotal > 0))
```

```
##                 Country TotalTransactions PercentageOfTotal
## 1             Australia              1259       0.232326830
## 2               Austria               401       0.073997664
## 3               Bahrain                19       0.003506124
## 4               Belgium              2069       0.381798420
## 5                Brazil                32       0.005905050
## 6                Canada               151       0.027864457
## 7       Channel Islands               758       0.139875883
## 8                Cyprus               622       0.114779419
## 9         Czech Republic                30       0.005535985
## 10               Denmark               389       0.071783270
## 11                  EIRE              8196       1.512431054
## 12   European Community                61       0.011256502
## 13               Finland               695       0.128250315
## 14                France              8557       1.579047405
## 15               Germany              9495       1.752139197
## 16                Greece               146       0.026941793
## 17             Hong Kong               288       0.053145454
## 18               Iceland               182       0.033584975
## 19                Israel               297       0.054806250
## 20                 Italy               803       0.148179860
## 21                 Japan               358       0.066062752
## 22               Lebanon                45       0.008303977
## 23             Lithuania                35       0.006458649
## 24                 Malta               127       0.023435669
## 25           Netherlands              2371       0.437527334
## 26                Norway              1086       0.200402651
## 27                Poland               341       0.062925694
## 28              Portugal              1519       0.280305365
## 29                   RSA                58       0.010702904
## 30          Saudi Arabia                10       0.001845328
## 31             Singapore               229       0.042258017
## 32                 Spain              2533       0.467421652
## 33                Sweden               462       0.085254166
```

```
## 34          Switzerland              2002       0.369434721
## 35 United Arab Emirates                68       0.012548232
## 36       United Kingdom            495478      91.431956288
## 37          Unspecified               446       0.082301641
## 38                  USA               291       0.053699053
```

```r
# 1
print(subset(countryCount, PercentageOfTotal > 1))
```

```
##           Country TotalTransactions PercentageOfTotal
## 11           EIRE              8196          1.512431
## 14         France              8557          1.579047
## 15        Germany              9495          1.752139
## 36 United Kingdom            495478         91.431956
```

```r
# 2
data$TransactionValue <- data$Quantity * data$UnitPrice
```

```r
# 3
countrySumTransactions <- aggregate(data$TransactionValue, by=list(data$Country), sum)
colnames(countrySumTransactions) <- c("Country", "TransactionValue")
print(subset(countrySumTransactions, TransactionValue > 0))
```

```
##                 Country TransactionValue
## 1             Australia        137077.27
## 2               Austria         10154.32
## 3               Bahrain           548.40
## 4               Belgium         40910.96
## 5                Brazil          1143.60
## 6                Canada          3666.38
## 7       Channel Islands         20086.29
## 8                Cyprus         12946.29
## 9        Czech Republic           707.72
## 10              Denmark         18768.14
## 11                EIRE        263276.82
## 12   European Community          1291.75
## 13              Finland         22326.74
## 14               France        197403.90
## 15              Germany        221698.21
## 16               Greece          4710.52
## 17            Hong Kong         10117.04
## 18              Iceland          4310.00
## 19               Israel          7907.82
## 20                Italy         16890.51
## 21                Japan         35340.62
## 22              Lebanon          1693.88
## 23            Lithuania          1661.06
## 24                Malta          2505.47
## 25          Netherlands        284661.54
## 26               Norway         35163.46
## 27               Poland          7213.14
## 28             Portugal         29367.02
```

```
## 29                  RSA          1002.31
## 30         Saudi Arabia           131.17
## 31            Singapore          9120.39
## 32                Spain         54774.58
## 33               Sweden         36595.91
## 34          Switzerland         56385.35
## 35 United Arab Emirates          1902.28
## 36       United Kingdom       8187806.36
## 37          Unspecified          4749.79
## 38                  USA          1730.92
```

```
# 3
print(subset(countrySumTransactions, TransactionValue > 130000))
```

```
##           Country TransactionValue
## 1       Australia         137077.3
## 11           EIRE         263276.8
## 14         France         197403.9
## 15        Germany         221698.2
## 25    Netherlands         284661.5
## 36 United Kingdom        8187806.4
```

```
# 4
Online_Retail <- data
Online_Retail$InvoiceDate <- format(Online_Retail$InvoiceDate, format="%m/%d/%Y %H:%M", tz="GMT")
Temp=strptime(Online_Retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')
Online_Retail$New_Invoice_Date <- as.Date(Temp)
Online_Retail$New_Invoice_Date[20000]- Online_Retail$New_Invoice_Date[10]
```

```
## Time difference of 8 days
```

```
Online_Retail$Invoice_Day_Week= weekdays(Online_Retail$New_Invoice_Date)
Online_Retail$New_Invoice_Hour = as.numeric(format(Temp, "%H"))
Online_Retail$New_Invoice_Month = as.numeric(format(Temp, "%m"))
Online_Retail$TransactionValue <- Online_Retail$Quantity * Online_Retail$UnitPrice
```

```
# 4 a
print(prop.table(table(Online_Retail$Invoice_Day_Week)) * 100)
```

```
##
##    Friday    Monday    Sunday  Thursday   Tuesday Wednesday
##  15.16731  17.55110  11.87930  19.16503  18.78692  17.45035
```

```
# 4 b
dayOfWeekVol <- aggregate(Online_Retail$TransactionValue, by=list(Online_Retail$Invoice_Day_Week), sum)
dayOfWeekVolPercent <- (dayOfWeekVol$x / sum(dayOfWeekVol$x))
print(data.frame(Day = dayOfWeekVol$Group.1,
                 TotalVolume = dayOfWeekVol$x,
                 Percent = dayOfWeekVolPercent * 100)
)
```

```
##         Day TotalVolume   Percent
## 1    Friday   1540610.8 15.804787
## 2    Monday   1588609.4 16.297194
## 3    Sunday    805678.9  8.265282
## 4  Thursday   2112519.0 21.671867
## 5   Tuesday   1966182.8 20.170636
## 6 Wednesday   1734147.0 17.790232
```

```
# 4 c
monthVol <- aggregate(Online_Retail$TransactionValue, by=list(Online_Retail$New_Invoice_Month), sum)
monthPercent <- (monthVol$x / sum(monthVol$x))
print(data.frame(Month = monthVol$Group.1,
                 TotalVolume = monthVol$x,
                 Percent = monthPercent * 100)
)
```

```
##    Month TotalVolume   Percent
## 1      1    560000.3  5.744919
## 2      2    498062.7  5.109515
## 3      3    683267.1  7.009487
## 4      4    493207.1  5.059703
## 5      5    723333.5  7.420519
## 6      6    691123.1  7.090080
## 7      7    681300.1  6.989308
## 8      8    682680.5  7.003469
## 9      9   1019687.6 10.460751
## 10    10   1070704.7 10.984123
## 11    11   1461756.2 14.995836
## 12    12   1182625.0 12.132290
```

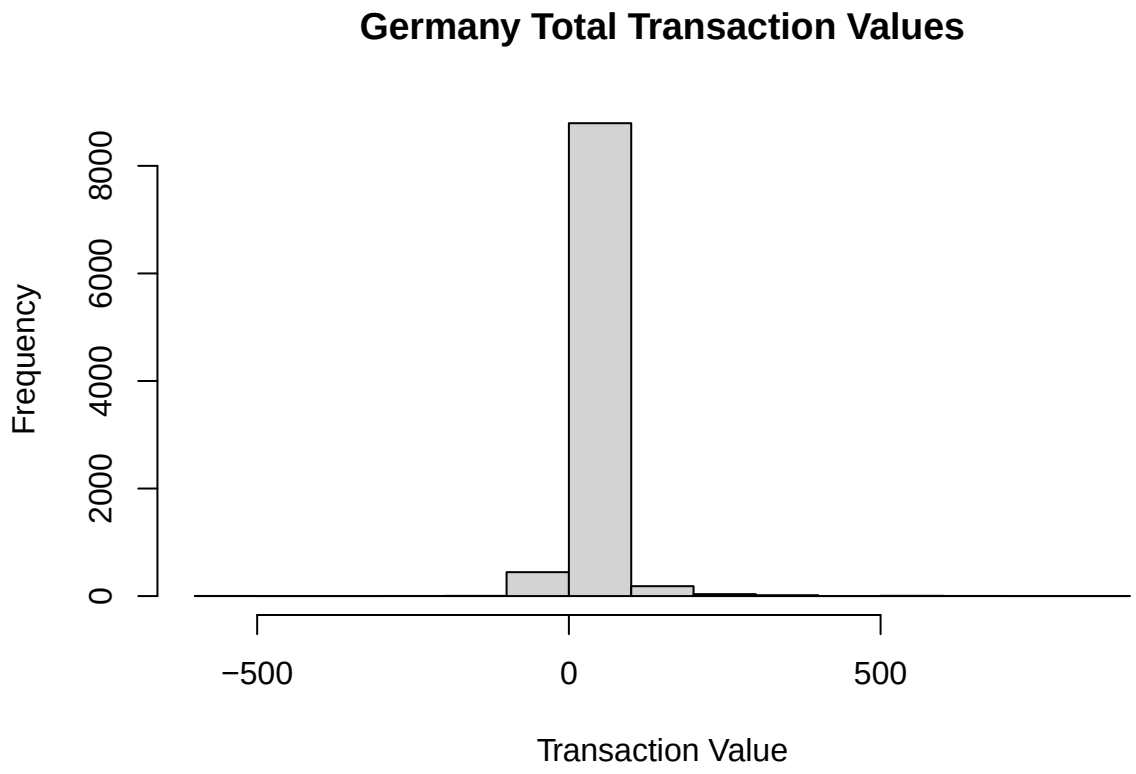
```
# 4 d
austrailiaTransactions <- table(Online_Retail[Online_Retail$Country == "Australia", ]$New_Invoice_Date)
highestTransactionDate <- names(which.max((austrailiaTransactions)))
cat("Date with most transactions: ", highestTransactionDate)
```

```
## Date with most transactions:  2011-06-15
```

```
# 4 e
between07and20 <- function(datetime) {
  if (inherits(datetime, 'POSIXct')) {
    hour <- as.numeric(format(datetime, "%H"))
    return(hour >= 7 && hour <= 20)
  }
}
tempVector <- unlist(lapply(Online_Retail$InvoiceDate, between07and20))
fltrd <- Online_Retail[tempVector, ]
print(fltrd)
```

```
##  [1] InvoiceNo         StockCode         Description       Quantity
##  [5] InvoiceDate       UnitPrice         CustomerID        Country
##  [9] TransactionValue  New_Invoice_Date  Invoice_Day_Week  New_Invoice_Hour
## [13] New_Invoice_Month
## <0 rows> (or 0-length row.names)
```

```r
# 5
germanyTransactions <- data[data$Country == "Germany", "TransactionValue"]
hist(germanyTransactions,
     main = "Germany Total Transaction Values",
     xlab = "Transaction Value",
     ylab = "Frequency"
     )
```




```r
# 6
cat("Customer with the most transactions is :", names(which.max(table(data$CustomerID))))
```

```
## Customer with the most transactions is : 17841
```

```r
cat("\nCustomer with the highest number of total transactional value is :", names(which.max(tapply(data$
```

```
##
## Customer with the highest number of total transactional value is : 14646
```

```r
# 7
percentMissing <- colMeans(is.na(data)) * 100
print(data.frame(Variable = names(percentMissing), MissingPercentage = percentMissing))
```

```
##                          Variable MissingPercentage
```

```
## InvoiceNo               InvoiceNo           0.00000
## StockCode               StockCode           0.00000
## Description           Description           0.00000
## Quantity                 Quantity           0.00000
## InvoiceDate           InvoiceDate           0.00000
## UnitPrice               UnitPrice           0.00000
## CustomerID             CustomerID          24.92669
## Country                   Country           0.00000
## TransactionValue TransactionValue           0.00000
```

```r
# 8
missingTransactions = data[is.na(data$CustomerID), ]
missingByCountry = table(missingTransactions$Country)
print(missingByCountry)
```

```
##
##        Bahrain           EIRE         France      Hong Kong         Israel
##              2            711             66            288             47
##       Portugal    Switzerland United Kingdom    Unspecified
##             39            125         133600            202
```

```r
# 9 - attempted number 4, 9 I was unable to complete this one.
```

```r
# 10
totalFranceTransactions <- nrow(data[data$Country == "France", ])
canceledFranceTransactions <- nrow(data[data$Country == "France" & substr(data$InvoiceNo, 1, 1) == 'C',
rr <- canceledFranceTransactions / totalFranceTransactions
cat("France rate of return is :", rr)
```

```
## France rate of return is : 0.01741264
```

```r
# 11
revenueByProduct <- aggregate(data$TransactionValue, by=list(data$StockCode), sum)
colnames(revenueByProduct) <- c("StockCode", "TotalRevenue")
bestProduct <- revenueByProduct[which.max(revenueByProduct$TotalRevenue), ]
cat("Product with the most revenue is:\n")
```

```
## Product with the most revenue is:
```

```r
print(bestProduct$StockCode[1])
```

```
## [1] "DOT"
```

```r
# 12
uniqueCustomers <- length(unique(data$CustomerID))
cat("Number of unique customers in the dataset: ", uniqueCustomers)
```

```
## Number of unique customers in the dataset:  4373
```