Mukhtar_Assignment1_MIS_64036

Contents

Problem Statement

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.

Data Preparation

```
getwd()
```

```
## [1] "C:/Users/Mukht/OneDrive/Desktop/Kent State University/College of Business Admin-Bus. Analytics I
```

```
setwd("C:\\Users\\Mukht\\OneDrive\\Desktop\\Kent State University\\College of Business Admin-Bus. Analyt

Assignment1Online<-read.csv("Assignt1_Online Retail.csv")
str(Assignment1Online)
```

```
## 'data.frame':    541909 obs. of  8 variables:
##  $ ï..InvoiceNo: chr  "536365" "536365" "536365" "536365" ...
##  $ StockCode   : chr  "85123A" "71053" "84406B" "84029G" ...
##  $ Description : chr  "WHITE HANGING HEART T-LIGHT HOLDER" "WHITE METAL LANTERN" "CREAM CUPID HEARTS
##  $ Quantity    : int  6 6 8 6 6 2 6 6 6 32 ...
##  $ InvoiceDate : chr  "12/1/2010 8:26" "12/1/2010 8:26" "12/1/2010 8:26" "12/1/2010 8:26" ...
##  $ UnitPrice   : num  2.55 3.39 2.75 3.39 3.39 7.65 4.25 1.85 1.85 1.69 ...
##  $ CustomerID  : int  17850 17850 17850 17850 17850 17850 17850 17850 17850 13047 ...
##  $ Country     : chr  "United Kingdom" "United Kingdom" "United Kingdom" "United Kingdom" ...
```

```
head(Assignment1Online)
```

```
##   ï..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
```

```
library(class)
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(lattice)
library(ggplot2)
library(ISLR)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
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
```

```
library(e1071)
```

```
summary(Assignment1Online)
```

```
##  ï..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
```

```
# Solution 1a. how many transactions are in the dataset for each country in total number
table(Assignment1Online$Country)
```

```
##
##           Australia             Austria             Bahrain
##                1259                 401                  19
##             Belgium              Brazil              Canada
##                2069                  32                 151
##     Channel Islands              Cyprus      Czech Republic
##                 758                 622                  30
##             Denmark                EIRE  European Community
##                 389                8196                  61
##             Finland              France             Germany
##                 695                8557                9495
##              Greece           Hong Kong             Iceland
##                 146                 288                 182
##              Israel               Italy               Japan
##                 297                 803                 358
##             Lebanon           Lithuania               Malta
##                  45                  35                 127
##         Netherlands              Norway              Poland
##                2371                1086                 341
##            Portugal                 RSA        Saudi Arabia
##                1519                  58                  10
##           Singapore               Spain              Sweden
##                 229                2533                 462
##         Switzerland United Arab Emirates      United Kingdom
##                2002                  68              495478
##         Unspecified                 USA
##                 446                 291
```

```
# Solution 1b. how many transactions are in the dataset for each country in proportion
prop.table(table(Assignment1Online$Country))
```

```
##
##           Australia             Austria             Bahrain
```

```
##         2.323268e-03         7.399766e-04         3.506124e-05
##              Belgium               Brazil               Canada
##         3.817984e-03         5.905050e-05         2.786446e-04
##      Channel Islands               Cyprus       Czech Republic
##         1.398759e-03         1.147794e-03         5.535985e-05
##              Denmark                 EIRE   European Community
##         7.178327e-04         1.512431e-02         1.125650e-04
##              Finland               France              Germany
##         1.282503e-03         1.579047e-02         1.752139e-02
##               Greece            Hong Kong              Iceland
##         2.694179e-04         5.314545e-04         3.358497e-04
##               Israel                Italy                Japan
##         5.480625e-04         1.481799e-03         6.606275e-04
##              Lebanon            Lithuania                Malta
##         8.303977e-05         6.458649e-05         2.343567e-04
##          Netherlands               Norway               Poland
##         4.375273e-03         2.004027e-03         6.292569e-04
##             Portugal                  RSA         Saudi Arabia
##         2.803054e-03         1.070290e-04         1.845328e-05
##            Singapore                Spain               Sweden
##         4.225802e-04         4.674217e-03         8.525417e-04
##          Switzerland United Arab Emirates       United Kingdom
##         3.694347e-03         1.254823e-04         9.143196e-01
##          Unspecified                  USA
##         8.230164e-04         5.369905e-04
```

```
Assignment1Online %>% group_by(Country) %>% summarize(n=n())
```

```
## # A tibble: 38 x 2
##    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
```

```
# Solution 1. Transactions that are in the dataset for each country in frequency
Assignment1Online %>% group_by(Country) %>% summarize(n=n())%>% mutate(freq=n/sum(n))
```

```
## # A tibble: 38 x 3
##    Country             n      freq
##    <chr>           <int>     <dbl>
##  1 Australia        1259 0.00232
##  2 Austria           401 0.000740
##  3 Bahrain            19 0.0000351
##  4 Belgium          2069 0.00382
```

```
##  5 Brazil             32 0.0000591
##  6 Canada            151 0.000279
##  7 Channel Islands   758 0.00140
##  8 Cyprus            622 0.00115
##  9 Czech Republic     30 0.0000554
## 10 Denmark           389 0.000718
## # ... with 28 more rows
```

```
# Solution 1. Transactions that are in the dataset for each country in percentage
Assignment1Online %>% group_by(Country) %>% summarize(n=n())%>% mutate(freq=n/sum(n)*100)
```

```
## # A tibble: 38 x 3
##    Country             n    freq
##    <chr>           <int>   <dbl>
##  1 Australia        1259 0.232
##  2 Austria           401 0.0740
##  3 Bahrain            19 0.00351
##  4 Belgium          2069 0.382
##  5 Brazil             32 0.00591
##  6 Canada            151 0.0279
##  7 Channel Islands   758 0.140
##  8 Cyprus            622 0.115
##  9 Czech Republic     30 0.00554
## 10 Denmark           389 0.0718
## # ... with 28 more rows
```

```
# Solution 1d. countries accounting for only more than 1% of total transactions
Assignment1Online %>% group_by(Country) %>% summarize(n=n())%>% mutate(freq=n/sum(n)*100)%>% filter(freq
```

```
## # A tibble: 4 x 3
##   Country             n  freq
##   <chr>           <int> <dbl>
## 1 EIRE             8196  1.51
## 2 France           8557  1.58
## 3 Germany          9495  1.75
## 4 United Kingdom 495478 91.4
```

```
# Solution 2.   A new variable 'TransactionValue' that is the product of the exising 'Quantity' and 'Un

Assignment1Online$TransactionValue<- (Assignment1Online$Quantity) * (Assignment1Online$UnitPrice)
colnames(Assignment1Online)
```

```
## [1] "ï..InvoiceNo"     "StockCode"        "Description"      "Quantity"
## [5] "InvoiceDate"      "UnitPrice"        "CustomerID"       "Country"
## [9] "TransactionValue"
```

```
# Solution 3. transaction values by countries i.e. how much money in total has been spent each country.

Sum_of_Transaction<-Assignment1Online %>%
  group_by(Country) %>%
  summarize(Sum_of_Transaction = sum(TransactionValue))%>%
  filter(Sum_of_Transaction >=130000)
Sum_of_Transaction
```

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

head(Assignment1Online) unique(Assignment1Online$Country)

```r
# Solution 4 Optional
Temp<-strptime(Assignment1Online$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"
```

```r
Assignment1Online$Invoice_Date_Week <- weekdays(Temp)
Assignment1Online$New_Invoice_Hour <- as.numeric(format(Temp, "%H"))
Assignment1Online$New_Invoice_Month <- as.numeric(format(Temp, "%m"))
head(Assignment1Online)
```

```
##   ï..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
##   Invoice_Date_Week New_Invoice_Hour New_Invoice_Month
## 1         Wednesday               8                12
## 2         Wednesday               8                12
## 3         Wednesday               8                12
## 4         Wednesday               8                12
## 5         Wednesday               8                12
## 6         Wednesday               8                12
```

```r
#a) The percentage of transactions (by numbers) by days of the week
Assignment1Online %>% group_by(Invoice_Date_Week) %>% summarise(Percentage_of_Trans = n()/nrow(Assignme
```

```
## # A tibble: 6 x 2
##   Invoice_Date_Week Percentage_of_Trans
##   <chr>                           <dbl>
## 1 Friday                          0.152
## 2 Monday                          0.176
## 3 Sunday                          0.119
## 4 Thursday                        0.192
## 5 Tuesday                         0.188
## 6 Wednesday                       0.175
```

```
#b) The percentage of transactions (by transaction volume) by days of the week
Assignment1Online %>% group_by(Invoice_Date_Week)%>% summarise(Volume_Percentage = sum(Quantity)/sum(As
```

```
## # A tibble: 6 x 2
##   Invoice_Date_Week Volume_Percentage
##   <chr>                         <dbl>
## 1 Friday                       0.153
## 2 Monday                       0.158
## 3 Sunday                       0.0904
## 4 Thursday                     0.226
## 5 Tuesday                      0.186
## 6 Wednesday                    0.187
```

```
#c) The percentage of transactions (by transaction volume) by month of the year
Assignment1Online %>% group_by(New_Invoice_Month) %>% summarise(Volume_Percentage = sum(Quantity)/sum(As
```

```
## # A tibble: 12 x 2
##    New_Invoice_Month Volume_Percentage
##                <dbl>             <dbl>
##  1                 1              5.97
##  2                 2              5.37
##  3                 3              6.80
##  4                 4              5.58
##  5                 5              7.35
##  6                 6              6.60
##  7                 7              7.56
##  8                 8              7.85
##  9                 9             10.6
## 10                10             11.0
## 11                11             14.3
## 12                12             11.0
```

```
#d) The date with the highest number of transactions from Australia?
Assignment1Online_AU<- Assignment1Online[Assignment1Online$Country == "Australia", ] %>%
  group_by(InvoiceDate) %>%
  summarise(Num_of_Trans = n())
Assignment1Online_AU[which.max(Assignment1Online_AU$Num_of_Trans),]
```

```
## # A tibble: 1 x 2
##   InvoiceDate     Num_of_Trans
##   <chr>                  <int>
## 1 6/15/2011 13:37          139
```

```r
#e) The company needs to shut down the website for two consecutive hours for maintenance. What would be
volumehr <- Assignment1Online %>% group_by(New_Invoice_Hour) %>% summarise(Volume = sum(abs(Quantity)))
volumehr <- volumehr[volumehr$New_Invoice_Hour >= 7 & volumehr$New_Invoice_Hour < 20,]
volumehr_2 <- volumehr[1:(nrow(volumehr)-1), ] + volumehr[2:nrow(volumehr), ]
volumehr_2[which.min(volumehr_2$Volume), ]
```

```
##    New_Invoice_Hour Volume
## 12               37 108185
```

```r
volumehr
```

```
## # A tibble: 13 x 2
##    New_Invoice_Hour Volume
##               <dbl>  <int>
##  1                7  15379
##  2                8 160111
##  3                9 614220
##  4               10 955388
##  5               11 718154
##  6               12 883915
##  7               13 745133
##  8               14 641677
##  9               15 693252
## 10               16 364937
## 11               17 234167
## 12               18  73999
## 13               19  34186
```

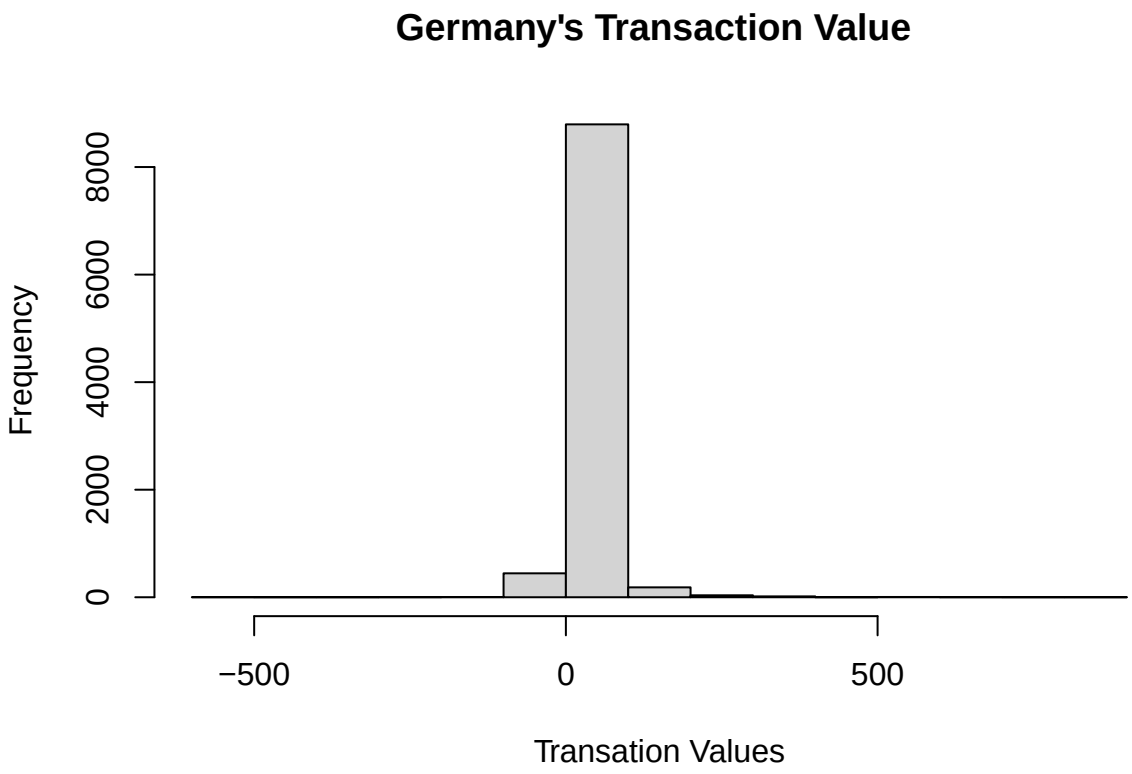
```r
volumehr_2
```

```
##    New_Invoice_Hour  Volume
## 1                15  175490
## 2                17  774331
## 3                19 1569608
## 4                21 1673542
## 5                23 1602069
## 6                25 1629048
## 7                27 1386810
## 8                29 1334929
## 9                31 1058189
## 10               33  599104
## 11               35  308166
## 12               37  108185
```

```r
#Based on the index we could conclude that the website volume is the lowest from 15:00 to 17:00. So the

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

Germany<- Assignment1Online %>% filter(Country == "Germany")
hist(Germany$TransactionValue, main = "Germany's Transaction Value ", xlab = "Transation Values")
```




```
# Solution 6a. Customer that had the highest number of transactions? Which customer is most valuable (i
#Cutomer with the highest number of transactions
High_Trans_Customer <- Assignment1Online %>% group_by(CustomerID) %>% summarise(Num_of_Transactions = n
High_Trans_Customer <- na.omit(High_Trans_Customer, col = "CustomerID")
High_Trans_Customer[which.max(High_Trans_Customer$CustomerID), ]
```

```
## # A tibble: 1 x 2
##   CustomerID Num_of_Transactions
##        <int>               <int>
## 1      18287                  70
```

```
# Solution 6b. Customer with the highest total sum of transactions
High_Trans_Customer<- Assignment1Online %>% group_by(CustomerID) %>% summarise(Total_Trans_Value = sum(A
High_Trans_Customer<- na.omit(High_Trans_Customer, col = "CustomerID")
High_Trans_Customer[which.max(High_Trans_Customer$Total_Trans_Value), ]
```

```
## # A tibble: 1 x 2
##   CustomerID Total_Trans_Value
##        <int>             <dbl>
## 1      12346          9747748.
```

```
# Solution 7.   The percentage of missing values for each variable in the dataset (5 marks). Hint colMe
colMeans(is.na(Assignment1Online))
```

```
##      ï..InvoiceNo         StockCode       Description          Quantity
##         0.0000000         0.0000000         0.0000000         0.0000000
##       InvoiceDate         UnitPrice        CustomerID           Country
##         0.0000000         0.0000000         0.2492669         0.0000000
##  TransactionValue Invoice_Date_Week  New_Invoice_Hour New_Invoice_Month
##         0.0000000         0.0000000         0.0000000         0.0000000
```

```
# Solution 8.   The number of transactions with missing CustomerID records by countries
Assignment1Online %>% group_by(Country) %>% summarize(Missing_Value = sum(is.na(CustomerID)))
```

```
## # A tibble: 38 x 2
##    Country         Missing_Value
##    <chr>                   <int>
##  1 Australia                   0
##  2 Austria                     0
##  3 Bahrain                     2
##  4 Belgium                     0
##  5 Brazil                      0
##  6 Canada                      0
##  7 Channel Islands             0
##  8 Cyprus                      0
##  9 Czech Republic              0
## 10 Denmark                     0
## # ... with 28 more rows
```

```
# Solution 9.   On average, the costumers comeback to the website for their next shopping? (i.e. what i
```

```
#13.    In the retail sector, it is very important to understand the return rate of the goods purchased
France <- Assignment1Online[Assignment1Online$Country == "France", ]
nrow(France[France$Quantity < 0, ])/nrow(France)
```

```
## [1] 0.01741264
```

```
#11.    The product that has generated the highest revenue for the retailer? (i.e. item with the highes
Revenue_by_item <- Assignment1Online %>% filter (Quantity > 0) %>% group_by(StockCode) %>% summarise(Re
Revenue_by_item[which.max(Revenue_by_item$Revenue), ]
```

```
## # A tibble: 1 x 2
##   StockCode Revenue
##   <chr>       <dbl>
## 1 DOT       206249.
```

```
#12.    The unique customers that are represented in the dataset
length(unique(Assignment1Online$CustomerID))
```

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