Assignment 1

Gordon Wall (gwall2)

##Import and read Online Retail dataset

Online\_Retail <- read.csv("C:/Users/Gordon/Dropbox/KSU MSBA Program/SEMESTER 4/BUSINESS ANALYTICS 64036/Module 4/Online\_Retail.csv")

##Libraries

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(caret)

## Loading required package: lattice

## Loading required package: ggplot2

##Question 1

Total\_Transactions = unique(Online\_Retail$InvoiceNo)  
Total = nlevels(Total\_Transactions)  
Total

## [1] 25900

There are **25900** Total unique transactions in this dataset.

#Transactions by Country (total number)

TransTotal = Online\_Retail %>% group\_by(Country, InvoiceNo) %>% summarise(n = n()) %>% count(Country, sort = TRUE, name = "TransTotal")  
subset(TransTotal, TransTotal > 0.01\*25900)

## # A tibble: 4 x 2  
## # Groups: Country [4]  
## Country TransTotal  
## <fct> <int>  
## 1 United Kingdom 23494  
## 2 Germany 603  
## 3 France 461  
## 4 EIRE 360

#Transactions by Country (percentage)

TransTotalP = summarise(TransTotal, PercTotal = TransTotal/25900\*100)  
subset(TransTotalP, PercTotal > 1)

## # A tibble: 4 x 2  
## Country PercTotal  
## <fct> <dbl>  
## 1 EIRE 1.39  
## 2 France 1.78  
## 3 Germany 2.33  
## 4 United Kingdom 90.7

##Question 2

Online\_Retail\_New = Online\_Retail %>% mutate(TransactionValue = Quantity\*UnitPrice)

##Question 3

SumTrans = aggregate(Online\_Retail\_New$TransactionValue, by = list(Country=Online\_Retail\_New$Country), FUN = sum)  
subset(SumTrans, x > 130000)

## Country x  
## 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

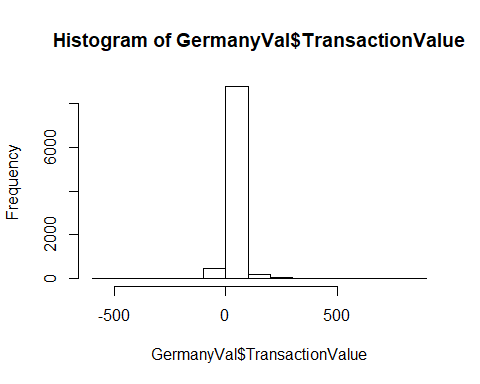
##Question 4 #Bonus

Temp=strptime(Online\_Retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')  
Online\_Retail$New\_Invoice\_Date <- as.Date(Temp)  
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"))

//NOTE: I tried several different attempts and methods to solve these, none of which worked sufficiently enough to turn in, but some of which were very close. Any explanation is appreciated!//

##Question 5

GermanyVal = Online\_Retail\_New %>% filter(TransactionValue, Country == "Germany")  
hist(GermanyVal$TransactionValue)



##Question 6

Online\_Retail\_New %>% group\_by(CustomerID, InvoiceNo) %>% summarise(n=n()) %>% count(CustomerID, sort = TRUE)

## # A tibble: 4,373 x 2  
## # Groups: CustomerID [4,373]  
## CustomerID n  
## <int> <int>  
## 1 NA 3710  
## 2 14911 248  
## 3 12748 224  
## 4 17841 169  
## 5 14606 128  
## 6 13089 118  
## 7 15311 118  
## 8 12971 89  
## 9 14527 86  
## 10 13408 81  
## # ... with 4,363 more rows

Excluding those CustomerID observations that have NA values, the highest number of transactions belongs to CustomerID **14911**.

SumCust = aggregate(Online\_Retail\_New$TransactionValue, by = list(Customer=Online\_Retail\_New$CustomerID), FUN = sum)  
SumCust[which.max(SumCust$x),]

## Customer x  
## 1704 14646 279489

CustomerID **14646** is the most valuable, with total transaction expenditures amounting to **$279,489**.

##Question 7

colMeans(is.na(Online\_Retail))

## InvoiceNo StockCode Description Quantity   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## InvoiceDate UnitPrice CustomerID Country   
## 0.0000000 0.0000000 0.2492669 0.0000000   
## New\_Invoice\_Date Invoice\_Day\_Week New\_Invoice\_Hour New\_Invoice\_Month   
## 0.0000000 0.0000000 0.0000000 0.0000000

It appears the CustomerID variable is the only column possessing missing values, with an average of about 1 in 4 being N/A (~25%).

##Question 8

CountryNA = Online\_Retail %>% group\_by(Country, Missing = is.na(CustomerID)) %>% summarise(n = n())  
filter(CountryNA, Missing == "TRUE")

## # A tibble: 9 x 3  
## # Groups: Country [9]  
## Country Missing n  
## <fct> <lgl> <int>  
## 1 Bahrain TRUE 2  
## 2 EIRE TRUE 711  
## 3 France TRUE 66  
## 4 Hong Kong TRUE 288  
## 5 Israel TRUE 47  
## 6 Portugal TRUE 39  
## 7 Switzerland TRUE 125  
## 8 United Kingdom TRUE 133600  
## 9 Unspecified TRUE 202

##Question 9 #Bonus //NOTE: (same as note in question 4)//

##Question 10

French = Online\_Retail\_New %>% group\_by(Country, InvoiceNo, Quantity)  
FrenchQuant = subset(French, Country == "France")  
FrenchNegative = subset(FrenchQuant, Quantity < 0)  
subset(FrenchNegative, !duplicated(InvoiceNo))

## # A tibble: 69 x 9  
## # Groups: Country, InvoiceNo, Quantity [69]  
## InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice  
## <fct> <fct> <fct> <int> <fct> <dbl>  
## 1 C537893 22450 SILK PURSE~ -1 12/9/2010 ~ 3.35  
## 2 C539104 22662 LUNCH BAG ~ -3 12/16/2010~ 1.65  
## 3 C539114 22667 "RECIPE BO~ -3 12/16/2010~ 2.95  
## 4 C540151 21843 RED RETROS~ -1 1/5/2011 1~ 11.0   
## 5 C540156 21843 RED RETROS~ -1 1/5/2011 1~ 11.0   
## 6 C540367 POST POSTAGE -3 1/6/2011 1~ 18   
## 7 C540853 22634 "CHILDS BR~ -1 1/12/2011 ~ 9.95  
## 8 C541688 22423 REGENCY CA~ -4 1/20/2011 ~ 11.0   
## 9 C541825 21877 HOME SWEET~ -1 1/21/2011 ~ 1.25  
## 10 C542088 22556 "PLASTERS ~ -12 1/25/2011 ~ 1.65  
## # ... with 59 more rows, and 3 more variables: CustomerID <int>,  
## # Country <fct>, TransactionValue <dbl>

Output shows 69 unique transactions (69 observations) that were cancelled and returned. Previous analysis performed in question 1 reveals that France has participated in 461 unique total transactions. Therefore, the return rate of French customers is 69/461, or **0.149674**, meaning about **15%**. //NOTE: I’m aware I did this the long way with several stored intermediate variables. I couldn’t make the piping method work absent error for this question. I arrived at the correct answer (I believe), but slowly, so any insight is appreciated.//

##Question 11

temprev = aggregate(Online\_Retail\_New$TransactionValue, by = list(Online\_Retail\_New$StockCode), FUN = sum, sort = TRUE)  
temprev[which.max(temprev$x),]

## Group.1 x  
## 4060 DOT 206246.5

It appears that the stockcode **DOT** pertains to the item that generated the highest total revenue across all sales. If this is sufficient, question is answered. If description of item is needed, below is the filtered list of transactions by stockcode DOT, showing us the complete description of this item.

filter(Online\_Retail\_New, StockCode == "DOT")

## InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice  
## 1 536544 DOT DOTCOM POSTAGE 1 12/1/2010 14:32 569.77  
## 2 536592 DOT DOTCOM POSTAGE 1 12/1/2010 17:06 607.49  
## 3 536862 DOT DOTCOM POSTAGE 1 12/3/2010 11:13 254.43  
## 4 536864 DOT DOTCOM POSTAGE 1 12/3/2010 11:27 121.06  
## 5 536865 DOT DOTCOM POSTAGE 1 12/3/2010 11:28 498.47  
## 6 536876 DOT DOTCOM POSTAGE 1 12/3/2010 11:36 887.52  
## 7 537237 DOT DOTCOM POSTAGE 1 12/6/2010 9:58 863.74  
## 8 537240 DOT DOTCOM POSTAGE 1 12/6/2010 10:08 940.87  
## 9 537434 DOT DOTCOM POSTAGE 1 12/6/2010 16:57 950.99

Stockcode DOT identifies the highest revenue item as **DOTCOM POSTAGE**.

##Question 12

Online\_Retail %>% summarise(n\_distinct(CustomerID))

## n\_distinct(CustomerID)  
## 1 4373

There are **4373** unique customers in this dataset.