Business Analytics- Online Retail Analytics

Steven Pavliga

2024-10-13

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.

```
OnlineRetail <- read.csv("C:/Users/Owner/Documents/Online_Retail.csv")

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(ISLR)

View(OnlineRetail)

Data <- as.data.frame(OnlineRetail)

class(Data$Country)
```

```
## [1] "character"
```

```
TransByCountry <- Data %>% group_by(Country) %>% summarise(TransactionCount = n())

print(TransByCountry, n = 38)
```

```
## # A tibble: 38 x 2
##    Country              TransactionCount
##    <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
## 11 EIRE                              8196
## 12 European Community                  61
## 13 Finland                            695
## 14 France                            8557
## 15 Germany                           9495
## 16 Greece                             146
## 17 Hong Kong                          288
## 18 Iceland                            182
## 19 Israel                             297
## 20 Italy                              803
## 21 Japan                              358
## 22 Lebanon                             45
## 23 Lithuania                           35
## 24 Malta                              127
## 25 Netherlands                       2371
## 26 Norway                            1086
## 27 Poland                             341
## 28 Portugal                          1519
## 29 RSA                                 58
## 30 Saudi Arabia                        10
## 31 Singapore                          229
## 32 Spain                             2533
## 33 Sweden                             462
## 34 Switzerland                       2002
## 35 USA                                291
## 36 United Arab Emirates                68
## 37 United Kingdom                  495478
## 38 Unspecified                        446
```

```
SumTransByCountry <- sum(TransByCountry$TransactionCount)

print(SumTransByCountry)
```

```
## [1] 541909
```

```
TransByCountryWPC <- TransByCountry %>% mutate(Percentage = (TransactionCount / SumTransByCountry) * 10

print(TransByCountryWPC)
```

```
## # A tibble: 38 x 3
##    Country         TransactionCount Percentage
##    <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
## # i 28 more rows
```

```
TotalTransByCountryPercentage <- TransByCountryWPC %>% filter(Percentage > 1)

print(TotalTransByCountryPercentage)
```

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

2. Create a new variable 'TransactionValue' that is the product of the exising 'Quantity' and 'UnitPrice' variables. Add this variable to the dataframe.

```
Data <- Data %>% mutate(TransactionValue = Data$Quantity * Data$UnitPrice)
```

3. 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.

```
TransValuesByCountry <- Data %>% group_by(Country) %>% summarise(TotalTrans = sum(TransactionValue, na.

print(TransValuesByCountry, n = 38)
```

```
## # A tibble: 38 x 2
##    Country              TotalTrans
##    <chr>                     <dbl>
##  1 Australia               137077.
##  2 Austria                  10154.
##  3 Bahrain                    548.
##  4 Belgium                  40911.
##  5 Brazil                    1144.
##  6 Canada                    3666.
##  7 Channel Islands          20086.
##  8 Cyprus                   12946.
##  9 Czech Republic             708.
## 10 Denmark                  18768.
## 11 EIRE                    263277.
## 12 European Community        1292.
## 13 Finland                  22327.
## 14 France                  197404.
```

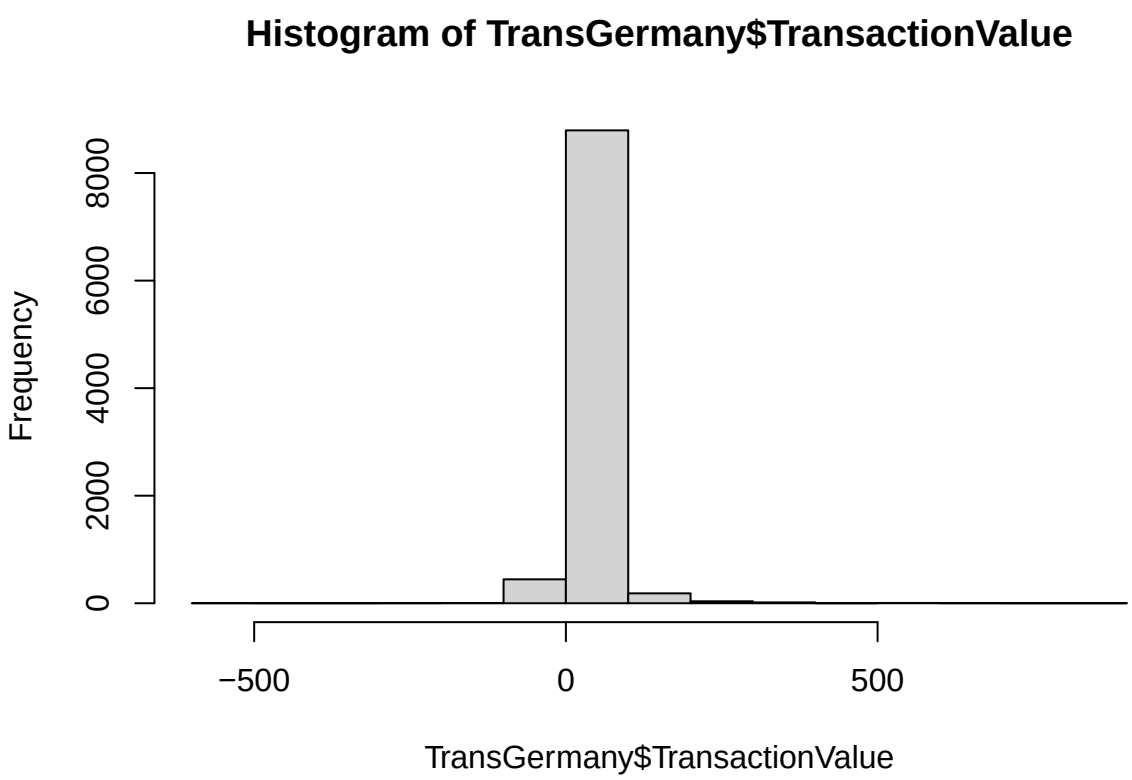
```
## 15 Germany                 221698.
## 16 Greece                    4711.
## 17 Hong Kong                10117.
## 18 Iceland                   4310
## 19 Israel                    7908.
## 20 Italy                    16891.
## 21 Japan                    35341.
## 22 Lebanon                   1694.
## 23 Lithuania                 1661.
## 24 Malta                     2505.
## 25 Netherlands             284662.
## 26 Norway                   35163.
## 27 Poland                    7213.
## 28 Portugal                 29367.
## 29 RSA                       1002.
## 30 Saudi Arabia               131.
## 31 Singapore                 9120.
## 32 Spain                    54775.
## 33 Sweden                   36596.
## 34 Switzerland              56385.
## 35 USA                       1731.
## 36 United Arab Emirates      1902.
## 37 United Kingdom         8187806.
## 38 Unspecified               4750.
```

```
CountriesFiltered <- TransValuesByCountry %>% filter(TotalTrans > 130000)

print(CountriesFiltered)
```

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

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

```
TransGermany <- Data %>% filter(Country == "Germany")

hist(TransGermany$TransactionValue)
```



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

```
TopTransCustomer <- Data %>% filter(!is.na(CustomerID)) %>% count(CustomerID, name = "TransCount") %>% 

print(TopTransCustomer)
```

```
##   CustomerID TransCount
## 1      17841       7983
```

```
TopCustomer <- Data %>% filter(!is.na(CustomerID)) %>% group_by(CustomerID) %>% summarise(TotalTrans = 

print(TopCustomer)
```

```
## # A tibble: 1 x 2
##   CustomerID TotalTrans
##        <int>      <dbl>
## 1      14646    279489.
```

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

```
colMeans(is.na(Data)) * 100
```

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

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

```
TransNoID <- Data %>% filter(is.na(CustomerID)) %>% group_by(Country) %>% summarise(MissingTrans = n())

print(TransNoID)
```

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

9.

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

```
FrenchTrans <- Data %>% filter(Country == "France")

FrenchTransTotal <- nrow(FrenchTrans)

print(FrenchTransTotal)
```

```
## [1] 8557
```

```
FrenchTransCancelled <- FrenchTrans %>% filter(Quantity < 0) %>% nrow()

print(FrenchTransCancelled)
```

```
## [1] 149
```

```
(FrenchTransCancelled / FrenchTransTotal) * 100
```

```
## [1] 1.741264
```

```
#1.74 out of 100 transactions are cancelled.
```

11. What is the product that has generated the highest revenue for the retailer?

```
TopRevenue <- Data %>% group_by(StockCode) %>% summarise(TotalRevenue = sum(TransactionValue, na.rm = TR

print(TopRevenue)
```

```
## # A tibble: 1 x 2
##   StockCode TotalRevenue
##   <chr>            <dbl>
## 1 DOT            206245.
```

12. How many unique customers are represented in the dataset?

```
length(unique(Data$CustomerID))
```

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