

For_OnlineRetail.R

vk0589

2022-11-19

```
setwd("C:/Users/vk0589/OneDrive - UNT System/Documents/UNT/INFO_5307/Datasets")
#install.packages("tidyverse")
#read.csv("OnlineRetail.csv")
library(readxl)
data1 <- read_excel("OnlineRetail.xlsx")
data2 <- as.data.frame(data1)
# head(data2)

# Transforming Data for Plotting #####
# Renaming the Country Names
#data2$Country[data2$Country == "United Kingdom"] = "UK"
#data2$Country[data2$Country == "EIRE"] = "Ireland"
#data2$Country[data2$Country == "RSA"] = "South Africa"
#data2$Country[data2$Country == "Hong Kong"] = "China"

# Changing characters to factors
data2$InvoiceNo <- as.factor(data2$InvoiceNo)
data2$Description <- as.factor(data2$Description)
data2$CustomerID <- as.factor(data2$CustomerID)
data2$Description <- as.factor(data2$Description)
data2$Country <- as.factor(data2$Country)
data2$StockCode <- as.factor(data2$StockCode)
# summary(data2)

# World Map #####
# install.packages("tidyverse")
# install.packages("maps")
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
```

```

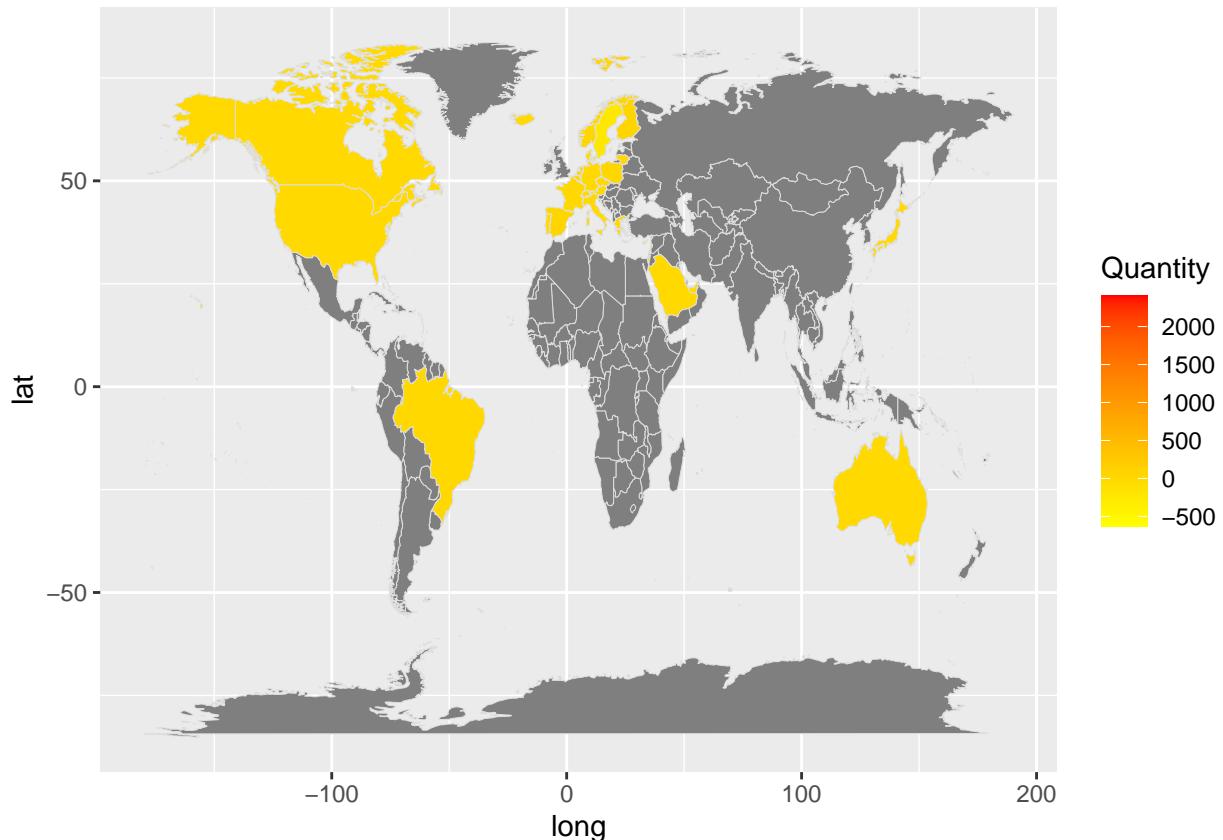
library(tidyr)
library(ggplot2)
library(maps)

## Warning: package 'maps' was built under R version 4.2.2

# creating variable for World Map
world <- map_data("world")

# Plotting Countries with ggplot on World Map
# Following Map indicates the data available in OnlineRetail.xlsx dataset
# Changing to factor
world %>%
  left_join(data2, by = c("region" = "Country")) %>%
  ggplot(aes(x = long, y = lat, group = group, fill = Quantity)) +
  geom_polygon(color = "#DFDFDF", size = 0.1) +
  scale_fill_gradient(low = "yellow", high = "red" )

```



```

# Findings #####
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.2 --
## v tibble 3.1.8     v stringr 1.4.1

```

```

## v readr  2.1.3      vforcats 0.5.2
## v purrr  0.3.5
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::map()   masks maps::map()

library(ggplot2)
library(dplyr)

# Observations summary based on Country (Count)
summary(data2$Country)

```

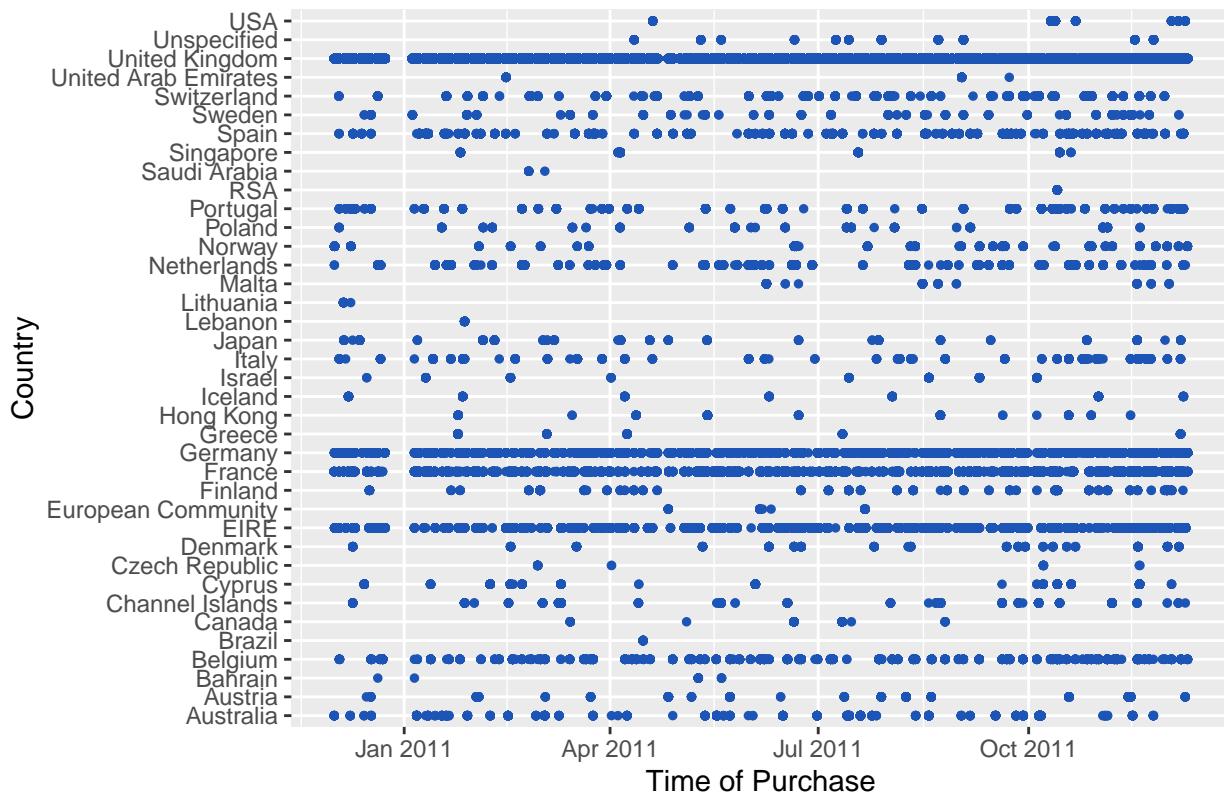
##	Australia	Austria	Bahrain
##	1259	401	19
##	Belgium	Brazil	Canada
##	2069	32	151
##	Channel Islands	Cyprus	Czech Republic
##	758	622	30
##	Denmark	EIRE	European Community
##	389	8196	61
##	Finland	France	Germany
##	695	8557	9495
##	Greece	Hong Kong	Iceland
##	146	288	182
##	Israel	Italy	Japan
##	297	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	

```

# Purchase Time v/s Country
data2%>%
  ggplot(aes(InvoiceDate, Country))+
  geom_point(color = "#1D55B6", size = 1) +
  labs(x = "Time of Purchase", y = "Country",
       title = "Purchase Frequency of All Countries")

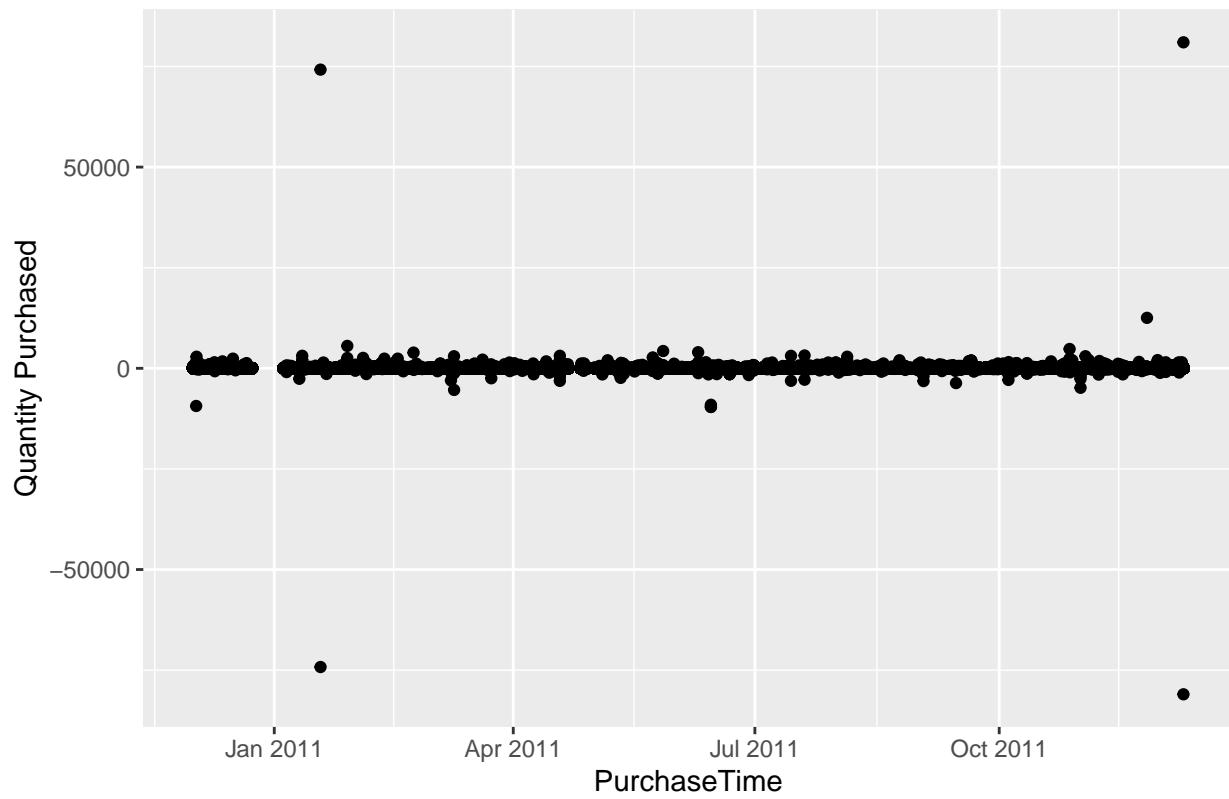
```

Purchase Frequency of All Countries



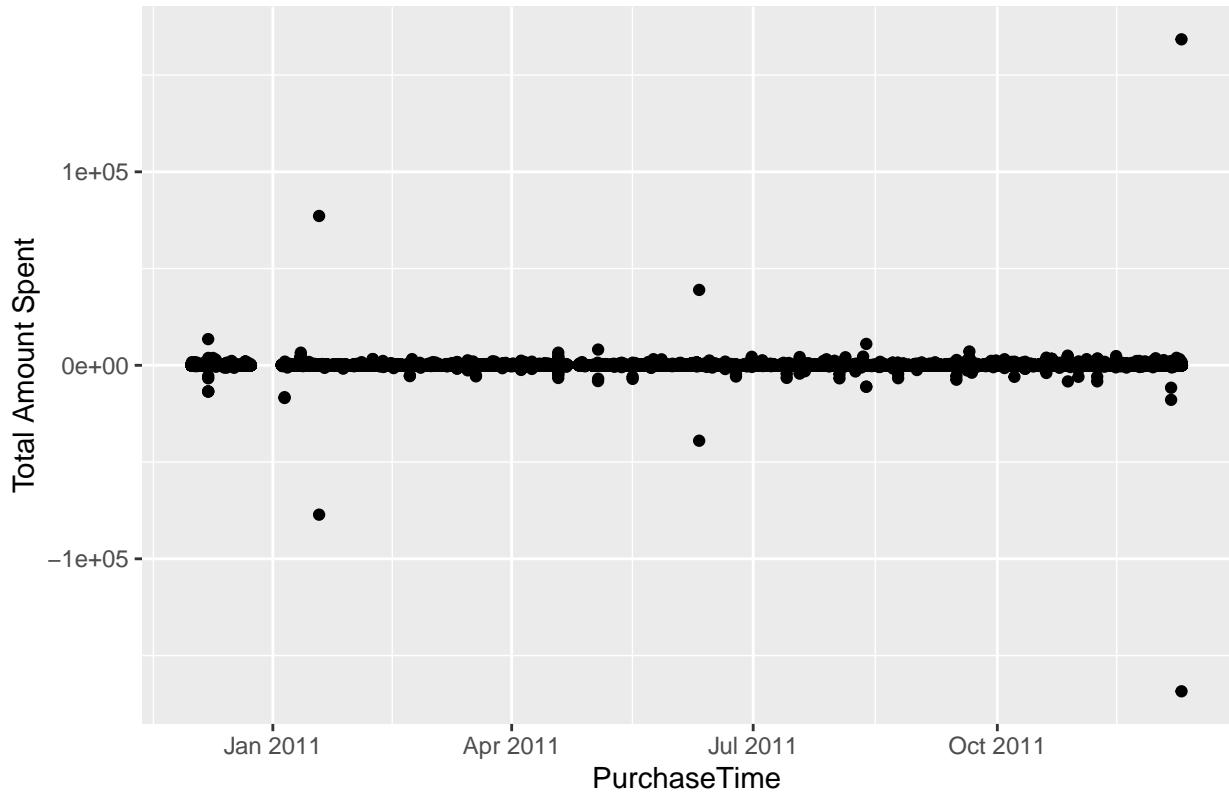
```
# Purchase Time v/s Quantity for whole Data:
qplot(x = InvoiceDate, y = Quantity,
       data = data2,
       xlab = "PurchaseTime",
       ylab = "Quantity Purchased",
       main = "PurchaseTime v/s Quantity Purchased (Whole Data)")
```

PurchaseTime v/s Quantity Purchased (Whole Data)



```
# Purchase Time v/s Price Spent for whole Data:  
qplot(x = InvoiceDate, y = Quantity*UnitPrice,  
      data = data2,  
      xlab = "PurchaseTime",  
      ylab = "Total Amount Spent",  
      main = "PurchaseTime v/s Total Amount (Whole Data)")
```

PurchaseTime v/s Total Amount (Whole Data)



```
# Tests based on Countries
# 1) USA
filter(data2, Country == "USA") -> us_data
summary(us_data)
```

```
##      InvoiceNo      StockCode          Description
## 570467 :101  22027 : 4 CARD DOLLY GIRL       : 4
## C570867:101  22712 : 4 TEA PARTY BIRTHDAY CARD   : 4
## 572215 : 24  21121 : 3 EMBROIDERED RIBBON REEL SUSIE : 3
## 550644 : 22  21122 : 3 PINK HAPPY BIRTHDAY BUNTING : 3
## 580553 : 21  21123 : 3 SET 2 PANTRY DESIGN TEA TOWELS : 3
## 580158 : 11  21124 : 3 SET OF 12 FAIRY CAKE BAKING CASES: 3
## (Other): 11  (Other):271 (Other)                   :271
##      Quantity      InvoiceDate        UnitPrice
## Min.    :-36.000  Min. 2011-04-19 16:19:00.00  Min.   : 0.420
## 1st Qu.:-10.000  1st Qu.:2011-10-10 16:06:00.00  1st Qu.: 0.850
## Median :  5.000   Median :2011-10-12 16:17:00.00  Median : 1.450
## Mean   :  3.553   Mean  :2011-10-07 08:22:58.97  Mean   : 2.216
## 3rd Qu.: 12.000   3rd Qu.:2011-10-12 16:17:00.00  3rd Qu.: 2.950
## Max.   : 72.000   Max.  :2011-12-08 10:14:00.00  Max.   :16.950
##
##      CustomerID      Country
## 12607 :202      USA     :291
## 12646 : 45      Australia: 0
## 12558 : 22      Austria : 0
## 12733 : 22      Bahrain : 0
```

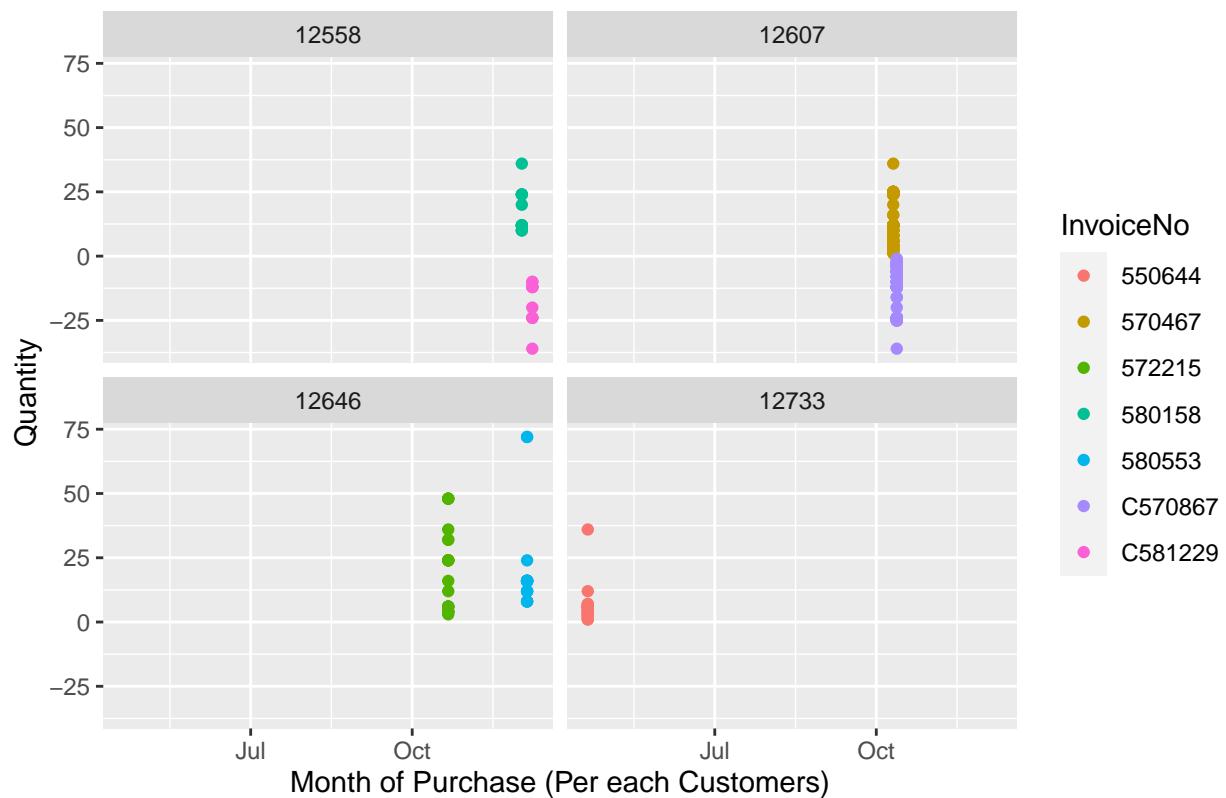
```

## 12346 : 0 Belgium : 0
## 12347 : 0 Brazil : 0
## (Other): 0 (Other) : 0

# qplot(x = InvoiceDate, y = Quantity, data = us_data, color = CustomerID)
# qplot(x = InvoiceDate, y = Quantity, data = us_data, color = InvoiceNo)
qplot(x = InvoiceDate, y = Quantity,
       data = us_data,
       xlab = "Month of Purchase (Per each Customers)",
       ylab = "Quantity",
       main = "Purchase Month v/s Quantity",
       color = InvoiceNo, # Coloring Points based on 'InvoiceNo'
       facets = ~ CustomerID) # Separated Graphs for CustomerID

```

Purchase Month v/s Quantity



```

# 2) France
filter(data2, Country == "France") -> frc_data
summary(frc_data)

```

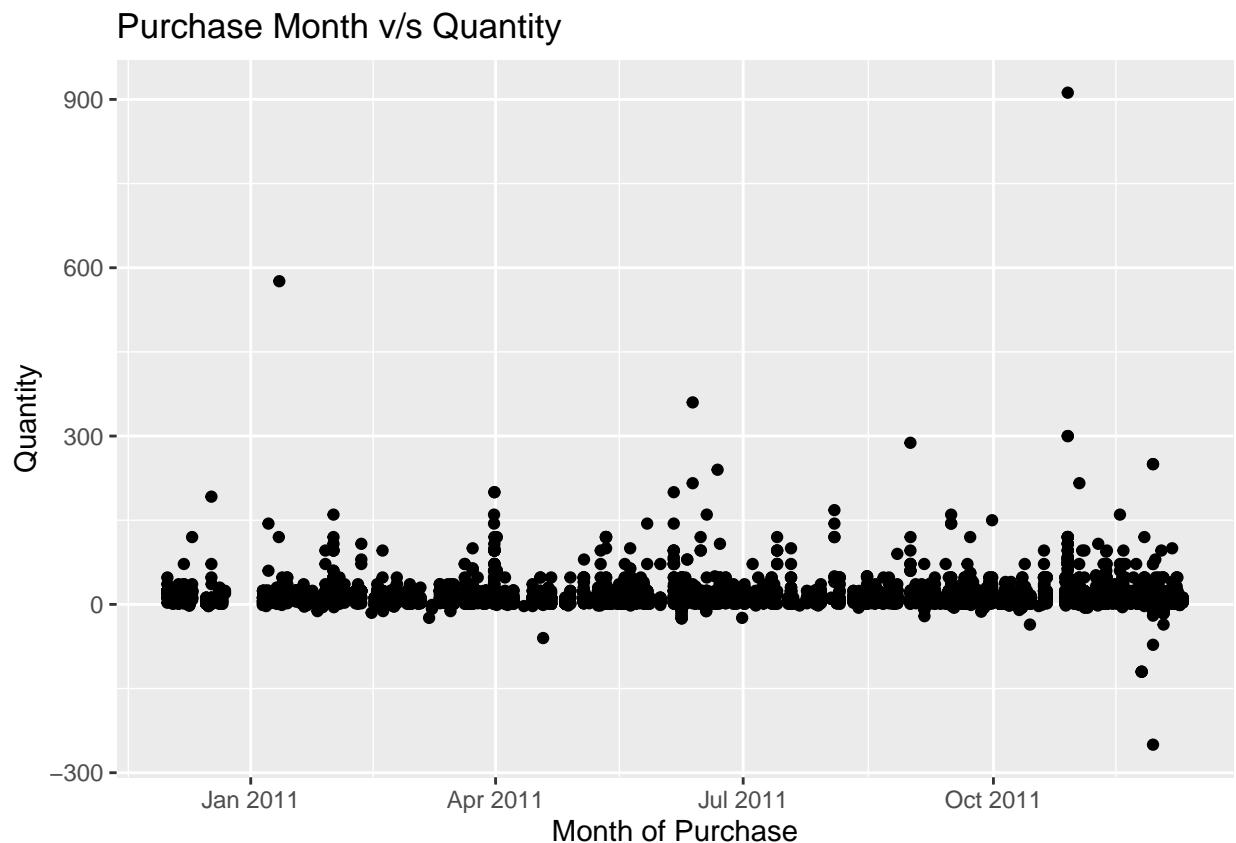
	InvoiceNo	StockCode	Description
##	570672	POST : 259	POSTAGE : 311
##	578541	23084 : 126	RABBIT NIGHT LIGHT : 75
##	569568	21731 : 89	RED TOADSTOOL LED NIGHT LIGHT : 72
##	576927	22554 : 81	PLASTERS IN TIN CIRCUS PARADE : 68
##	569332	22556 : 77	PLASTERS IN TIN WOODLAND ANIMALS : 68
##	552826	22326 : 75	ROUND SNACK BOXES SET OF4 WOODLAND: 65

```

##  (Other):7850  (Other):7898  (Other) :7898
##    Quantity      InvoiceDate          UnitPrice
##  Min.   :-250.00  Min.   :2010-12-01 08:45:00.00  Min.   : 0.000
##  1st Qu.:  5.00  1st Qu.:2011-04-07 13:07:00.00  1st Qu.: 1.250
##  Median : 10.00  Median :2011-08-17 08:50:00.00  Median : 1.790
##  Mean   : 12.91  Mean   :2011-07-13 01:19:17.42  Mean   : 5.029
##  3rd Qu.: 12.00  3rd Qu.:2011-10-19 13:49:00.00  3rd Qu.: 3.750
##  Max.   : 912.00  Max.   :2011-12-09 12:50:00.00  Max.   :4161.060
##
##    CustomerID      Country
##  12681   : 646  France   :8557
##  12682   : 525  Australia: 0
##  12567   : 463  Austria  : 0
##  12637   : 394  Bahrain  : 0
##  12683   : 362  Belgium  : 0
##  (Other):6101  Brazil   : 0
##  NA's    : 66   (Other) : 0

qplot(x = InvoiceDate, y = Quantity,
       xlab = "Month of Purchase",
       ylab = "Quantity",
       main = "Purchase Month v/s Quantity",
       data = frc_data)

```



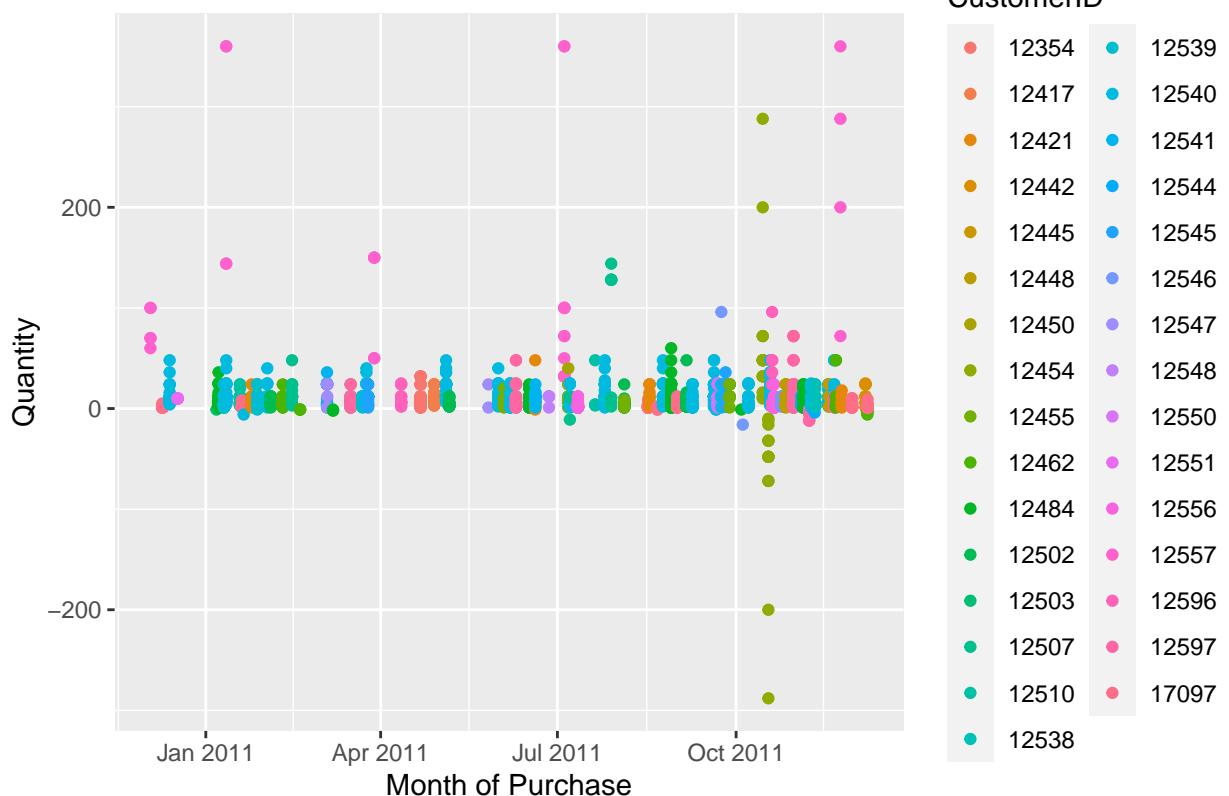
```
#color = CustomerID - FAILURE - Lot of Customers - Inappropriate Graph

# 3) Spain
filter(data2, Country == "Spain") -> spn_data
summary(spn_data)
```

```
##   InvoiceNo      StockCode          Description
## 564734 : 122    POST    : 62    POSTAGE           : 62
## 573362 : 116    22423   : 25    REGENCY CAKESTAND 3 TIER     : 25
## 540469 : 98     22960   : 16    JAM MAKING SET WITH JARS   : 16
## 540550 : 93     22077   : 15    6 RIBBONS RUSTIC CHARM   : 15
## 542303 : 93     22553   : 13    PLASTERS IN TIN SKULLS    : 13
## 559665 : 67     22326   : 12    ASSORTED COLOUR BIRD ORNAMENT: 12
## (Other):1944   (Other):2390  (Other)                   :2390
##   Quantity      InvoiceDate          UnitPrice
## Min.   :-288.00  Min.   :2010-12-03 12:20:00.00  Min.   : 0.000
## 1st Qu.:  3.00   1st Qu.:2011-03-16 14:00:00.00  1st Qu.: 1.250
## Median :  6.00   Median :2011-07-11 13:35:00.00  Median : 2.080
## Mean   : 10.59   Mean   :2011-06-29 16:39:39.26  Mean   : 4.987
## 3rd Qu.: 12.00   3rd Qu.:2011-10-18 12:59:00.00  3rd Qu.: 4.250
## Max.   :360.00   Max.   :2011-12-07 17:05:00.00  Max.   :1715.850
##
##   CustomerID      Country
## 12540  :481     Spain    :2533
## 12484  :350     Australia: 0
## 12539  :274     Austria  : 0
## 12597  :214     Bahrain  : 0
## 17097  :213     Belgium  : 0
## 12502  :147     Brazil   : 0
## (Other):854   (Other)   : 0
```

```
qplot(x = InvoiceDate, y = Quantity,
      data = spn_data,
      xlab = "Month of Purchase",
      ylab = "Quantity",
      main = "Purchase Month v/s Quantity",
      color = CustomerID) #color = CustomerID - MANAGABLE - 31 Customers
```

Purchase Month v/s Quantity



```
# 4) Japan
filter(data2, Country == "Japan") -> jpn_data
summary(jpn_data)
```

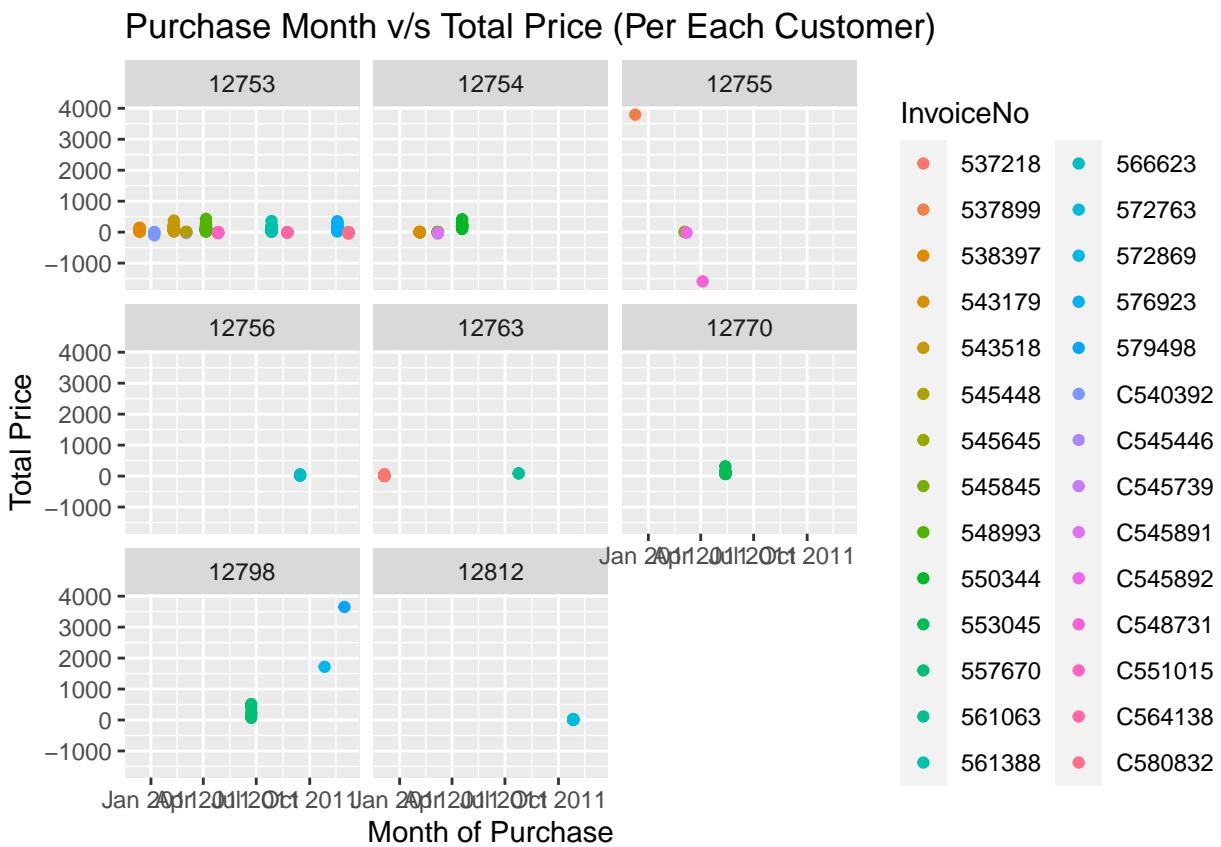
```
##   InvoiceNo      StockCode          Description
## 538397 : 48    21218   : 7    RED SPOTTY BISCUIT TIN      : 7
## 543518 : 48    22661   : 6    CHARLOTTE BAG DOLLY GIRL DESIGN: 6
## 543179 : 37    22489   : 5    LUNCH BAG DOLLY GIRL DESIGN  : 5
## 561388 : 37    22662   : 5    PACK OF 12 TRADITIONAL CRAYONS : 5
## 548993 : 32    23084   : 5    RABBIT NIGHT LIGHT       : 5
## 576923 : 31    21210   : 4    BASKET OF TOADSTOOLS     : 4
## (Other):125   (Other):326   (Other)                      :326
##   Quantity      InvoiceDate        UnitPrice
## Min.   :-624.00  Min.   :2010-12-05 15:46:00.00  Min.   : 0.210
## 1st Qu.:  4.00   1st Qu.:2011-02-04 10:32:00.00  1st Qu.: 0.850
## Median : 36.00   Median :2011-04-05 00:04:00.00  Median : 1.650
## Mean   : 70.44   Mean   :2011-04-23 10:54:33.51  Mean   : 2.276
## 3rd Qu.: 72.00   3rd Qu.:2011-07-27 09:32:00.00  3rd Qu.: 2.550
## Max.   :2040.00   Max.   :2011-12-06 11:40:00.00  Max.   :45.570
##
##   CustomerID      Country
## 12753 : 230    Japan   :358
## 12754 : 65     Australia: 0
## 12763 : 18     Austria : 0
## 12812 : 15     Bahrain: 0
```

```

## 12770 : 12 Belgium : 0
## 12798 : 8 Brazil : 0
## (Other): 10 (Other) : 0

# qplot(x = InvoiceDate, y = Quantity*UnitPrice, data = jpn_data, color = CustomerID)
# 8 Customers
qplot(x = InvoiceDate, y = Quantity*UnitPrice,
       data = jpn_data,
       xlab = "Month of Purchase",
       ylab = "Total Price",
       main = "Purchase Month v/s Total Price (Per Each Customer)",
       color = InvoiceNo,
       facets = ~ CustomerID)

```



```

# TotalPrice (UnitPrice x Quantity)

# 5) Italy
filter(data2, Country == "Italy") -> ity_data
summary(ity_data)

```

```

##   InvoiceNo StockCode Description
## 1 577609    POST    POSTAGE          : 18
## 2 541115   22720 SET OF 3 CAKE TINS PANTRY DESIGN : 13
## 3 571670   22960    JAM MAKING SET WITH JARS      : 11
## 4 562528   22847 BREAD BIN DINER STYLE IVORY        : 8

```

```

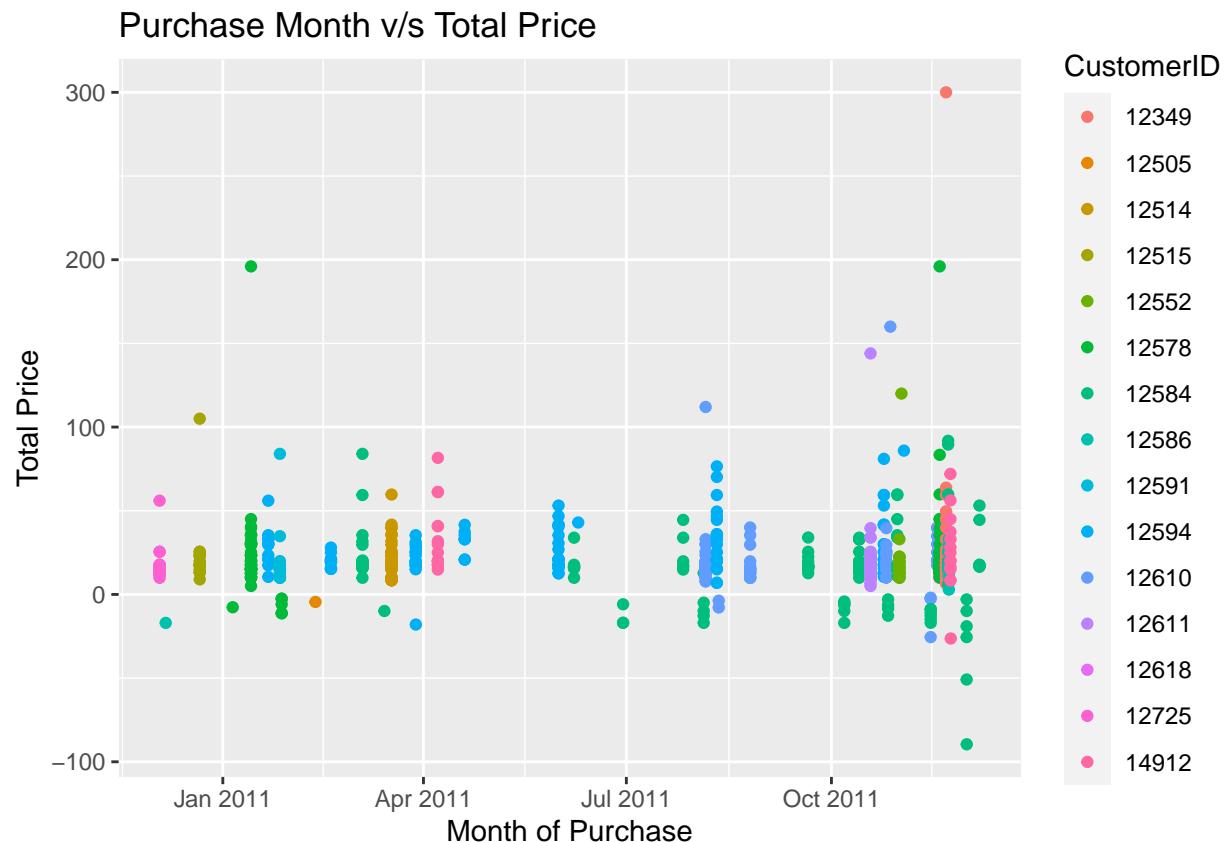
## 546875 : 51  23243   : 8   SET OF TEA COFFEE SUGAR TINS PANTRY: 8
## 577316 : 44  22139   : 7   RETROSPOT TEA SET CERAMIC 11 PC    : 7
## (Other):446 (Other):738 (Other)                           :738
##      Quantity          InvoiceDate            UnitPrice
## Min.   :-12.000   Min.   :2010-12-03 15:45:00.00   Min.   : 0.120
## 1st Qu.:  4.000   1st Qu.:2011-03-17 16:38:00.00   1st Qu.: 1.570
## Median :  6.000   Median :2011-08-25 12:57:00.00   Median : 2.550
## Mean   :  9.961   Mean   :2011-07-16 16:29:56.25   Mean   : 4.831
## 3rd Qu.: 12.000   3rd Qu.:2011-10-31 12:11:00.00   3rd Qu.: 4.950
## Max.   :200.000   Max.   :2011-12-06 09:35:00.00   Max.   :300.000
##
##      CustomerID        Country
## 12584   :126   Italy     :803
## 12578   :120   Australia: 0
## 12594   :119   Austria   : 0
## 12610   :111   Bahrain   : 0
## 12349   : 73   Belgium   : 0
## 12611   : 65   Brazil    : 0
## (Other):189 (Other)   : 0

```

```

qplot(x = InvoiceDate, y = Quantity*UnitPrice,
       data = ity_data,
       xlab = "Month of Purchase",
       ylab = "Total Price",
       main = "Purchase Month v/s Total Price",
       color = CustomerID)

```



```

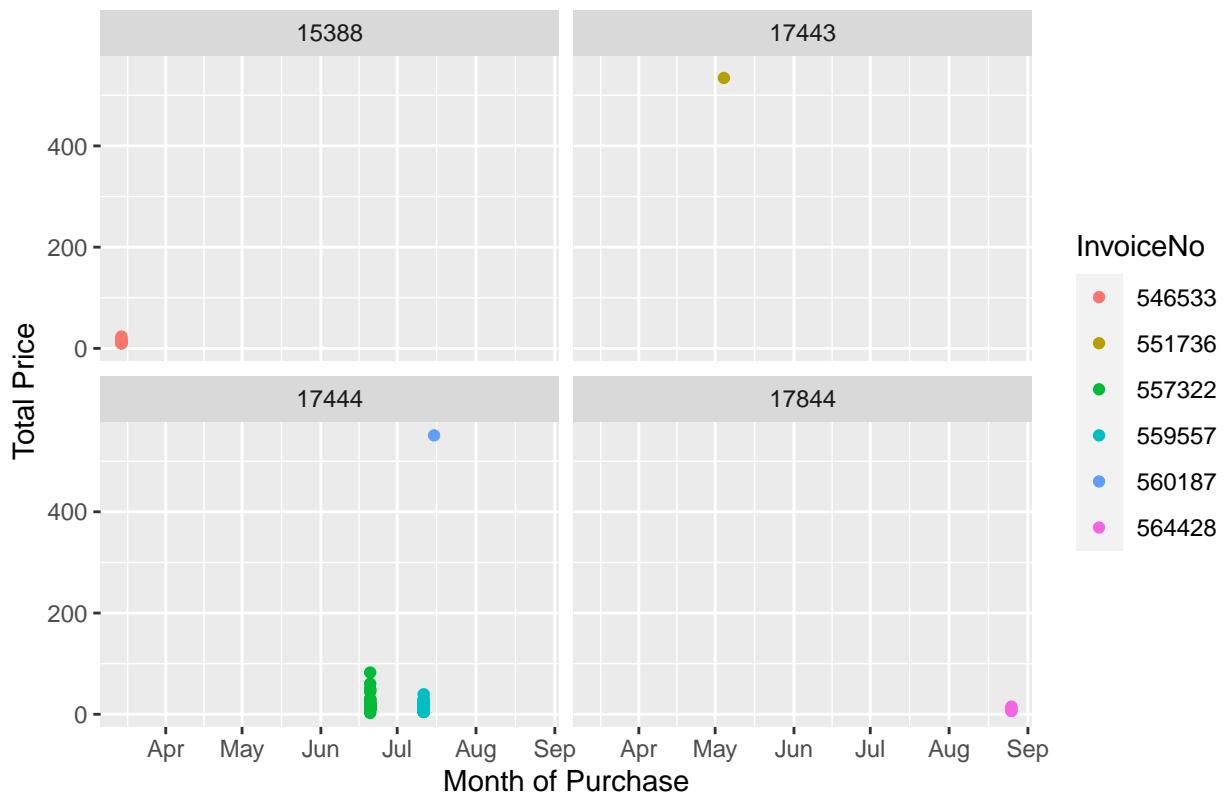
# 6) Canada
filter(data2, Country == "Canada") -> cnd_data
summary(cnd_data)

##      InvoiceNo      StockCode          Description
## 559557 :77    10133 : 2 COLOURING PENCILS BROWN TUBE      : 3
## 557322 :57    23190 : 2 BUNDLE OF 3 ALPHABET EXERCISE BOOKS: 2
## 546533 :10    23192 : 2 BUNDLE OF 3 SCHOOL EXERCISE BOOKS : 2
## 564428 : 5    79030D : 2 10 COLOUR SPACEBOY PEN           : 1
## 551736 : 1    10135 : 1 12 PENCILS TALL TUBE POSY         : 1
## 560187 : 1    15044A : 1 4 TRADITIONAL SPINNING TOPS       : 1
## (Other): 0    (Other):141 (Other)                      :141
##      Quantity      InvoiceDate        UnitPrice
## Min.   : 1.0   Min.   :2011-03-14 13:53:00.00  Min.   : 0.10
## 1st Qu.: 6.0   1st Qu.:2011-06-20 09:04:00.00  1st Qu.: 0.83
## Median :12.0   Median :2011-07-11 10:33:00.00  Median : 1.65
## Mean   :18.3   Mean   :2011-06-26 16:27:03.57  Mean   : 6.03
## 3rd Qu.:20.0   3rd Qu.:2011-07-11 10:33:00.00  3rd Qu.: 2.95
## Max.   :504.0   Max.   :2011-08-25 11:27:00.00  Max.   :550.94
##
##      CustomerID      Country
## 17444  :135    Canada   :151
## 15388  : 10    Australia: 0
## 17844  : 5     Austria  : 0
## 17443  : 1     Bahrain  : 0
## 12346  : 0     Belgium  : 0
## 12347  : 0     Brazil   : 0
## (Other): 0    (Other)  : 0

# qplot(x = InvoiceDate, y = Quantity*UnitPrice, data = cnd_data, color = CustomerID)
qplot(x = InvoiceDate, y = Quantity*UnitPrice,
       data = cnd_data,
       xlab = "Month of Purchase",
       ylab = "Total Price",
       main = "Purchase Month v/s Total Price (Per Each Customer)",
       color = InvoiceNo,
       facets = ~ CustomerID)

```

Purchase Month v/s Total Price (Per Each Customer)



```
# 7) Unspecified
filter(data2, Country == "Unspecified") -> unsp_data
summary(unsp_data)
```

```
##   InvoiceNo StockCode          Description
## 561658 :83  22150 : 4  3 STRIPEY MICE FELTCRAFT : 4
## 559521 :72  20983 : 3  12 PENCILS TALL TUBE RED RETROSPOT: 3
## 565303 :66  21124 : 3  4 TRADITIONAL SPINNING TOPS : 3
## 561661 :51  21591 : 3  ASSORTED COLOUR BIRD ORNAMENT : 3
## 552695 :47  21888 : 3  BINGO SET : 3
## 578539 :34  21889 : 3  CHILDRENS CUTLERY DOLLY GIRL : 3
## (Other):93 (Other):427 (Other) :427
##   Quantity      InvoiceDate        UnitPrice
## Min.   : 1.000  Min.   :2011-04-11 13:29:00.00  Min.   : 0.19
## 1st Qu.: 1.000  1st Qu.:2011-07-08 16:26:00.00  1st Qu.: 0.85
## Median : 3.000  Median :2011-07-28 16:06:00.00  Median : 1.65
## Mean   : 7.399  Mean   :2011-07-30 15:13:21.66  Mean   : 2.70
## 3rd Qu.:12.000  3rd Qu.:2011-09-02 12:17:00.00  3rd Qu.: 3.35
## Max.   :48.000  Max.   :2011-11-24 14:55:00.00  Max.   :16.95
##
##   CustomerID      Country
## 12743 :134  Unspecified:446
## 16320 : 56   Australia   : 0
## 14265 : 31   Austria    : 0
## 12363 : 23   Bahrain    : 0
```

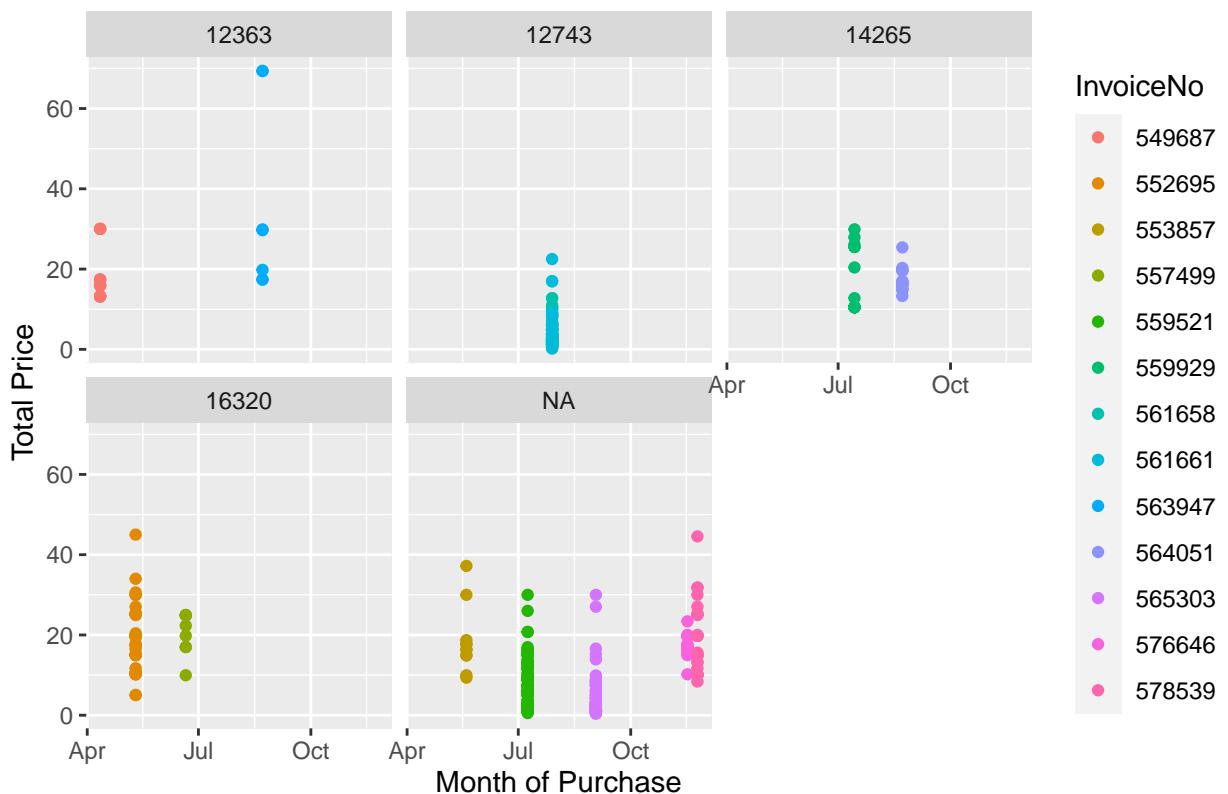
```

## 12346 : 0 Belgium : 0
## (Other): 0 Brazil : 0
## NA's :202 (Other) : 0

# qplot(x = InvoiceDate, y = Quantity*UnitPrice, data = unsp_data, color = CustomerID)
qplot(x = InvoiceDate, y = Quantity*UnitPrice,
       data = unsp_data,
       xlab = "Month of Purchase",
       ylab = "Total Price",
       main = "Purchase Month v/s Total Price (Per Each Customer)",
       color = InvoiceNo,
       facets = ~ CustomerID)

```

Purchase Month v/s Total Price (Per Each Customer)



```
# Last graph consists of the data of the Invoices where Country name was not mentioned
```