Data Set

Nikunj Gupta

August 2018

1 Data Set information

It is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 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 Set Characteristics: Multivariate, Sequential, Time-Series
- Number of Instances: 541909
- Attribute Characteristics: Integer, Real
- Number of Attributes: 8

2 Attributes

2.1 InvoiceNo

Invoice number.
Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
Thoughts: Can be used to segregate orders. Considering a set of items as a single order.

2.2 StockCode

StockCode
Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
Thoughts: Itemtype.

2.3 Description

Product (item) name. Nominal. Thoughts: Names are too specific. Can be reduced to some common form. Other data preprocessing can also be done here. (Actually, Stock codes are enough for the missing values here.)

2.4 Quantity

The quantities of each product (item) per transaction. Numeric.
Thoughts: Can used to figure out the general demand amount per transaction.

2.5 InvoiceDate

Invoice Date and time. Numeric, the day and time when each transaction was generated.
Thoughts: The data is for a period of 1 year. So, only seasonal demands can be forecasted.

2.6 UnitPrice

Unit price. Numeric, Product price per unit in sterling. Thoughts: ItemSize would have been more useful perhaps.

2.7 CustomerID

Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Thoughts: Can be useful if we want to do Collaborative Filtering (Recommendation Systems).

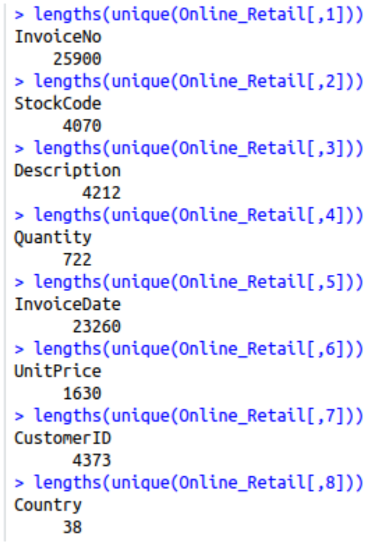
2.8 Country

Country name. Nominal, the name of the country where each customer resides.
Thoughts: Can be used to when we have many warehouses spread across a country/continent and we want our RL agent to figure out individual demands of these locations and accordingly manage the supply and placement of products.

3 Data Visualization

3.1 Unique Elements

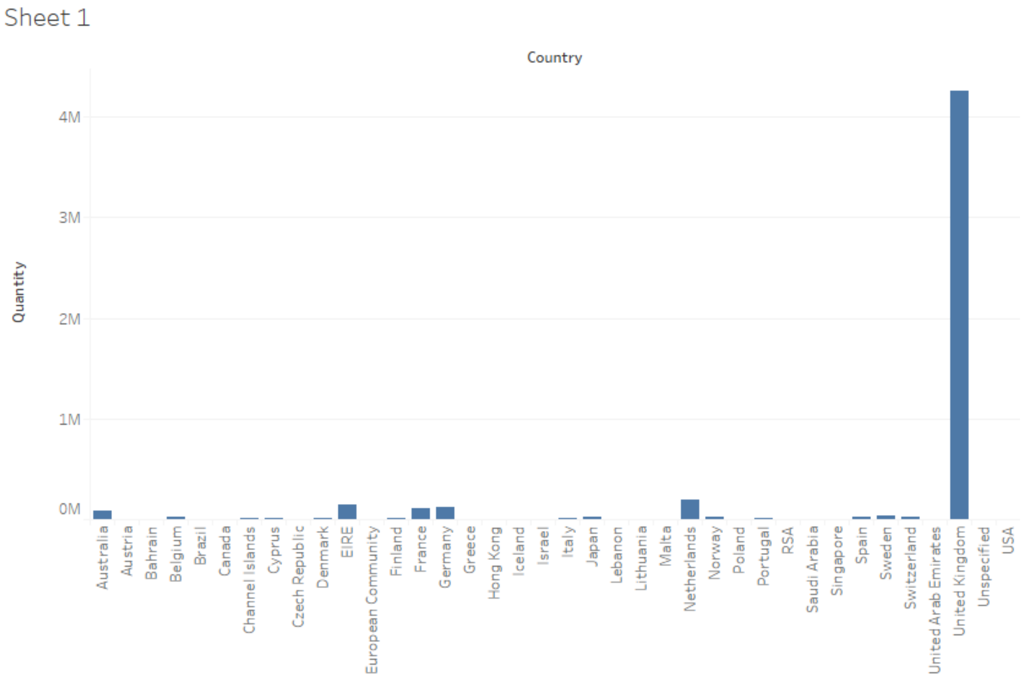
The following are the number of unique elements present in the dataset. For example, we have 4212 different products (Description) and 38 unique countries' data.



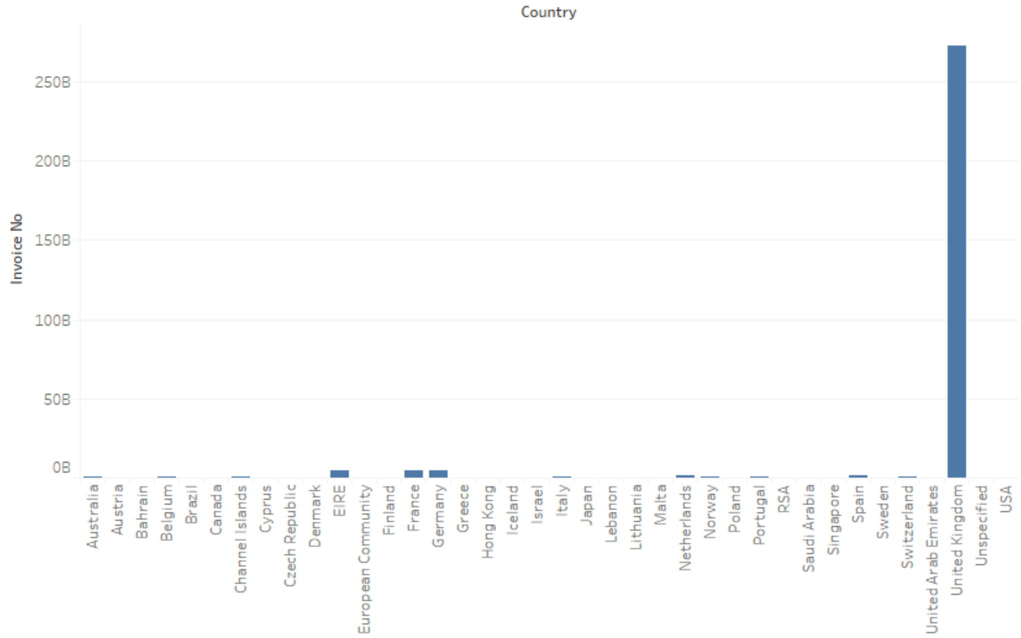
```
> lengths(unique(Online_Retail[,1]))
InvoiceNo
    25900
> lengths(unique(Online_Retail[,2]))
StockCode
     4070
> lengths(unique(Online_Retail[,3]))
Description
       4212
> lengths(unique(Online_Retail[,4]))
Quantity
     722
> lengths(unique(Online_Retail[,5]))
InvoiceDate
      23260
> lengths(unique(Online_Retail[,6]))
UnitPrice
     1630
> lengths(unique(Online_Retail[,7]))
CustomerID
      4373
> lengths(unique(Online_Retail[,8]))
Country
     38
```

3.2 Country vs Quantity/InvoiceNo

This plot shows that UK has the most number of items ordered over the year. Here, the country attribute probably can be dropped and all products could be considered from the same place.

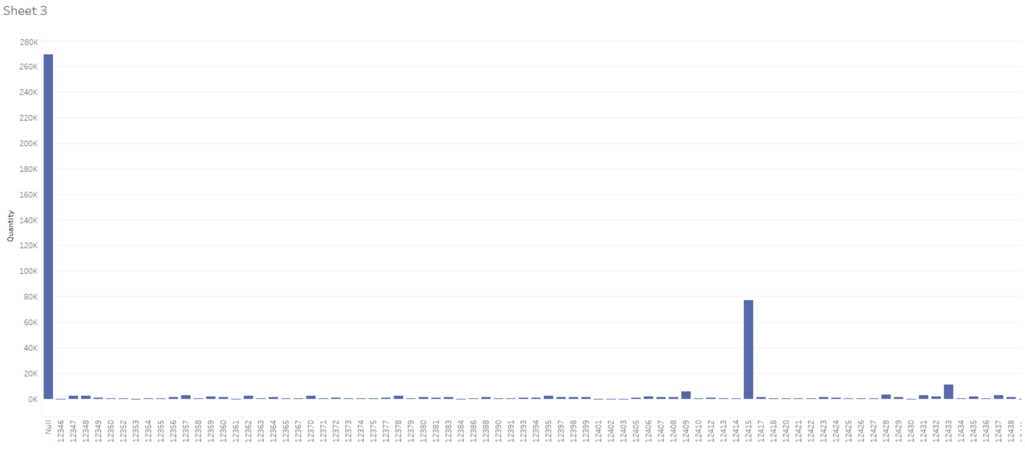


Sheet 1



3.3 CustomerID vs Quantity

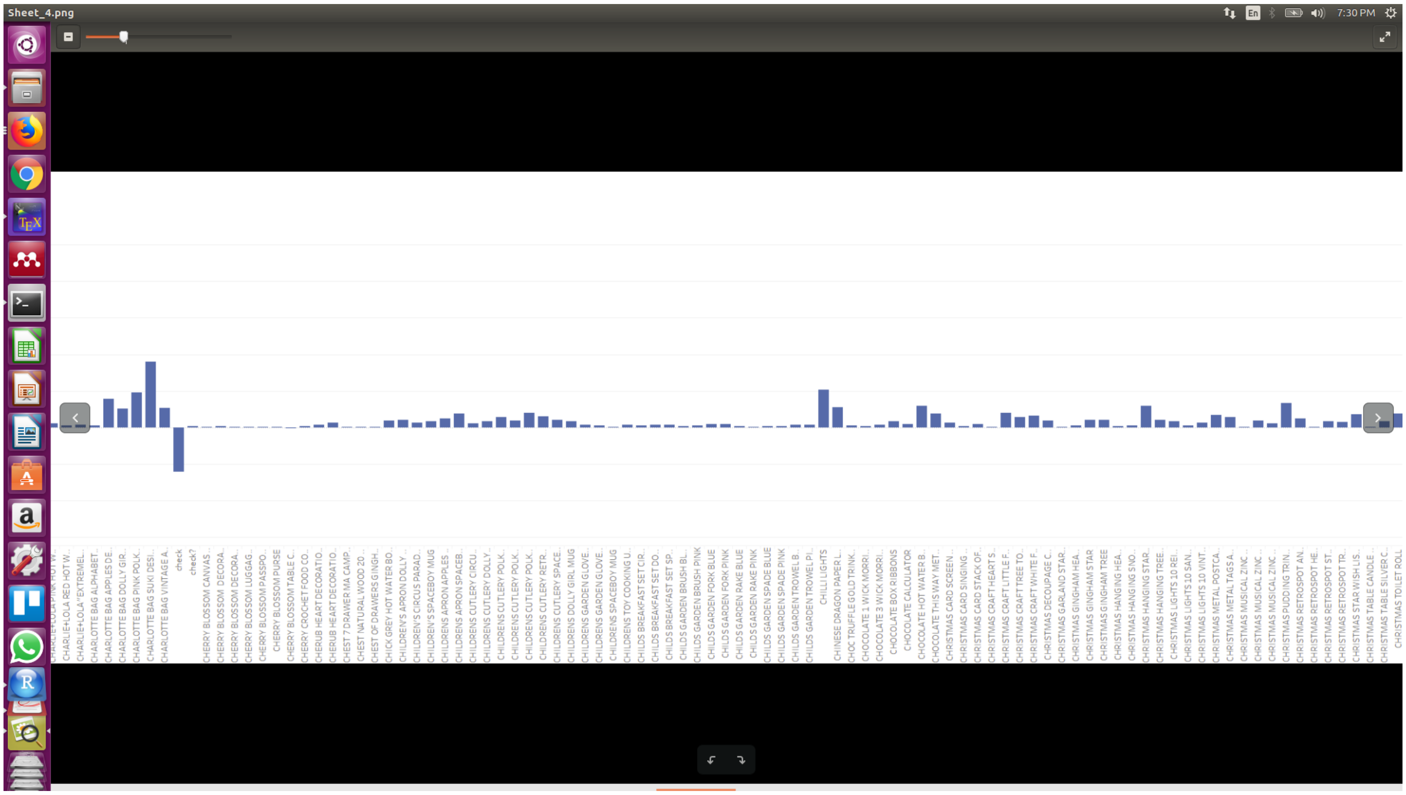
[redacted]



[redacted]

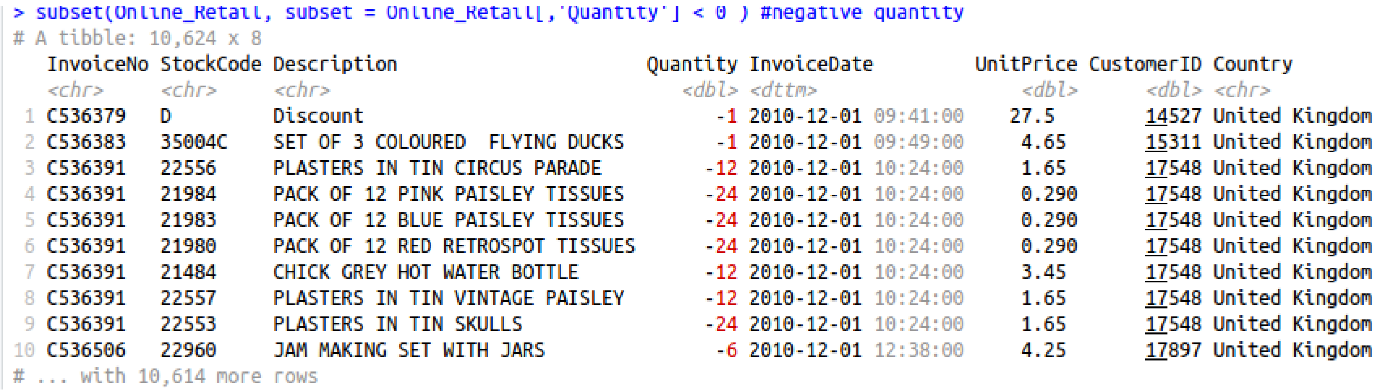
Majority of the CustomerIDs were NULL. So, we must omit this attribute.

3.4 Description vs Quantity


Sheet_4.png
En
7:30 PM
CHARLIE+LOLA PINK HOT W...
CHARLIE+LOLA RED HOT W...
CHARLIE+LOLA"EXTREMEL...
CHARLOTTE BAG ALPHABET...
CHARLOTTE BAG APPLES DE...
CHARLOTTE BAG DOLLY GIR...
CHARLOTTE BAG PINK POLK...
CHARLOTTE BAG SUKI DESI...
CHARLOTTE BAG VINTAGE A...
check
check?
CHERRY BLOSSOM CANVAS ...
CHERRY BLOSSOM DECORA...
CHERRY BLOSSOM DECORA...
CHERRY BLOSSOM LUGGAG...
CHERRY BLOSSOM PASSPO...
CHERRY BLOSSOM PURSE
CHERRY BLOSSOM TABLE C...
CHERRY CROCHET FOOD CO...
CHERUB HEART DECORATIO...
CHERUB HEART DECORATIO...
CHEST 7 DRAWER MA CAMP...
CHEST NATURAL WOOD 20 ...
CHEST OF DRAWERS GINGH...
CHICK GREY HOT WATER BO...
CHILDREN'S APRON DOLLY ...
CHILDREN'S CIRCUS PARAD...
CHILDREN'S SPACEBOY MUG
CHILDRENS APRON APPLES ...
CHILDRENS APRON SPACEB...
CHILDRENS CUTLERY CIRCU...
CHILDRENS CUTLERY DOLLY ...
CHILDRENS CUTLERY POLK...
CHILDRENS CUTLERY POLK...
CHILDRENS CUTLERY POLK...
CHILDRENS CUTLERY RETR...
CHILDRENS CUTLERY SPACE...
CHILDRENS DOLLY GIRL MUG
CHILDRENS GARDEN GLOVE...
CHILDRENS GARDEN GLOVE...
CHILDRENS SPACEBOY MUG
CHILDRENS TOY COOKING U...
CHILDS BREAKFAST SET CIR...
CHILDS BREAKFAST SET DO...
CHILDS BREAKFAST SET SP...
CHILDS GARDEN BRUSH BL...
CHILDS GARDEN BRUSH PINK
CHILDS GARDEN FORK BLUE
CHILDS GARDEN FORK PINK
CHILDS GARDEN RAKE BLUE
CHILDS GARDEN RAKE PINK
CHILDS GARDEN SPADE BLUE
CHILDS GARDEN SPADE PINK
CHILDS GARDEN TROWEL B...
CHILDS GARDEN TROWEL PI...
CHILLI LIGHTS
CHINESE DRAGON PAPER L...
CHOC TRUFFLE GOLD TRINK...
CHOCOLATE 1 WICK MORRI...
CHOCOLATE 3 WICK MORRI...
CHOCOLATE BOX RIBBONS
CHOCOLATE CALCULATOR
CHOCOLATE HOT WATER B...
CHOCOLATE THIS WAY MET...
CHRISTMAS CARD SCREEN ...
CHRISTMAS CARD SINGING ...
CHRISTMAS CARD STACK OF...
CHRISTMAS CRAFT HEART S...
CHRISTMAS CRAFT LITTLE F...
CHRISTMAS CRAFT TREE TO...
CHRISTMAS CRAFT WHITE F...
CHRISTMAS DECOUPAGE C...
CHRISTMAS GARLAND STAR...
CHRISTMAS GINGHAM HEA...
CHRISTMAS GINGHAM STAR
CHRISTMAS GINGHAM TREE
CHRISTMAS HANGING HEA...
CHRISTMAS HANGING SNO...
CHRISTMAS HANGING STAR...
CHRISTMAS HANGING TREE...
CHRISTMAS LIGHTS 10 REI...
CHRISTMAS LIGHTS 10 SAN...
CHRISTMAS LIGHTS 10 VINT...
CHRISTMAS METAL POSTCA...
CHRISTMAS METAL TAGS A...
CHRISTMAS MUSICAL ZINC ...
CHRISTMAS MUSICAL ZINC ...
CHRISTMAS MUSICAL ZINC ...
CHRISTMAS PUDDING TRIN...
CHRISTMAS RETROSPOT AN...
CHRISTMAS RETROSPOT HE...
CHRISTMAS RETROSPOT ST...
CHRISTMAS RETROSPOT TR...
CHRISTMAS STAR WISH LIS...
CHRISTMAS TABLE CANDLE ...
CHRISTMAS TABLE SILVER C...
CHRISTMAS TOILET ROLL

This is a snapshot of some of the 4000+ products. Point to be noticed here is that some prodcuts have negative quantity here. The data needs preprocessing.

3.5 Negative Quantity



```
> subset(Online_Retail, subset = Online_Retail[,'Quantity'] < 0 ) #negative quantity
# A tibble: 10,624 x 8
   InvoiceNo StockCode Description                        Quantity InvoiceDate         UnitPrice CustomerID Country
   <chr>     <chr>     <chr>                                 <dbl> <dttm>                  <dbl>      <dbl> <chr>
 1 C536379   D         Discount                                 -1 2010-12-01 09:41:00    27.5        14527 United Kingdom
 2 C536383   35004C    SET OF 3 COLOURED  FLYING DUCKS          -1 2010-12-01 09:49:00     4.65       15311 United Kingdom
 3 C536391   22556     PLASTERS IN TIN CIRCUS PARADE           -12 2010-12-01 10:24:00     1.65       17548 United Kingdom
 4 C536391   21984     PACK OF 12 PINK PAISLEY TISSUES         -24 2010-12-01 10:24:00     0.290      17548 United Kingdom
 5 C536391   21983     PACK OF 12 BLUE PAISLEY TISSUES         -24 2010-12-01 10:24:00     0.290      17548 United Kingdom
 6 C536391   21980     PACK OF 12 RED RETROSPOT TISSUES        -24 2010-12-01 10:24:00     0.290      17548 United Kingdom
 7 C536391   21484     CHICK GREY HOT WATER BOTTLE             -12 2010-12-01 10:24:00     3.45       17548 United Kingdom
 8 C536391   22557     PLASTERS IN TIN VINTAGE PAISLEY         -12 2010-12-01 10:24:00     1.65       17548 United Kingdom
 9 C536391   22553     PLASTERS IN TIN SKULLS                  -24 2010-12-01 10:24:00     1.65       17548 United Kingdom
10 C536506   22960     JAM MAKING SET WITH JARS                 -6 2010-12-01 12:38:00     4.25       17897 United Kingdom
# ... with 10,614 more rows
```