Phase 4

Data Analytics Component

Anuradha Nitin Bhave
ab5890@rit.edu

Asmita Hari
ah1743@rit.edu

Himanshi Chetwani
hc9165@rit.edu

Ishika Prasad
ip1262@rit.edu

Poornima Sapkal
ps5067@rit.edu

ABSTRACT

1. DATA EXPLORATION AND PRE-PROCESSING/CLEANING

1.1 Data Processing to pick important attributes

- To find the association rules we first filtered on the invoice number to check if multiple products had the same invoice number.
- We filtered the data by product description to make sure that the product description was consistent with all the invoice numbers.
- Attributes that we're concerned with - invoice number and product description
- Group the product description by invoice number
- This generates a set of items that were bought together
- Perform Market Basket Analysis on this data

1.2 Description of the Data Set

We have chosen the online retail data set for our data mining phase. This is a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based online retail store. The company mainly sells unique all-occasion gifts.

Before data Exploration and pre-processing/cleaning, the data set contains the following attributes :

- InvoiceNo: Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Nominal, the name of the country where each customer resides.

After data Exploration and pre-processing/cleaning the data set contains the following attributes :

- InvoiceNo: Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- Description: Product (item) name. Nominal

1.3 Summary of each column

```
> data = read.csv("Online_Retail.csv", header = T, na.strings="?")
> summary(data)
   InvoiceNo         StockCode                                Description
 573585 :  1114   85123A :  2313   WHITE HANGING HEART T-LIGHT HOLDER:  2369
 581219 :   749   22423  :  2203   REGENCY CAKESTAND 3 TIER          :  2200
 581492 :   731   85099B :  2159   JUMBO BAG RED RETROSPOT           :  2159
 580729 :   721   47566  :  1727   PARTY BUNTING                     :  1727
 558475 :   705   20725  :  1639   LUNCH BAG RED RETROSPOT           :  1638
 579777 :   687   84879  :  1502   (Other)                           :531769
 (Other):537202   (Other):530366   NA's                              :    47
    Quantity                 InvoiceDate       UnitPrice           CustomerID
 Min.   :-80995.00   10/31/11 14:41:  1114   Min.   :-11062.06   Min.   :12346
 1st Qu.:     1.00   12/8/11 9:28   :   749   1st Qu.:     1.25   1st Qu.:13953
 Median :     3.00   12/9/11 10:03  :   731   Median :     2.08   Median :15152
 Mean   :     9.55   12/5/11 17:24  :   721   Mean   :     4.61   Mean   :15288
 3rd Qu.:    10.00   6/29/11 15:58  :   705   3rd Qu.:     4.13   3rd Qu.:16791
 Max.   : 80995.00   11/30/11 15:13:   687   Max.   : 38970.00   Max.   :18287
                     (Other)        :537202                       NA's   :135080
         Country
 United Kingdom:495478
 Germany       :  9495
 France        :  8557
 EIRE          :  8196
 Spain         :  2533
 Netherlands   :  2371
 (Other)       : 15279
```

Figure 1: Summary of data set without any data being cleaned

Refer to Figure 1 - Original Data Summary The data set describes a purchase of items from the online retail store

through the following attributes: Invoice Number, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country.In this data set, invoice number is used to identify a bill, and stock code is used to identify an item. Thus, multiple items, that have been purchased in the same transaction, have a unique stock code but the same invoice number.

```
 i..InvoiceNo   StockCode                          Description    Quantity
540458 : 149   POST   : 383   POSTAGE                      : 383   Min.   :-288.00
555383 : 134   22326  : 120   ROUND SNACK BOXES SET OF4 WOODLAND : 120   1st Qu.:   5.00
571328 :  99   22423  :  81   REGENCY CAKESTAND 3 TIER     :  81   Median :  10.00
564856 :  90   22328  :  78   ROUND SNACK BOXES SET OF 4 FRUITS :  78   Mean   :  12.37
557466 :  83   22554  :  67   PLASTERS IN TIN WOODLAND ANIMALS  :  67   3rd Qu.:  12.00
571824 :  80   20719  :  59   WOODLAND CHARLOTTE BAG       :  59   Max.   : 600.00
(Other):8860   (Other):8707   (Other)                    :8707
       InvoiceDate     UnitPrice          CustomerID       Country
1/7/2011 12:28  : 149   Min.   :    0.000   Min.   :12426   Germany:9495
6/2/2011 15:13  : 134   1st Qu.:    1.250   1st Qu.:12480
10/17/2011 11:27:  99   Median :    1.950   Median :12592
8/31/2011 9:11  :  90   Mean   :    3.967   Mean   :12646
6/20/2011 13:08 :  83   3rd Qu.:    3.750   3rd Qu.:12662
10/19/2011 11:49:  80   Max.   :  599.500   Max.   :14335
(Other)         :8860
```

Figure 2: Summary of data set with only Germany's data

```
 InvoiceNo                          Description
540458 : 149   POSTAGE                      : 383
555383 : 125   ROUND SNACK BOXES SET OF4 WOODLAND : 119
571328 :  99   REGENCY CAKESTAND 3 TIER     :  81
564856 :  90   ROUND SNACK BOXES SET OF 4 FRUITS :  78
557466 :  83   PLASTERS IN TIN WOODLAND ANIMALS  :  66
571824 :  80   WOODLAND CHARLOTTE BAG       :  59
(Other):8840   (Other)                      :8680
>
```

Refer to Figure 2 - Represents summary of the data set which contains all records of just Germany.

Figure 3: Summary of data set with only Germany's data

Refer to figure 3 above - We are cleaning data using -

- Removing missing values
- Deleting duplicate entries

Cleaning of the data set included considering only necessary attributes, removing null values, and removal of duplicate elements. We executed the following commands in the given order for cleaning our data: Data <- read.csv("OnlineRetail Germany.csv", header = T) which reads the csv file into a data frame called Data Data.StockCode <- NULL which sets attribute StockCode to null Similarly, we set Invoice Date, Quantity, Country, Unit Price and CustomerID to NULL. Then, in order to handle data duplication, we executed the following commands: uData = unique(Data) unique is an R command whose input is our data frame, and output is the same data frame with duplicate values removed. dupData=Data[!duplicated(Data),] The dupData data frame consists of data that does not have any duplicate values, and is further cleaned to remove null values using the following command: dupData=na.omit(dupData) na.omit() takes a data frame as its argument and returns the data frame with the omission of null values. The execution of these commands gives us data that consists of just the invoice number and product description, and is in atomic form, where one row consists of the invoice number and the description for one product.
Figure below shows steps to remove Duplicate Data

```
> #Removing duplicate data from Germany dataset
> duData=unique(Data)
> dupData=Data[!duplicated(Data), ]
> summary(dupData)
   InvoiceNo      StockCode
 540458 : 149   POST   : 383
 555383 : 132   22326  : 119
 571328 :  99   22423  :  81
 564856 :  90   22328  :  78
 557466 :  83   22554  :  66
 571824 :  80   20719  :  59
 (Other):8847   (Other):8694
                              Description      Quantity
 POSTAGE                             : 383   Min.   :-288.00
 ROUND SNACK BOXES SET OF4 WOODLAND : 119   1st Qu.:   5.00
 REGENCY CAKESTAND 3 TIER            :  81   Median :  10.00
 ROUND SNACK BOXES SET OF 4 FRUITS   :  78   Mean   :  12.38
 PLASTERS IN TIN WOODLAND ANIMALS    :  66   3rd Qu.:  12.00
 WOODLAND CHARLOTTE BAG              :  59   Max.   : 600.00
 (Other)                             :8694
       InvoiceDate     UnitPrice          CustomerID
 1/7/11 12:28  : 149   Min.   :    0.00   Min.   :12426
 6/2/11 15:13  : 132   1st Qu.:    1.25   1st Qu.:12480
 10/17/11 11:27:  99   Median :    1.95   Median :12592
 8/31/11 9:11  :  90   Mean   :    3.97   Mean   :12646
 6/20/11 13:08 :  83   3rd Qu.:    3.75   3rd Qu.:12662
 10/19/11 11:49:  80   Max.   :  599.50   Max.   :14335
 (Other)         :8847
     Country
 Germany  :9480
 Australia:   0
 Austria  :   0
 Bahrain  :   0
 Belgium  :   0
 Brazil   :   0
 (Other)  :   0
```

Figure 4: Summary of data after removing duplicate data

Figure below shows steps to remove missing values

```
> #Removing the missing values from Germany dataset
> dupData=na.omit(dupData)
> summary(dupData)
   InvoiceNo       StockCode
 540458 : 149   POST    : 383
 555383 : 132   22326   : 119
 571328 :  99   22423   :  81
 564856 :  90   22328   :  78
 557466 :  83   22554   :  66
 571824 :  80   20719   :  59
 (Other):8847   (Other):8694
                                Description
 POSTAGE                                 : 383
 ROUND SNACK BOXES SET OF4 WOODLAND      : 119
 REGENCY CAKESTAND 3 TIER                :  81
 ROUND SNACK BOXES SET OF 4 FRUITS       :  78
 PLASTERS IN TIN WOODLAND ANIMALS        :  66
 WOODLAND CHARLOTTE BAG                  :  59
 (Other)                                 :8694
    Quantity              InvoiceDate      UnitPrice
 Min.   :-288.00   1/7/11 12:28  : 149   Min.    :  0.00
 1st Qu.:   5.00   6/2/11 15:13  : 132   1st Qu.:  1.25
 Median :  10.00   10/17/11 11:27:  99   Median :  1.95
 Mean   :  12.38   8/31/11 9:11  :  90   Mean   :  3.97
 3rd Qu.:  12.00   6/20/11 13:08 :  83   3rd Qu.:  3.75
 Max.   : 600.00   10/19/11 11:49:  80   Max.   :599.50
                   (Other)       :8847
   CustomerID         Country
 Min.   :12426   Germany  :9480
 1st Qu.:12480   Australia:   0
 Median :12592   Austria  :   0
 Mean   :12646   Bahrain  :   0
 3rd Qu.:12662   Belgium  :   0
 Max.   :14335   Brazil   :   0
                 (Other)  :   0
```

Figure 5: Removing Missing Values

Figure below shows steps to remove attributes not required

```
> #Removing the attributes which is not required
> dupData$StockCode <- NULL
> dupData$InvoiceDate <- NULL
> dupData$Quantity <- NULL
> dupData$Country<- NULL
> dupData$UnitPrice <- NULL
> summary(dupData)
   InvoiceNo                                Description
 540458 : 149   POSTAGE                             : 383
 555383 : 132   ROUND SNACK BOXES SET OF4 WOODLAND  : 119
 571328 :  99   REGENCY CAKESTAND 3 TIER            :  81
 564856 :  90   ROUND SNACK BOXES SET OF 4 FRUITS   :  78
 557466 :  83   PLASTERS IN TIN WOODLAND ANIMALS    :  66
 571824 :  80   WOODLAND CHARLOTTE BAG              :  59
 (Other):8847   (Other)                             :8694
   CustomerID
 Min.   :12426
 1st Qu.:12480
 Median :12592
 Mean   :12646
 3rd Qu.:12662
 Max.   :14335
```

Figure 6: Remove Attributes Not Required

1.4 Visualizing individual attributes

Figure 4 shows the histogram representing the Quantity of the item bought in a particular invoice or can be explained as bought at a given time.

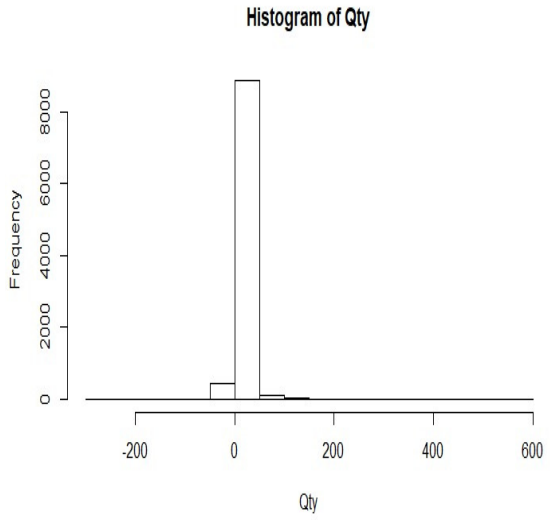


Figure 7: Histogram of column Quantity

Figure 5 shows the histogram representing the Unit Price of the item when it was bought.

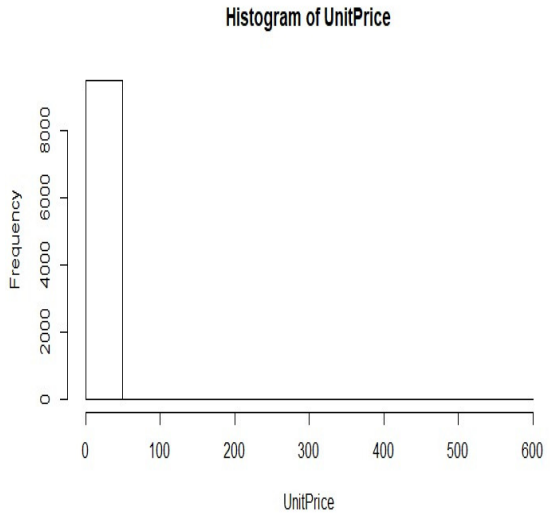


Figure 8: Histogram of column Unit Price

Figure 6 shows the histogram representing the Customer ID of the ID of the customer buying the items, which may be repeated for multiple items

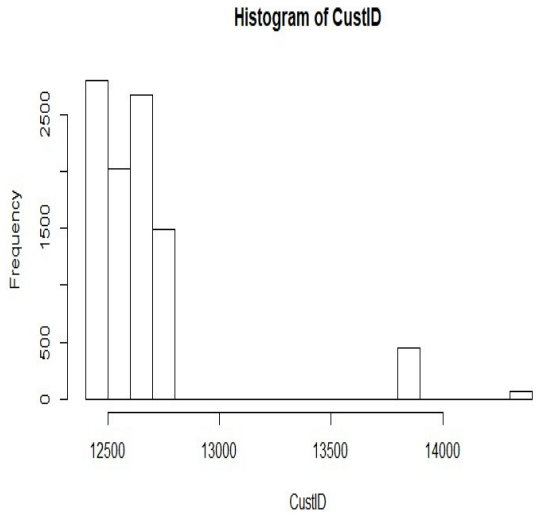


Figure 9: Histogram of Customer ID

Figure 7 shows the pie-chart representing the fact that we have only Germany's data now.

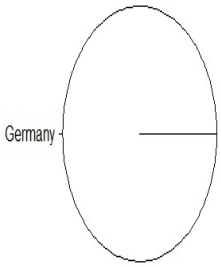


Figure 10: Pie Chart of Country - Only Germany

Figure 8 shows the histogram of the frequency of the invoice number representing the number of items associated with the invoice.

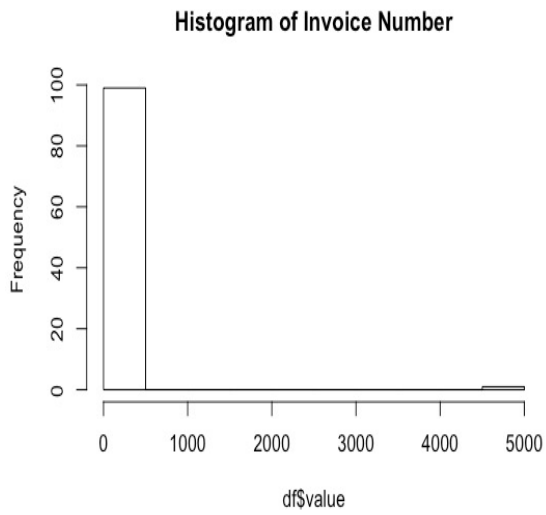


Figure 11: Histogram of the frequency of Invoice Number

Figure 9 shows the histogram of the frequency of the description representing the number of times the item was bought.

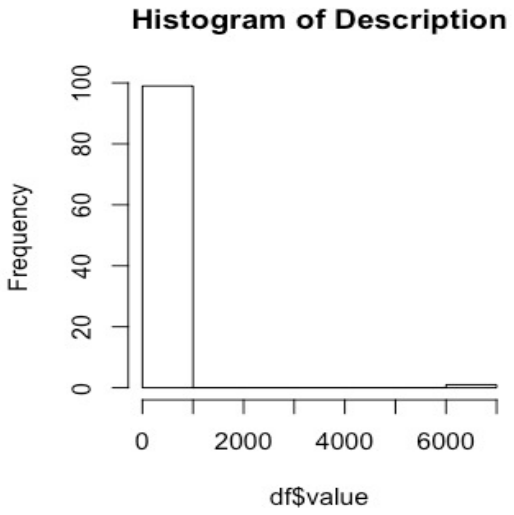


Figure 12: Histogram of the frequency of Description

Figure 10 shows the histogram of the frequency of the stock code

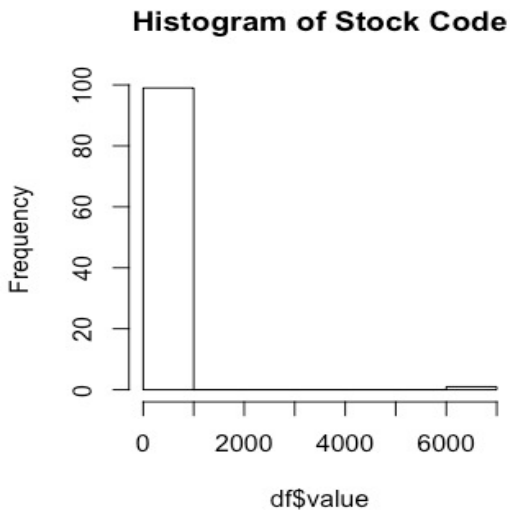


Figure 13: Histogram of the frequency of Stock

2. DATA MINING TECHNIQUE

2.1 Data Mining Technique - Association

Choice of Data Mining Technique - Association Rules

2.2 Data Analytics Task

For our data mining component, we decided to perform a Market Basket Analysis of our online retail data set. We chose this data mining task over classification and clustering because it enabled us to obtain knowledge about which items are frequently purchased together. We believe that this garnered knowledge will be beneficial to the retailer since it will be instrumental in determining which items should be marketed and recommended together. Of the available mining tasks, we focused on market basket analysis because it helped us generate relations between a large set of items. Classification and clustering would have enabled us to determine if a particular data point belonged to a particular class label, which would have been less informative than a market basket analysis.

2.3 What is Market Basket Analysis

What is Market Basket Analysis Market Basket Analysis is a method of detecting associations between items that are frequently purchased together. The output of a market basket analysis is a set of association rules, which establish a relation between a set of items. For our project component, we decided to perform a market basket analysis for all transactions of Germany present in the data set. We're making use of the Apriori Algorithm for generating association rules.

2.4 What is Apriori Algorithm

The Apriori Algorithm is frequently used to generate association rules from transactional data sets. This algorithm works on the principle of generating subsets of item sets, and generating rules based on the frequency of these subsets compared against a thresh hold value. The association rules are generated on the basis of two factors - support and confidence. For our mining component, we have considered 603 transactions, which we obtained after processing the data. We are using the arules package available in R for implementing apriori algorithm and arulesViz package for visualizing the results of the algorithm.

2.5 Support and Confidence

Support and Confidence Support is defined as the percentage of transactions that include all the items of an item set. A higher measure of support indicates a higher frequency of the items occurring together in one transaction. The computation of support requires the computation of support count, which is evaluated by measuring the frequency of occurrence of an item set. Confidence is a measure of probability. Confidence measures how probable it is for a customer to have purchased items on the right hand side of a rule, if they have purchased all items in the item set on the left hand side of the rule. Higher values of support and confidence lead to the generation of stronger rules. A high support means that a particular rule occurs very frequently, and high confidence indicates a higher probability of the rule occurring in a transaction.

3. PROCESS FOLLOWED TO PERFORM MARKET BASKET ANALYSIS

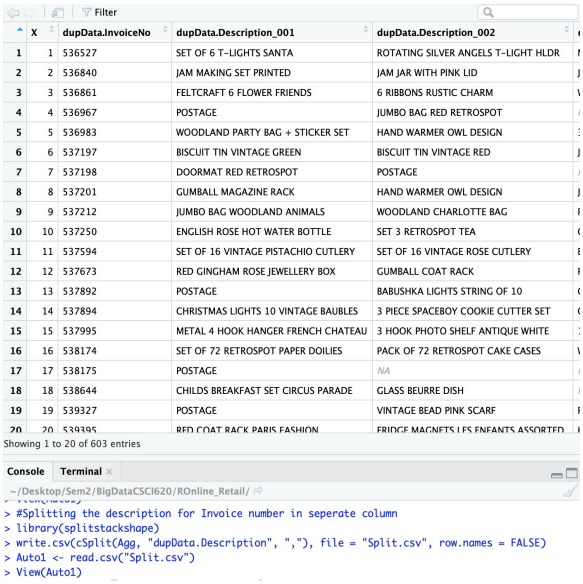
Once the data cleaning is done, steps to perform Market Basket Analysis are:

- Aggregation of the Invoice and product description Figure 11 shows the steps to perform the same

```
> #Aggregation of the Invoice and Product description
> write.csv(x = aggregate(dupData$Description~dupData$InvoiceNo,FUN=toString),file = "Aggregation.csv")
> Agg <- read.csv("Aggregation.csv")
> summary(Agg)
       X        dupData.InvoiceNo
 Min.   :  1.0   536527 :  1
 1st Qu.:151.5   536840 :  1
 Median :302.0   536861 :  1
 Mean   :302.0   536967 :  1
 3rd Qu.:452.5   536983 :  1
 Max.   :603.0   537197 :  1
                 (Other):597
                                                                                                  dupData.Description
 POSTAGE                                                                                                        : 14
 Manual                                                                                                         :  9
 CHILDS BREAKFAST SET CIRCUS PARADE, GLASS  BEURRE DISH                                                         :  3
 REGENCY CAKESTAND 3 TIER                                                                                       :  3
 CAKE STAND 3 TIER MAGIC GARDEN, SWEETHEART 3 TIER CAKE STAND , CAKE STAND LOVEBIRD 2 TIER WHITE:  2
 GUMBALL COAT RACK                                                                                              :  2
 (Other)                                                                                                        :570
```

Figure 14: Aggregation of the Invoice and product description

- Splitting the description of Invoice Number in seperate column Figure 12 shows the steps to perform the same



	X	dupData.InvoiceNo	dupData.Description_001	dupData.Description_002
1	1	536527	SET OF 6 T-LIGHTS SANTA	ROTATING SILVER ANGELS T-LIGHT HLDR
2	2	536840	JAM MAKING SET PRINTED	JAM JAR WITH PINK LID
3	3	536861	FELTCRAFT 6 FLOWER FRIENDS	6 RIBBONS RUSTIC CHARM
4	4	536967	POSTAGE	JUMBO BAG RED RETROSPOT
5	5	536983	WOODLAND PARTY BAG + STICKER SET	HAND WARMER OWL DESIGN
6	6	537197	BISCUIT TIN VINTAGE GREEN	BISCUIT TIN VINTAGE RED
7	7	537198	DOORMAT RED RETROSPOT	POSTAGE
8	8	537201	GUMBALL MAGAZINE RACK	HAND WARMER OWL DESIGN
9	9	537212	JUMBO BAG WOODLAND ANIMALS	WOODLAND CHARLOTTE BAG
10	10	537250	ENGLISH ROSE HOT WATER BOTTLE	SET 3 RETROSPOT TEA
11	11	537594	SET OF 16 VINTAGE PISTACHIO CUTLERY	SET OF 16 VINTAGE ROSE CUTLERY
12	12	537673	RED GINGHAM ROSE JEWELLERY BOX	GUMBALL COAT RACK
13	13	537892	POSTAGE	BABUSHKA LIGHTS STRING OF 10
14	14	537894	CHRISTMAS LIGHTS 10 VINTAGE BAUBLES	3 PIECE SPACEBOY COOKIE CUTTER SET
15	15	537995	METAL 4 HOOK HANGER FRENCH CHATEAU	3 HOOK PHOTO SHELF ANTIQUE WHITE
16	16	538174	SET OF 72 RETROSPOT PAPER DOILIES	PACK OF 72 RETROSPOT CAKE CASES
17	17	538175	POSTAGE	NA
18	18	538644	CHILDS BREAKFAST SET CIRCUS PARADE	GLASS BEURRE DISH
19	19	539327	POSTAGE	VINTAGE BEAD PINK SCARF
20	20	539395	RED COAT RACK PARIS FASHION	FRIDGE MAGNETS LES ENFANTS ASSORTED

Showing 1 to 20 of 603 entries

Console Terminal

~/Desktop/Sem2/BigDataCSCI620/ROnline_Retail/

```
> #Splitting the description for Invoice number in seperate column
> library(splitstackshape)
> write.csv(cSplit(Agg, "dupData.Description", ","), file = "Split.csv", row.names = FALSE)
> Auto1 <- read.csv("Split.csv")
> View(Auto1)
```

Figure 15: Splitting the description of Invoice Number in seperate column

- Performing the market basket analysis Figure 13 shows the steps to perform the same

```
> bs_rules <- apriori(Data1, parameter = list(supp = 0.1, conf = 0.8, minlen =2, maxlen = 10))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE       5     0.1      2     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 60

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[3074 item(s), 604 transaction(s)] done [0.01s].
sorting and recoding items ... [5 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [2 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> #Removing the duplicate rules
> redundant_rules<-is.redundant(bs_rules)
> bs_rules <- bs_rules[!redundant_rules]
> inspect(bs_rules)
    lhs                                rhs       support    confidence lift     count
[1] {ROUND SNACK BOXES SET OF 4 FRUITS}  => {POSTAGE} 0.1142384 0.8846154  1.395059  69
[2] {ROUND SNACK BOXES SET OF4 WOODLAND} => {POSTAGE} 0.1705298 0.8655462  1.364987 103
```

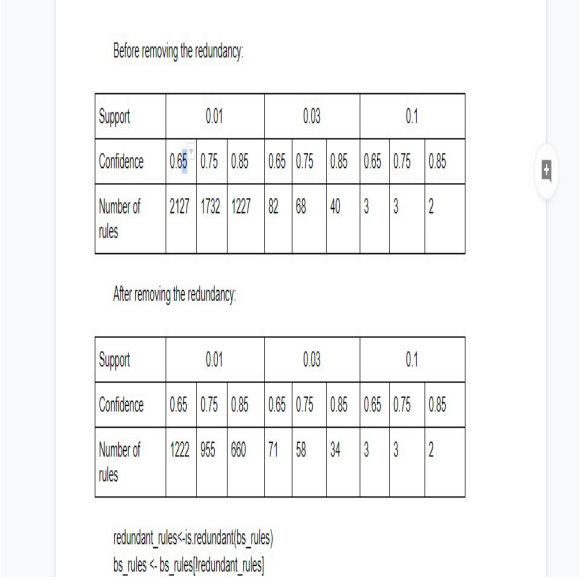
Figure 16: Market Basket Analysis Using Apriori Algorithm

4. CONCLUSION

4.1 Knowledge Gained

After performing Market Basket Analysis on the online retail data, we have gained knowledge about items that were frequently bought together. After careful observation of the data, we have come up with the following threshold values: Support = 0.1 Confidence = 0.85

The reason for choosing the above mentioned support and confidence is depicted in the figure shown below which explains the rules obtained for varing support and confidence :



Before removing the redundancy:

Support	0.01			0.03			0.1		
Confidence	0.65	0.75	0.85	0.65	0.75	0.85	0.65	0.75	0.85
Number of rules	2127	1732	1227	82	68	40	3	3	2

After removing the redundancy:

Support	0.01			0.03			0.1		
Confidence	0.65	0.75	0.85	0.65	0.75	0.85	0.65	0.75	0.85
Number of rules	1222	955	660	71	58	34	3	3	2

```
redundant_rules<-is.redundant(bs_rules)
bs_rules <- bs_rules[!redundant_rules]
```

Figure 17: Market Basket Analysis Using Apriori Algorithm

We obtained 2 rules that fit the threshold criteria. The observed lift values for these 2 rules is > 1 and hence they are considered as good rules. Association rules are as follows:

- $\{\text{ROUND SNACK BOXES SET OF 4 FRUITS}\} \rightarrow \{POSTAGE\}$
  $\{ROUNDSNACKBOXESSETOF4WOODLAND\} \rightarrow \{POSTAGE\}$
  Our association rules tell us that customers usually bought these items . After performing Market Basket Analysis on the online retail data, we have gained knowledge about items that were frequently bought together. After careful observation of the data, we have come up with the following threshold values: Support = 0.1 Confidence = 0.85 We obtained 2 rules that fit the threshold criteria. The observed lift values for these 2 rules is > 1 and hence they are considered as good rules. Association rules are as follows:
- $\{\text{ROUND SNACK BOXES SET OF 4 FRUITS}\} \rightarrow \{POSTAGE\}$
  $\{ROUNDSNACKBOXESSETOF4WOODLAND\} \rightarrow \{POSTAGE\}$
  Our association rules tell us that customers usually bought these items

5. WORKLOAD DISTRIBUTION

Cleaning the data - Himanshi, Ishika, Asmita, Poornima and Anuradha

Aggregation of the product description against the invoice numbers - Ishika, Asmita and Poornima

Converting the data to transactional data - Ishika, Asmita and Poornima

Performing Market Basket Analysis using Apriori Algorithm - Himanshi, Ishika, Asmita, Poornima and Anuradha

Removing the duplicate rules - Himanshi, Ishika, Asmita, Poornima and Anuradha

Visualization of the data - Himanshi, Ishika, Asmita, Poornima and Anuradha

Project Report - Himanshi and Anuradha

6. CHALLENGES

Selection of Data Mining Task : Selection of association over classification and clustering, and determining why it would provide us the most relevant information.
Data exploration and visualization of data: Visualization and modeling of textual attributes
Market Basket Analysis: Conversion of input atomic data to transactional data in order to run the Apriori algorithm. Determination of support and confidence to generate significant rules