1. Problem Statement:

The retailer MyShop needs to send physical / e-catalog to its customers for boosting their sales. But the catalogs have to be sent in such a way that they wouldn't invoke any negative sentiments. In addition, the cost of production and distribution of catalogs is high, hence they need to be targeted to customers so that the probability of inducing a purchase is high.

2. Aim of the Solution:

- MyShop employs Cross-Selling strategy where the retailer would try to sell a second product once an existing customer has purchased their first product. MyShop intends to achieve this by sending the required catalog.
- The algorithm that can be used here to solve the problem is Association Rule Mining.
    - Here various market baskets are analyzed to identify the antecedents and consequents pairs. This would tell which catalog to be sent to customers

3. Importing, Understanding and Processing the Data:

- The data is imported from the given file with read.table function with separator as ',' (csv file) and header TRUE. The tail of the data is viewed to understand the structure.

```
> data <- read.table('MyShopData.csv',sep=',',header=TRUE) # Importing data: .csv format
> tail(data) # Extra row at the end (contains summary data)
     Customer.Number Clothing.Division Housewares.Division Health.Products.Division Automotive.Division
4994       337441034                 0                    0                       1                    0
4995       337498968                 0                    1                       1                    0
4996       337516678                 0                    0                       1                    0
4997       337528921                 0                    0                       1                    0
4998       337534044                 0                    0                       1                    0
4999              NA               165                 1967                    4998                  674
     Personal.Electronics.Division Computers.Division Garden.Division Novelty.Gift.Division
4994                             0                   0                0                     0
4995                             0                   0                1                     0
4996                             1                   0                0                     0
4997                             0                   0                0                     0
4998                             0                   0                0                     0
4999                          2336                 234             1360                  1137
     Jewelry.Division     X
4994                0    NA
4995                1    NA
4996                1    NA
4997                0    NA
4998                0    NA
4999             1784 14655
```

- The given dataset has 4999 rows and 11 columns
    - The first column has customer number which is unique to each of the 4998 customers
    - The next 9 columns signify 9 categories in which a customer could have made a purchase. If the customer has made a purchase in a specific category, they are marked as 1 and if not then they are marked as 0
    - The last row and last column contains just the summary details (sum of the 9 rows). Hence, this row and column needs to be removed.

```
> data <- data[-4999,-11] # Removing the last row & column from the data
> tail(data)
     Customer.Number Clothing.Division Housewares.Division Health.Products.Division Automotive.Division
4993       337438295                0                    0                        1                   0
4994       337441034                0                    0                        1                   0
4995       337498968                0                    1                        1                   0
4996       337516678                0                    0                        1                   0
4997       337528921                0                    0                        1                   0
4998       337534044                0                    0                        1                   0
     Personal.Electronics.Division Computers.Division Garden.Division Novelty.Gift.Division
4993                             0                  0               0                     0
4994                             0                  0               0                     0
4995                             0                  0               1                     0
4996                             1                  0               0                     0
4997                             0                  0               0                     0
4998                             0                  0               0                     0
     Jewelry.Division
4993                0
4994                0
4995                1
4996                1
4997                0
4998                0
```

- For Association Rule Mining, the data needs to be in the Single (Transaction) format. But the data is currently in a matrix format. Hence this needs to be transformed into the Single format.
    - Library tidyr is used for this transformation process
    - The function used here is **pivot_longer** which is used to convert the data from matrix to transaction format. Further, rows with purchase value of 1 are alone selected

```
> data_v1 <- as.data.frame(pivot_longer(data,-Customer.Number,names_to = "category",values_to = "purchase"))
 # Pivoting the data
> head(data_v1) # inspecting the data frame
  Customer.Number                  category purchase
1           11569          Clothing.Division        0
2           11569       Housewares.Division        1
3           11569   Health.Products.Division        1
4           11569      Automotive.Division        1
5           11569 Personal.Electronics.Division        1
6           11569       Computers.Division        0
> # data_v1 (transformed data) will have data for each customer for all categories: 0 and 1
> # Removing the rows with purchase value 0(purchase not made)
> data_v2 <- subset(data_v1,purchase==1) # subsetting data with purchase value 1
> head(data_v2) # inspecting the data frame
   Customer.Number                  category purchase
2           11569       Housewares.Division        1
3           11569   Health.Products.Division        1
4           11569      Automotive.Division        1
5           11569 Personal.Electronics.Division        1
8           11569     Novelty.Gift.Division        1
11          13714       Housewares.Division        1
> # Removing the purchase column; converting to single format
> data_v3 <- data_v2[,1:2]
> head(data_v3)
   Customer.Number                  category
2           11569       Housewares.Division
3           11569   Health.Products.Division
4           11569      Automotive.Division
5           11569 Personal.Electronics.Division
8           11569     Novelty.Gift.Division
11          13714       Housewares.Division
```

- The data is written into a new csv file - MyShopData_single.csv and imported again in the Single format with the function read.transactions and format as single

4. Association Rule Mining:

- Library arules and the function **apriori** inside it are used for association rule mining. The single format data that was imported earlier is passed on to the function

```
> # apriori for generating the association rules
> ass_rules<-apriori(data)
> print(inspect(sort(ass_rules,by="support")))
     lhs                                rhs                                support confidence  coverage   lift count
[1]  {}                              => {Health.Products.Division} 0.9998000     0.9998 1.0000000 1.0000  4998
[2]  {Personal.Electronics.Division} => {Health.Products.Division} 0.4672935     1.0000 0.4672935 1.0002  2336
[3]  {Housewares.Division}           => {Health.Products.Division} 0.3934787     1.0000 0.3934787 1.0002  1967
[4]  {Jewelry.Division}              => {Health.Products.Division} 0.3568714     1.0000 0.3568714 1.0002  1784
[5]  {Garden.Division}               => {Health.Products.Division} 0.2720544     1.0000 0.2720544 1.0002  1360
[6]  {Housewares.Division,
      Personal.Electronics.Division} => {Health.Products.Division} 0.2354471     1.0000 0.2354471 1.0002  1177
[7]  {Novelty.Gift.Division}         => {Health.Products.Division} 0.2274455     1.0000 0.2274455 1.0002  1137
[8]  {Jewelry.Division,
      Personal.Electronics.Division} => {Health.Products.Division} 0.1974395     1.0000 0.1974395 1.0002   987
[9]  {Housewares.Division,
      Jewelry.Division}              => {Health.Products.Division} 0.1948390     1.0000 0.1948390 1.0002   974
[10] {Novelty.Gift.Division,
```

- The generated rules are inspected using the **inspect** function by passing the ass_rules object and sorting the data by value of Support
    - Support denotes the percentage of transactions that contain both the antecedent and the consequent items in the overall transaction list
    - Confidence is the support of the selected basket divided by the support of antecedent item in it. This signifies how often the selected consequent appears in transaction with the antecedent.
- As a rule of thumb, minimum Support value is taken 0.4 and minimum Confidence value is taken as 0.6. This results in only one association rule,

```
> # Cutoffs for support and confidence given (0.4 and 0.6 respectively)
> ass_rules_v2 <- apriori(data,parameter=list(sup=0.4,conf=0.6,minlen=2))
> print(inspect(ass_rules_v2))
    lhs                                rhs                                support   confidence coverage  lift
[1] {Personal.Electronics.Division} => {Health.Products.Division} 0.4672935 1          0.4672935 1.0002
    count
[1] 2336
```

    - Association rule with a basket with two items: Personal Electronics Division and Health Products Division. The former is the antecedent and latter is consequent. The Support for this rule is 0.4673 which means, 46.73% of the all the transactions contains items from these two categories.
      The confidence for this rule is 1.00 which means in all the transaction where a customer has purchased from Personal Electronics Division, they have also purchased from Heath Products Division.
    - In fact, this is the case with all the transactions in the data set. All the 4998 customers have purchased products from the Health Products Division category, hence for all the association rules with consequent as Health Products Division, the confidence would be equal to 1.00
    - This can be taken in two ways,
        - It is evident that all the existing customers purchase from Health Products Division. Based on this, the retailer MyShop can include Health

Products Division catalog for all its **future new customers** who will be shopping from the store.
    - But the case specifically mentions to decide on the catalogs to be sent to its **existing customers only**
  - Hence, the Health Products Division category makes no significance to the existing customers as they already purchase from the category.

- To deal with this, the columns Health Products Division is removed from the data frame and once again the association rules are generated.
  - Library dplyr and function **filter** are used for filtering out the rows from the previous data frame which had Health Products Division category.

```
> library(dplyr)
> data_v4 <- filter(data_v3,category!='Health.Products.Division')
> head(data_v4)
  Customer.Number                      category
1           11569          Housewares.Division
2           11569          Automotive.Division
3           11569 Personal.Electronics.Division
4           11569         Novelty.Gift.Division
5           13714          Housewares.Division
6           13714          Automotive.Division
```

- The resulting data frame is once again written into a new csv file and imported in the single format
- Association Rule Mining function apriori is applied to this new dataset. The minimum Support constraint is given as 0.2 and minimum Confidence is given as 0.5 for mining the rules
  - The rationale behind reducing the support value from 0.4 to 0.2 is that, the dataset is not a small one. With 4998 customers, 20% comes around 1000 customers which is a considerable size. Hence, the minimum support value is given as 0.2
- The apriori function with the new data set generates 6 association rules

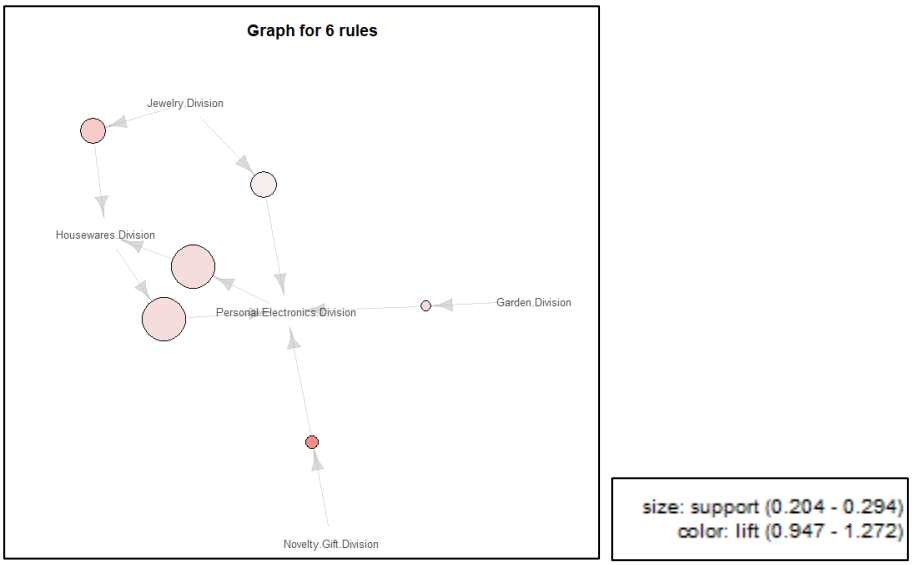
```
> # Cutoffs for support and confidence given (0.4 and 0.6 respectively)
> ass_rules_new_v2 <- apriori(data_new,parameter=list(sup=0.2,conf=0.5,minlen=2))
```

```
> print(inspect(ass_rules_new_v2))
    lhs                                rhs                                support   confidence coverage
[1] {Novelty.Gift.Division}         => {Personal.Electronics.Division} 0.2114086 0.7431838  0.2844633
[2] {Garden.Division}               => {Personal.Electronics.Division} 0.2041531 0.6000000  0.3402552
[3] {Jewelry.Division}              => {Housewares.Division}           0.2436828 0.5459641  0.4463348
[4] {Jewelry.Division}              => {Personal.Electronics.Division} 0.2469352 0.5532511  0.4463348
[5] {Housewares.Division}           => {Personal.Electronics.Division} 0.2944709 0.5983732  0.4921191
[6] {Personal.Electronics.Division} => {Housewares.Division}           0.2944709 0.5038527  0.5844383
    lift      count
[1] 1.2716206  845
[2] 1.0266267  816
[3] 1.1094146  974
[4] 0.9466373  987
[5] 1.0238431 1177
[6] 1.0238431 1177
```

- The Lift ratio is also considered here and the minimum value is taken to be 1.00. Lift ratio compares the confidence of the rule against a benchmark value which is nothing but the fraction of the consequent item in the dataset

- Hence from the result, rule numbers 1, 2, 3, 5 and 6 are considered as meaningful association rules

S No	Antecedent	Consequent	Support	Confidence	Lift	Count
1	Novelty Gift Division	Personal Electronics Division	0.211	0.743	1.271	845
2	Garden Division	Personal Electronics Division	0.204	0.6	1.027	816
3	Jewelry Division	Housewares Division	0.244	0.546	1.109	974
4	Housewares Division	Personal Electronics Division	0.294	0.598	1.024	1177
5	Personal Electronics Division	Housewares Division	0.294	0.504	1.024	1177



- The associations can also be represented graphically using the library arulesViz and **plot** function.
    - Here the direction of arrows determines if an item is antecedent or consequent. Outward arrows start from the antecedent element whereas, inward arrows end with the consequent element.
    - The size and color of the association circle denote the support and lift respectively

5. Interpreting the Business Value of the Result:

- From the result we can see that there are 2 consequents: Personal Electronics Division and Housewares Division
    - Novelty Gift, Garden and Housewares Division are the antecedent for the consequent Personal Electronics Division
    - Jewelry and Personal Electronics Division are the antecedent for the consequent Housewares Division

- Hence for the Cross Selling Strategy employed by MyShop (selling the second product once the customer has purchased the first product), we can determine the following based on the Association Rules mined,
    - Number of existing customers who currently buy the antecedent but do not buy the consequent.
    - The catalog for the consequent products can be sent to these customers
- The list of customers is extracted and written into a new data frame. This can be used by the MyShop retailer for sending the required catalog to the customers. Library dplyr and **filter** function are used for filtering the data

```
> raw_data <- read.table('MyShopData.csv',sep=',',header=TRUE) # Importing data: .csv format
> raw_data <- raw_data[-4999,-11]
> ass_rule_1 <- filter(raw_data,(Novelty.Gift.Division==1 |Housewares.Division==1 | Garden.Division==1),
+                  Personal.Electronics.Division!=1)
> ass_rule_2 <- filter(raw_data,(Personal.Electronics.Division == 1 | Jewelry.Division == 1),
+                      Housewares.Division!=1)
> ass_rule_1$Catalog.To.Send <- 'Personal.Electronics.Division'
> ass_rule_2$Catalog.To.Send <- 'Housewares.Division.Division'
> Final_List <- rbind(ass_rule_1,ass_rule_2)
> Final_List <- Final_List[,c(1,11)]
> head(Final_List)
  Customer.Number              Catalog.To.Send
1          422147 Personal.Electronics.Division
2          435925 Personal.Electronics.Division
3          469171 Personal.Electronics.Division
4          521682 Personal.Electronics.Division
5          681833 Personal.Electronics.Division
6          690370 Personal.Electronics.Division
> tail(Final_List)
     Customer.Number                Catalog.To.Send
2887       337292882 Housewares.Division.Division
2888       337310965 Housewares.Division.Division
2889       337377519 Housewares.Division.Division
2890       337401095 Housewares.Division.Division
2891       337407837 Housewares.Division.Division
2892       337516678 Housewares.Division.Division
> write.csv(Final_List,"CustomerList.csv",row.names=F)
```

- The Final_List file would contain the customer number and the name of the catalog to be sent to that particular customer.

6. Conclusion:

Thus by using the Association Rule Mining algorithm, the retailer is able to extract the associations between different categories. Subsequently, the required catalog for Cross Selling purpose can be sent to them based on the Customers identified.