Café Store Promoting Strategy Analytics

Mingchi Zhang

Content

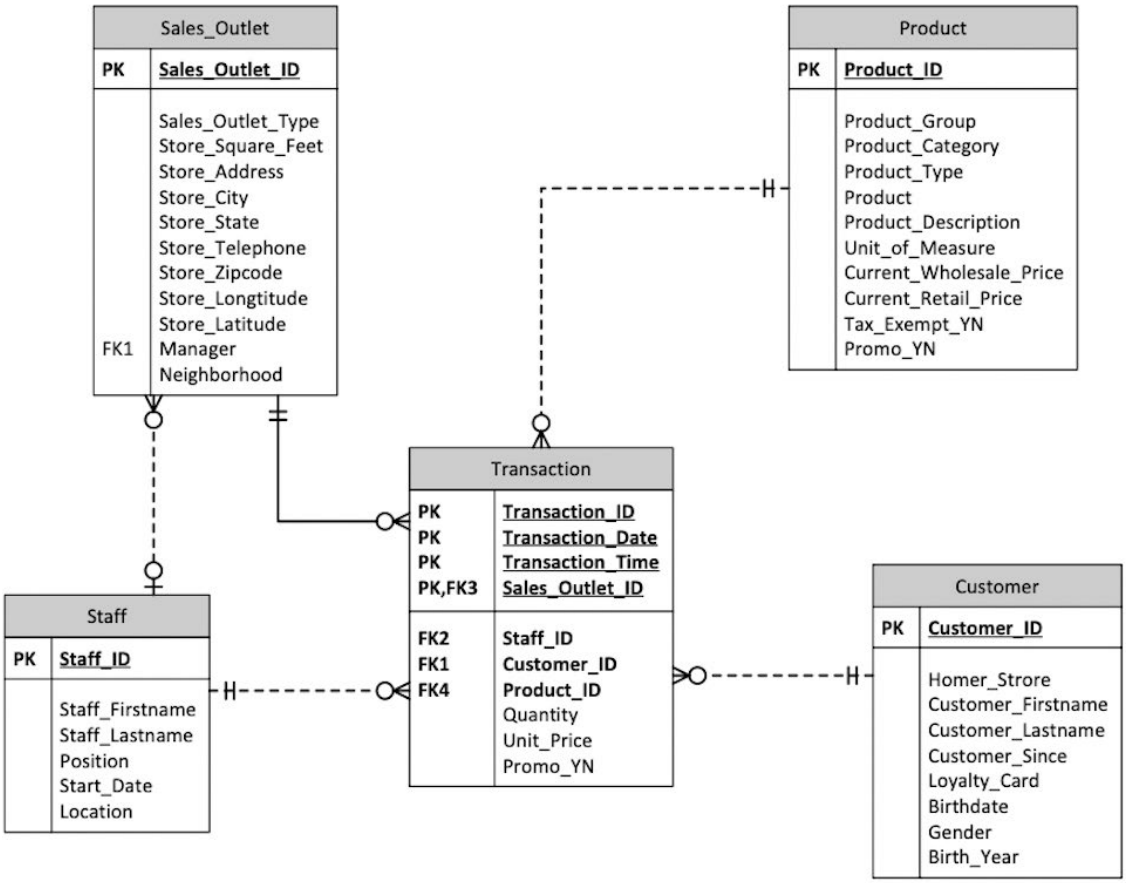
1. Executive Summary, Project Motivation/Background

Our client is a coffee chain with three locations in New York City. Their main business is to sell coffee beans, tea, chocolate, and some bakery, etc. The ultimate purpose of this project is to use the R model we learned to analyze the historical sales data and customer data of the café store. We hope to help the owner to：

a. Improve Café Store Promoting Strategy based on purchase frequency of different products to improve sales performance and achieve sales targets
b. Analysis of the market segment to promote to customers more strategically and deliver the targeted marketing efficiently
c. Drive other relevant business decision making via R outputs

2. Data Description

This dataset contains representative retail data from a coffee chain in three locations in New York City in April 2019. It contains a dataset of over 50000 transactions, over 2246 customer records, 89 products, 3 stores, and 55 staff.



3. Problem Statement

3.1 Product promotion

Based on the current sales data, not all stores met the sales target yet. Applying the **Association Rules**, we try to figure out whether the current promotion product is appropriate and would like to find better product sale bundles or promotion policy to reduce the percentage of waste and improve sales per order.

3.2 Marketing segment, target promotion, and membership policy

Analyze how much sales amount was made by members and are it profitable to develop more members. Using **Clustering** to analyze current members based on their buying recency, frequency, and monetary. We also plan to explore our customer profile (age and membership) to deliver targeted marketing or advertising through local media and mail ads.

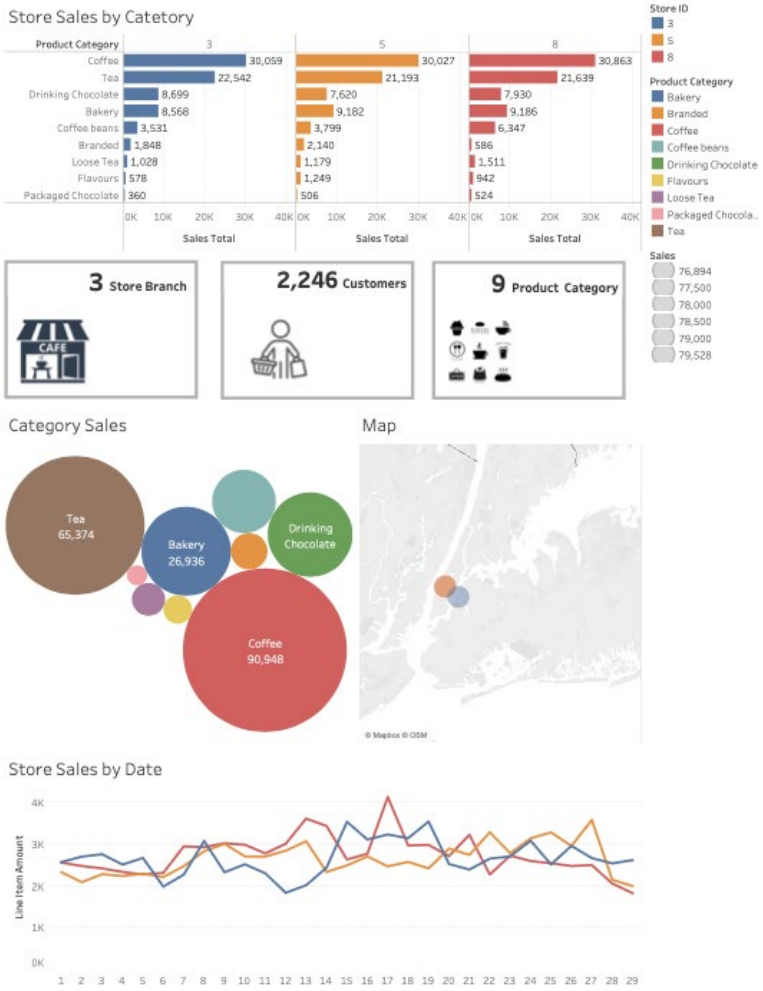
3.4 predict the product demand and staff arrangement

Regression Classification: Regression analysis can be performed during the exploratory relation of date(weekday), sales, category, store num, staff id, the promotion which can help predict the product demand and reduce the average waste for better inventory management.

4. Data Exploration

4.1 Variable Exploration

- It contains 50000 transaction data, over 8000 customer records, 89 products, 3 stores data, and 55 staff's data.
- All the stores are open from 1 am to 8 pm.
- There are 4 promotion products and 2 new products.



4.2 Dataset Cleaning

Dealing with missing value: We use R searching for null/0 value.

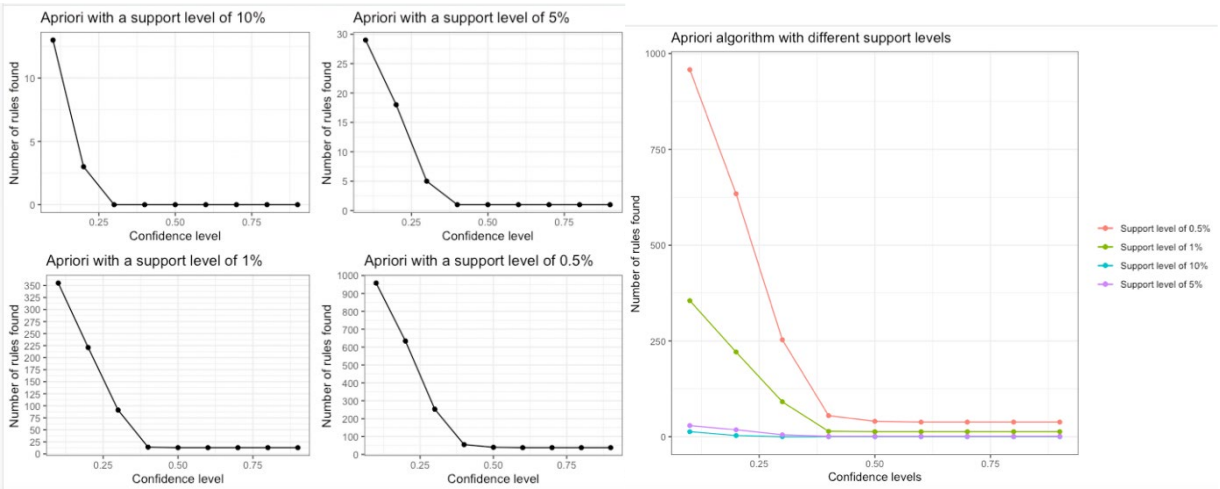
- We regard customer records with id 0 as customers with no membership.

- Remove customer(membership) record without age and membership start date (There is only one record of missing value.)
- Removing records with no instore/online information when doing instore or online analytics.
- Regarding the transaction with sales amount 0 as giveaways promotion.
- Dealing with outliers: Apply boxplots to find the outliers.
- There seems to be no extreme outlier with no meaning in our dataset

5. Business Intelligence Models

5.1 Market Basket Analysis (Association rules)

5.1.1 Model Results

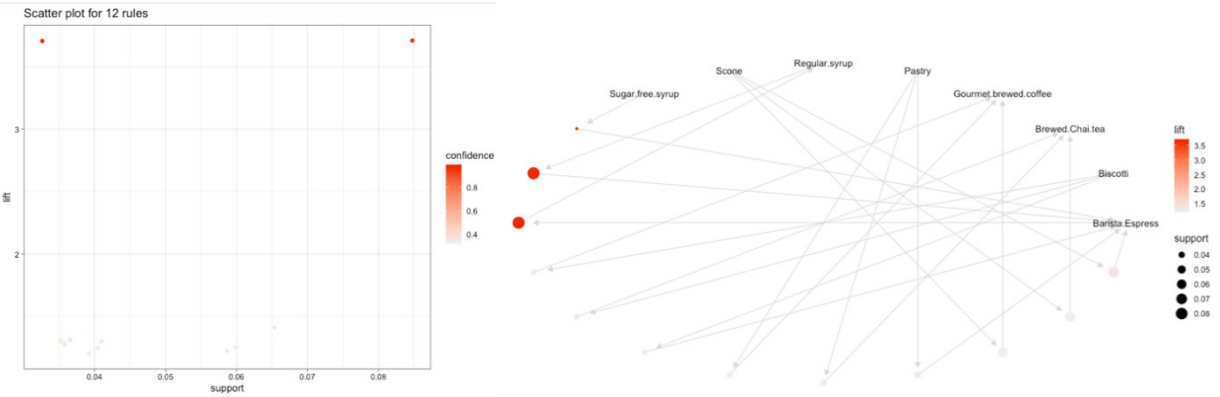


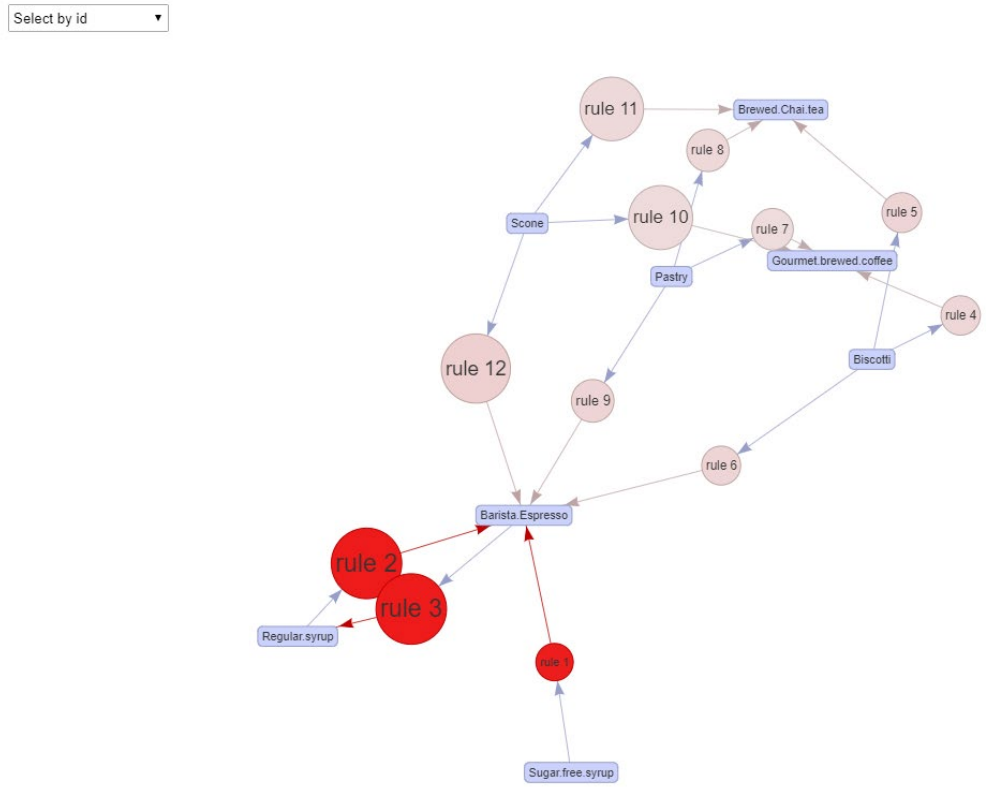
After visualizing the number of rules found with different settings of support and confidence levels, we chose the thresholds for support to be 3% and confidence levels to be 30%. The 12 valid association rules are exhibited below.

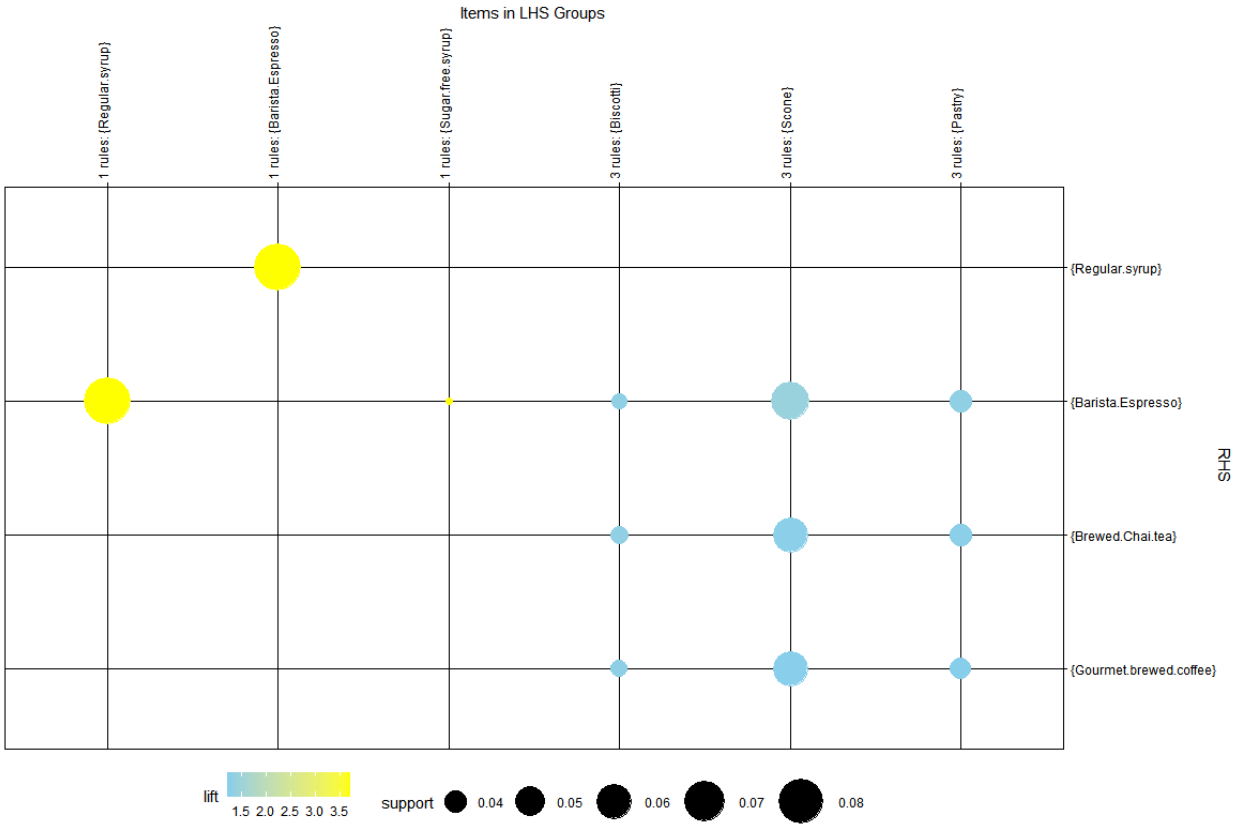
```
     lhs                    rhs                    support    confidence coverage   lift     count
[1]  {Regular.syrup}     => {Barista.Espresso}     0.08476569 0.9960526  0.08510162 3.709445 1514
[2]  {Barista.Espresso}  => {Regular.syrup}        0.08476569 0.3156797  0.26851800 3.709445 1514
[3]  {Sugar.free.syrup}  => {Barista.Espresso}     0.03264095 0.9948805  0.03280891 3.705080  583
[4]  {Scone}             => {Barista.Espresso}     0.06533789 0.3785274  0.17261072 1.409691 1167
[5]  {Biscotti}          => {Brewed.Chai.tea}      0.03656010 0.3637883  0.10049829 1.313182  653
[6]  {Biscotti}          => {Barista.Espresso}     0.03521639 0.3504178  0.10049829 1.305007  629
[7]  {Pastry}            => {Barista.Espresso}     0.04098315 0.3492366  0.11735065 1.300608  732
[8]  {Biscotti}          => {Gourmet.brewed.coffee} 0.03577627 0.3559889 0.10049829 1.279854  639
[9]  {Scone}             => {Brewed.Chai.tea}      0.05990706 0.3470645  0.17261072 1.252813 1070
[10] {Pastry}            => {Brewed.Chai.tea}      0.04047926 0.3449427  0.11735065 1.245154  723
[11] {Scone}             => {Gourmet.brewed.coffee} 0.05867533 0.3399286 0.17261072 1.222115 1048
[12] {Pastry}            => {Gourmet.brewed.coffee} 0.03924752 0.3344466 0.11735065 1.202405  701
```

We also generated several graphs to visualize these 12 top association rules.

5.1.2 Rules Visualization




Select by id
rule 11
Brewed.Chai.tea
rule 8
rule 5
Scone
rule 10
rule 7
Gourmet.brewed.coffee
Pastry
rule 4
rule 12
Biscotti
rule 9
rule 6
Barista.Espresso
rule 2
rule 3
Regular.syrup
rule 1
Sugar.free.syrup


Items in LHS Groups
1 rules: {Regular.syrup}
1 rules: {Barista.Espresso}
1 rules: {Sugar.free.syrup}
3 rules: {Biscotti}
3 rules: {Scone}
3 rules: {Pastry}
{Regular.syrup}
{Barista.Espresso}
{Brewed.Chai.tea}
{Gourmet.brewed.coffee}
RHS
lift
1.5 2.0 2.5 3.0 3.5
support
0.04
0.05
0.06
0.07
0.08

5.1.3 Relevant Findings

Percentage of Transaction:

99.60% of the customers who add Regular Syrup into Barista Espresso,

99.48% of the customers who add Sugar-Free Syrup into Barista Espresso,

37.85% of the customers who bought a Scone also bought Barista Espresso,

35.04% of the customers who bought a Biscotti also bought Barista Espresso,

34.92% of the customers who bought a Pastry also bought Barista Espresso.

According to the Top 3 association rules:

If a customer buys Regular Syrup, he or she is 3.71 times more likely to also buy Barista Espresso if we know nothing about him or her.

If a customer buys Sugar-Free Syrup, he or she is 3.70 times more likely to also buy Barista Espresso if we know nothing about him or her.

5.1.4 Identifying the Business Opportunities

The top 3 association rules are related to Barista Espresso and Regular syrup or Sugar-free syrup, as we can see, this is the most common combination mode customers ordered. So "Barista Espresso" can always be the **signature product** (coffee) in our café stores, and we believe it is the best way to introduce our products to any new customers.

Then when customers ordered dessert (like scone, biscotti, and pastry), there was more likely (from 1.2 times to 1.4 times) they would also order Barista Espresso or Gourmet brewed coffee or Brewed Chai tea. So, **bundling dessert with drinks proved** to be an effective way to increase sales. Moreover, we can give coupons when customers order one of them; for example, when customers purchase one of the desserts, they can get 10% off for an extra cup of coffee/tea.

However, with the consideration of the high waste rate of all the three kinds of desserts (about 50% to 65%), we may give a **higher discount** (like 20% off) to any customers who purchased Barista Espresso or Gourmet brewed coffee or Brewed Chai tea.

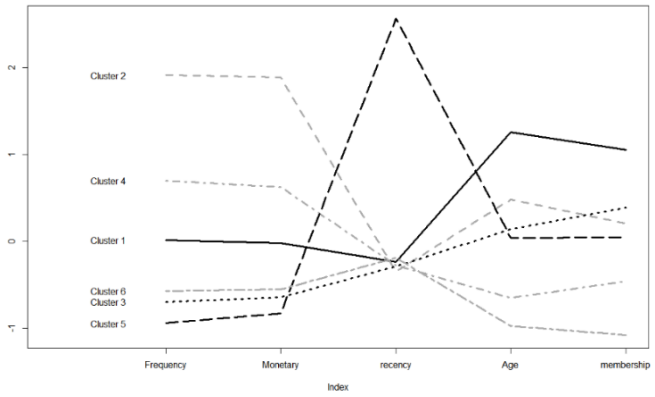
We also performed the "apriori" algorithm based on three different stores separately. Since the results (not shown here) are similar, these business strategies applied to any of the three stores.

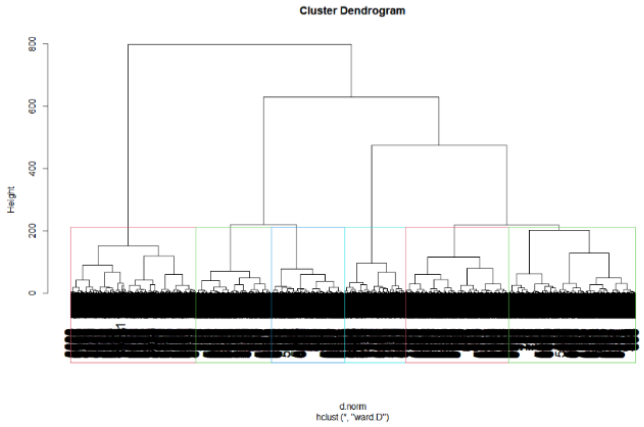
5.2 Clustering Analysis

5.2.1 Model Results

We obtained age and membership information from the customer table and calculated the values of recency, frequency, and monetary from the transaction table. Both Partitioning and Hierarchical clustering analyses were performed.

```
> km$centers
   Frequency    Monetary    recency         Age  membership
1  0.0151632 -0.01971847 -0.2384820  1.25414089  1.04815865
2  1.9108181  1.88764344 -0.3514166  0.48136577  0.20488678
3 -0.6966653 -0.64257848 -0.2880859  0.14028057  0.38810520
4  0.6973649  0.62521161 -0.2763767 -0.65421398 -0.45835611
5 -0.9441288 -0.82879202  2.5606578  0.03583481  0.04305959
6 -0.5708173 -0.55120128 -0.1936640 -0.97220936 -1.08203303
```





5.2.2 Relevant Findings

Line features:	
Cluster1	Higher age and Membership length
Cluster2	High Frequency and Monetary
Cluster3	Lower frequency, monetary and middle age, membership length
Cluster4	Middle frequency, monetary and lower age, and membership length
Cluster5	Higher recency
Cluster6	All low

5.2.3 Identifying the Business Opportunities

The clustering analysis results figured out a group of members with high frequency and monetary, middle age, and membership length (Cluster2), who should be our **target customers**

to keep. We can give them the golden membership card and 10% off for each order to keep their consuming capabilities. In the meantime, we can also try some packages based on their age features.

Besides, cluster 4 grouped **new customers** of lower age and membership length, but they have middle levels of purchasing frequency and monetary. We may do some surveys or questionnaires to acquire their interests and favorite flavor, and also launch new products to this group of customers.

New Customer Development: Target at Young generation which our cafe store lacked: 1) Leverage our social media and consider offering first-time customers a discount on their next drink if they leave a review online. 2) Providing App with the ability to order online.3) Our furniture arrangement will clearly define our relaxed setting, encouraging customers to sit down and take pictures.

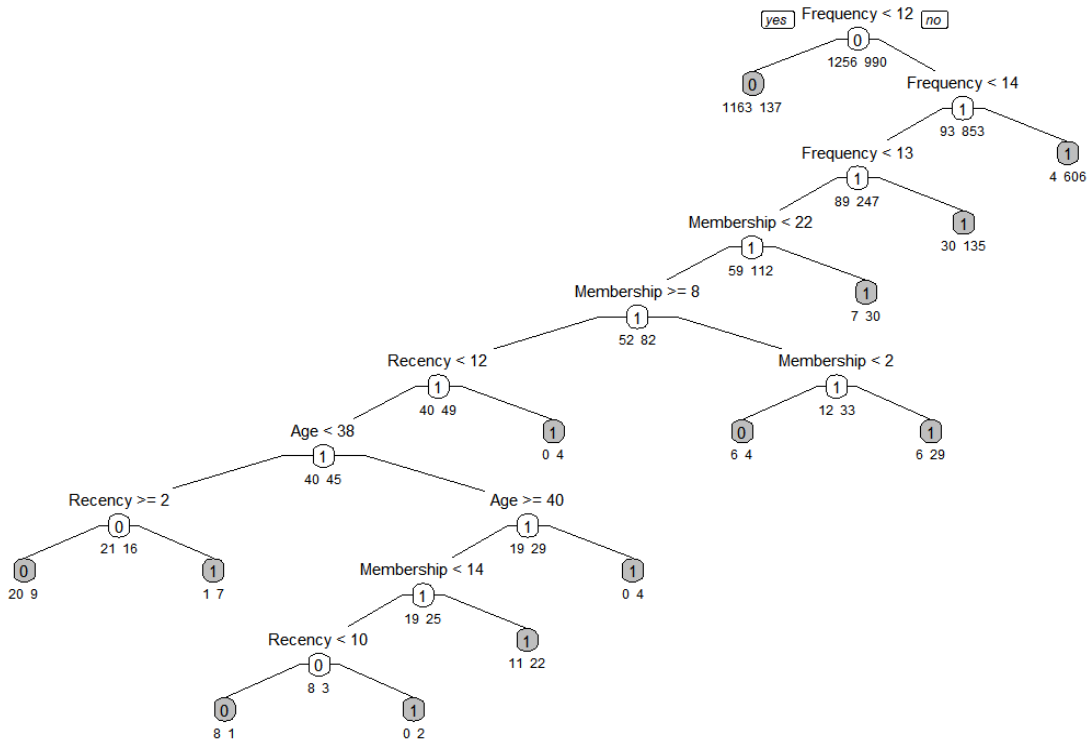
5.3 Decision Tree

5.3.1 Model Results

We assume the customers with consuming amount (Monetary) more than the average value to be “Heavy Buyers”, and we aim to predict if a customer would be a “heavy buyer” based on his/her age, membership length, recency and frequency.

With 60% of the data to be selected randomly as the training dataset, we built the model and validated it with the rest of 40% dataset.

The accuracy rate for both the training and validation dataset were as high as 90%.



confusion matrix for training data

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 717  92
         1  35 503

               Accuracy : 0.9057
                 95% CI : (0.8888, 0.9208)
    No Information Rate : 0.5583
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8069

 Mcnemar's Test P-Value : 6.723e-07

            Sensitivity : 0.9535
            Specificity : 0.8454
         Pos Pred Value : 0.8863
         Neg Pred Value : 0.9349
             Prevalence : 0.5583
         Detection Rate : 0.5323
   Detection Prevalence : 0.6006
      Balanced Accuracy : 0.8994

       'Positive' Class : 0
```

confusion matrix for validation data

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 480  59
         1  24 336

               Accuracy : 0.9077
                 95% CI : (0.8868, 0.9258)
    No Information Rate : 0.5606
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8108

 Mcnemar's Test P-Value : 0.00019

            Sensitivity : 0.9524
            Specificity : 0.8506
         Pos Pred Value : 0.8905
         Neg Pred Value : 0.9333
             Prevalence : 0.5606
         Detection Rate : 0.5339
   Detection Prevalence : 0.5996
      Balanced Accuracy : 0.9015

       'Positive' Class : 0
```

5.3.2 Relevant Findings

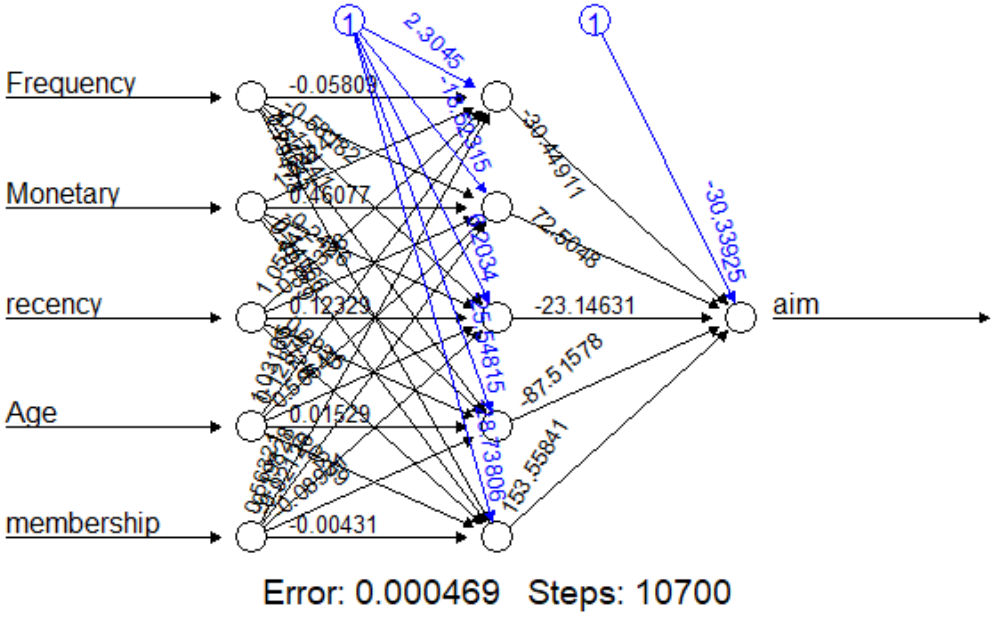
With this model, we can predict who would be the next heavy buyers (target customers) after collecting the information on his/her age, membership length, recency and frequency, then we may provide them customized services to develop our business.

5.4 Neural Network

5.4.1 Model Results

The 5.3 Decision Tree brings the characteristic insight of the “Heavy Buyers”. In order to get a better result, we also applied Neural Network to the same problem.

Segmented the customer into 2 groups “target (above the average)” and “non-target (below the average)” we use the mean of all customer's historical purchasing amounts. We structured training data and validation data by 60% and 40% for the customer- transactions data. Setting the number of hidden neurons to 5, the stepmax to 200000, and the learning rate to 0.5 we got the model as followed.



Then a confusion matrix is applied to the validation dataset to validate the performance of the model. The accuracy reaches 99.67%.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 502   1
         1   2 394

               Accuracy : 0.9967
                 95% CI : (0.9903, 0.9993)
    No Information Rate : 0.5606
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.9932

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9960
            Specificity : 0.9975
         Pos Pred Value : 0.9980
         Neg Pred Value : 0.9949
             Prevalence : 0.5606
         Detection Rate : 0.5584
   Detection Prevalence : 0.5595
      Balanced Accuracy : 0.9968

       'Positive' Class : 0
```

5.4.2 Relevant Findings

From the previous model we got the hint of the characteristic of the ‘Heavy buyer’. As a corroboration of the previous model, although our result cannot be used for different levels of the membership policy setting, we can predict who would be our target customers with a higher accuracy rate.

5.5 Logistic Regression

5.5.1 Model Results

Logistic regressions were applied to the data. We structured training data and validation data by 60% and 40% for transactions. Barista. Espresso was set up as the response variable, with all the other items as the explanatory variables.

```
> summary(logit.reg.vt)

Call:
glm(formula = Barista.Espresso ~ ., family = "binomial", data = valid.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5424  -0.6117  -0.5026   0.0274   2.7738

Coefficients:
                     Estimate Std. Error z value             Pr(>|z|)
(Intercept)          -1.08533    0.06305 -17.213 < 0.0000000000000002 ***
Biscotti              0.54319    0.10958   4.957 0.000000715859522731 ***
Black.tea             0.15056    0.49715   0.303              0.76201
Brewed.Black.tea     -0.33911    0.08259  -4.106 0.000040227810355800 ***
Brewed.Chai.tea      -0.42407    0.06899  -6.146 0.000000000792323085 ***
Brewed.Green.tea     -0.37188    0.11659  -3.190              0.00142 **
Brewed.herbal.tea    -0.52961    0.08805  -6.015 0.000000001801064208 ***
Chai.tea              0.24950    0.37493   0.665              0.50575
Clothing             -0.01993    0.64040  -0.031              0.97517
Drinking.Chocolate    0.58829    0.52319   1.124              0.26082
Drip.coffee          -0.49582    0.09764  -5.078 0.000000381861190064 ***
Espresso.Beans        0.62255    0.45969   1.354              0.17565
Gourmet.Beans        -0.03826    0.50105  -0.076              0.93914
Gourmet.brewed.coffee -0.52523   0.07350  -7.146 0.000000000000891435 ***
Green.beans          -0.67247    1.02892  -0.654              0.51339
Green.tea            -0.52060    0.76597  -0.680              0.49671
Herbal.tea            0.12327    0.46697   0.264              0.79181
Hot.chocolate        -0.56588    0.08723  -6.487 0.000000000087420783 ***
House.blend.Beans     1.37769    0.51972   2.651              0.00803 **
Housewares           -0.56865    0.52874  -1.075              0.28215
Organic.Beans         0.12145    0.39301   0.309              0.75730
Organic.brewed.coffee -0.64298   0.10144  -6.339 0.000000000231908919 ***
Organic.Chocolate     0.91091    0.60438   1.507              0.13177
Pastry                0.25598    0.09901   2.585              0.00973 **
Premium.Beans         0.77837    0.41412   1.880              0.06016 .
Premium.brewed.coffee -0.59207   0.10038  -5.898 0.000000003674617831 ***
Regular.syrup         8.20580    1.00487   8.166 0.000000000000000319 ***
Scone                 1.00560    0.07546  13.326 < 0.0000000000000002 ***
Sugar.free.syrup      6.74529    0.72742   9.273 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5.5.2 Relevant Findings and interpretation

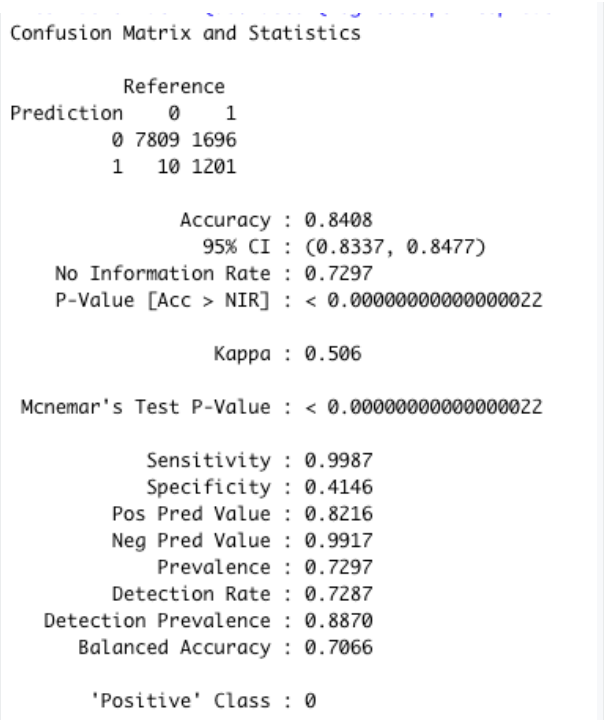
From the output, we found that two of the variables, Sugar.free.syrup and Regular.syrup, had a strong positive linear relationship with Barista.Espresso. we could make the probability interpretation hereunder:

“Customer buy Barista.Espresso is 6.86 times more likely to buy Regular.syrup”

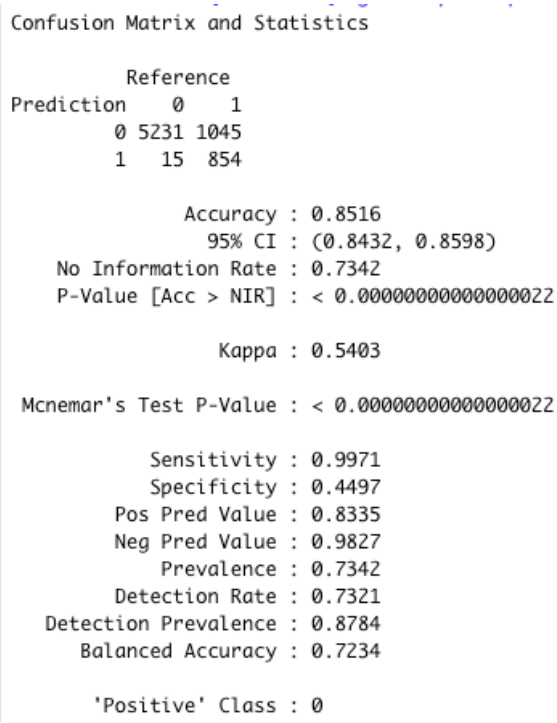
“Customer buy Barista.Espresso is 7.29 times more likely to buy Sugar.free.syrup”

Which highly matches the explanation with the associate rules analysis.

Then, we constructed percent correct predictions and formed a confusion matrix of the training dataset. The accuracy rate of the training dataset was 84.08%. Same as the steps shown above, we applied logistic regression in the validation dataset to review our model



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7809 1696
         1   10 1201

               Accuracy : 0.8408
                 95% CI : (0.8337, 0.8477)
    No Information Rate : 0.7297
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.506

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9987
            Specificity : 0.4146
         Pos Pred Value : 0.8216
         Neg Pred Value : 0.9917
             Prevalence : 0.7297
         Detection Rate : 0.7287
   Detection Prevalence : 0.8870
      Balanced Accuracy : 0.7066

       'Positive' Class : 0
```



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5231 1045
         1   15  854

               Accuracy : 0.8516
                 95% CI : (0.8432, 0.8598)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5403

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9971
            Specificity : 0.4497
         Pos Pred Value : 0.8335
         Neg Pred Value : 0.9827
             Prevalence : 0.7342
         Detection Rate : 0.7321
   Detection Prevalence : 0.8784
      Balanced Accuracy : 0.7234

       'Positive' Class : 0
```

When we interpreted the coefficient estimates in the validation dataset, the regression was consistent with the ones in the training dataset. Next, we formed a confusion matrix of the validation dataset. The accuracy rate of the validation dataset was 85.16%.

5.5.3 Managerial implication, conclusions

With the findings from logistic regressions, we are able to verify again that there is a strong correlation between our signature coffee and syrup choice, also the store can prevent from organizing the signature coffee with other items together for any future promotions.

5.6 Linear regression

5.6.1 Model Analysis

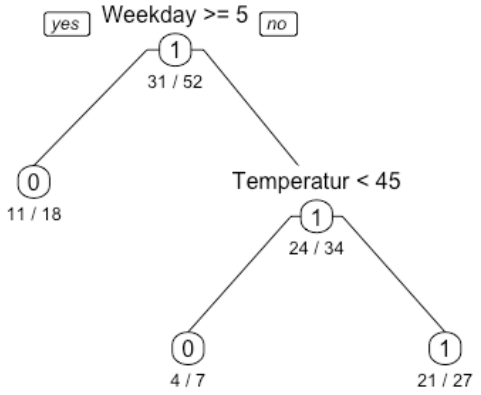
Linear regression was applied to our Sales data, we try to figure out if the daily weather will have some impact on our café business. So we structured training data and validation data by 60% and 40% for transactions. We set the dependent variable “daily sales >average sales”, 1 means that the sales of that day are higher than the monthly average, 0 means that the sales are lower than the monthly average.

```
> summary(income.reg)

Call:
lm(formula = Sales.Average ~ ., data = income.df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9007 -0.5511  0.2588  0.3730  0.5145

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.404403   0.428878   0.943   0.3484
Temperature  0.008308   0.008546   0.972   0.3338
Weekday     -0.052861   0.027438  -1.927   0.0574 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4769 on 84 degrees of freedom
Multiple R-squared:  0.04278,   Adjusted R-squared:  0.01999
F-statistic: 1.877 on 2 and 84 DF,  p-value: 0.1594
```

5.7 Decision Tree

5.7.1 Model Analysis

We also try to draw decision trees based on income training data set and validation data set.




```
> confusionMatrix(default.vt.point.pred.train, as.factor(valid.df.in$Sales.Average))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  0  0
         1 10 25

               Accuracy : 0.7143
                 95% CI : (0.537, 0.8536)
    No Information Rate : 0.7143
    P-Value [Acc > NIR] : 0.584223

                  Kappa : 0

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.7143
             Prevalence : 0.2857
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

5.7.2 Relevant Findings and interpretation

According to the decision trees, we find the weekday is the most influential factor for daily sales that is higher than average. Weekends are the busiest period within the whole week. Furthermore, these decision model predictions have more than 71.43% accuracy by an applied confusion matrix.

5.7.3 Managerial implication, conclusions

From the conclusions we get from decision tree analysis, several inputs include temperature and variables as incoming sales values prediction. The result of this analysis can be useful in terms of store operations for management. Depending on the results shown above, the store manager is able to predict the usage of the food/inventory per day and make the appropriate decisions to restore them. For example, Saturdays are the busiest day out of the week, and food preparation for that day should be sufficient to meet the daily needs. Also, employee scheduling should be arranged based on the results shown above. If not enough or more than enough employees are scheduled to work on that day, it will impact heavily on the gross income for the store owners.

6. Summary

By processing the findings of BI models on our dataset, our team members are all learning how to put BI models into practice. What's more, we can realize how important business intelligence could impact our daily life in every aspect. And the store employees/ owners will get some insight into the internal business and effective operations of their assets.