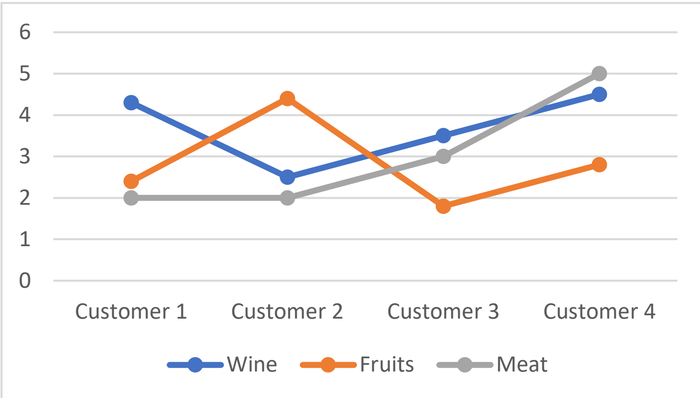


Customer Personality Analysis

DATA ANALYSIS & VISUALIZATION FINAL PROJECT REPORT

Nihit Parikh
MSM '22 | MCCOMBS SCHOOL OF BUSINESS

Data Analysis & Visualization Final Project Report – Nihit Parikh

One-page Summary:

Project Title: Exploratory Data Analysis on Customer Personality Analysis Dataset

This is an exploratory data analysis (EDA) on the customer personality analysis dataset. Businesses perform customer personality analysis to better understand their target group, and how they can modify their products to better fit the needs of their customers.

Initial Goal: to look at customer data for 4 Ps of Marketing: people, product, place, and promotion and come up with insights and recommendations for the marketing mix

Updated Goal: to analyze customer data for 4 Ps of Marketing: people, product, place, and promotion and generate insights using market basket analysis for a smaller sample

Summary in brief: After successfully picking a dataset that was a good mix of marketing and data, a much challenging task was to figure out what story to tell using this data. The dataset talked about variables related to 4 Ps and beginner's goal was to analyze the effect of these variables on customer purchase behavior. After plotting the distribution of each of the people & product variables, it was evident that customer's personality depended on many factors such as age, income, marital status, etc. To determine what products get bought together or are in close proximity, Market Basket Analysis was used. Associate Rule mining was used as an approach to identify strong rules discovered in this Kaggle dataset using some measures of interestingness. This report takes the curtain off the consumer demographics impacting buying behavior at grocery stores and aims to provide customized marketing/advertising recommendations to better serve customers based on their personalities.

Preliminary analysis of specific data to be used

Description of the sources for the data

Source: Kaggle (Customer Personality Analysis Dataset)

The dataset has in total 2240 rows. Each row represents a single customer, and the columns describe the customer in further detail. The columns can generally be separated into four distinct topics; what they all mean is summarized in the following:

People Columns

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if customer complained in the last 2 years, 0 otherwise

Products Columns

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years

- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion Columns

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place Columns

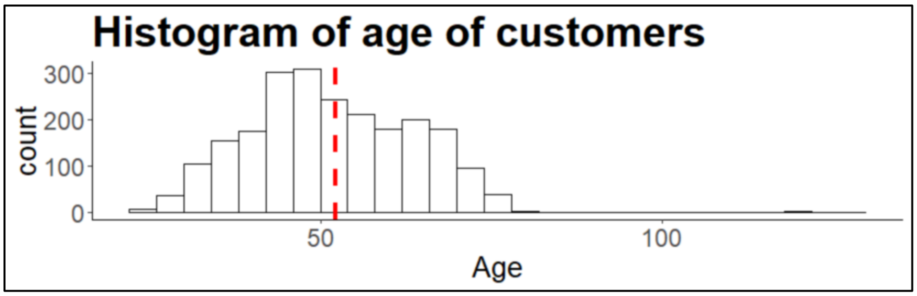
- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's web site in the last month

First, let's try to get a general understanding of these 4 Ps, and then we can further analyze it by looking for interesting interactions between the columns.

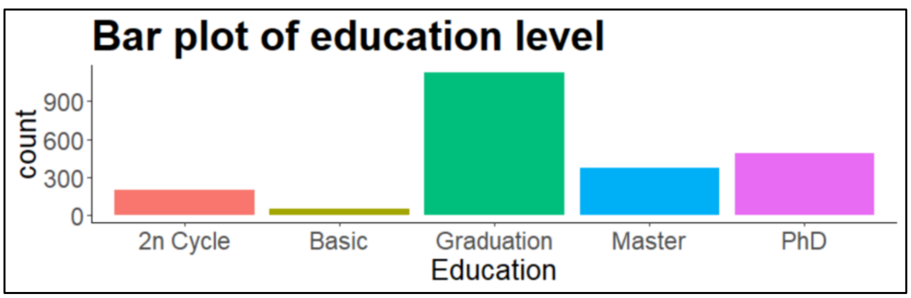
People:

Let's first get a picture of a typical customer.

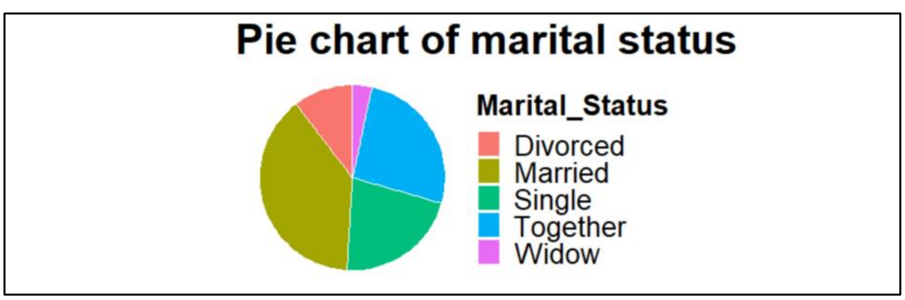
Age:



Education:



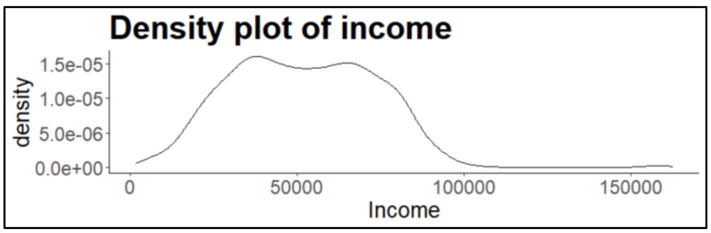
Marital Status:



As you might have noticed from the code, the pie chart does **not** contain all the Marital Statuses, instead the ones with less than 6 have been omitted. While being looked at them, they rather seem to be noise or untruthful answers, and make the chart look worse:

```
> marketing_campaign %>% count(Marital_Status) %>% filter(n<6)
  Marital_Status n
1         Absurd 2
2          Alone 3
3           YOLO 2
```

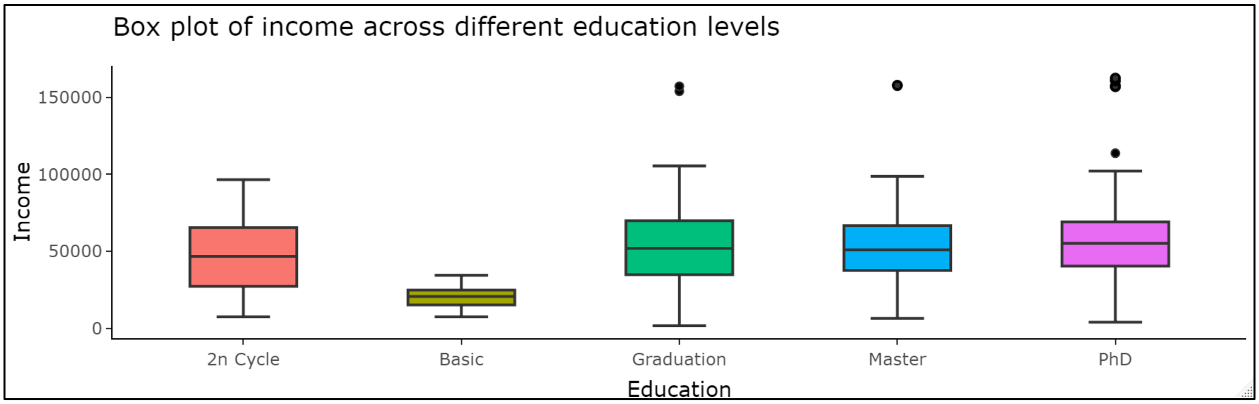
Income:



In the plot above, we filter out one outlier that has his income stated as $666.666, which seems random and is probably just wrong. To prove this, the top 5 Incomes from the dataset are displayed with the following code-snippet.

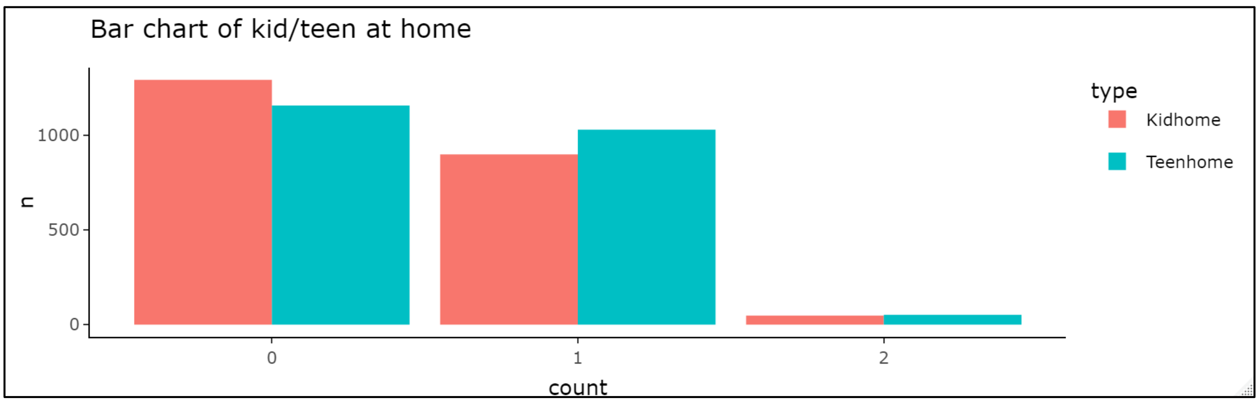
```
> head(sort(marketing_campaign$Income, decreasing=T))
[1] 666666 162397 160803 157733 157243 157146
```

Income distribution across customers with different education levels:



Click here to see the interactive chart

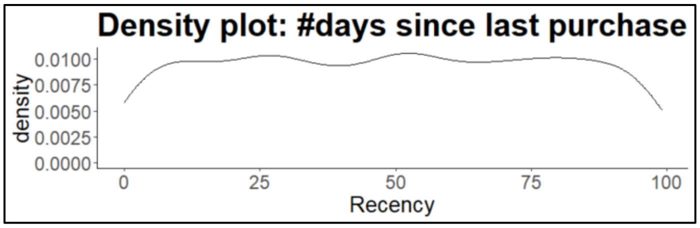
Number of kids/teens at home:



Note: This visualization can be a bit misleading, since the two columns 'Kidhome' and 'Teenhome' are correlated in one way or another and we cannot really tell how many children a customer has on average (will be covered in the next code snippet). Still, I wanted to show it here, since it can still give us some information about the customer.

```
> marketing_campaign %>% summarize(avg_number_of_kids=mean(Kidhome) + mean(Teenhome))
  avg_number_of_kids
1          0.9504464
```

Recency:

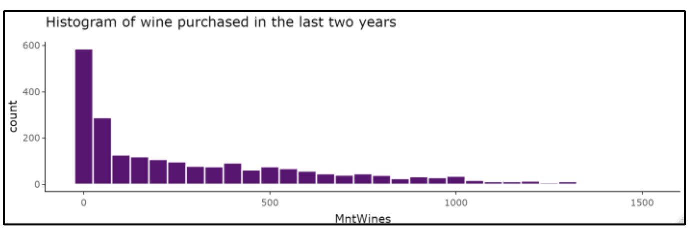


Findings about customer from the 1st P -> People:

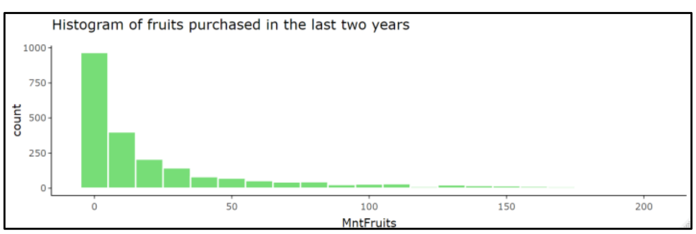
- **Is around 50 years old**
- **Has probably graduated or even a higher degree**
- **Earns around 50k per year**
- **Has 1 child**

Product:

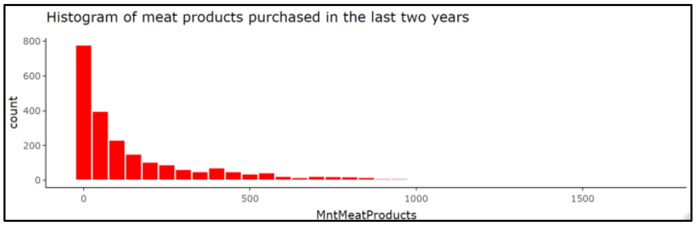
Wine:



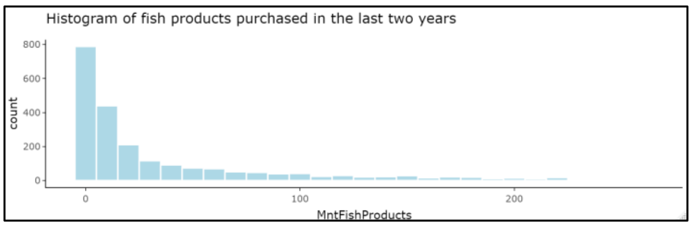
Fruits:



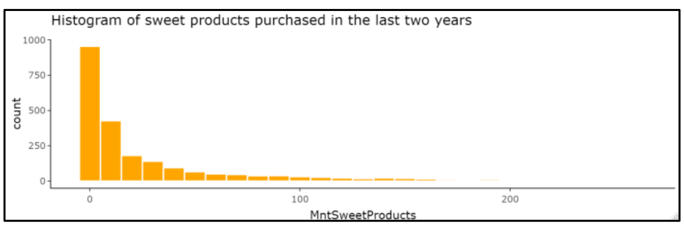
Meat:



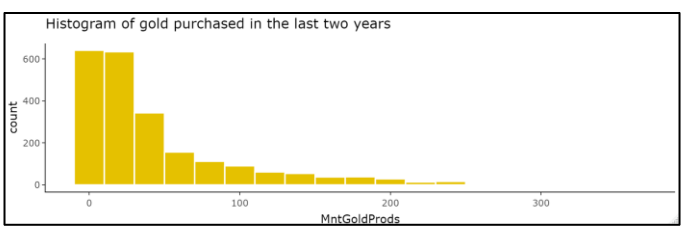
Fish:



Sweets:

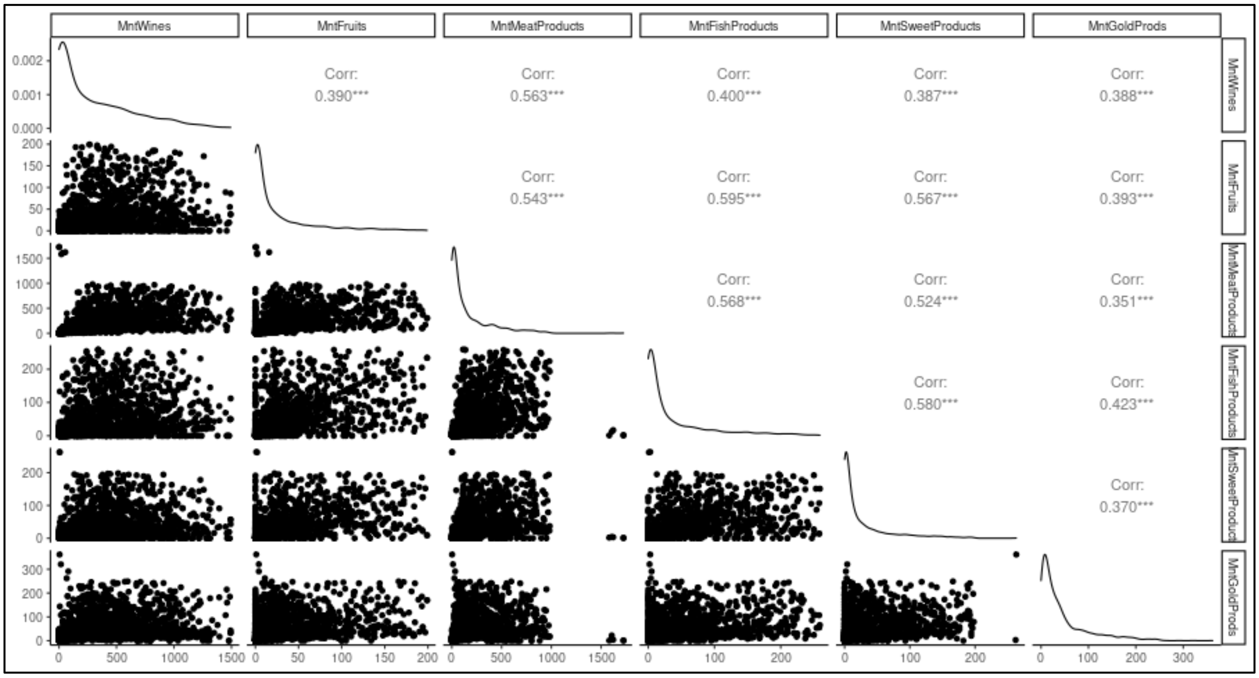


Gold:



An observation while working on this analysis was why are there normal goods like fish, meat, fruits etc. in the dataset, and then out of nowhere, the amount of gold purchased is recorded. Is this a grocery store that sells gold?

Correlation of Products:



Findings about customer from the 2nd P -> Product:

- **Wine seems to go well with meat**
- **Gold is correlated the most with fish**
- **Fish and fruits have the highest correlation**

Promotion & Place:

Research Question: How effective are the marketing campaigns in shaping the customer purchase decision-making journey?

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -2.38504    0.08208 -29.056  < 2e-16 ***
AcceptedCmp1  1.32148    0.21919   6.029 1.65e-09 ***
AcceptedCmp2  0.95780    0.48937   1.957  0.05032 .  
AcceptedCmp3  1.89691    0.18668  10.162  < 2e-16 ***
AcceptedCmp4  0.69863    0.22419   3.116  0.00183 ** 
AcceptedCmp5  1.59388    0.20675   7.709 1.27e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of Logistic Regression:

- If the customer accepts campaign 1, then there will be a 1.32% increase in his probability of responding to the offer.
- Campaign 3 is the most effective and campaign 4 is the least effective in making an impact on consumer purchase behavior.

<u>Research Question:</u> *How effective are the marketing campaigns in persuading the customer to make a purchase with discount offered (number of deals purchased by the customer)?*

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.37111    0.04508  52.597  < 2e-16 ***
AcceptedCmp1 -0.67398    0.18280  -3.687 0.000232 ***
AcceptedCmp2 -0.29438    0.36907  -0.798 0.425175    
AcceptedCmp3 -0.08854    0.16061  -0.551 0.581491    
AcceptedCmp4  0.64160    0.16769   3.826 0.000134 ***
AcceptedCmp5 -1.46959    0.17749  -8.280  < 2e-16 ***
Response      0.44737    0.12473   3.587 0.000342 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.887 on 2233 degrees of freedom
Multiple R-squared:  0.04924,	Adjusted R-squared:  0.04668 
F-statistic: 19.27 on 6 and 2233 DF,  p-value: < 2.2e-16
```

Interpretation of Linear Regression:

- Customer responding to the offer (campaign) has an impact in him/her making a purchase in real.
- Only campaign 4 has (least effective based on previous regression) a positive impact on customer purchasing the deal (this was a bit surprising!)

<u>Research Question:</u> *Do only customers with higher income and/or education levels complain?*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.02422 -0.01389 -0.01027 -0.00403  0.99619

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.501e-02  7.924e-03   3.156  0.00162 **
EducationBasic      -2.287e-02  1.502e-02  -1.522  0.12803
EducationGraduation -6.920e-03  7.447e-03  -0.929  0.35281
EducationMaster     -1.397e-02  8.529e-03  -1.637  0.10171
EducationPhD        -1.703e-02  8.177e-03  -2.082  0.03744 *
Income              -1.051e-07  8.375e-08  -1.255  0.20949
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of Linear Regression:

- Neither Education level nor Income are statistically significant. Therefore, this result means that they don’t influence the type of customers complaining.
- This may not be a result that can be generalized for a realistic larger population as education level and income seem to be factors that may impact consumer awareness and consumer complaints.

<u>Research Question:</u> *What is the effect of a particular product on the recency of customer purchases?*

```
Residuals:
    Min      1Q  Median      3Q     Max
-54.526 -24.607   0.527  25.279  53.127

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      48.4849617  0.9016501  53.774   <2e-16 ***
MntWines          0.0002147  0.0022788   0.094    0.925
MntFruits        -0.0233739  0.0211295  -1.106    0.269
MntMeatProducts   0.0038475  0.0038521   0.999    0.318
MntFishProducts  -0.0121422  0.0158666  -0.765    0.444
MntSweetProducts  0.0221953  0.0198995   1.115    0.265
MntGoldProds      0.0087809  0.0136207   0.645    0.519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of Linear Regression:

- All the model parameters are statistically insignificant meaning that the quantity of grocery item doesn't impact the number of days since a customer has made a purchase.
- For anyone interpreting this for the first time, this may not make sense. But, once one goes into the depth of it, one can say that recency should depend on the quantity of an item (i.e., 7 bananas) bought by the customer the previous time, the perishability of that grocery item, and number of family members of the household consuming this product. The model doesn't really depend on the total quantity purchased by the customers (this is quite debatable!)

<u>Research Question:</u> *What is the effect of number of monthly web visits on number of purchases made through website?*

```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.4248 -2.0411 -0.3609  1.7670 22.6391 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        4.42481    0.14123  31.330   <2e-16 ***
NumWebVisitsMonth -0.06395    0.02417  -2.646   0.0082 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation of Linear Regression:

- The result of this model is a bit surprising as normally assumed in digital marketing that increase in website visits would result in more website purchases as eventually the website visit would lead to call-to-action.
- Why the beta coefficient of number of web visits per month may be negative is unclear.

<u>Research Question:</u> *What is the effect of number of monthly web visits on number of catalog purchases?*

```
Residuals:
    Min      1Q  Median      3Q     Max 
-5.9946 -1.6068 -0.6068  1.1395 22.6323 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        5.99457    0.12707   47.18   <2e-16 ***
NumWebVisitsMonth -0.62682    0.02174  -28.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of Linear Regression:

- The result of this model is a bit surprising as normally assumed in digital marketing that increase in website visits would result in more catalog purchases as eventually the website visit would lead to call-to-action.
- Another way of interpreting this model is that more customers would start purchasing from the website that would negatively impact the catalog purchases.

<u>Research Question:</u> *What is the effect of number of monthly web visits on number of store purchases?*

```
Residuals:
    Min      1Q  Median      3Q     Max 
-8.8420 -2.2498 -0.8238  2.0281  9.3242 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        8.84197    0.14953   59.13   <2e-16 ***
NumWebVisitsMonth -0.57402    0.02559  -22.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of Linear Regression:

- The result of this model is a bit surprising as normally assumed in digital marketing that increase in website visits would result in more store purchases as eventually the website visit would lead to call-to-action.

- Another way of interpreting this model is that more customers would start purchasing from the website that would negatively impact the in-store purchases.

<u>Research Question:</u> Do customers' number of monthly web visits have an effect on their recency of making a purchase?

```
Residuals:
    Min      1Q  Median      3Q     Max
-50.214 -24.678   0.554  25.130  50.577

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       50.4701     1.4740  34.240   <2e-16 ***
NumWebVisitsMonth -0.2559     0.2522  -1.015     0.31
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of Linear Regression:

- The beta coefficient of the explanatory variable is statistically insignificant.

Market Basket Analysis:

Market Basket Analysis is commonly used to make product recommendations by identifying products that are frequently bought together or are in close proximity. **Association rule mining** is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. In any given transaction with a variety of items, association rules are meant to discover the rules that determine how or why certain items are connected.

A rule is a notation that represents which item(s) is frequently bought with what item(s). It has a RHS & LHS, and it can be represented as follows:

itemset A => itemset B (This means, the item/s on the right were frequently purchased along with items on the left.)

How to measure the strength of a rule?

The apriori( ) generates the most relevant set of rules from a given transaction data. It also shows the *support*, *confidence* and *lift* of those rules. These three measures can be used to decide the relative strength of the rules. So, what do these terms mean? Let's consider the rule **A => B** in order to compute these metrics.

$$Support = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions} = P\left(A \cap B\right)$$

$$Confidence = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} = \frac{P\left(A \cap B\right)}{P\left(A\right)}$$

$$ExpectedConfidence = \frac{Number\ of\ transactions\ with\ B}{Total\ number\ of\ transactions} = P\left(B\right)$$

$$Lift = \frac{Confidence}{Expected\ Confidence} = \frac{P\left(A \cap B\right)}{P\left(A\right).P\left(B\right)}$$

Let's look at 10 customers for market basket analysis:

```
> groceries
   Wines Fruits Meat Fish Sweet Gold
1    635     88  546  172    88   88
2     11      1    6    2     1    6
3    426     49  127  111    21   42
4     11      4   20   10     3    5
5    173     43  118   46    27   15
6    520     42   98    0    42   14
7    235     65  164   50    49   27
8     76     10   56    3     1   23
9     14      0   24    3     3    2
10    28      0    6    1     1   16
```

Which are the most frequently bought products on an individual customer basis?

```
> inspect(frequentItems)
     items                                           support count
[1]  {Wines=[11,28), Meat=[24,118), Fish=[3,46)}     0.1     1
[2]  {Wines=[11,28), Meat=[24,118), Sweet=[3,27)}    0.1     1
[3]  {Wines=[11,28), Meat=[24,118), Gold=[2,14)}     0.1     1
[4]  {Wines=[11,28), Fruits=[0,4), Meat=[24,118)}    0.1     1
[5]  {Fruits=[0,4), Meat=[24,118), Fish=[3,46)}      0.1     1
[6]  {Fruits=[0,4), Meat=[24,118), Sweet=[3,27)}     0.1     1
[7]  {Fruits=[0,4), Meat=[24,118), Gold=[2,14)}      0.1     1
[8]  {Wines=[235,635], Meat=[24,118), Fish=[0,3)}    0.1     1
[9]  {Meat=[24,118), Fish=[0,3), Sweet=[27,88]}      0.1     1
[10] {Meat=[24,118), Fish=[0,3), Gold=[14,23)}       0.1     1
```

```
[228] {Meat=[118,546], Gold=[23,88]}                      0.3     3
[229] {Fish=[46,172], Gold=[23,88]}                       0.3     3
[230] {Sweet=[27,88], Gold=[23,88]}                       0.2     2
[231] {Wines=[235,635], Fish=[46,172], Sweet=[27,88]}     0.2     2
[232] {Fruits=[43,88], Fish=[46,172], Sweet=[27,88]}      0.3     3
[233] {Meat=[118,546], Fish=[46,172], Sweet=[27,88]}      0.3     3
[234] {Wines=[235,635], Meat=[118,546], Sweet=[27,88]}    0.2     2
[235] {Fruits=[43,88], Meat=[118,546], Sweet=[27,88]}     0.3     3
[236] {Wines=[235,635], Fruits=[43,88], Sweet=[27,88]}    0.2     2
[237] {Wines=[235,635], Sweet=[27,88]}                    0.3     3
[238] {Fruits=[43,88], Sweet=[27,88]}                     0.3     3
[239] {Meat=[118,546], Sweet=[27,88]}                     0.3     3
[240] {Fish=[46,172], Sweet=[27,88]}                      0.3     3
[241] {Wines=[235,635], Meat=[118,546], Fish=[46,172]}    0.3     3
[242] {Fruits=[43,88], Meat=[118,546], Fish=[46,172]}     0.4     4
[243] {Wines=[235,635], Fruits=[43,88], Fish=[46,172]}    0.3     3
[244] {Wines=[235,635], Fish=[46,172]}                    0.3     3
[245] {Fruits=[43,88], Fish=[46,172]}                     0.4     4
[246] {Meat=[118,546], Fish=[46,172]}                     0.4     4
```

Product Recommendation Rules:

1) Higher Confidence Rules

```
> inspect(head(rules_conf))
    lhs                   rhs               support confidence coverage lift     count
[1] {Gold=[2,14)}     => {Wines=[11,28)}  0.3     1          0.3      3.333333 3
[2] {Wines=[11,28)}   => {Gold=[2,14)}    0.3     1          0.3      3.333333 3
[3] {Fish=[46,172]}   => {Meat=[118,546]} 0.4     1          0.4      2.500000 4
[4] {Meat=[118,546]}  => {Fish=[46,172]}  0.4     1          0.4      2.500000 4
[5] {Fish=[46,172]}   => {Fruits=[43,88]} 0.4     1          0.4      2.500000 4
[6] {Fruits=[43,88]}  => {Fish=[46,172]}  0.4     1          0.4      2.500000 4
```

2) Higher Lift Rules

```
> inspect(head(rules_lift))
    lhs                               rhs              support confidence coverage lift     count
[1] {Gold=[2,14)}                  => {Wines=[11,28)} 0.3     1          0.3      3.333333 3
[2] {Wines=[11,28)}                => {Gold=[2,14)}   0.3     1          0.3      3.333333 3
[3] {Meat=[24,118), Gold=[14,23)} => {Fruits=[4,43)} 0.1     1          0.1      3.333333 1
[4] {Fruits=[4,43), Gold=[14,23)} => {Meat=[24,118)} 0.1     1          0.1      3.333333 1
[5] {Meat=[24,118), Gold=[14,23)} => {Fish=[0,3)}    0.1     1          0.1      3.333333 1
[6] {Meat=[24,118), Fish=[0,3)}   => {Gold=[14,23)}  0.1     1          0.1      3.333333 1
```

The rules with confidence of 1 (see rules_conf above) imply that, whenever the LHS item was purchased, the RHS item was also purchased 100% of the time.

A rule with a lift of 18 (see rules_lift above) imply that, the items in LHS and RHS are 18 times more likely to be purchased together compared to the purchases when they are assumed to be unrelated.

How do I control the number of rules in the output?

```
> rules <- apriori(groceries, parameter = list (supp = 0.001, conf = 0.5, maxlen=2)) # maxlen = 2 limits the el
ements in a rule to 2
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.5    0.1    1 none FALSE            TRUE       5   0.001      1      2  rules TRUE

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

Absolute minimum support count: 0

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[18 item(s), 10 transaction(s)] done [0.00s].
sorting and recoding items ... [18 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 done [0.00s].
writing ... [76 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

How do I remove the redundant rules?

```
> ## Removing the redundant rules
> # get subset rules in vector
> subsetRules <- which(colSums(is.subset(rules, rules)) > 1)
> length(subsetRules)   #> 76
[1] 76
```

How do I find out what customers had purchased before product X?

This can be achieved by modifying the appearance parameter in the apriori( ) function. To find out what customers had purchased before buying 'Sweets'. This will help you understand the patterns that led to the purchase of 'Sweets'.

```
rules <- apriori (data=groceries, parameter=list (supp=0.001, conf = 0.08), appearance = list (default="lhs",rhs="Sweets"), control = list (verbose=F)) # high confidence rules

rules_conf <- sort (rules, by="confidence", decreasing=TRUE)

inspect(head(rules_conf))
```

How do I find out what customers had purchased after/along with product X?

```
rules <- apriori (data=groceries, parameter=list (supp=0.001, conf = 0.15, minlen=2), appearance = list(default="rhs",lhs="Fish"), control = list (verbose=F)) # those who bought 'fish' also bought..

rules_conf <- sort (rules, by="confidence", decreasing=TRUE) # 'high-confidence' rules.

inspect(head(rules_conf))
```

Marketing Recommendations:

- Wine & Gold are in closer proximity and influence customers to buy both when they buy one of them. A good marketing strategy would be to give discount on weekly or monthly discount on wine if they make a gold purchase.
- For meat buyers, giving a discount coupon on fish purchase or vice-versa (after looking at the pricing model) would be a good idea.
- Designing an advertising strategy to target health-conscious individuals, the firm should do email marketing campaigns to promote fish (protein) and fruits (vitamins) together and offer promotional discounts. Influencer marketing could also be a good idea if the grocery retail chain is planning for a nationwide campaign of a substantial campaign budget.
- Looking into more depth of analysis for Gold as a product should be the future goal for the marketing analyst as from the regressions, plots and market basket analysis it seems like that

for an expensive product like that, there is more information needed before the marketing team takes a business decision.

References:

- https://stats.idre.ucla.edu/r/dae/logit-regression/
- http://r-statistics.co/Association-Mining-With-R.html
- https://www.rdocumentation.org/packages/arules/versions/1.7-1/topics/itemFrequencyPlot
- https://rstudio-pubs-static.s3.amazonaws.com/267119_9a033b870b9641198b19134b7e61fe56.html
- https://www.datacamp.com/community/tutorials/market-basket-analysis-r
- http://www.salemmarafi.com/code/market-basket-analysis-with-r/comment-page-1/
- https://en.wikipedia.org/wiki/Association_rule_learning