**MKTA Assignment**

Parthipan V

Roll no. D18025

**Table of Contents:**

1. D3M Case Analysis
2. RFM Analysis
3. Market Basket Analysis
4. Dimensionality reduction
5. Cluster Analysis

**1 D3M Case Analysis:**

Below is the Revenue for two stores:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1. AVERAGE TRANSACTION VALUE** | | | NOTE: Average transaction value = Total revenue / total number of bills | |
|  | **FY - 2015-16** | **(in Rs)** | **Total no of bills** | **Average transaction value** |
|  | REVENUE - STORE 1 | **298000000** | **7,22,439** | **412.4915737** |
|  | REVENUE - STORE 2 | **131000000** | **4,38,422** | **298.7988741** |

Here, the total no of bills for store 1 is nearly 1.6 times that of store 2. And, Average transaction value for store 1 is 1.3 times that of store 2. So, total no of bills and Average transaction value contribute for higher profit in store 1 than store 2.

*Actions for store 2: We can concentrate in increasing the total no of bills through higher customer acquisition.*

Below is the monthly wise footfall and no of bills:

NOTE: Conversion rate = (No. of bills / Footfall)

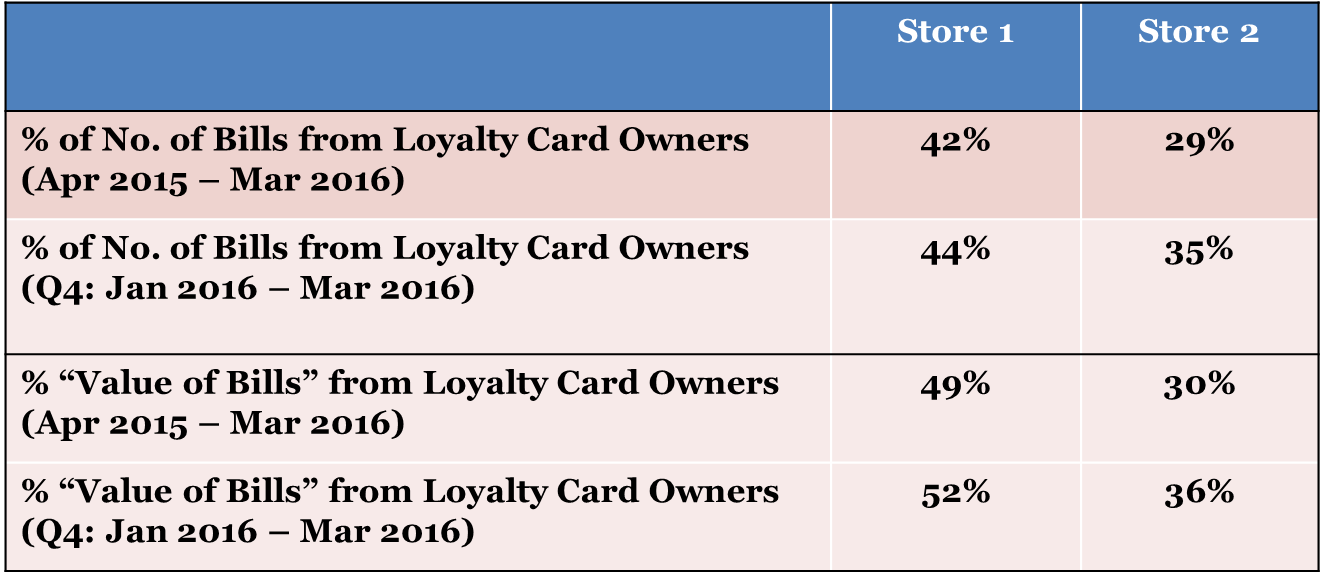
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2. CONVERTION RATES - ON MONTHLY BASIS** | | | |  |
|  | **FOOTFALL AND BILLS DATA - STORE 1** | | |  |
|  |  |  |  |  |
|  | **Months (FY 2015-16)** | **Footfall** | **No. of Bills** | **Convertion Rate** |
|  | Apr | 87630 | 56,908 | **64.94%** |
|  | May | 88976 | 58,976 | **66.28%** |
|  | June | 88874 | 60,432 | **68.00%** |
|  | July | 90685 | 60,420 | **66.63%** |
|  | August | 88976 | 58,118 | **65.32%** |
|  | September | 91877 | 60,998 | **66.39%** |
|  | October | 92765 | 61,980 | **66.81%** |
|  | November | 91744 | 61,654 | **67.20%** |
|  | December | 92870 | 62,080 | **66.85%** |
|  | Jan | 90685 | 62,012 | **68.38%** |
|  | Feb | 88976 | 59,821 | **67.23%** |
|  | March | 90013 | 59,040 | **65.59%** |

|  |  |  |  |
| --- | --- | --- | --- |
| **FOOTFALL AND BILLS DATA - STORE 2** | | |  |
|  |  |  |  |
| **Months (FY 2015-16)** | **Footfall** | **No. of Bills** | **Convertion Rate** |
| Apr | 60019 | 36,690 | **61.13%** |
| May | 60654 | 36,679 | **60.47%** |
| June | 61432 | 36,690 | **59.72%** |
| July | 61891 | 36,177 | **58.45%** |
| August | 61580 | 37,980 | **61.68%** |
| September | 62890 | 38,436 | **61.12%** |
| October | 64581 | 36,122 | **55.93%** |
| November | 64200 | 36,030 | **56.12%** |
| December | 66360 | 36,780 | **55.42%** |
| Jan | 65339 | 35,012 | **53.59%** |
| Feb | 66230 | 35,876 | **54.17%** |
| March | 68682 | 35,950 | **52.34%** |

Here, the conversion rate of store 1 is almost constant for the year. But for store 2, it is decreasing and almost it is decreased from 61.13% in Apr’15 to 52% in Mar’16. This difference in store 2 also one of the key reasons of poor performance in store 2.

*Actions for store 2: We can concentrate in increasing the footfall by increased advertisement campaigns as well as need to improve conversion rates by providing attractive offers such as in store offers and coupons.*

Loyalty card data of stores 1 and 2:



Here, it is clearly seen that the % of No. of Bills from Loyalty Card Owners of store 1 is pretty much higher than store 2 and also store 1 maintains % “Value of Bills” from Loyalty Card Owners as 49% which is nearly half of the revenue comes from Loyal card owners in store 1. While in the case of store 2, both these metrics of Loyal card owners is low and this is also one of the reasons of less revenue in store 2.

*Actions for store 2: We have to concentrate on increasing the number of loyalty customers as they are the one who contributes majorly in increasing revenue that the other customers.*

Customer profile data:

|  |  |  |  |
| --- | --- | --- | --- |
| **Profile of Loyalty Customers (Apr 2015 - Mar 2016)** | | | |
| **DEMOGRAPHICS** | **STORE 1** | **STORE 2** | **National Avg of All Stores** |
| **Gender** |  |  |  |
| Male | 50% | 70% | 53% |
| Female | 50% | 30% | 47% |
| **Age Groups** |  |  |  |
| Below 25yrs | 10% | 25% | 15% |
| 25 - 34yrs | 42% | 30% | 40% |
| 35 - 44yrs | 28% | 15% | 17% |
| 45 - 55yrs | 12% | 15% | 17% |
| 55yr+ | 8% | 15% | 10% |
| **Occupation** |  |  |  |
| Students/Unemployed | 14% | 28% | 20% |
| Housewives | 32% | 24% | 30% |
| Working Executives | 24% | 20% | 21% |
| Self-Employed Professionals | 7% | 12% | 11% |
| Businessmen | 23% | 16% | 18% |

Here, comparing with the national average of all stores with the store 1, almost all the

*Actions to store 2:*

*Gender: Target the Female to increase from 30% to 50% as the national average is also 47%.*

*Age Group: Target in increasing the customers of age 25 – 34yrs as they are the ones with may be high frequent buyers as the national average also 40%.*

*Occupation: Store 2 need to target more on Businessmen, Working executives and Housewives as they are the high valued customers than students/unemployed.*

**STORE 1 - Customer Transactions for Loyalty Card Users**

Average sales per transaction = (total bill value / no of transactions which is 100 here)

**1115.43**

|  |  |  |
| --- | --- | --- |
| **Average sales of category** | | |
| No of txn of FMCG is : 89 | |  |
| Average sales of FMCG category = (total bill value of FMCG category / no of transactions containing FMCG which is 89 here) | | |
| 459.595506 |  |  |
| No of txn of APPAREL is : 49 | |  |
| Average sales of APPAREL category = (total bill value of APPAREL category / no of transactions containing APPAREL which is 49 here) | | |
| 947.040816 |  |  |
| No of txn of Others : 93 | |  |
| Average sales of OTHERS category = (total bill value of OTHERS category / no of transactions containing OTHERS which is 93 here) | | |
| 260.580645 |  |  |

**STORE 2 - Customer Transactions for Loyalty Card Users**

Average sales per transaction = (total bill value / no of transactions which is 100 here)

**715.4**

|  |  |  |
| --- | --- | --- |
| **Average sales of category** | | |
| No of txn of FMCG is : 79 | |  |
| Average sales of FMCG category = (total bill value of FMCG category / no of transactions containing FMCG which is 79 here) | | |
| 409.367089 |  |  |
| No of txn of APPAREL is : 25 | |  |
| Average sales of APPAREL category = (total bill value of APPAREL category / no of transactions containing APPAREL which is 25 here) | | |
| 1088.28 |  |  |
| No of txn of Others : 72 | |  |
| Average sales of OTHERS category = (total bill value of OTHERS category / no of transactions containing OTHERS which is 72 here) | | |
| 166.569444 |  |  |

From the above results from store 1 and 2 on transaction data for loyal card users, it is clearly visible that the Average sales per transaction is higher in store 1 than store 2. Also, in each category wise, the average sales is higher in store 1.

*Actions to store 2:*

*It is seen that the average sales for APPAREL category in store 2 is higher than the store 1, but the frequency of customer for this category is only 25 while in store 1 it is 49. So, it is advised to increase the frequency of customers to buy apparel in store 2 so that it will improve the revenue.*

|  |  |  |  |
| --- | --- | --- | --- |
| **MYSTERY SHOPPING FEEDBACK** | |  |  |
|  |  |  |  |
|  | **Store 1** | **Store 2** | **Avg Store X** |
| **Store Ratings** |  |  |  |
| Temperature |  |  |  |
| 5 (%Top Box) | 52% | 74% | 61% |
| 4 | 21% | 14% | 21% |
| 3 | 16% | 8% | 14% |
| 2 | 9% | 2% | 3% |
| 1 | 2% | 2% | 1% |
| Lighting |  |  |  |
| 5 (%Top Box) | 66% | 74% | 64% |
| 4 | 16% | 10% | 23% |
| 3 | 16% | 13% | 11% |
| 2 | 2% | 2% | 1% |
| 1 | 0% | 1% | 1% |
| Music |  |  |  |
| 5 (%Top Box) | 58% | 67% | 60% |
| 4 | 20% | 17% | 21% |
| 3 | 16% | 12% | 13% |
| 2 | 4% | 2% | 3% |
| 1 | 2% | 2% | 3% |
| **Staff Ratings** |  |  |  |
| Friendliness |  |  |  |
| 5 (%Top Box) | 62% | 73% | 68% |
| 4 | 17% | 20% | 26% |
| 3 | 21% | 5% | 5% |
| 2 | 0% | 2% | 1% |
| 1 | 0% | 0% | 0% |
| Product Knowledge |  |  |  |
| 5 (%Top Box) | 66% | 69% | 68% |
| 4 | 23% | 21% | 23% |
| 3 | 11% | 7% | 7% |
| 2 | 0% | 2% | 1% |
| 1 | 0% | 1% | 1% |

While seeing the Mystery Shopping feedback data, all the characteristics of store and staff, the ratings are higher for store 2 than the store 1 but though poor performance of store 2 is not dependent on this feedback. So, the store 2 can concentrate on other features as well.

**2 RFM Analysis:**

The RFM analysis is done using Python codes. The given dataset has 25613 entries with unique customers’ count of 13787 and order dates ranging from 2014-01-01 to 2014-03-31.

Recency , Frequency and Monetary (RFM) values are calculated and then divided into 5 parts as quintiles. Each of the RFM has scores ranging from 1 to 5. Then concatenating the scores to get the RFM score. The RFM score table is exported as csv file.

Quick analysis 1: Who are the top 5 best customers? by RFM Class (111), buy often and spend the most with recent transaction?

| **recency** | **frequency** | **monetary** | **R** | **F** | **M** | **RFM Score** |
| --- | --- | --- | --- | --- | --- | --- |
| **customer** |  |  |  |  |  |  |  |
| **Brenna Boehm** | 88 | 1 | 39 | 1 | 1 | 1 | 111 |
| **Emory Koepp** | 81 | 1 | 39 | 1 | 1 | 1 | 111 |
| **Odis Schoen** | 88 | 1 | 39 | 1 | 1 | 1 | 111 |
| **Noelle Turcotte** | 80 | 1 | 39 | 1 | 1 | 1 | 111 |
| **Nile Wiegand** | 84 | 1 | 39 | 1 | 1 | 1 | 111 |

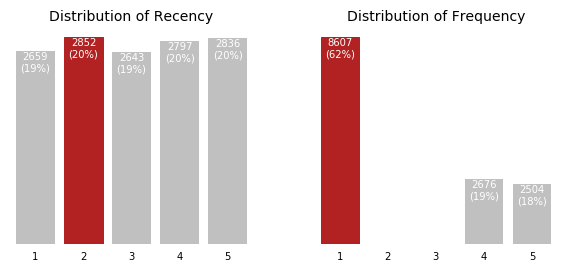
The RFM score gives us 5 \* 5 \* 5 = 125 segments which is very tough to interpret the results. So, the RFM score is then divided into 10 segments based on R and F scores as:

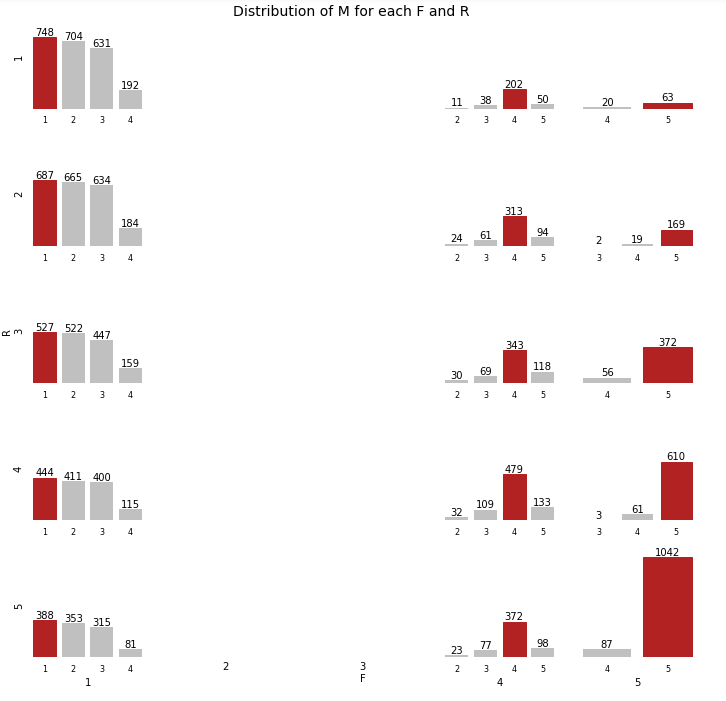
|  |  |
| --- | --- |
| **Segment** | **Description** |
| Champions | Bought recently, buy often and spend the most |
| Loyal Customers | Buy on a regular basis. Responsive to promotions. |
| Potential Loyalist | Recent customers with average frequency. |
| Recent Customers | Bought most recently, but not often. |
| Promising | Recent shoppers, but haven’t spent much. |
| Customers Needing Attention | Above average recency, frequency and monetary values. May not have bought very recently though. |
| About To Sleep | Below average recency and frequency. Will lose them if not reactivated. |
| At Risk | Purchased often but a long time ago. Need to bring them back! |
| Can’t Lose Them | Used to purchase frequently but haven’t returned for a long time. |
| Hibernating | Last purchase was long back and low number of orders. May be lost. |

The resulting matrix will look like this:



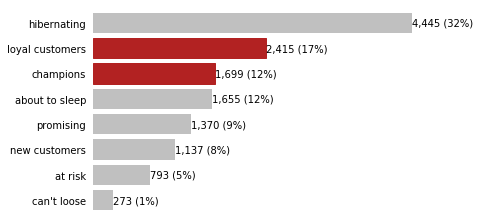
***Visualizing the RFM***:





When looking at the monetary value, we see that the customers spending the most are those with the highest activity (R and F of 4-5).

Plotting the number of customers in each segments:



We have a lot of customers who don't buy frequently from us (32% are hibernating). However, 29% of our customers are either champions or loyal customers.

**3 Market Basket Analysis:**

The given retail dataset of groceries contains 9835 transactions with the basket size of 12 items. The total unique items are 169 items. Also, there are 2 transactions are duplicated and removed it.

Then the dataset is transformed into in such a way that the columns as 169 items along with the transaction id column and the rows as 9833 transactions with the values as 0 or 1 where 0 as the customer didn’t purchase the item and 1 as the customer did purchase the item irrespective of the quantity of items purchased.

Below is the list of 169 unique items:

['citrus fruit', 'tropical fruit', 'whole milk', 'pip fruit',

'other vegetables', 'rolls/buns', 'pot plants', 'beef',

'frankfurter', 'chicken', 'butter', 'fruit/vegetable juice',

'packaged fruit/vegetables', 'chocolate', 'specialty bar',

'butter milk', 'bottled water', 'yogurt', 'sausage', 'brown bread',

'hamburger meat', 'root vegetables', 'pork', 'pastry',

'canned beer', 'berries', 'coffee', 'misc. beverages', 'ham',

'turkey', 'curd cheese', 'red/blush wine',

'frozen potato products', 'flour', 'sugar', 'frozen meals',

'herbs', 'soda', 'detergent', 'grapes', 'processed cheese', 'fish',

'sparkling wine', 'newspapers', 'curd', 'pasta', 'popcorn',

'finished products', 'beverages', 'bottled beer', 'dessert',

'dog food', 'specialty chocolate', 'condensed milk', 'cleaner',

'white wine', 'meat', 'ice cream', 'hard cheese', 'cream cheese ',

'liquor', 'pickled vegetables', 'liquor (appetizer)', 'UHT-milk',

'candy', 'onions', 'hair spray', 'photo/film', 'domestic eggs',

'margarine', 'shopping bags', 'salt', 'oil', 'whipped/sour cream',

'frozen vegetables', 'sliced cheese', 'dish cleaner',

'baking powder', 'specialty cheese', 'salty snack',

'Instant food products', 'pet care', 'white bread',

'female sanitary products', 'cling film/bags', 'soap',

'frozen chicken', 'house keeping products', 'spread cheese',

'decalcifier', 'frozen dessert', 'vinegar', 'nuts/prunes',

'potato products', 'frozen fish', 'hygiene articles',

'artif. sweetener', 'light bulbs', 'canned vegetables',

'chewing gum', 'canned fish', 'cookware', 'semi-finished bread',

'cat food', 'bathroom cleaner', 'prosecco', 'liver loaf',

'zwieback', 'canned fruit', 'frozen fruits', 'brandy',

'baby cosmetics', 'spices', 'napkins', 'waffles', 'sauces', 'rum',

'chocolate marshmallow', 'long life bakery product', 'bags',

'sweet spreads', 'soups', 'mustard', 'specialty fat',

'instant coffee', 'snack products', 'organic sausage',

'soft cheese', 'mayonnaise', 'dental care', 'roll products ',

'kitchen towels', 'flower soil/fertilizer', 'cereals',

'meat spreads', 'dishes', 'male cosmetics', 'candles', 'whisky',

'tidbits', 'cooking chocolate', 'seasonal products', 'liqueur',

'abrasive cleaner', 'syrup', 'ketchup', 'cream', 'skin care',

'rubbing alcohol', 'nut snack', 'cocoa drinks', 'softener',

'organic products', 'cake bar', 'honey', 'jam', 'kitchen utensil',

'flower (seeds)', nan, 'rice', 'tea', 'salad dressing',

'specialty vegetables', 'pudding powder', 'ready soups',

'make up remover', 'toilet cleaner', 'preservation products',

'sound storage medium']

Then, the frequent item sets with minimum support of 0.01 are generated using apriori method. Also, the rules are generated as with the metric as “lift” with minimum lift value of 1.

Then the association rules are filtered with lift value greater than or equal to 2 and confidence scores of greater than equal to 0.3 and the results are given as:

| **antecedents** | **consequents** | **antecedent support** | **consequent support** | **support** | **confidence** | **lift** | **leverage** | **conviction** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (pip fruit) | (tropical fruit) | 0.075460 | 0.104648 | 0.019933 | 0.264151 | 2.524195 | 0.012036 | 1.216761 |
| (beef) | (root vegetables) | 0.052375 | 0.107800 | 0.014136 | 0.269903 | 2.503731 | 0.008490 | 1.222029 |
| (citrus fruit) | (tropical fruit) | 0.082681 | 0.104648 | 0.019729 | 0.238622 | 2.280247 | 0.011077 | 1.175964 |
| (whole milk, other vegetables) | (root vegetables) | 0.066918 | 0.107800 | 0.015865 | 0.237082 | 2.199272 | 0.008651 | 1.169457 |
| (curd) | (yogurt) | 0.052578 | 0.135767 | 0.015458 | 0.294004 | 2.165498 | 0.008320 | 1.224132 |
| (root vegetables, whole milk) | (other vegetables) | 0.038849 | 0.189464 | 0.015865 | 0.408377 | 2.155432 | 0.008504 | 1.370021 |
| (whipped/sour cream) | (yogurt) | 0.070070 | 0.135767 | 0.020441 | 0.291727 | 2.148729 | 0.010928 | 1.220197 |
| (onions) | (other vegetables) | 0.031018 | 0.189464 | 0.012611 | 0.406557 | 2.145829 | 0.006734 | 1.365820 |
| (cream cheese ) | (yogurt) | 0.039154 | 0.135767 | 0.011187 | 0.285714 | 2.104441 | 0.005871 | 1.209926 |
| (root vegetables) | (other vegetables) | 0.107800 | 0.189464 | 0.041696 | 0.386792 | 2.041508 | 0.021272 | 1.321797 |
| (other vegetables) | (root vegetables) | 0.189464 | 0.107800 | 0.041696 | 0.220075 | 2.041508 | 0.021272 | 1.143956 |

Interpretation of the above for few:

Here, few rules with a high lift value which means that it occurs more frequently than would be expected given the number of transaction and product combinations and also confidence is high as well.

It is seen that for the items pip fruit and tropical fruit:

Support = P(pip fruit AND tropical fruit) = 0.019933

i.e., Probability of buying both the pip and tropical fruits is 0.019.

Confidence = P(tropical fruit | pip fruit) = 0.264151

i.e., Probability of buying tropical fruit given that pip fruit is already bought is 0. 264151.

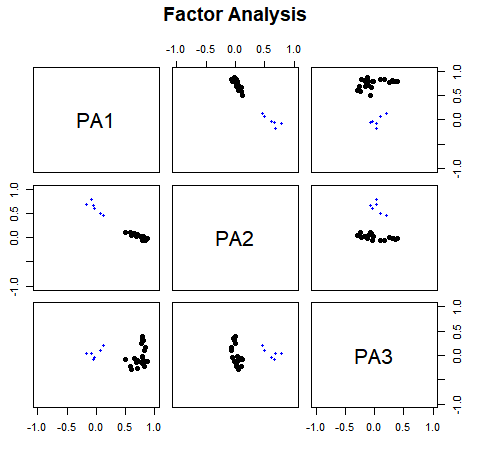
Lift = P(pip fruit AND tropical fruit) / P(pip fruit) \* P(tropical fruit) = 2.524195

Lift is interpreted as that “Customers who buy pip fruit are 2.524195 times as likely to buy tropical fruit as customers from the entire data set”.

**4 Dimensionality reduction**

Here, for the given data, the factor analysis need to be done to explain what each factors are meant for. The EFA is considered here without any rotation with number of factors as 3:

The plot is:



And the results:

Factor Analysis using method = pa

Call: fa(r = cor(df), nfactors = 3, rotate = "none", fm = "pa")

Standardized loadings (pattern matrix) based upon correlation matrix

PA1 PA2 PA3 h2 u2 com

q31ab 0.70 0.51 0.49 1.1

q31ac 0.61 0.45 0.55 1.4

q31ad 0.69 0.50 0.50 1.1

q31ae 0.70 0.56 0.44 1.3

q31af 0.50 0.27 0.73 1.1

q31ag 0.46 0.27 0.73 1.5

q31ah 0.58 0.40 0.60 1.4

q31ba 0.82 0.68 0.32 1.0

q31bb 0.77 0.61 0.39 1.1

q31bc 0.87 0.77 0.23 1.0

q31bd 0.82 0.73 0.27 1.1

q31be 0.80 0.68 0.32 1.1

q31bf 0.78 0.61 0.39 1.0

q31bg 0.79 0.64 0.36 1.0

q31bh 0.66 0.45 0.55 1.0

q31bi 0.78 0.62 0.38 1.0

q31bj 0.66 0.44 0.56 1.1

q31bk 0.79 0.63 0.37 1.0

q31bl 0.61 0.37 0.63 1.0

q31ca 0.77 0.66 0.34 1.2

q31cb 0.81 0.31 0.76 0.24 1.3

q31cc 0.83 0.70 0.30 1.0

q31cd 0.69 0.51 0.49 1.1

q31ce 0.84 0.74 0.26 1.1

q31cf 0.79 0.39 0.77 0.23 1.5

q31cg 0.79 0.35 0.75 0.25 1.4

q31ch 0.79 0.32 0.73 0.27 1.3

q31ci 0.50 0.27 0.73 1.2

PA1 PA2 PA3

SS loadings 12.61 2.41 1.04

Proportion Var 0.45 0.09 0.04

Cumulative Var 0.45 0.54 0.57

Proportion Explained 0.79 0.15 0.06

Cumulative Proportion 0.79 0.94 1.00

Mean item complexity = 1.2

Test of the hypothesis that 3 factors are sufficient.

The degrees of freedom for the null model are 378 and the objective function was 21.73

The degrees of freedom for the model are 297 and the objective function was 2.77

The root mean square of the residuals (RMSR) is 0.04

The df corrected root mean square of the residuals is 0.05

Fit based upon off diagonal values = 0.99

Measures of factor score adequacy

PA1 PA2 PA3

Correlation of (regression) scores with factors 0.99 0.91 0.87

Multiple R square of scores with factors 0.97 0.82 0.76

Minimum correlation of possible factor scores 0.95 0.65 0.52

**5 Cluster Analysis:**

The given dataset is about PC User groups who are more or less likely to use digital devices. It is read in pandas as excel file with the required sheet. The data has 2158 row / observations and 45 attributes of each customer.

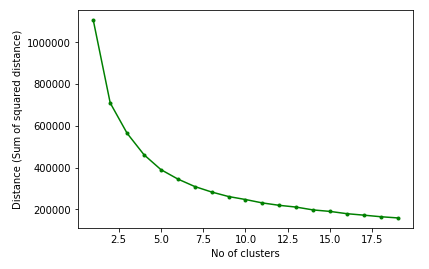
Then the data set is converted into dummies data set for all the nominal dataset. As per the requirement, there are three questions need to be covered:

**Do a cluster analysis. Which type of cluster analysis is suitable here?**

Here, I have considered the usage of K-Means cluster analysis. The distance measure of Euclidean distance (sum of squared) is used in this cluster analysis.

**What should be the ideal number of cluster or clusters? Give reasons behind your statement.**

As per the graph below plotted between sum of squared distances and number of clusters:

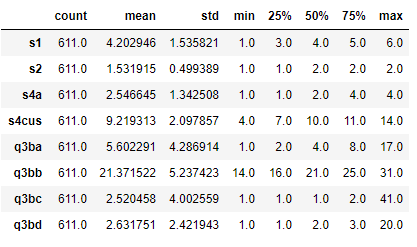


Here, the optimal cluster might be the one with number of clusters to be considered is 6.

**Do a profiling for the variables s1, s2, s4a, s4cus, q3ba, q3bb, q3bc, q3bd corresponding to each of these clusters and comment.**

After creating 6 clusters with all the 45 columns, we need to do profiling on these 8 columns: ['s1', 's2', 's4a', 's4cus', 'q3ba', 'q3bb', 'q3bc', 'q3bd']

Cluster 1 has description of:



And mode as:

| **s1** | **s2** | **s4a** | **s4cus** | **q3ba** | **q3bb** | **q3bc** | **q3bd** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 5 | 2 | 4 | 7 | 3 | 21 | 1 | 1 |