CAPSTONE PROJECT
CUSTOMER'S SENTIMENT ANALYSIS

Dataset Link : https://www.kaggle.com/imakash3011/customer-personality-analysis

Description:

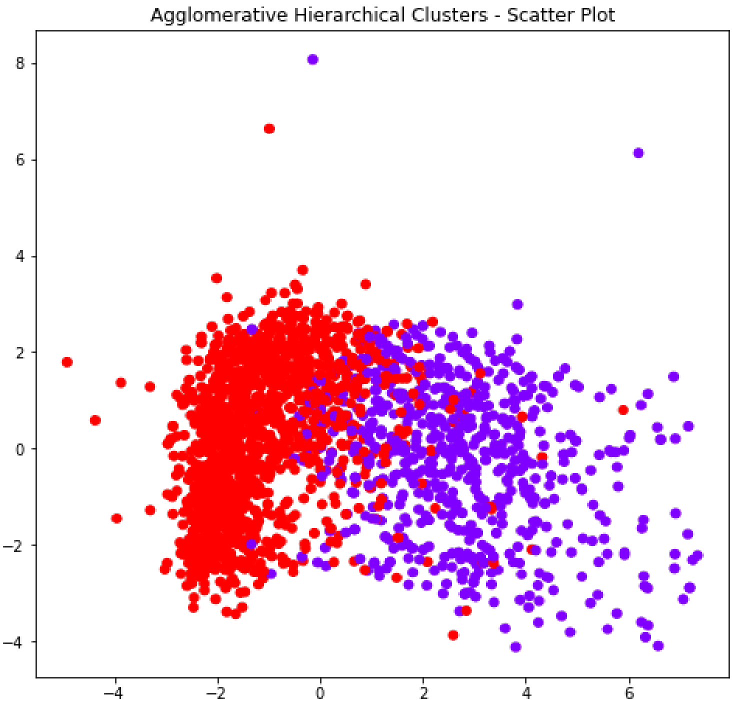
The dataset is about the company's customers who oftenly purchase products from them through different purchase channels

Customer Personality Analysis is a detailed analysis of a company's customers. It helps a business to better understand its customers' behaviour and makes it easier for them to modify products according to customers needs, behaviors and concerns.

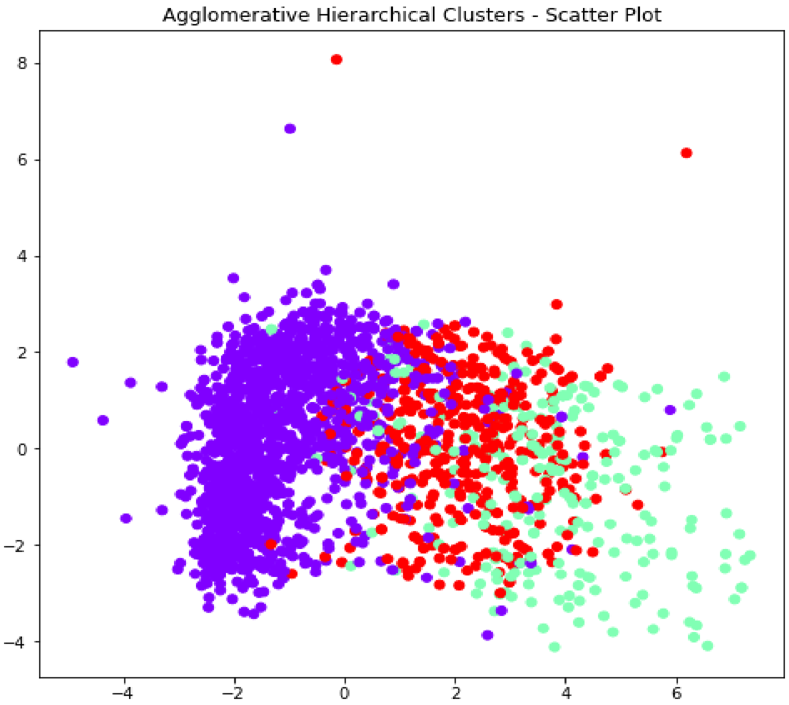
Customer personality analysis helps a business to modify its product based on its target customers from different types of customers segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

ANALYSIS:

1. To perform Hierarchical clustering in order to group the customers according to their behaviour using different parameters



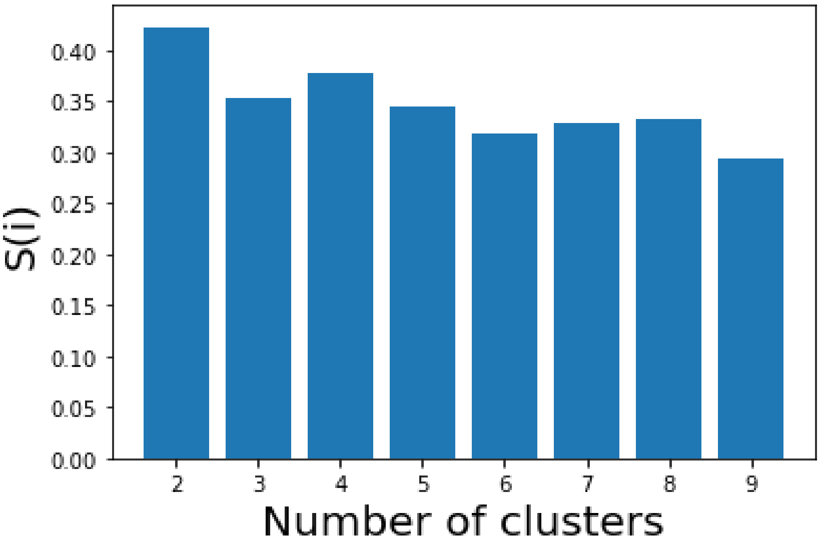
Clusters=2

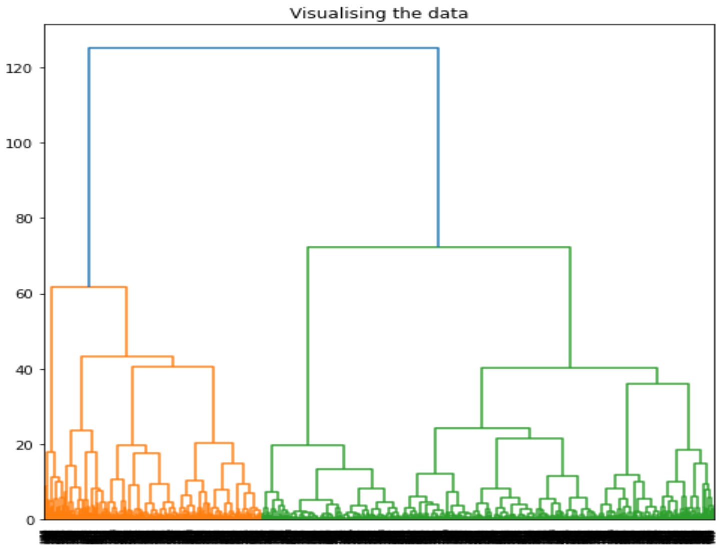


Clusters=3

We will select no of clusters=2 as we are getting more specific groups as compared to 3 clusters as in 3 clusters there is more overlapping of points

2. Using silhouette score and dendrogram we conclude that no of clusters should be 2
   As silhouette score is maximum for n=2





And for dendrogram if we draw a horizontal line it cut the dendrogram at two points without cutting any block .so,from here also it depicts no of clusters should be 2

REGRESSION:

For the same dataset we used machine learning\regression technique also to see whether the variables we have in our data set have their impact on the income of the customer or not
.
This analysis can be further used to predict the income of the customers if we have data of other variables available as usually customers hesitate to share the information about their income

1. First regression technique we use was **Decision Tree** and the result we got
   Train accuracy: 0.5689509837374063
   Test accuracy: 0.6278543788000224

Which is not so good then we use **Random Forest** to see whether there is any improvement in result or not

2. **Random Forest**
   Train accuracy 0.7272938306049124
   Test accuracy: 0.7306002321736864

There is quite balance in train and test accuracy so we can check about the cost function

test;
Mean Absolute Error: 6831.9937694746195
Mean Absolute Percentage Error: 26.201479053992085
Mean Squared Error: 122676793.70871527
Root Mean Squared Error: 11075.955656678807

3. **Light GBM**
   Train accuracy 0.7804394045358937
   Test accuracy: 0.7752969496944222

test:
Mean Absolute Error: 7030.873447315789
Mean Absolute Percentage Error: 24.097385589061762
Mean Squared Error: 102323212.7127466
Root Mean Squared Error: 10115.493695947154

Among the Above three regression models based on their performance Light Gbm performs much better than other two in terms of accuracy and cost function also

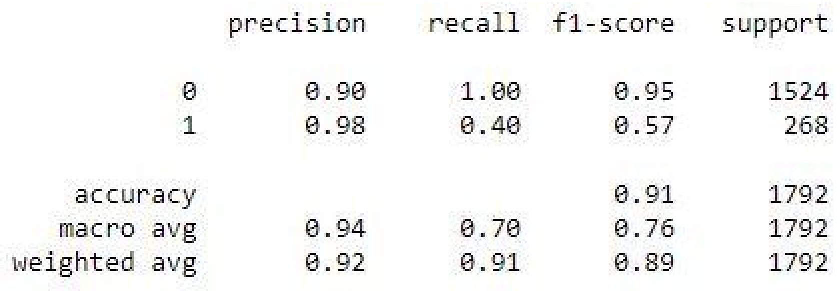
CLASSIFICATION:

For the same dataset we used machine learning\classification technique also to see whether the variables we have in our data set have their impact on the response of the customer or not .
This analysis can be further used to classify the customers on the basis that whether they response or not and then we can use that model to predict whether the new customers engage with us will respond to us in future or not so that we can decide whether to contact that customer or not

1.First classification technique we used was Random Forest and result we got

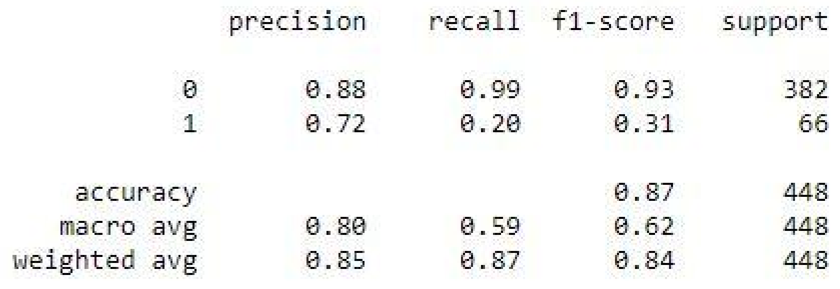
Train Accuracy 0.9095982142857143
Test Accuracy 0.8705357142857143

Train Classification Report:



```
              precision    recall  f1-score   support

           0       0.90      1.00      0.95      1524
           1       0.98      0.40      0.57       268

    accuracy                           0.91      1792
   macro avg       0.94      0.70      0.76      1792
weighted avg       0.92      0.91      0.89      1792
```

Test Classification Report:



```
              precision    recall  f1-score   support

           0       0.88      0.99      0.93       382
           1       0.72      0.20      0.31        66

    accuracy                           0.87       448
   macro avg       0.80      0.59      0.62       448
weighted avg       0.85      0.87      0.84       448
```

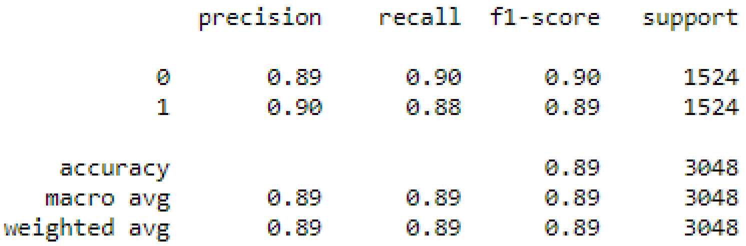
From confusion matrix it can be seen that there is huge variation in recall for 0 and 1 class because of huge imbalance in data

So, in order to overcome this problem we use smote technique for train data so as to balance between classes is maintained

And the result we got after doing smote is

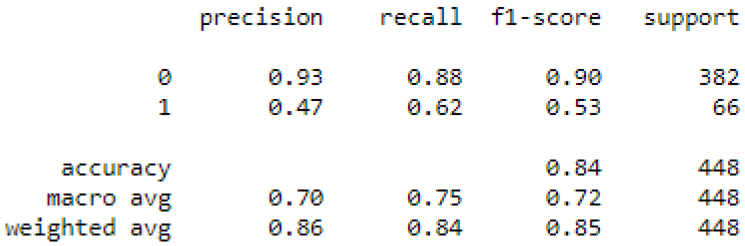
Train accuracy 0.8943569553805775
Test accuracy: 0.8392857142857143

Train Classification Report After Smote:



```
              precision    recall  f1-score   support

           0       0.89      0.90      0.90      1524
           1       0.90      0.88      0.89      1524

    accuracy                           0.89      3048
   macro avg       0.89      0.89      0.89      3048
weighted avg       0.89      0.89      0.89      3048
```

Test Classification Report After Smote:

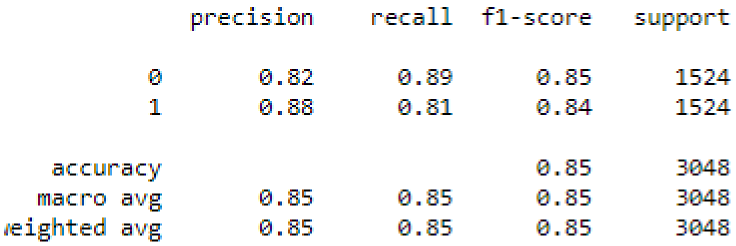


```
              precision    recall  f1-score   support

           0       0.93      0.88      0.90       382
           1       0.47      0.62      0.53        66

    accuracy                           0.84       448
   macro avg       0.70      0.75      0.72       448
weighted avg       0.86      0.84      0.85       448
```

2. Light GBM Classifier:

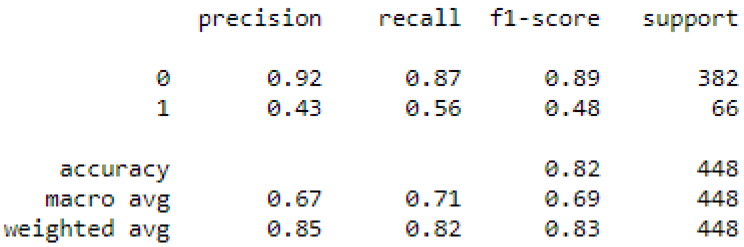
Train accuracy 0.8471128608923885
Test accuracy: 0.8236607142857143

Train Classification Report:



```
              precision    recall  f1-score   support

           0       0.82      0.89      0.85      1524
           1       0.88      0.81      0.84      1524

    accuracy                           0.85      3048
   macro avg       0.85      0.85      0.85      3048
ieighted avg       0.85      0.85      0.85      3048
```

Test Classification Report:



```
              precision    recall  f1-score   support

           0       0.92      0.87      0.89       382
           1       0.43      0.56      0.48        66

    accuracy                           0.82       448
   macro avg       0.67      0.71      0.69       448
weighted avg       0.85      0.82      0.83       448
```

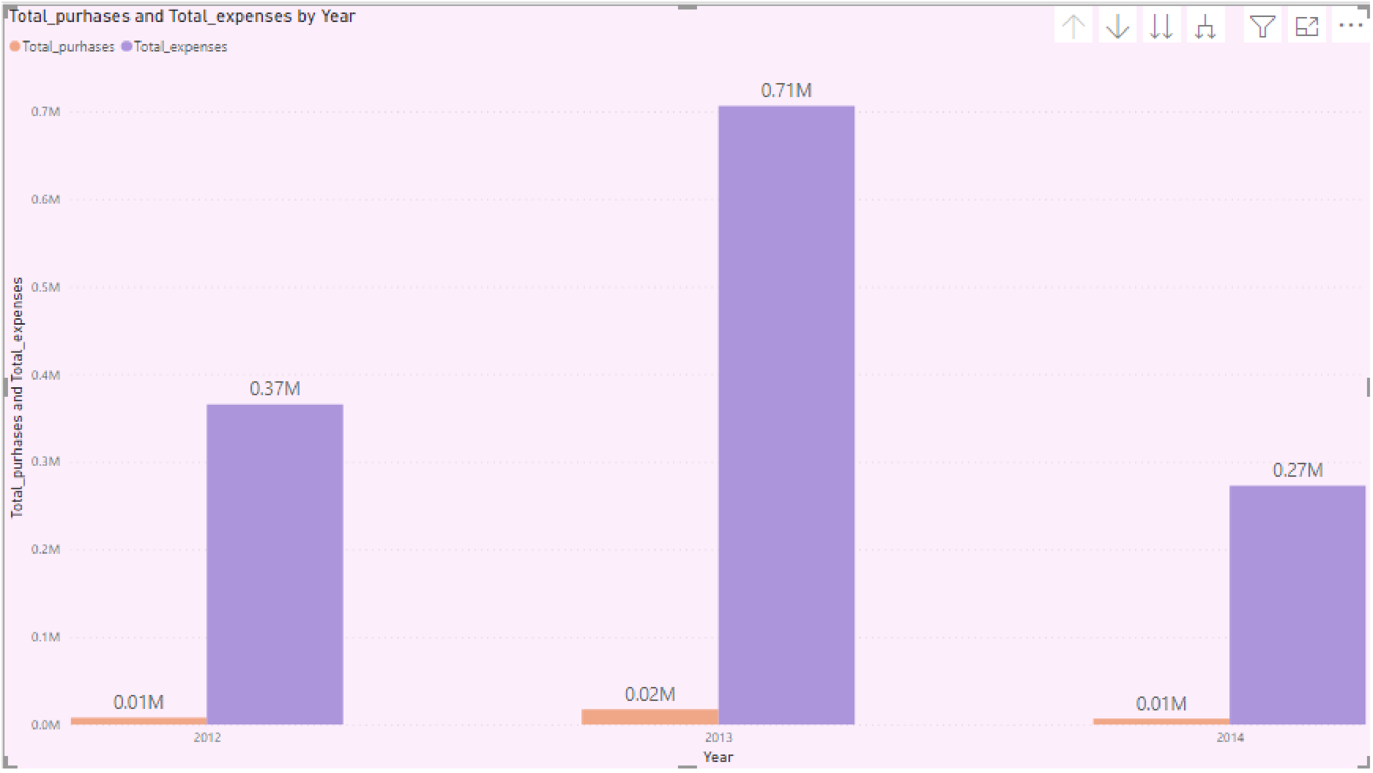
There is still variation in recall for both the classes ,the reason behind this is imbalance in the data

If we compare the performance, in this case Random Forest works better then Light GBM

Visualisation Using Power BI

The visualisation we did for various features to understand the relation and impact of variables on each other

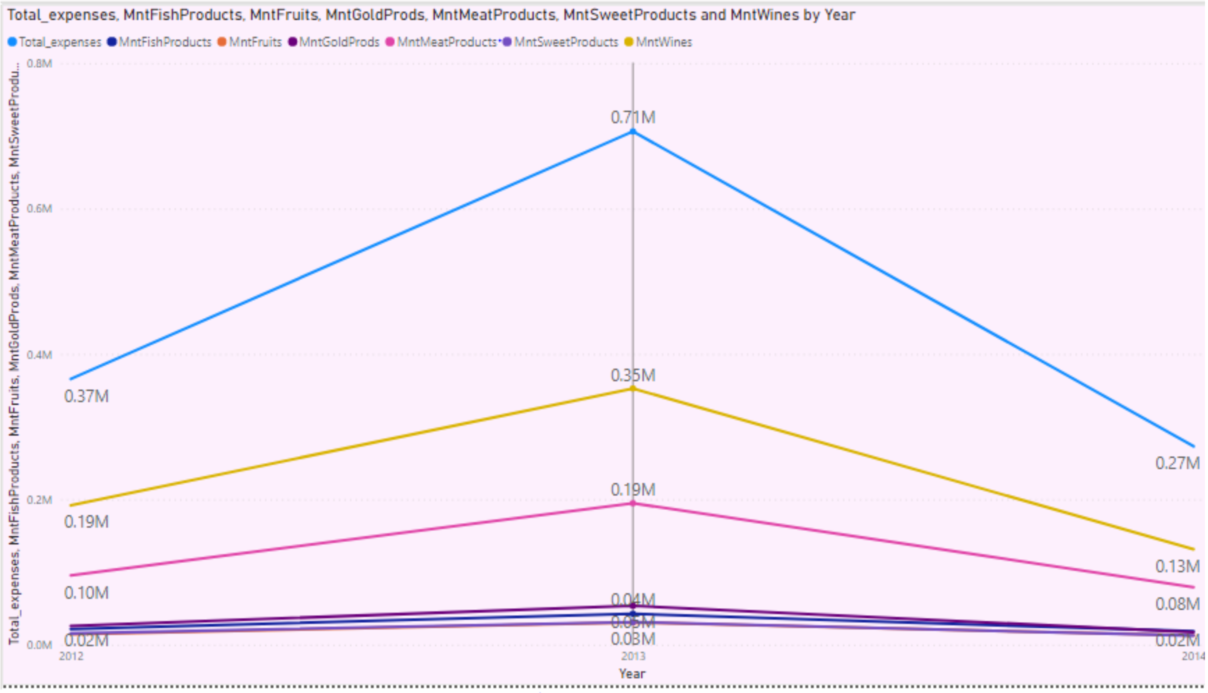
1. Total Purchase and Total Expenses by Year


Total_purhases and Total_expenses by Year
● Total_purhases ● Total_expenses
Total_purhases and Total_expenses
0.7M
0.6M
0.5M
0.4M
0.3M
0.2M
0.1M
0.0M
0.01M
0.37M
0.02M
0.71M
0.01M
0.27M
2012
2013
2014
Year

This Graph shows Year wise Total Purchase and Total Expense by the customers
Depicts how much purchases done by the customers in a year and what will be the total cost for it.

Maximum purchases done in the year 2013 and thus maximum expenses was also in the same year but if we compare between year 2012 and 2014, in both the year Total Purchases are same but there is difference in total expenses which shows in 2012 more costly items were purchased as compared to 2014 which leads to the difference of 0.10 million in Total Expenses

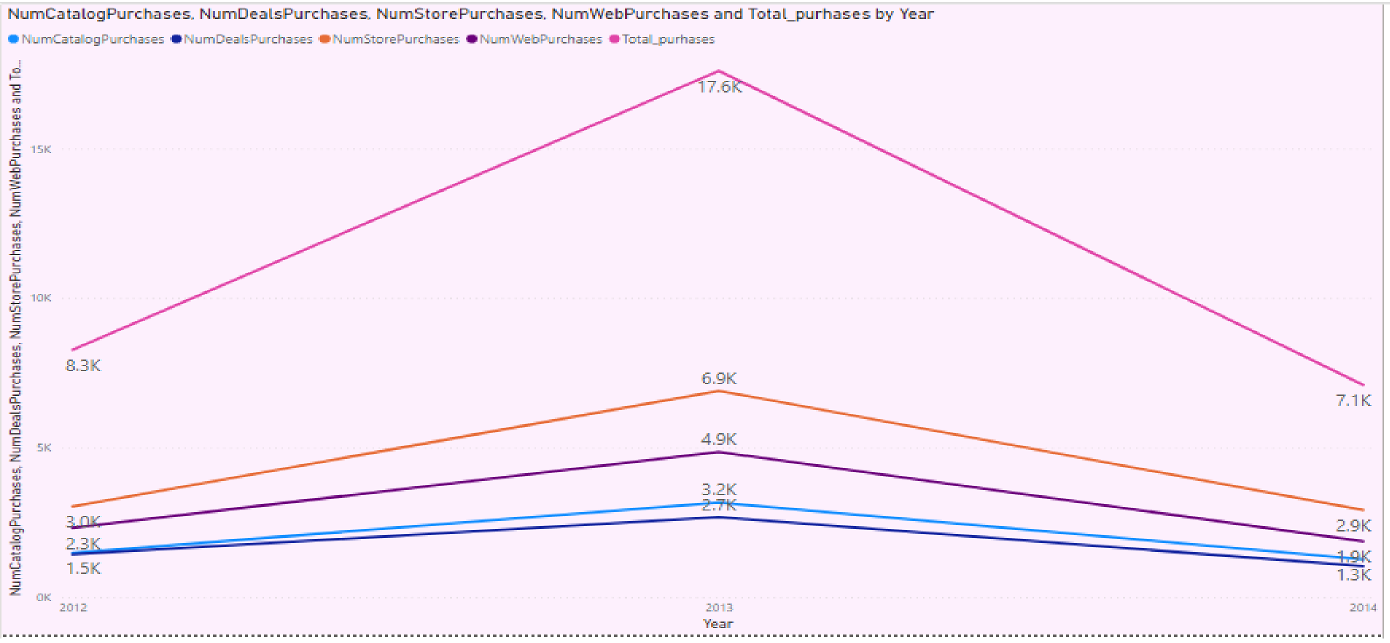
2. Year Wise Total Expenses By Different Products



From this it can be seen that major part of expense is done on **Fruits** only in all the years with respect to other products followed by **Meat** products

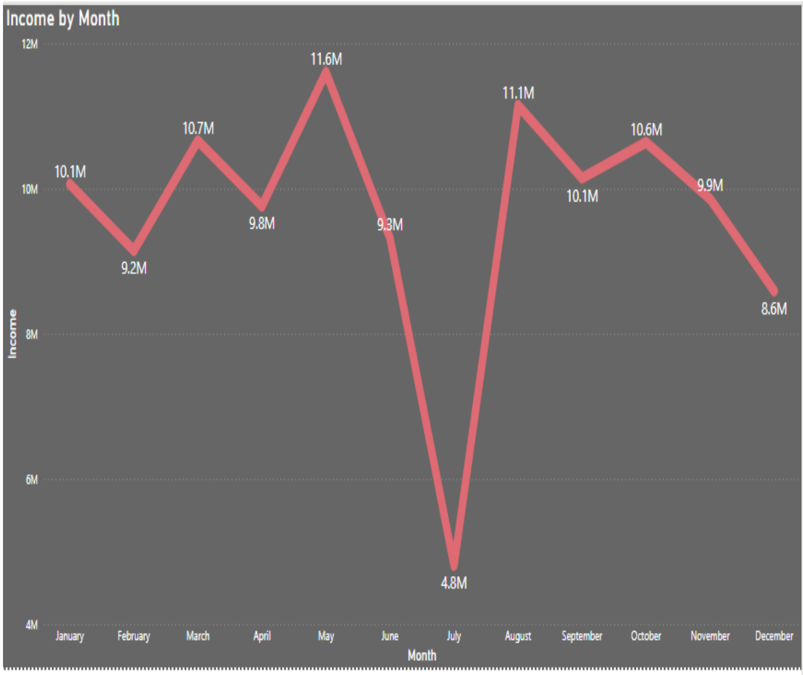
Thus company can analyse the condition why there is less demand or purchase of other products and make that items on sale or provide with offer so that there is increase in their sale also and make available more stock of fruit and meat products

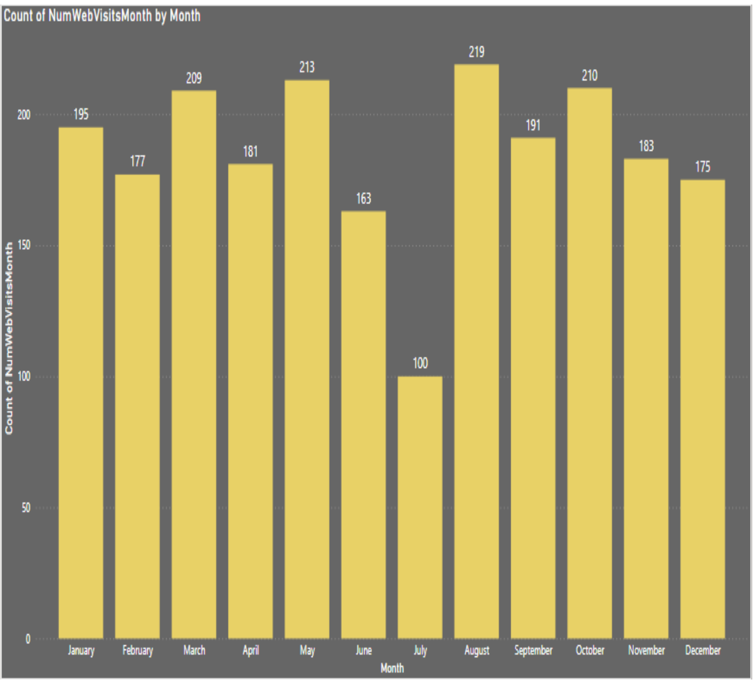
3. Year Wise Total Purchases By Different Channels



Majority of purchases are done from stores only followed by web purchases. Thus company should invest more resources towards these two channels of purchases so as to maximise their profit

4. Month wise Income of customer's and Website visitors

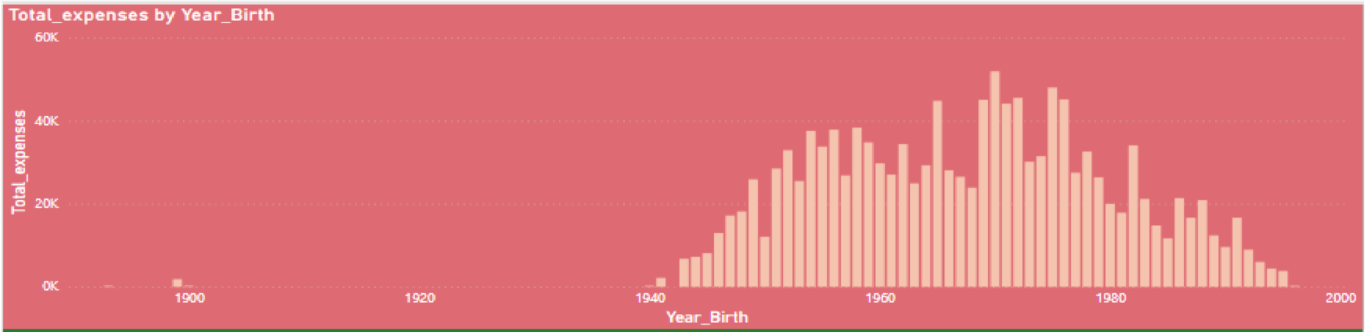


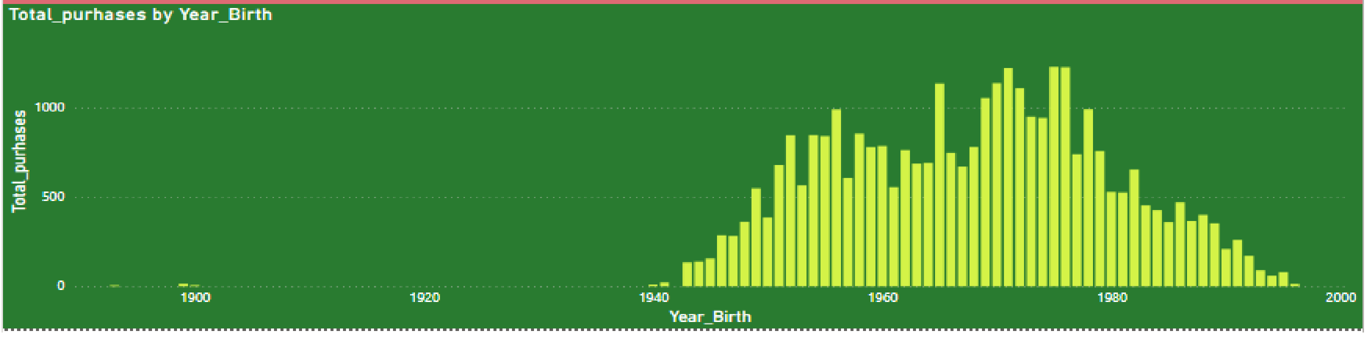


This shows the Good Positive relation between Income and Website Visitors as with the increase in income no of visitors also increases and vice-versa which depicts Income also an important aspect from company's point of view

As there is comparatively very less income in july thus leads to less number of visitors

5. Impact of Birth Year on Purchases and Expenses

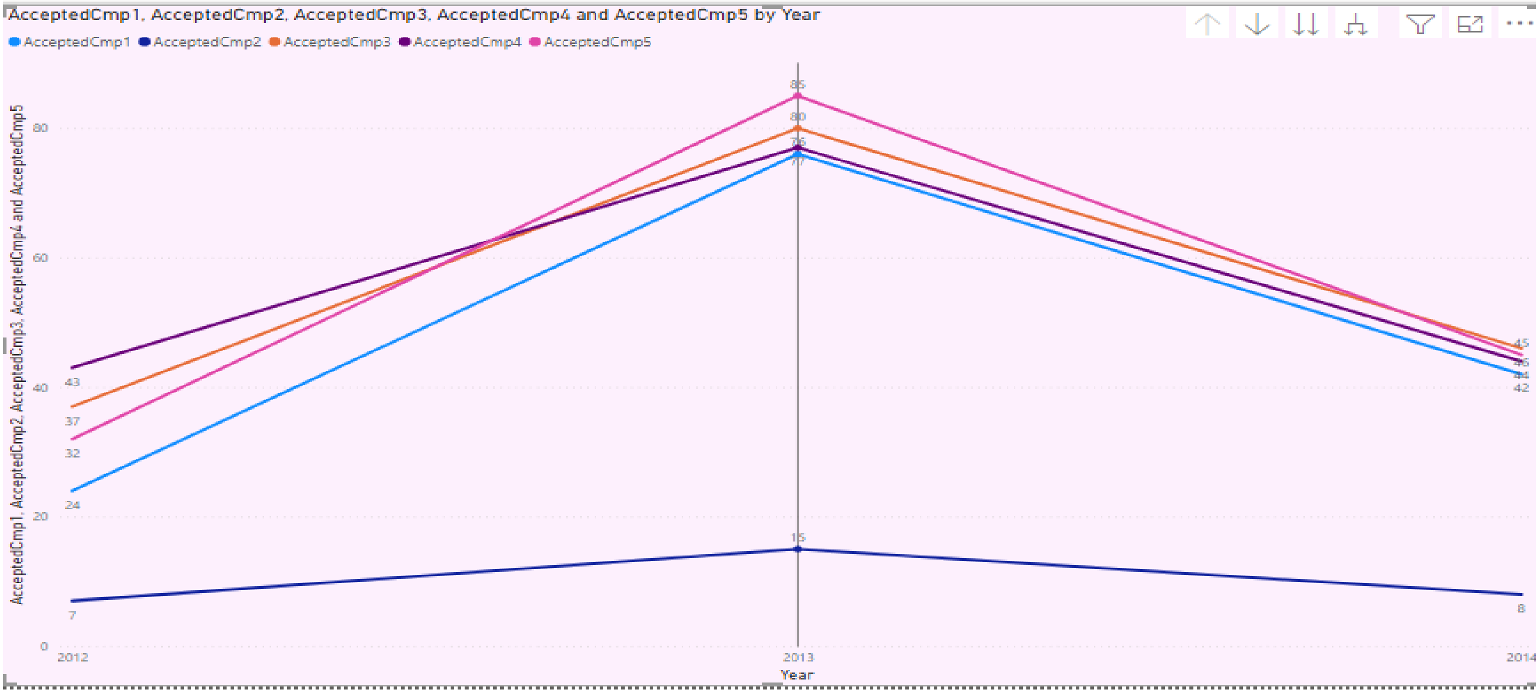




This graph depicts that the distribution of Total_expenses and Total_purchases seems to be normal

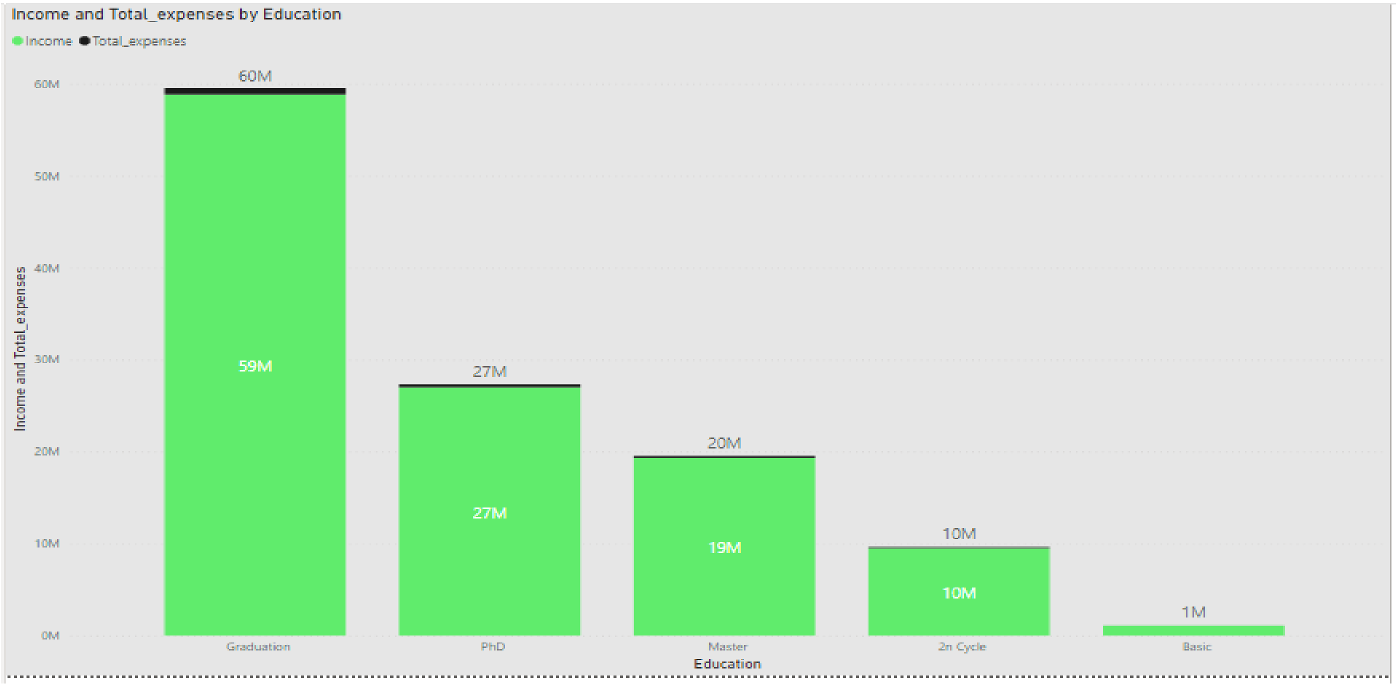
People with birth year from 1940-1980 people who born in 1940-1960 comparatively purchased less items then those who born between 1960-1980 , then after 1980 there is again decrease in Total_expenses and Total_purchases because the people of this age group are mostly dependent

6. *Offer Accepted with respect to No of Campaign*



As from Graph, we can see that Majority of offers accepted at last campaign Due to which the company can bear losses also . so , company should work on campaign part so that offers will be accepted at first campaign only

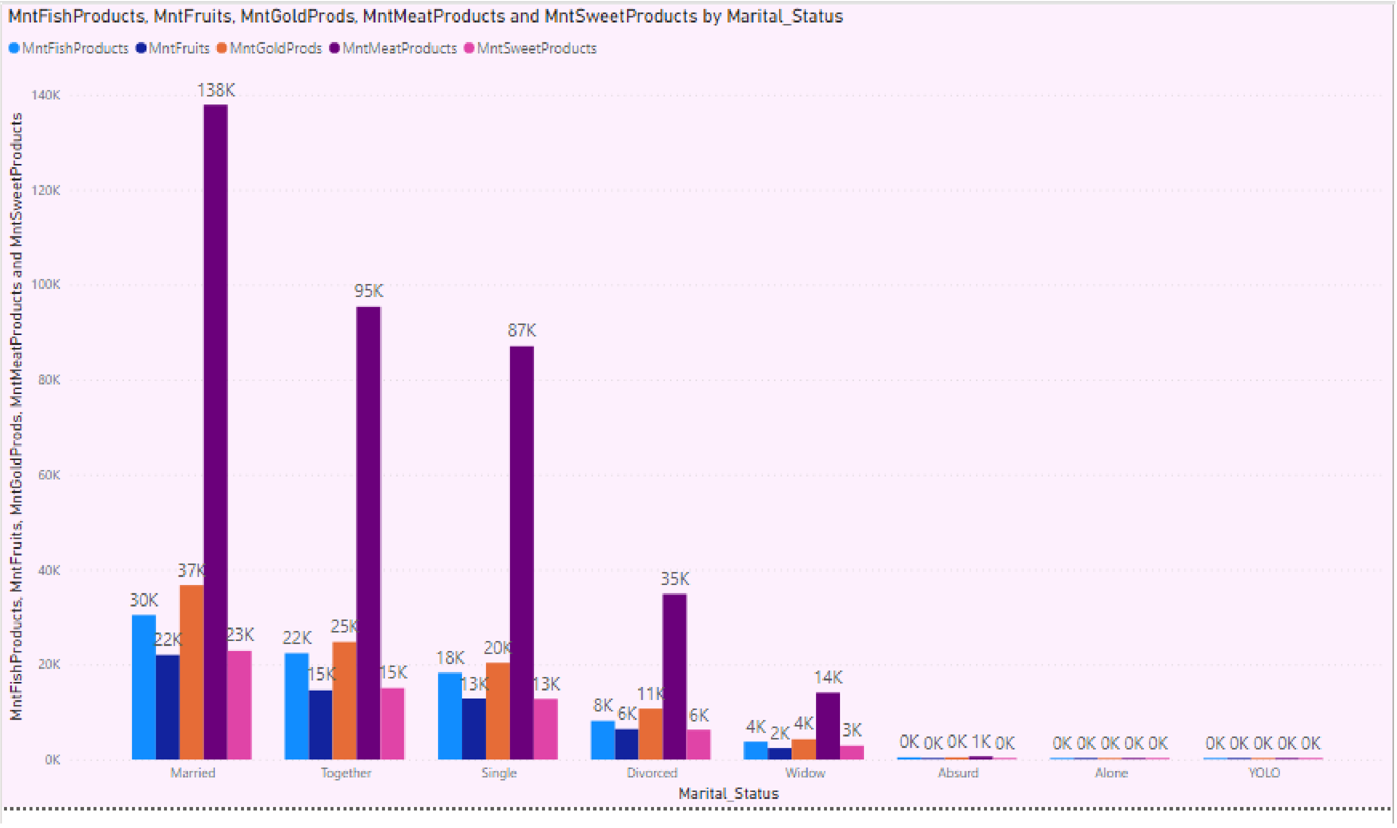
7. Impact of Education on Purchases and hence on Expenses



In terms of income if we see people with Graduation earns Total of 59M i.e with maximum among all the education group this may be because of number of people with graduation degree are way more than people with other degrees followed by Phd because of the prestigious degree they have , it definitely have Positive impact on their income i.e their income is more than other education group

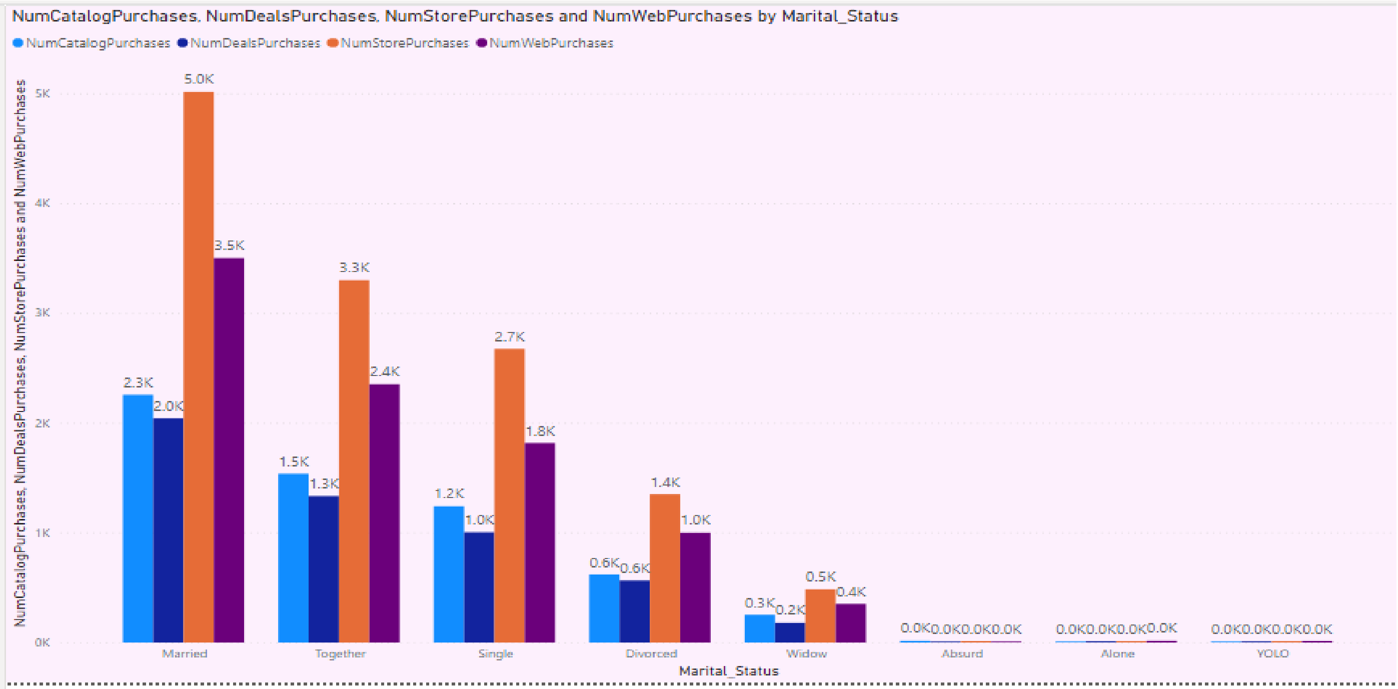
But in terms of expenses if we see in comparison to income their expenditure is very very less . so, company can organise a campaign in which they try to emphasize the customers so that they start purchasing things by knowing their area of interest

8. Impact of Marital Status on Purchasing Taste


MntFishProducts, MntFruits, MntGoldProds, MntMeatProducts and MntSweetProducts by Marital_Status

The graph clearly shows people who are married purchase more as compared to other category people and majority of items purchased are **fruit items .** so, company can open a Separate store for fruits only and try to connect more to people who are married as compared to other people

9. Impact of marital status on Purchasing Channel


NumCatalogPurchases, NumDealsPurchases, NumStorePurchases and NumWebPurchases by Marital_Status

From the graph and the previous ones, it can be seen that people with respect to different features usually prefers to buy from stores followed by Web purchases so company should prefer more on store buildings and website in order to increase the customers engagement and hence the profit

CONCLUSION:

Firstly company can see the behaviour of customers on the basis of cluster to which they belong and can design the campaign according to the cluster differently (as the people from different cluster behaves differently) so as the company can increase its profit in terms of engagement of customers

Also from regression technique we can predict the income of the person whose income is Missing as we saw income has its major impact on Total_Purchases

And from classification, since we classify the response of the customers , we can categories the people who response certainly or not and try to connect to those customers who didn’t respond frequently and can arrange a separate campaign for them and try to give them some offer’s based on their purchasing taste so that they will not leave us

GROUP MEMBERS:

ARUSHI KATHURIA
SHIVALI SHARMA