CS-626(1+2)

Data Mining & Warehousing

MCS Final Year (Evening)
Batch 2019
Department Of Computer Science,
University Of Karachi

Group Project

Title: Marketing Campaign

Course Supervisor: Dr.Tahseen Ahmed Jilani

Group Members	
Muhammad Hammad Hassan	EP20101030
Summaiya Muneer	EP20101051
Noor Sahar	EP20101044

Contents:

Key:

Green & Bold Text – Fields in our dataset

Light Blue, Bold & Italic – Values within Fields

Grey Highlighted Green – Package or module for Python

1- Introduction

1.1- Introduction to Dataset:

Original Source: The dataset for this project is provided by Dr. Omar Romero-Hernandez.
Source on Kaggle[0]: here
This dataset encompasses a marketing campaign movement initiated by an anonymous company. The Dataset includes details about:

- **People**; several details about the customer and whether they were interested or bought any of our packages or not among other details.
- **Product**; Types of products we were in charge of selling and promoting to our customers
- **Promotion**; Details such as types of promotion, number of deals purchased etc
- **Place**; Domains through which purchases or queries were being entertained

Further Details on fields are present in Section 3

1.2- Introduction to Working Activities Performed:

Our work was mainly focused on **clustering** and **categorization** of the customer base. For this purpose, we performed several tests to see which model best suits our interests. This includes both Supervised and Unsupervised models.

We had to perform some tweaking in order to shape our dataset and make it ready for proper data mining. This step became necessary since there were many potential noise and redundant dimensions, which needed to be taken care of before proceeding to our main task.

We also performed EDA and found out many useful insights which answered some of the important questions we had such as:

- What type of people constitute a big (if not huge) part of our customer base?
- Are we doing good in our campaign?
- If so, Then how good/bad is it?
- What type of people seem more interested in our offers?
- What type of people are more likely to reject our offers?
- What are points we should improve upon to get better results?

More details on our work is present in further sections 4 till 8 of our document

2- Literature Review

Since we made combined use of [1]Matthews Correlation Coefficient(as MCC) and Accuracy Score(as ACC) in our workings and more specifically in "Supervised Predictions " section of our notebook. This was done for the purpose of evaluating our models performance and also for comparison and selecting a best model.
We made use of sklearn.metrics to import both accuracy_score and matthews_corrcoef. We studied the importance and of it by referring to a towardsdatascience article on MCC as an alternative to F1 score.
Towards Data Science Inc. is a corporation registered in Canada.

Our activity in applying several models and comparing the results were presentable in our notebook due to a youtube channel named [2]CodeBasics. This is a Youtube channel owned by Dhaval Patel, a Data Scientist by occupation who teaches us how to use python and other tools for Data Mining and Data Science. His work encompases many topics in Data Science and related software and tools. We learned several techniques from it such as train_test_split, KFold.cross_val_score, confusion_matrix etc

[3]**Kaggle courses** helped in the data cleaning and reduction tasks as we studied the Data Cleaning course which helped in our cleaning and shaping the dataset we were going to use.
We also studied some notebooks in Kaggle as well to better understand how to proceed with our work.

We also referred to some of [4]**Analytics Vidhya** free courses to help us in our data exploration tasks. We made use of this knowledge in our Univariate, Bivariate and Multivariate analysis. This helped us study the variables more closely and understand the data a bit more clearly.

Lastly we referred to [5]**sklearn's and** [6]**Pandas' own documentation** for proper use of its library and to better understand the parameters provided in each function as well as how best can we make use of those models.

3- Dataset Details (Marketing Campaign)

Following are the description of each column[0]:

Table 3.1 – Data Description List

Fields	Description	Datatype
People		
ID	Customers Unique Identifier	int
Year_Birth	Customer's birth year	Int
Education	Customer's education level	String
Marital_Status	Customer's marital status	String
Income	Customer's yearly household income	Float
Kidhome	Number of children in customer's household	Int
Teenhome	Number of teenagers in customer's household	Int
Dt_Customer	Date of customer's enrollment with the company	String
Recency	Number of days since customer's last purchase	Int
Complain	1 if the customer complained in the last 2 years, 0 otherwise	Int
Products		
MntWines	Amount spent on wine in last 2 years	Int
MntFruits	Amount spent on fruits in last 2 years	Int
MntMeatProducts	Amount spent on meat in last 2 years	Int
MntFishProducts	Amount spent on fish in last 2 years	Int
MntSweetProducts	Amount spent on sweets in last 2 years	Int
MntGoldProds	Amount spent on gold in last 2 years	Int
Promotion		
NumDealsPurchases	Number of purchases made with a discount	Int
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise	Int
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise	Int
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise	Int
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise	Int
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise	Int
Response	1 if customer accepted the offer in the last campaign, 0 otherwise	Int
Place		
NumWebPurchases	Number of purchases made through the company’s website	Int
NumCatalogPurchases	Number of purchases made using a catalogue	Int
NumStorePurchases	Number of purchases made directly in stores	Int
NumWebVisitsMonth	Number of visits to company’s website in the last month	Int

4- Data Overview & Preprocessing

In this section we are analyzing what our data looks like and checking if there are any inconsistencies in our dataset. Upon finding such inconsistencies, we will attempt to either purge them or adjust them according some normalization techniques.

In the raw form (the initial state of dataset when we first got our hand on it) our dataset had 29 different fields and 2240 entries for each Customer. There were 25 Integer Values, 3 strings datatypes and only 1 float value present. It was evident that we mostly had integer values to work with.

4.1- Checking for Null Values

We only found 24 Null Values in the **Income** Column. Since the Null values were only in the small quantities, therefore we filled the empty spaces with average values.

This inconsistency may be due to people being uncomfortable with sharing their income status. This is common among people as society tends to judge people who have low income.
But instead of assuming low wages for those missing entries we will go with common method for normalization i.e. take average of the whole column and fill in the Null values within that column

4.2- Checking for Duplicate Values

Our Dataset didn't have any duplicates, therefore no further explanation is needed here.

4.3- Checking for Unique Values

Columns **Z_revenue** and **Z_costContact** had only 1 unique value, therefore we simply dropped it. It wasn't going to affect any response variable since it was only a constant at that point.

ID and **Dt_Customer** were of no importance in our calculations, therefore we dropped them as well.

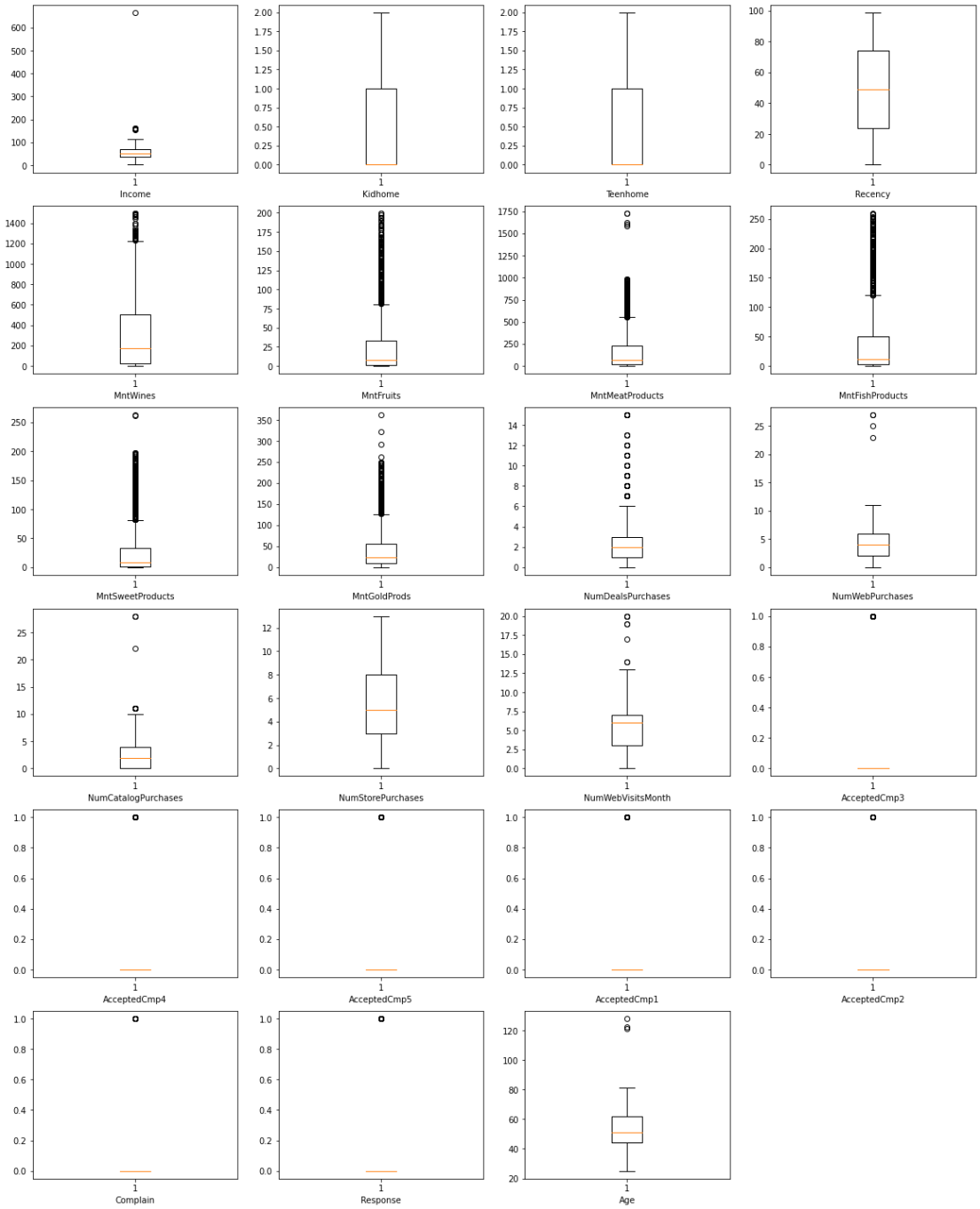
We also needed to recategorize the **Marital_Status** column since ***Alone***, ***YOLO*** and ***Absurd*** can be easily grouped into ***Single*** category

We also recategorized the **Education** column. We simply merged ***2n_Cycle*** into ***Masters*** and ***Graduation*** into ***Bachelors***.

4.3- Checking for Outliers

We used the box plot to identify the outliers in each column/field. Since we were sure about outliers in **Age** and **Income** pose unnecessary plots, therefore we can remove them. As for other fields, we were not sure, therefore we will let them be.

Fig 4.3.1 – *Box whiskers plot to detect outliers in each column*



4.4- Final Preprocessed Data

After preprocessing our dataset had 25 columns but the number of records didn't change since did not need to remove the Null entries. Therefore all 2240 entries were intact, except for **Income** column which had Null values and were filled with Average of the whole column

5- Exploratory Data Analysis

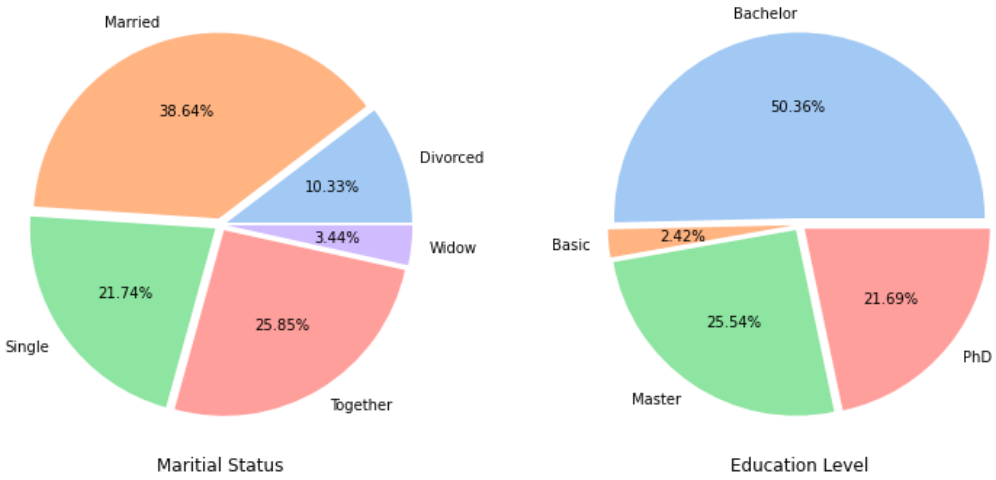
Here we are just trying to visualize the dataset using histograms and charts to better understand how the values within each field is distributed.

5.1- Univariate Analysis

By constructing a Pie Chart for the **Marital_Status** and Education fields we can conclude that more ***Married*** couples, less people with ***Together*** and ***Single*** status and very scarce with ***Widow*** and ***Divorced*** status were encountered during our Campaign as shown by fig 5.1.1

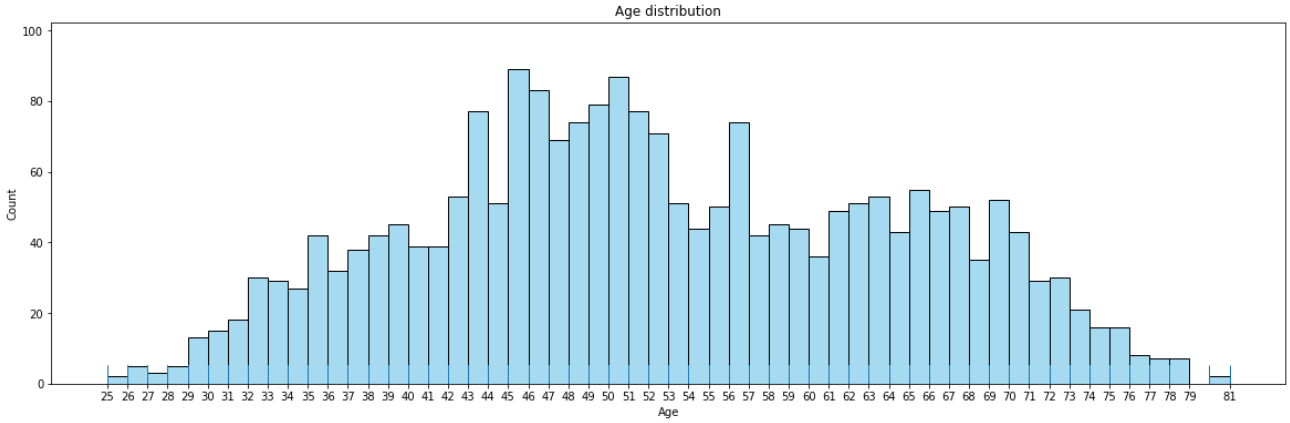
Whereas, People holding ***Bachelor*** degree constituted half of the total population we encountered in our campaign. People with Basic education level were scarcely encountered.

Fig 5.1.1 – *Pie Chart of Marital_Status and Education Level of Customers*



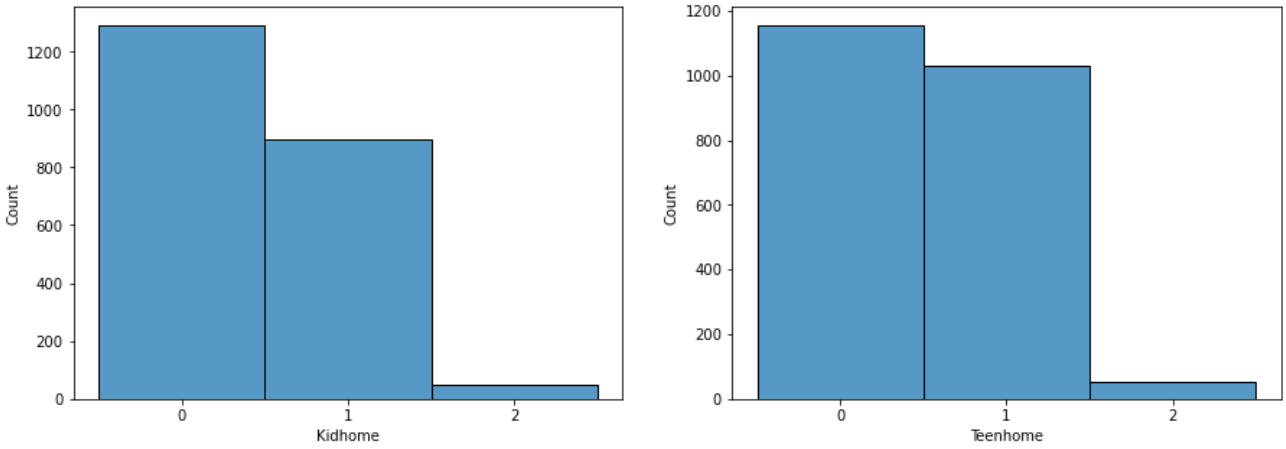
By visualizing a histogram of the **Age** field, we saw that most of our customers were people having age of 45 to 60 years.

Fig 5.1.2 – Histogram of Age field



We also plotted a bar plot for **Kidhome** and **Teenhome** separately. We found out that most customers had no kids or no teens at home. Very few people were seen to have more than one kid or teen.

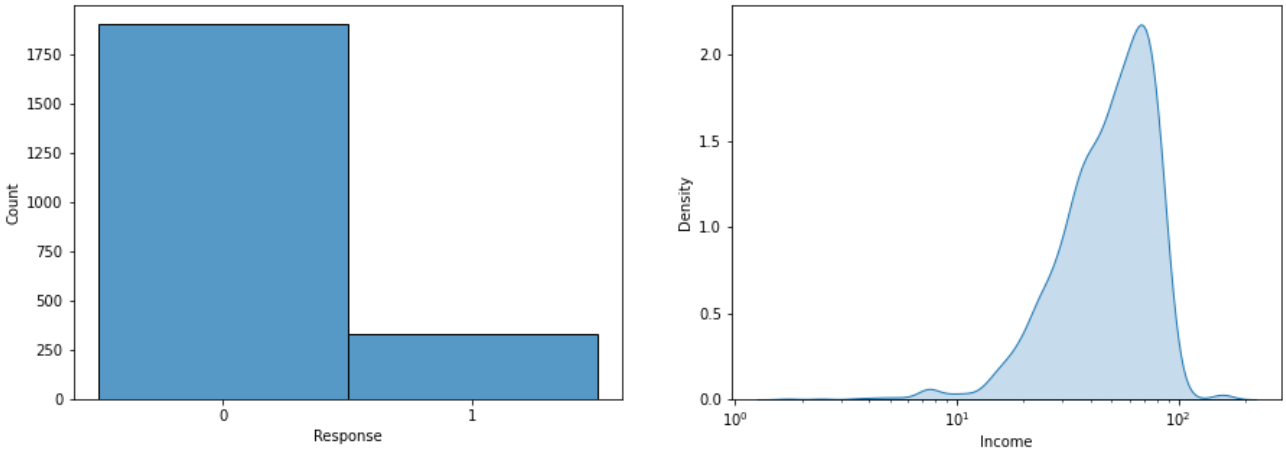
Fig 5.1.3 – *Kidhome and Teenhome bar plot*



By observing the **Income** status we found that most people had an income range of 10,000 to 100,000 units (preferably dollar). Other income categories were very scarce.

The **Responses** of people were not in our favor since above 1700 responses notified rejection of offers and less than 350 responses said that they are interested.

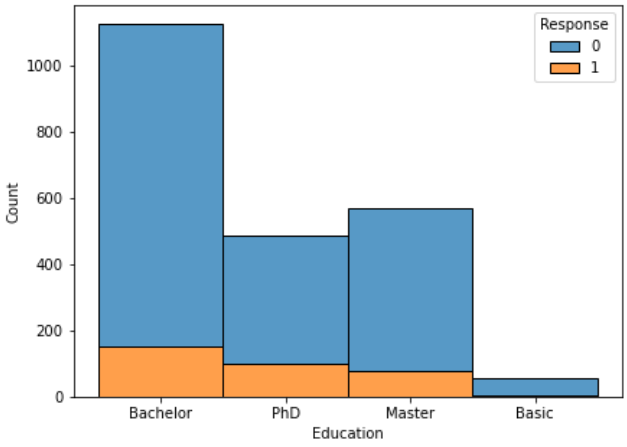
Fig 5.1.4 – *Income and Responses plots*

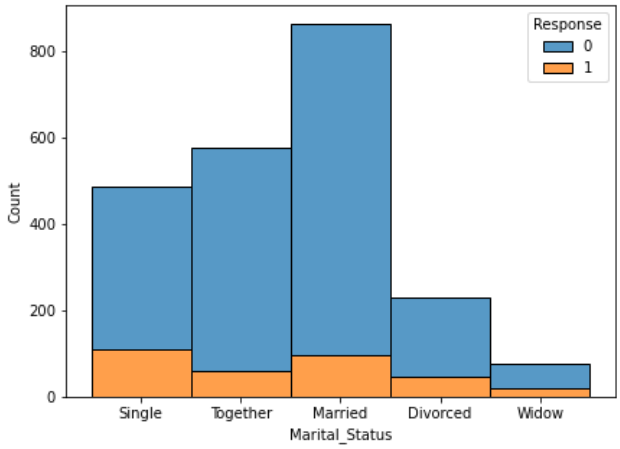


5.2- Bivariate Analysis

We plotted barplots to compare **Responses** with **Education** level and **Marital_Status** and found out that mostly ***Bachelors*** were encountered but less than a quarter of their population responded positively. Whereas ***Married*** couples were largely encountered but ***Single*** people responded positively on a higher rate.

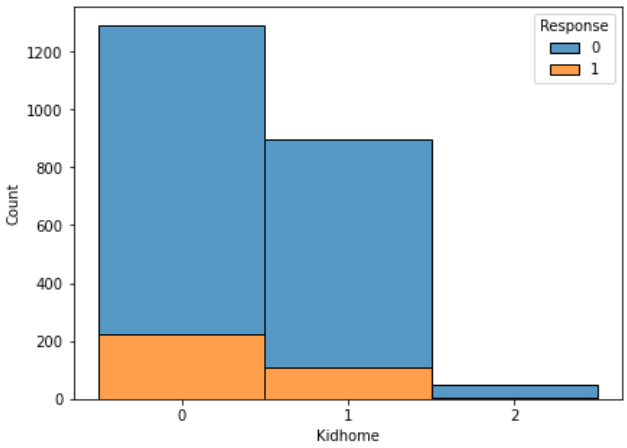
Fig 5.2.1 – *Education and Marital_status comparison with Responses*

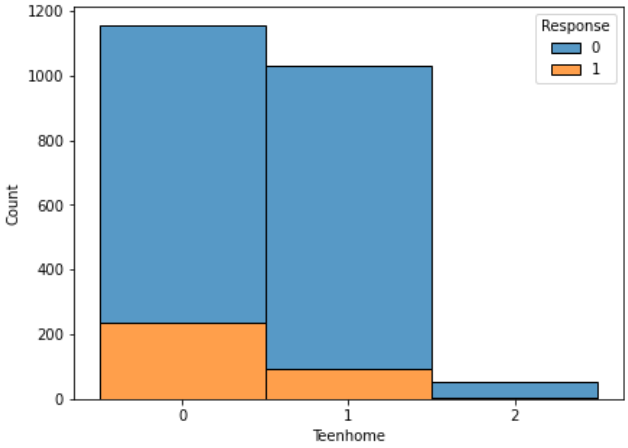




Comparing **Responses** with **Kidhome** and **Teenhome** we can see that people having no kids or teens at home were more active with positive responses than the people with one or more kids or teens.

Fig 5.2.2 – *Teenhome and Kidhome comparison with Responses*

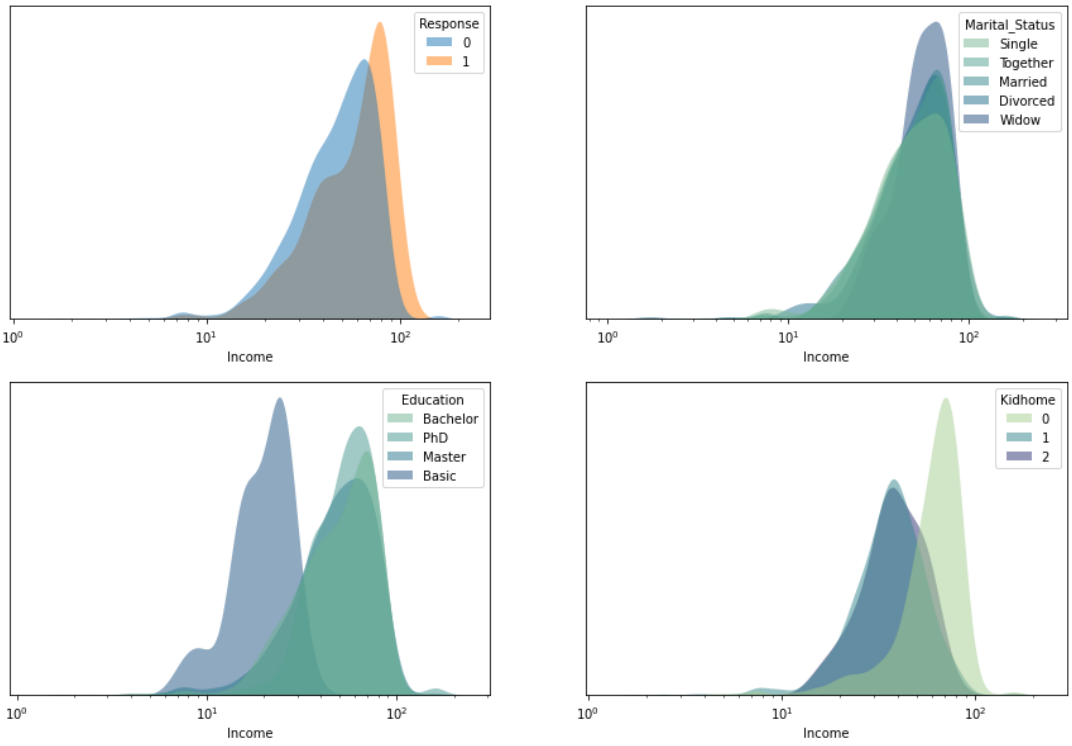




Observing the effects of **Income** on **Marital_status**, **Responses**, **Education** and **Kidhome** fields, we concluded the following:

- People with more Income are more likely to accept our offers. We can say that people with income 14-15,000 or less don't seem that much interested.
- Different Marital Status does not seem to be the cause of positive or negative response to our marketing campaign.
- People having lower level of education have less income. Those Bachelors, Masters or Ph.D degrees do not have clear difference between their incomes.
- Those who do not have kids at home have higher income.

Fig 5.2.3 – *Income vs Marital_Status, Responses, Education, Kidhome*



5.3- Multivariate Analysis

Here we tried to find a relationship between all fields using a heat map. Unfortunately, we didn't find any high correlation among any field.

Fig 5.3.1 – *Heatmap of all fields*

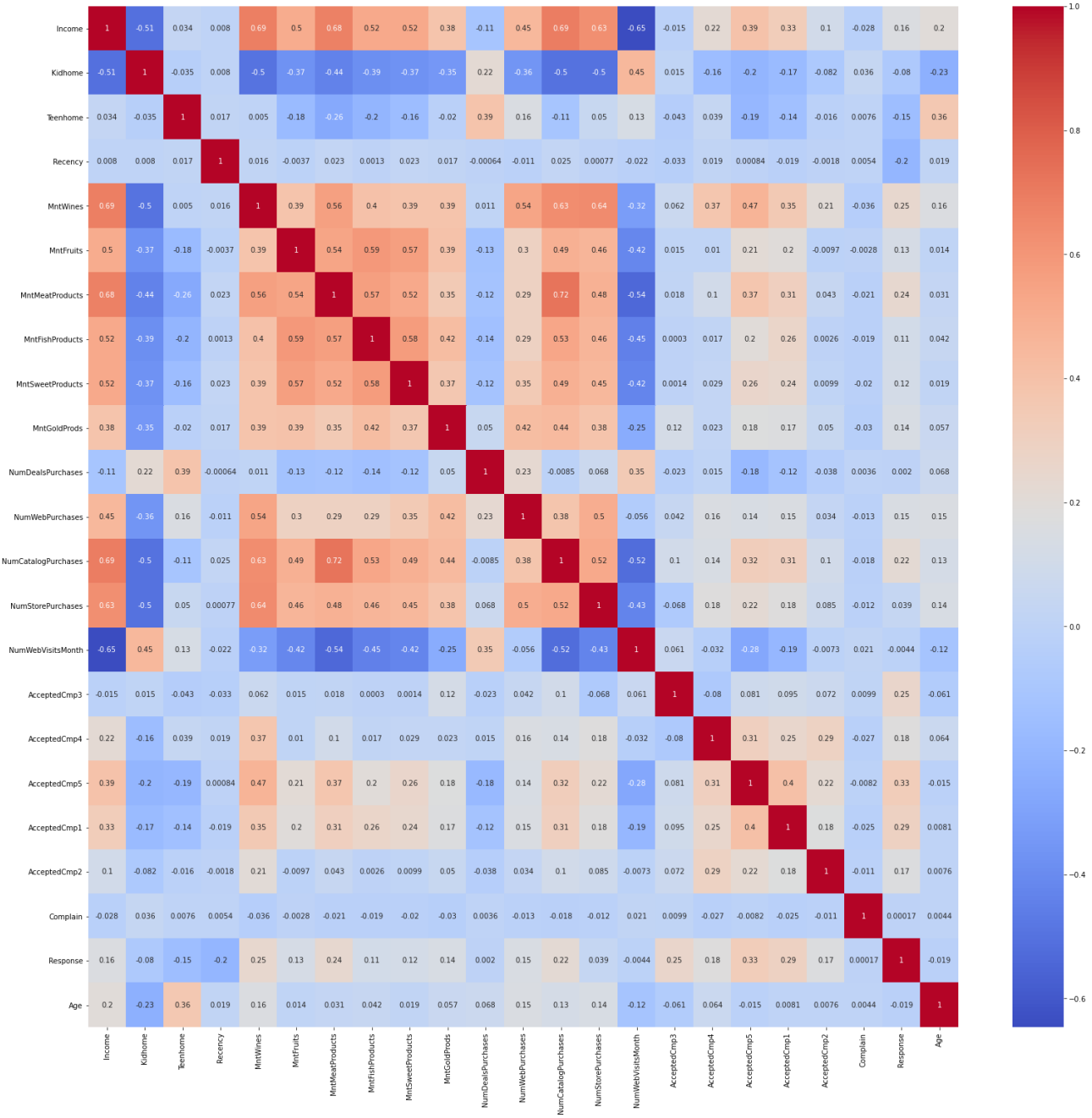
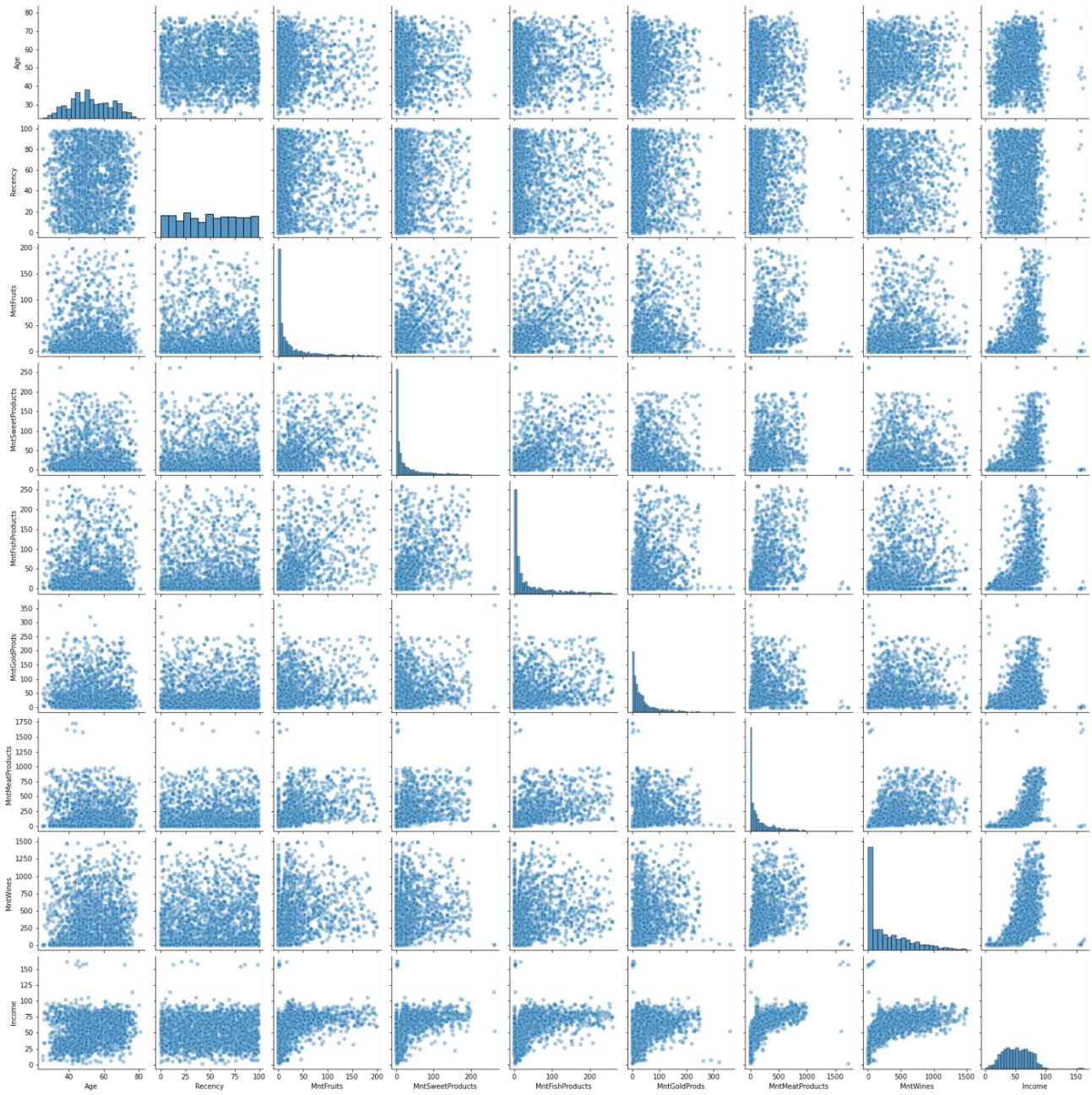


Fig 5.3.2 – *Pair plot of all fields*



6- Feature Selection And Dimensionality Reduction

6.1- Preparing Sample for Prediction Models

First we converted the **Marital_Status** field into machine readable form by using Dummy Encoding. The encoding was done as follows:

- Marital_Status_Divorced
- Marital_Status_Married
- Marital_Status_Single
- Marital_Status_Together
- Marital_Status_Widow

Also, the encoding was applied on all other String Values datatypes, i.e. **Education**:

- Basic 0
- Bachelors 1
- Masters 2
- PhD 3

This constitutes our **Raw Dataset**

6.2- Feature Selection By Random Forest

Random Forest falls under the category of **Supervised Learning Model** that implements both **decision trees** and **bagging** method. But it has another function as a **Feature Selector** which utilizes a process called **Bootstrap**.

The way how this works is that each sample in Random Forest iteration contains a random number(subset) of raw columns and is employed to fit it to a decision tree. There is a separate decision tree for each iteration. Therefore, in this scenario the number of decision tree models which employees some numbers of columns/features are considered as **Hyperparameters** which are to be optimized. In the last step of this feature selection method, all the results/prediction of decision tree model are mixed together and are either utilized in calculation of the regression mean (y bar) or uses soft voting for classification.

Here we are using the [5]Random Forest algorithm (obtained from sklearn library) and feature_importances_ attribute. Turns out that to top 5 important features in our dataset Fig 6.2.1 are:

1. Recency
2. MntWines
3. Income
4. MntMeatProduct
5. MntGoldProducts

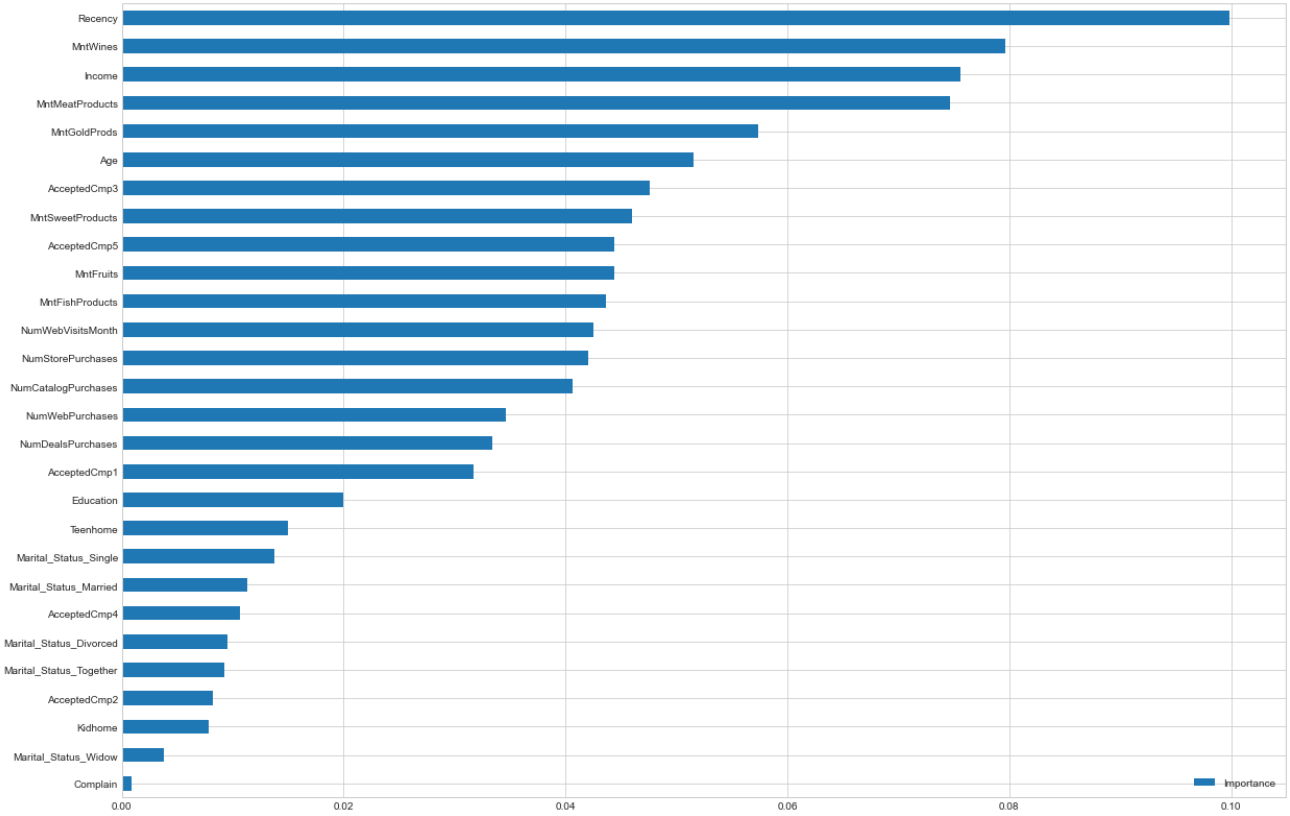
We decided to keep 90% of the importance ratio which included the following (total 18) fields:

Recency	MntFruits
MntWines	MntFishProducts
Income	NumWebVisitsMonth
MntMeatProducts	NumStorePurchases
MntGoldProds	NumCatalogPurchases
Age	NumWebPurchases
AcceptedCmp3	NumDealsPurchases
MntSweetProducts	AcceptedCmp1
AcceptedCmp5	Education

(Refer to section 3 of this document for details on these fields.)

The total importance ratio of these fields is calculated to be **0.9097738142686715 %**

Fig 6.2.1 – *Feature Importance plot by Random Forest Classifier*



This constitutes our **Feature Selected Dataset.**

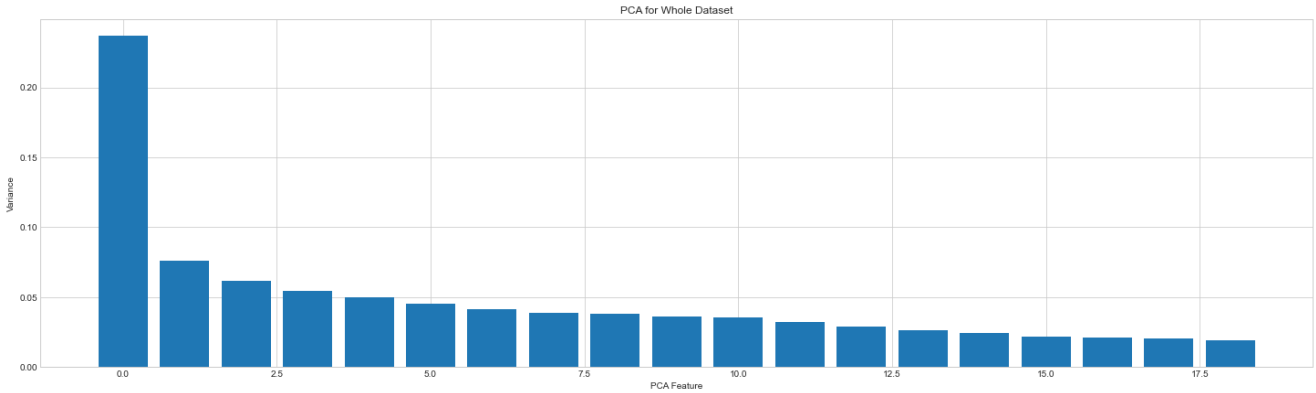
6.3- PCA Transformation

PCA is well known as an orthogonal linear transformation of features which enables us the visualize features in such a way that we can easily detect which feature contributes the most to the spread of the dataset(variance). In this method, the feature having the greatest contributions to the variance

comes out as the first coordinate and is better known as **Principal Component 1**. The feature having second largest contribution to the variance is determined during the second iteration and is termed as **Principal Component 2**. This process goes on until all of the features have been tested.

Using PCA we have graphed the variance for each feature in our dataset. Implementing what we observed in section 6.2 we can now see that our dataset has 19 columns in total and 2236 entries.

Fig 6.3.1 – *variance of each feature*



This constitutes our **PCA Dataset**

7- Supervised Predictions

7.1- Preparing Sample for Classification Models

We used SMOTE here since we are going to create a model that predict Response classification based on other Customer data. But the problem here is that our responses are in an imbalanced ratio i.e there are many negative responses and too less positive ones. This is not a good case for our models.

[7]The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on, the minority class, although typically it is performance on the minority class that is most important.

By using SMOTE balanced our data and made it doable for our models which we are going to implement.

We further split our dataset into two sets i.e. training and testing dataset and tested it with SMOTE. Before resampling with SMOTE we had 2000 entries in training dataset which is {0: 1698, 1: 302}. This too imbalanced. After resampling with SMOTE we can see equal level of each set of responses {0: 1698, 1: 1698}.

Here is the summary:

```
Before SMOTE
Train:  2000
Test:  236
```

```
N/P Sample:  {0: 1698, 1: 302}

After SMOTE
Train:  3396
Test:  236
N/P Sample:  {1: 1698, 0: 1698}
```

In all implemented models ahead we have used Matthews Correlation Coefficient (MCC). [1]The Matthews correlation coefficient (MCC), is a more reliable (compared to F1) statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.

For each model we have set the Kfold = 8 and we used two scoring methods consecutively ACC and MCC. We used a cross_val_score to determine the scores with both ACC and MCC.

7.2- Logistic Regression

By using Logistic Regression model the results were as follows:

> **ACC:** 0.749420 mean (0.015291 std)
> **MCC:** 0.499386 mean (0.030593 std)

7.3- Boosting Tree

By using Boosting Tree model the results were as follows:

> **ACC:** 0.903476 mean (0.098889 std)
> **MCC:** 0.819825 mean (0.173745 std)

7.4- Support Vector Machine (SVM)

By using SVM model the results were as follows:

> **ACC:** 0.80215 mean (0.023620 std)
> **MCC:** 0.644576 mean (0.039894 std)

7.5- Neural Networks

Using this model required the use of models and layers from keras. Our neural network trained for 8 folds(as we had already set).
By using Neural Network model the results were as follows:

> **ACC:** 0.758902 mean (0.077483 std)
> **MCC:** 0.557649 mean (0.124293 std)

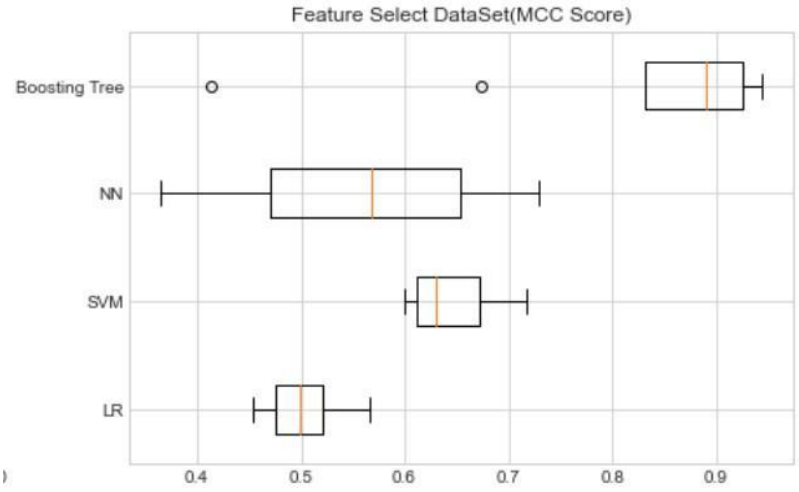
7.6- Performance Comparison Among all 4 Models

Comparing the results of all them applied models we can see that Boosting Tree has the most ACC and MCC(in all three datasets.

Boosting Tree has some outliers in "Feature-Selected Dataset" and "Raw Dataset". This shows that this model might not work well in these datasets. Boosting tree may have outliers but that is comparable with other models' performances

In conclusion we are using Boosting Tree dataset + Feature-Selected dataset to achieve the best MCC score.

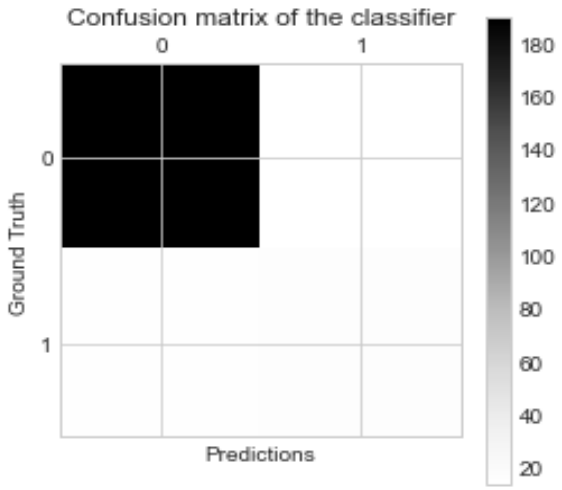
Fig 7.6.1 – *all models working in Feature Selected Dataset*



7.7- Final Model Performance

Finally we can draw a score report using confusion matrix:

Fig 7.7.1 – *Confusion Matrix*




```
Train MCC: 0.9894020385748943
Test MCC:  0.46887717596730055
Test ACC:  0.8771186440677966

              precision    recall  f1-score   support

           0       0.93      0.93      0.93       204
           1       0.55      0.53      0.54        32

    accuracy                           0.88       236
   macro avg       0.74      0.73      0.73       236
weighted avg       0.88      0.88      0.88       236
```

The overall test accuracy of the model is 0.88. But dive deep into the score report, the model performs quite good in recognizing negative samples(0), but not good in positive samples (precision: 0.55, recall: 0.55).

The test MCC is 0.469, which indicates that the model may not good at finding positive samples in test set. While we find the Train MCC is 0.98, this result shows there might exist overfitting problem in the model. But the fact is even I tried a lot to simplify the model and the train MCC decreases a lot, the test mcc still can not show much improvements. This result might indicate the predictors we use in this dataset might not predict 'Response' very well.

8- Unsupervised Learning

8.1- Feature Engineering And Clustering

For **Marital_Status**, our aim to classify them as in ***relationship*** or ***single***, therefore we will engineer those values into proper representations.

```
Relationship    1442
Single           794
```

As for **Kidhome** and **Teenhome**, we will combine them into a new category called **Children**.

We will further add a new column called **MntTotal**, to represent the total purchasing amount of all products for all column names starting with “Mnt”.

We will add **NumTotal** to represent total purchasing number for different purchasing types for all column whose name contain “Purchases” in it.

We don’t care about which campaign the customer will participate in, instead we care about the total participation times. Therefore, we will introduce a new column called **TotalAccepted**, instead of **AcceptedCmp1**, **AcceptedCmp2**, **AcceptedCmp3**, **AcceptedCmp4**, **AcceptedCmp5**.

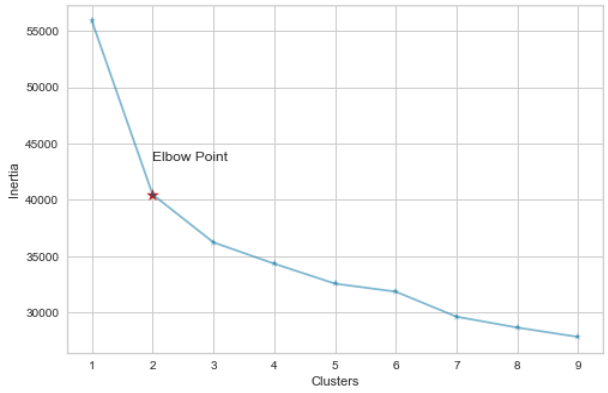
Finally, we used LabelEncoder to encode **Eduation** and **Marital_Status** into respective digits.

8.2- K-Means

K-means is a **clustering algorithm** which falls under the category of **unsupervised learning** ML models. It finds groups in a scatter plot of data point which are not labeled. This is best for detecting hidden patterns among the dataset which are not easily seen during normal statistical analysis of complex dataset.

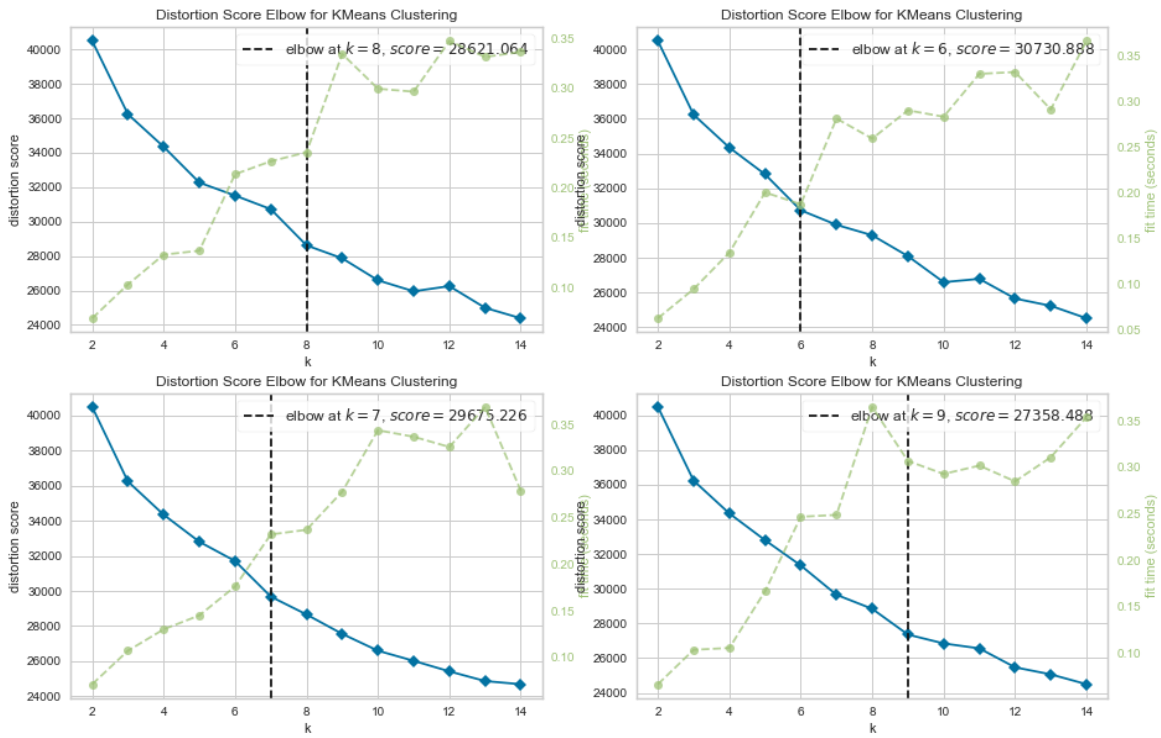
For K-Means we used KElbowVisualizer from yellowbrick to determine the number of clusters we should aim for. The Elbow value was found to be 2

Fig 8.2.1 – *Elbow plot*



We then performed K-Means clustering 4 times with different initial clusters and plotted graph for them, which is as follows:

Fig 8.2.2 – *K-Means Distortion score elbow*



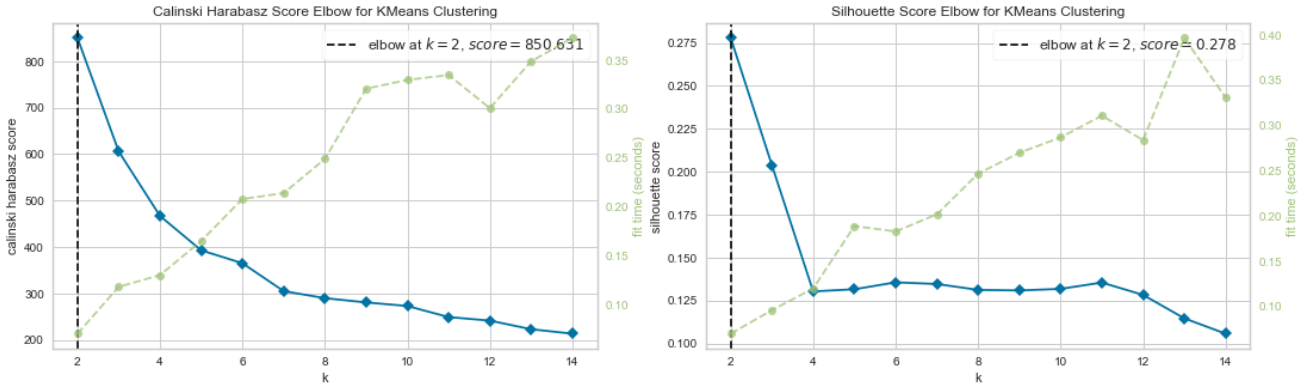
We selected the scoring parameter as “distortion”. This makes calculates the sum of squared distances from each point to it center(set by current iteration)

Here the Silhouette Score determines the Silhouette Coefficient of all samples.

Whereas, the Calinski Harabasz Score determines the ratio of dispersion between and within clusters.

I also use these 2 metrics to get clusters.

Fig 8.2.3 – *Calinski Harabasz Score vs Sillhouette Score*



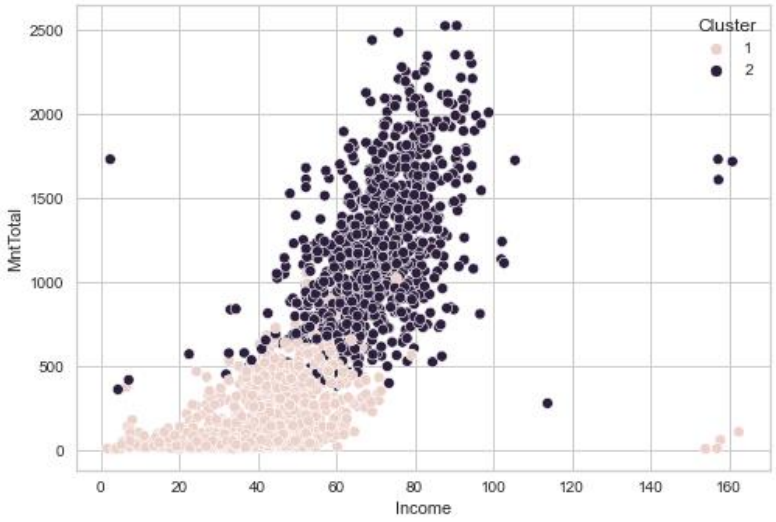
We set random_state variable and by using model.inertia_ we found and Elbow value of 2. This was verified by the Calinski Harabasz and Silhouette Scorers.

Considering all the clutering results we reject the notion of Distortion Score(K=7), and we finally choose K=2 to cluster the customers.

We then fitted the K-Means model with k=2, and inspect the value counts of clusters to find:

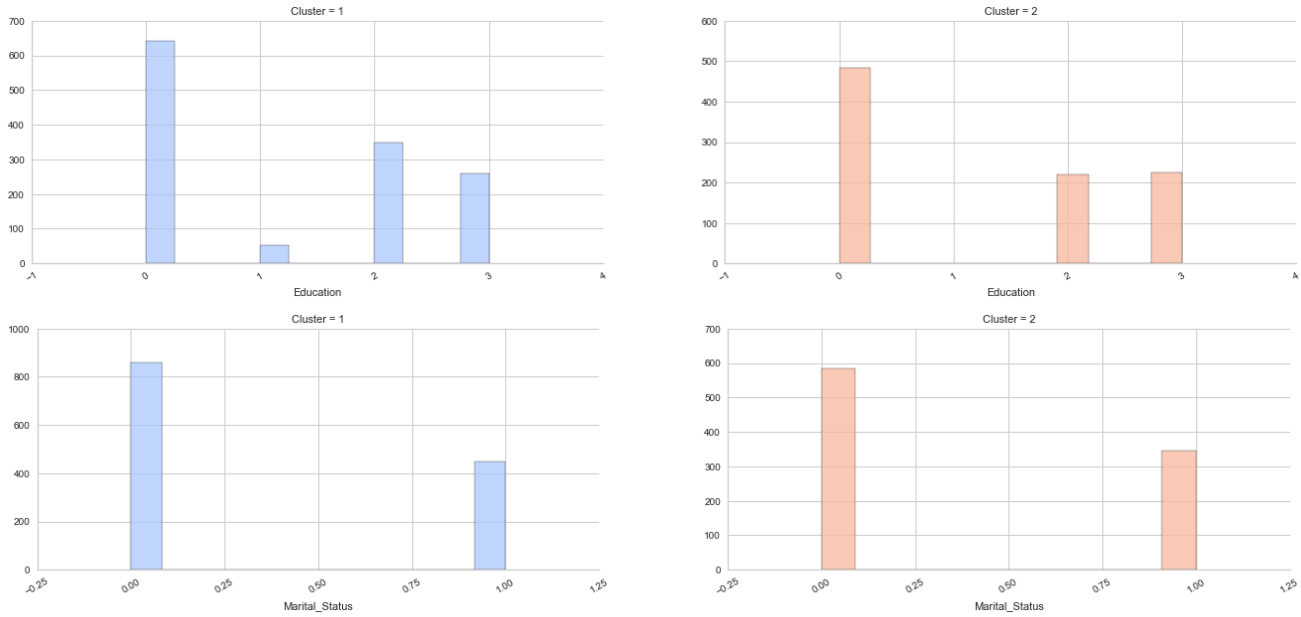
1	1306
2	930

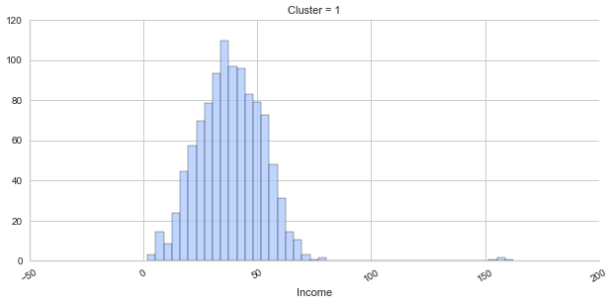
Fig 8.2.4 – *plotting a scatter plot for MntTotal and Income*

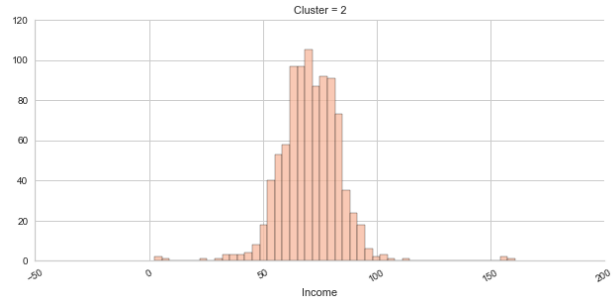


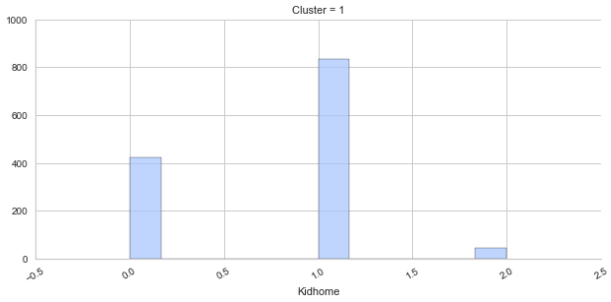
Now we will plot bar chart for each column separated by difference in clusters.

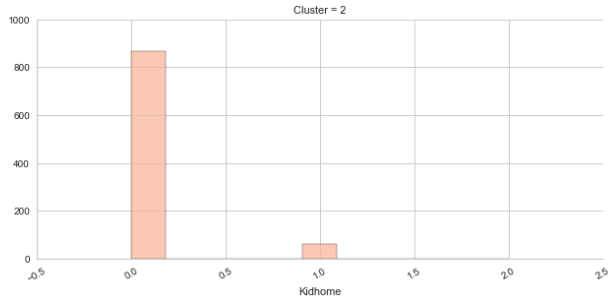
Fig 8.2.5 – *bar chart for each column w.r.t clusters*

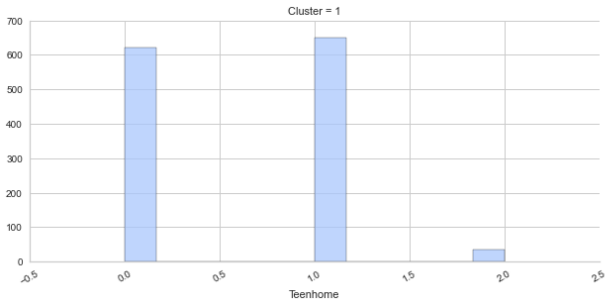


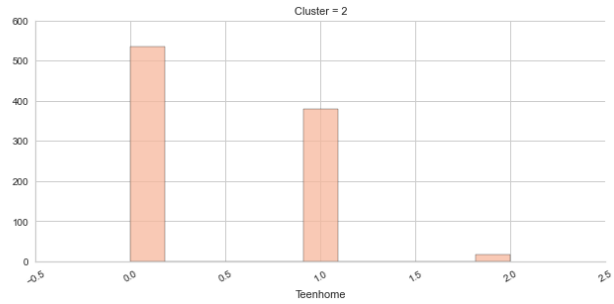

Cluster = 1
Income

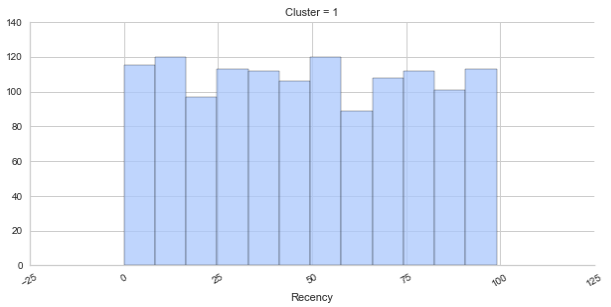

Cluster = 2
Income

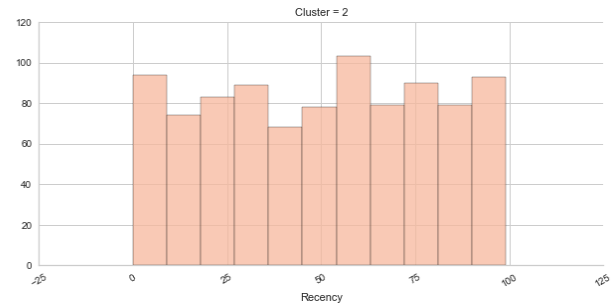

Cluster = 1
Kidhome

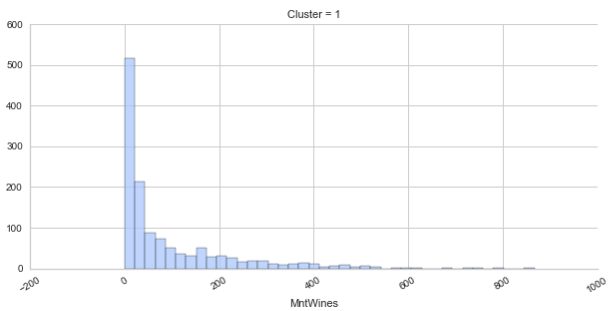

Cluster = 2
Kidhome

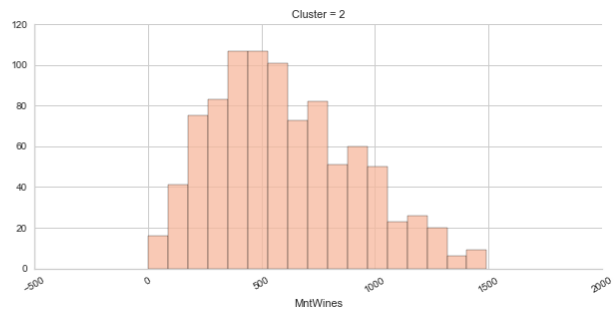

Cluster = 1
Teenhome

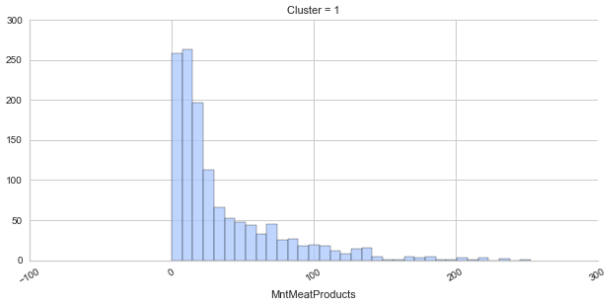

Cluster = 2
Teenhome

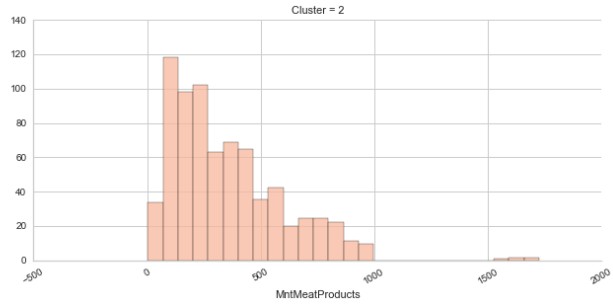

Cluster = 1
Recency

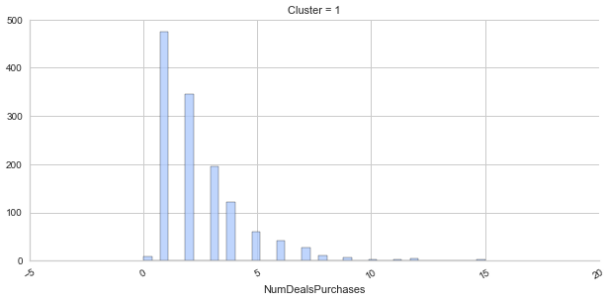

Cluster = 2
Recency

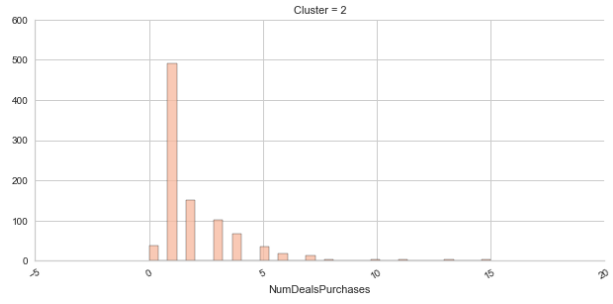

Cluster = 1
MntWines

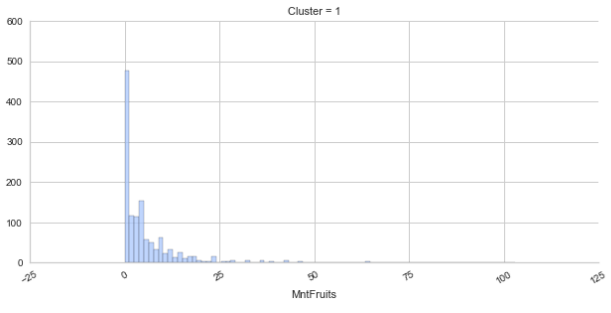

Cluster = 2
MntWines

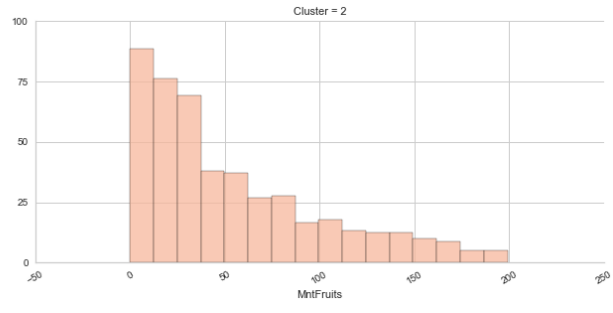

Cluster = 1
MntMeatProducts

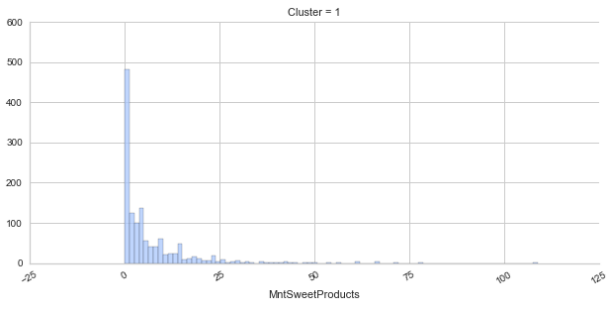

Cluster = 2
MntMeatProducts

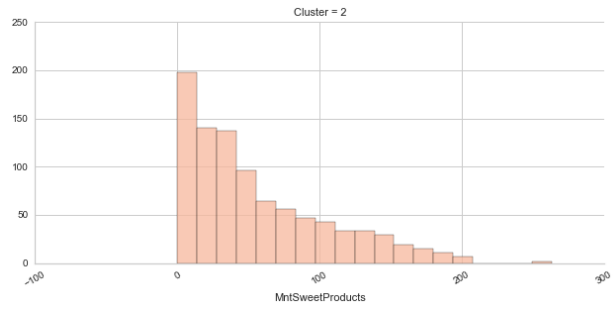

Cluster = 1
NumDealsPurchases

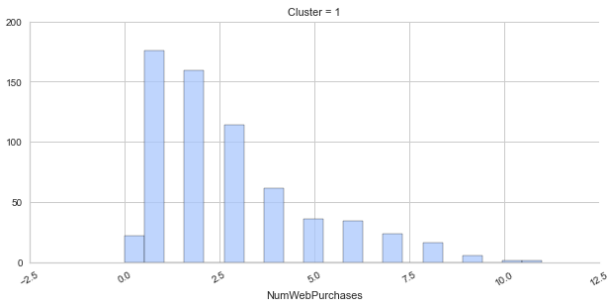

Cluster = 2
NumDealsPurchases

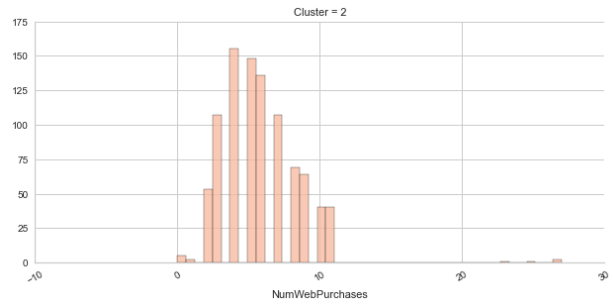

Cluster = 1
MntFruits

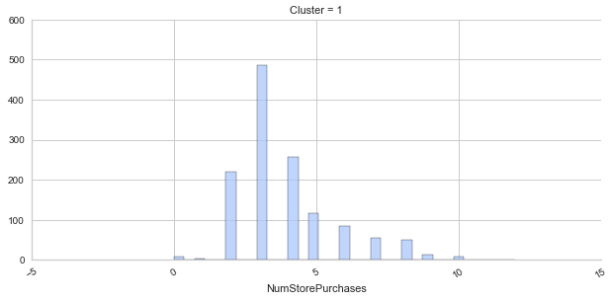

Cluster = 2
MntFruits

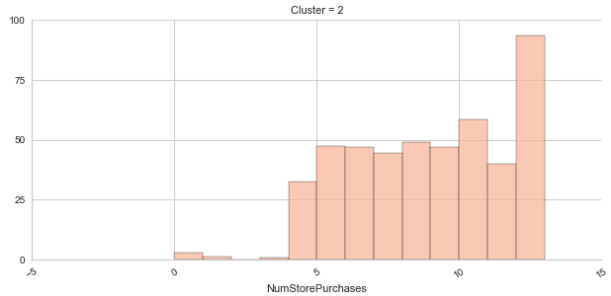

Cluster = 1
MntSweetProducts


Cluster = 2
MntSweetProducts

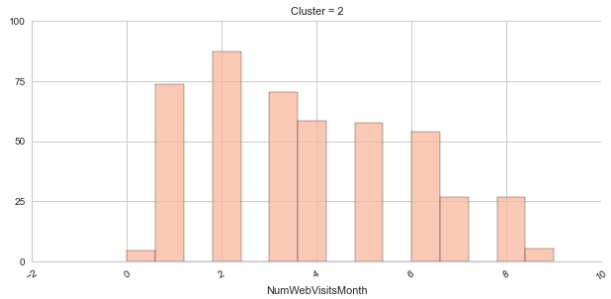

Cluster = 1
NumWebPurchases

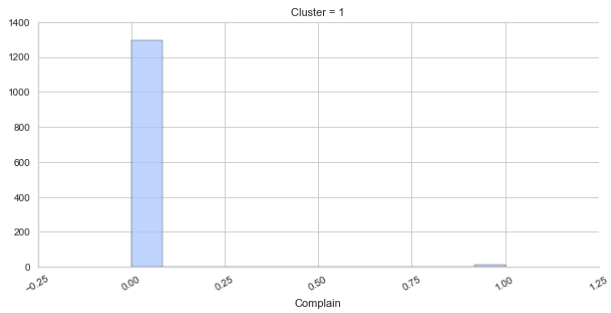

Cluster = 2
NumWebPurchases

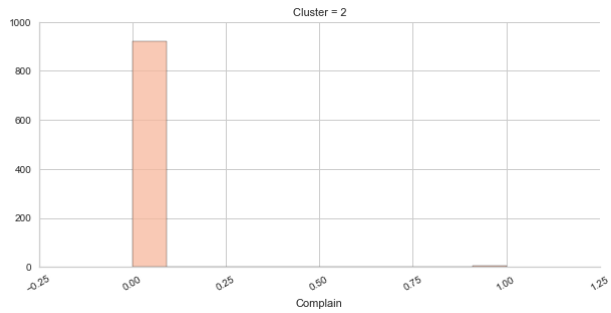

Cluster = 1
NumStorePurchases

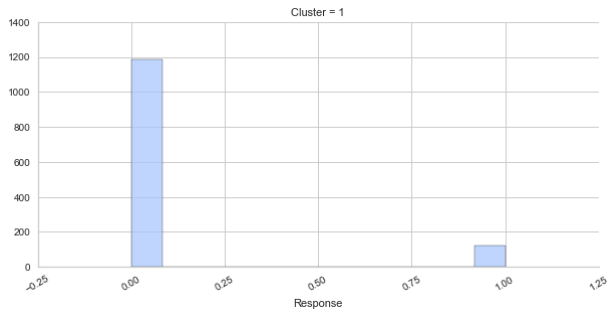

Cluster = 2
NumStorePurchases

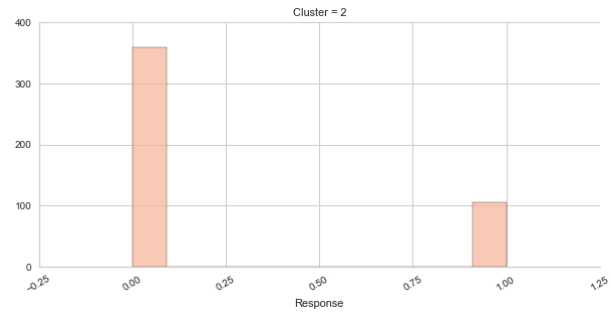

Cluster = 1
NumWebVisitsMonth

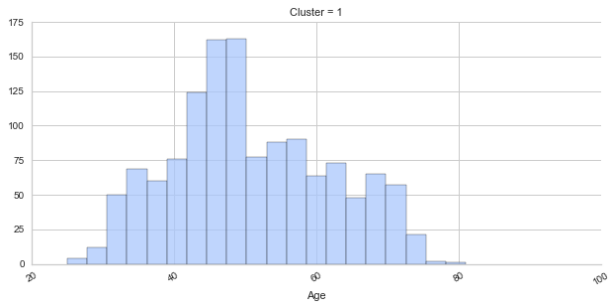

Cluster = 2
NumWebVisitsMonth

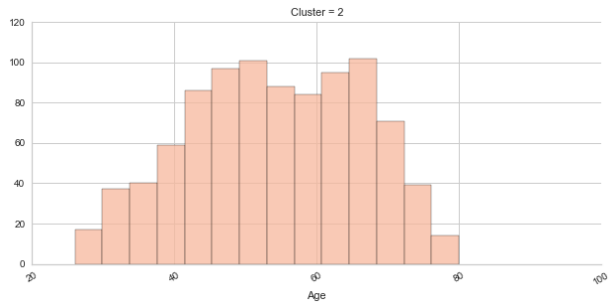

Cluster = 1
Complain

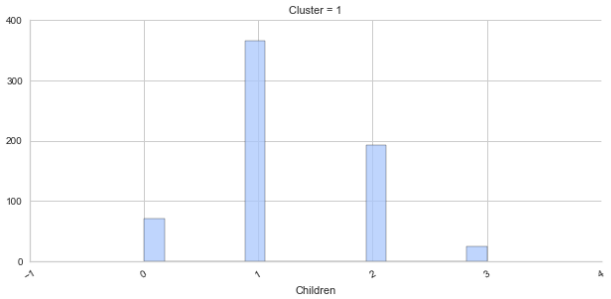

Cluster = 2
Complain

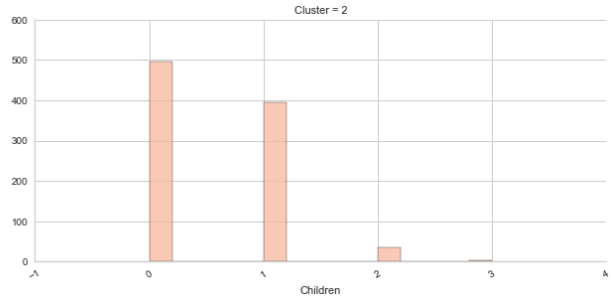

Cluster = 1
Response

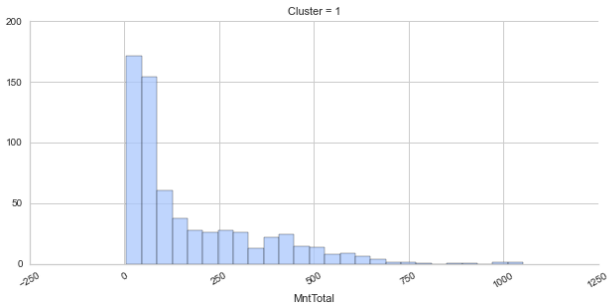

Cluster = 2
Response

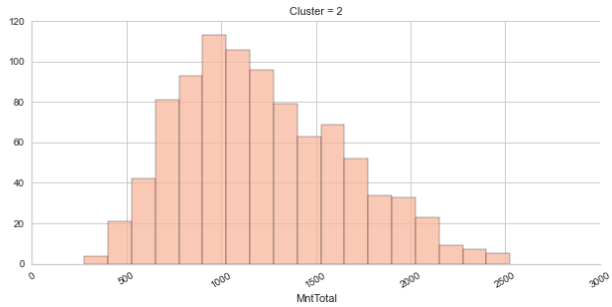

Cluster = 1
Age

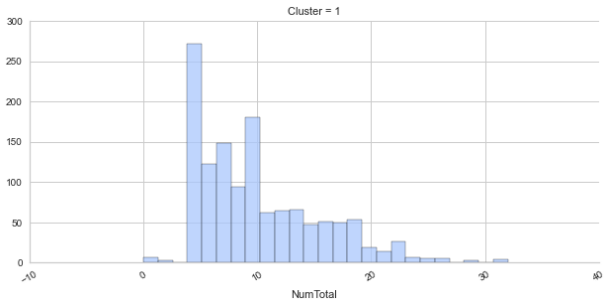

Cluster = 2
Age

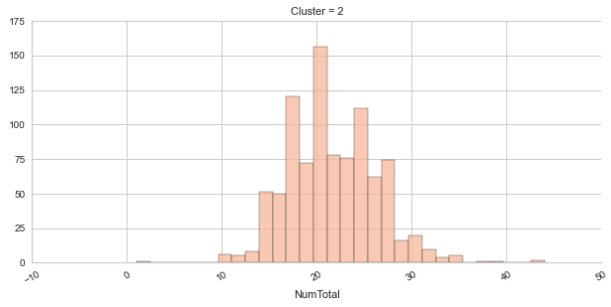

Cluster = 1
Children

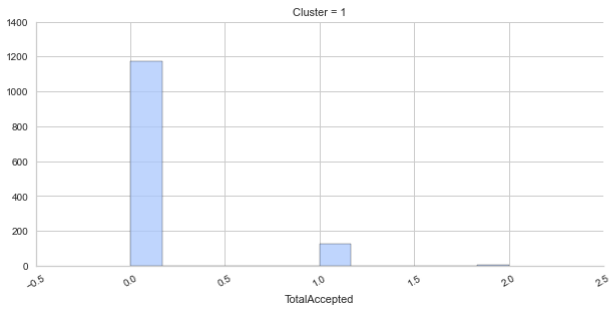

Cluster = 2
Children

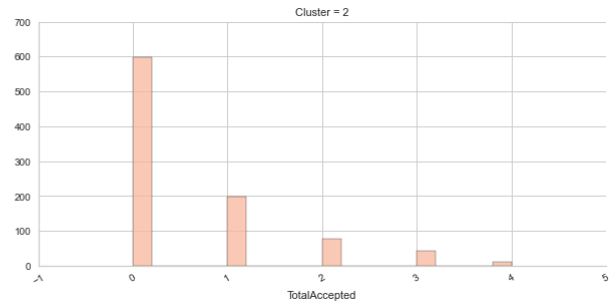

Cluster = 1
MntTotal

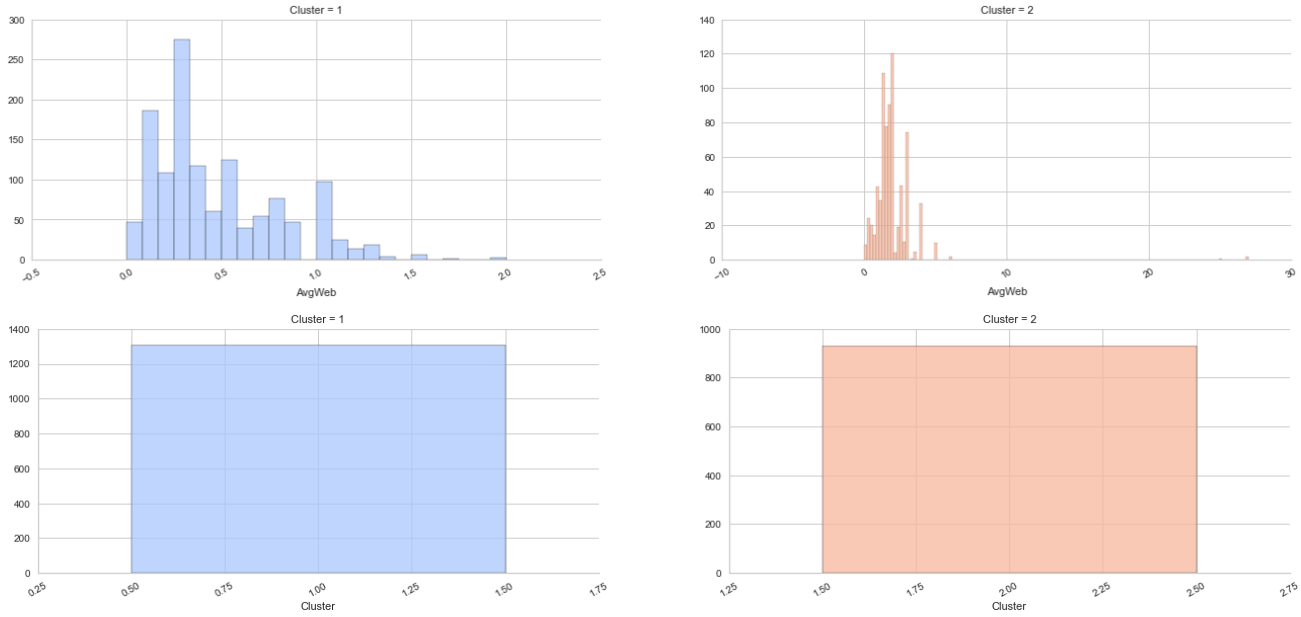

Cluster = 2
MntTotal


Cluster = 1
NumTotal


Cluster = 2
NumTotal


Cluster = 1
TotalAccepted


Cluster = 2
TotalAccepted



Both the above clusters found by KMeans have distinct differences shown as:

Cluster 1 (blue)	Cluster 2 (orange)
Higher Income	Lower Income
Less or no Kids/Children. About 5% of this population has 1 Kid/Child	Have either 1 or more Kids/Children
Buys more products but cheaper ones	Buys less product but expensive ones
Have more consumption power	Have less consumption Power
More likely to buy goods from us	Not that likely to buy our product as compared to Cluster 1
Accepts more offers than Cluster 2	Accepts less offers than Cluster 1

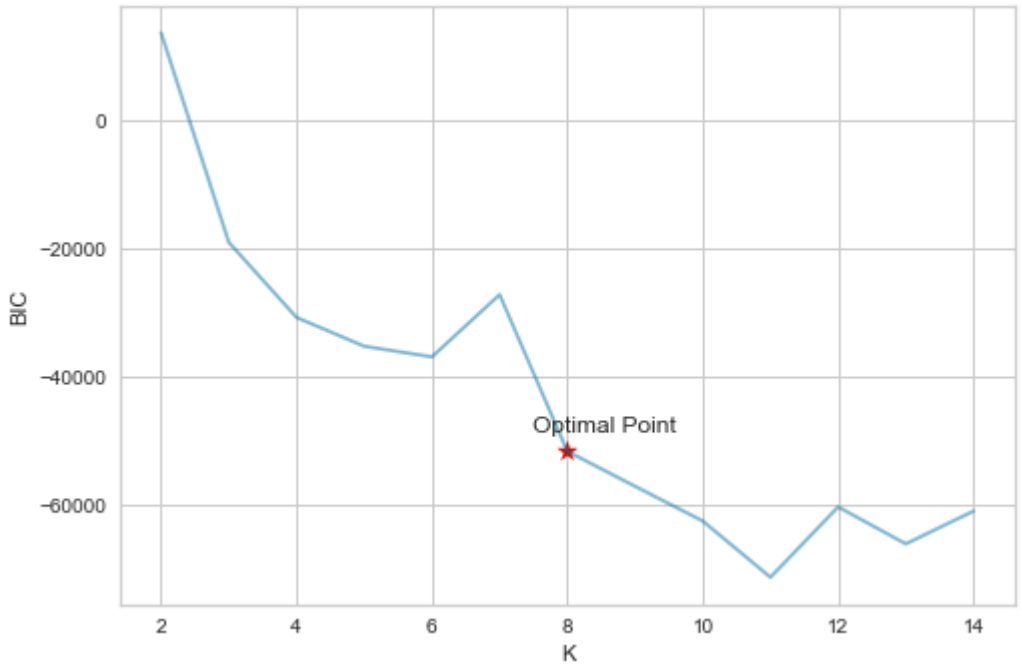
Another observation which should be brought to light is that Cluster 2 have some unusual amount of products purchased at the campaign. This might signify that those purchases were for a festival or other similar reason. Checking the purchase date might verify this.

8.3- Gaussian Mixture Model

[8]Gaussian Mixture Model is a clustering algorithm which works on simple principles but with unerring accuracy. It is soft version of K-Means which calculates the sample probability to different clusters. One advantage of using this model is that this model does not require to know which sample a data point belongs to. This is instead learnt automatically.

By using random state=100 (other inputs remain constant) we found out the optimal value for K as shown in fig 8.3.1.

Fig 8.3.1 – *Finding the optimal value for K in GMM*

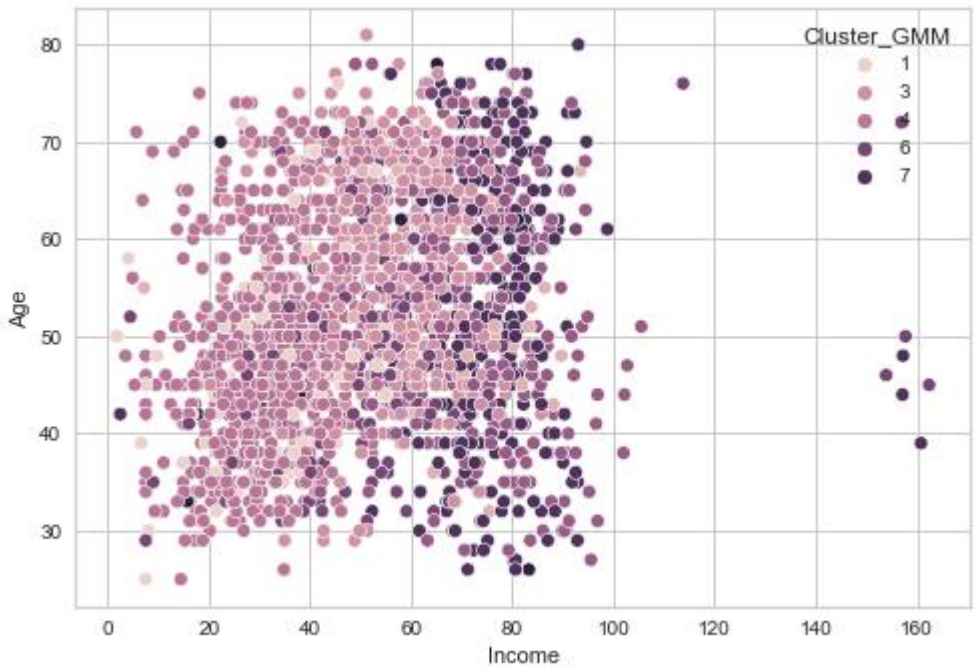


Here, we use BIC to evaluate the effectiveness of clustering. When K=8, the BIC score comes to the balanced point (will not show much improvement when increasing K), so we choose 8 as the final clustering result.

Afterwards we built and fitted the clustering model and tried to inspect the clusters formation. We found out the following results(shown w.r.t importance):

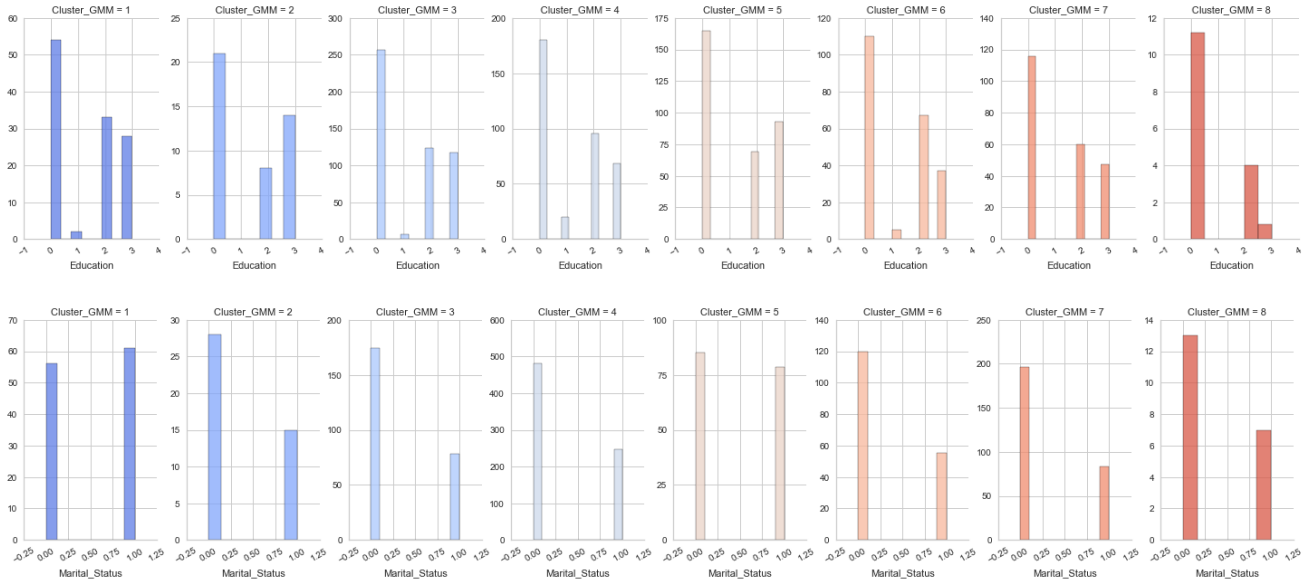
```
4    727
3    504
5    327
7    279
6    219
1    117
2     43
8     20
Name: Cluster_GMM, dtype: int64
```

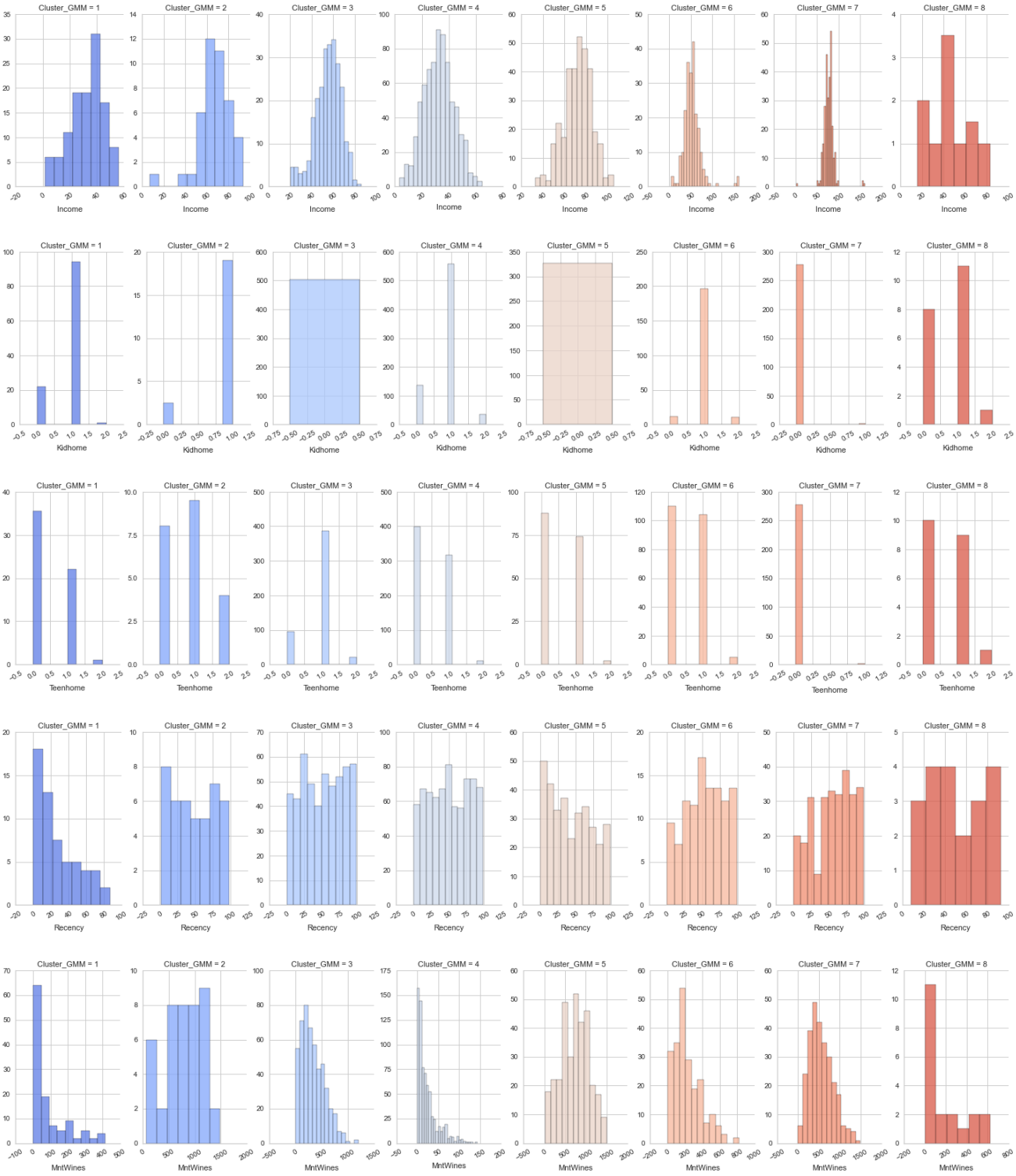
Fig 8.3.2 – *Inspecting the cluster difference*

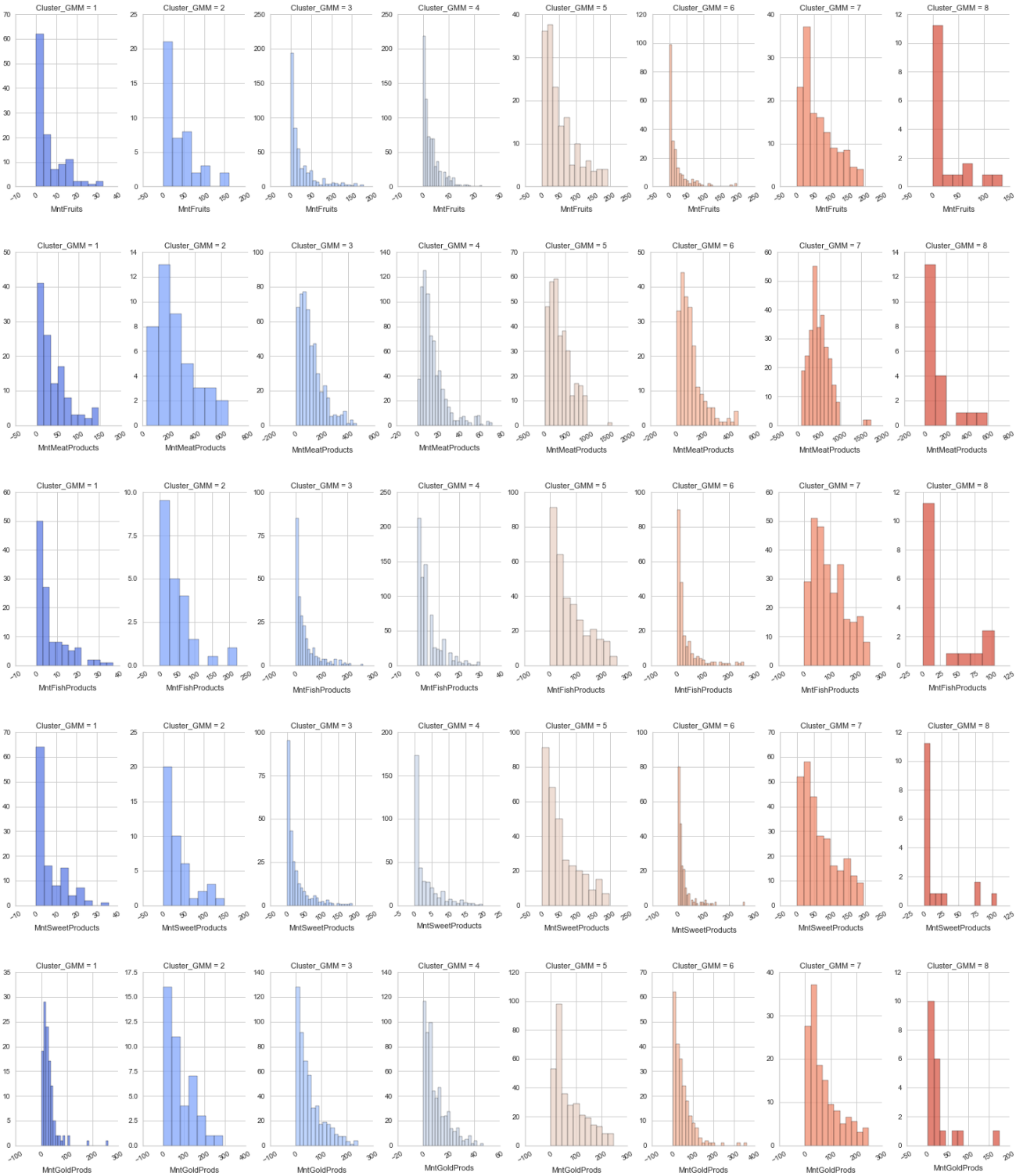


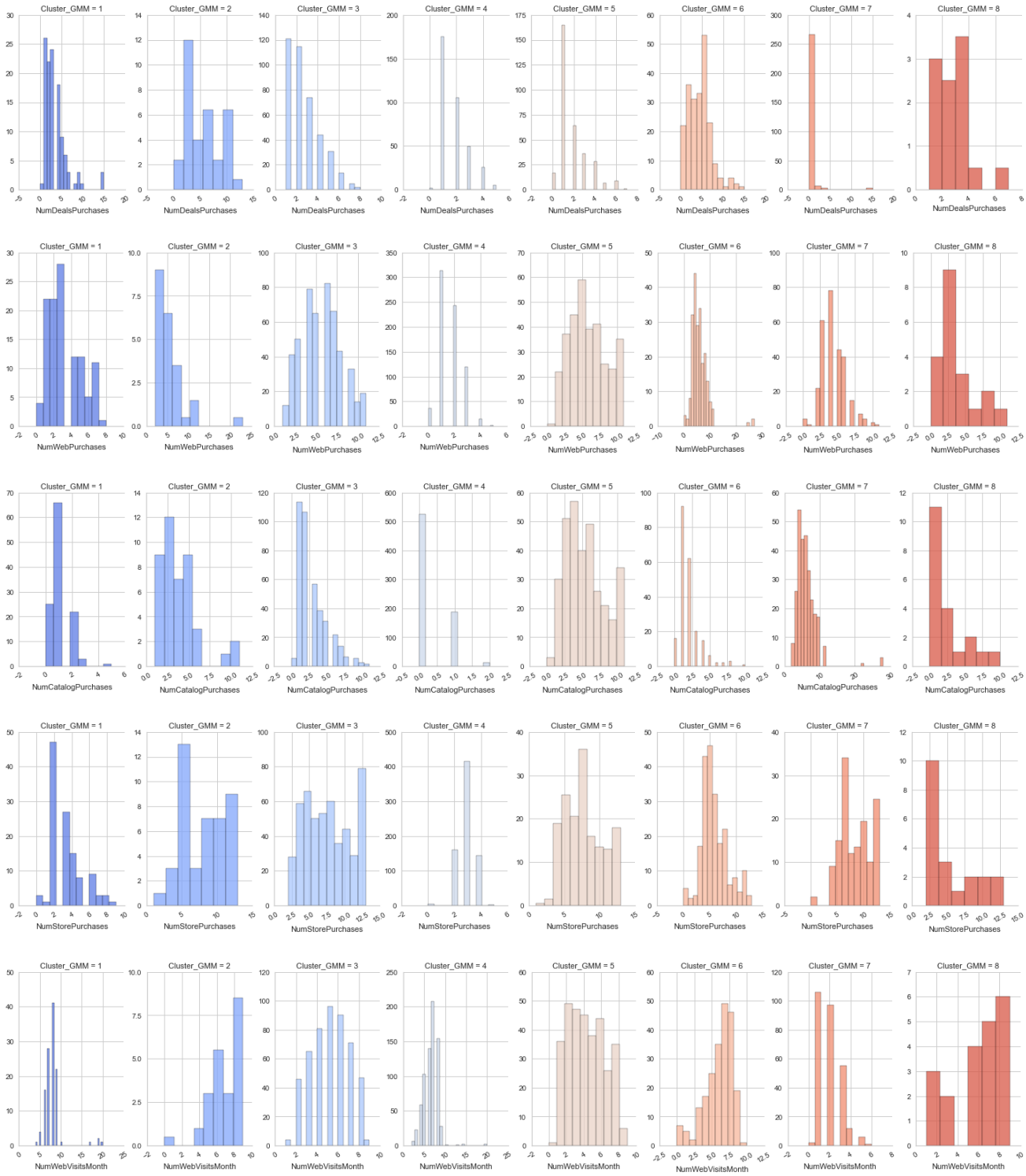
Now we plot values for each field in our dataset with respect to each cluster just like we did in the K-Means:

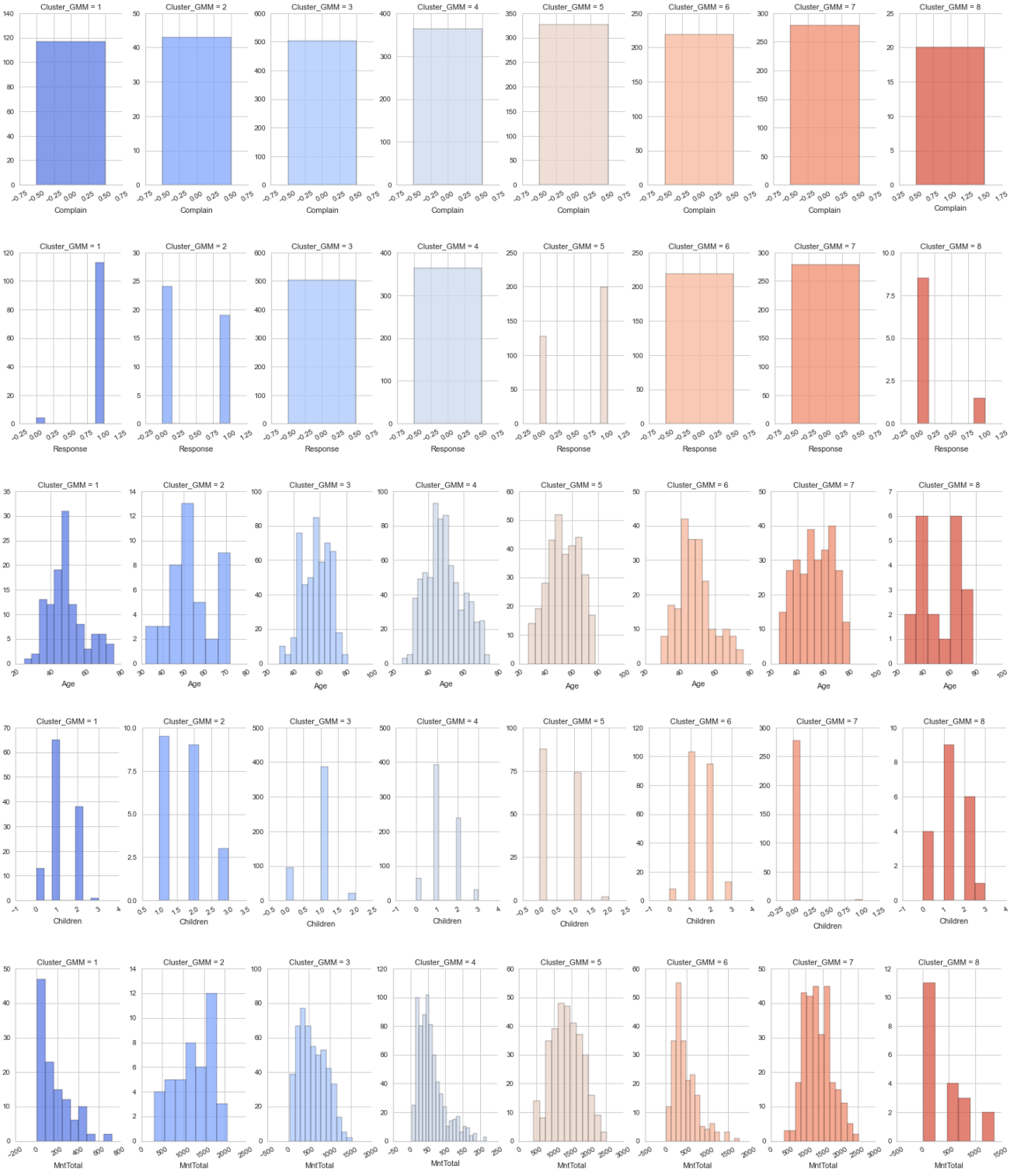
Fig 8.3.3 – *Plots for each column w.r.t clusters*

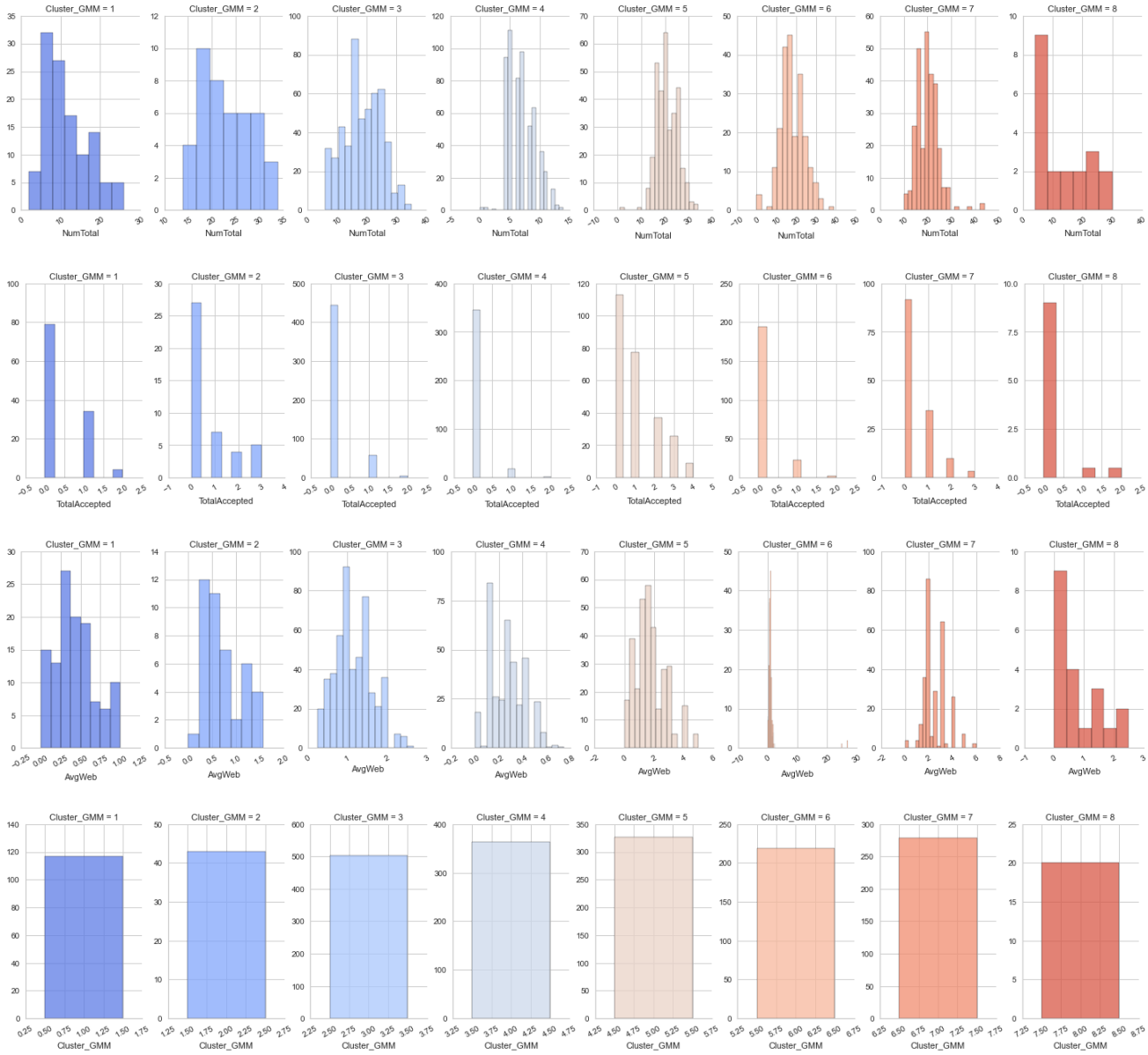



Cluster_GMM = 1
Cluster_GMM = 2
Cluster_GMM = 3
Cluster_GMM = 4
Cluster_GMM = 5
Cluster_GMM = 6
Cluster_GMM = 7
Cluster_GMM = 8
Income
Kidhome
Teenhome
Recency
MntWines


Cluster_GMM = 1
Cluster_GMM = 2
Cluster_GMM = 3
Cluster_GMM = 4
Cluster_GMM = 5
Cluster_GMM = 6
Cluster_GMM = 7
Cluster_GMM = 8
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
MntGoldProds


Cluster_GMM = 1
Cluster_GMM = 2
Cluster_GMM = 3
Cluster_GMM = 4
Cluster_GMM = 5
Cluster_GMM = 6
Cluster_GMM = 7
Cluster_GMM = 8
NumDealsPurchases
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth


Cluster_GMM = 1
Cluster_GMM = 2
Cluster_GMM = 3
Cluster_GMM = 4
Cluster_GMM = 5
Cluster_GMM = 6
Cluster_GMM = 7
Cluster_GMM = 8
Complain
Response
Age
Children
MntTotal



There are 8 different clusters, which is difficult to describe, but we could see clear difference in their basic information, family condition and consumption power

9- Summary

9.1- Customer Related Summary

- Most of our Customer base belongs to those who are commited in a relationship(Married Couples or Together) or people who hold a Bachelors Degree. We are not popular among people with Basic education and among those who used to be in a relationship and now they are not.
- Most of our Customers are around 45-50 years old.
- Most of them have either one or no kids/teens at home. Those with more than 1 kid/Teen are either not familiar with our business or are not interested.
- Our Marketing Campaign yeilded exponentially more Negative responses than Positive ones.
- Those whom we approached with our Campaign earned 10,000 to 100,000 units of currency.
- Those with higher income were more interested, while those with income lower than 14-15,000 were either not interested or they rejected our offers.
- PhD or Masters Degree holders were more likely to accept our offers. Those with lower levels of education were hardly interested.
- Those who have lower level of education have less income, while those with Bachelors, Masters or PhD degree do not have clear distinction between their income.
- Those with no kids at home had higher income.
- Single people seemed more likely to accept our offers as compared to those commited in a relationship. But as shown by KDE plot, different Marital Status does not seem to be the cause of positive or negative response to our marketing campaign.
- No Linear Relationships are exhibited by the Customer data

9.2- Supervised Prediction Summary

Algorithms	Mathews Correlation Coefficient	Accuracy Score
Logistic Regression	0.499386 mean (0.030593 std)	0.749420 mean (0.015291 std)
Boosting Tree	0.821840 mean (0.169042 std)	0.904652 mean (0.096071 std)
SVM	0.644576 mean (0.039894 std)	0.802105 mean (0.023620 std)
Neural Network	0.631383 mean (0.086735 std)	0.805347 mean (0.050553 std)

- Boosting Tree works really well among all the tested models.
- Feature-Selected Dataset shows usability in the models
- Therefore, we are using **Boosting Tree dataset + Feature-Selected dataset** to achieve the best MCC score.

Overall Model Performance;

	Precision	Recall	F1 score	Support
0	0.93	0.93	0.93	204
1	0.55	0.53	0.54	32
Accuracy			88	236
Macro Avg	0.74	0.73	0.73	236
Weigted Avg	0.88	0.88	0.88	236

- The overall test accuracy of the model is 0.88. But by observing the score report more closely, we find that the model performs quite good in recognizing negative samples(0), but does not perform well in recognizing positive samples (precision: 0.55, recall: 0.55).
- The test MCC is 0.469, which indicates that the model may not good at finding positive samples in test set. While we find the Train MCC is 0.98, this result shows there might exist overfitting problem in the model. (tried and couldn't solve it)

9.3- Unsupervised Prediction Summary

K Means

- Cluster 1 have obvious higher income than Cluster 2.
- Cluster 1 have either none or at least 1(about 5% of them) kids at home. Whereas Cluster 2 have 1 or more.
- Cluster 1 customers buy much more products than cluster 2 customers. Almost every products shows the same trend in each cluster. This implies that cluster 1 have more consumption power, and they are more likely to purchase goods from the company.
- The cluster with more consumption power(cluster 1) are shown to accept more offers than the other.
- Also, people in cluster 1 are observed to have overall more purchases with respect to quantity. Among all these places, they may prefer to buy products in real store
- Cluster 2 have some extreme situations in product purchasing amount. Some customers in Cluster 2 are observed to have purchased an unusual amount of products overtime.

Gaussian Mixture Model

- Working with GMM, we found that it created 8 different clusters.
- Apparently the 8 clusters have a range of lowest to highest in each field (such as education, no. of kids at home etc)

10- References

[0] Customer Personality Analysis, data source – Kaggle.com

[1] Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of (Nov, 22, 2019) – *towards data science* by Boaz Shmueli

[2] Machine Learning playlist - *Code Basics* by Dhaval Patel

[3] Kaggle courses – *Kaggle.com*

[4] Analytics Vidhya courses - *Analytics Vidhya*

[5] Sklearn's documentation - *scikit-learn.org*

[6] Pandas documentation – *pandas.pydata.org*

[7] SMOTE for Imbalanced Classification with python (Mar 17, 2021) – *machinelearningmastery.com* by Jason Browniee

[8] Gaussian Mixture Model (Oct 31,2019) - *analyticsvidhya.com* by Aishwarya Singh

END OF DOCUMENT