**CUSTOMER PERSONALITY ANALYSIS PROJECT**

**Problem Statement:-** A grocery shop wants to segment the customers into the groups that reflect similarities among customers in each cluster. **It also helps the business to cater to the concerns of different types of customers.**

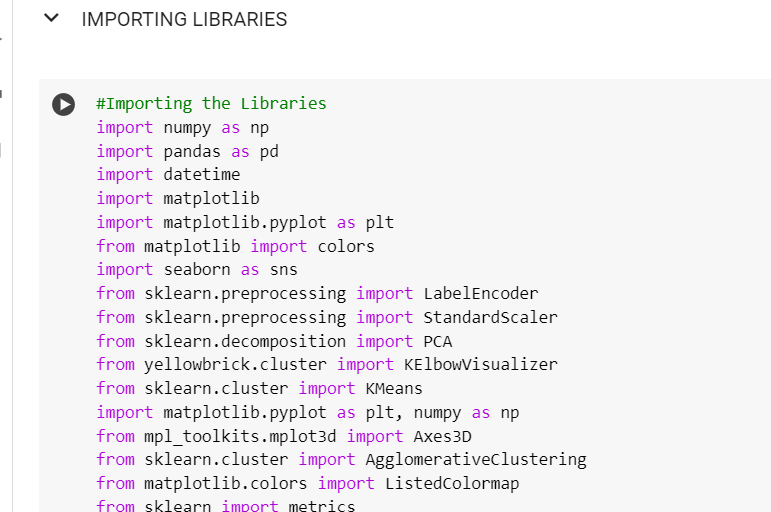
Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company’s database, a company can analyze which customer segment is most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

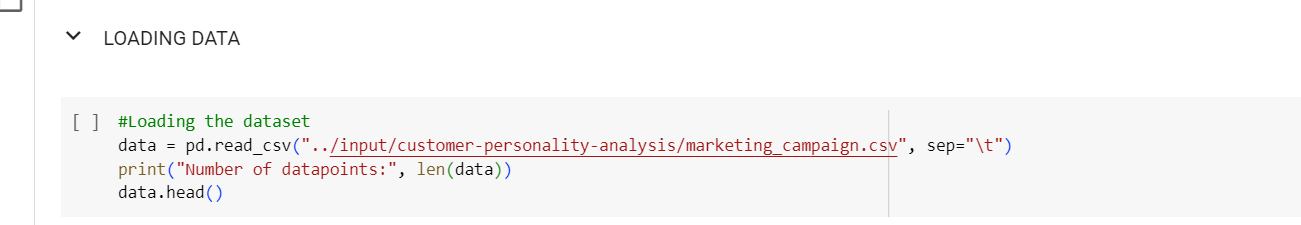
The most important part of a customer personality analysis is getting the answers to questions such as:

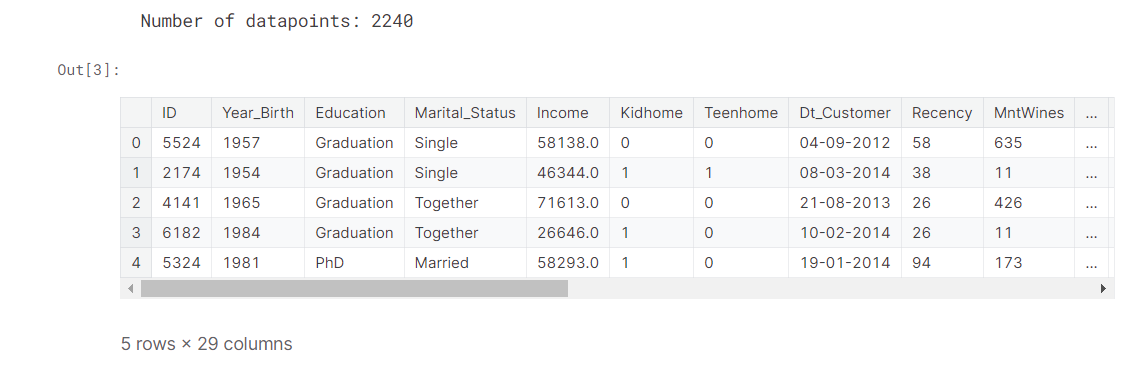
1. What people say about your product: what gives customers’ attitude towards the product.
2. What people do: which reveals what people are doing rather than what they are saying about your product.

**Approach:-** It is an unsupervised learning problem i.e there is no mapping between the input and the output variables.

So for clustering we will be using the **K-Means clustering Algorithm and finding the optimal number of clusters using the elbow method . I will divide customers into segments to optimize the significance of each customer to the business.**

****

****

****

**About the data set:-**

**Rows:- 2240 Attributes:- 29**

1. **Customer’s Information:-**

ID

Year\_Birth

Education

Marital\_status

Income

KidHome

TeenHome

Dt\_Customer

Recency

Complain

1. **Products**

Amount spent on different products on last 2 years

MnWines

MnFruits

MnFishProducts

MnMeatProducts

MtSweetProducts

MnGoldProducts

1. **Promotion**

NumDealsPurchases

AcceptedCmp1

AcceptedCmp2

AcceptedCmp3

AcceptedCmp4

AcceptedCmp5

Response

1. **Place**

NumWebPurchases

NumCatalogPurchases

NumStorePurchases

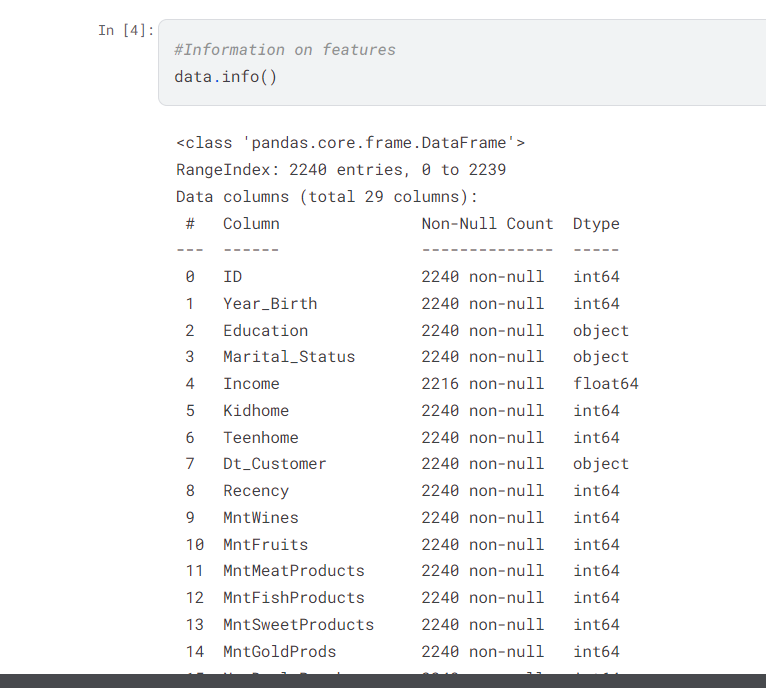
NumWebVisitsMonth

**DATA CLEANING**

In this section

* Data Cleaning
* Feature Engineering

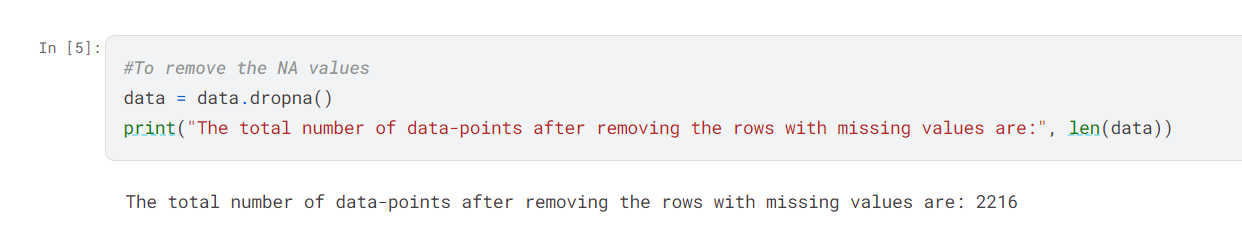
Let us have a look at the information of the data:-



From the above we conclude that:-

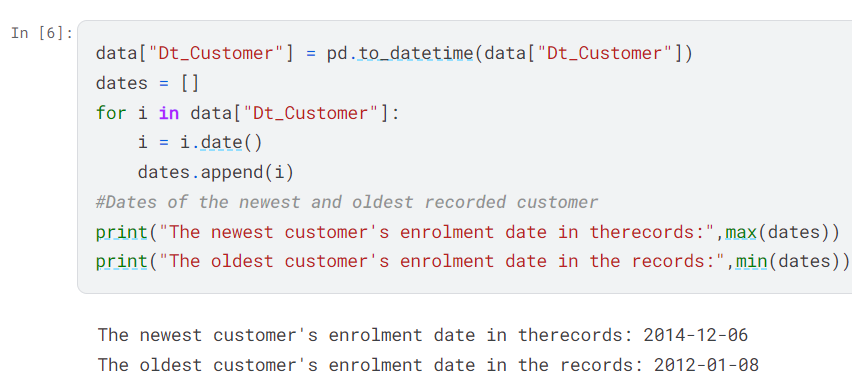
1. There are **24 missing values i.e 1.071%** of the data in the **income**
2. Dt\_customer that indicates the date on which the customer joined the database is **not parsed as datetime rather than as an object.**
3. There are some categorical data such as education, marital\_status. So we need to encode it in numerical data.

As the **missing values are very less** we will **delete all the rows** containing the missing values

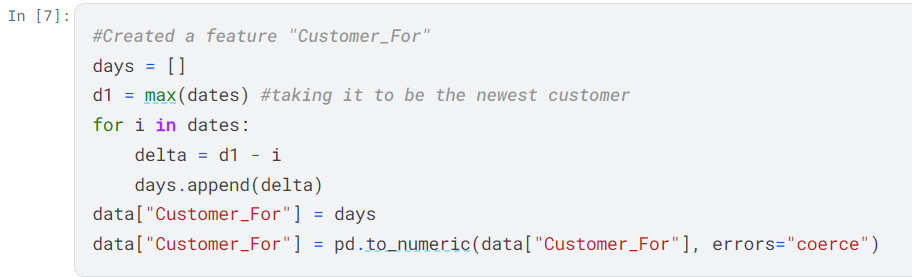


Now after removing there are **2216 data points.**

In the next step, I am going to create a feature out of "Dt\_Customer" that indicates the number of days a customer is registered in the firm's database. However, in order to keep it simple, I am taking this value relative to the most recent customer in the record.

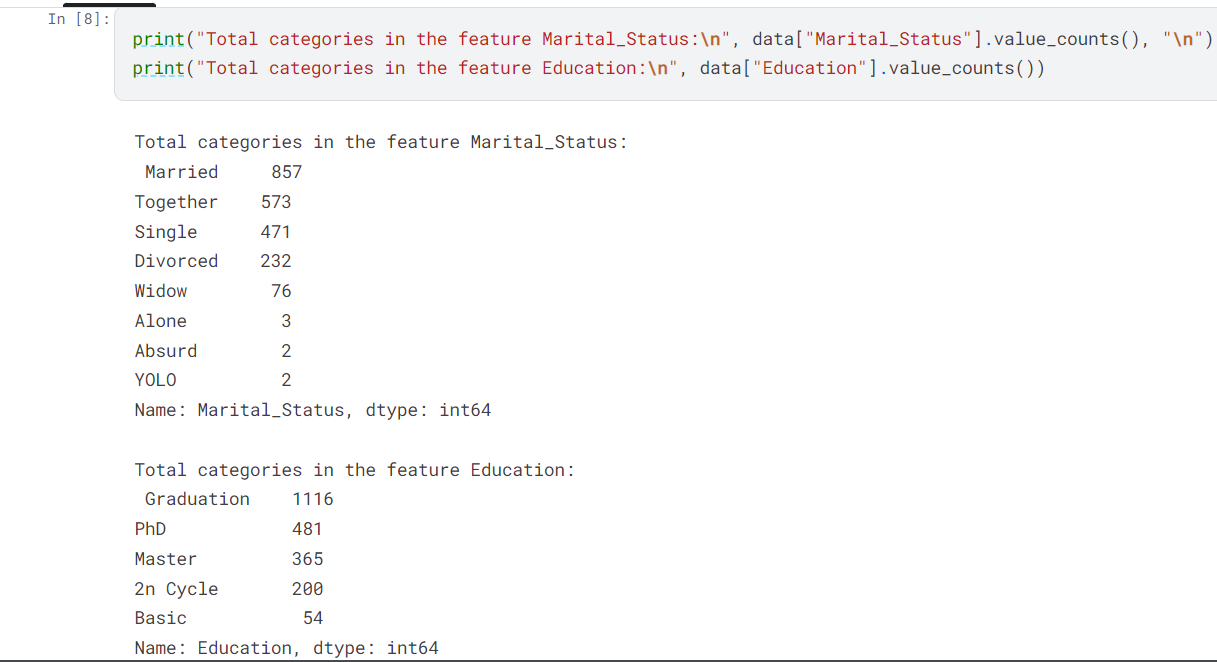


Creating a feature **Customer\_for** of the number of days the customers started to shop in the store relative to the last recorded data



Now we will be exploring the unique values in the categorical features to get the clear idea of the data

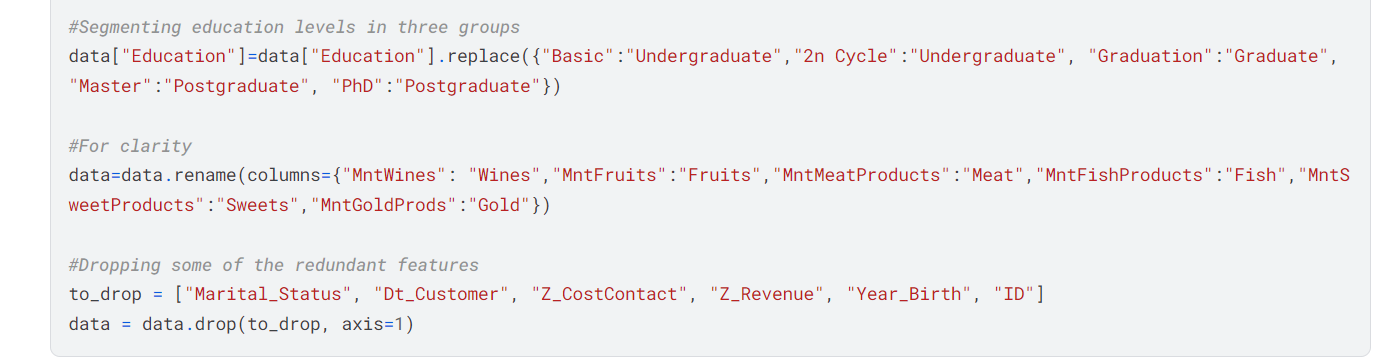
**Martial\_Status and Education**

****

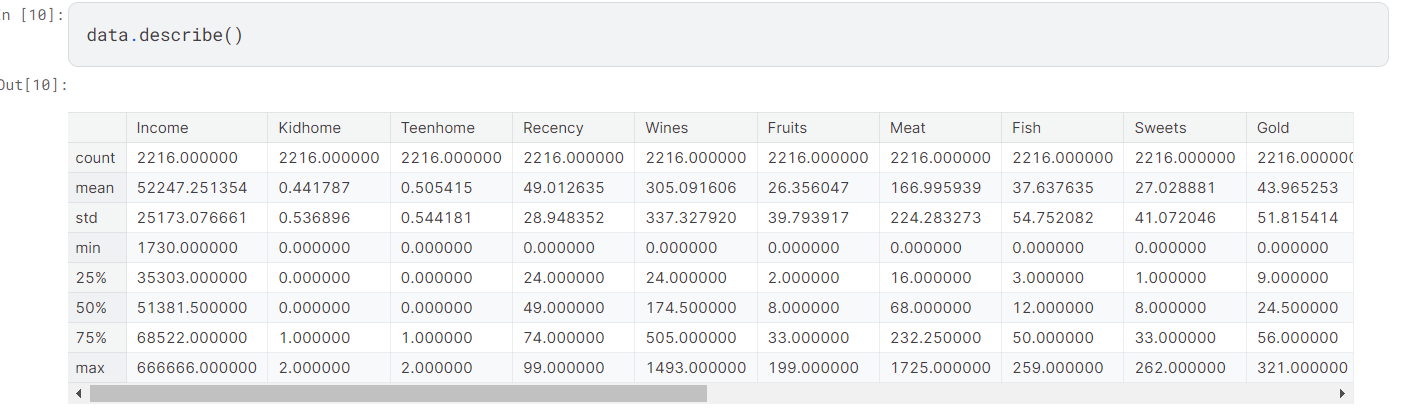
**In the next bit, I will be performing the following steps to engineer some new features:**

* Extract the **"Age"** of a customer by the **"Year\_Birth"** indicating the birth year of the respective person.
* Create another feature **"Spent"** indicating the **total amount spent** by the customer in various categories over the span of two years.
* Create another feature **"Living\_With"** out of **"Marital\_Status"** to extract the living situation of couples.
* Create a feature **"Children"** to indicate total children in a household, i.e **kids and teenagers.**
* To get further clarity of household, Creating feature indicating **"Family\_Size"**
* Create a feature **"Is\_Parent"** to indicate parenthood status
* Lastly, I will create three categories in the **"Education"** by simplifying its **value counts.**
* Dropping some of the redundant features

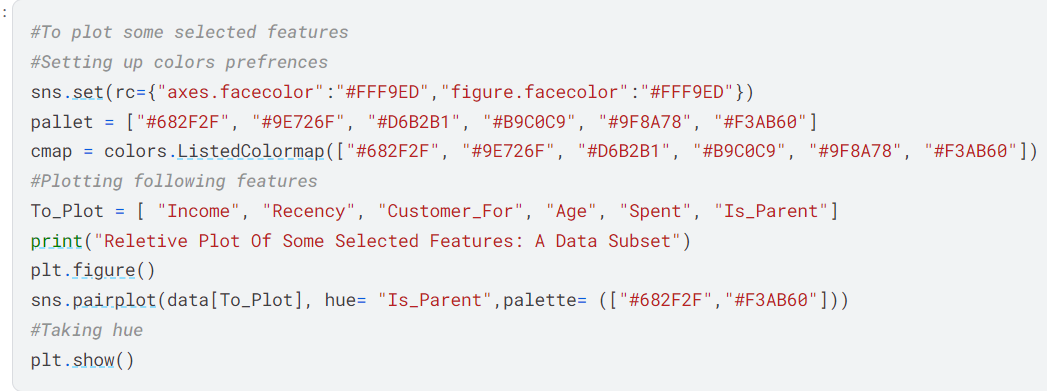


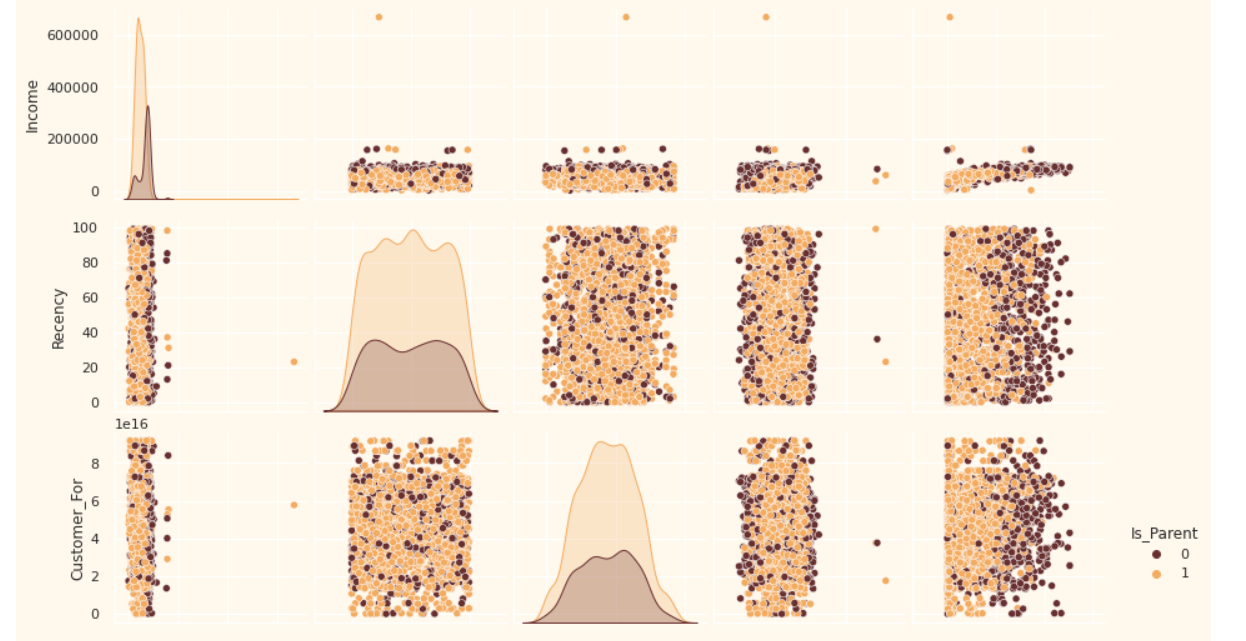


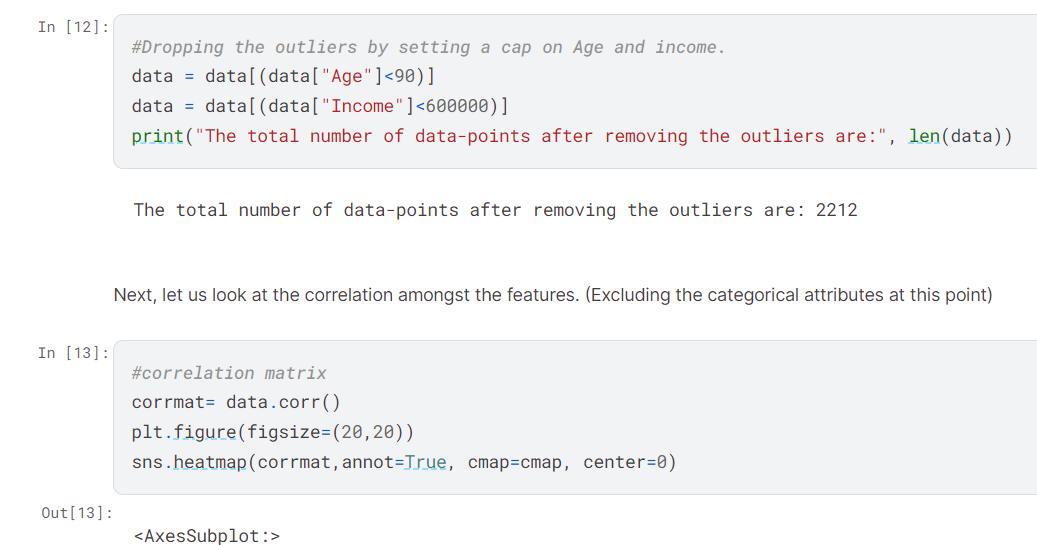
Now that we have some new features let’s have a look at the data stats

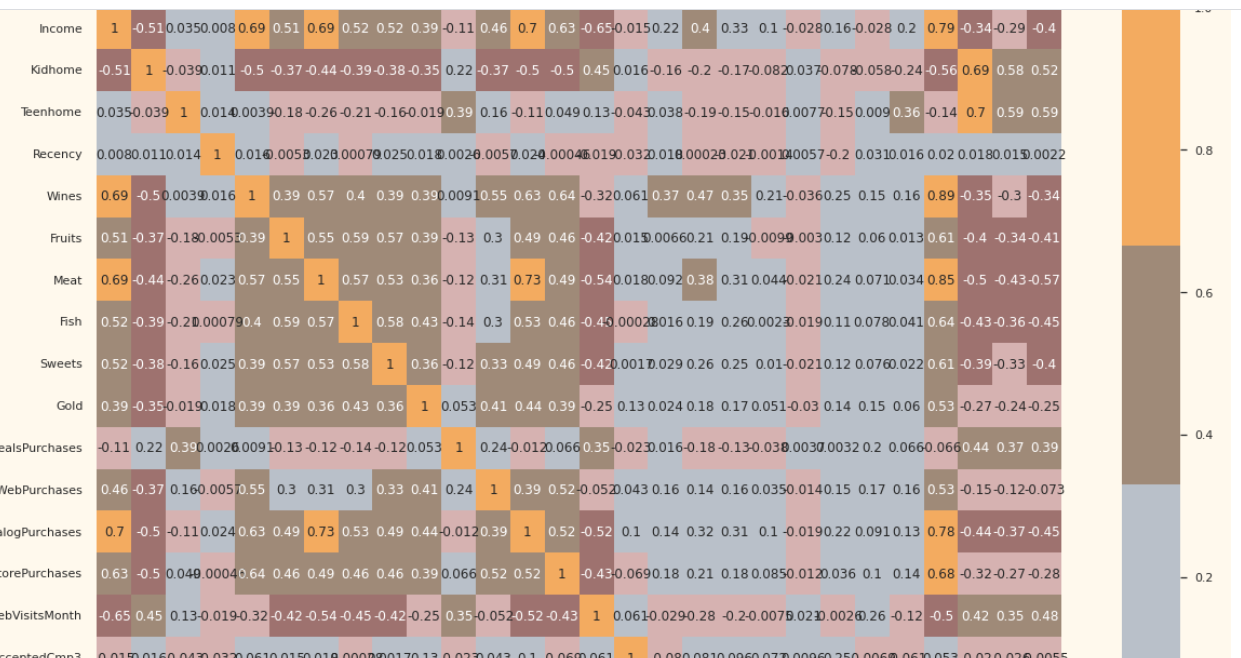


The above stats show some discrepancies in mean Income and Age and max Income and age. Do note **that max-age is 128 years,** As I calculated the age that would be today (i.e. 2021) and the data is old. I must take a look at the broader view of the data. I will plot some of the selected features.





Clearly, there are a few outliers in the Income and Age features. I will be deleting the outliers in the data as they are 2 i.e 0.09% of the total data



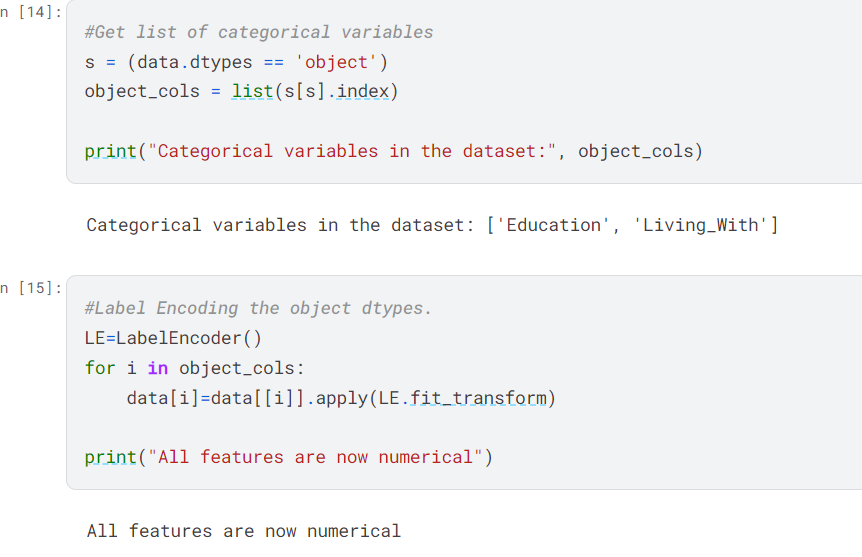
The data is quite clean and the new features have been included. I will proceed to the next step. That is, preprocessing the data.

**Data Processing**

In this section, I will be preprocessing the data to perform clustering operations.The following steps are applied to preprocess the data:

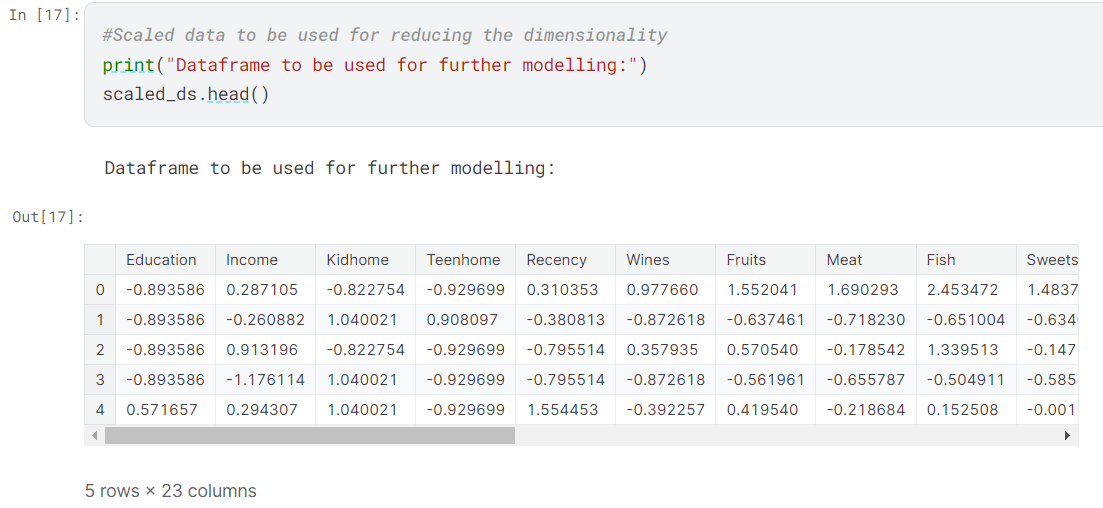
* Label encoding the categorical features
* Scaling the features using the standard scaler
* Creating a subset dataframe for dimensionality reduction

**Encoding the categorical data using Label Encoding**



**Scaling all the columns :** As K-Means Clustering is a euclidean distance based





**DIMENSIONALITY REDUCTION**

In this problem, there are many factors on the basis of which the final classification will be done. These factors are basically attributes or features. The higher the number of features, the harder it is to work with it. Many of these features are correlated, and hence redundant. This is why I will be performing dimensionality reduction on the selected features before putting them through a classifier.  
*Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.*

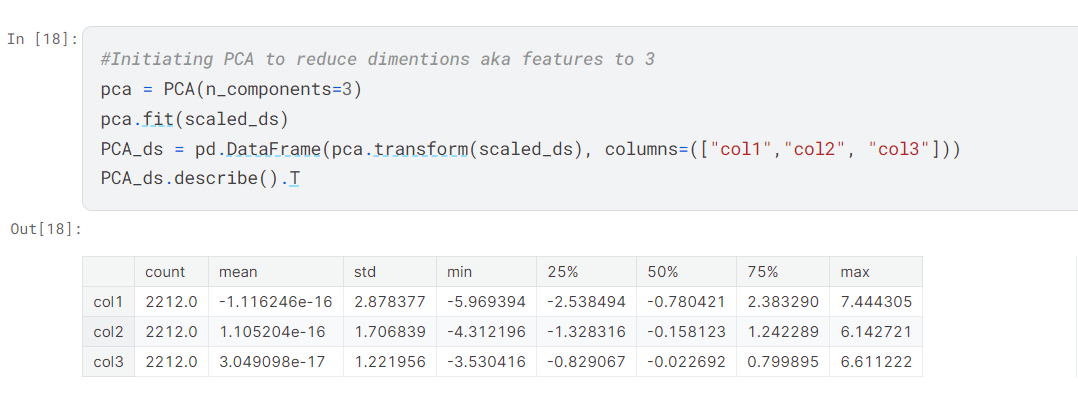
**Principal component analysis (PCA)** is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.

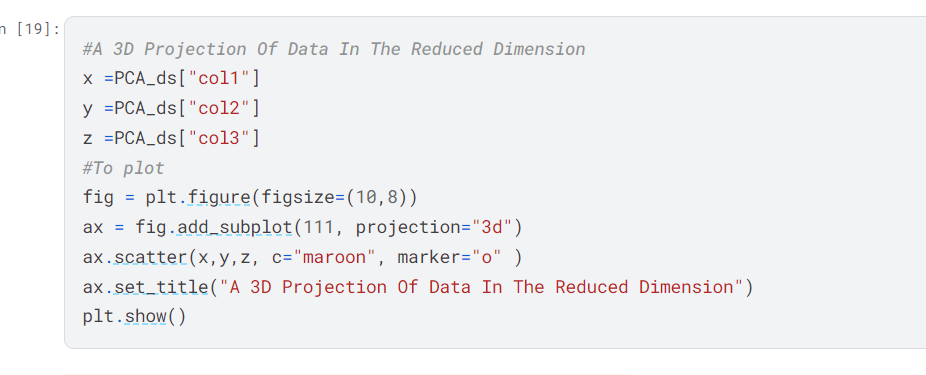
**Steps in this section:**

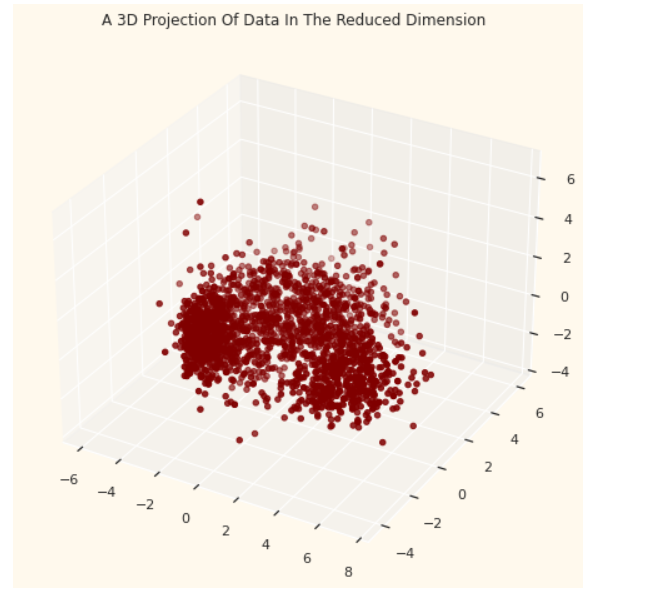
* Dimensionality reduction with PCA
* Plotting the reduced dataframe

**Dimensionality reduction with PCA**

For this project, I will be reducing the dimensions to 3.





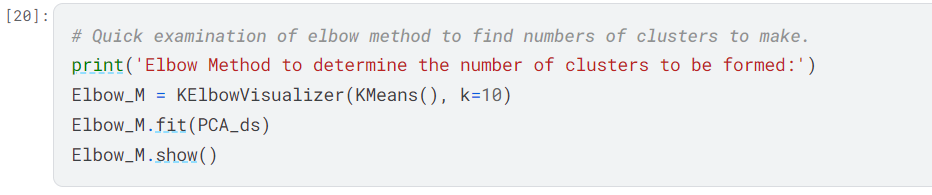


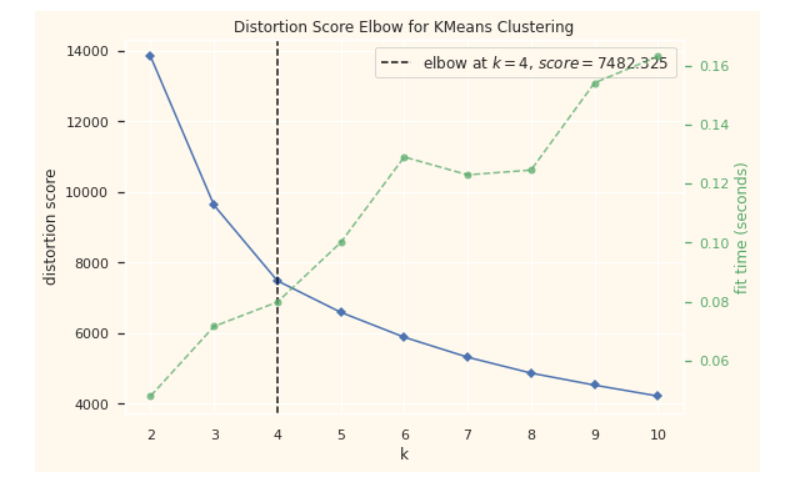
**CLUSTERING**

Now that I have reduced the attributes to three dimensions, I will be performing clustering via K means clustering.

**Steps involved in the Clustering**

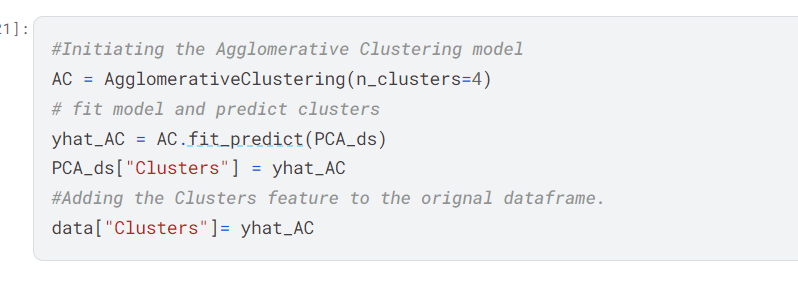
* Elbow Method to determine the number of clusters to be formed
* Clustering via K- means Clustering
* Examining the clusters formed via scatter plot





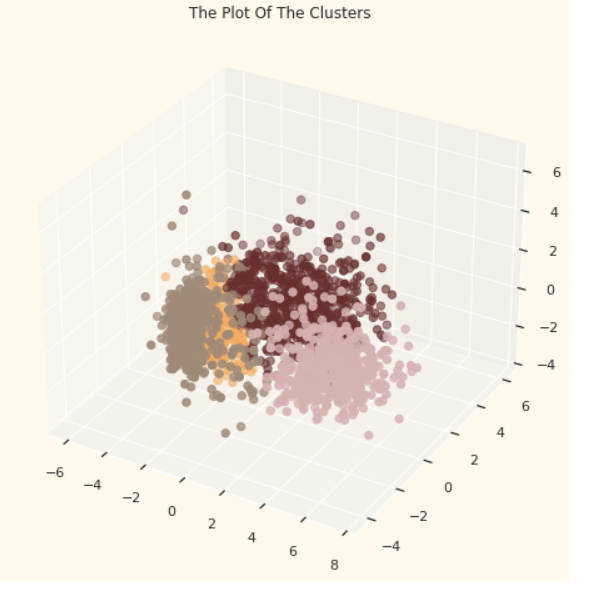
This curve determines that there should be 4 will be an optimal number of the clusters.

We will be using **K-Means Clustering** Algorithm as data had no outlier but the shape of the data was not spherical we then used **DBscan Clustering** Algorithm



To examine the clusters let’s have a look at the 3-D distribution of the clusters.



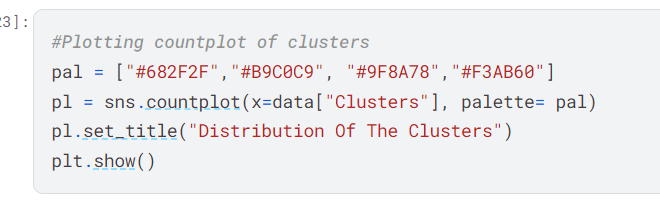


**FINDING PURCHASING PATTERN IN 4 CLUSTERS**

Since this is unsupervised clustering. We do not have a tagged feature to evaluate or score our model. The purpose of this section is to study the patterns in the clusters formed and determine the nature of the clusters' patterns.

For that, we will be having a look at the data in light of clusters via exploratory data analysis and drawing conclusions.

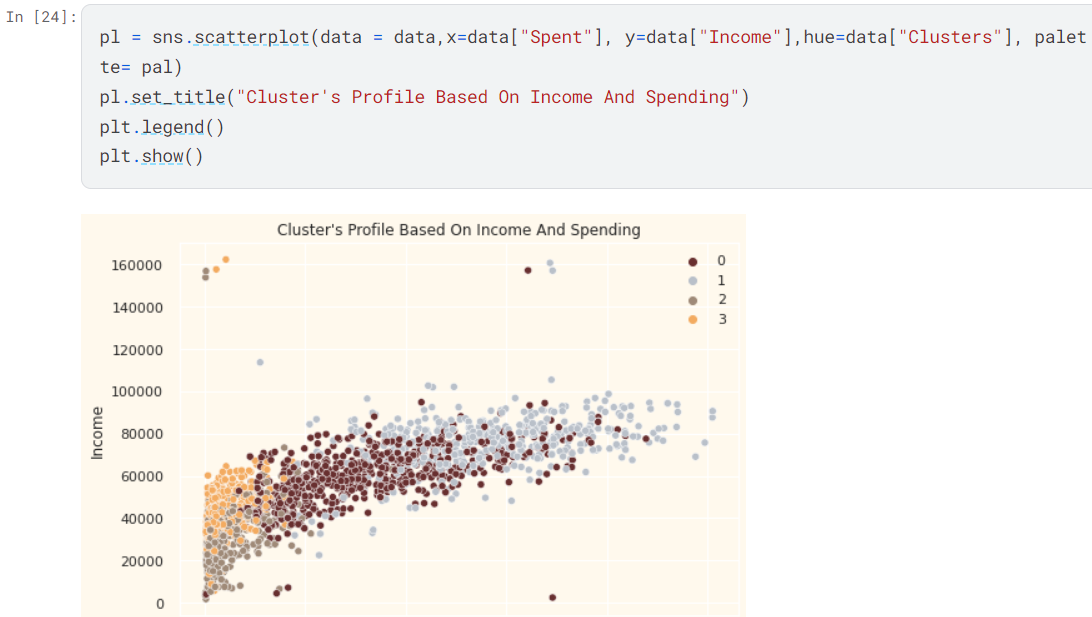
**Firstly, let us have a look at the group distribution of clustering**

****

****

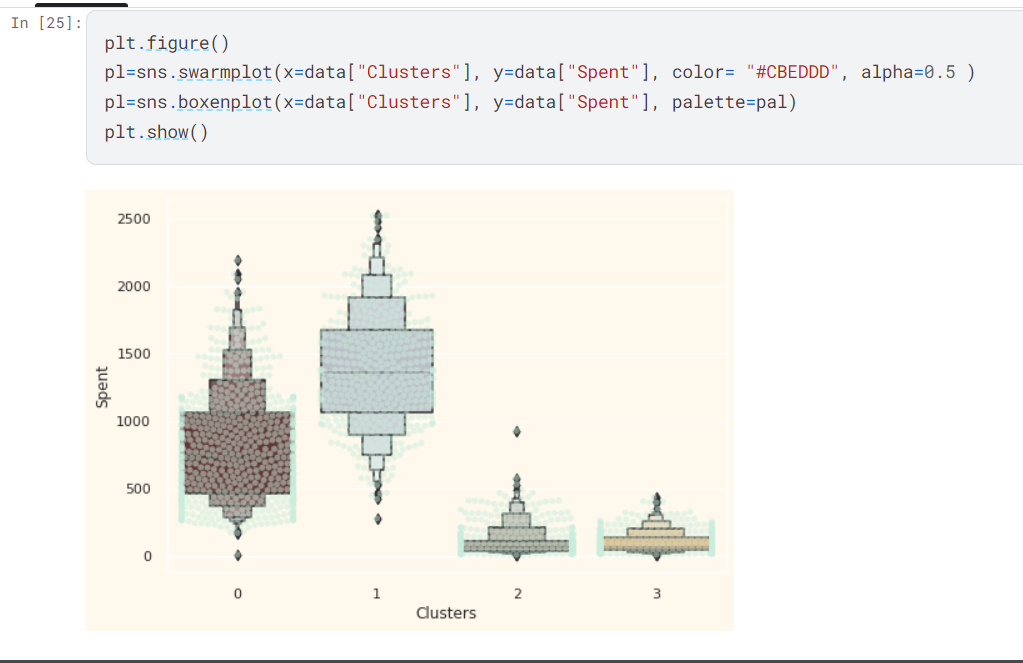
The clusters seems ro be fairly distributed

Now Analyzing the data on the basis of the Spendings and the income.



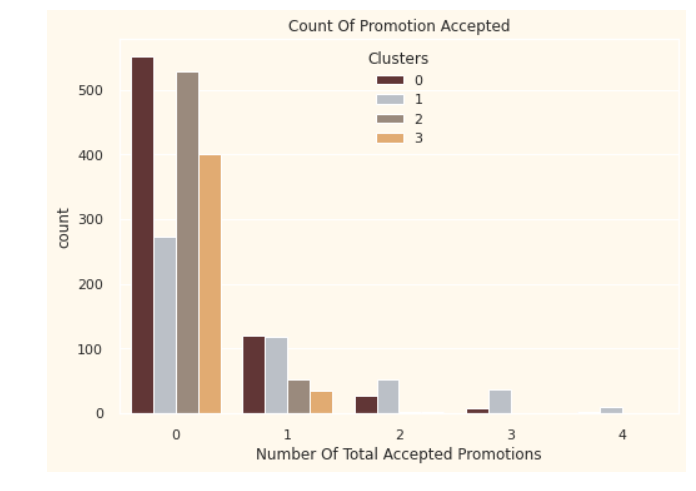
**Income vs spending plot shows the clusters pattern**

* group 0: high spending & average income
* group 1: high spending & high income
* group 2: low spending & low income
* group 3: high spending & low income

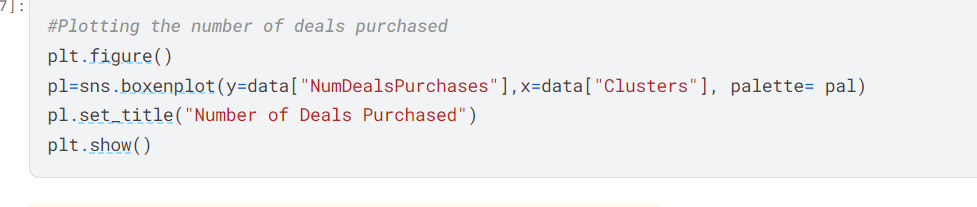


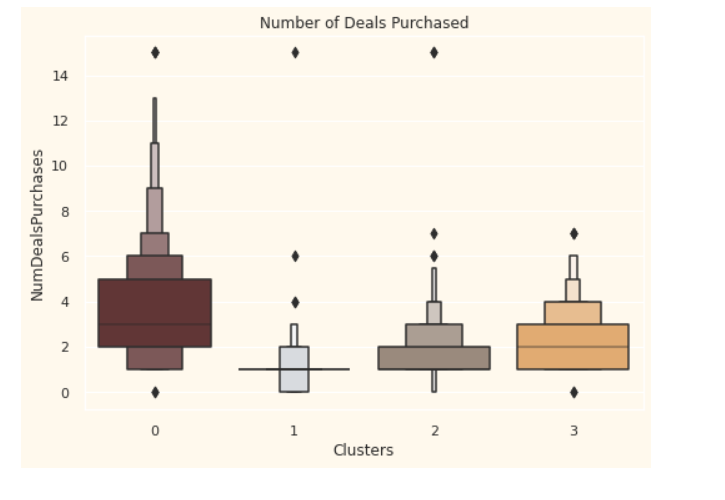
From the above plot, it can be clearly seen that cluster 1 is our biggest set of customers closely followed by cluster 0. We can explore what each cluster is spending on for the targeted marketing strategies.





There has not been an overwhelming response to the campaigns so far. Very few participants overall. Moreover, no one took part in all 5 of them. Perhaps better-targeted and well-planned campaigns are required to boost sales.





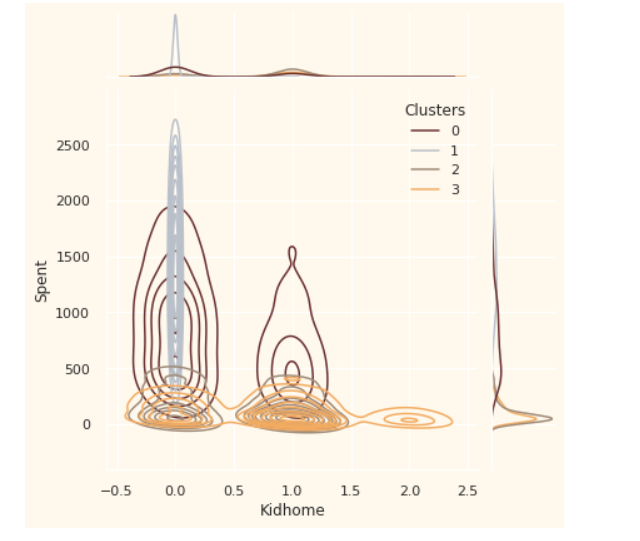
Unlike campaigns, the deals offered did well. It has the best outcome with cluster 0 and cluster 3. However, our star customers cluster 1 are not much into the deals. Nothing seems to attract cluster 2 overwhelmingly.

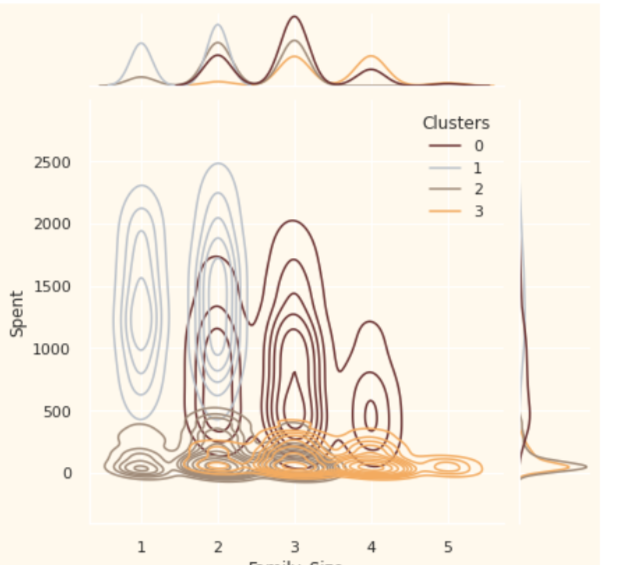
**PROFILING**

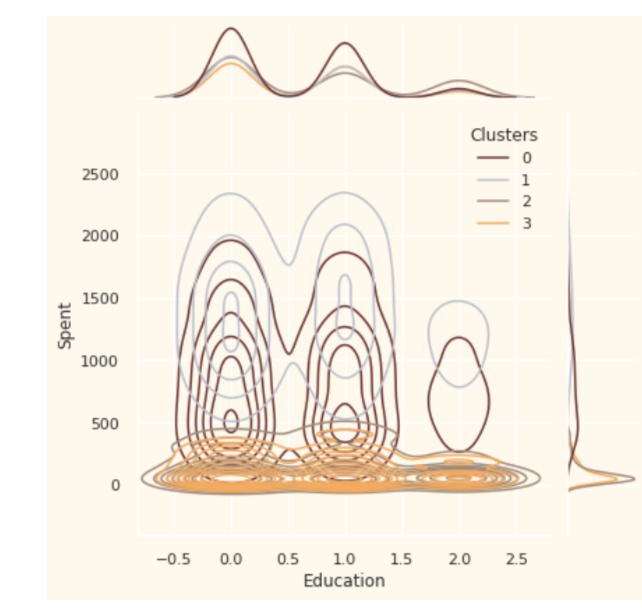
**Let us see who all are there in the clusters.**

Now that we have formed the clusters and looked at their purchasing habits. Let us see who all are there in these clusters. For that, we will be profiling the clusters formed and come to a conclusion about who is our star customer and who needs more attention from the retail store's marketing team.

To decide that I will be plotting some of the features that are indicative of the customer's personal traits in light of the cluster they are in. On the basis of the outcomes, I will be arriving at the conclusions.



****

****

**Profiling the clusters**

**GROUP 0:- High spending and average Income**

1. Are definitely a parent
2. At the max have 4 members in the family and at least 2
3. Single parents are the subset of this group
4. Most have a teenager at home

**GROUP 1:- High spending and High Income**

1. Are definitely not a parent
2. At the max are only 2 members in the family
3. A slight majority of couples over single people
4. Span of all ages
5. A high income groups

**GROUP 2:- Low spending and Low Income**

1. The majority of these people are parents
2. At the max there are 3 members in the family
3. The majority have one kid
4. Relatively younger

**GROUP 3:- High spending and Low Income**

1. They are definitely a parent
2. At the max are 5 members in the family and at least 2
3. Majority of them have a teenager at home
4. Relatively older
5. A lower income group

# **Silhouette Score came out to be 0.789**

# Silhouette analysis refers to a method of interpretation and validation of consistency within clusters of data.The Silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other Clusters (Separation).

**Calculation of Silhouette Value –**If the Silhouette index value is high, the object is well-matched to its own cluster and poorly matched to neighboring clusters. The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient is defined as –

**S(i) = ( b(i) – a(i) ) / ( max { ( a(i), b(i) ) }**

Where,

* a(i) is the average dissimilarity of ith object to all other objects in the same cluster
* b(i) is the average dissimilarity of ith object with all objects in the closest cluster.

**Range of Silhouette Value –**

Now, obviously S(i) will lie between **[-1, 1]** –

1. If silhouette value is close to 1, sample is well-clustered and already assigned to a very appropriate cluster.
2. If silhouette value is about to 0, sample could be assign to another cluster closest to it and the sample lies equally far away from both the clusters. That means it indicates overlapping clusters
3. If silhouette value is close to –1, sample is misclassified and is merely placed somewhere in between the clusters.

**CONCLUSION**

In this project, I performed unsupervised clustering. I did use dimensionality reduction followed by **K-means Clustering and DBScan** . I came up with 4 clusters and further used them in profiling customers in clusters according to their family structures and income/spending. This can be used in planning better marketing strategies.

**HOW PRINCIPAL COMPONENT ANALYSIS**

The main goal of Principal Component Analysis (PCA) is to reduce the dimensionality of a dataset while preserving the most important patterns or relationships between the variables without any prior knowledge of the target variables.

**Steps in PCA**

1. Standardization
2. Covariance matrix Computation
3. Compute Eigenvalues and EigenVectors of the covariance matrix to determine principal components [ |A - lambda x|]
4. Calculate the Explained Variance which is (cumulative Sum of Eigenvalues)/ Sum of Eigenvalues
5. Now find the number of components required for the

n\_components **=** np.argmax(explained\_var >**=** 0.50) **+** 1

And for my data it came out to be **3**

The eigenvectors of the covariance matrix of the data are referred to as the principal axes of the data, and the projection of the data instances onto these principal axes are called the principal components. Dimensionality reduction is then obtained by only retaining those axes (dimensions) that account for the most of the variance, and discarding all others.