Table of Contents

1. INTRODUCTION

The study is an individual assignment to examine a dataset using data mining techniques. For the coursework, customer personality analysis was utilized to examine and comprehend the relationship between the customer and the product sold by the business. The dataset analyses help supermarkets that employ customer personality analysis to understand their customers' requirements and personalities in order to stay up with demand.

1.1. Aim

The aim of the project is to critically analyze a selected dataset and apply data mining techniques to generate a report based on it.

1.2. Objective

- Go Through datasets from open data mining resources.
- Select a suitable dataset.
- Apply data mining techniques on the selected dataset.
- Generate a Report based on the findings and the steps

1.3. Big Data

Big Data is a vast collection of data that is growing at an exponential rate. It is a data collection that is so large and complex that traditional data management solutions are incapable of storing or analyzing it properly. Big data is a term used to describe data that is extraordinarily enormous in size. Volume, Velocity, Variety, Veracity, and Value are the five pillars of big data. (Taylor, 2021).

Big Data is classified into three kinds. There are three types of data: structured data, unstructured data, and semi-structured data. Structured data is information that can be stored, retrieved, and processed in a consistent fashion. Unstructured data is data that has an undefined shape or organization. Semi structured data can include both structured and semi structured data. (Taylor, 2021)

1.4. Data Mining

Data mining is the process of searching through big data sets for patterns and correlations that may be used to address business problems through data analysis. Data mining techniques and technologies help businesses to forecast future trends and make better business decisions (Stedman, n.d.).The key features of data mining are to automatically predict pattern predictions based on trends and analysis , to predict based on likely outcomes, creation of decision oriented information , to focus on large datasets and databases for analysis and clustering based on the findings and visually represent and document it. (The Economic Times, 2019)

2. CASE STUDY

2.1. Overview

A comprehensive investigation of a company's ideal clients is known as customer personality analysis. It enables a company to better understand its consumers and makes it simpler to change goods to meet the individual wants, habits, and concerns of various sorts of customers.

Customer personality analysis enables a company to change its product based on its target customers from various customer categories. Instead of wasting money marketing a new product to every client in the firm's database, a corporation may determine which customer group is most likely to buy the product and then sell the product just to that specific segment.

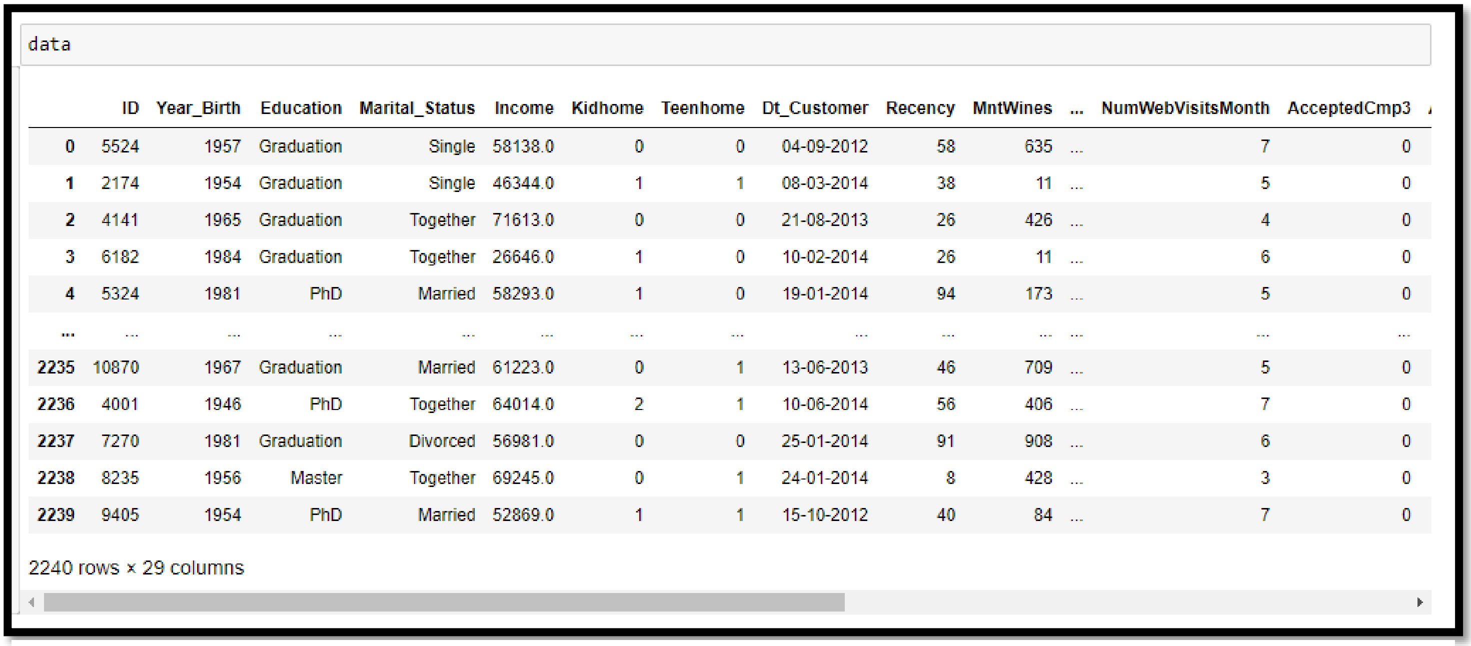
Data mining can be found in a variety of industries and is valuable in a variety of research initiatives, but it is most commonly employed in business to create predictions based on the product or service being sold.

In this coursework, the dataset is utilized with data mining techniques and tools to better understand the customer's personality and demand. The dataset's challenges include the fact that the data is dispersed and has to be aggregated in order to ease data mining visualizations. Furthermore, visualization might be a challenge since modifications to the dataset are required for the visualization to be statistically accepted.

2.2. Dataset

The dataset collected for the coursework is from kaggle and it is to identify the company’s ideal customers through the customer’s personality dataset

The initial dataset had 2240 number of instances and 29 number of attributes.



```
data
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2235	10870	1967	Graduation	Married	61223.0	0	1	13-06-2013	46	709	...	5	0
2236	4001	1946	PhD	Together	64014.0	2	1	10-06-2014	56	406	...	7	0
2237	7270	1981	Graduation	Divorced	56981.0	0	0	25-01-2014	91	908	...	6	0
2238	8235	1956	Master	Together	69245.0	0	1	24-01-2014	8	428	...	3	0
2239	9405	1954	PhD	Married	52869.0	1	1	15-10-2012	40	84	...	7	0

2240 rows × 29 columns

Figure 1the dataset

The attributes of the dataset are:

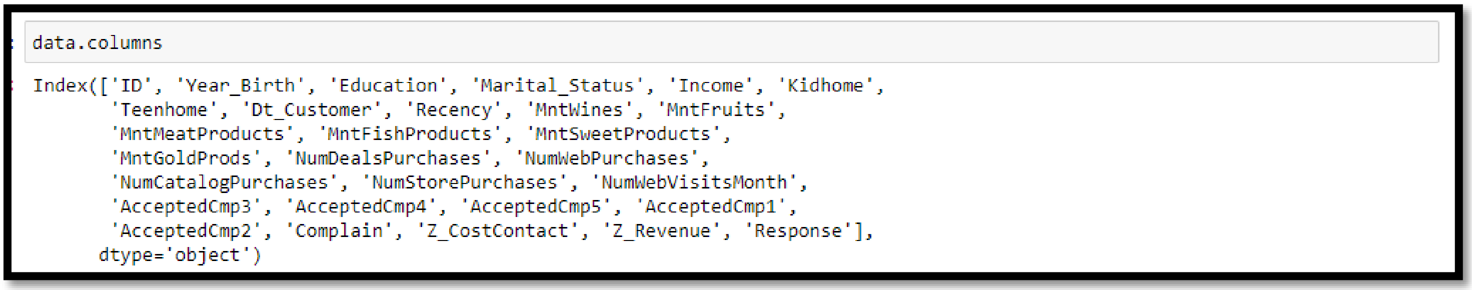
- ID: Customer's unique identifying number
- Year_Birth: Customer's year of birth
- Education: Customer's level of education
- Marital_Status: marital status of the customer
- Income: yearly household income of the customer
- Kidhome: Number of children the customer has
- Teenhome: Number of teenagers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

- MntWines: Total Amount spent on wine in last 2 years
- MntFruits: Total Amount spent on fruits in last 2 years
- MntMeatProducts: Total Amount spent on meat in last 2 years
- MntFishProducts: Total Amount spent on fish in last 2 years
- MntSweetProducts: Total Amount spent on sweets in last 2 years
- MntGoldProds: Total Amount spent on gold in last 2 years

- NumDealsPurchases: Number of purchases made with discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, otherwise 0

- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, otherwise 0
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, otherwise 0
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, otherwise 0
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, otherwise 0
- Response: 1 if customer accepted the offer in the last campaign, otherwise 0

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

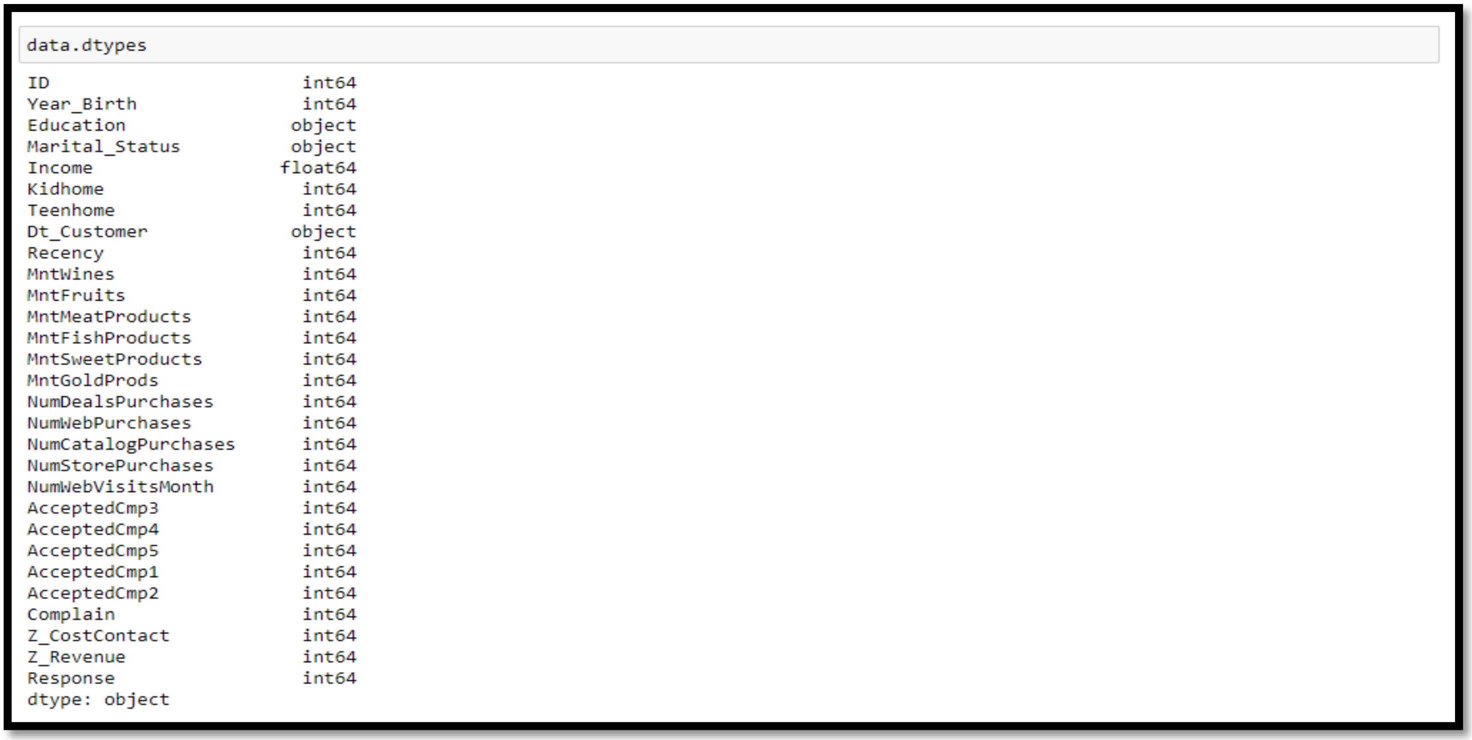


```
data.columns
```

```
Index(['ID', 'Year_Birth', 'Education', 'Marital_Status', 'Income', 'Kidhome',
       'Teenhome', 'Dt_Customer', 'Recency', 'MntWines', 'MntFruits',
       'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts',
       'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases',
       'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth',
       'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1',
       'AcceptedCmp2', 'Complain', 'Z_CostContact', 'Z_Revenue', 'Response'],
      dtype='object')
```

Figure 2 Data columns before data mining

- The data types of the dataset



```
data.dtypes
```

```
ID                       int64
Year_Birth               int64
Education               object
Marital_Status          object
Income                 float64
Kidhome                  int64
Teenhome                 int64
Dt_Customer             object
Recency                  int64
MntWines                 int64
MntFruits                int64
MntMeatProducts          int64
MntFishProducts          int64
MntSweetProducts         int64
MntGoldProds             int64
NumDealsPurchases        int64
NumWebPurchases          int64
NumCatalogPurchases      int64
NumStorePurchases        int64
NumWebVisitsMonth        int64
AcceptedCmp3             int64
AcceptedCmp4             int64
AcceptedCmp5             int64
AcceptedCmp1             int64
AcceptedCmp2             int64
Complain                 int64
Z_CostContact            int64
Z_Revenue                int64
Response                 int64
dtype: object
```

Figure 3 dataset data types

2.3. Data mining tools

Data mining tools are software applications that aid in the development and execution of data mining techniques for the creation and testing of data models.

There are several tools available on the market, both open source and commercial, with differing degrees of effectiveness. At its core, each tool aids in the implementation of a data mining plan, but the distinction is in the amount of complexity you, the software's customer, require (Mayuresh, 2020).

For the coursework Jupyter Notebook IDE (Integrated Development Environment) is used to mine the data in the dataset using python version 3.0.

2.3.1. Jupyter Notebook IDE

Jupyter Notebook is an open-source online application that provides a computing environment that is interactive. It generates papers (notebooks) by combining inputs (code) and outputs into a single file. It provides a single document that includes visualization, mathematical equations, statistical modelling etc.

Jupyter was selected as It follows a single document approach that helps users work more understandable, shareable and repeatable. Jupyter Notebooks support over 40 programming languages and has a major focus on Python. For the coursework Python and its libraries are the main focus for data mining (Wickramasinghe, 2021).

2.3.2. Python

Python is a free and open-source programming language with a short learning curve. Python is an excellent tool for enterprises who want the software they use to be custom designed to their requirements because of its capacity as a general-purpose language and its big library of packages that assist develop a system for generating data models from scratch.

Python provides the flexibility for anyone to pick up and construct their own environment using graphical interfaces of their choice. Python is also supported by a strong online community of package authors who guarantee the packages available are sturdy and safe. Python is well-known for one of its characteristics is powerful on the fly visualization features it offers (Mayuresh, 2020).

Python was selected due to simplicity in its code and the various libraries present in it which makes data mining and visualization easy to use and understand and due to the code modifications are further possible

2.3.3. Libraries

- Pandas

  Pandas is an analysis software package used for data visualization and created for the Python computer language. It helps manipulating numerical tables and time series in particular.
- Seaborn

  Seaborn is a free and open-source Python module based on matplotlib. It is used for exploratory data analysis and data visualization. Seaborn is simple to use with dataframes and the Pandas library. The generated graphs can also be readily altered. The following are some of the advantages of data visualization.
- Matplotlib.pyplot

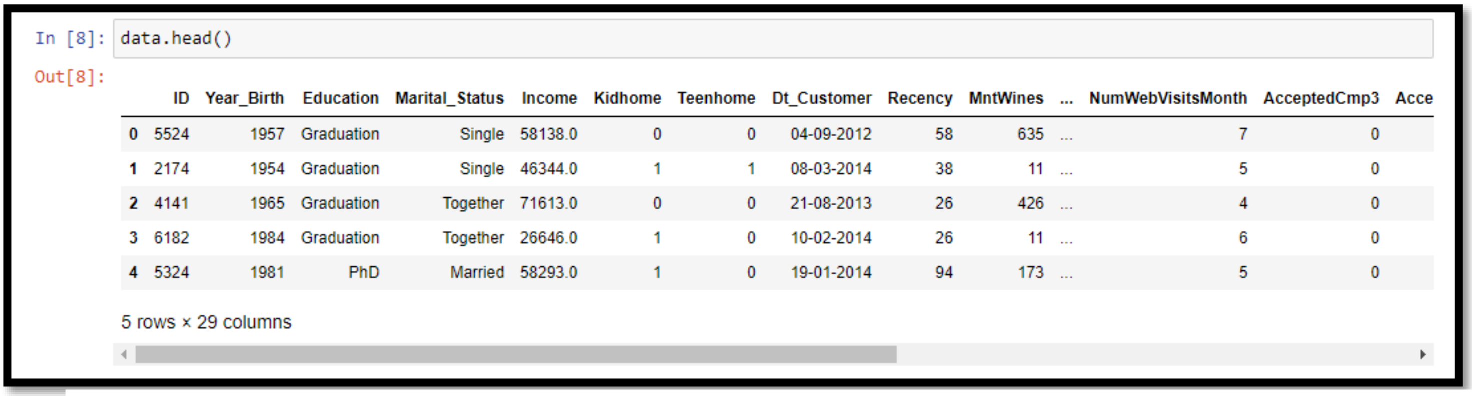
  Matplotlib is a data visualization and graphical plotting package for Python and its numerical extension NumPy that is cross-platform. As such, it provides an open source alternative to MATLAB. Developers may also incorporate plots in GUI programs by using matplotlib's APIs (Application Programming Interfaces).

2.4. Data mining techniques

2.4.1. Data Cleaning and Preparation

The first data mining technique is to clean and prepare the data as the dataset collected is a raw dataset and needs to be formatted for analytical purposes. This step is a very important step and as without it the quality of the data can be unreliable or meaningless.

- First the data is viewed with the head() function as shown in figure 4:



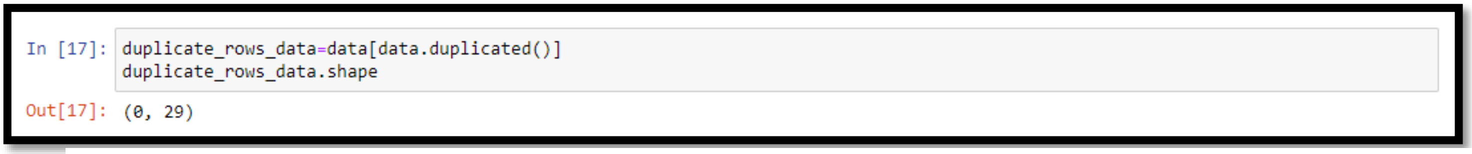
```
In [8]: data.head()
Out[8]:
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	Acce
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0	
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0	
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0	
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0	
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0	

5 rows × 29 columns

Figure 4 data. Head ()

- Identify and delete any duplicates in the dataset.



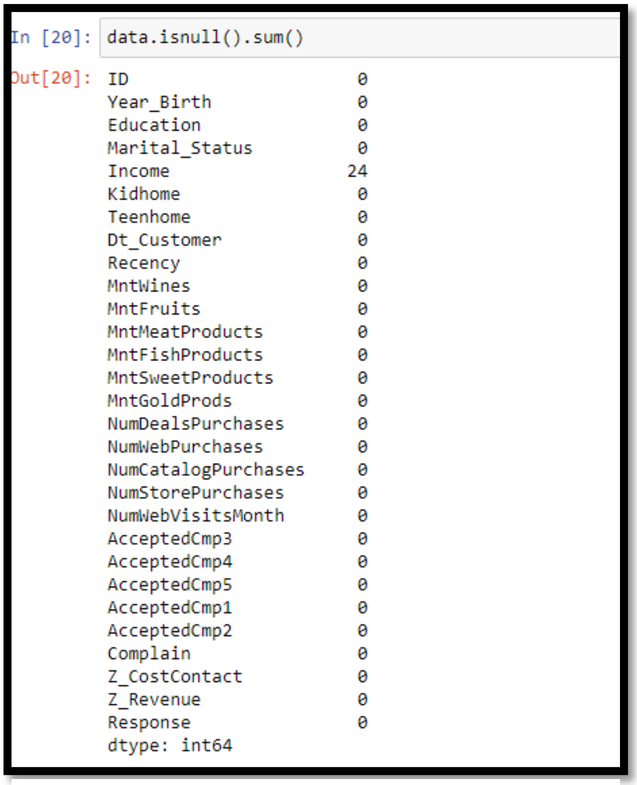
```
In [17]: duplicate_rows_data=data[data.duplicated()]
         duplicate_rows_data.shape

Out[17]: (0, 29)
```

Figure 5 identifying duplicates in the dataset

The dataset has no duplicate values and thus there is no need to delete any redundant or duplicate values from the dataset

- Identifying null or empty values

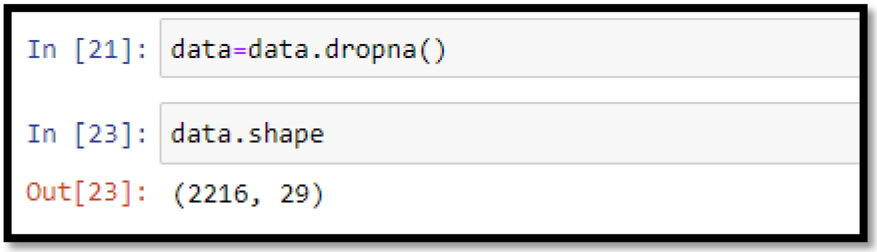


```
In [20]: data.isnull().sum()

Out[20]: ID                      0
         Year_Birth              0
         Education               0
         Marital_Status          0
         Income                 24
         Kidhome                 0
         Teenhome                0
         Dt_Customer             0
         Recency                 0
         MntWines                0
         MntFruits               0
         MntMeatProducts         0
         MntFishProducts         0
         MntSweetProducts        0
         MntGoldProds            0
         NumDealsPurchases       0
         NumWebPurchases         0
         NumCatalogPurchases     0
         NumStorePurchases       0
         NumWebVisitsMonth       0
         AcceptedCmp3            0
         AcceptedCmp4            0
         AcceptedCmp5            0
         AcceptedCmp1            0
         AcceptedCmp2            0
         Complain                0
         Z_CostContact           0
         Z_Revenue               0
         Response                0
         dtype: int64
```

Figure 6 null values in columns

There only 24 null values in the income column which will be removed as shown in figure 7



```
In [21]: data=data.dropna()

In [23]: data.shape

Out[23]: (2216, 29)
```

Figure 7 dropping null values and data shape

Now the dataset has 2216 rows or instances and has 29 rows.

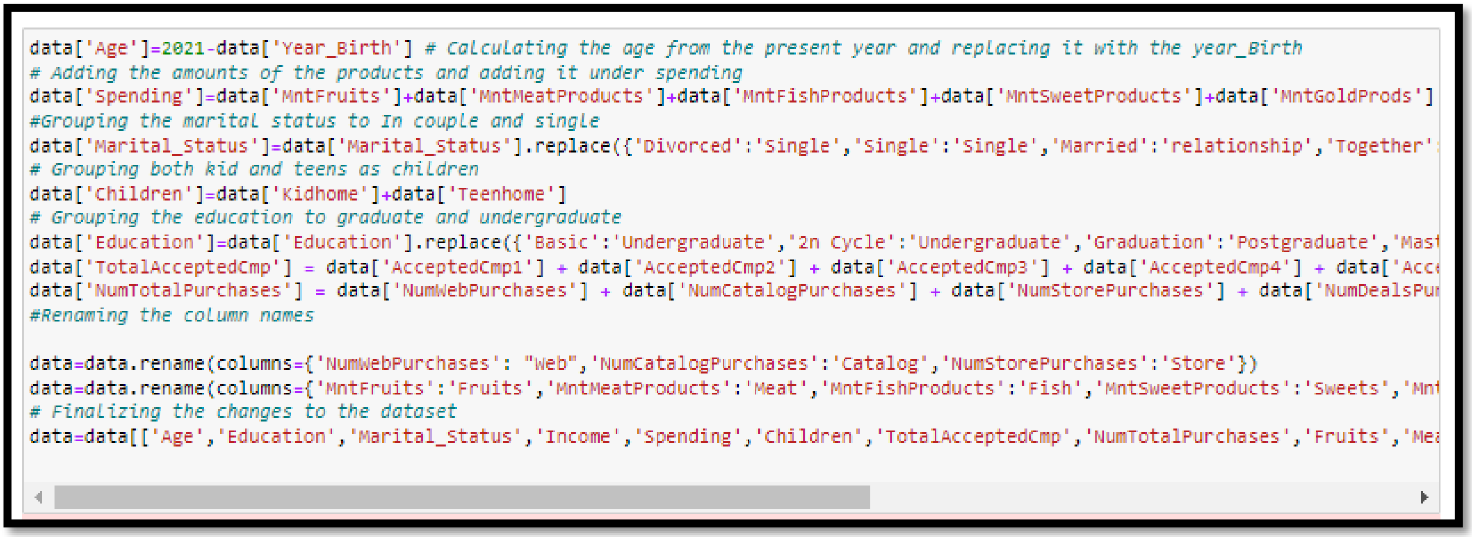
- Feature extraction.

  In this section we identify columns and make changes to them like data type, name etc. and also delete any unnecessary columns.

  When analysing the dataset it was noticed that there were many similar columns that could be 'grouped' and that certain data in the dataset can be simplified. The Year_Birth column can be converted to age instead of the year.

  The following is the code used for feature extraction.

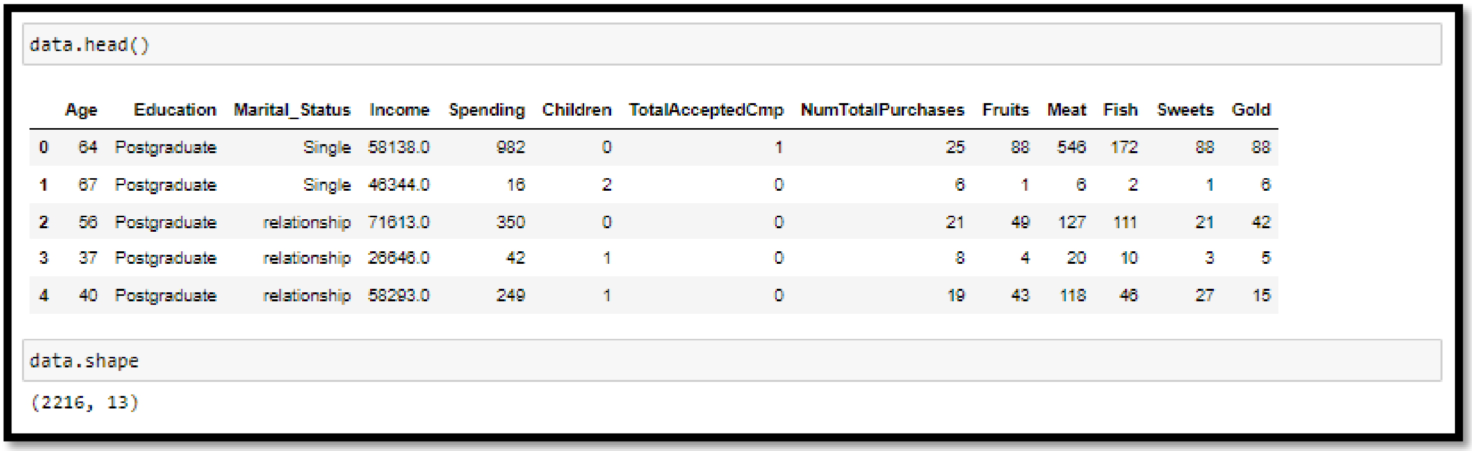
After all the data preparing and cleaning the dataset has now been simplified to 2216 instances and 13 columns or attributes which can be used for data mining.



```
data['Age']=2021-data['Year_Birth'] # Calculating the age from the present year and replacing it with the year_Birth
# Adding the amounts of the products and adding it under spending
data['Spending']=data['MntFruits']+data['MntMeatProducts']+data['MntFishProducts']+data['MntSweetProducts']+data['MntGoldProds']
#Grouping the marital status to In couple and single
data['Marital_Status']=data['Marital_Status'].replace({'Divorced':'Single','Single':'Single','Married':'relationship','Together':
# Grouping both kid and teens as children
data['Children']=data['Kidhome']+data['Teenhome']
# Grouping the education to graduate and undergraduate
data['Education']=data['Education'].replace({'Basic':'Undergraduate','2n Cycle':'Undergraduate','Graduation':'Postgraduate','Mast
data['TotalAcceptedCmp'] = data['AcceptedCmp1'] + data['AcceptedCmp2'] + data['AcceptedCmp3'] + data['AcceptedCmp4'] + data['Acce
data['NumTotalPurchases'] = data['NumWebPurchases'] + data['NumCatalogPurchases'] + data['NumStorePurchases'] + data['NumDealsPur
#Renaming the column names

data=data.rename(columns={'NumWebPurchases': "Web",'NumCatalogPurchases':'Catalog','NumStorePurchases':'Store'})
data=data.rename(columns={'MntFruits':'Fruits','MntMeatProducts':'Meat','MntFishProducts':'Fish','MntSweetProducts':'Sweets','Mnt
# Finalizing the changes to the dataset
data=data[['Age','Education','Marital_Status','Income','Spending','Children','TotalAcceptedCmp','NumTotalPurchases','Fruits','Mea
```

Figure 8 feature extraction



```
data.head()
```

	Age	Education	Marital_Status	Income	Spending	Children	TotalAcceptedCmp	NumTotalPurchases	Fruits	Meat	Fish	Sweets	Gold
0	64	Postgraduate	Single	58138.0	982	0	1	25	88	546	172	88	88
1	67	Postgraduate	Single	46344.0	16	2	0	6	1	6	2	1	6
2	56	Postgraduate	relationship	71613.0	350	0	0	21	49	127	111	21	42
3	37	Postgraduate	relationship	26646.0	42	1	0	8	4	20	10	3	5
4	40	Postgraduate	relationship	58293.0	249	1	0	19	43	118	46	27	15

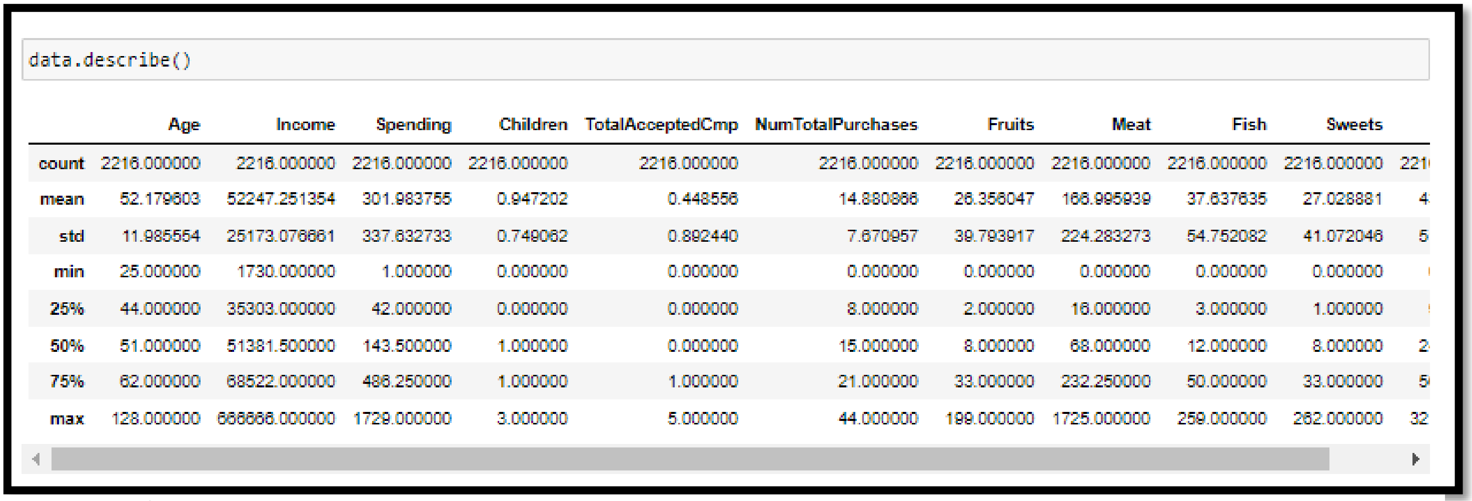
```
data.shape
```

```
(2216, 13)
```

Figure 4 Data after preprocessing

2.4.2. Identifying patterns

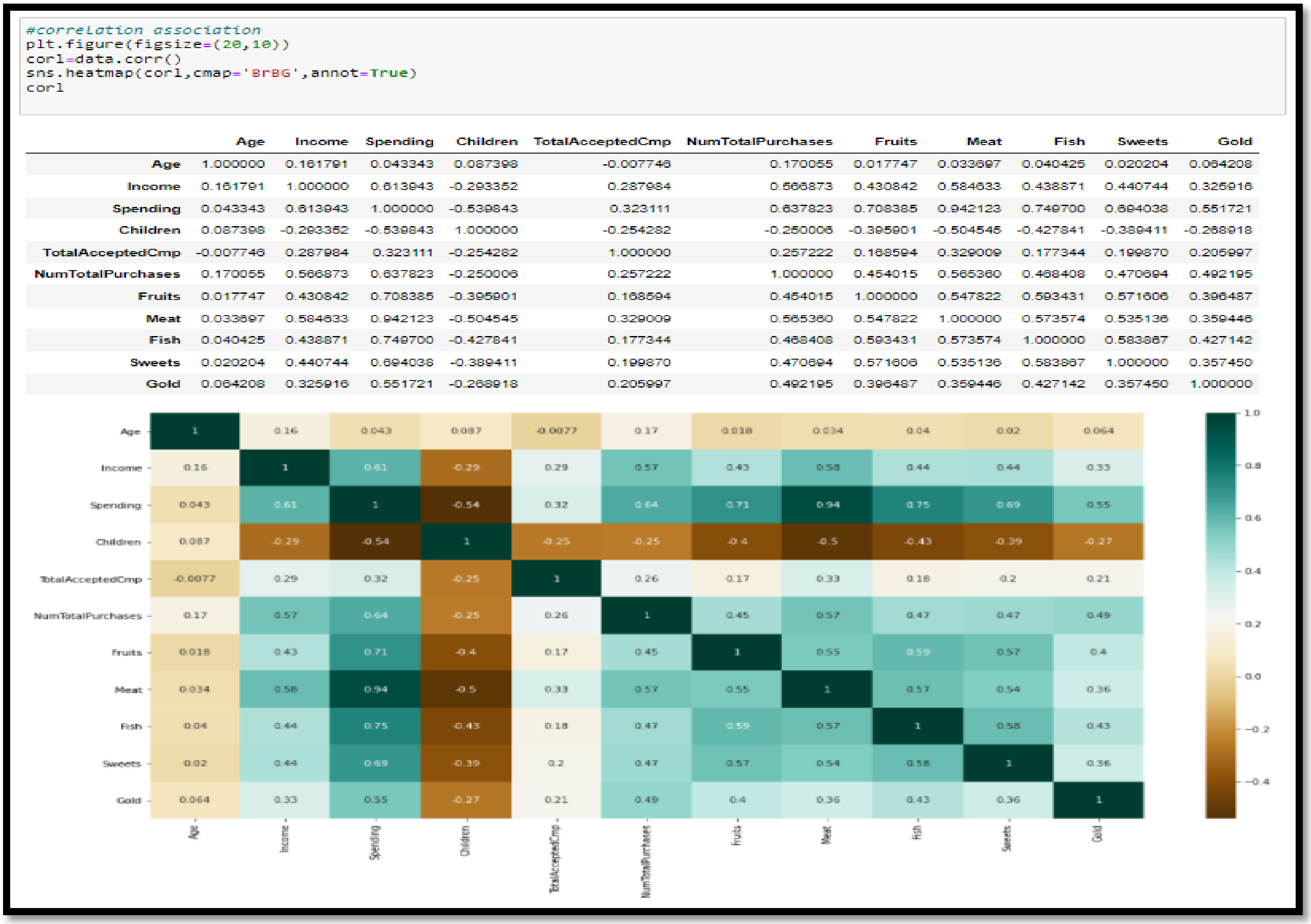
In this section we use visualization to identify any patterns in the new dataset that can further help in classification or clustering.



```
data.describe()
```

```
              Age          Income     Spending     Children  TotalAcceptedCmp  NumTotalPurchases       Fruits         Meat         Fish       Sweets
count  2216.000000    2216.000000  2216.000000  2216.000000       2216.000000        2216.000000  2216.000000  2216.000000  2216.000000  2216.000000  221
mean     52.179603   52247.251354   301.983755     0.947202          0.448556          14.880866    26.356047   166.995939    37.637635    27.028881    4
 std     11.985554   25173.076661   337.632733     0.749062          0.892440           7.670957    39.793917   224.283273    54.752082    41.072046    5
 min     25.000000    1730.000000     1.000000     0.000000          0.000000           0.000000     0.000000     0.000000     0.000000     0.000000
 25%     44.000000   35303.000000    42.000000     0.000000          0.000000           8.000000     2.000000    16.000000     3.000000     1.000000
 50%     51.000000   51381.500000   143.500000     1.000000          0.000000          15.000000     8.000000    68.000000    12.000000     8.000000    2
 75%     62.000000   68522.000000   486.250000     1.000000          1.000000          21.000000    33.000000   232.250000    50.000000    33.000000    5
 max    128.000000  666666.000000  1729.000000     3.000000          5.000000          44.000000   199.000000  1725.000000   259.000000   262.000000   32
```

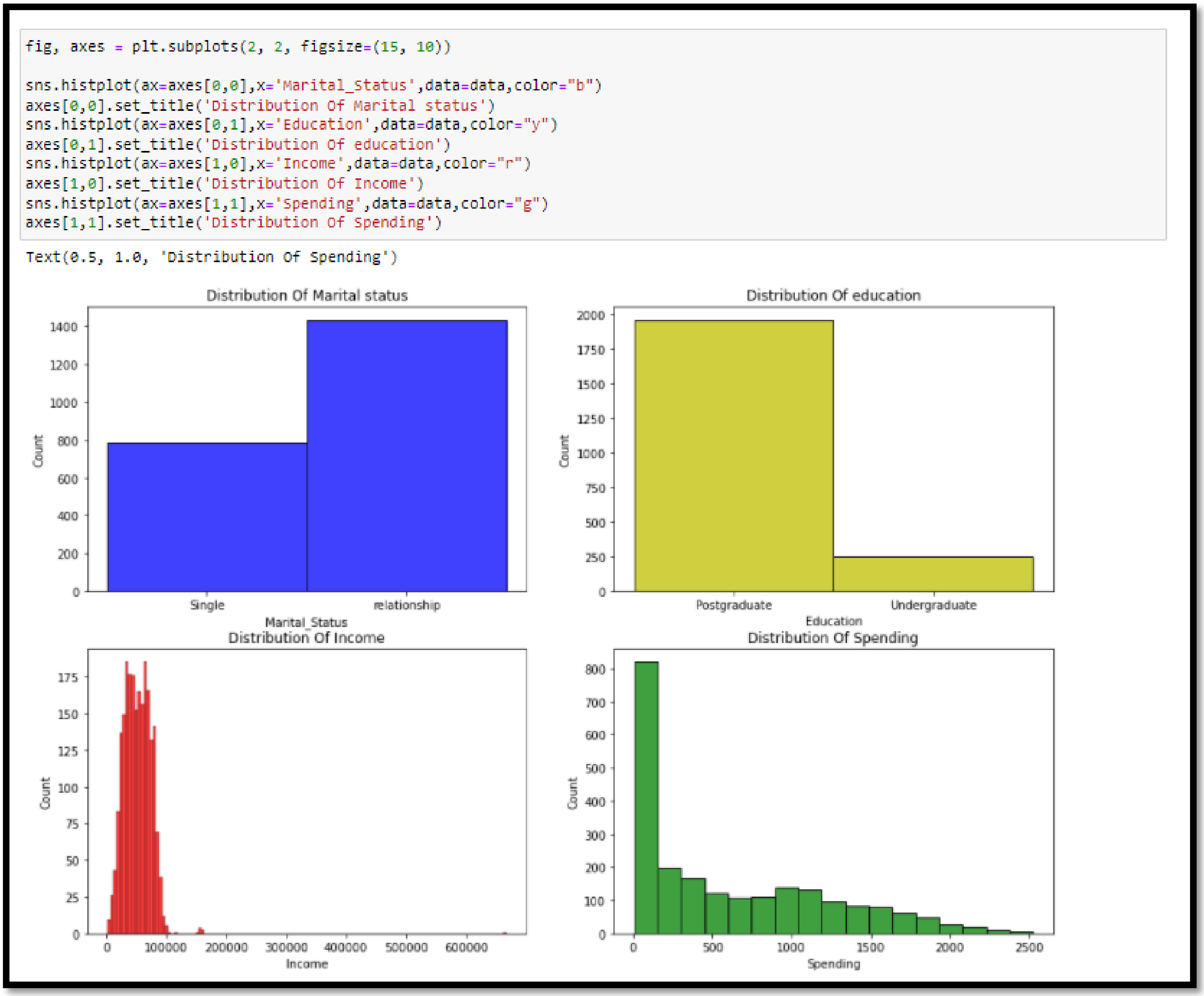
Figure 9 Data description



```
#correlation association
plt.figure(figsize=(20,10))
corl=data.corr()
sns.heatmap(corl,cmap='BrBG',annot=True)
corl
```

```
                        Age    Income  Spending  Children  TotalAcceptedCmp  NumTotalPurchases    Fruits      Meat      Fish    Sweets      Gold
              Age  1.000000  0.161791  0.043343  0.087398         -0.007746           0.170055  0.017747  0.033697  0.040425  0.020204  0.064208
           Income  0.161791  1.000000  0.613943 -0.293352          0.287984           0.566873  0.430842  0.584633  0.438871  0.440744  0.325916
         Spending  0.043343  0.613943  1.000000 -0.539843          0.323111           0.637823  0.708385  0.942123  0.749700  0.694038  0.551721
         Children  0.087398 -0.293352 -0.539843  1.000000         -0.254282          -0.250006 -0.395901 -0.504545 -0.427841 -0.389411 -0.268918
 TotalAcceptedCmp -0.007746  0.287984  0.323111 -0.254282          1.000000           0.257222  0.168594  0.329009  0.177344  0.199870  0.205997
NumTotalPurchases  0.170055  0.566873  0.637823 -0.250006          0.257222           1.000000  0.454015  0.565360  0.468408  0.470694  0.492195
           Fruits  0.017747  0.430842  0.708385 -0.395901          0.168594           0.454015  1.000000  0.547822  0.593431  0.571606  0.396487
             Meat  0.033697  0.584633  0.942123 -0.504545          0.329009           0.565360  0.547822  1.000000  0.573574  0.535136  0.359446
             Fish  0.040425  0.438871  0.749700 -0.427841          0.177344           0.468408  0.593431  0.573574  1.000000  0.583867  0.427142
           Sweets  0.020204  0.440744  0.694038 -0.389411          0.199870           0.470694  0.571606  0.535136  0.583867  1.000000  0.357450
             Gold  0.064208  0.325916  0.551721 -0.268918          0.205997           0.492195  0.396487  0.359446  0.427142  0.357450  1.000000
```

Figure 10 correlation diagram

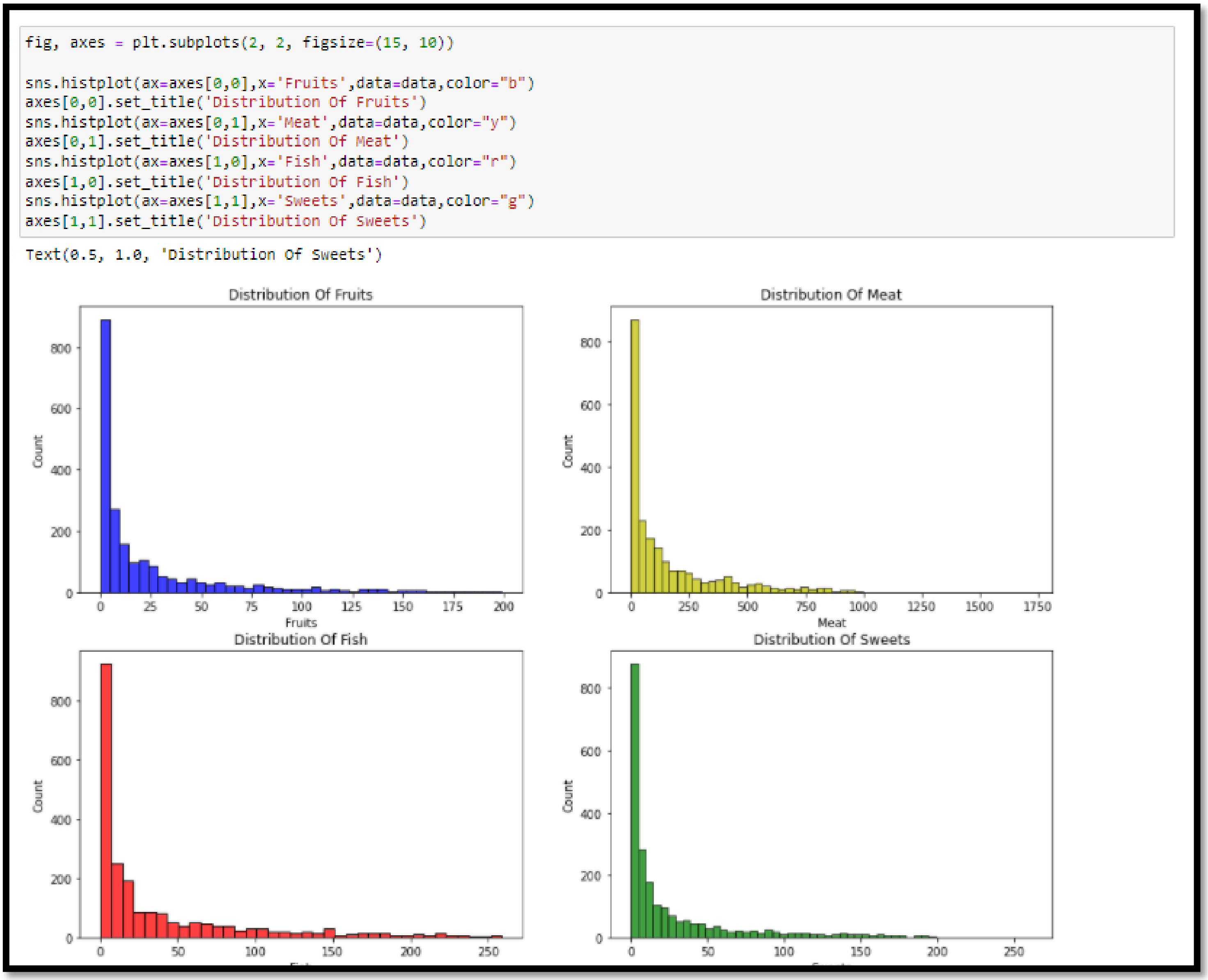


```
fig, axes = plt.subplots(2, 2, figsize=(15, 10))

sns.histplot(ax=axes[0,0],x='Marital_Status',data=data,color="b")
axes[0,0].set_title('Distribution Of Marital status')
sns.histplot(ax=axes[0,1],x='Education',data=data,color="y")
axes[0,1].set_title('Distribution Of education')
sns.histplot(ax=axes[1,0],x='Income',data=data,color="r")
axes[1,0].set_title('Distribution Of Income')
sns.histplot(ax=axes[1,1],x='Spending',data=data,color="g")
axes[1,1].set_title('Distribution Of Spending')
```

```
Text(0.5, 1.0, 'Distribution Of Spending')
```

Figure 11 Distribution diagram of education, spending ,Income and Spending

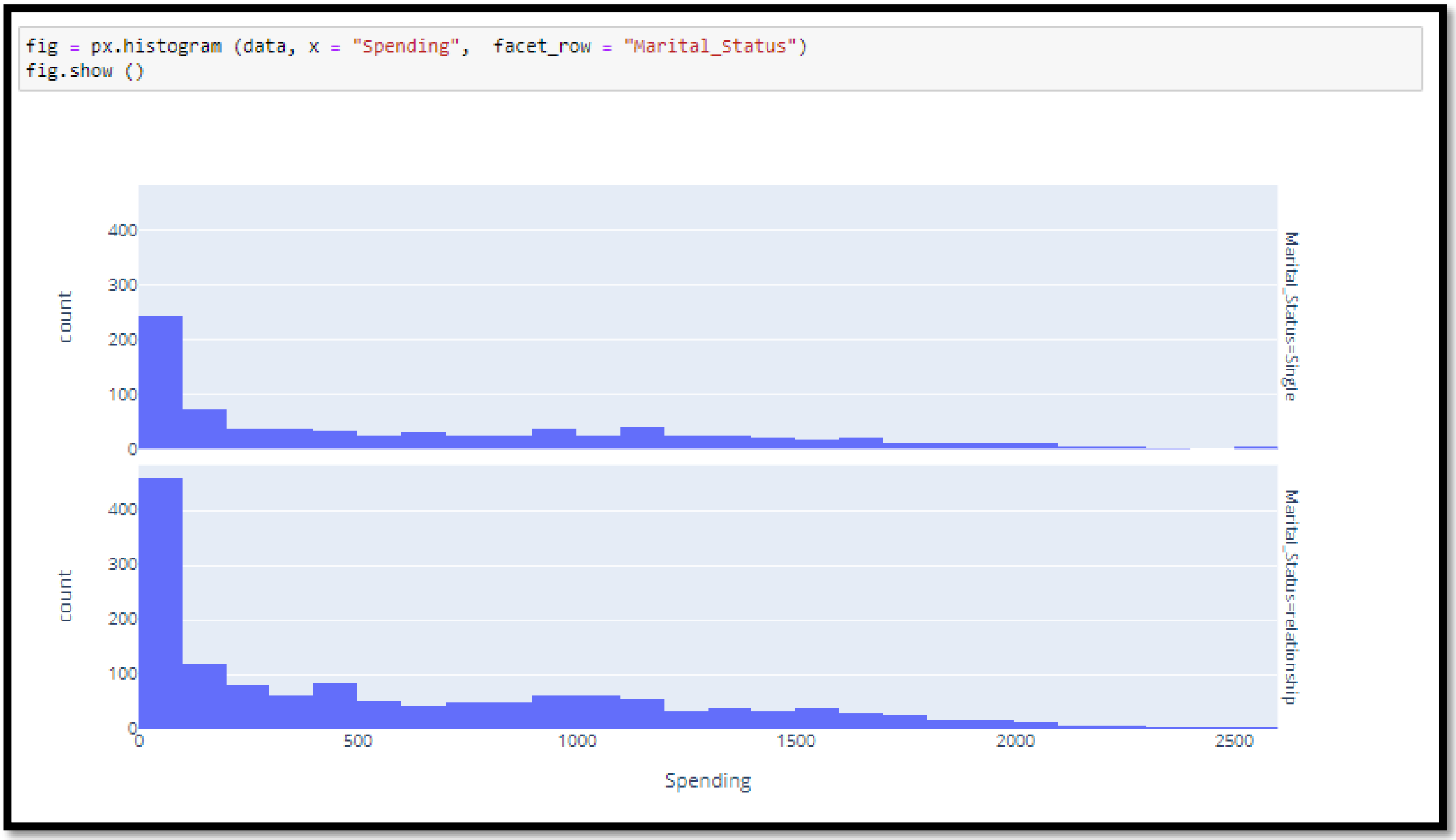


```
fig, axes = plt.subplots(2, 2, figsize=(15, 10))

sns.histplot(ax=axes[0,0],x='Fruits',data=data,color="b")
axes[0,0].set_title('Distribution Of Fruits')
sns.histplot(ax=axes[0,1],x='Meat',data=data,color="y")
axes[0,1].set_title('Distribution Of Meat')
sns.histplot(ax=axes[1,0],x='Fish',data=data,color="r")
axes[1,0].set_title('Distribution Of Fish')
sns.histplot(ax=axes[1,1],x='Sweets',data=data,color="g")
axes[1,1].set_title('Distribution Of Sweets')
```

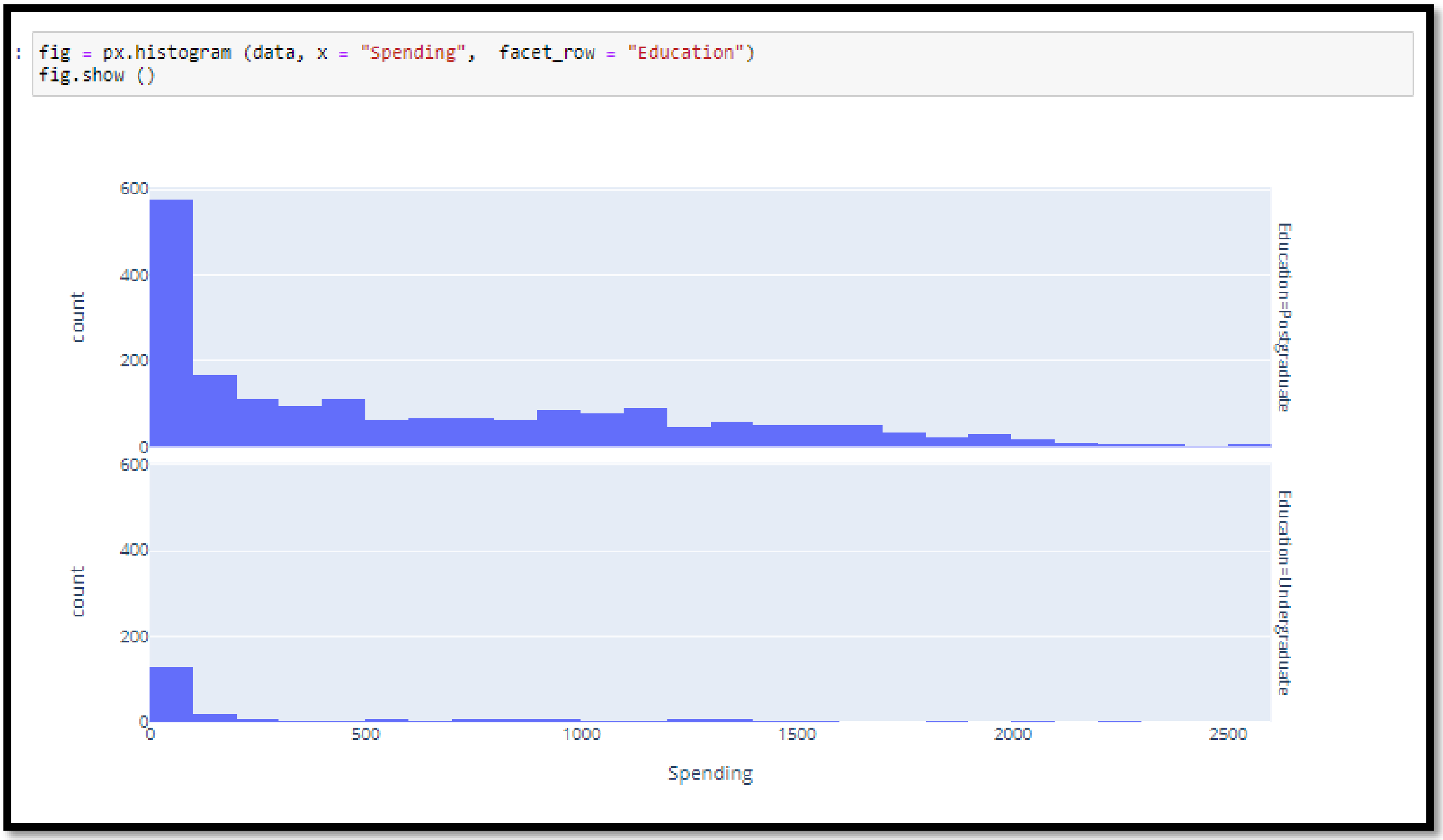
```
Text(0.5, 1.0, 'Distribution Of Sweets')
```

Figure 12 Distribution of fruits, meat, fish, and sweets



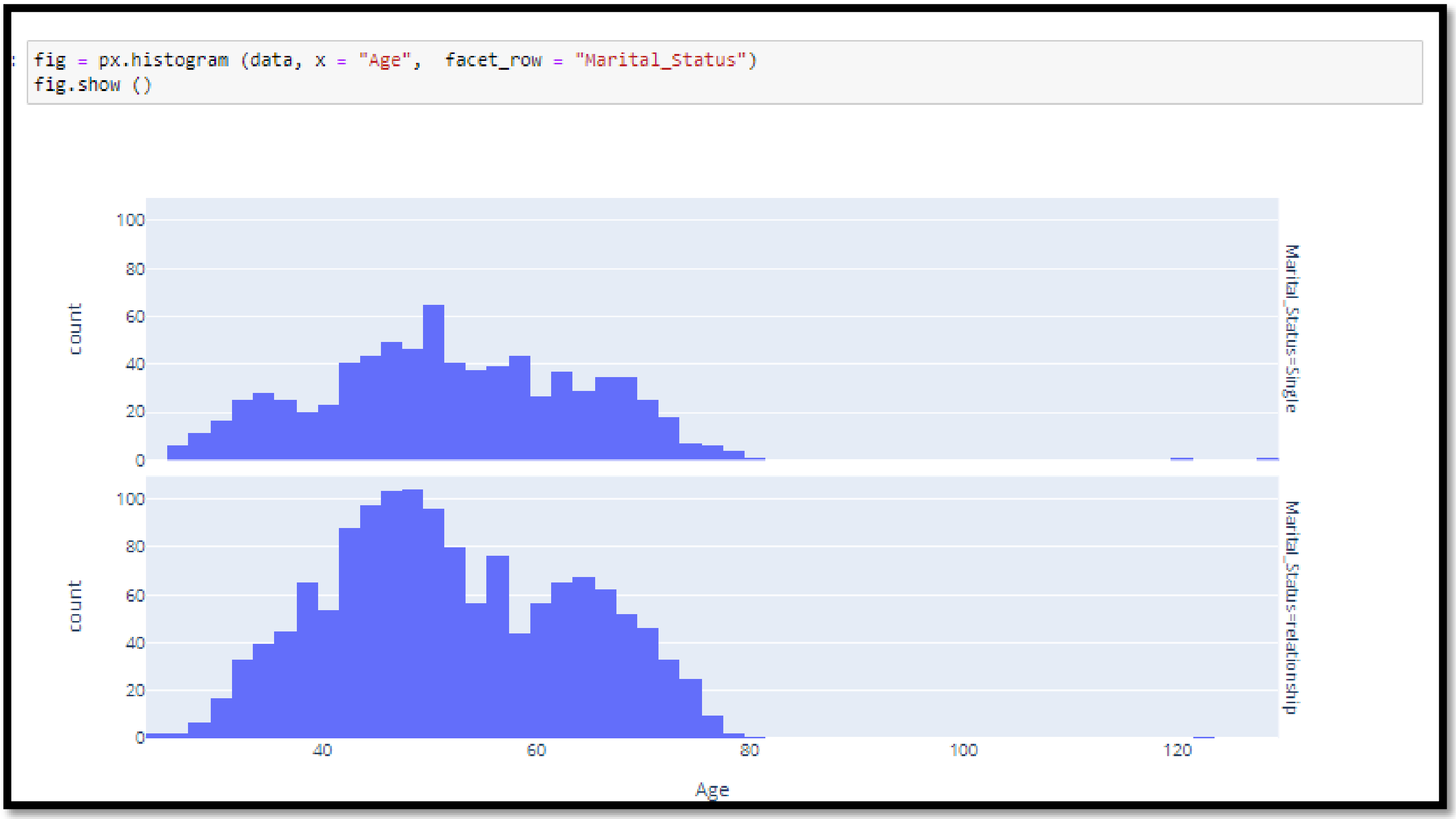
```
fig = px.histogram (data, x = "Spending",  facet_row = "Marital_Status")
fig.show ()
```

Figure 13 Relationship between Education and Marital status



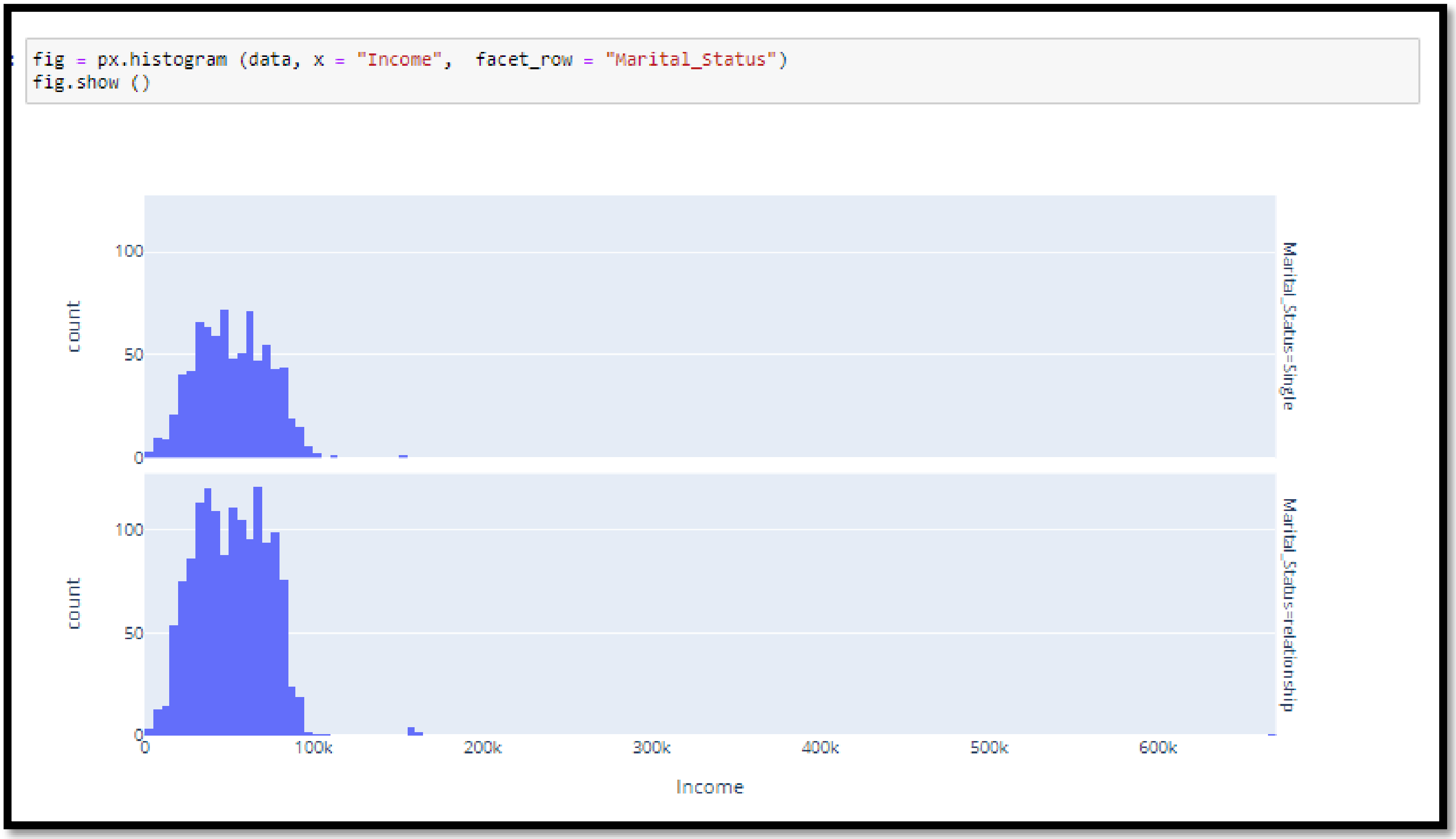
```
fig = px.histogram (data, x = "Spending",  facet_row = "Education")
fig.show ()
```

Figure 14Distribution between education and spending



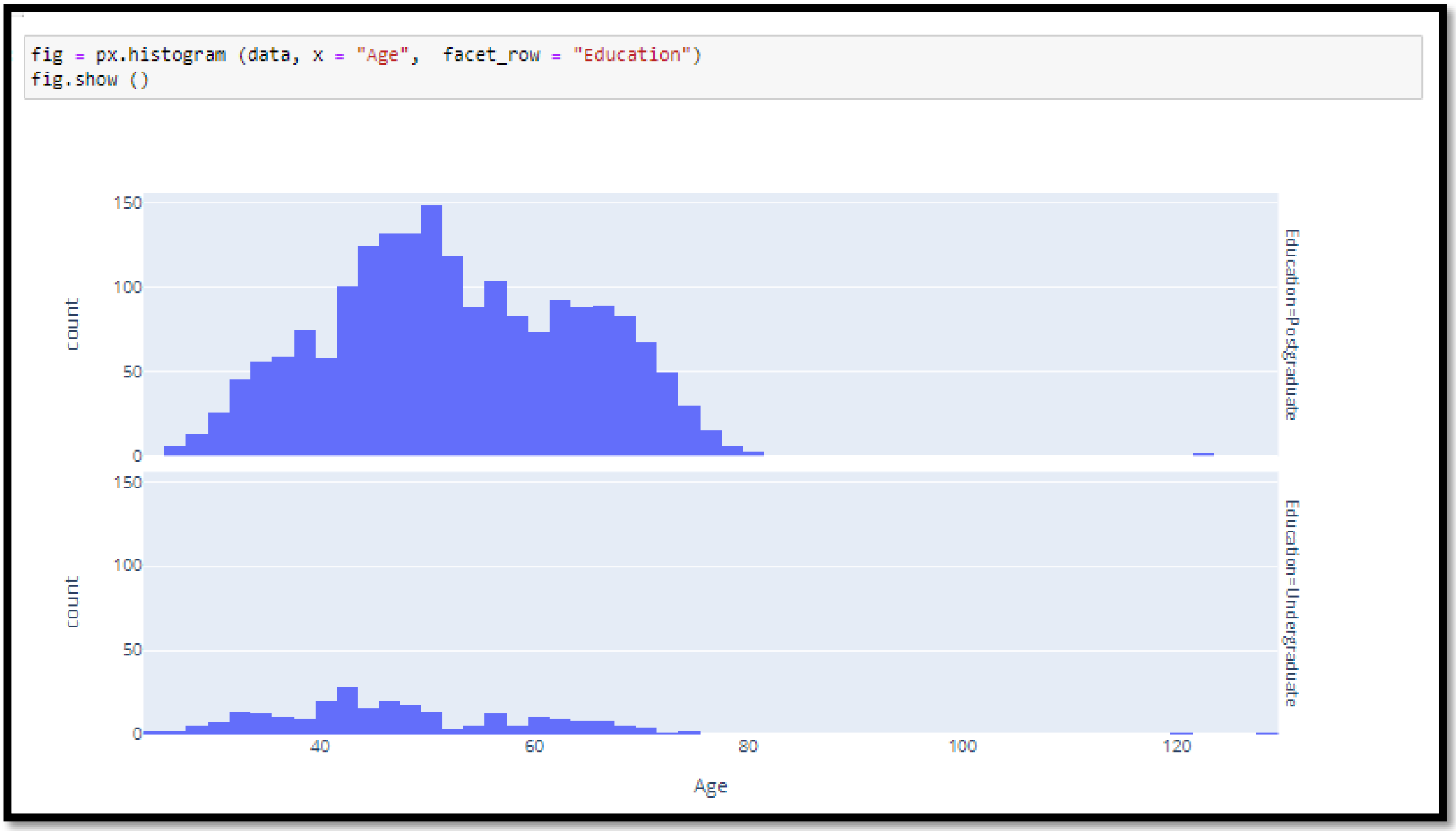
```
fig = px.histogram (data, x = "Age",  facet_row = "Marital_Status")
fig.show ()
```

Figure 15 Distribution between Age and Marital Status



```
fig = px.histogram (data, x = "Income",  facet_row = "Marital_Status")
fig.show ()
```

Figure 16 Distribution between Income and Marital status



```
fig = px.histogram (data, x = "Age",  facet_row = "Education")
fig.show ()
```

Figure 17 Distribution between Age and education

Most of the data relationship and correlation was easily understandable from the correlation matrix diagram but from the above mentioned diagrams we can see patterns such as strong relationship between spending in meat and fish while also that graduates spend more when compared to undergraduates. Moreover the number of children doesn’t affect the purchases. This also highlights that more adults purchase goods when compared to other age groups. Further data analysis will be mentioned in the Dataset analysis summary section.

2.4.3. Classification

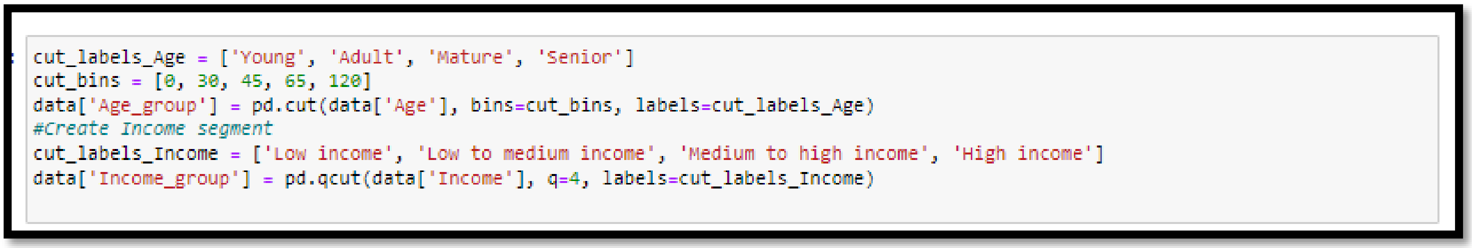
For the dataset classification will be done based on the income where we classify the customers to four categories. They are:

- Low income
- Low to medium income
- Medium to high income
- High income

This was classified by coding and dividing the income into four parts.

The dataset was also classified Age based of four categories which are:

- Young
- Adult
- Mature
- Senior



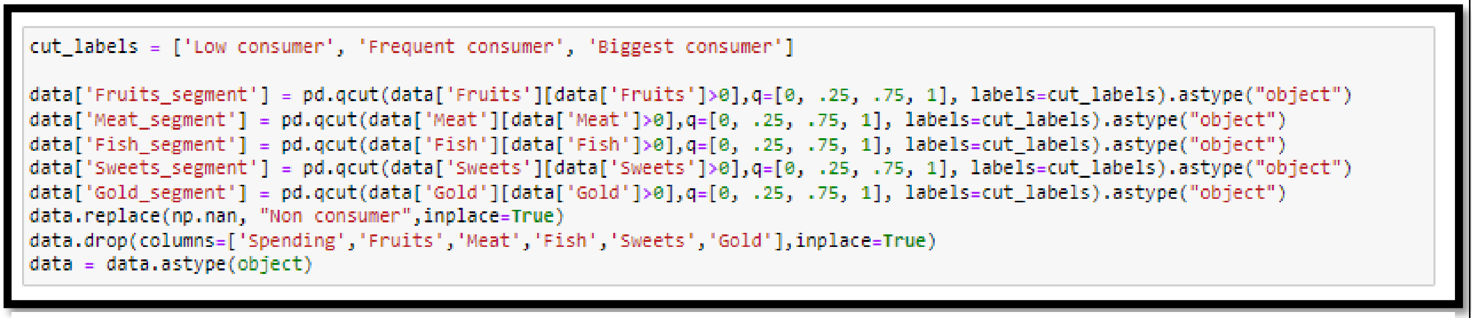
```
cut_labels_Age = ['Young', 'Adult', 'Mature', 'Senior']
cut_bins = [0, 30, 45, 65, 120]
data['Age_group'] = pd.cut(data['Age'], bins=cut_bins, labels=cut_labels_Age)
#Create Income segment
cut_labels_Income = ['Low income', 'Low to medium income', 'Medium to high income', 'High income']
data['Income_group'] = pd.qcut(data['Income'], q=4, labels=cut_labels_Income)
```

Figure 18 Classification of Age and Income

All the above classifications were done mainly to classify which customer buys which category of the product from the business more. Thus, the purchase of products in fruits, meat, Fish, sweets and gold were further classified into:

- Low consumer
- Frequent consumer
- Biggest consumer
- Non Consumer

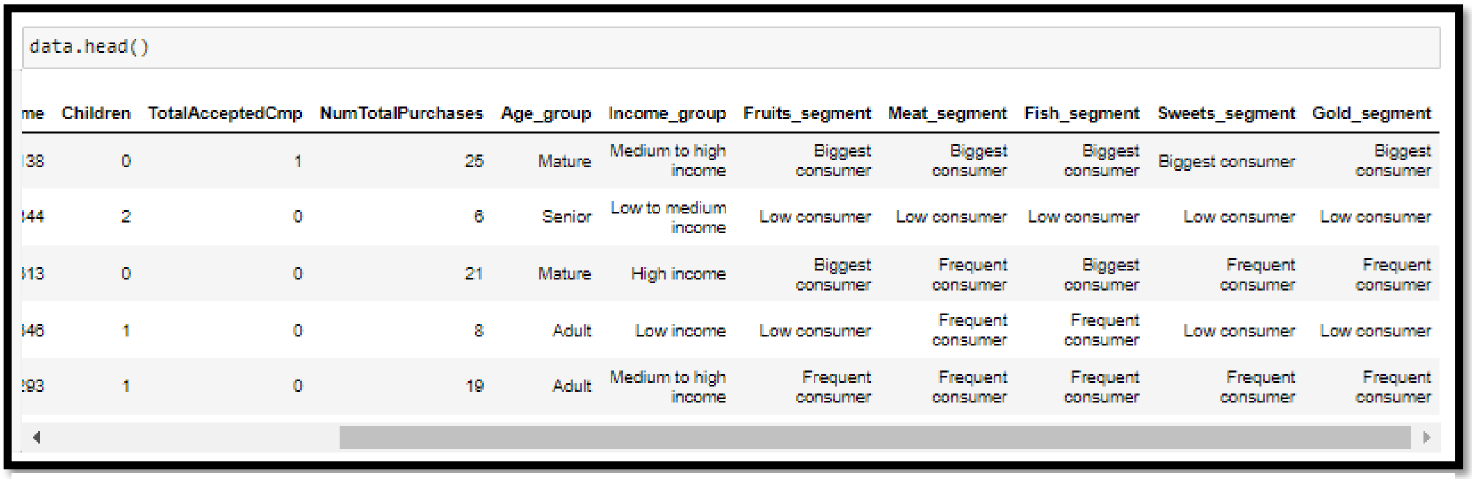
Through the final classification we were able to understand that meat is the highest purchase among the products and has no Non Consumer and the least purchases were made in fruits.



```
cut_labels = ['Low consumer', 'Frequent consumer', 'Biggest consumer']

data['Fruits_segment'] = pd.qcut(data['Fruits'][data['Fruits']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")
data['Meat_segment'] = pd.qcut(data['Meat'][data['Meat']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")
data['Fish_segment'] = pd.qcut(data['Fish'][data['Fish']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")
data['Sweets_segment'] = pd.qcut(data['Sweets'][data['Sweets']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")
data['Gold_segment'] = pd.qcut(data['Gold'][data['Gold']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")
data.replace(np.nan, "Non consumer",inplace=True)
data.drop(columns=['Spending','Fruits','Meat','Fish','Sweets','Gold'],inplace=True)
data = data.astype(object)
```

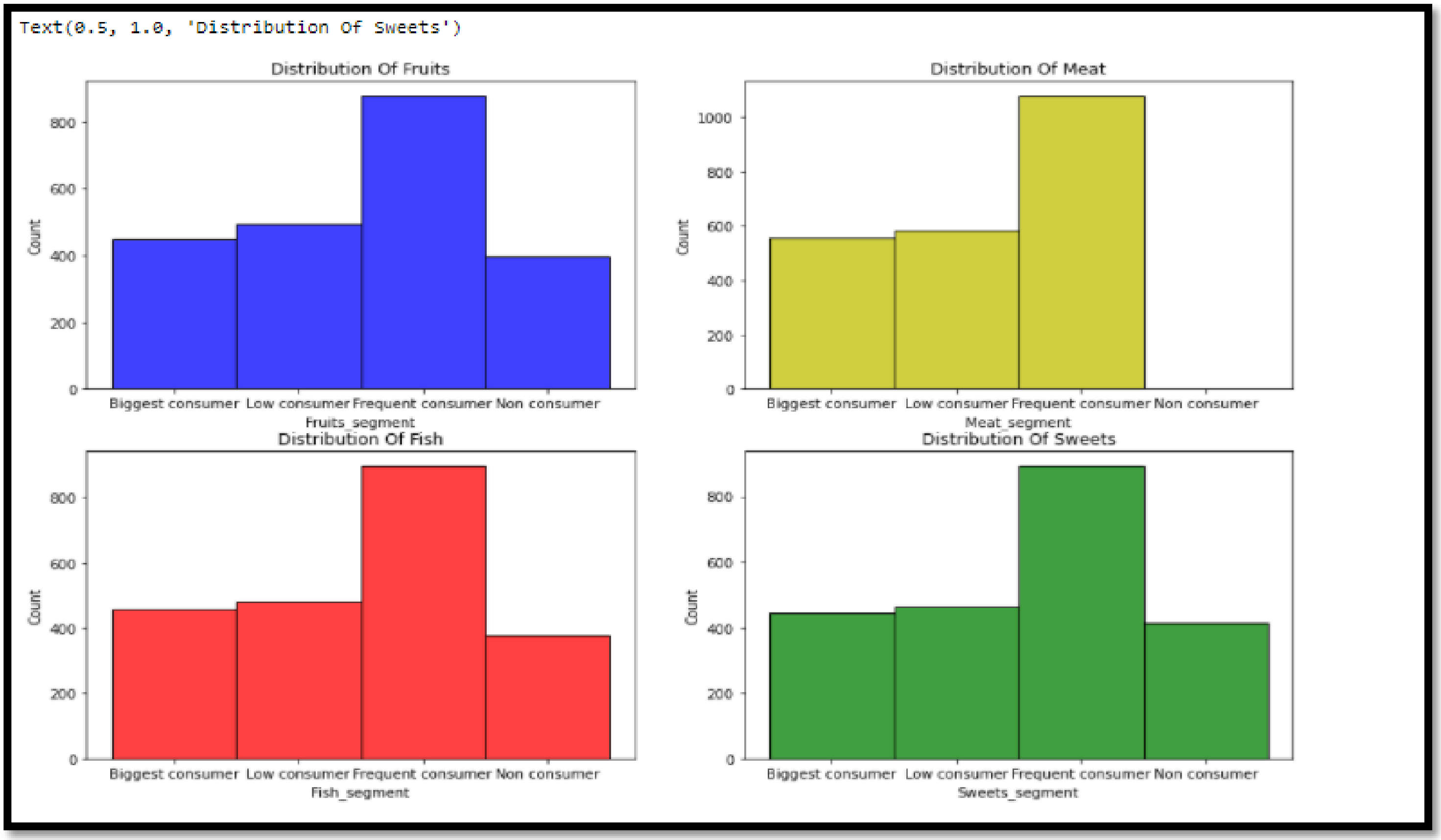
Figure 19 Classification of the products based on the consumer



```
data.head()
```

me	Children	TotalAcceptedCmp	NumTotalPurchases	Age_group	Income_group	Fruits_segment	Meat_segment	Fish_segment	Sweets_segment	Gold_segment
38	0	1	25	Mature	Medium to high income	Biggest consumer	Biggest consumer	Biggest consumer	Biggest consumer	Biggest consumer
44	2	0	6	Senior	Low to medium income	Low consumer	Low consumer	Low consumer	Low consumer	Low consumer
13	0	0	21	Mature	High income	Biggest consumer	Frequent consumer	Biggest consumer	Frequent consumer	Frequent consumer
46	1	0	8	Adult	Low income	Low consumer	Frequent consumer	Frequent consumer	Low consumer	Low consumer
93	1	0	19	Adult	Medium to high income	Frequent consumer	Frequent consumer	Frequent consumer	Frequent consumer	Frequent consumer

Figure 20 Dataset after classification



```
Text(0.5, 1.0, 'Distribution Of Sweets')
```

Figure 21 Distribution of classification of fruits, meat fish and sweets

2.5. Dataset Analysis Summary

The dataset collected was a raw dataset that was pre-processed and prepared for further data analysis or data mining. From trying to understand the patterns, most of the correlation and relationship between the products and spending were easily understood from the correlation diagram in figure 10. Through the dataset and other diagrams for identifying any more patterns, it was noticed that there is a strong connection between the income of the customer and spending and that the higher the income of the customer, the higher the spending from them. There is also a correlation between the education of the customer and their spending. From the patterns, it’s understood that customers who are graduates purchase more from the company when compared to undergraduates. Thus, based on the income and spending habits of the user, further classification of the data can be done. There is also a good correlation between the customers and the products sold, such as fruits, meat, fish, sweets, etc. From the patterns, it’s also understood that most of the customers with higher incomes and spending tend to buy more meat from the store compared to other products.

In this classification, age and income were divided into four groups and meat, fruits, fish, and sweets were also classified based on how much the consumer was purchasing based on how much they spent. Based on this data, conclusions were made on which products were purchased most and which products were purchased less. Such as, through the final classification, we were able to understand that meat is the highest product purchased among the products and has no non-consumer, and the least purchases were made in fruits.

3. CONCLUSION

In conclusion, the coursework's met with data mining techniques and tools were used to extract valuable information about the dataset, which in this case was Customer personality analysis, which is very important for businesses to understand their customers' personalities in relation to their demands and needs and thus help businesses make decisions. The dataset had been preprocessed and cleaned, and several visualizations had been performed to discover any patterns in the dataset. Finally, the commodities or products were categorized based on customer spending in order to make and get further insights about consumer purchases and demand.

4. REFERENCES

Taylor, D. (2021). *What is BIG DATA? Introduction, Types, Characteristics, Example*. [online] www.guru99.com. Available at: https://www.guru99.com/what-is-big-data.html#6 [Accessed 23 Nov. 2021].

The Economic Times (2019). *Definition of Data Mining | What Is Data Mining ? Data Mining Meaning - the Economic Times*. [online] The Economic Times. Available at: https://economictimes.indiatimes.com/definition/data-mining [Accessed 22 Nov. 2021].

Stedman, C. (n.d.). *What is Data Mining?* [online] SearchBusinessAnalytics. Available at: https://searchbusinessanalytics.techtarget.com/definition/data-mining [Accessed 22 Nov. 2021].

Mayuresh (2020). *Top 10 Data Mining Tools*. [online] Jigsaw. Available at: https://www.jigsawacademy.com/blogs/data-science/data-mining-tools/#Python [Accessed 22 Nov. 2021].

Wickramasinghe, S. (2021). *Jupyter Notebooks for Data Analytics: A Beginner's Guide*. [online] BMC Blogs. Available at: https://www.bmc.com/blogs/installing-jupyter-for-big-data-and-analytics/ [Accessed 22 Nov. 2021].