Project – Course 1
(Exploratory Data Analysis)

Overview:

Here in the data there are some features about customers' personality about certain product. In this report, three hypothesis are created from some basic Exploratory Data Analysis (EDA) and statistical test has been performed for one of them. In the whole thing some feature engineering is also performed. In first section **'About Data'** description of data has been shared. Next all EDAs and Feature Engineering details are elaborately discussed in the **'EDA, Feature Engineering and Hypothesis'** section. At the end this report concluded with **'Conclusion'** section.

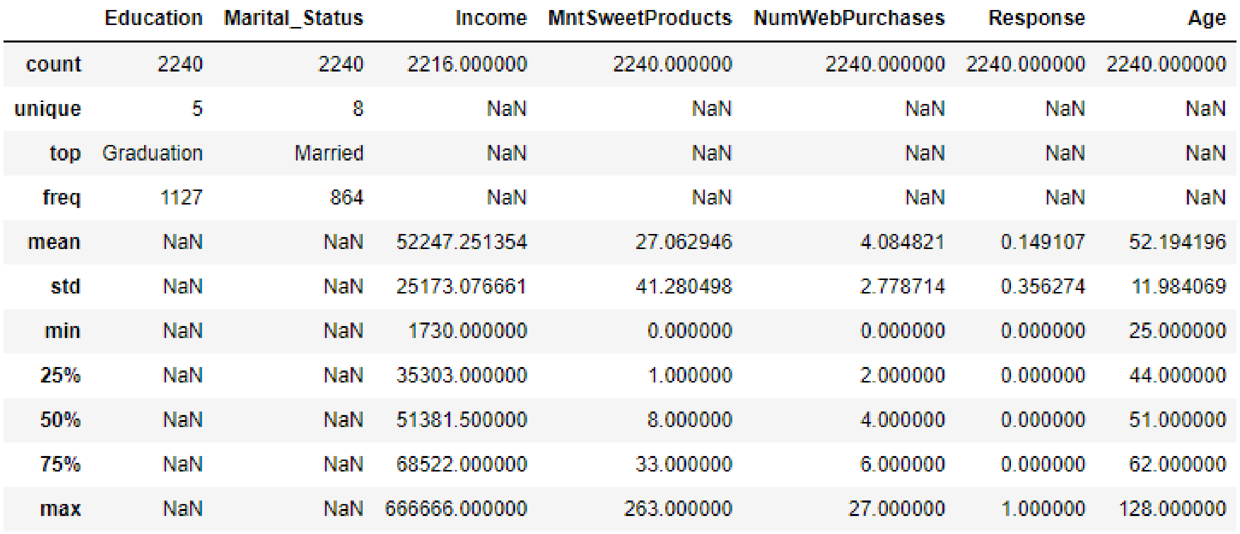
About Data:

Original data is collected from Original Data Source. Here few features have been taken from actual dataset.

Below is description of attributes (taken from original data source link above):

- ID: Customer's unique identifier (type – int64)
- Year_Birth: Customer's birth year (type - int64)
- Education: Customer's education level (type – object)
- Marital_Status: Customer's marital status (type – object)
- Income: Customer's yearly household income (type - int64)
- MntSweetProducts: Amount spent on sweets in last 2 years (type - int64)
- NumWebPurchases: Number of purchases made through the company's website (type - int64)
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise (type - float64)

Here are **8 features** with **2240 data points**. Data types are also mentioned with each variable.



	Education	Marital_Status	Income	MntSweetProducts	NumWebPurchases	Response	Age
count	2240	2240	2216.000000	2240.000000	2240.000000	2240.000000	2240.000000
unique	5	8	NaN	NaN	NaN	NaN	NaN
top	Graduation	Married	NaN	NaN	NaN	NaN	NaN
freq	1127	864	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	52247.251354	27.062946	4.084821	0.149107	52.194196
std	NaN	NaN	25173.076661	41.280498	2.778714	0.356274	11.984069
min	NaN	NaN	1730.000000	0.000000	0.000000	0.000000	25.000000
25%	NaN	NaN	35303.000000	1.000000	2.000000	0.000000	44.000000
50%	NaN	NaN	51381.500000	8.000000	4.000000	0.000000	51.000000
75%	NaN	NaN	68522.000000	33.000000	6.000000	0.000000	62.000000
max	NaN	NaN	666666.000000	263.000000	27.000000	1.000000	128.000000

Fig : 1

Figure 1 gives the description of all the variables (both numeric and objective).

Plan for Data Analysis:

My plan (more or less) will be:

- To check if there is any null value present.
- If null values are present, the method to impute it.
- During this if needed and possible, I will create new feature(s), drop unnecessary feature(s), do some EDA and other feature engineering.
- After imputation I will again do further EDA and feature engineering to do remaining work.
- In the whole process till last step above, if I can set hypothesis at any stage, I will take as my hypothesis.
- If some more hypothesis needed, I will do more analysis, set hypothesis, perform formal statistical test for one of them.

EDA, Feature Engineering and Hypothesis:

First of all we got that here ‘Income’ variable contains null values. We impute them later. First we have seen some EDA over it.

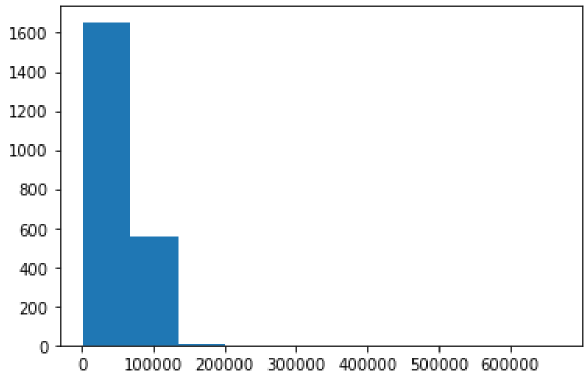


Fig: 2.1

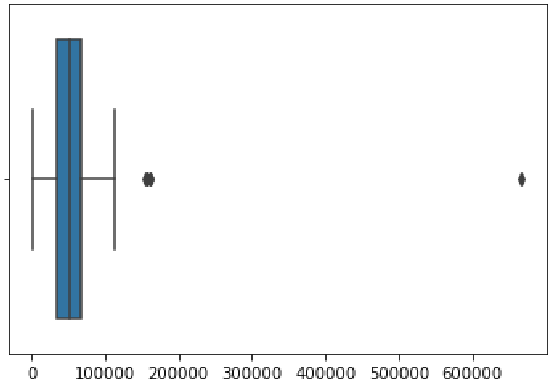


Fig: 2.2

Figure 2.1 and 2.2 shows the histogram and boxplot of income with outliers before imputation (in the portion where they are not null).

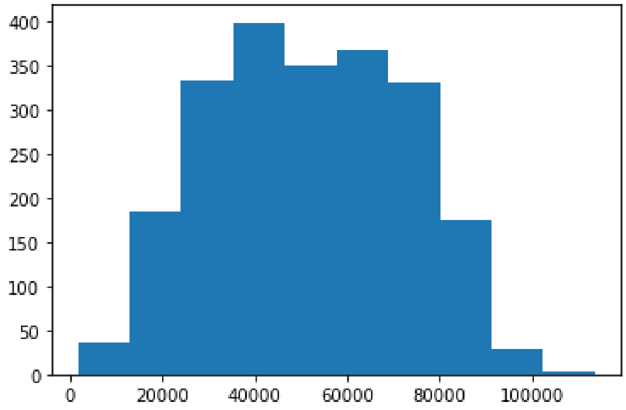


Fig: 2.3

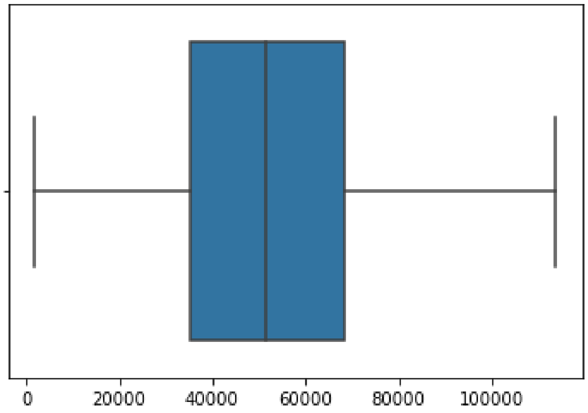


Fig: 2.4

Figure 2.3 and 2.4 shows the histogram and boxplot of income without outliers before imputation (in the portion where they are not null).

From the above diagram distribution of Income seems quite normal without extreme value (in the below 150000 portion) (obviously oin non-null portion). We can think of it as hypothesis. Thus our first hypothesis would be:

Hypothesis-1:

H_0: Income distribution is normal for income below 150000.

H_1: Income distribution is not normal for income below 150000.

Let us perform Shapiro-wilk test to check it statistically. We will reject null hypothesis at α level of significance if test-statistic value will be greater than p-value at level of significance α.

Value of test statistic is = 231.546 and corresponding p-value = 0.000000. So both 5% and 1% level of significance, we can conclude that based on given information, null hypothesis can be rejected that is Income (without extreme values and which portion is available i.e. non-null) is not normally distributed.

Now, two new features 'Age' and 'Age_labels' are created here:

- 'Age' is created by simply subtracting their birth year from current year.
- 'Age_labels' is created by following rule:

  We got minimum and maximum age as 25 and 128 respectively and the rule for creating feature is as follows:

Age Boundary	Value Assigned
25<= age <35	1
35<= age <50	2
50<= age <60	3
60<= age <80	4
age >=80	5

We impute missing income values by calculating median income from each group of combinations from Age labels, Educational Status and Marital Status.

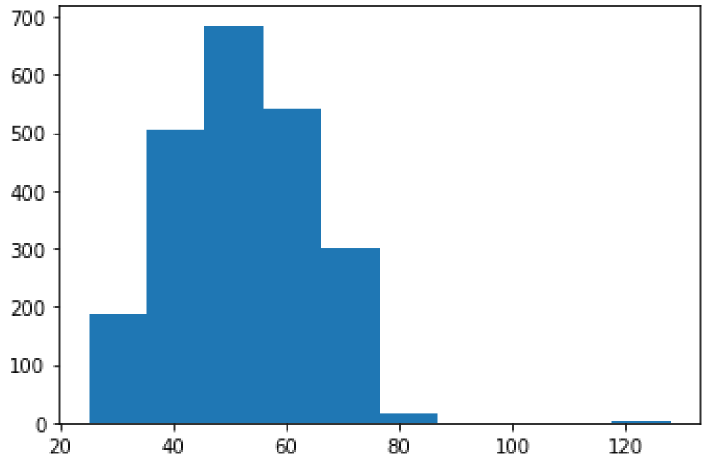


Fig: 3.1

Figure 3.1 shows histogram of Ages of customers in the given data. It seems that very few values around age 120. So let's remove and see the distribution of the major portion.

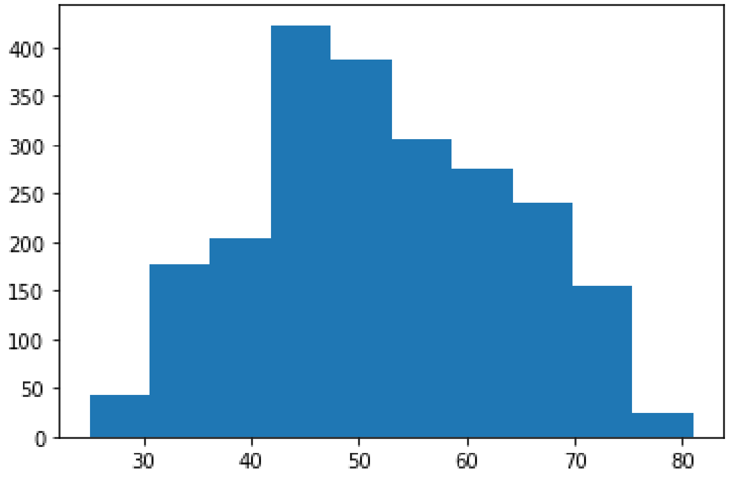


Fig: 3.2

Figure 3.2 shows the distribution of ages of most customers.

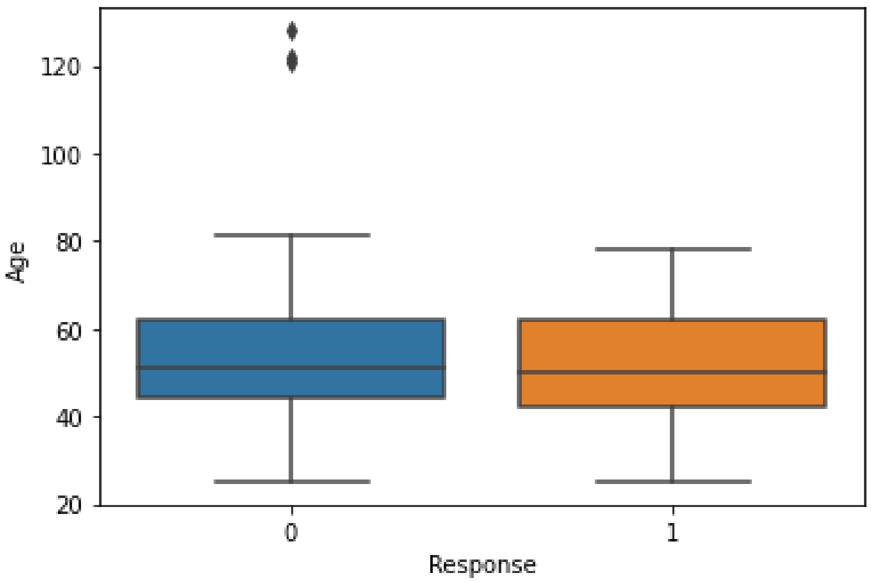


Fig: 3.3

Figure 3.3 shows boxplot between variables Response and Age.

It seems from the diagram that median age of both responses are nearly same. We can think of hypothesis that:

Hypothesis-2:

H_0: Median age is same for both responses.

H_1: Median age is different for both responses.

Next look at the diagram below:

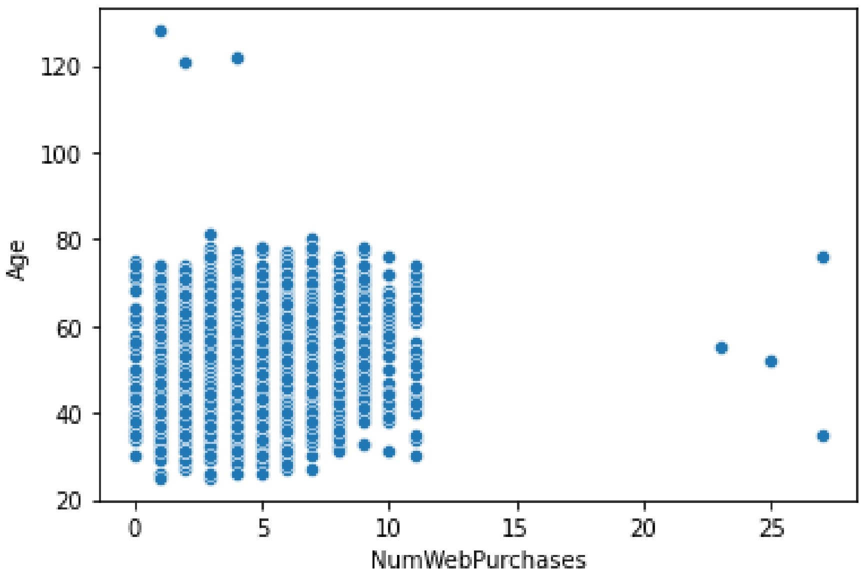


Fig: 4

Figure 4 shows scatter plot between Number of purchases made through the company's website (NumWebPurchases) and Age for customers.

From the diagram it seems that there is no correlation between Number of Web Purchases and age. We can think of hypothesis.

Hypothesis-3:

H_0: Age and Number of Web Purchases (NumWebPurchases) have no correlation.

H_1: Age and Number of Web Purchases (NumWebPurchases) are correlated.

Conclusion:

Here data description is given clearly; Exploratory Data Analysis and Feature Engineering has been performed; 3 hypothesis has been created and among them 1 has been checked via formal statistical tests and all necessary diagrams, results and values are also attached.

Suggestions for Next Steps in Analyzing the Data:

Here in this data next we can:

- Check for outliers in each variable, check for distributions of each variable.
- Calculate percentages of different groups like percentages of different education groups for both kind of responses, percentages of different marital status for both kind of responses.
- Analyze the dependence between different variables, we can also check that if there is any effect of different educational groups or marital status over other variables.

In this way we can also do some other analysis.

Quality of Data:

Data contained:

- Missing values.
- Both numeric (int64, float64) and categorical (object) variables.

Also dataset contains 2240 data points which is not very less. All these things help us to get our hands dirty with raw data analysis.

So, it can be said that data is of quite good quality.