**Data Wrangling Report.**

The data is downloaded from Kaggle website.
Pandas Library is used to read the csv file into Dataframe.
The dataset contains about 537577 observations and 12 variable.

**DATATYPES**:

While exploring datatypes for each variable following information about datatype is observed and columns are transormed to the appropriate category to make the more sense about the data.

```
User_ID                     537577 non-null int64
Product_ID                  537577 non-null object
Gender                      537577 non-null object
Age                         537577 non-null object
Occupation                  537577 non-null int64
City_Category               537577 non-null object
Stay_In_Current_City_Years  537577 non-null object
Marital_Status              537577 non-null int64
Product_Category_1          537577 non-null int64
Product_Category_2          370591 non-null float64
Product_Category_3          164278 non-null float64
Purchase                    537577 non-null int64
```

Following code describes the steps taken to transform the datatypes to appropriate types to help process the data without any typecasting  error while using Numpy/Pandas/Scikit-learn etc.

```
# categorical variables

df['Product_ID'] = df.Product_ID.astype('category')
df['Gender'] = df.Gender.astype('category')
df['Age'] = df.Age.astype('category')
df['Occupation'] = df.Occupation.astype('category')
df['City_Category'] = df.City_Category.astype('category')
df['Marital_Status'] = df.Marital_Status.astype('category')
df['Product_Category_2'] = df.Marital_Status.astype('int')
df['Product_Category_3'] = df.Marital_Status.astype('int')
```

**REMOVAL OF SPECIAL CHARACTER FROM COLUMN VALUES**

Also removed the special character from variable values of columns
‘'Stay_In_Current_City_Years'

The code snippet is as below:

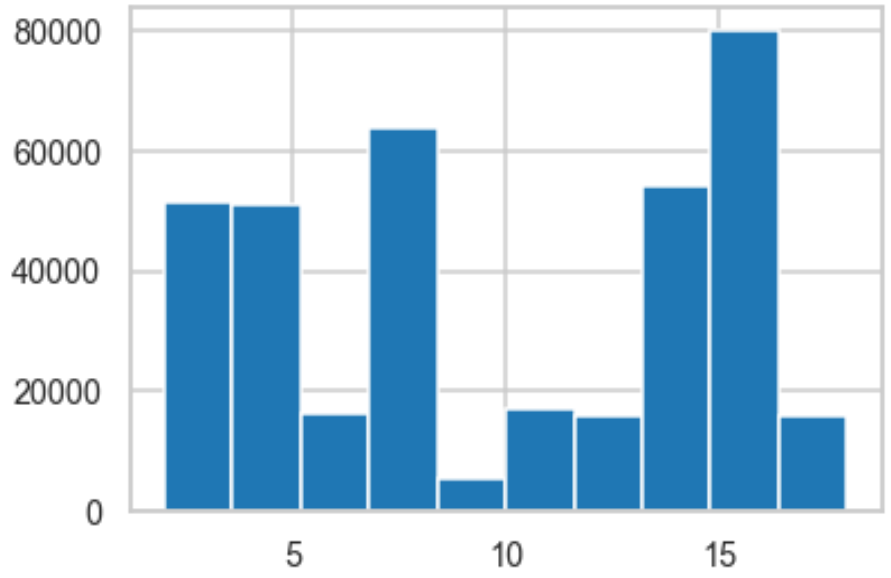
```
df['Stay_In_Current_City_Years'] = (df['Stay_In_Current_City_Years'].str.strip('+').astype(int))
```

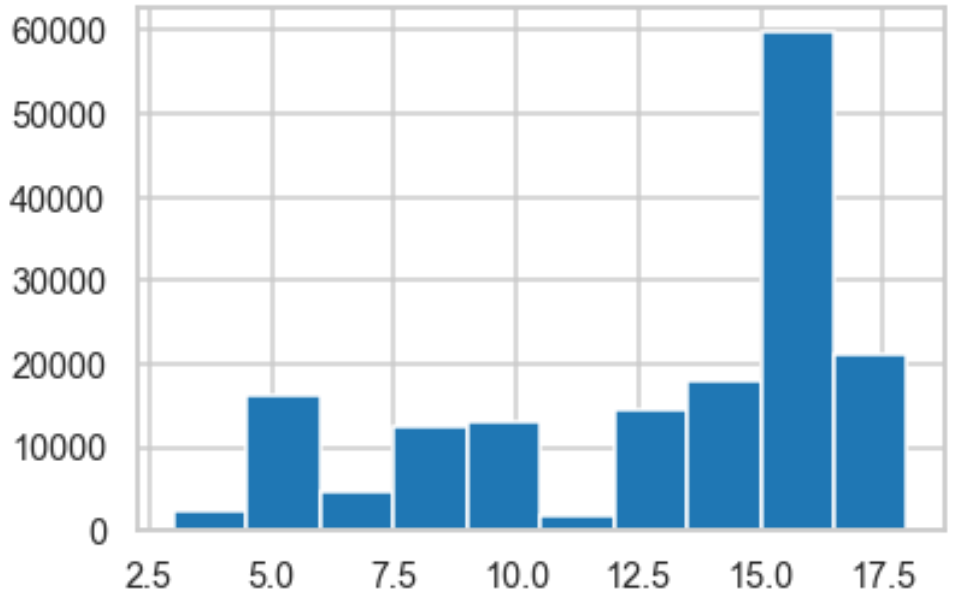
**MISSING VALUES:**

It is observed that there are missing values in both of the columns as mentioned below:

| | |
|---|---|
| Product_Category_2 | 0.310627 |
| Product_Category_3 | 0.694410 |

Product_Category_2 :






Product_Category 3

As per above histogram distribution for Product Category is non-normal distribution.However while imputing missing value ,we can replace the 'nan' value with 0 ,considering non-purchase of perticular item from the category.