**Data Import and Initial Validation**

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', 'Response', 'Complain', 'Country'],

dtype='object')

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Number of duplicated rows in the data is : 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2240 entries, 0 to 2239

Data columns (total 28 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 ID 2240 non-null int64

1 Year\_Birth 2240 non-null int64

2 Education 2240 non-null object

3 Marital\_Status 2240 non-null object

4 Income 2216 non-null object

5 Kidhome 2240 non-null int64

6 Teenhome 2240 non-null int64

7 Dt\_Customer 2240 non-null object

8 Recency 2240 non-null int64

9 MntWines 2240 non-null int64

10 MntFruits 2240 non-null int64

11 MntMeatProducts 2240 non-null int64

12 MntFishProducts 2240 non-null int64

13 MntSweetProducts 2240 non-null int64

14 MntGoldProds 2240 non-null int64

15 NumDealsPurchases 2240 non-null int64

16 NumWebPurchases 2240 non-null int64

17 NumCatalogPurchases 2240 non-null int64

18 NumStorePurchases 2240 non-null int64

19 NumWebVisitsMonth 2240 non-null int64

20 AcceptedCmp3 2240 non-null int64

21 AcceptedCmp4 2240 non-null int64

22 AcceptedCmp5 2240 non-null int64

23 AcceptedCmp1 2240 non-null int64

24 AcceptedCmp2 2240 non-null int64

25 Response 2240 non-null int64

26 Complain 2240 non-null int64

27 Country 2240 non-null object

dtypes: int64(23), object(5)

memory usage: 490.1+ KB

None

# 2. Data type cnoversion

<class 'pandas.core.series.Series'>

RangeIndex: 2240 entries, 0 to 2239

Series name: Dt\_Customer

Non-Null Count Dtype

-------------- -----

2240 non-null datetime64[ns]

dtypes: datetime64[ns](1)

memory usage: 17.6 KB

None

ID int64

Year\_Birth int64

Education category

Marital\_Status category

Income float64

Kidhome int64

Teenhome int64

Dt\_Customer datetime64[ns]

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

Response int64

Complain int64

Country category

dtype: object

# 3. Handling missing values of columns

Total number of missing value for each column

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

Response 0

Complain 0

Country 0

dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Total number of missing value after filling for each column

ID 0

Year\_Birth 0

Education 0

Marital\_Status 0

Income 0

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

Response 0

Complain 0

Country 0

dtype: int64

# 4. Define new columns and calculate values of the columns

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*print 5 rows of data (number of children,age, toal spending and total purchase for customers)

ID Year\_Birth Education Marital\_Status Income Kidhome Teenhome \

0 1826 1970 Graduation Divorced 84835.0 0 0

1 1 1961 Graduation Single 57091.0 0 0

2 10476 1958 Graduation Married 67267.0 0 1

3 1386 1967 Graduation Together 32474.0 1 1

4 5371 1989 Graduation Single 21474.0 1 0

Dt\_Customer Recency MntWines ... AcceptedCmp5 AcceptedCmp1 \

0 2014-06-16 0 189 ... 0 0

1 2014-06-15 0 464 ... 0 0

2 2014-05-13 0 134 ... 0 0

3 2014-05-11 0 10 ... 0 0

4 2014-04-08 0 6 ... 0 0

AcceptedCmp2 Response Complain Country Number\_children Age \

0 0 1 0 SP 0 55

1 1 1 0 CA 0 64

2 0 0 0 US 1 67

3 0 0 0 AUS 2 58

4 0 1 0 SP 1 36

total\_spending total\_purchases

0 1190 14

1 577 17

2 251 10

3 11 3

4 91 6

[5 rows x 32 columns]

# 5. Encode categorical variables using ordinal encoding and one-hot encoding.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*shows first five rows of the datasets after converting the columns (Education, Marital\_Status, Country)\*\*\*\*\*\*\*\*\*\*\*\*

ID Year\_Birth Income Kidhome Teenhome Dt\_Customer Recency \

0 1826 1970 84835.0 0 0 2014-06-16 0

1 1 1961 57091.0 0 0 2014-06-15 0

2 10476 1958 67267.0 0 1 2014-05-13 0

3 1386 1967 32474.0 1 1 2014-05-11 0

4 5371 1989 21474.0 1 0 2014-04-08 0

MntWines MntFruits MntMeatProducts ... Marital\_Status\_Widow \

0 189 104 379 ... 0.0

1 464 5 64 ... 0.0

2 134 11 59 ... 0.0

3 10 0 1 ... 0.0

4 6 16 24 ... 0.0

Marital\_Status\_YOLO Country\_AUS Country\_CA Country\_GER Country\_IND \

0 0.0 0.0 0.0 0.0 0.0

1 0.0 0.0 1.0 0.0 0.0

2 0.0 0.0 0.0 0.0 0.0

3 0.0 1.0 0.0 0.0 0.0

4 0.0 0.0 0.0 0.0 0.0

Country\_ME Country\_SA Country\_SP Country\_US

0 0.0 0.0 1.0 0.0

1 0.0 0.0 0.0 0.0

2 0.0 0.0 0.0 1.0

3 0.0 0.0 0.0 0.0

4 0.0 0.0 1.0 0.0

[5 rows x 46 columns]

Number of rows in the converted dataset: 2240

Number of columns in the converted dataset: 46