Data Cleaning & Preprocessing

- In terms of cleaning data, checking on ***missing values*** and ***duplicate*** rows are done
  - Before executing EDA, some missing values had been addressed as such 'Year_Birth' and 'Income'.
  - After thorough checking, 'Conversion_Rate' seemed to have 11 missing values as the calculation was supposed to be 0. Hence, filling in missing values with 0 is done.
  - No duplicates detected.

```
rows_with_na = df[df['Conversion_Rate'].isna()]
print(rows_with_na)
df['Conversion_Rate'] = df['Conversion_Rate'].fillna(0)


duplicate_row = df[df.duplicated(keep=False)]
duplicate_row
```

For more information, see jupyter notebook here

Data Cleaning & Preprocessing

As some new features are made, those features that wont be used for machine learning will be drop and only features below are retained:

- NumSpendingTotal
- TotalChild
- NumPurchaseTotal
- AcceptedCmpTotal
- NumWebVisitsMonth
- Age_Category
- Conversion Rate
- Complain
- Recency
- Education
- Marital_Status
- Income
- Response (target variable)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Education          2240 non-null   object
 1   Marital_Status     2240 non-null   object
 2   Income             2240 non-null   float64
 3   Recency            2240 non-null   int64
 4   NumWebVisitsMonth  2240 non-null   int64
 5   Complain           2240 non-null   int64
 6   Response           2240 non-null   int64
 7   Conversion_Rate    2240 non-null   float64
 8   Age_Category       2240 non-null   object
 9   AcceptedCmpTotal   2240 non-null   int64
 10  NumPurchaseTotal   2240 non-null   int64
 11  TotalChild         2240 non-null   int64
 12  NumSpendingTotal   2240 non-null   int64
dtypes: float64(2), int64(8), object(3)
memory usage: 227.6+ KB
```

For more information, see jupyter notebook here

Data Cleaning & Preprocessing

- Next for data preprocessing, ***feature encoding*** and ***feature standardisation*** are done.
  - Feature encoding for categorical features include:
    - Label encoding for ordinal data (Education, Age_Category)
    - One hot encoding for non-ordinal data (Marital_Status)
  - Feature standardisation using StandardScaler, features has been checked for standardised
    - Check if features were not encoded before
    - Check if features have different scales with other features

For more information, see jupyter notebook here