Task 01 – Data Cleaning and Preprocessing

Dataset: Customer Personality Analysis

In [1]:
```
# Step 1: Import libraries
import pandas as pd
```

```
C:\Users\pramo\anaconda3\Lib\site-packages\pandas\core\arrays\masked.py:60: UserWarning: Pandas requires version '1.3.6' or newer of 'bottleneck' (version '1.3.5' currently installed).
  from pandas.core import (
```

In [2]:
```
# Step 2: Load the dataset
df = pd.read_csv("marketing_campaign.csv", sep="\t")  # Adjust path as needed
df.head()
```

Out[2]:

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recenc
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	5
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	3
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	2
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	2
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	9

5 rows × 29 columns

In [3]:
```
# Step 3: Check for missing values
df.isnull().sum()
```

Out[3]:
```
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
```

In [4]:
```
# Step 4: Fill missing values or drop them
df = df.fillna("Unknown")
```

In [5]:
```
# Step 5: Remove duplicates
df = df.drop_duplicates()
df.shape
```

Out[5]:
```
(2240, 29)
```

In [6]:
```
# Step 6: Standardize text fields
df['Education'] = df['Education'].str.lower().str.strip()
df['Marital_Status'] = df['Marital_Status'].str.lower().str.strip()
```

In [7]:
```
# Step 7: Convert date formats
df['Dt_Customer'] = pd.to_datetime(df['Dt_Customer'], errors='coerce')
```

In [8]:
```
# Step 8: Rename columns
df.columns = [col.strip().lower().replace(" ", "_") for col in df.columns]
```

In [9]:
```
# Step 9: Check data types
df.dtypes
```

Out[9]:
```
id                             int64
year_birth                     int64
education                     object
marital_status                object
income                        object
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
complain                       int64
z_costcontact                  int64
z_revenue                      int64
response                       int64
dtype: object
```

In [10]:
```
# Step 10: Save cleaned data
df.to_csv("cleaned_customer_personality.csv", index=False)
```

In [ ]:
```

```