**1. Import cleaned data and necessary libraries**

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

RangeIndex: 2240 entries, 0 to 2239

Data columns (total 46 columns):

# Column Non-Null Count Dtype

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

0 ID 2240 non-null int64

1 Year\_Birth 2240 non-null int64

2 Income 2240 non-null float64

3 Kidhome 2240 non-null int64

4 Teenhome 2240 non-null int64

5 Dt\_Customer 2240 non-null datetime64[ns]

6 Recency 2240 non-null int64

7 MntWines 2240 non-null int64

8 MntFruits 2240 non-null int64

9 MntMeatProducts 2240 non-null int64

10 MntFishProducts 2240 non-null int64

11 MntSweetProducts 2240 non-null int64

12 MntGoldProds 2240 non-null int64

13 NumDealsPurchases 2240 non-null int64

14 NumWebPurchases 2240 non-null int64

15 NumCatalogPurchases 2240 non-null int64

16 NumStorePurchases 2240 non-null int64

17 NumWebVisitsMonth 2240 non-null int64

18 AcceptedCmp3 2240 non-null int64

19 AcceptedCmp4 2240 non-null int64

20 AcceptedCmp5 2240 non-null int64

21 AcceptedCmp1 2240 non-null int64

22 AcceptedCmp2 2240 non-null int64

23 Response 2240 non-null int64

24 Complain 2240 non-null int64

25 Number\_children 2240 non-null int64

26 Age 2240 non-null int64

27 total\_spending 2240 non-null int64

28 total\_purchases 2240 non-null int64

29 Education 2240 non-null float64

30 Marital\_Status\_Absurd 2240 non-null float64

31 Marital\_Status\_Alone 2240 non-null float64

32 Marital\_Status\_Divorced 2240 non-null float64

33 Marital\_Status\_Married 2240 non-null float64

34 Marital\_Status\_Single 2240 non-null float64

35 Marital\_Status\_Together 2240 non-null float64

36 Marital\_Status\_Widow 2240 non-null float64

37 Marital\_Status\_YOLO 2240 non-null float64

38 Country\_AUS 2240 non-null float64

39 Country\_CA 2240 non-null float64

40 Country\_GER 2240 non-null float64

41 Country\_IND 2240 non-null float64

42 Country\_ME 2240 non-null float64

43 Country\_SA 2240 non-null float64

44 Country\_SP 2240 non-null float64

45 Country\_US 2240 non-null float64

dtypes: datetime64[ns](1), float64(18), int64(27)

memory usage: 805.1 KB

None

# 2. Histograms & boxplots to detect outliers.



# 3. Interquartile Range (IQR) method is used to detect the possible outliers.

Possible outliers listed below:

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

513 11004 1893 60182.0 0 1 2014-05-17 23

827 1150 1899 83532.0 0 0 2013-09-26 36

2233 7829 1900 36640.0 1 0 2013-09-26 99

325 4931 1977 157146.0 0 0 2013-04-29 13

497 1501 1982 160803.0 0 0 2012-08-04 21

527 9432 1977 666666.0 1 0 2013-06-02 23

731 1503 1976 162397.0 1 1 2013-06-03 31

853 5336 1971 157733.0 1 0 2013-06-04 37

1826 5555 1975 153924.0 0 0 2014-02-07 81

1925 11181 1949 156924.0 0 0 2013-08-29 85

2204 8475 1973 157243.0 0 1 2014-03-01 98

671 5735 1991 90638.0 0 0 2014-02-13 29

672 5350 1991 90638.0 0 0 2014-02-13 29

1404 1763 1988 87679.0 0 0 2013-07-27 62

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

513 8 0 5 ... 0.0

827 755 144 562 ... 0.0

2233 15 6 8 ... 0.0

325 1 0 1725 ... 0.0

497 55 16 1622 ... 0.0

527 9 14 18 ... 0.0

731 85 1 16 ... 0.0

853 39 1 9 ... 0.0

1826 1 1 1 ... 0.0

1925 2 1 2 ... 0.0

2204 20 2 1582 ... 0.0

671 1156 120 915 ... 0.0

672 1156 120 915 ... 0.0

1404 1259 172 815 ... 0.0

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

513 0.0 0.0 0.0 0.0 0.0

827 0.0 0.0 0.0 0.0 0.0

2233 0.0 0.0 0.0 0.0 1.0

325 0.0 0.0 0.0 0.0 0.0

497 0.0 0.0 0.0 0.0 0.0

527 0.0 0.0 0.0 0.0 0.0

731 0.0 0.0 0.0 0.0 0.0

853 0.0 0.0 0.0 0.0 0.0

1826 0.0 0.0 0.0 0.0 0.0

1925 0.0 0.0 1.0 0.0 0.0

2204 0.0 0.0 0.0 0.0 1.0

671 0.0 0.0 0.0 0.0 0.0

672 0.0 0.0 0.0 0.0 0.0

1404 0.0 0.0 1.0 0.0 0.0

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

513 0.0 1.0 0.0 0.0

827 0.0 0.0 1.0 0.0

2233 0.0 0.0 0.0 0.0

325 0.0 1.0 0.0 0.0

497 0.0 0.0 0.0 1.0

527 0.0 1.0 0.0 0.0

731 0.0 0.0 1.0 0.0

853 0.0 0.0 1.0 0.0

1826 0.0 0.0 1.0 0.0

1925 0.0 0.0 0.0 0.0

2204 0.0 0.0 0.0 0.0

671 0.0 0.0 1.0 0.0

672 0.0 1.0 0.0 0.0

1404 0.0 0.0 0.0 0.0

[14 rows x 46 columns]

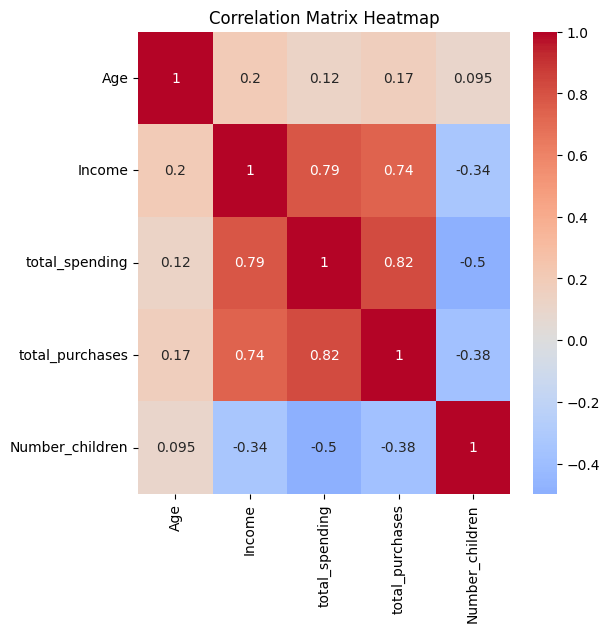
Number of outliers: 14

# 4. Drop outliers from the dataset

Number of row (data points) before oulier removal: 2240

Number of row (data points) after oulier removal:2235

# 5. Correlation heatmap to analyze relationships



# 6. Check statistical distribution of the variable for being normal



Results of skewness and kurtosis analysis:

Skewness of Age distribution 0.09318005141809128 and excess kurtosis -0.7960573867959697

Skewness of income distribution 0.34962227620452396 and excess kurtosis 0.7603032490741608

Skewness of total\_spending distribution 0.8555656955614988 and excess kurtosis -0.35936897061620776

Skewness of total\_purchases 0.29608524206744147 and excess kurtosis -1.1190277763730458

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

Results of Shapiro–Wilk test:

P-value for age distribution is 1.6729978229338104e-15

P-value for income distribution is 5.884611736613058e-19

P-value for total\_spending distribution is 3.3094740822001374e-40

P-value for total\_purchases distribution is 3.3573367503886564e-30