In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
```

In [3]:
```
df=pd.read_csv('/Users/4star/Desktop/Data Analysis/ifood_df.csv')
```

In [5]:
```
df.head()
```

Out[5]:

	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts
0	58138.0	0	0	58	635	88	546
1	46344.0	1	1	38	11	1	6
2	71613.0	0	0	26	426	49	127
3	26646.0	1	0	26	11	4	20
4	58293.0	1	0	94	173	43	118

5 rows × 39 columns

In [6]:
```
#Shape of dataset
df.shape
```

Out[6]:
```
(2205, 39)
```

In [7]:
```
#information related dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2205 entries, 0 to 2204
Data columns (total 39 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Income                2205 non-null   float64
 1   Kidhome               2205 non-null   int64  
 2   Teenhome              2205 non-null   int64  
 3   Recency               2205 non-null   int64  
 4   MntWines              2205 non-null   int64  
 5   MntFruits             2205 non-null   int64  
 6   MntMeatProducts       2205 non-null   int64  
 7   MntFishProducts       2205 non-null   int64  
 8   MntSweetProducts      2205 non-null   int64  
 9   MntGoldProds          2205 non-null   int64  
 10  NumDealsPurchases     2205 non-null   int64  
 11  NumWebPurchases       2205 non-null   int64  
 12  NumCatalogPurchases   2205 non-null   int64  
 13  NumStorePurchases     2205 non-null   int64  
 14  NumWebVisitsMonth     2205 non-null   int64  
 15  AcceptedCmp3          2205 non-null   int64  
 16  AcceptedCmp4          2205 non-null   int64  
 17  AcceptedCmp5          2205 non-null   int64  
 18  AcceptedCmp1          2205 non-null   int64  
 19  AcceptedCmp2          2205 non-null   int64  
 20  Complain              2205 non-null   int64  
 21  Z_CostContact         2205 non-null   int64  
 22  Z_Revenue             2205 non-null   int64  
 23  Response              2205 non-null   int64  
 24  Age                   2205 non-null   int64  
 25  Customer_Days         2205 non-null   int64  
 26  marital_Divorced      2205 non-null   int64  
 27  marital_Married       2205 non-null   int64  
 28  marital_Single        2205 non-null   int64  
 29  marital_Together      2205 non-null   int64  
 30  marital_Widow         2205 non-null   int64  
 31  education_2n Cycle    2205 non-null   int64  
 32  education_Basic       2205 non-null   int64  
 33  education_Graduation  2205 non-null   int64  
 34  education_Master      2205 non-null   int64  
 35  education_PhD         2205 non-null   int64  
 36  MntTotal              2205 non-null   int64  
 37  MntRegularProds       2205 non-null   int64  
 38  AcceptedCmpOverall    2205 non-null   int64  
dtypes: float64(1), int64(38)
memory usage: 672.0 KB
```

In [9]:
```
#checking dor the null values
df.isna().sum()
```

Out[9]:
```
Income                 0
Kidhome                0
Teenhome               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
Age                    0
Customer_Days          0
marital_Divorced       0
marital_Married        0
marital_Single         0
marital_Together       0
marital_Widow          0
education_2n Cycle     0
education_Basic        0
education_Graduation   0
education_Master       0
education_PhD          0
MntTotal               0
MntRegularProds        0
AcceptedCmpOverall     0
dtype: int64
```

In [10]:
```
df.describe()
```

Out[10]:

	Income	Kidhome	Teenhome	Recency	MntWines	M
count	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205
mean	51622.094785	0.442177	0.506576	49.009070	306.164626	2
std	20713.063826	0.537132	0.544380	28.932111	337.493839	39
min	1730.000000	0.000000	0.000000	0.000000	0.000000	0
25%	35196.000000	0.000000	0.000000	24.000000	24.000000	2
50%	51287.000000	0.000000	0.000000	49.000000	178.000000	8
75%	68281.000000	1.000000	1.000000	74.000000	507.000000	33
max	113734.000000	2.000000	2.000000	99.000000	1493.000000	199

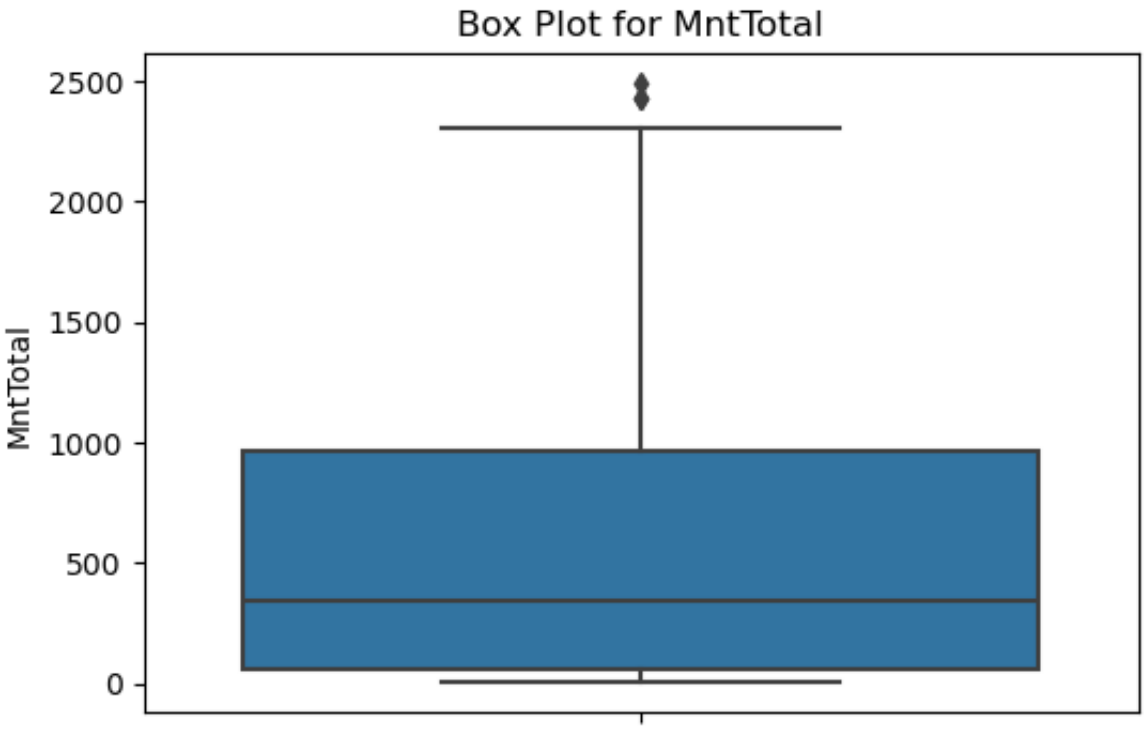
8 rows × 39 columns

In [11]:
```
#check the unique values in each column
df.nunique()
```

Out[11]:
```
Income                  1963
Kidhome                    3
Teenhome                   3
Recency                  100
MntWines                 775
MntFruits                158
MntMeatProducts          551
MntFishProducts          182
MntSweetProducts         176
MntGoldProds             212
NumDealsPurchases         15
NumWebPurchases           15
NumCatalogPurchases       13
NumStorePurchases         14
NumWebVisitsMonth         16
AcceptedCmp3               2
AcceptedCmp4               2
AcceptedCmp5               2
AcceptedCmp1               2
AcceptedCmp2               2
Complain                   2
Z_CostContact              1
Z_Revenue                  1
Response                   2
Age                       56
Customer_Days            662
marital_Divorced           2
marital_Married            2
marital_Single             2
marital_Together           2
marital_Widow              2
education_2n Cycle         2
education_Basic            2
education_Graduation       2
education_Master           2
education_PhD              2
MntTotal                 897
MntRegularProds          974
AcceptedCmpOverall         5
dtype: int64
```

In [13]:
```
df.drop(columns=['Z_CostContact','Z_Revenue'],inplace =True)
```

In [14]:
```
#Boxplot will help us to find outliers if any.
plt.figure(figsize=(6, 4))
sns.boxplot(data=df, y='MntTotal')
plt.title('Box Plot for MntTotal')
plt.ylabel('MntTotal')
plt.show()
```



In [18]:
```
#Outliers
#The box plot spotted a few outliers in the MntTotal. Let's take a closer

Q1 = df['MntTotal'].quantile(0.25)
Q3 = df['MntTotal'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df['MntTotal'] < lower_bound) | (df['MntTotal'] > upper_bo
outliers.head()
```

Out[18]:

	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProd
1159	90638.0	0	0	29	1156	120	
1467	87679.0	0	0	62	1259	172	
1547	90638.0	0	0	29	1156	120	

3 rows × 37 columns

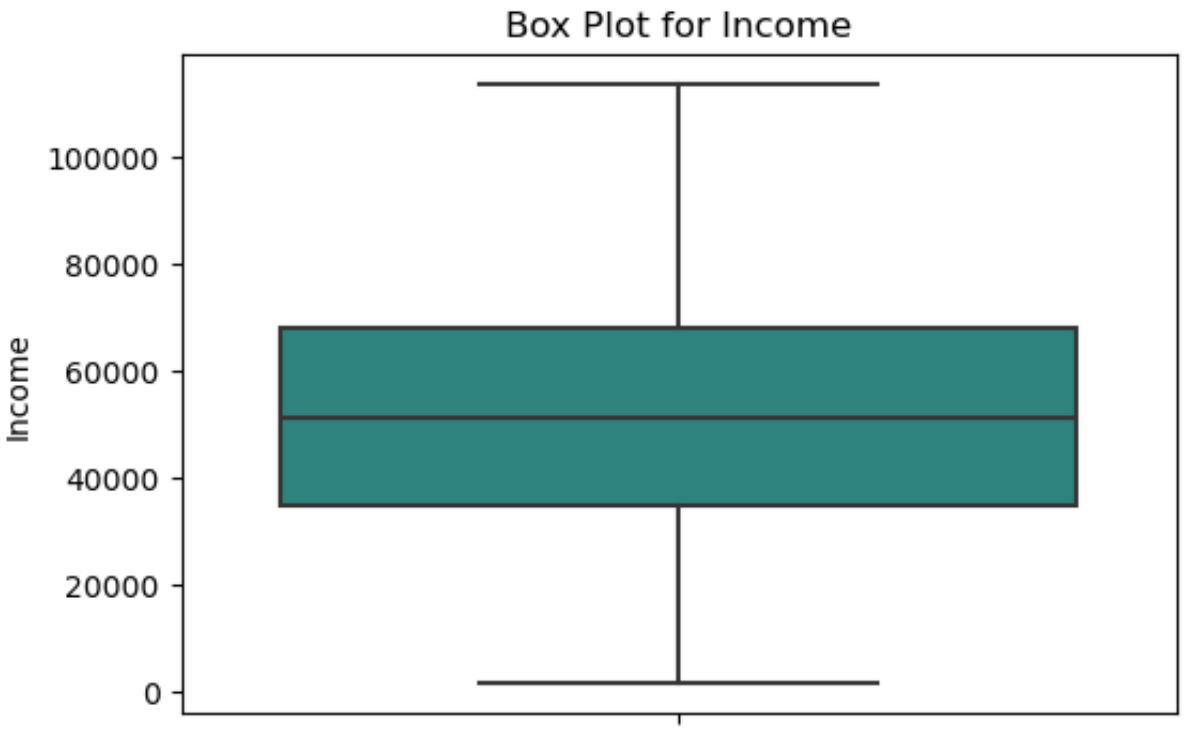
In [19]:
```
#Outliers removal
data = df[(df['MntTotal'] > lower_bound) & (df['MntTotal'] < upper_bound)
data.describe()
```

Out[19]:

	Income	Kidhome	Teenhome	Recency	MntWines	M
count	2202.000000	2202.000000	2202.000000	2202.000000	2202.000000	220
mean	51570.283379	0.442779	0.507266	49.021344	304.960036	26
std	20679.438848	0.537250	0.544429	28.944211	336.135586	3
min	1730.000000	0.000000	0.000000	0.000000	0.000000	0
25%	35182.500000	0.000000	0.000000	24.000000	24.000000	2
50%	51258.500000	0.000000	0.000000	49.000000	176.500000	8
75%	68146.500000	1.000000	1.000000	74.000000	505.000000	33
max	113734.000000	2.000000	2.000000	99.000000	1493.000000	199

8 rows × 37 columns

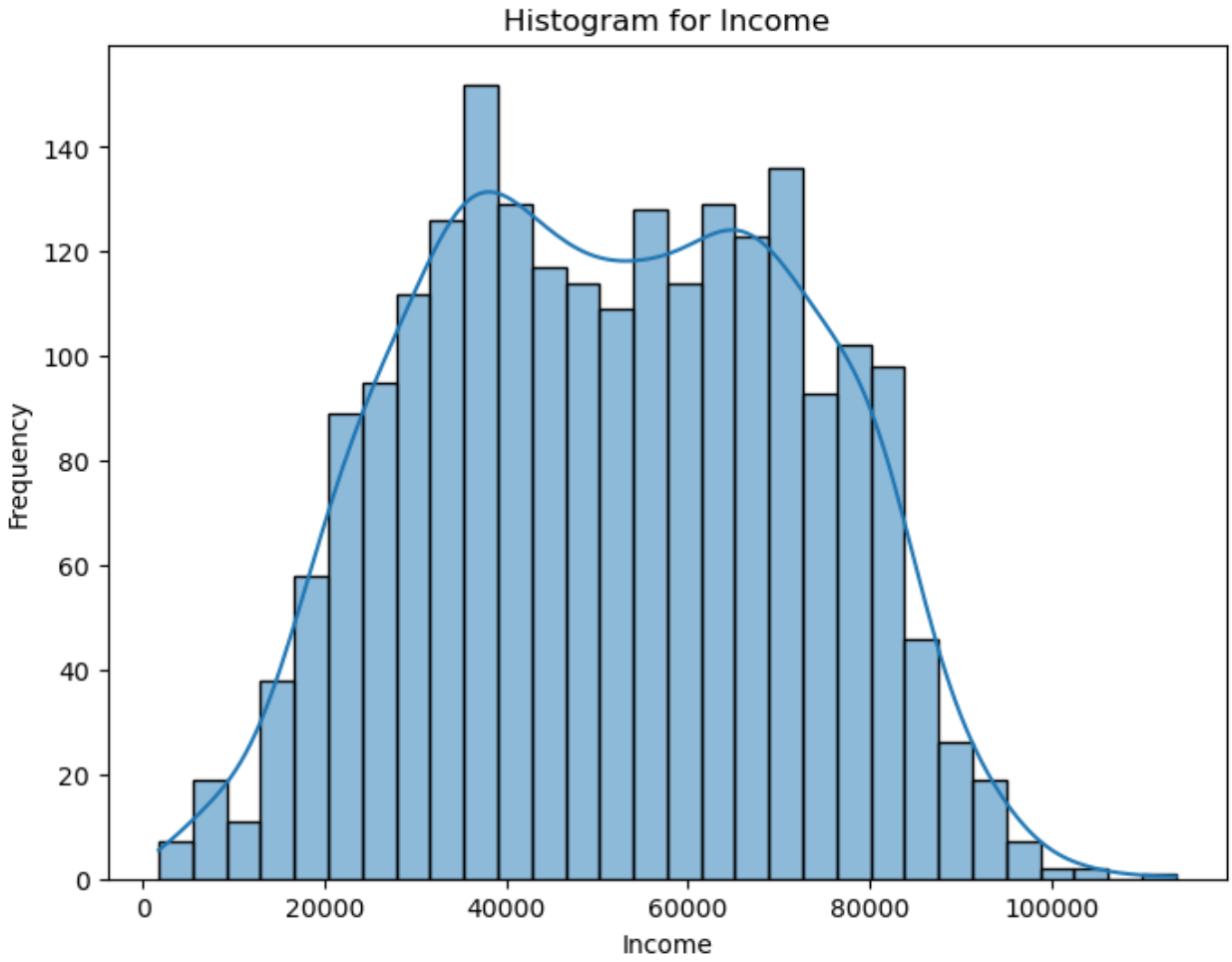
In [20]:
```
#Box plot and histogram for income
plt.figure(figsize=(6, 4))
sns.boxplot(data=data, y='Income', palette='viridis')
plt.title('Box Plot for Income')
plt.ylabel('Income')
plt.show()
```



In [21]:
```
plt.figure(figsize=(8, 6))
sns.histplot(data=data, x='Income', bins=30, kde=True)
plt.title('Histogram for Income')
plt.xlabel('Income')
plt.ylabel('Frequency')
```

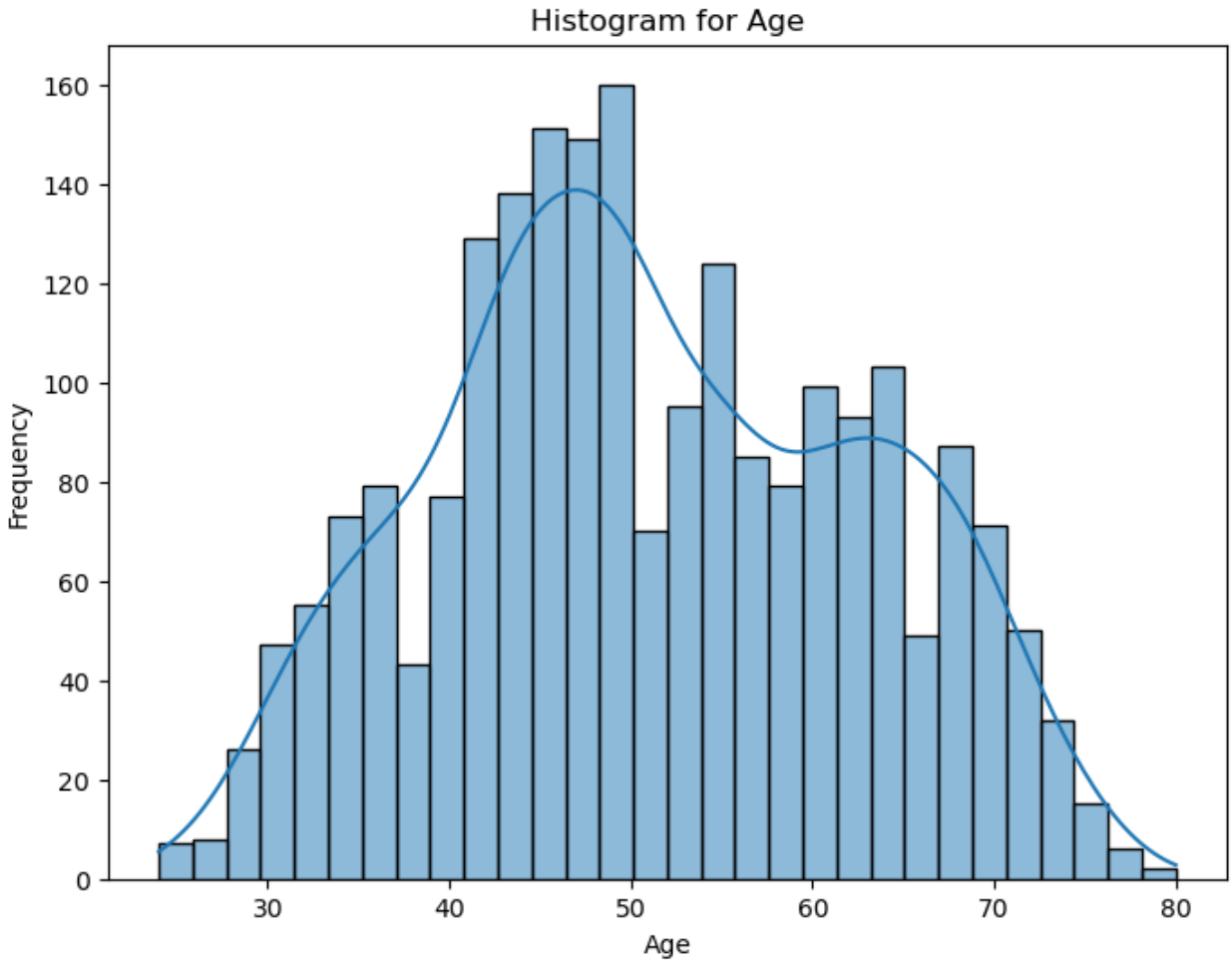
```
plt.show()
```

```
/opt/anaconda3/lib/python3.11/site-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



In [22]:
```
#Histogram for age
plt.figure(figsize=(8, 6))
sns.histplot(data=data, x='Age', bins=30, kde=True)
plt.title('Histogram for Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

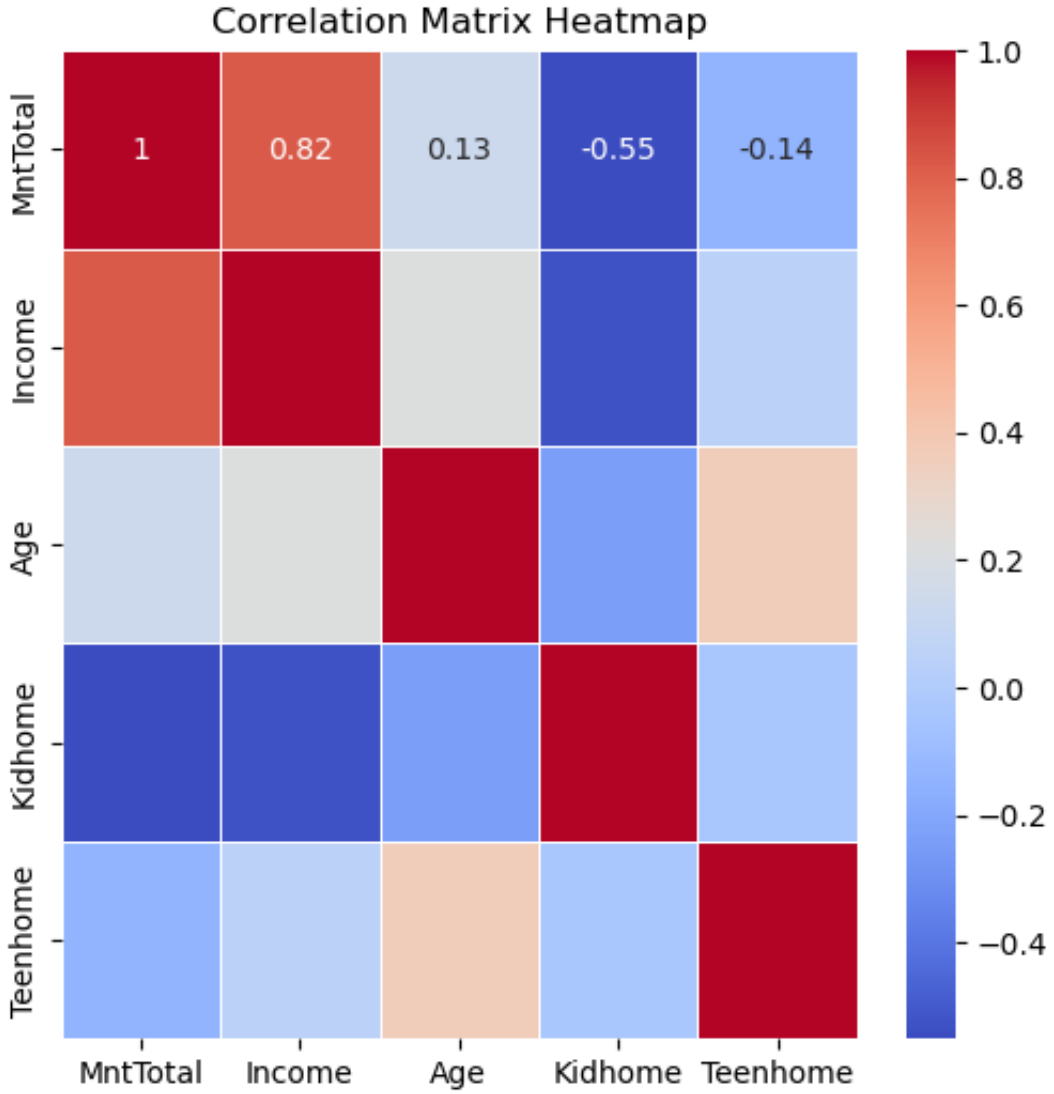
```
/opt/anaconda3/lib/python3.11/site-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



In [23]:
```
print("Skewness: %f" % data['Age'].skew())
print("Kurtosis: %f" % data['Age'].kurt())
```

```
Skewness: 0.091227
Kurtosis: -0.796125
```

In [25]:
```
cols_demographics = ['Income','Age']
cols_children = ['Kidhome', 'Teenhome']
cols_marital = ['marital_Divorced', 'marital_Married','marital_Single', '
cols_mnt = ['MntTotal', 'MntRegularProds','MntWines', 'MntFruits', 'MntMe
cols_communication = ['Complain', 'Response', 'Customer_Days']
cols_campaigns = ['AcceptedCmpOverall', 'AcceptedCmp1', 'AcceptedCmp2', '
cols_source_of_purchase = ['NumDealsPurchases', 'NumWebPurchases','NumCat
cols_education = ['education_2n Cycle', 'education_Basic', 'education_Gra
corr_matrix = data[['MntTotal']+cols_demographics+cols_children].corr()
plt.figure(figsize=(6,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix Heatmap')
plt.show()
```


Correlation Matrix Heatmap
MntTotal
Income
Age
Kidhome
Teenhome
1
0.82
0.13
-0.55
-0.14
MntTotal
Income
Age
Kidhome
Teenhome
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

In [ ]: