In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
demo = pd.read_sas("Downloads/NHANES_2017-2020/P_DEMO.xpt", format='xport')
diet_day1 = pd.read_sas("Downloads/NHANES_2017-2020/P_DR1TOT.xpt", format='x
diet_day2 = pd.read_sas("Downloads/NHANES_2017-2020/P_DR2TOT.xpt", format='x
blood_pressure = pd.read_sas("Downloads/NHANES_2017-2020/P_BPXO.xpt", format
body_measures = pd.read_sas("Downloads/NHANES_2017-2020/P_BMX.xpt", format='
insulin = pd.read_sas("Downloads/NHANES_2017-2020/P_INS.xpt", format='xport'
glucose = pd.read_sas("Downloads/NHANES_2017-2020/P_GLU.xpt", format='xport'
lipids = pd.read_sas("Downloads/NHANES_2017-2020/P_TRIGLY.xpt", format='xpor
hdl = pd.read_sas("Downloads/NHANES_2017-2020/P_HDL.xpt", format='xport')
inflammation = pd.read_sas("Downloads/NHANES_2017-2020/P_HSCRP.xpt", format=
fasting_qst = pd.read_sas("Downloads/NHANES_2017-2020/P_FASTQX.xpt", format=
physical_activity = pd.read_sas("Downloads/NHANES_2017-2020/P_PAQ.xpt", form
prescriptions = pd.read_sas("Downloads/NHANES_2017-2020/P_RXQ_RX.xpt", forma
bpq_qst = pd.read_sas("Downloads/NHANES_2017-2020/P_BPQ.xpt", format='xport'
diabetes_qst = pd.read_sas("Downloads/NHANES_2017-2020/P_DIQ.xpt", format='x
diet_behavior = pd.read_sas("Downloads/NHANES_2017-2020/P_DBQ.xpt", format='
income = pd.read_sas("Downloads/NHANES_2017-2020/P_INQ.xpt", format='xport')
insurance = pd.read_sas("Downloads/NHANES_2017-2020/P_HIQ.xpt", format='xpor
```

In [3]:
```
demo.columns
```

Out[3]:
```
Index(['SEQN', 'SDDSRVYR', 'RIDSTATR', 'RIAGENDR', 'RIDAGEYR', 'RIDAGEMN',
       'RIDRETH1', 'RIDRETH3', 'RIDEXMON', 'DMDBORN4', 'DMDYRUSZ', 'DMDEDUC
2',
       'DMDMARTZ', 'RIDEXPRG', 'SIALANG', 'SIAPROXY', 'SIAINTRP', 'FIALAN
G',
       'FIAPROXY', 'FIAINTRP', 'MIALANG', 'MIAPROXY', 'MIAINTRP', 'AIALANG
A',
       'WTINTPRP', 'WTMECPRP', 'SDMVPSU', 'SDMVSTRA', 'INDFMPIR'],
      dtype='object')
```

In [4]:
```
demo_keep = ['SEQN', 'RIAGENDR', 'RIDAGEYR', 'RIDRETH3', 'DMDBORN4', 'DMDEDU
# RIAGENDR: Gender
# RIDAGEYR: Age
# RIDRETH3: Race/ethnicity
# DMDBORN4: Country of birth
# DMDEDUC2: Education level
# INDFMPIR: Income-to-poverty ratio
```

In [5]:
```
# drop pregnant particpants
demo = demo[demo['RIDEXPRG'] != 1.0]
demo_filtered = demo[demo_keep]
demo_filtered
```

Out[5]:

	SEQN	RIAGENDR	RIDAGEYR	RIDRETH3	DMDBORN4	DMDEDUC2	INDFMI
0	109263.0	1.0	2.0	6.0	1.0	NaN	4
1	109264.0	2.0	13.0	1.0	1.0	NaN	0
2	109265.0	1.0	2.0	3.0	1.0	NaN	3
3	109266.0	2.0	29.0	6.0	2.0	5.0	5
4	109267.0	2.0	21.0	2.0	2.0	4.0	5
...	...	...	...	...	...	...	
15555	124818.0	1.0	40.0	4.0	1.0	5.0	3
15556	124819.0	1.0	2.0	4.0	1.0	NaN	0
15557	124820.0	2.0	7.0	3.0	1.0	NaN	1
15558	124821.0	1.0	63.0	4.0	1.0	2.0	3
15559	124822.0	1.0	74.0	2.0	2.0	3.0	N

15473 rows × 7 columns

In [6]:
```
demo_filtered.isnull().sum()
```

Out[6]:
```
SEQN           0
RIAGENDR       0
RIDAGEYR       0
RIDRETH3       0
DMDBORN4       0
DMDEDUC2    6328
INDFMPIR    2187
dtype: int64
```

In [7]:
```
diet_keep = [
    'SEQN', 'DR1TKCAL', 'DR1TPROT', 'DR1TCARB', 'DR1TSUGR',
    'DR1TFIBE', 'DR1TTFAT', 'DR1TSFAT', 'DR1TCHOL',
    'DR1TALCO', 'DR1TCAFF', 'DR1TSODI', 'DR1TPOTA'
]

diet_day1_filtered = diet_day1[diet_keep]
diet_day1_filtered
```

Out[7]:

	SEQN	DR1TKCAL	DR1TPROT	DR1TCARB	DR1TSUGR	DR1TFIBE	DR1TTFA
0	109263.0	1402.0	52.79	187.65	73.42	9.4	48.8
1	109264.0	1046.0	55.55	121.68	27.86	8.2	37.6
2	109265.0	1926.0	57.47	246.53	157.08	7.6	80.6
3	109266.0	1698.0	52.58	217.69	94.20	20.7	73.8
4	109269.0	1251.0	24.96	159.99	84.83	5.1	57.8
...	...	...	...	...	...	...	.
14295	124818.0	3868.0	58.05	512.14	278.84	15.0	183.5
14296	124819.0	1749.0	67.43	197.26	108.36	8.3	77.7
14297	124820.0	1204.0	39.32	157.81	65.77	12.3	48.7
14298	124821.0	1698.0	138.10	110.59	50.57	6.7	76.0
14299	124822.0	NaN	NaN	NaN	NaN	NaN	Na

14300 rows × 13 columns

In [8]:
```
diet_day1_filtered.isnull().sum()
```

Out[8]:
```
SEQN          0
DR1TKCAL    1908
DR1TPROT    1908
DR1TCARB    1908
DR1TSUGR    1908
DR1TFIBE    1908
DR1TTFAT    1908
DR1TSFAT    1908
DR1TCHOL    1908
DR1TALCO    1908
DR1TCAFF    1908
DR1TSODI    1908
DR1TPOTA    1908
dtype: int64
```

In [9]:
```
diet2_keep = [
    'SEQN', 'DR2TKCAL', 'DR2TPROT', 'DR2TCARB', 'DR2TSUGR',
    'DR2TFIBE', 'DR2TTFAT', 'DR2TSFAT', 'DR2TCHOL',
    'DR2TALCO', 'DR2TCAFF', 'DR2TSODI', 'DR2TPOTA'
]
```

```
diet_day2_filtered = diet_day2[diet2_keep]
diet_day2_filtered
```

Out[9]:

	SEQN	DR2TKCAL	DR2TPROT	DR2TCARB	DR2TSUGR	DR2TFIBE	DR2TTF
0	109263.0	1133.0	34.45	192.65	82.92	4.3	23.
1	109264.0	1932.0	74.78	251.58	89.08	17.2	74
2	109265.0	1551.0	48.81	194.87	125.10	9.5	66
3	109266.0	1896.0	62.92	275.62	65.71	18.7	61
4	109269.0	847.0	27.86	99.55	78.67	1.4	38.
...	...	...	...	...	...	...	
14295	124818.0	2926.0	115.08	353.60	220.20	22.3	121.
14296	124819.0	1143.0	25.70	151.15	58.31	9.9	49.
14297	124820.0	1662.0	48.49	209.74	92.60	9.9	72.
14298	124821.0	NaN	NaN	NaN	NaN	NaN	N
14299	124822.0	NaN	NaN	NaN	NaN	NaN	N

14300 rows × 13 columns

In [10]:
```
diet_day2_filtered.isnull().sum()
```

Out[10]:
```
SEQN          0
DR2TKCAL    3673
DR2TPROT    3673
DR2TCARB    3673
DR2TSUGR    3673
DR2TFIBE    3673
DR2TTFAT    3673
DR2TSFAT    3673
DR2TCHOL    3673
DR2TALCO    3673
DR2TCAFF    3673
DR2TSODI    3673
DR2TPOTA    3673
dtype: int64
```

In [11]:
```
blood_pressure_filtered = blood_pressure.drop(columns=['BPAOARM', 'BPAOCSZ'])
blood_pressure_filtered
```

Out[11]:

	SEQN	BPXOSY1	BPXODI1	BPXOSY2	BPXODI2	BPXOSY3	BPXODI3	BPXC
0	109264.0	109.0	67.0	109.0	68.0	106.0	66.0	
1	109266.0	99.0	56.0	99.0	55.0	99.0	52.0	
2	109270.0	123.0	73.0	124.0	77.0	127.0	70.0	
3	109271.0	102.0	65.0	108.0	68.0	111.0	68.0	
4	109273.0	116.0	68.0	110.0	66.0	115.0	68.0	
...	...	...	...	...	...	...	...	
11651	124815.0	141.0	87.0	138.0	86.0	141.0	87.0	
11652	124817.0	111.0	69.0	112.0	67.0	113.0	66.0	
11653	124818.0	106.0	67.0	108.0	65.0	95.0	70.0	
11654	124821.0	121.0	66.0	122.0	67.0	129.0	67.0	
11655	124822.0	122.0	55.0	121.0	56.0	123.0	58.0	

11656 rows × 10 columns

In [12]:
```
blood_pressure_filtered.isnull().sum()
```

Out[12]:
```
SEQN          0
BPXOSY1    1304
BPXODI1    1304
BPXOSY2    1329
BPXODI2    1329
BPXOSY3    1370
BPXODI3    1370
BPXOPLS1   2185
BPXOPLS2   2208
BPXOPLS3   2244
dtype: int64
```

In [46]:
```
blood_pressure_filtered.columns
```

Out[46]:
```
Index(['SEQN', 'BPXOSY1', 'BPXODI1', 'BPXOSY2', 'BPXODI2', 'BPXOSY3',
       'BPXODI3', 'BPXOPLS1', 'BPXOPLS2', 'BPXOPLS3'],
      dtype='object')
```

In [13]:
```
body_keep = ['SEQN', 'BMXWT', 'BMXHT', 'BMXBMI', 'BMXWAIST']
# BMXWT: Weight
# BMXHT: Height
# BMXBMI: BMI
# BMXWAIST: Waist circumference
body_measures_filtered = body_measures[body_keep]
body_measures_filtered
```

Out[13]:

	SEQN	BMXWT	BMXHT	BMXBMI	BMXWAIST
0	109263.0	NaN	NaN	NaN	NaN
1	109264.0	42.2	154.7	17.6	63.8
2	109265.0	12.0	89.3	15.0	41.2
3	109266.0	97.1	160.2	37.8	117.9
4	109269.0	13.6	NaN	NaN	NaN
...	...	...	...	...	...
14295	124818.0	108.8	168.7	38.2	114.7
14296	124819.0	15.4	93.7	17.5	48.4
14297	124820.0	22.9	123.3	15.1	57.5
14298	124821.0	79.5	176.4	25.5	97.1
14299	124822.0	59.7	167.5	21.3	86.9

14300 rows × 5 columns

In [14]:
```
body_measures_filtered.isnull().sum()
```

Out[14]:
```
SEQN           0
BMXWT        225
BMXHT       1143
BMXBMI      1163
BMXWAIST    1726
dtype: int64
```

In [15]:
```
insulin_keep = ['SEQN', 'LBXIN']
# LBXIN: Fasting insulin
insulin_filtered = insulin[insulin_keep]
insulin_filtered
```

Out[15]:

	SEQN	LBXIN
0	109264.0	6.05
1	109271.0	16.96
2	109274.0	13.52
3	109277.0	6.44
4	109282.0	7.49
...	...	...
5085	124813.0	8.19
5086	124814.0	7.27
5087	124815.0	7.10
5088	124821.0	7.75
5089	124822.0	4.45

5090 rows × 2 columns

In [16]:
```
insulin_filtered.isnull().sum()
```

Out[16]:
```
SEQN       0
LBXIN    465
dtype: int64
```

In [17]:
```
glucose_keep = ['SEQN', 'LBXGLU']
# LBXGLU: Fasting glucose
glucose_filtered = glucose[glucose_keep]
glucose_filtered
```

Out[17]:

	SEQN	LBXGLU
0	109264.0	97.0
1	109271.0	103.0
2	109274.0	154.0
3	109277.0	92.0
4	109282.0	95.0
...	...	...
5085	124813.0	98.0
5086	124814.0	105.0
5087	124815.0	102.0
5088	124821.0	125.0
5089	124822.0	96.0

5090 rows × 2 columns

In [18]:
```
glucose_filtered.isnull().sum()
```

Out[18]:
```
SEQN        0
LBXGLU    346
dtype: int64
```

In [19]:
```
lipids_keep = ['SEQN', 'LBXTR', 'LBDLDL']
# LBXTR: Triglycerides
# LBDLDL: LDL cholesterol
lipids_filtered = lipids[lipids_keep]
lipids_filtered
```

Out[19]:

	SEQN	LBXTR	LBDLDL
0	109264.0	40.0	86.0
1	109271.0	84.0	97.0
2	109274.0	133.0	49.0
3	109277.0	24.0	64.0
4	109282.0	132.0	164.0
...	...	...	...
5085	124813.0	45.0	96.0
5086	124814.0	74.0	160.0
5087	124815.0	38.0	128.0
5088	124821.0	51.0	101.0
5089	124822.0	75.0	91.0

5090 rows × 3 columns

In [20]:
```
lipids_filtered.isnull().sum()
```

Out[20]:
```
SEQN        0
LBXTR     440
LBDLDL    473
dtype: int64
```

In [21]:
```
hdl_keep = ['SEQN', 'LBDHDD']
# LBDHDD: HDL cholesterol
hdl_filtered = hdl[hdl_keep]
hdl_filtered
```

Out[21]:

	SEQN	LBDHDD
0	109264.0	72.0
1	109266.0	56.0
2	109270.0	47.0
3	109271.0	33.0
4	109273.0	42.0
...	...	...
12193	124817.0	60.0
12194	124818.0	50.0
12195	124820.0	64.0
12196	124821.0	44.0
12197	124822.0	65.0

12198 rows × 2 columns

In [22]:
```
hdl_filtered.isnull().sum()
```

Out[22]:
```
SEQN         0
LBDHDD    1370
dtype: int64
```

In [23]:
```
inflammation_keep = ['SEQN', 'LBXHSCRP']
# LBXHSCRP: C-reactive protein
inflammation_filtered = inflammation[inflammation_keep]
inflammation_filtered
```

Out[23]:

	SEQN	LBXHSCRP
0	109263.0	NaN
1	109264.0	0.11
2	109265.0	0.31
3	109266.0	0.72
4	109269.0	0.73
...	...	...
13767	124818.0	2.04
13768	124819.0	NaN
13769	124820.0	NaN
13770	124821.0	0.51
13771	124822.0	0.82

13772 rows × 2 columns

In [24]:
```
inflammation_filtered.isnull().sum()
```

Out[24]:
```
SEQN           0
LBXHSCRP    2158
dtype: int64
```

In [25]:
```
fasting_keep = ['SEQN', 'PHAFSTHR']
# PHAFSTHR: Hours fasted before blood draw
fasting_filtered = fasting_qst[fasting_keep]
fasting_filtered
```

Out[25]:

	SEQN	PHAFSTHR
0	109263.0	15.0
1	109264.0	10.0
2	109265.0	6.0
3	109266.0	16.0
4	109269.0	12.0
...	...	...
13767	124818.0	6.0
13768	124819.0	3.0
13769	124820.0	4.0
13770	124821.0	10.0
13771	124822.0	12.0

13772 rows × 2 columns

In [26]:
```
fasting_filtered.isnull().sum()
```

Out[26]:
```
SEQN          0
PHAFSTHR    467
dtype: int64
```

In [27]:
```
physical_activity_keep = [
    'SEQN', 'PAQ605', 'PAQ620', 'PAQ635', 'PAQ650', 'PAQ665',
]
physical_activity_filtered = physical_activity[physical_activity_keep]
physical_activity_filtered
```

Out[27]:

	SEQN	PAQ605	PAQ620	PAQ635	PAQ650	PAQ665
0	109266.0	2.0	2.0	2.0	1.0	1.0
1	109267.0	2.0	2.0	2.0	1.0	2.0
2	109268.0	1.0	1.0	2.0	2.0	2.0
3	109271.0	2.0	1.0	2.0	2.0	2.0
4	109273.0	1.0	2.0	2.0	2.0	1.0
...	...	...	...	...	...	...
9688	124815.0	1.0	2.0	1.0	1.0	1.0
9689	124817.0	1.0	2.0	2.0	2.0	2.0
9690	124818.0	2.0	2.0	2.0	2.0	2.0
9691	124821.0	1.0	2.0	2.0	2.0	2.0
9692	124822.0	2.0	1.0	2.0	1.0	1.0

9693 rows × 6 columns

In [28]:
```
physical_activity_filtered.isnull().sum()
```

Out[28]:
```
SEQN      0
PAQ605    0
PAQ620    0
PAQ635    0
PAQ650    0
PAQ665    0
dtype: int64
```

In [29]:
```
prescriptions_keep = ['SEQN', 'RXDUSE', 'RXDDRUG', 'RXDDRGID', 'RXDRSC1']
prescriptions_filtered = prescriptions[prescriptions_keep]
prescriptions_filtered['RXDUSE'] = prescriptions_filtered['RXDUSE'].replace(
for col in ['RXDDRUG', 'RXDDRGID', 'RXDRSC1']:
    prescriptions_filtered[col] = prescriptions_filtered[col].apply(
        lambda x: x.decode('utf-8') if isinstance(x, bytes) and x else 'NONE
    )
prescriptions_filtered
```

```
/var/folders/z1/w_njtpr52ss65yn2_5wp9xfm0000gn/T/ipykernel_4717/2211785616.p
y:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  prescriptions_filtered['RXDUSE'] = prescriptions_filtered['RXDUSE'].replac
e(9.0, 2.0)
/var/folders/z1/w_njtpr52ss65yn2_5wp9xfm0000gn/T/ipykernel_4717/2211785616.p
y:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  prescriptions_filtered[col] = prescriptions_filtered[col].apply(
```

Out[29]:

	SEQN	RXDUSE	RXDDRUG	RXDDRGID	RXDRSC1
0	109263.0	2.0	NONE	NONE	NONE
1	109264.0	2.0	NONE	NONE	NONE
2	109265.0	2.0	NONE	NONE	NONE
3	109266.0	2.0	NONE	NONE	NONE
4	109267.0	1.0	99999	NONE	NONE
...	...	...	...	...	...
32957	124821.0	1.0	MELOXICAM	d04532	M06.9
32958	124821.0	1.0	METOPROLOL	d00134	I21.P
32959	124821.0	1.0	TAMSULOSIN	d04121	N40
32960	124822.0	1.0	ASPIRIN	d00170	I21.P
32961	124822.0	1.0	ATORVASTATIN	d04105	E78.0

32962 rows × 5 columns

In [30]:
```
prescriptions_filtered.isnull().sum()
```

Out[30]:
```
SEQN        0
RXDUSE      0
RXDDRUG     0
RXDDRGID    0
RXDRSC1     0
dtype: int64
```

In [31]:
```
bpq_keep = ['SEQN', 'BPQ020']
# BPQ020: Has been told they had high blood pressure
bpq_filtered = bpq_qst[bpq_keep]
bpq_filtered
```

Out[31]:

	SEQN	BPQ020
0	109266.0	2.0
1	109267.0	2.0
2	109268.0	2.0
3	109271.0	2.0
4	109273.0	2.0
...	...	...
10190	124815.0	2.0
10191	124817.0	2.0
10192	124818.0	2.0
10193	124821.0	1.0
10194	124822.0	2.0

10195 rows × 2 columns

In [32]:
```
bpq_filtered.isnull().sum()
```

Out[32]:
```
SEQN      0
BPQ020    0
dtype: int64
```

In [33]:
```
diabetes_keep = [
    'SEQN',
    'DIQ010',   # Diagnosed diabetes
]
diabetes_filtered = diabetes_qst[diabetes_keep]
diabetes_filtered
```

Out[33]:

	SEQN	DIQ010
0	109263.0	2.0
1	109264.0	2.0
2	109265.0	2.0
3	109266.0	2.0
4	109267.0	2.0
...	...	...
14981	124818.0	2.0
14982	124819.0	2.0
14983	124820.0	2.0
14984	124821.0	3.0
14985	124822.0	2.0

14986 rows × 2 columns

In [34]:
```
diabetes_filtered.isnull().sum()
```

Out[34]:
```
SEQN      0
DIQ010    0
dtype: int64
```

In [35]:
```
diet_behavior_keep = [
    'SEQN',
    'DBQ700',   # Trying to lose weight
    'DBQ197',   # Eat out frequency
]
diet_behavior_filtered = diet_behavior[diet_behavior_keep]
diet_behavior_filtered
```

Out[35]:

	SEQN	DBQ700	DBQ197
0	109263.0	NaN	3.000000e+00
1	109264.0	NaN	3.000000e+00
2	109265.0	NaN	3.000000e+00
3	109266.0	3.0	2.000000e+00
4	109267.0	1.0	5.397605e-79
...	...	...	...
15555	124818.0	4.0	1.000000e+00
15556	124819.0	NaN	1.000000e+00
15557	124820.0	NaN	3.000000e+00
15558	124821.0	2.0	1.000000e+00
15559	124822.0	3.0	2.000000e+00

15560 rows × 3 columns

In [36]:
```
diet_behavior_filtered.isnull().sum()
```

Out[36]:
```
SEQN         0
DBQ700    5365
DBQ197     574
dtype: int64
```

In [37]:
```
insurance_keep = [
    'SEQN',     # unique ID
    'HIQ011',   # covered by any health insurance (Yes/No)
]
insurance_filtered = insurance[insurance_keep]
insurance_filtered
```

Out[37]:

	SEQN	HIQ011
0	109263.0	1.0
1	109264.0	1.0
2	109265.0	1.0
3	109266.0	1.0
4	109267.0	1.0
...	...	...
15555	124818.0	1.0
15556	124819.0	1.0
15557	124820.0	1.0
15558	124821.0	2.0
15559	124822.0	1.0

15560 rows × 2 columns

In [38]:
```
insurance_filtered.isnull().sum()
```

Out[38]:
```
SEQN      0
HIQ011    0
dtype: int64
```

In [39]:
```
# Start with the base DataFrame: demo_filtered
merged_df = demo_filtered.copy()

# List of all filtered DataFrames to merge
dataframes_to_merge = [
    diet_day1_filtered,
    diet_day2_filtered,
    blood_pressure_filtered,
    body_measures_filtered,
    insulin_filtered,
    glucose_filtered,
    lipids_filtered,
    hdl_filtered,
    inflammation_filtered,
    fasting_filtered,
    physical_activity_filtered,
    prescriptions_filtered,
    bpq_filtered,
    diabetes_filtered,
    diet_behavior_filtered,
    insurance_filtered
]

# Merge each one on SEQN
for df in dataframes_to_merge:
    merged_df = pd.merge(merged_df, df, on='SEQN', how='inner')
```

```
# View result
merged_df
```

Out[39]:

	SEQN	RIAGENDR	RIDAGEYR	RIDRETH3	DMDBORN4	DMDEDUC2	INDFMF
0	109271.0	1.0	49.0	3.0	1.0	2.0	N
1	109271.0	1.0	49.0	3.0	1.0	2.0	N
2	109271.0	1.0	49.0	3.0	1.0	2.0	N
3	109274.0	1.0	68.0	7.0	1.0	4.0	1.
4	109274.0	1.0	68.0	7.0	1.0	4.0	1.
...	...	...	...	...	...	...	
12138	124821.0	1.0	63.0	4.0	1.0	2.0	3
12139	124821.0	1.0	63.0	4.0	1.0	2.0	3
12140	124821.0	1.0	63.0	4.0	1.0	2.0	3
12141	124822.0	1.0	74.0	2.0	2.0	3.0	N
12142	124822.0	1.0	74.0	2.0	2.0	3.0	N

12143 rows × 65 columns

In [40]:

```
missing_dict = merged_df.isna().sum().to_dict()
for k, v in missing_dict.items():
    if v > 0:
        print(f"{k}: {v}")
```

```
DMDEDUC2: 221
INDFMPIR: 1682
DR1TKCAL: 1195
DR1TPROT: 1195
DR1TCARB: 1195
DR1TSUGR: 1195
DR1TFIBE: 1195
DR1TTFAT: 1195
DR1TSFAT: 1195
DR1TCHOL: 1195
DR1TALCO: 1195
DR1TCAFF: 1195
DR1TSODI: 1195
DR1TPOTA: 1195
DR2TKCAL: 2510
DR2TPROT: 2510
DR2TCARB: 2510
DR2TSUGR: 2510
DR2TFIBE: 2510
DR2TTFAT: 2510
DR2TSFAT: 2510
DR2TCHOL: 2510
DR2TALCO: 2510
DR2TCAFF: 2510
DR2TSODI: 2510
DR2TPOTA: 2510
BPXOSY1: 1292
BPXODI1: 1292
BPXOSY2: 1312
BPXODI2: 1312
BPXOSY3: 1366
BPXODI3: 1366
BPXOPLS1: 2144
BPXOPLS2: 2164
BPXOPLS3: 2214
BMXWT: 335
BMXHT: 353
BMXBMI: 374
BMXWAIST: 949
LBXIN: 999
LBXGLU: 724
LBXTR: 928
LBDLDL: 1052
LBDHDD: 913
LBXHSCRP: 1008
PHAFSTHR: 120
```

In [41]:
```
merged_df.dropna(inplace=True)
```

In [42]:
```
merged_df
```

Out[42]:

	SEQN	RIAGENDR	RIDAGEYR	RIDRETH3	DMDBORN4	DMDEDUC2	INDFMF
16	109290.0	2.0	68.0	4.0	1.0	5.0	
17	109290.0	2.0	68.0	4.0	1.0	5.0	
18	109290.0	2.0	68.0	4.0	1.0	5.0	
19	109290.0	2.0	68.0	4.0	1.0	5.0	
29	109300.0	2.0	54.0	6.0	2.0	5.0	
...	...	...	...	...	...	...	
12131	124812.0	2.0	62.0	2.0	1.0	4.0	
12132	124812.0	2.0	62.0	2.0	1.0	4.0	
12133	124812.0	2.0	62.0	2.0	1.0	4.0	
12135	124814.0	1.0	64.0	4.0	1.0	3.0	
12136	124815.0	1.0	52.0	4.0	1.0	4.0	

6160 rows × 65 columns

In [43]:
```
merged_df.columns
```

Out[43]:
```
Index(['SEQN', 'RIAGENDR', 'RIDAGEYR', 'RIDRETH3', 'DMDBORN4', 'DMDEDUC2',
       'INDFMPIR', 'DR1TKCAL', 'DR1TPROT', 'DR1TCARB', 'DR1TSUGR', 'DR1TFIB
E',
       'DR1TTFAT', 'DR1TSFAT', 'DR1TCHOL', 'DR1TALCO', 'DR1TCAFF', 'DR1TSOD
I',
       'DR1TPOTA', 'DR2TKCAL', 'DR2TPROT', 'DR2TCARB', 'DR2TSUGR', 'DR2TFIB
E',
       'DR2TTFAT', 'DR2TSFAT', 'DR2TCHOL', 'DR2TALCO', 'DR2TCAFF', 'DR2TSOD
I',
       'DR2TPOTA', 'BPXOSY1', 'BPXODI1', 'BPXOSY2', 'BPXODI2', 'BPXOSY3',
       'BPXODI3', 'BPXOPLS1', 'BPXOPLS2', 'BPXOPLS3', 'BMXWT', 'BMXHT',
       'BMXBMI', 'BMXWAIST', 'LBXIN', 'LBXGLU', 'LBXTR', 'LBDLDL', 'LBDHD
D',
       'LBXHSCRP', 'PHAFSTHR', 'PAQ605', 'PAQ620', 'PAQ635', 'PAQ650',
       'PAQ665', 'RXDUSE', 'RXDDRUG', 'RXDDRGID', 'RXDRSC1', 'BPQ020',
       'DIQ010', 'DBQ700', 'DBQ197', 'HIQ011'],
      dtype='object')
```

In [44]:
```
merged_df['RXDUSE'].value_counts()
```

Out[44]:
```
RXDUSE
1.0    5283
2.0     877
Name: count, dtype: int64
```

In [45]:
```
merged_df.to_csv('clean.csv', index=False)
```