# Data Clearning - Medical Data Set

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
```

## Code used to import data set and review data information.

In [2]:

```
data= pd.read_csv('C:/Users/cynth/OneDrive/Documents/MS Data Analytics/medical_clean.cs
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   CaseOrder           10000 non-null  int64  
 1   Customer_id         10000 non-null  object 
 2   Interaction         10000 non-null  object 
 3   UID                 10000 non-null  object 
 4   City                10000 non-null  object 
 5   State               10000 non-null  object 
 6   County              10000 non-null  object 
 7   Zip                 10000 non-null  int64  
 8   Lat                 10000 non-null  float64
 9   Lng                 10000 non-null  float64
 10  Population          10000 non-null  int64  
 11  Area                10000 non-null  object 
 12  TimeZone            10000 non-null  object 
 13  Job                 10000 non-null  object 
 14  Children            10000 non-null  int64  
 15  Age                 10000 non-null  int64  
 16  Income              10000 non-null  float64
 17  Marital             10000 non-null  object 
 18  Gender              10000 non-null  object 
 19  ReAdmis             10000 non-null  object 
 20  VitD_levels         10000 non-null  float64
 21  Doc_visits          10000 non-null  int64  
 22  Full_meals_eaten    10000 non-null  int64  
 23  vitD_supp           10000 non-null  int64  
 24  Soft_drink          10000 non-null  object 
 25  Initial_admin       10000 non-null  object 
 26  HighBlood           10000 non-null  object 
 27  Stroke              10000 non-null  object 
 28  Complication_risk   10000 non-null  object 
 29  Overweight          10000 non-null  object 
 30  Arthritis           10000 non-null  object 
 31  Diabetes            10000 non-null  object 
 32  Hyperlipidemia      10000 non-null  object 
 33  BackPain            10000 non-null  object 
 34  Anxiety             10000 non-null  object 
 35  Allergic_rhinitis   10000 non-null  object 
 36  Reflux_esophagitis  10000 non-null  object 
 37  Asthma              10000 non-null  object 
 38  Services            10000 non-null  object 
 39  Initial_days        10000 non-null  float64
 40  TotalCharge         10000 non-null  float64
 41  Additional_charges  10000 non-null  float64
 42  Item1               10000 non-null  int64
 43  Item2               10000 non-null  int64
 44  Item3               10000 non-null  int64
 45  Item4               10000 non-null  int64
 46  Item5               10000 non-null  int64
 47  Item6               10000 non-null  int64
 48  Item7               10000 non-null  int64
 49  Item8               10000 non-null  int64
dtypes: float64(7), int64(16), object(27)
memory usage: 3.8+ MB
```

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

```
Out[3]: Index(['CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State',
               'County', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'TimeZone', 'Job',
               'Children', 'Age', 'Income', 'Marital', 'Gender', 'ReAdmis',
               'VitD_levels', 'Doc_visits', 'Full_meals_eaten', 'vitD_supp',
               'Soft_drink', 'Initial_admin', 'HighBlood', 'Stroke',
               'Complication_risk', 'Overweight', 'Arthritis', 'Diabetes',
               'Hyperlipidemia', 'BackPain', 'Anxiety', 'Allergic_rhinitis',
               'Reflux_esophagitis', 'Asthma', 'Services', 'Initial_days',
               'TotalCharge', 'Additional_charges', 'Item1', 'Item2', 'Item3', 'Item4',
               'Item5', 'Item6', 'Item7', 'Item8'],
              dtype='object')
```

# Checking for Duplicated Data

```
In [4]: data.loc[data.duplicated()]
```

Out[4]:

| | CaseOrder | Customer_id | Interaction | UID | City | State | County | Zip | Lat | Lng | ... | TotalCharge | Addit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

0 rows × 50 columns

# Categorical Data Preparation: converting categorical variables to type categorical and converting to dummy variables as necessary.

```
In [5]: data['Area'].unique()
```

```
Out[5]: array(['Suburban', 'Urban', 'Rural'], dtype=object)
```

```
In [6]: data['Area'].value_counts()
```

```
Out[6]: Rural       3369
        Suburban    3328
        Urban       3303
        Name: Area, dtype: int64
```

```
In [7]: data['Area'] = pd.Categorical(data['Area'],['Urban', 'Suburban', 'Rural'] )
In [8]:  Area_dummies = pd.get_dummies(data.Area, prefix='Area').iloc[:, 1:]
```

```
In [9]:  data = data.drop(['Area'], axis=1)
```

```
In [10]: data = pd.concat([data, Area_dummies], axis=1)
```

```
In [11]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 51 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   CaseOrder           10000 non-null  int64  
 1   Customer_id         10000 non-null  object 
 2   Interaction         10000 non-null  object 
 3   UID                 10000 non-null  object 
 4   City                10000 non-null  object 
 5   State               10000 non-null  object 
 6   County              10000 non-null  object 
 7   Zip                 10000 non-null  int64  
 8   Lat                 10000 non-null  float64
 9   Lng                 10000 non-null  float64
 10  Population          10000 non-null  int64  
 11  TimeZone            10000 non-null  object 
 12  Job                 10000 non-null  object 
 13  Children            10000 non-null  int64  
 14  Age                 10000 non-null  int64  
 15  Income              10000 non-null  float64
 16  Marital             10000 non-null  object 
 17  Gender              10000 non-null  object 
 18  ReAdmis             10000 non-null  object 
 19  VitD_levels         10000 non-null  float64
 20  Doc_visits          10000 non-null  int64  
 21  Full_meals_eaten    10000 non-null  int64  
 22  vitD_supp           10000 non-null  int64  
 23  Soft_drink          10000 non-null  object 
 24  Initial_admin       10000 non-null  object 
 25  HighBlood           10000 non-null  object 
 26  Stroke              10000 non-null  object 
 27  Complication_risk   10000 non-null  object 
 28  Overweight          10000 non-null  object 
 29  Arthritis           10000 non-null  object 
 30  Diabetes            10000 non-null  object 
 31  Hyperlipidemia      10000 non-null  object 
 32  BackPain            10000 non-null  object 
 33  Anxiety             10000 non-null  object 
 34  Allergic_rhinitis   10000 non-null  object 
 35  Reflux_esophagitis  10000 non-null  object 
 36  Asthma              10000 non-null  object 
 37  Services            10000 non-null  object 
 38  Initial_days        10000 non-null  float64
 39  TotalCharge         10000 non-null  float64
 40  Additional_charges  10000 non-null  float64
 41  Item1               10000 non-null  int64  
 42  Item2               10000 non-null  int64  
 43  Item3               10000 non-null  int64  
 44  Item4               10000 non-null  int64  
 45  Item5               10000 non-null  int64  
 46  Item6               10000 non-null  int64  
 47  Item7               10000 non-null  int64  
 48  Item8               10000 non-null  int64  
 49  Area_Suburban       10000 non-null  uint8  
 50  Area_Rural          10000 non-null  uint8
dtypes: float64(7), int64(16), object(26), uint8(2)
memory usage: 3.8+ MB
```

In [12]:
```
data['Marital'].unique()
```

Out[12]:
```
array(['Divorced', 'Married', 'Widowed', 'Never Married', 'Separated'],
      dtype=object)
```

In [13]:
```
data['Marital'].value_counts()
```

Out[13]:
```
Widowed          2045
Married          2023
Separated        1987
Never Married    1984
Divorced         1961
Name: Marital, dtype: int64
```

In [14]:
```
data['Marital'] = pd.Categorical(data['Marital'], ['Widowed', 'Married', 'Separated', '
```

In [15]:
```
Marital_dummies = pd.get_dummies(data.Marital, prefix='Marital').iloc[:, 1:]
```

In [16]:
```
data = data.drop(['Marital'], axis=1)
```

In [17]:
```
data = pd.concat([data, Marital_dummies], axis=1)
```

In [18]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 54 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   CaseOrder              10000 non-null  int64
 1   Customer_id            10000 non-null  object
 2   Interaction            10000 non-null  object
 3   UID                    10000 non-null  object
 4   City                   10000 non-null  object
 5   State                  10000 non-null  object
 6   County                 10000 non-null  object
 7   Zip                    10000 non-null  int64
 8   Lat                    10000 non-null  float64
 9   Lng                    10000 non-null  float64
 10  Population             10000 non-null  int64
 11  TimeZone               10000 non-null  object
 12  Job                    10000 non-null  object
 13  Children               10000 non-null  int64
 14  Age                    10000 non-null  int64
 15  Income                 10000 non-null  float64
 16  Gender                 10000 non-null  object
 17  ReAdmis                10000 non-null  object
 18  VitD_levels            10000 non-null  float64
 19  Doc_visits             10000 non-null  int64
 20  Full_meals_eaten       10000 non-null  int64
 21  vitD_supp              10000 non-null  int64
 22  Soft_drink             10000 non-null  object
 23  Initial_admin          10000 non-null  object
 24  HighBlood              10000 non-null  object
 25  Stroke               10000 non-null  object 
 26  Complication_risk    10000 non-null  object 
 27  Overweight           10000 non-null  object 
 28  Arthritis            10000 non-null  object 
 29  Diabetes             10000 non-null  object 
 30  Hyperlipidemia       10000 non-null  object 
 31  BackPain             10000 non-null  object 
 32  Anxiety              10000 non-null  object 
 33  Allergic_rhinitis    10000 non-null  object 
 34  Reflux_esophagitis   10000 non-null  object 
 35  Asthma               10000 non-null  object 
 36  Services             10000 non-null  object 
 37  Initial_days         10000 non-null  float64
 38  TotalCharge          10000 non-null  float64
 39  Additional_charges   10000 non-null  float64
 40  Item1                10000 non-null  int64  
 41  Item2                10000 non-null  int64  
 42  Item3                10000 non-null  int64  
 43  Item4                10000 non-null  int64  
 44  Item5                10000 non-null  int64  
 45  Item6                10000 non-null  int64  
 46  Item7                10000 non-null  int64  
 47  Item8                10000 non-null  int64  
 48  Area_Suburban        10000 non-null  uint8  
 49  Area_Rural           10000 non-null  uint8  
 50  Marital_Married      10000 non-null  uint8  
 51  Marital_Separated    10000 non-null  uint8  
 52  Marital_Never Married  10000 non-null  uint8  
 53  Marital_Divorced     10000 non-null  uint8  
dtypes: float64(7), int64(16), object(25), uint8(6)
memory usage: 3.7+ MB
```

In [19]:
```
data['Gender'].unique()
```

Out[19]:
```
array(['Male', 'Female', 'Nonbinary'], dtype=object)
```

In [20]:
```
data['Gender'].value_counts()
```

Out[20]:
```
Female       5018
Male         4768
Nonbinary     214
Name: Gender, dtype: int64
```

In [21]:
```
data['Gender'] = pd.Categorical(data['Gender'], ['Male', 'Female', 'Nonbinary'])
```

In [22]:
```
Gender_dummies = pd.get_dummies(data.Gender, prefix='Gender').iloc[:, 1:]
Gender_dummies.value_counts()
```

Out[22]:
```
Gender_Female  Gender_Nonbinary
1              0                   5018
0              0                   4768
               1                    214
dtype: int64
```

In [23]:
```
data = data.drop(['Gender'], axis=1)
```

In [24]:
```
data = pd.concat([data, Gender_dummies], axis=1)
```

In [25]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 55 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   CaseOrder            10000 non-null  int64  
 1   Customer_id          10000 non-null  object 
 2   Interaction          10000 non-null  object 
 3   UID                  10000 non-null  object 
 4   City                 10000 non-null  object 
 5   State                10000 non-null  object 
 6   County               10000 non-null  object 
 7   Zip                  10000 non-null  int64  
 8   Lat                  10000 non-null  float64
 9   Lng                  10000 non-null  float64
 10  Population           10000 non-null  int64  
 11  TimeZone             10000 non-null  object 
 12  Job                  10000 non-null  object 
 13  Children             10000 non-null  int64  
 14  Age                  10000 non-null  int64  
 15  Income               10000 non-null  float64
 16  ReAdmis              10000 non-null  object 
 17  VitD_levels          10000 non-null  float64
 18  Doc_visits           10000 non-null  int64  
 19  Full_meals_eaten     10000 non-null  int64  
 20  vitD_supp            10000 non-null  int64  
 21  Soft_drink           10000 non-null  object 
 22  Initial_admin        10000 non-null  object 
 23  HighBlood            10000 non-null  object 
 24  Stroke               10000 non-null  object 
 25  Complication_risk    10000 non-null  object 
 26  Overweight           10000 non-null  object 
 27  Arthritis            10000 non-null  object 
 28  Diabetes             10000 non-null  object 
 29  Hyperlipidemia       10000 non-null  object 
 30  BackPain             10000 non-null  object 
 31  Anxiety              10000 non-null  object 
 32  Allergic_rhinitis    10000 non-null  object 
 33  Reflux_esophagitis   10000 non-null  object 
 34  Asthma               10000 non-null  object 
 35  Services             10000 non-null  object 
 36  Initial_days         10000 non-null  float64
 37  TotalCharge          10000 non-null  float64
 38  Additional_charges   10000 non-null  float64
 39  Item1                10000 non-null  int64  
 40  Item2                10000 non-null  int64  
 41  Item3                10000 non-null  int64  
 42  Item4                10000 non-null  int64  
 43  Item5                10000 non-null  int64  
 44  Item6                10000 non-null  int64  
 45  Item7                10000 non-null  int64  
 46  Item8                10000 non-null  int64  
 47  Area_Suburban        10000 non-null  uint8  
 48  Area_Rural           10000 non-null  uint8  
 49  Marital_Married      10000 non-null  uint8  
 50  Marital_Separated    10000 non-null  uint8  
 51  Marital_Never Married  10000 non-null  uint8  
 52  Marital_Divorced     10000 non-null  uint8  
 53  Gender_Female        10000 non-null  uint8  
 54  Gender_Nonbinary     10000 non-null  uint8  
dtypes: float64(7), int64(16), object(24), uint8(8)
memory usage: 3.7+ MB
```

In [26]:
```
data['ReAdmis'].unique()
```

Out[26]:
```
array(['No', 'Yes'], dtype=object)
```

In [27]:
```
data['ReAdmis'].value_counts()
```

Out[27]:
```
No     6331
Yes    3669
Name: ReAdmis, dtype: int64
```

In [28]:
```
data['ReAdmis'] = pd.Categorical(data['ReAdmis'], ['No', 'Yes'])
```

In [29]:
```
data['ReAdmis']= data['ReAdmis'].cat.codes
data['ReAdmis'].value_counts()
```

Out[29]:
```
0    6331
1    3669
Name: ReAdmis, dtype: int64
```

In [30]:
```
data['Soft_drink'].unique()
```

Out[30]:
```
array(['No', 'Yes'], dtype=object)
```

In [31]:
```
data['Soft_drink'].value_counts()
```

Out[31]:
```
No     7425
Yes    2575
Name: Soft_drink, dtype: int64
```

In [32]:
```
data['Soft_drink'] = pd.Categorical(data['Soft_drink'], ['No', 'Yes'])
```

In [33]:
```
data['Soft_drink']= data['Soft_drink'].cat.codes
data['Soft_drink'].value_counts()
```

Out[33]:
```
0    7425
1    2575
Name: Soft_drink, dtype: int64
```

In [34]:
```
data['Initial_admin'].unique()
```

Out[34]:
```
array(['Emergency Admission', 'Elective Admission',
       'Observation Admission'], dtype=object)
```

In [35]:
```
data['Initial_admin'].value_counts()
```

Out[35]:
```
Emergency Admission      5060
Elective Admission       2504
Observation Admission    2436
Name: Initial_admin, dtype: int64
```

In [36]:

```
data['Initial_admin'] = pd.Categorical(data['Initial_admin'], ['Emergency Admission', '
              'Observation Admission'])
```

In [37]:

```
Initial_admin_dummies = pd.get_dummies(data.Initial_admin, prefix='Initial_admin').iloc
Initial_admin_dummies.value_counts()
```

Out[37]:

```
Initial_admin_Elective Admission  Initial_admin_Observation Admission
0                                 0                                      5060
1                                 0                                      2504
0                                 1                                      2436
dtype: int64
```

In [38]:

```
data = data.drop(['Initial_admin'], axis=1)
```

In [39]:

```
data = pd.concat([data, Initial_admin_dummies], axis=1)
```

In [40]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 56 columns):
 #   Column                                    Non-Null Count  Dtype  
---  ------                                    --------------  -----  
 0   CaseOrder                                 10000 non-null  int64  
 1   Customer_id                               10000 non-null  object 
 2   Interaction                               10000 non-null  object 
 3   UID                                       10000 non-null  object 
 4   City                                      10000 non-null  object 
 5   State                                     10000 non-null  object 
 6   County                                    10000 non-null  object 
 7   Zip                                       10000 non-null  int64  
 8   Lat                                       10000 non-null  float64
 9   Lng                                       10000 non-null  float64
 10  Population                                10000 non-null  int64  
 11  TimeZone                                  10000 non-null  object 
 12  Job                                       10000 non-null  object 
 13  Children                                  10000 non-null  int64  
 14  Age                                       10000 non-null  int64  
 15  Income                                    10000 non-null  float64
 16  ReAdmis                                   10000 non-null  int8   
 17  VitD_levels                               10000 non-null  float64
 18  Doc_visits                                10000 non-null  int64  
 19  Full_meals_eaten                          10000 non-null  int64  
 20  vitD_supp                                 10000 non-null  int64  
 21  Soft_drink                                10000 non-null  int8   
 22  HighBlood                                 10000 non-null  object 
 23  Stroke                                    10000 non-null  object 
 24  Complication_risk                         10000 non-null  object 
 25  Overweight                                10000 non-null  object 
 26  Arthritis                                 10000 non-null  object 
 27  Diabetes                                  10000 non-null  object 
 28  Hyperlipidemia                            10000 non-null  object 
 29  BackPain                                  10000 non-null  object 
 30  Anxiety                                   10000 non-null  object 
 31  Allergic_rhinitis                         10000 non-null  object 
 32  Reflux_esophagitis                        10000 non-null  object 
 33  Asthma                                    10000 non-null  object 
 34  Services                                  10000 non-null  object 
 35  Initial_days                              10000 non-null  float64
 36  TotalCharge                               10000 non-null  float64
 37  Additional_charges                    10000 non-null  float64
 38  Item1                                 10000 non-null  int64
 39  Item2                                 10000 non-null  int64
 40  Item3                                 10000 non-null  int64
 41  Item4                                 10000 non-null  int64
 42  Item5                                 10000 non-null  int64
 43  Item6                                 10000 non-null  int64
 44  Item7                                 10000 non-null  int64
 45  Item8                                 10000 non-null  int64
 46  Area_Suburban                         10000 non-null  uint8
 47  Area_Rural                            10000 non-null  uint8
 48  Marital_Married                       10000 non-null  uint8
 49  Marital_Separated                     10000 non-null  uint8
 50  Marital_Never Married                 10000 non-null  uint8
 51  Marital_Divorced                      10000 non-null  uint8
 52  Gender_Female                         10000 non-null  uint8
 53  Gender_Nonbinary                      10000 non-null  uint8
 54  Initial_admin_Elective Admission      10000 non-null  uint8
 55  Initial_admin_Observation Admission   10000 non-null  uint8
dtypes: float64(7), int64(16), int8(2), object(21), uint8(10)
memory usage: 3.5+ MB
```

In [41]:

```
data['HighBlood'].unique()
```

Out[41]:

```
array(['Yes', 'No'], dtype=object)
```

In [42]:

```
data['HighBlood'].value_counts()
```

Out[42]:

```
No     5910
Yes    4090
Name: HighBlood, dtype: int64
```

In [43]:

```
data['HighBlood'] = pd.Categorical(data['HighBlood'], ['No', 'Yes'])
```

In [44]:

```
data['HighBlood']= data['HighBlood'].cat.codes
data['HighBlood'].value_counts()
```

Out[44]:

```
0    5910
1    4090
Name: HighBlood, dtype: int64
```

In [45]:

```
data['Stroke'].unique()
```

Out[45]:

```
array(['No', 'Yes'], dtype=object)
```

In [46]:

```
data['Stroke'].value_counts()
```

Out[46]:

```
No     8007
Yes    1993
Name: Stroke, dtype: int64
```

In [47]:

```
data['Stroke'] = pd.Categorical(data['Stroke'], ['No', 'Yes'])
```

In [48]:

```
data['Stroke']= data['Stroke'].cat.codes
```

```
data['Stroke'].value_counts()
```

Out[48]:
```
0    8007
1    1993
Name: Stroke, dtype: int64
```

In [49]:
```
data['Complication_risk'].unique()
```

Out[49]:
```
array(['Medium', 'High', 'Low'], dtype=object)
```

In [50]:
```
data['Complication_risk'].value_counts()
```

Out[50]:
```
Medium    4517
High      3358
Low       2125
Name: Complication_risk, dtype: int64
```

In [51]:
```
data['Complication_risk'] = pd.Categorical(data['Complication_risk'], ['Medium', 'High'
```

In [52]:
```
Complication_risk_dummies = pd.get_dummies(data.Complication_risk, prefix='Complication
Complication_risk_dummies.value_counts()
```

Out[52]:
```
Complication_risk_High  Complication_risk_Low
0                       0                        4517
1                       0                        3358
0                       1                        2125
dtype: int64
```

In [53]:
```
data = data.drop(['Complication_risk'], axis=1)
```

In [54]:
```
data = pd.concat([data, Complication_risk_dummies], axis=1)
```

In [55]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 57 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   CaseOrder                             10000 non-null  int64  
 1   Customer_id                           10000 non-null  object 
 2   Interaction                           10000 non-null  object 
 3   UID                                   10000 non-null  object 
 4   City                                  10000 non-null  object 
 5   State                                 10000 non-null  object 
 6   County                                10000 non-null  object 
 7   Zip                                   10000 non-null  int64  
 8   Lat                                   10000 non-null  float64
 9   Lng                                   10000 non-null  float64
 10  Population                            10000 non-null  int64  
 11  TimeZone                              10000 non-null  object 
 12  Job                                   10000 non-null  object 
 13  Children                              10000 non-null  int64  
 14  Age                                   10000 non-null  int64  
 15  Income                                10000 non-null  float64
 16  ReAdmis                               10000 non-null  int8   
 17  VitD_levels                           10000 non-null  float64
 18  Doc_visits                            10000 non-null  int64
 19  Full_meals_eaten                      10000 non-null  int64
 20  vitD_supp                             10000 non-null  int64
 21  Soft_drink                            10000 non-null  int8
 22  HighBlood                             10000 non-null  int8
 23  Stroke                                10000 non-null  int8
 24  Overweight                            10000 non-null  object
 25  Arthritis                             10000 non-null  object
 26  Diabetes                              10000 non-null  object
 27  Hyperlipidemia                        10000 non-null  object
 28  BackPain                              10000 non-null  object
 29  Anxiety                               10000 non-null  object
 30  Allergic_rhinitis                     10000 non-null  object
 31  Reflux_esophagitis                    10000 non-null  object
 32  Asthma                                10000 non-null  object
 33  Services                              10000 non-null  object
 34  Initial_days                          10000 non-null  float64
 35  TotalCharge                           10000 non-null  float64
 36  Additional_charges                    10000 non-null  float64
 37  Item1                                 10000 non-null  int64
 38  Item2                                 10000 non-null  int64
 39  Item3                                 10000 non-null  int64
 40  Item4                                 10000 non-null  int64
 41  Item5                                 10000 non-null  int64
 42  Item6                                 10000 non-null  int64
 43  Item7                                 10000 non-null  int64
 44  Item8                                 10000 non-null  int64
 45  Area_Suburban                         10000 non-null  uint8
 46  Area_Rural                            10000 non-null  uint8
 47  Marital_Married                       10000 non-null  uint8
 48  Marital_Separated                     10000 non-null  uint8
 49  Marital_Never Married                 10000 non-null  uint8
 50  Marital_Divorced                      10000 non-null  uint8
 51  Gender_Female                         10000 non-null  uint8
 52  Gender_Nonbinary                      10000 non-null  uint8
 53  Initial_admin_Elective Admission      10000 non-null  uint8
 54  Initial_admin_Observation Admission   10000 non-null  uint8
 55  Complication_risk_High                10000 non-null  uint8
 56  Complication_risk_Low                 10000 non-null  uint8
dtypes: float64(7), int64(16), int8(4), object(18), uint8(12)
memory usage: 3.3+ MB
```

In [56]:

```
data['Overweight'].unique()
```

Out[56]:

```
array(['No', 'Yes'], dtype=object)
```

In [57]:

```
data['Overweight'].value_counts()
```

Out[57]:

```
Yes    7094
No     2906
Name: Overweight, dtype: int64
```

In [58]:

```
data['Overweight'] = pd.Categorical(data['Overweight'], ['No', 'Yes'])
```

In [59]:

```
data['Overweight']= data['Overweight'].cat.codes
data['Overweight'].value_counts()
```

Out[59]:

```
1    7094
```

```
0    2906
Name: Overweight, dtype: int64
```

In [60]:

```
data['Arthritis'].unique()
```

Out[60]:

```
array(['Yes', 'No'], dtype=object)
```

In [61]:

```
data['Arthritis'].value_counts()
```

Out[61]:

```
No     6426
Yes    3574
Name: Arthritis, dtype: int64
```

In [62]:

```
data['Arthritis'] = pd.Categorical(data['Arthritis'], ['No', 'Yes'])
```

In [63]:

```
data['Arthritis']= data['Arthritis'].cat.codes
data['Arthritis'].value_counts()
```

Out[63]:

```
0    6426
1    3574
Name: Arthritis, dtype: int64
```

In [64]:

```
data['Diabetes'].unique()
```

Out[64]:

```
array(['Yes', 'No'], dtype=object)
```

In [65]:

```
data['Diabetes'].value_counts()
```

Out[65]:

```
No     7262
Yes    2738
Name: Diabetes, dtype: int64
```

In [66]:

```
data['Diabetes'] = pd.Categorical(data['Diabetes'], ['No', 'Yes'])
```

In [67]:

```
data['Diabetes']= data['Diabetes'].cat.codes
data['Diabetes'].value_counts()
```

Out[67]:

```
0    7262
1    2738
Name: Diabetes, dtype: int64
```

In [68]:

```
data['Hyperlipidemia'].unique()
```

Out[68]:

```
array(['No', 'Yes'], dtype=object)
```

In [69]:

```
data['Hyperlipidemia'].value_counts()
```

Out[69]:

```
No     6628
Yes    3372
Name: Hyperlipidemia, dtype: int64
```

In [70]:

```
data['Hyperlipidemia'] = pd.Categorical(data['Hyperlipidemia'], ['No', 'Yes'])
```

In [71]:

```
data['Hyperlipidemia']= data['Hyperlipidemia'].cat.codes
data['Hyperlipidemia'].value_counts()
```

Out[71]:

```
0    6628
1    3372
Name: Hyperlipidemia, dtype: int64
```

In [72]:

```
data['BackPain'].unique()
```

Out[72]:

```
array(['Yes', 'No'], dtype=object)
```

In [73]:

```
data['BackPain'].value_counts()
```

Out[73]:

```
No     5886
Yes    4114
Name: BackPain, dtype: int64
```

In [74]:

```
data['BackPain'] = pd.Categorical(data['BackPain'], ['No', 'Yes'])
```

In [75]:

```
data['BackPain']= data['BackPain'].cat.codes
data['BackPain'].value_counts()
```

Out[75]:

```
0    5886
1    4114
Name: BackPain, dtype: int64
```

In [76]:

```
data['Anxiety'].unique()
```

Out[76]:

```
array(['Yes', 'No'], dtype=object)
```

In [77]:

```
data['Anxiety'].value_counts()
```

Out[77]:

```
No     6785
Yes    3215
Name: Anxiety, dtype: int64
```

In [78]:

```
data['Anxiety'] = pd.Categorical(data['Anxiety'], ['No', 'Yes'])
```

In [79]:

```
data['Anxiety']= data['Anxiety'].cat.codes
data['Anxiety'].value_counts()
```

Out[79]:

```
0    6785
1    3215
Name: Anxiety, dtype: int64
```

In [80]:

```
data['Allergic_rhinitis'].unique()
```

```
Out[80]: array(['Yes', 'No'], dtype=object)
```

```
In [81]: data['Allergic_rhinitis'].value_counts()
```

```
Out[81]: No     6059
         Yes    3941
         Name: Allergic_rhinitis, dtype: int64
```

```
In [82]: data['Allergic_rhinitis'] = pd.Categorical(data['Allergic_rhinitis'], ['No', 'Yes'])
```

```
In [83]: data['Allergic_rhinitis']= data['Allergic_rhinitis'].cat.codes
         data['Allergic_rhinitis'].value_counts()
```

```
Out[83]: 0    6059
         1    3941
         Name: Allergic_rhinitis, dtype: int64
```

```
In [84]: data['Reflux_esophagitis'].unique()
```

```
Out[84]: array(['No', 'Yes'], dtype=object)
```

```
In [85]: data['Reflux_esophagitis'].value_counts()
```

```
Out[85]: No     5865
         Yes    4135
         Name: Reflux_esophagitis, dtype: int64
```

```
In [86]: data['Reflux_esophagitis'] = pd.Categorical(data['Reflux_esophagitis'], ['No', 'Yes'])
```

```
In [87]: data['Reflux_esophagitis']= data['Reflux_esophagitis'].cat.codes
         data['Reflux_esophagitis'].value_counts()
```

```
Out[87]: 0    5865
         1    4135
         Name: Reflux_esophagitis, dtype: int64
```

```
In [88]: data['Asthma'].unique()
```

```
Out[88]: array(['Yes', 'No'], dtype=object)
```

```
In [89]: data['Asthma'].value_counts()
```

```
Out[89]: No     7107
         Yes    2893
         Name: Asthma, dtype: int64
```

```
In [90]: data['Asthma'] = pd.Categorical(data['Asthma'], ['No', 'Yes'])
```

```
In [91]: data['Asthma']= data['Asthma'].cat.codes
         data['Asthma'].value_counts()
```

Out[91]:
```
0    7107
1    2893
Name: Asthma, dtype: int64
```

In [92]:
```
data['Services'].unique()
```

Out[92]:
```
array(['Blood Work', 'Intravenous', 'CT Scan', 'MRI'], dtype=object)
```

In [93]:
```
data['Services'].value_counts()
```

Out[93]:
```
Blood Work     5265
Intravenous    3130
CT Scan        1225
MRI             380
Name: Services, dtype: int64
```

In [94]:
```
data['Services'] = pd.Categorical(data['Services'], ['Blood Work', 'Intravenous', 'CT S
```

In [95]:
```
Services_dummies = pd.get_dummies(data.Services, prefix='Services').iloc[:, 1:]
Services_dummies.value_counts()
```

Out[95]:
```
Services_Intravenous  Services_CT Scan  Services_MRI
0                     0                 0               5265
1                     0                 0               3130
0                     1                 0               1225
                      0                 1                380
dtype: int64
```

In [96]:
```
data = data.drop(['Services'], axis=1)
```

In [97]:
```
data = pd.concat([data, Services_dummies], axis=1)
```

In [98]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 59 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   CaseOrder                           10000 non-null  int64
 1   Customer_id                         10000 non-null  object
 2   Interaction                         10000 non-null  object
 3   UID                                 10000 non-null  object
 4   City                                10000 non-null  object
 5   State                               10000 non-null  object
 6   County                              10000 non-null  object
 7   Zip                                 10000 non-null  int64
 8   Lat                                 10000 non-null  float64
 9   Lng                                 10000 non-null  float64
 10  Population                          10000 non-null  int64
 11  TimeZone                            10000 non-null  object
 12  Job                                 10000 non-null  object
 13  Children                            10000 non-null  int64
 14  Age                                 10000 non-null  int64
 15  Income                              10000 non-null  float64
 16  ReAdmis                             10000 non-null  int8
 17  VitD_levels                           10000 non-null  float64
 18  Doc_visits                            10000 non-null  int64
 19  Full_meals_eaten                      10000 non-null  int64
 20  vitD_supp                             10000 non-null  int64
 21  Soft_drink                            10000 non-null  int8
 22  HighBlood                             10000 non-null  int8
 23  Stroke                                10000 non-null  int8
 24  Overweight                            10000 non-null  int8
 25  Arthritis                             10000 non-null  int8
 26  Diabetes                              10000 non-null  int8
 27  Hyperlipidemia                        10000 non-null  int8
 28  BackPain                              10000 non-null  int8
 29  Anxiety                               10000 non-null  int8
 30  Allergic_rhinitis                     10000 non-null  int8
 31  Reflux_esophagitis                    10000 non-null  int8
 32  Asthma                                10000 non-null  int8
 33  Initial_days                          10000 non-null  float64
 34  TotalCharge                           10000 non-null  float64
 35  Additional_charges                    10000 non-null  float64
 36  Item1                                 10000 non-null  int64
 37  Item2                                 10000 non-null  int64
 38  Item3                                 10000 non-null  int64
 39  Item4                                 10000 non-null  int64
 40  Item5                                 10000 non-null  int64
 41  Item6                                 10000 non-null  int64
 42  Item7                                 10000 non-null  int64
 43  Item8                                 10000 non-null  int64
 44  Area_Suburban                         10000 non-null  uint8
 45  Area_Rural                            10000 non-null  uint8
 46  Marital_Married                       10000 non-null  uint8
 47  Marital_Separated                     10000 non-null  uint8
 48  Marital_Never Married                 10000 non-null  uint8
 49  Marital_Divorced                      10000 non-null  uint8
 50  Gender_Female                         10000 non-null  uint8
 51  Gender_Nonbinary                      10000 non-null  uint8
 52  Initial_admin_Elective Admission      10000 non-null  uint8
 53  Initial_admin_Observation Admission   10000 non-null  uint8
 54  Complication_risk_High                10000 non-null  uint8
 55  Complication_risk_Low                 10000 non-null  uint8
 56  Services_Intravenous                  10000 non-null  uint8
 57  Services_CT Scan                      10000 non-null  uint8
 58  Services_MRI                          10000 non-null  uint8
dtypes: float64(7), int64(16), int8(13), object(8), uint8(15)
memory usage: 2.6+ MB
```

In [99]:

```
data.columns
```

Out[99]:

```
Index(['CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State',
       'County', 'Zip', 'Lat', 'Lng', 'Population', 'TimeZone', 'Job',
       'Children', 'Age', 'Income', 'ReAdmis', 'VitD_levels', 'Doc_visits',
       'Full_meals_eaten', 'vitD_supp', 'Soft_drink', 'HighBlood', 'Stroke',
       'Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain',
       'Anxiety', 'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma',
       'Initial_days', 'TotalCharge', 'Additional_charges', 'Item1', 'Item2',
       'Item3', 'Item4', 'Item5', 'Item6', 'Item7', 'Item8', 'Area_Suburban',
       'Area_Rural', 'Marital_Married', 'Marital_Separated',
       'Marital_Never Married', 'Marital_Divorced', 'Gender_Female',
       'Gender_Nonbinary', 'Initial_admin_Elective Admission',
       'Initial_admin_Observation Admission', 'Complication_risk_High',
       'Complication_risk_Low', 'Services_Intravenous', 'Services_CT Scan',
       'Services_MRI'],
      dtype='object')
```

# Removing unnecessary columns

```
In [100…  data = data.drop(['CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State',
                 'County', 'Zip', 'TimeZone', 'Job'], axis=1)
```

```
In [101…  data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 49 columns):
 #   Column                                 Non-Null Count  Dtype  
---  ------                                 --------------  -----  
 0   Lat                                    10000 non-null  float64
 1   Lng                                    10000 non-null  float64
 2   Population                             10000 non-null  int64  
 3   Children                               10000 non-null  int64  
 4   Age                                    10000 non-null  int64  
 5   Income                                 10000 non-null  float64
 6   ReAdmis                                10000 non-null  int8   
 7   VitD_levels                            10000 non-null  float64
 8   Doc_visits                             10000 non-null  int64  
 9   Full_meals_eaten                       10000 non-null  int64  
 10  vitD_supp                              10000 non-null  int64  
 11  Soft_drink                             10000 non-null  int8   
 12  HighBlood                              10000 non-null  int8   
 13  Stroke                                 10000 non-null  int8   
 14  Overweight                             10000 non-null  int8   
 15  Arthritis                              10000 non-null  int8   
 16  Diabetes                               10000 non-null  int8   
 17  Hyperlipidemia                         10000 non-null  int8   
 18  BackPain                               10000 non-null  int8   
 19  Anxiety                                10000 non-null  int8   
 20  Allergic_rhinitis                      10000 non-null  int8   
 21  Reflux_esophagitis                     10000 non-null  int8   
 22  Asthma                                 10000 non-null  int8   
 23  Initial_days                           10000 non-null  float64
 24  TotalCharge                            10000 non-null  float64
 25  Additional_charges                     10000 non-null  float64
 26  Item1                                  10000 non-null  int64  
 27  Item2                                  10000 non-null  int64  
 28  Item3                                  10000 non-null  int64  
 29  Item4                                  10000 non-null  int64  
 30  Item5                                  10000 non-null  int64  
 31  Item6                                  10000 non-null  int64  
 32  Item7                                  10000 non-null  int64  
 33  Item8                                  10000 non-null  int64  
 34  Area_Suburban                          10000 non-null  uint8  
 35  Area_Rural                             10000 non-null  uint8  
 36  Marital_Married                        10000 non-null  uint8  
 37  Marital_Separated                      10000 non-null  uint8  
 38  Marital_Never Married                  10000 non-null  uint8  
 39  Marital_Divorced                       10000 non-null  uint8  
 40  Gender_Female                          10000 non-null  uint8  
 41  Gender_Nonbinary                       10000 non-null  uint8  
 42  Initial_admin_Elective Admission       10000 non-null  uint8  
 43  Initial_admin_Observation Admission    10000 non-null  uint8  
 44  Complication_risk_High                 10000 non-null  uint8  
 45  Complication_risk_Low                  10000 non-null  uint8  
 46  Services_Intravenous                   10000 non-null  uint8  
 47  Services_CT Scan                       10000 non-null  uint8  
 48  Services_MRI                           10000 non-null  uint8  
```

```
dtypes: float64(7), int64(14), int8(13), uint8(15)
memory usage: 1.9 MB
```

In [102…

```
data.columns
```

Out[102…

```
Index(['Lat', 'Lng', 'Population', 'Children', 'Age', 'Income', 'ReAdmis',
       'VitD_levels', 'Doc_visits', 'Full_meals_eaten', 'vitD_supp',
       'Soft_drink', 'HighBlood', 'Stroke', 'Overweight', 'Arthritis',
       'Diabetes', 'Hyperlipidemia', 'BackPain', 'Anxiety',
       'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma', 'Initial_days',
       'TotalCharge', 'Additional_charges', 'Item1', 'Item2', 'Item3', 'Item4',
       'Item5', 'Item6', 'Item7', 'Item8', 'Area_Suburban', 'Area_Rural',
       'Marital_Married', 'Marital_Separated', 'Marital_Never Married',
       'Marital_Divorced', 'Gender_Female', 'Gender_Nonbinary',
       'Initial_admin_Elective Admission',
       'Initial_admin_Observation Admission', 'Complication_risk_High',
       'Complication_risk_Low', 'Services_Intravenous', 'Services_CT Scan',
       'Services_MRI'],
      dtype='object')
```

# Identifying and Removing Outliers

In [103…

```
# Removing Outliers
for col in data[['Lat', 'Lng', 'Population', 'Children', 'Age', 'Income', 'ReAdmis',
        'VitD_levels', 'Doc_visits', 'Full_meals_eaten', 'vitD_supp',
        'Soft_drink', 'HighBlood', 'Stroke', 'Overweight', 'Arthritis',
        'Diabetes', 'Hyperlipidemia', 'BackPain', 'Anxiety',
        'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma', 'Initial_days',
        'TotalCharge', 'Additional_charges', 'Item1', 'Item2', 'Item3', 'Item4',
        'Item5', 'Item6', 'Item7', 'Item8']]:
    col_Z= col +'_Z'
    data[col_Z]= stats.zscore(data[col], axis = 0)
```

In [104…

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 83 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   Lat                                  10000 non-null  float64
 1   Lng                                  10000 non-null  float64
 2   Population                           10000 non-null  int64
 3   Children                             10000 non-null  int64
 4   Age                                  10000 non-null  int64
 5   Income                               10000 non-null  float64
 6   ReAdmis                              10000 non-null  int8
 7   VitD_levels                          10000 non-null  float64
 8   Doc_visits                           10000 non-null  int64
 9   Full_meals_eaten                     10000 non-null  int64
 10  vitD_supp                            10000 non-null  int64
 11  Soft_drink                           10000 non-null  int8
 12  HighBlood                            10000 non-null  int8
 13  Stroke                               10000 non-null  int8
 14  Overweight                           10000 non-null  int8
 15  Arthritis                            10000 non-null  int8
 16  Diabetes                             10000 non-null  int8
 17  Hyperlipidemia                       10000 non-null  int8
 18  BackPain                             10000 non-null  int8
 19  Anxiety                              10000 non-null  int8   
 20  Allergic_rhinitis                    10000 non-null  int8   
 21  Reflux_esophagitis                   10000 non-null  int8   
 22  Asthma                               10000 non-null  int8   
 23  Initial_days                         10000 non-null  float64
 24  TotalCharge                          10000 non-null  float64
 25  Additional_charges                   10000 non-null  float64
 26  Item1                                10000 non-null  int64  
 27  Item2                                10000 non-null  int64  
 28  Item3                                10000 non-null  int64  
 29  Item4                                10000 non-null  int64  
 30  Item5                                10000 non-null  int64  
 31  Item6                                10000 non-null  int64  
 32  Item7                                10000 non-null  int64  
 33  Item8                                10000 non-null  int64  
 34  Area_Suburban                        10000 non-null  uint8  
 35  Area_Rural                           10000 non-null  uint8  
 36  Marital_Married                      10000 non-null  uint8  
 37  Marital_Separated                    10000 non-null  uint8  
 38  Marital_Never Married                10000 non-null  uint8  
 39  Marital_Divorced                     10000 non-null  uint8  
 40  Gender_Female                        10000 non-null  uint8  
 41  Gender_Nonbinary                     10000 non-null  uint8  
 42  Initial_admin_Elective Admission     10000 non-null  uint8  
 43  Initial_admin_Observation Admission  10000 non-null  uint8  
 44  Complication_risk_High               10000 non-null  uint8  
 45  Complication_risk_Low                10000 non-null  uint8  
 46  Services_Intravenous                 10000 non-null  uint8  
 47  Services_CT Scan                     10000 non-null  uint8  
 48  Services_MRI                         10000 non-null  uint8  
 49  Lat_Z                                10000 non-null  float64
 50  Lng_Z                                10000 non-null  float64
 51  Population_Z                         10000 non-null  float64
 52  Children_Z                           10000 non-null  float64
 53  Age_Z                                10000 non-null  float64
 54  Income_Z                             10000 non-null  float64
 55  ReAdmis_Z                            10000 non-null  float64
 56  VitD_levels_Z                        10000 non-null  float64
 57  Doc_visits_Z                         10000 non-null  float64
 58  Full_meals_eaten_Z                   10000 non-null  float64
 59  vitD_supp_Z                          10000 non-null  float64
 60  Soft_drink_Z                         10000 non-null  float64
 61  HighBlood_Z                          10000 non-null  float64
 62  Stroke_Z                             10000 non-null  float64
 63  Overweight_Z                         10000 non-null  float64
 64  Arthritis_Z                          10000 non-null  float64
 65  Diabetes_Z                           10000 non-null  float64
 66  Hyperlipidemia_Z                     10000 non-null  float64
 67  BackPain_Z                           10000 non-null  float64
 68  Anxiety_Z                            10000 non-null  float64
 69  Allergic_rhinitis_Z                  10000 non-null  float64
 70  Reflux_esophagitis_Z                 10000 non-null  float64
 71  Asthma_Z                             10000 non-null  float64
 72  Initial_days_Z                       10000 non-null  float64
 73  TotalCharge_Z                        10000 non-null  float64
 74  Additional_charges_Z                 10000 non-null  float64
 75  Item1_Z                              10000 non-null  float64
 76  Item2_Z                              10000 non-null  float64
 77  Item3_Z                              10000 non-null  float64
 78  Item4_Z                              10000 non-null  float64
 79  Item5_Z                              10000 non-null  float64
 80  Item6_Z                              10000 non-null  float64
 81  Item7_Z                              10000 non-null  float64
 82  Item8_Z                              10000 non-null  float64
```

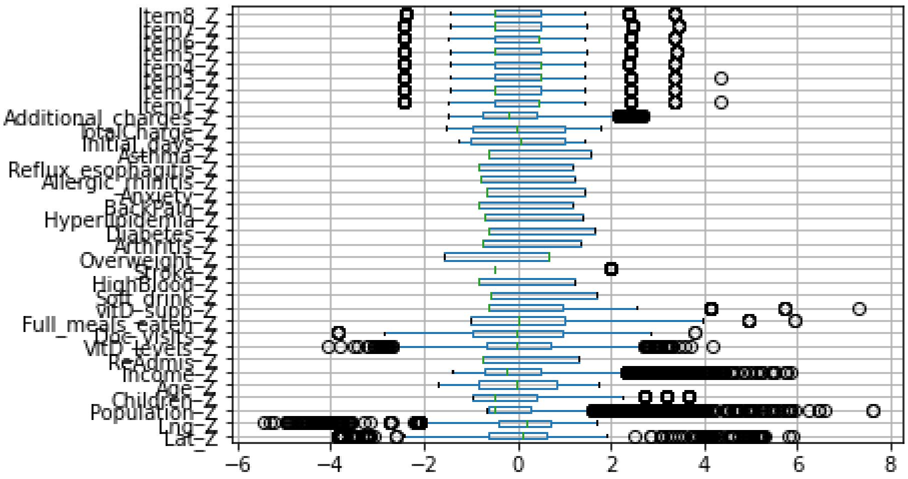
```
dtypes: float64(41), int64(14), int8(13), uint8(15)
memory usage: 4.5 MB
```

In [105…

```
data.iloc[:, 49:83].boxplot(vert=False)
```

Out[105…

```
<AxesSubplot:>
```



In [106…

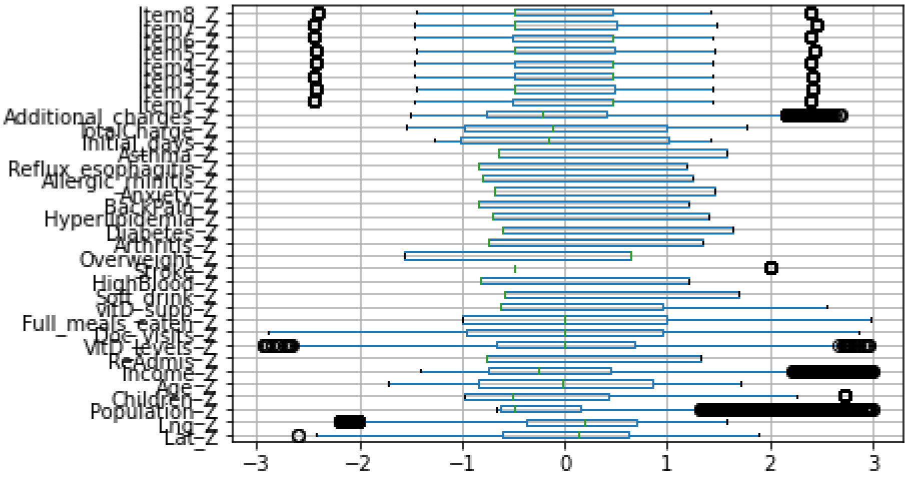
```
#Trimming outliers

for col in data.iloc[:, 49:83]:
        data = data.loc[(data[col] <= 3) & (data[col] >= -3)]
```

In [107…

```
data.iloc[:, 49:83].boxplot(vert=False)
```

Out[107…

```
<AxesSubplot:>
```



In [108…

```
data.columns
```

Out[108…

```
Index(['Lat', 'Lng', 'Population', 'Children', 'Age', 'Income', 'ReAdmis',
       'VitD_levels', 'Doc_visits', 'Full_meals_eaten', 'vitD_supp',
       'Soft_drink', 'HighBlood', 'Stroke', 'Overweight', 'Arthritis',
       'Diabetes', 'Hyperlipidemia', 'BackPain', 'Anxiety',
       'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma', 'Initial_days',
       'TotalCharge', 'Additional_charges', 'Item1', 'Item2', 'Item3', 'Item4',
       'Item5', 'Item6', 'Item7', 'Item8', 'Area_Suburban', 'Area_Rural',
       'Marital_Married', 'Marital_Separated', 'Marital_Never Married',
       'Marital_Divorced', 'Gender_Female', 'Gender_Nonbinary',
       'Initial_admin_Elective Admission',
       'Initial_admin_Observation Admission', 'Complication_risk_High',
       'Complication_risk_Low', 'Services_Intravenous', 'Services_CT Scan',
       'Services_MRI', 'Lat_Z', 'Lng_Z', 'Population_Z', 'Children_Z', 'Age_Z',
       'Income_Z', 'ReAdmis_Z', 'VitD_levels_Z', 'Doc_visits_Z',
       'Full_meals_eaten_Z', 'vitD_supp_Z', 'Soft_drink_Z', 'HighBlood_Z',
       'Stroke_Z', 'Overweight_Z', 'Arthritis_Z', 'Diabetes_Z',
       'Hyperlipidemia_Z', 'BackPain_Z', 'Anxiety_Z', 'Allergic_rhinitis_Z',
       'Reflux_esophagitis_Z', 'Asthma_Z', 'Initial_days_Z', 'TotalCharge_Z',
       'Additional_charges_Z', 'Item1_Z', 'Item2_Z', 'Item3_Z', 'Item4_Z',
       'Item5_Z', 'Item6_Z', 'Item7_Z', 'Item8_Z'],
      dtype='object')
```

In [109…

```
data = data.drop(['Lat_Z', 'Lng_Z', 'Population_Z', 'Children_Z', 'Age_Z',
       'Income_Z', 'ReAdmis_Z', 'VitD_levels_Z', 'Doc_visits_Z',
       'Full_meals_eaten_Z', 'vitD_supp_Z', 'Soft_drink_Z', 'HighBlood_Z',
       'Stroke_Z', 'Overweight_Z', 'Arthritis_Z', 'Diabetes_Z',
       'Hyperlipidemia_Z', 'BackPain_Z', 'Anxiety_Z', 'Allergic_rhinitis_Z',
       'Reflux_esophagitis_Z', 'Asthma_Z', 'Initial_days_Z', 'TotalCharge_Z',
       'Additional_charges_Z', 'Item1_Z', 'Item2_Z', 'Item3_Z', 'Item4_Z',
       'Item5_Z', 'Item6_Z', 'Item7_Z', 'Item8_Z'], axis=1)
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9120 entries, 0 to 9999
Data columns (total 49 columns):
 #   Column                               Non-Null Count  Dtype  
---  ------                               --------------  -----  
 0   Lat                                  9120 non-null   float64
 1   Lng                                  9120 non-null   float64
 2   Population                           9120 non-null   int64  
 3   Children                             9120 non-null   int64  
 4   Age                                  9120 non-null   int64  
 5   Income                               9120 non-null   float64
 6   ReAdmis                              9120 non-null   int8   
 7   VitD_levels                          9120 non-null   float64
 8   Doc_visits                           9120 non-null   int64  
 9   Full_meals_eaten                     9120 non-null   int64  
 10  vitD_supp                            9120 non-null   int64  
 11  Soft_drink                           9120 non-null   int8   
 12  HighBlood                            9120 non-null   int8   
 13  Stroke                               9120 non-null   int8   
 14  Overweight                           9120 non-null   int8   
 15  Arthritis                            9120 non-null   int8   
 16  Diabetes                             9120 non-null   int8   
 17  Hyperlipidemia                       9120 non-null   int8   
 18  BackPain                             9120 non-null   int8   
 19  Anxiety                              9120 non-null   int8   
 20  Allergic_rhinitis                    9120 non-null   int8   
 21  Reflux_esophagitis                   9120 non-null   int8   
 22  Asthma                               9120 non-null   int8   
 23  Initial_days                         9120 non-null   float64
 24  TotalCharge                          9120 non-null   float64
 25  Additional_charges                   9120 non-null   float64
 26  Item1                                9120 non-null   int64  
 27  Item2                                9120 non-null   int64  
 28  Item3                                9120 non-null   int64  
 29  Item4                                9120 non-null   int64  
 30  Item5                                9120 non-null   int64  
 31  Item6                                9120 non-null   int64  
 32  Item7                                9120 non-null   int64  
 33  Item8                               9120 non-null   int64
 34  Area_Suburban                       9120 non-null   uint8
 35  Area_Rural                          9120 non-null   uint8
 36  Marital_Married                     9120 non-null   uint8
 37  Marital_Separated                   9120 non-null   uint8
 38  Marital_Never Married               9120 non-null   uint8
 39  Marital_Divorced                    9120 non-null   uint8
 40  Gender_Female                       9120 non-null   uint8
 41  Gender_Nonbinary                    9120 non-null   uint8
 42  Initial_admin_Elective Admission    9120 non-null   uint8
 43  Initial_admin_Observation Admission 9120 non-null   uint8
 44  Complication_risk_High              9120 non-null   uint8
 45  Complication_risk_Low               9120 non-null   uint8
 46  Services_Intravenous                9120 non-null   uint8
 47  Services_CT Scan                    9120 non-null   uint8
 48  Services_MRI                        9120 non-null   uint8
dtypes: float64(7), int64(14), int8(13), uint8(15)
memory usage: 1.8 MB
```

In [110…

```
#Exporting Clean Data Set
```

In [111…

```
### 5.  Provide a copy of the prepared data set.
data.to_csv("Cleaned_Medical_Dataset.csv")
```