

# Eric Yarger, D208 Task 2

## Design research question, Select Variables from model

What variables from the data set are the most effective at predicting if a patient will be Readmitted?

In [1]:

```
#Import Libraries
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
import missingno as msno
from scipy import stats
from scipy.stats import zscore
```

In [2]:

```
# Read in medical_clean datafile
df = pd.read_csv('C:/Users/ericy/Desktop/medical_clean.csv')
```

## Environment Details

In [3]:

```
# Jupyter environment version
!jupyter --version
```

```
jupyter core      : 4.6.3
jupyter-notebook  : 6.0.3
qtconsole         : 4.7.2
ipython           : 7.13.0
ipykernel         : 5.1.4
jupyter client    : 6.1.2
jupyter lab       : 1.2.6
nbconvert         : 5.6.1
ipywidgets        : 7.5.1
nbformat          : 5.0.4
traitlets         : 4.3.3
```

In [4]:

```
# Python Environment version
import platform
print(platform.python_version())
```

3.7.7

## Cleaning, Preparation, Manipulation

In [5]:

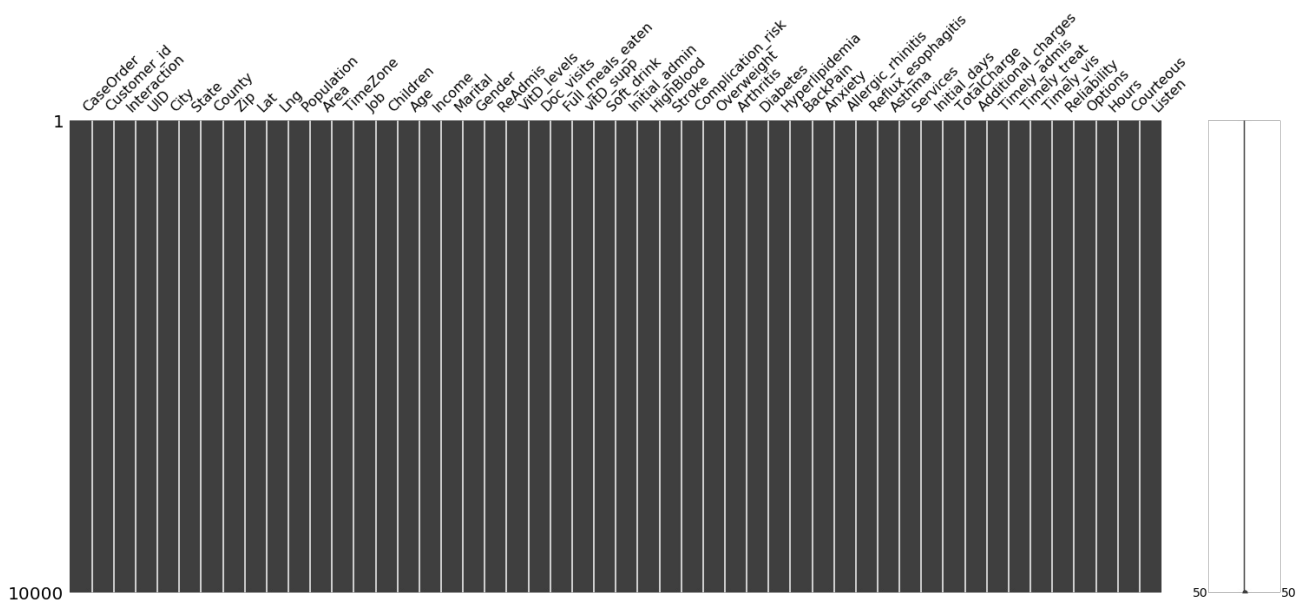
```
#Rename columns for dataset cohesiveness and readability
df.rename(columns={'Item1':'Timely_admis','Item2':'Timely_treat','Item3':'Timely_vis','Item4':'Reliability','Item5':'Options','Item6':'Hours','Item7':'Courteous','Item8':'Listen'},inplace=True)
```

In [6]:

```
# Check for null values
msno.matrix(df)
```

Out[6]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe44118c88>



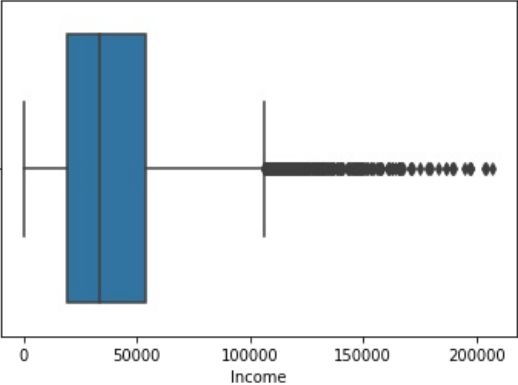
In [7]:

```
# Overview of data set - type, null count, variable names
df.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  Timely_admis            10000 non-null  int64
43  Timely_treat            10000 non-null  int64
44  Timely_vis              10000 non-null  int64
45  Reliability              10000 non-null  int64
46  Options                 10000 non-null  int64
47  Hours                   10000 non-null  int64
48  Courteous               10000 non-null  int64
49  Listen                  10000 non-null  int64
dtypes: float64(7), int64(16), object(27)
memory usage: 3.8+ MB
```

```
In [8]:
sns.boxplot(df['Income'])
```

```
Out[8]:
<matplotlib.axes._subplots.AxesSubplot at 0x1fe448f8688>
```

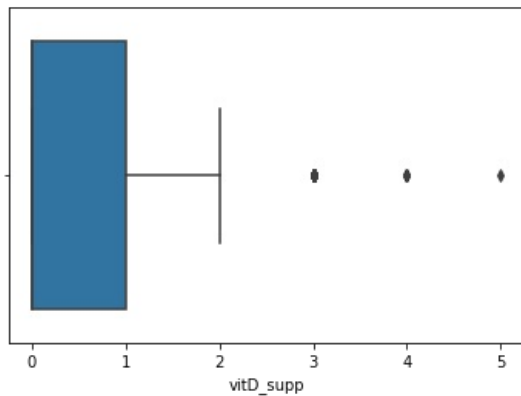


In [9]:

```
sns.boxplot(df['vitD_supp'])
```

Out[9]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe4351f348>

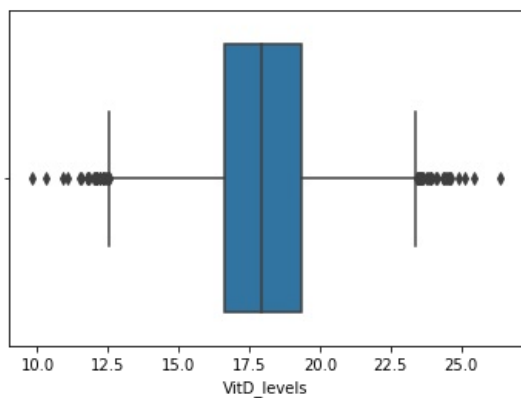


In [10]:

```
sns.boxplot(df['VitD_levels'])
```

Out[10]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe4358ca48>

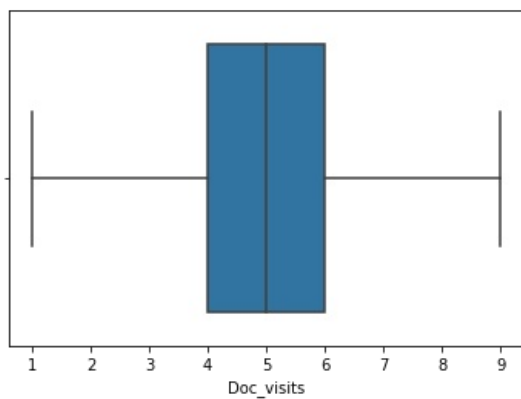


In [11]:

```
sns.boxplot(df['Doc_visits'])
```

Out[11]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe435efd08>

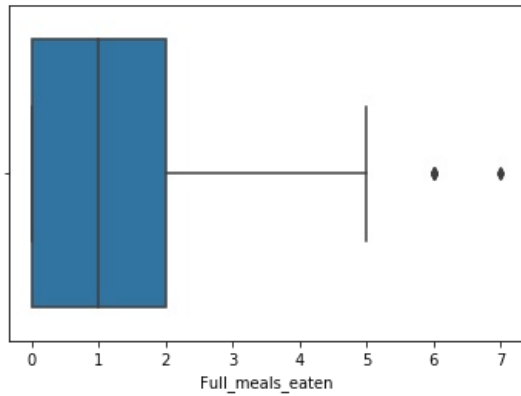


In [12]:

```
sns.boxplot(df['Full_meals_eaten'])
```

Out[12]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe43653608>

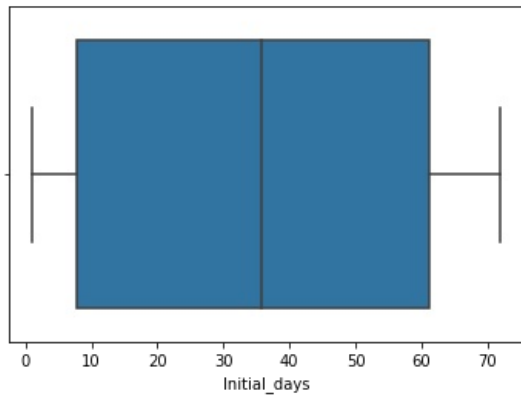


In [13]:

```
sns.boxplot(df['Initial_days'])
```

Out[13]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe436c3408>

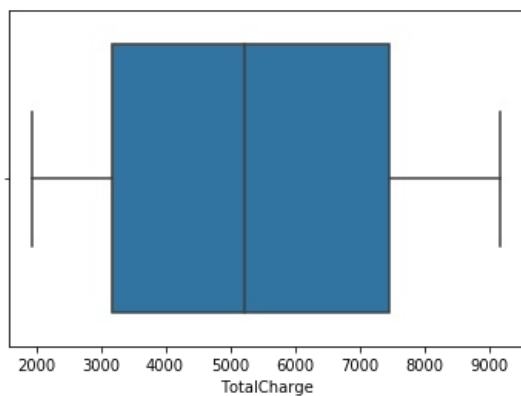


In [14]:

```
sns.boxplot(df['TotalCharge'])
```

Out[14]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe43737248>

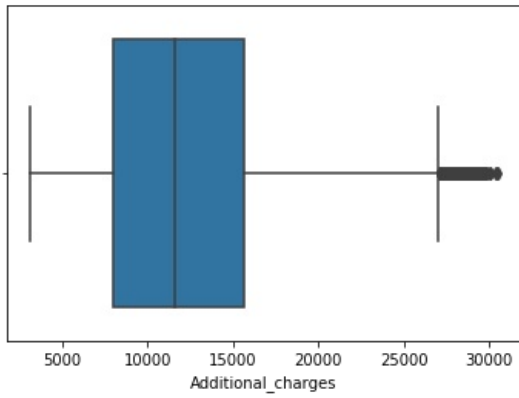


In [15]:

```
sns.boxplot(df['Additional_charges'])
```

Out[15]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe4377ab48>

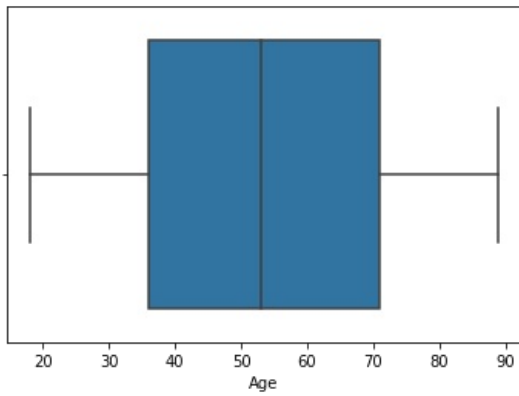


In [16]:

```
sns.boxplot(df['Age'])
```

Out[16]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe437ff308>

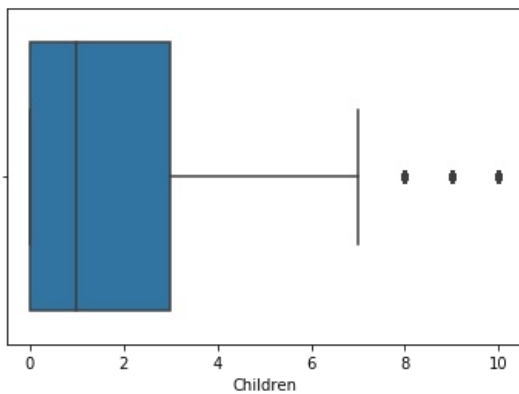


In [17]:

```
sns.boxplot(df['Children'])
```

Out[17]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe43860088>

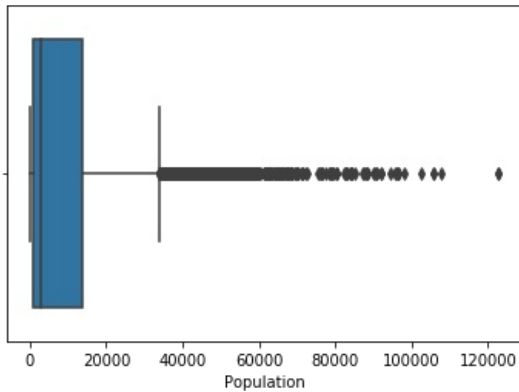


In [18]:

```
sns.boxplot(df['Population'])
```

Out[18]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe438bf608>



In [19]:

```
# Outlier removal method via Z-score, Code reference (Bushmanov, 2019)
num_data = df.select_dtypes(include=['number'])
cat_data = df.select_dtypes(exclude=['number'])
```

In [20]:

```
idx = np.all(stats.zscore(num_data) <3, axis=1)
```

In [21]:

```
df = pd.concat([num_data.loc[idx], cat_data.loc[idx]], axis=1)
```

In [22]:

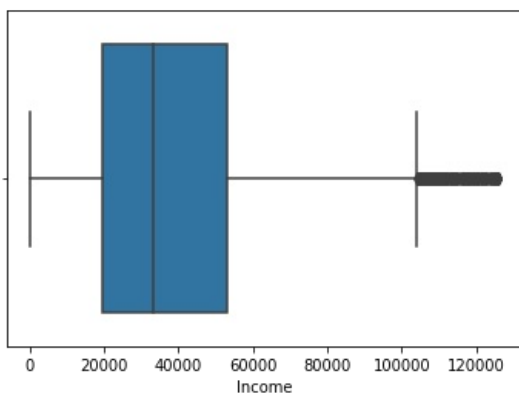
```
# Target variable, change responses to numerical binary
df['ReAdmis'].replace(('Yes','No'), (1,0), inplace=True)
```

In [23]:

```
sns.boxplot(df['Income'])
```

Out[23]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe448ee808>

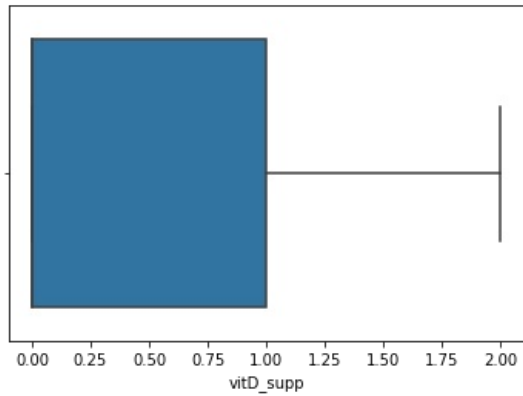


In [24]:

```
sns.boxplot(df['vitD_supp'])
```

Out[24]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe42fb2c48>

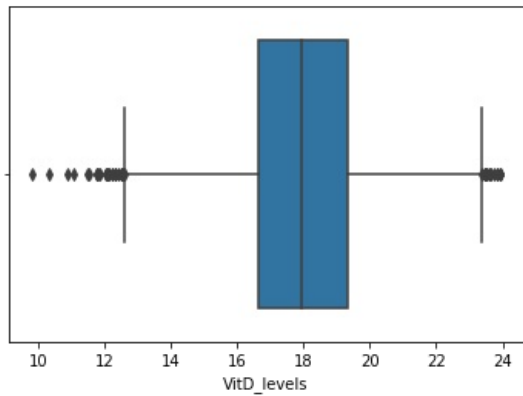


In [25]:

```
sns.boxplot(df['VitD_levels'])
```

Out[25]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe4302a9c8>

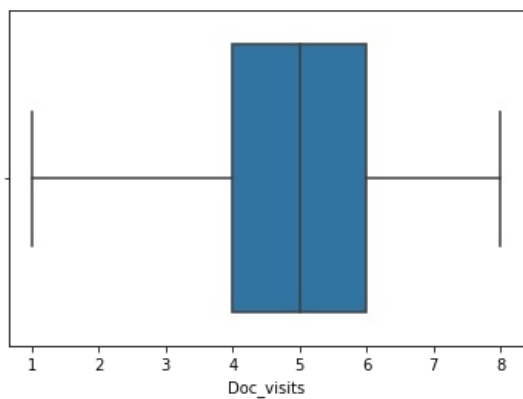


In [26]:

```
sns.boxplot(df['Doc_visits'])
```

Out[26]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe4308a708>



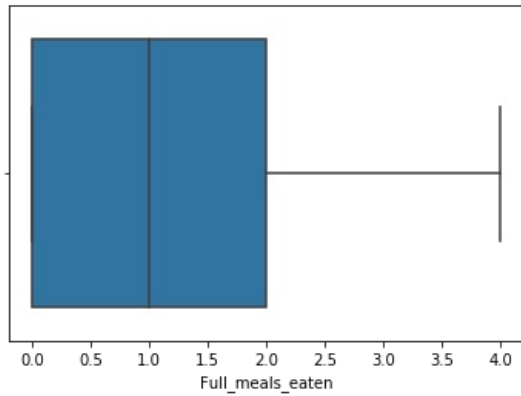


In [27]:

```
sns.boxplot(df['Full_meals_eaten'])
```

Out[27]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe430f5448>

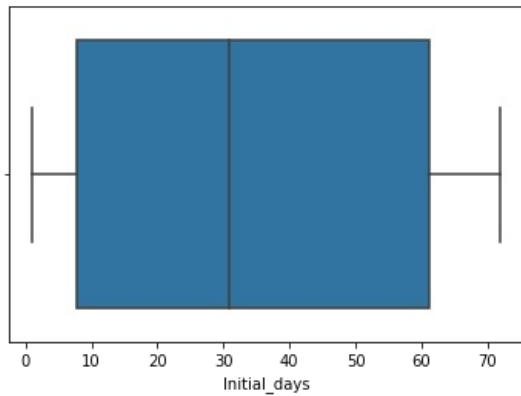


In [28]:

```
sns.boxplot(df['Initial_days'])
```

Out[28]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe43105708>

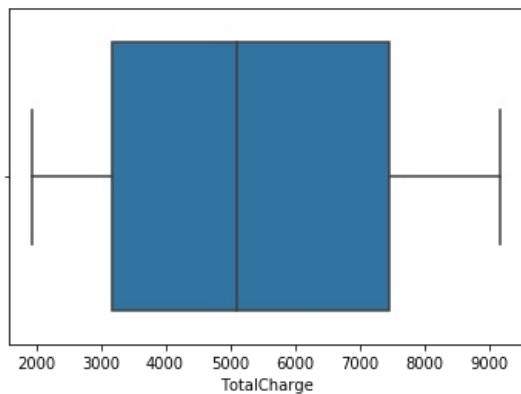


In [29]:

```
sns.boxplot(df['TotalCharge'])
```

Out[29]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe45fedd88>

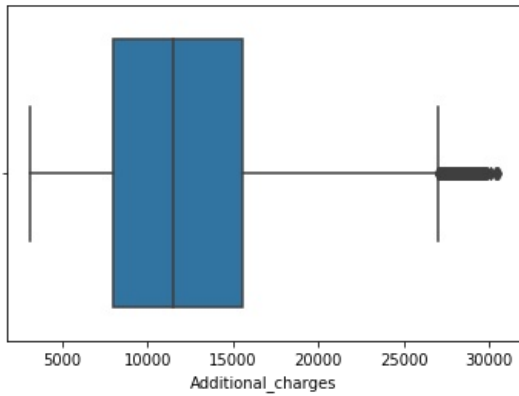


In [30]:

```
sns.boxplot(df['Additional_charges'])
```

Out[30]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe46056208>

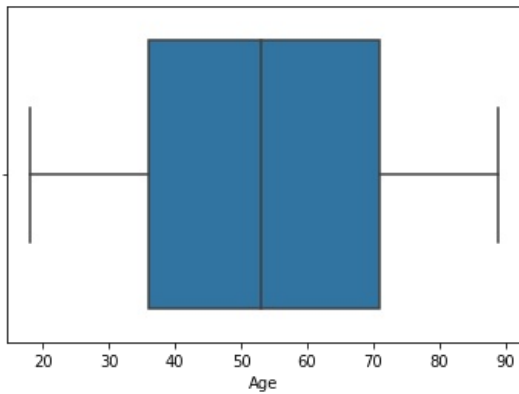


In [31]:

```
sns.boxplot(df['Age'])
```

Out[31]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe460b5ec8>

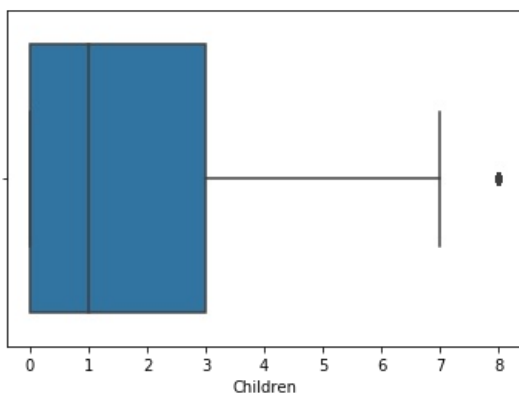


In [32]:

```
sns.boxplot(df['Children'])
```

Out[32]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe46124108>

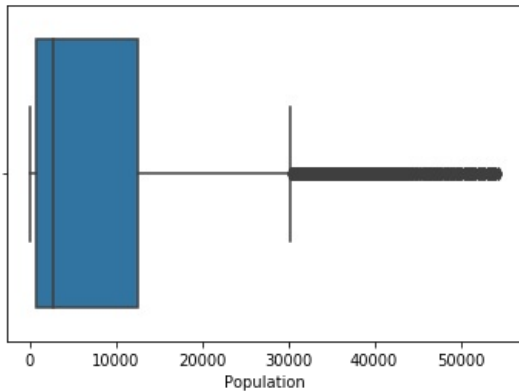


In [33]:

```
sns.boxplot(df['Population'])
```

Out[33]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x1fe46189288>



## Univariate Visualization

### Histograms

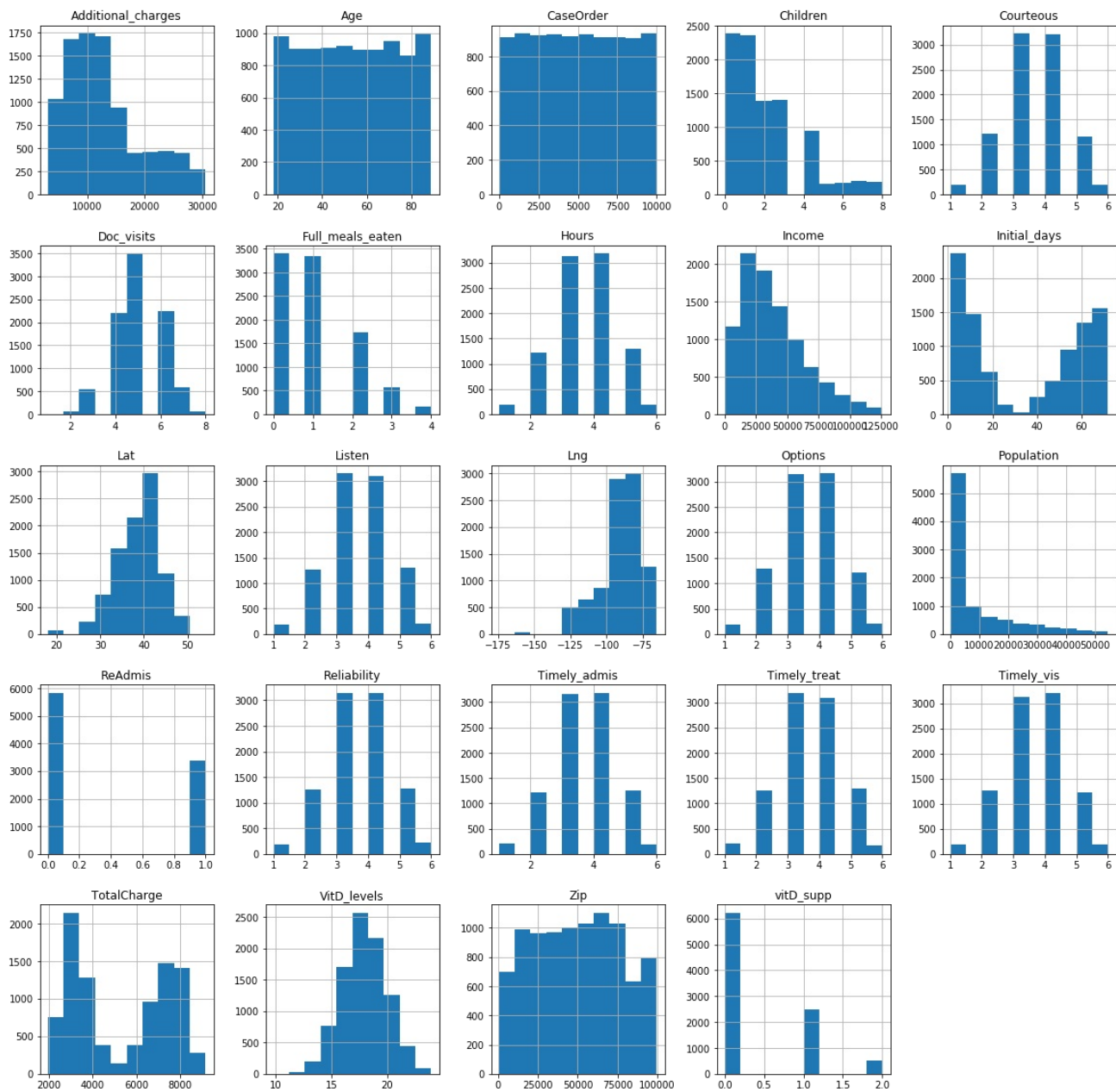
#### Identify Feature Distribution and Normality

In [34]:

```
df.hist(figsize=(20,20))
```

Out[34]:

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001FE4623C888>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46267B48>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE462A33C8>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE462DC448>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46314588>],
 [<matplotlib.axes._subplots.AxesSubplot object at 0x000001FE4634C608>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46387708>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE4650D848>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46519448>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46553608>],
 [<matplotlib.axes._subplots.AxesSubplot object at 0x000001FE465B9AC8>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE465EFBC8>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE4662ACC8>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46661E08>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE4669BF08>],
 [<matplotlib.axes._subplots.AxesSubplot object at 0x000001FE466D3FC8>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46712148>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46749248>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46782348>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE467BA488>],
 [<matplotlib.axes._subplots.AxesSubplot object at 0x000001FE467F5588>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE4682C608>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE46865748>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE4689E888>,
   <matplotlib.axes._subplots.AxesSubplot object at 0x000001FE468D9988>]],
      dtype=object)
```



## Bivariate Visualization

### Scatterplots with

**X-Axis = ReAdmis**

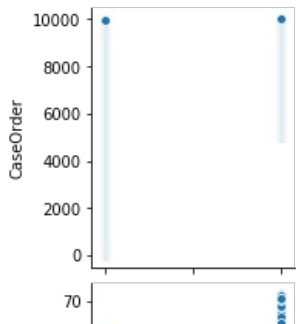
**Y-Axis = Independent feature**

In [35]:

```
sns.pairplot(df, x_vars=['ReAdmis'], y_vars=['CaseOrder','Initial_days','VitD_levels','Doc_visits','vitD_supp','Soft_drink','Initial_admin','HighBlood','Stroke','Complication_risk','Overweight','Arthritis','Diabetes','Hyperlipidemia','BackPain','Anxiety','Allergic_rhinitis','Reflux_esophagitis','Asthma','Services','TotalCharge','Additional_charges'])
```

Out[35]:

<seaborn.axisgrid.PairGrid at 0x1fe47635fc8>



Initial\_admin

Observation Admission

Elective Admission

Emergency Admission

HighBlood

Stroke

Soft\_drink

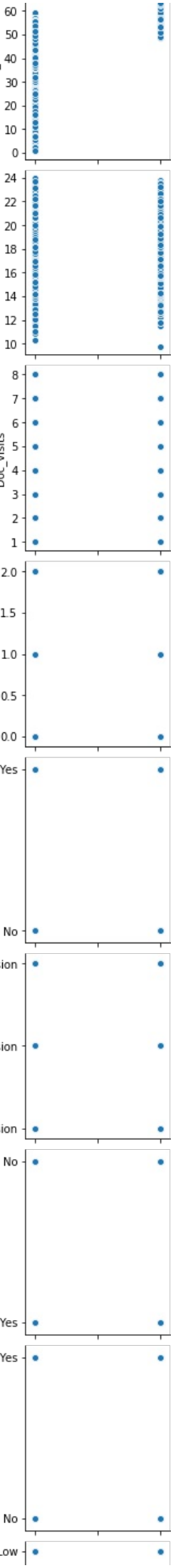
vitD\_supp

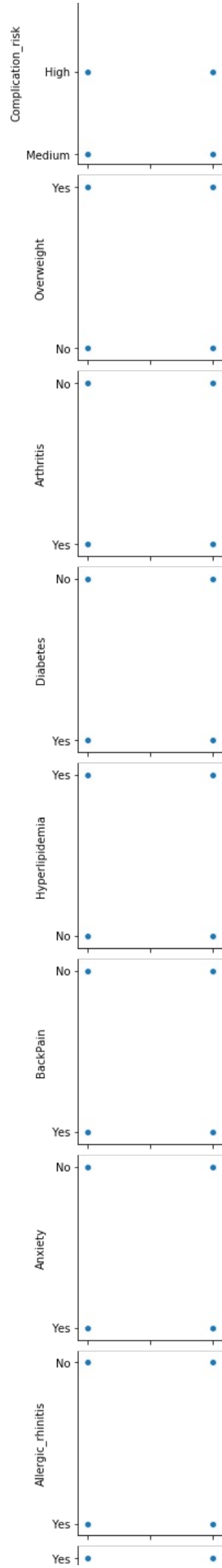
Doc\_visits

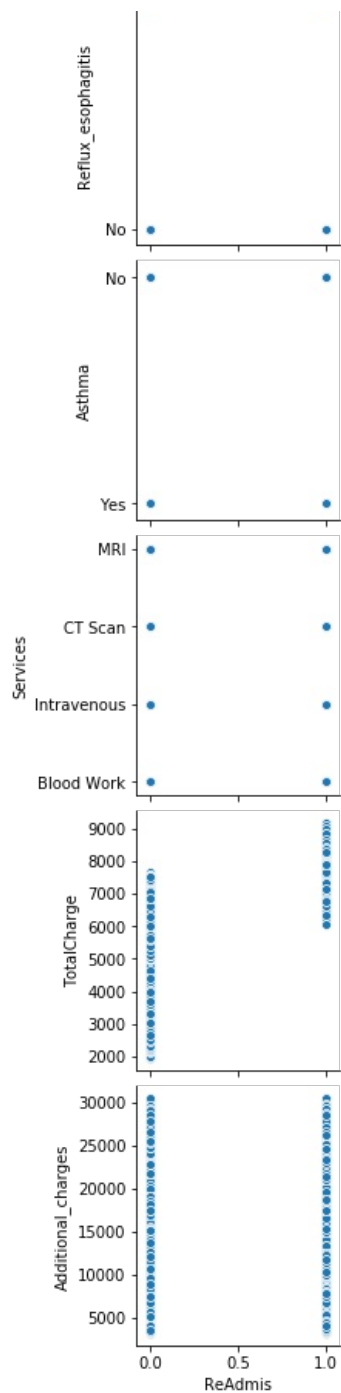
VitD\_levels

Initial\_days

Initial\_days







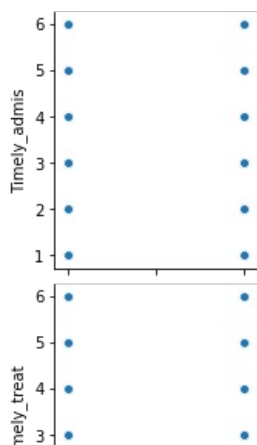
In [ ]:

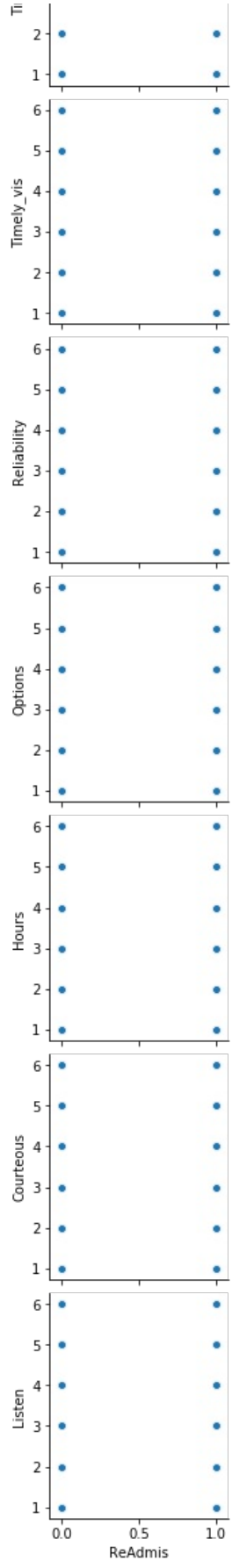
In [36]:

```
sns.pairplot(df, x_vars=['ReAdmis'], y_vars=['Timely_admis', 'Timely_treat', 'Timely_vis', 'Reliability', 'Options', 'Hours', 'Courteous', 'Listen'])
```

Out[36]:

<seaborn.axisgrid.PairGrid at 0x1fe486c1cc8>







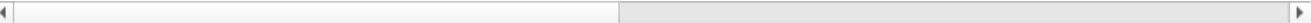
In [37]:

```
df.corr()
```

Out[37]:

	CaseOrder	Zip	Lat	Lng	Population	Children	Age	Income	VitD_levels	Doc_visits	...	A
CaseOrder	1.000000	0.010465	-0.012946	-0.012081	0.001489	0.017027	-0.003011	-0.012265	-0.015026	-0.006920	...	
Zip	0.010465	1.000000	-0.084258	-0.913573	0.012947	0.014307	-0.003327	0.010507	-0.010747	0.000257	...	
Lat	-0.012946	-0.084258	1.000000	0.001062	-0.187334	0.005874	-0.000132	-0.015414	-0.005158	0.004689	...	
Lng	-0.012081	-0.913573	0.001062	1.000000	-0.018263	-0.014141	0.002780	-0.008175	0.000931	0.002417	...	
Population	0.001489	0.012947	-0.187334	-0.018263	1.000000	0.007810	-0.018884	0.002162	0.004719	0.016088	...	
Children	0.017027	0.014307	0.005874	-0.014141	0.007810	1.000000	0.006050	0.003951	0.006542	-0.003467	...	
Age	-0.003011	-0.003327	-0.000132	0.002780	-0.018884	0.006050	1.000000	-0.003218	0.008795	0.010819	...	
Income	-0.012265	0.010507	-0.015414	-0.008175	0.002162	0.003951	-0.003218	1.000000	-0.015684	0.011179	...	
VitD_levels	-0.015026	-0.010747	-0.005158	0.000931	0.004719	0.006542	0.008795	-0.015684	1.000000	0.010297	...	
Doc_visits	-0.006920	0.000257	0.004689	0.002417	0.016088	-0.003467	0.010819	0.011179	0.010297	1.000000	...	
Full_meals_eaten	-0.020805	0.013077	-0.001353	-0.013120	-0.025711	-0.005112	0.008499	-0.012628	0.032606	-0.004586	...	
vitD_supp	0.026011	0.009348	0.005225	-0.001817	0.004134	-0.010125	0.009336	0.001478	-0.015671	0.002755	...	
Initial_days	0.831426	0.011103	-0.009938	-0.006659	0.004435	0.022122	0.009943	-0.006543	-0.007267	-0.008363	...	
TotalCharge	0.821397	0.010493	-0.012843	-0.005866	0.004758	0.022909	0.010785	-0.008523	-0.004403	-0.005363	...	
Additional_charges	-0.003178	0.001545	-0.001433	0.003290	-0.011835	0.014076	0.716409	-0.005190	0.006120	0.014611	...	
Timely_admis	-0.016607	-0.008630	0.008075	0.011933	0.004194	0.004097	0.005614	-0.004194	0.010499	0.003984	...	
Timely_treat	-0.005508	-0.002475	0.009184	-0.002521	0.016837	0.006169	0.004382	-0.012371	0.003697	0.004377	...	
Timely_vis	-0.006320	-0.010277	0.010924	0.002614	-0.004754	-0.002485	0.006990	-0.007394	-0.011930	-0.003794	...	
Reliability	-0.016204	0.001231	-0.011577	0.000283	-0.008892	-0.001091	0.003407	-0.003532	-0.016650	-0.006303	...	
Options	-0.004709	0.006290	0.000179	-0.002771	0.013720	0.003409	-0.013980	-0.005088	0.007878	-0.011124	...	
Hours	-0.006087	-0.001406	0.009542	-0.004637	0.007970	-0.002796	0.003434	0.003083	0.004610	0.009226	...	
Courteous	0.005102	-0.004203	0.009071	0.002070	0.010529	0.015894	0.009339	0.008516	-0.007461	0.005322	...	
Listen	-0.012319	-0.010159	0.004348	0.003871	-0.005522	-0.011509	0.002873	0.020238	-0.024347	0.006145	...	
ReAdmis	0.661462	0.009519	-0.012324	-0.004241	0.007563	0.023890	0.011880	-0.008669	0.002858	-0.002226	...	

24 rows × 24 columns



Dummies & Renaming

In [38]:

```
df.drop('CaseOrder',axis=1, inplace=True)
```

In [39]:

```
#Get dummies code reference (Pandas.get_dummies, N.d.)
df = pd.get_dummies(df, columns=['Area','Marital','Gender','Doc_visits','vitD_supp','Soft_drink','Initial_admin',
'HighBlood','Stroke','Complication_risk','Overweight','Arthritis','Diabetes','Hyperlipidemia','BackPain','Anxiety',
'Allergic_rhinitis','Reflux_esophagitis','Asthma','Services'], drop_first=True)
```

In [40]:

```
df = pd.get_dummies(df, columns=['Timely_admis','Timely_treat','Timely_vis','Reliability','Options','Hours','Courteous','Listen'],drop_first=True)
```

In [41]:

```
df = pd.get_dummies(df, columns=['ReAdmis'],drop_first=True)
```

In [42]:

```
df = pd.get_dummies(df, columns=['Children'],drop_first=True)
```

In [43]:

```
#Rename features with spaces in name for future analysis
df.rename(columns={'Services_CT Scan':'Services_CT_Scan','Marital_Never Married':'Marital_Never_Married','Initial_admin_Emergency Admission':'Initial_admin_Emergency_Admission','Initial_admin_Observation Admission':'Initial_admin_Observation_Admission'},inplace=True)
```

In [44]:

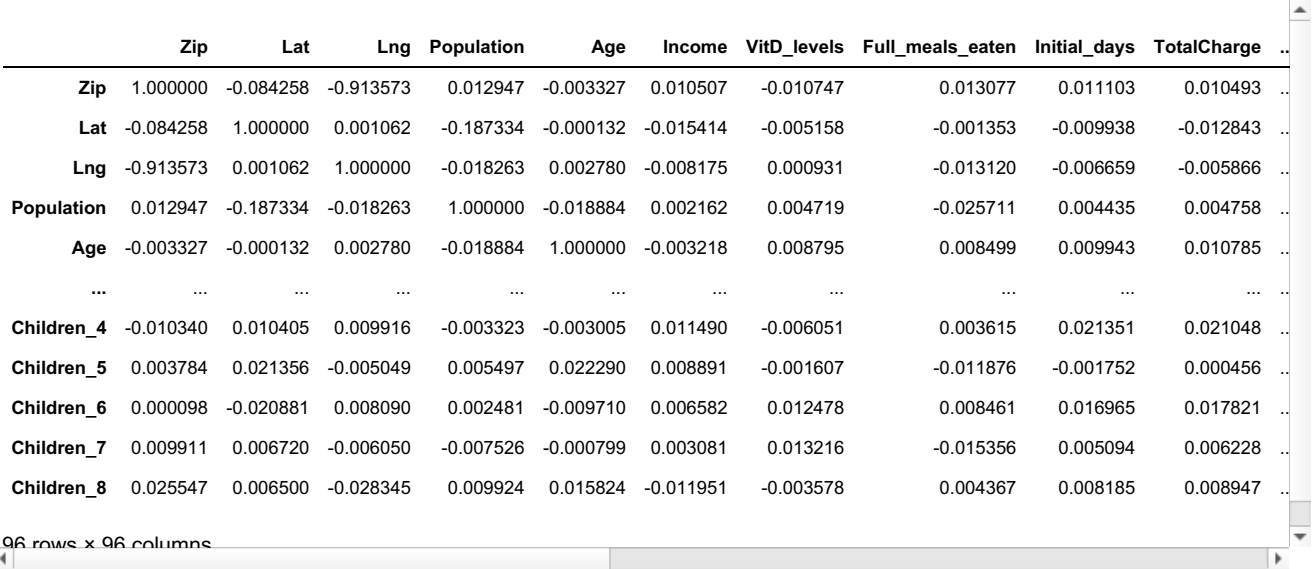
```
df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 9206 entries, 0 to 9999
Columns: 104 entries, Zip to Children_8
dtypes: float64(7), int64(4), object(8), uint8(85)
memory usage: 2.2+ MB
```

In [45]:

```
df.corr()
```

Out[45]:



In [46]:

```
# Heatmap code reference (Seaborn.heatmap, N.d.)
import matplotlib
matplotlib.pyplot.figure(figsize=(20,20))
heatmap = sns.heatmap(df.corr()[['ReAdmis_1']].sort_values(by='ReAdmis_1', ascending=False), vmin=-1, vmax=1, annot=True, cmap='BrBG')
heatmap.set_title('Variables correlating with ReAdmis Heatmap',pad=12)
```

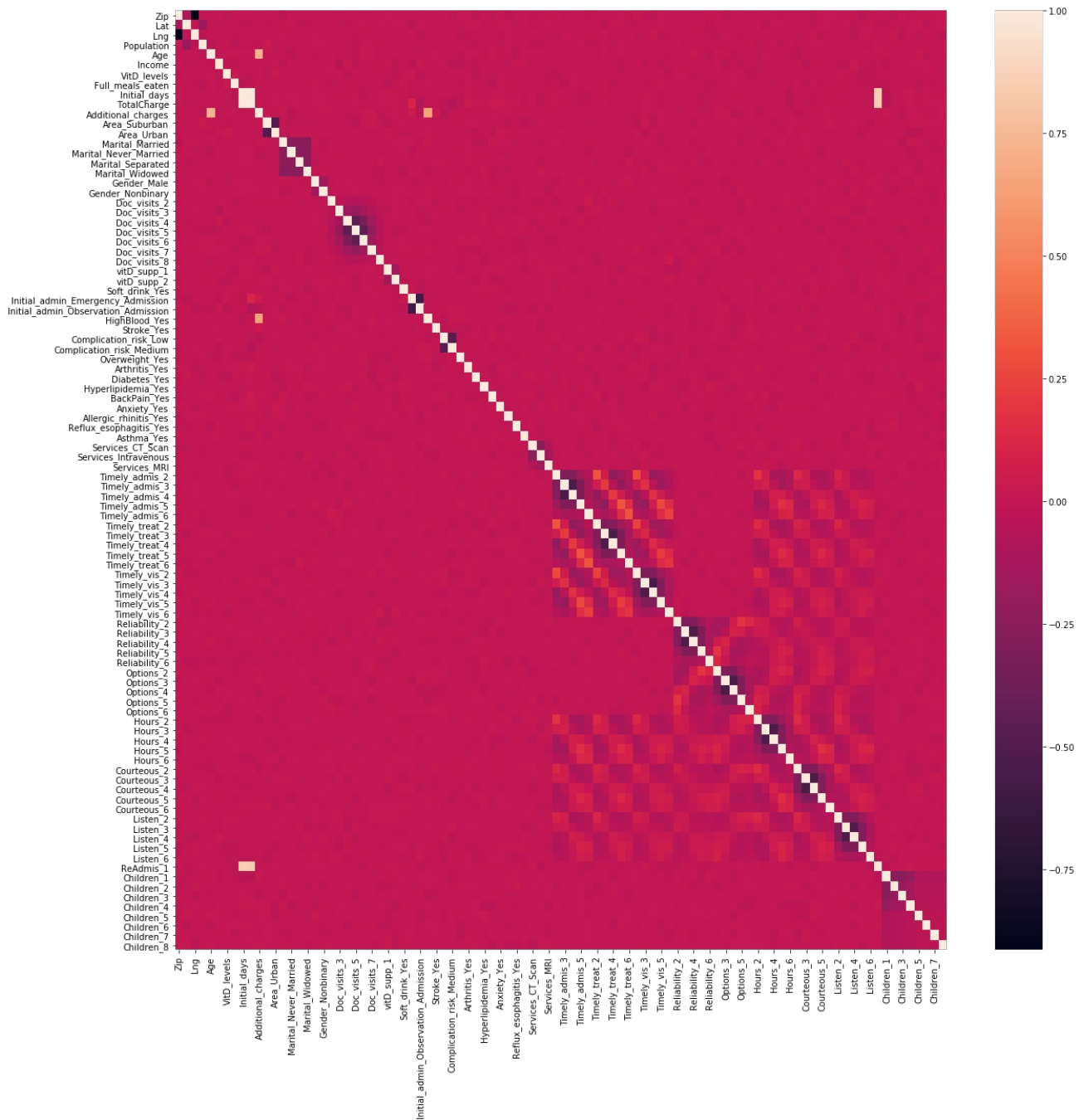
Out[46]:

Text(0.5, 1, 'Variables correlating with ReAdmis Heatmap')



In [47]:

```
fig_dims = (20, 20)
fig, ax = plt.subplots(figsize=fig_dims)
sns.heatmap(df.corr(), ax=ax)
plt.show()
```



```
In [48]:
df.to_excel('C:/Users/eric/Desktop/D208.2.full.xlsx', index=False)
```

## Initial Feature Selection

Target variable is ReAdmis\_1 Correlation > .02 for explanatory variable selection

```
In [49]:
abs(df.corr()["ReAdmis_1"][abs(df.corr()["ReAdmis_1"])>=0.02].drop('ReAdmis_1')).index.tolist()
```

```
Out[49]:
['Initial_days',
 'TotalCharge',
 'Services_CT_Scan',
 'Timely_admis_6',
 'Timely_vis_3',
 'Hours_6',
 'Listen_6',
 'Children_1',
 'Children_4']
```

```
In [50]:
# VIF technique code reference (Zach, 2020)
```

In [51]:

```
from patsy import dmatrices
from statsmodels.stats.outliers_influence import variance_inflation_factor
y, X = dmatrices('ReAdmis_1 ~ Initial_days+TotalCharge+Children_1+Children_4+Services_CT_Scan+Timely_vis_3+Timely_admis_6+Listen_6+Hours_6', data=df, return_type='dataframe')
```

In [52]:

```
vif = pd.DataFrame()
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['variable'] = X.columns
```

In [53]:

```
vif
```

Out[53]:

	VIF	variable
0	57.215523	Intercept
1	40.824852	Initial_days
2	40.825038	TotalCharge
3	1.041991	Children_1
4	1.041378	Children_4
5	1.001029	Services_CT_Scan
6	1.011761	Timely_vis_3
7	1.023459	Timely_admis_6
8	1.004894	Listen_6
9	1.017869	Hours_6

In [54]:

```
y, X = dmatrices('ReAdmis_1 ~ Initial_days+Children_1+Children_4+Services_CT_Scan+Timely_vis_3+Timely_admis_6+Listen_6+Hours_6', data=df, return_type='dataframe')
```

In [55]:

```
vif = pd.DataFrame()
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['variable'] = X.columns
```

In [56]:

```
vif
```

Out[56]:

	VIF	variable
0	4.053441	Intercept
1	1.002430	Initial_days
2	1.041870	Children_1
3	1.041375	Children_4
4	1.000610	Services_CT_Scan
5	1.011730	Timely_vis_3
6	1.023351	Timely_admis_6
7	1.004721	Listen_6
8	1.017866	Hours_6

C2: Summary Statistics for target variable & all predictor variables for inital model

In [57]:

```
dfi = df[['ReAdmis_1',
'Children_1',
'Children_4',
'Initial_days',
'Services_CT_Scan',
'Timely_vis_3',
'Timely_admis_6',
'Listen_6',
'Hours_6']]
```

In [58]:

```
dfi.to_excel('C:/Users/ericy/Desktop/D208.2.Selected.xlsx', index=False)
```

In [59]:

```
dfi['ReAdmis_1'].unique()
```

Out[59]:

```
array([0, 1], dtype=uint8)
```

In [60]:

```
dfi.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9206 entries, 0 to 9999
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   ReAdmis_1             9206 non-null   uint8
1   Children_1            9206 non-null   uint8
2   Children_4            9206 non-null   uint8
3   Initial_days          9206 non-null   float64
4   Services_CT_Scan      9206 non-null   uint8
5   Timely_vis_3          9206 non-null   uint8
6   Timely_admis_6        9206 non-null   uint8
7   Listen_6              9206 non-null   uint8
8   Hours_6               9206 non-null   uint8
dtypes: float64(1), uint8(8)
memory usage: 215.8 KB
```

In [61]:

```
dfi.describe()
```

Out[61]:

	ReAdmis_1	Children_1	Children_4	Initial_days	Services_CT_Scan	Timely_vis_3	Timely_admis_6	Listen_6	Hours_6
count	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000
mean	0.366935	0.256137	0.10189	34.399945	0.122855	0.339561	0.020747	0.022268	0.021399
std	0.481995	0.436522	0.30252	26.325319	0.328288	0.473586	0.142545	0.147562	0.144718
min	0.000000	0.000000	0.00000	1.001981	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.00000	7.881412	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.00000	30.841461	0.000000	0.000000	0.000000	0.000000	0.000000
75%	1.000000	1.000000	0.00000	61.157838	0.000000	1.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.00000	71.981490	1.000000	1.000000	1.000000	1.000000	1.000000

In [62]:

```
dfi.corr()
```

Out[62]:

	ReAdmis_1	Children_1	Children_4	Initial_days	Services_CT_Scan	Timely_vis_3	Timely_admis_6	Listen_6	Hours_6
ReAdmis_1	1.000000	-0.025936	0.022959	0.852064	0.026087	0.021399	-0.023851	-0.020195	-0.02380
Children_1	-0.025936	1.000000	-0.197647	-0.035131	0.004024	0.001216	0.005373	-0.000857	-0.00594
Children_4	0.022959	-0.197647	1.000000	0.021351	0.004116	-0.008726	-0.001161	-0.002160	-0.01754
Initial_days	0.852064	-0.035131	0.021351	1.000000	0.010723	0.013412	-0.013688	-0.012797	-0.02178
Services_CT_Scan	0.026087	0.004024	0.004116	0.010723	1.000000	-0.016801	-0.005723	0.008555	0.00639
Timely_vis_3	0.021399	0.001216	-0.008726	0.013412	-0.016801	1.000000	-0.094715	-0.028930	-0.05055
Timely_admis_6	-0.023851	0.005373	-0.001161	-0.013688	-0.005723	-0.094715	1.000000	0.045175	0.11539
Listen_6	-0.020195	-0.000857	-0.002160	-0.012797	0.008555	-0.028930	0.045175	1.000000	0.04890
Hours_6	-0.023807	-0.005948	-0.017549	-0.021784	0.006397	-0.050554	0.115398	0.048904	1.00000

In [63]:

```
dfi.mean()
```

Out[63]:

```
ReAdmis_1      0.366935
Children_1      0.256137
Children_4      0.101890
Initial_days    34.399945
Services_CT_Scan 0.122855
Timely_vis_3    0.339561
Timely_admis_6  0.020747
Listen_6        0.022268
Hours_6         0.021399
dtype: float64
```

In [64]:

```
dfi.median()
```

Out[64]:

```
ReAdmis_1      0.000000
Children_1      0.000000
Children_4      0.000000
Initial_days    30.841461
Services_CT_Scan 0.000000
Timely_vis_3    0.000000
Timely_admis_6  0.000000
Listen_6        0.000000
Hours_6         0.000000
dtype: float64
```

In [65]:

```
dfi.mode()
```

Out[65]:

	ReAdmis_1	Children_1	Children_4	Initial_days	Services_CT_Scan	Timely_vis_3	Timely_admis_6	Listen_6	Hours_6
0	0.0	0.0	0.0	67.42139	0.0	0.0	0.0	0.0	0.0
1	NaN	NaN	NaN	70.32542	NaN	NaN	NaN	NaN	NaN

In [66]:

```
# C5: Prepared Dataset
dfi.to_csv('C:/Users/ericy/Desktop/D208.2_prepared.csv', index=False)
```

## Initial Logistic Regression Model

In [67]:

```
# D1: Initial Logistic Regression with logit. Code Reference (Cosine1509, 2022).
from statsmodels.formula.api import logit
```

In [68]:

```
re_log = logit('ReAdmis_1 ~ Initial_days + Children_1 + Children_4 + Services_CT_Scan + Timely_vis_3 + Timely_admis_6 + Listen_6 + Hours_6', data=df).fit()
```

Optimization terminated successfully.  
Current function value: 0.048161  
Iterations 13

In [69]:

```
print(re_log.summary())
```

```

                    Logit Regression Results
=====
Dep. Variable:          ReAdmis_1      No. Observations:          9206
Model:                  Logit          Df Residuals:              9197
Method:                 MLE           Df Model:                  8
Date:                  Tue, 05 Jul 2022      Pseudo R-squ.:          0.9267
Time:                  14:09:23             Log-Likelihood:         -443.37
converged:              True             LL-Null:                -6051.1
Covariance Type:        nonrobust          LLR p-value:            0.000
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
Intercept      -54.2497       2.606     -20.815     0.000     -59.358     -49.141
Initial_days       0.9959       0.048     20.902     0.000       0.903       1.089
Children_1      -0.3136       0.207     -1.516     0.129      -0.719       0.092
Children_4       0.1378       0.277       0.497     0.619      -0.405       0.681
Services_CT_Scan  0.9881       0.277     3.561     0.000       0.444       1.532
Timely_vis_3      0.4035       0.185     2.179     0.029       0.040       0.767
Timely_admis_6   -0.1088       0.669     -0.163     0.871      -1.420       1.202
Listen_6        -1.1492       0.556     -2.068     0.039      -2.238      -0.060
Hours_6         -0.4155       0.584     -0.711     0.477      -1.560       0.729
=====
```

Possibly complete quasi-separation: A fraction 0.74 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

## Reduced Feature Selection

In [70]:

```
## K Nearest Neighbors & correlation for feature selection
# Code Reference (Feely, 2020), starting at 12:30 in video - going to minute 16:00
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import KFold
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import cross_val_predict
from sklearn.linear_model import LinearRegression
from math import sqrt
```

## K Nearest Neighbors & correlation for feature selection

### Code Reference (Feely, 2020), starting at 12:30 in video - going to minute 16:00

In [71]:

```
X=df[['Initial_days','Children_1','Children_4','Services_CT_Scan','Timely_vis_3','Timely_admis_6','Listen_6','Hours_6']]
y = df.ReAdmis_1
B=df[['ReAdmis_1','Initial_days','Children_1','Children_4','Services_CT_Scan','Timely_vis_3','Timely_admis_6','Listen_6','Hours_6']]
```

In [72]:

```
cv = KFold(n_splits=10, random_state=0, shuffle=True)
classifier_pipeline = make_pipeline(StandardScaler(), KNeighborsRegressor(n_neighbors=10))
y_pred = cross_val_predict(classifier_pipeline, X, y, cv=cv)
print("RMSE: " + str(round(sqrt(mean_squared_error(y,y_pred)),2)))
print("Pseudo_R_squared: " + str(round(r2_score(y,y_pred),2)))
```

RMSE: 0.15  
Pseudo\_R\_squared: 0.91



In [73]:

```
vals = [0.02,.022,.025,.03,.025,.04,.045,0.05,0.08,0.1,0.2]
for val in vals:
    features = abs(B.corr()["ReAdmis_1"])[abs(df.corr()["ReAdmis_1"]>val).drop('ReAdmis_1').index.tolist()]

    X = B.drop(columns='ReAdmis_1')
    X=X[features]

    print(features)

    y_pred = cross_val_predict(classifier_pipeline, X, y, cv=cv)
    print("RMSE: " + str(round(sqrt(mean_squared_error(y,y_pred)),2)))
```

```
['Initial_days', 'Children_1', 'Children_4', 'Services_CT_Scan', 'Timely_vis_3', 'Timely_admis_6', '
Listen_6', 'Hours_6']
RMSE: 0.15
['Initial_days', 'Children_1', 'Children_4', 'Services_CT_Scan', 'Timely_admis_6', 'Hours_6']
RMSE: 0.14
['Initial_days', 'Children_1', 'Services_CT_Scan']
RMSE: 0.13
['Initial_days']
RMSE: 0.13
['Initial_days', 'Children_1', 'Services_CT_Scan']
RMSE: 0.13
['Initial_days']
RMSE: 0.13
['Initial_days']
RMSE: 0.13
['Initial_days']
RMSE: 0.13
['Initial_days']
RMSE: 0.13
['Initial_days']
RMSE: 0.13
['Initial_days']
RMSE: 0.13
['Initial_days']
RMSE: 0.13
```

## E2 Initai Model evalutaion metrics, including confusion matrix

In [74]:

```
#Code Reference (Sklearn.metrics.confusion_matrix, N.d.), (Sklearn.metrics.accuracy_score, N.d.)
#Code Reference (Zach, 2021) for confusion matrix.
Xtest = df[['Initial_days','Children_1','Children_4','Services_CT_Scan','Timely_vis_3','Timely_admis_6','Listen_6
','Hours_6']]
ytest = df['ReAdmis_1']

p = re_log.predict(Xtest)
prediction = list(map(round, p))
```

In [75]:

```
#Confusion Matrix Initial Model
from sklearn.metrics import (confusion_matrix, accuracy_score)
conf_mat = confusion_matrix(ytest, prediction)
print ('Confusion Matrix : \n', conf_mat)
print ('Test Accuracy is ', accuracy_score(ytest, prediction))
```

```
Confusion Matrix :
[[5727 101]
 [ 94 3284]]
Test Accuracy is  0.9788181620682164
```

In [76]:

```
# Prediction Classification Report, Initial Model
from sklearn.metrics import classification_report
print(classification_report(ytest, prediction))
```

	precision	recall	f1-score	support
0	0.98	0.98	0.98	5828
1	0.97	0.97	0.97	3378
accuracy			0.98	9206
macro avg	0.98	0.98	0.98	9206
weighted avg	0.98	0.98	0.98	9206

```
# Predictions from the initial model used to perform the analysis
print(prediction)
```

[illegible]

[illegible]

[illegible]

[illegible]

## Reduced model

**Created from feature selection  $P(z) < .05$  and RMSE minimization**

In [78]:

```
#reduced_log = logit('ReAdmis_1 ~ Initial_days', data=df).fit()
reduced_log = logit('ReAdmis_1 ~ Initial_days+Services CT Scan', data=df).fit()
```

```
Optimization terminated successfully.  
Current function value: 0.048910  
Iterations 13
```

In [79]:

```
print(reduced_log.summary())
```

Logit Regression Results						
=====						
Dep. Variable:	ReAdmis_1	No. Observations:	9206			
Model:	Logit	Df Residuals:	9203			
Method:	MLE	Df Model:	2			
Date:	Tue, 05 Jul 2022	Pseudo R-squ.:	0.9256			
Time:	14:09:36	Log-Likelihood:	-450.27			
converged:	True	LL-Null:	-6051.1			
Covariance Type:	nonrobust	LLR p-value:	0.000			
=====						
	coef	std err	z	P> z	[0.025	0.975]
-----						
Intercept	-53.6643	2.561	-20.955	0.000	-58.684	-48.645
Initial_days	0.9858	0.047	21.022	0.000	0.894	1.078
Services_CT_Scan	0.9831	0.276	3.557	0.000	0.441	1.525
-----						

Possibly complete quasi-separation: A fraction 0.73 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

## E2 Reduced Model Evaluation Metrics, including confusion matrix

In [80]:

```
#Xtest1 = df[['Initial_days']]
Xtest1 = df[['Initial_days', 'Services_CT_Scan']]
ytest1 = df[['ReAdmis_1']]
```

```
# Confusion Matrix for Reduced Model
g = reduced_log.predict(Xtest1)
reduced_prediction = list(map(round, g))
```

```
from sklearn.metrics import (confusion_matrix, accuracy_score)
conf_mat1 = confusion_matrix(ytest1, reduced_prediction)
print ('Confusion Matrix : \n', conf_mat1)
print ('Test Accuracy is ', accuracy_score(ytest1, reduced_prediction))
```

In [83]:

```
from sklearn.metrics import classification_report
```

```
# # Prediction Classification Report, Reduced Model
print(classification_report(ytest1, prediction))
```

In [85]:

```
# Predictions from the reduced model used to perform the analysis
print(reduced_prediction)
```

[illegible]

[illegible]

[illegible]



In [ ]: