Eric Yarger, D208 Task 1: Multiple Regression

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 and Preparation

Initial Feature Selection

Outliers with Zscore

Necessary feature renaming

Dummy Varibles, k-1 number of variables

Univariate & Bivariate Visualization to check for normality

Heatmaps for correlation visualization

Statistical initial feature selection >.02 correlation

In [5]:

```
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  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 [ ]:

```

```

In [ ]:

```

```

In [6]:

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

In [7]:

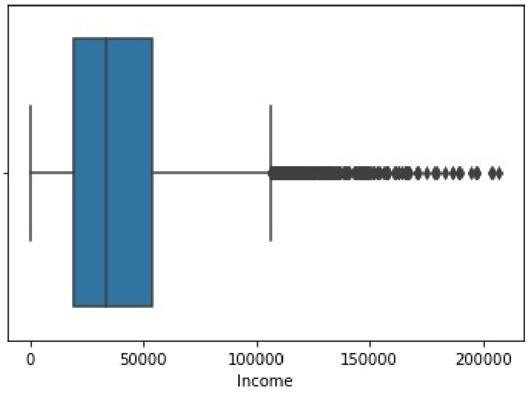
```
#Z-score before boxplots and histograms
```

In [8]:

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

Out[8]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5577188>
```

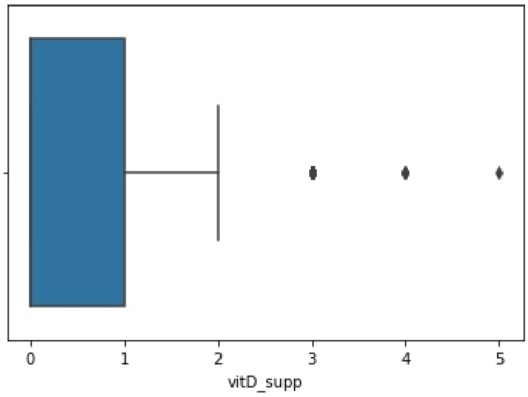


In [9]:

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

Out[9]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5c21948>
```

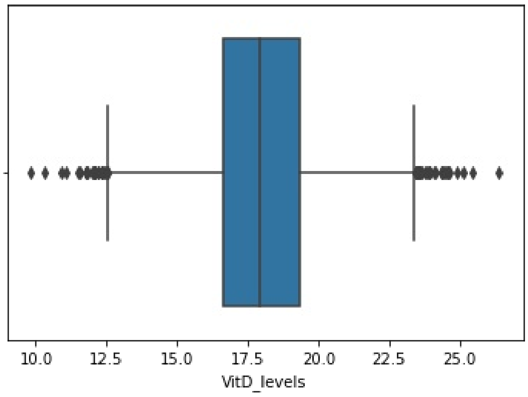


In [10]:

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

Out[10]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5cb9908>
```

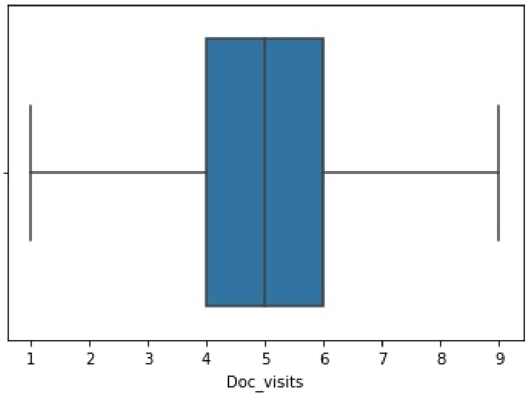


In [11]:

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

Out[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5d2cac8>
```

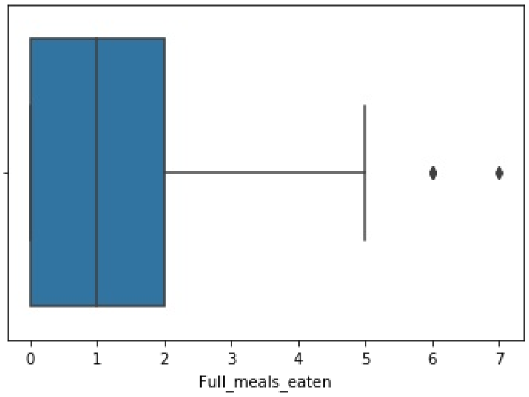


In [12]:

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

Out[12]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5daa888>
```

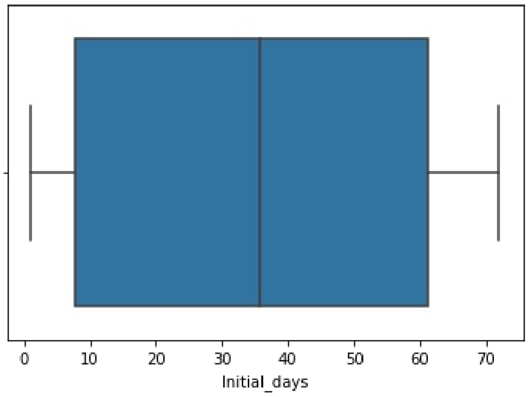


In [13]:

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

Out[13]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5e0e108>
```

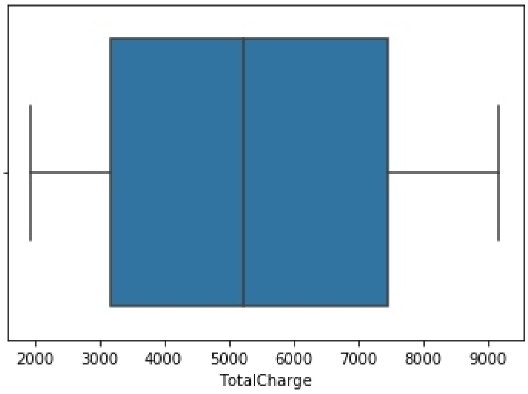


In [14]:

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

Out[14]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5e74c08>
```

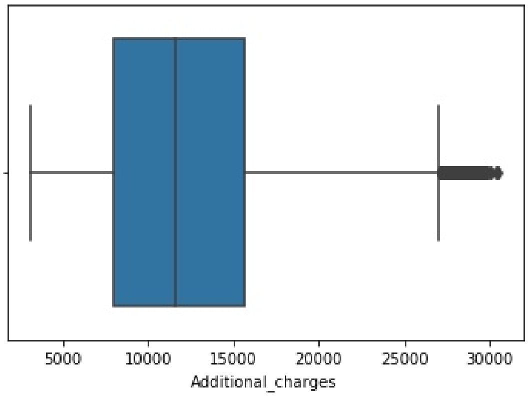


In [15]:

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

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5eeee08>
```

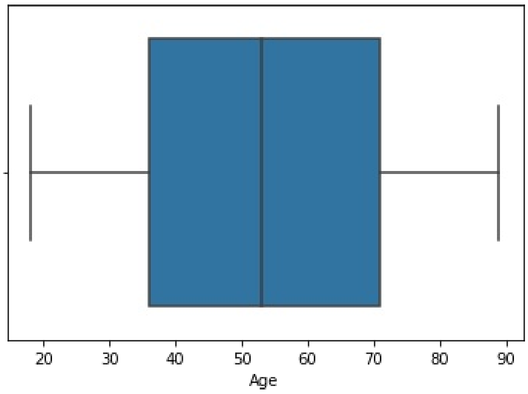


In [16]:

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

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5f46388>
```

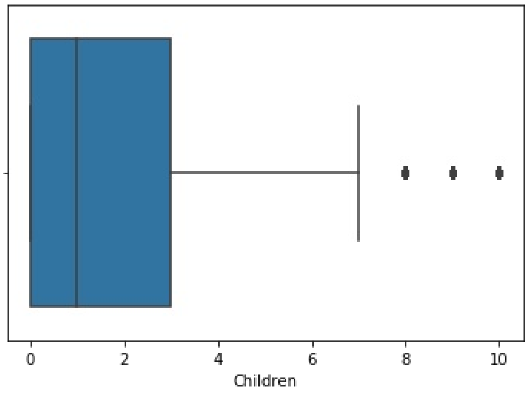


In [17]:

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

Out[17]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5fbc9c8>
```

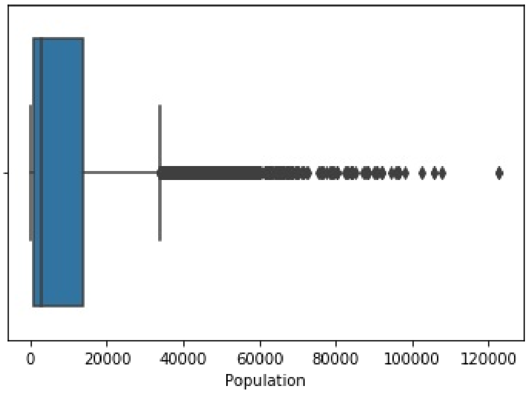


In [18]:

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

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5fbcf88>
```



Z-score calculation and removal of cases >3

Code reference (Bushmanov, 2019)

In [19]:

```
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]:

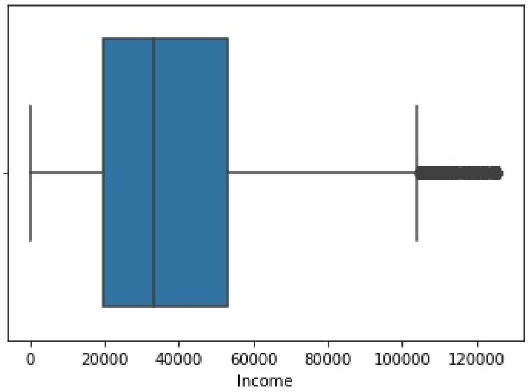
```
#Z-score after boxplots and histograms
```

In [23]:

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

Out[23]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5e15788>
```

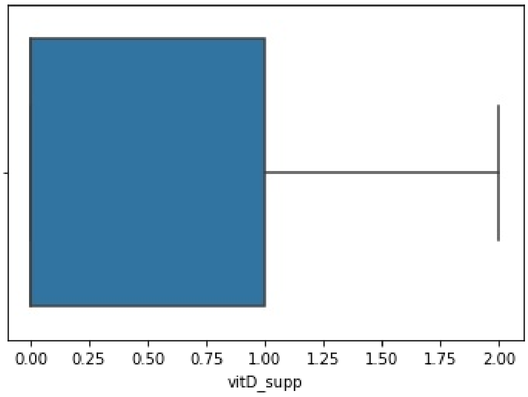


In [24]:

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

Out[24]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea4ebcd88>
```

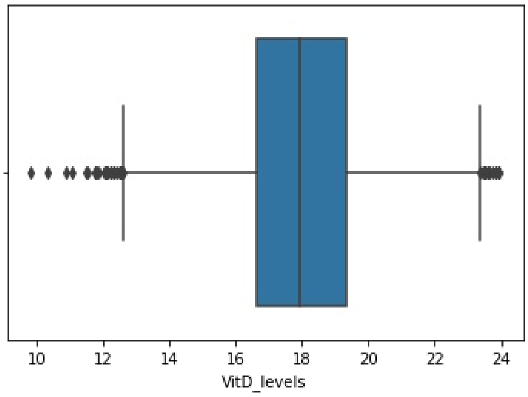


In [25]:

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

Out[25]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea4f25448>
```

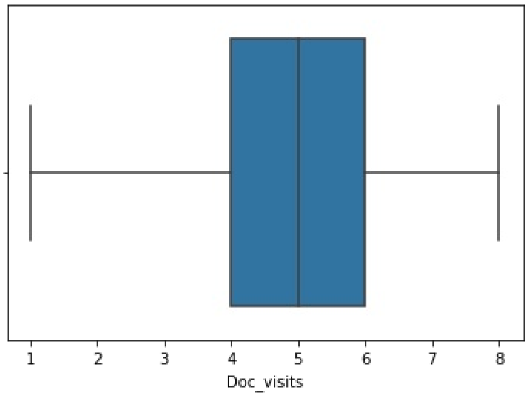


In [26]:

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

Out[26]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea4f948c8>
```

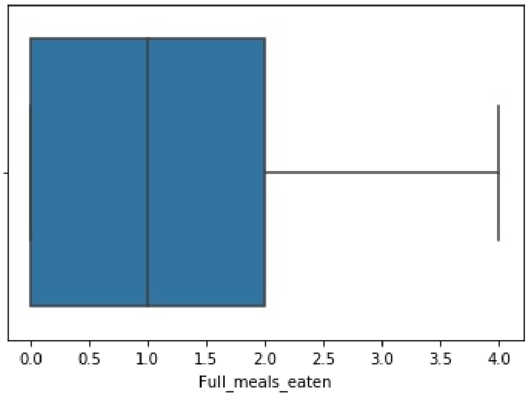


In [27]:

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

Out[27]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5000588>
```

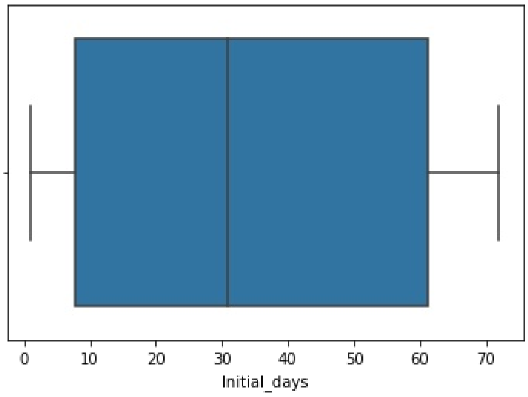


In [28]:

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

Out[28]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea507c088>
```

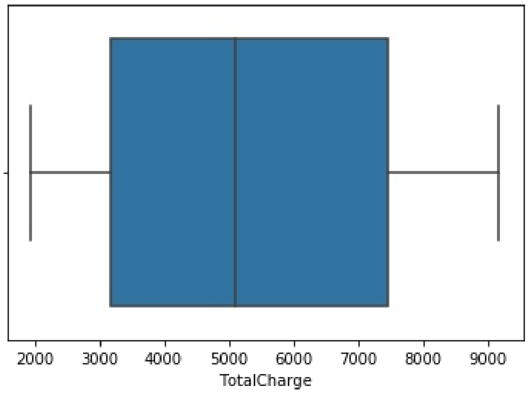


In [29]:

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

Out[29]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea50e2288>
```

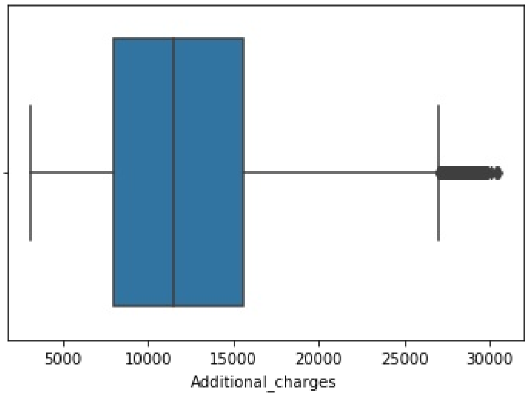


In [30]:

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

Out[30]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea5152c88>
```

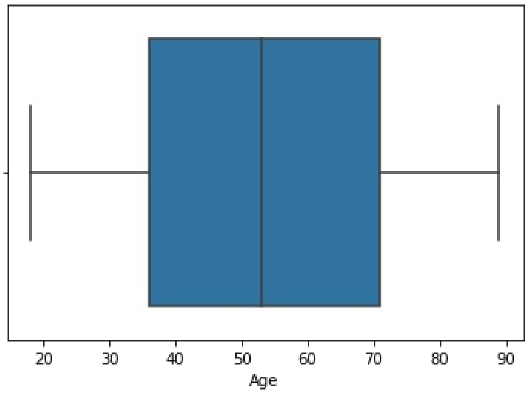


In [31]:

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

Out[31]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea64402c8>
```

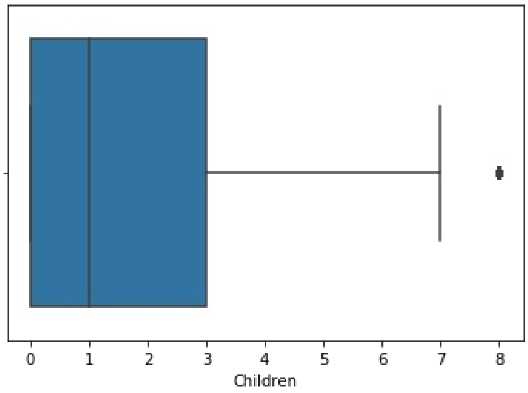


In [32]:

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

Out[32]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea64ad5c8>
```

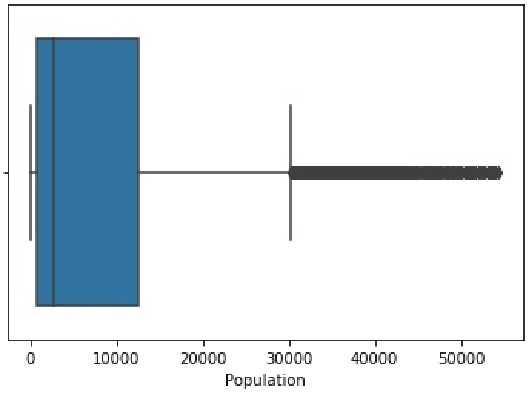


In [33]:

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

Out[33]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x24ea6516d08>
```



In [34]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9206 entries, 0 to 9999
Data columns (total 50 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   CaseOrder           9206 non-null   int64  
 1   Zip                 9206 non-null   int64  
 2   Lat                 9206 non-null   float64
 3   Lng                 9206 non-null   float64
 4   Population          9206 non-null   int64  
 5   Children            9206 non-null   int64  
 6   Age                 9206 non-null   int64  
 7   Income              9206 non-null   float64
 8   VitD_levels         9206 non-null   float64
 9   Doc_visits          9206 non-null   int64  
 10  Full_meals_eaten    9206 non-null   int64  
 11  vitD_supp           9206 non-null   int64  
 12  Initial_days        9206 non-null   float64
 13  TotalCharge         9206 non-null   float64
 14  Additional_charges  9206 non-null   float64
 15  Timely_admis        9206 non-null   int64  
 16  Timely_treat        9206 non-null   int64  
 17  Timely_vis          9206 non-null   int64  
 18  Reliability         9206 non-null   int64  
 19  Options             9206 non-null   int64  
 20  Hours               9206 non-null   int64  
 21  Courteous           9206 non-null   int64  
 22  Listen              9206 non-null   int64  
 23  Customer_id         9206 non-null   object 
 24  Interaction         9206 non-null   object 
 25  UID                 9206 non-null   object 
 26  City                9206 non-null   object 
 27  State               9206 non-null   object 
 28  County              9206 non-null   object 
 29  Area                9206 non-null   object 
 30  TimeZone            9206 non-null   object 
 31  Job                 9206 non-null   object 
 32  Marital             9206 non-null   object 
 33  Gender              9206 non-null   object 
 34  ReAdmis             9206 non-null   object 
 35  Soft_drink          9206 non-null   object 
 36  Initial_admin       9206 non-null   object 
 37  HighBlood           9206 non-null   object 
 38  Stroke              9206 non-null   object 
 39  Complication_risk   9206 non-null   object 
 40  Overweight          9206 non-null   object 
 41  Arthritis           9206 non-null   object 
 42  Diabetes            9206 non-null   object 
 43  Hyperlipidemia      9206 non-null   object 
 44  BackPain            9206 non-null   object 
 45  Anxiety             9206 non-null   object 
 46  Allergic_rhinitis   9206 non-null   object 
 47  Reflux_esophagitis  9206 non-null   object 
 48  Asthma              9206 non-null   object 
 49  Services            9206 non-null   object 
dtypes: float64(7), int64(16), object(27)
memory usage: 3.6+ MB
```

Univariate Visualization

Histograms to look at

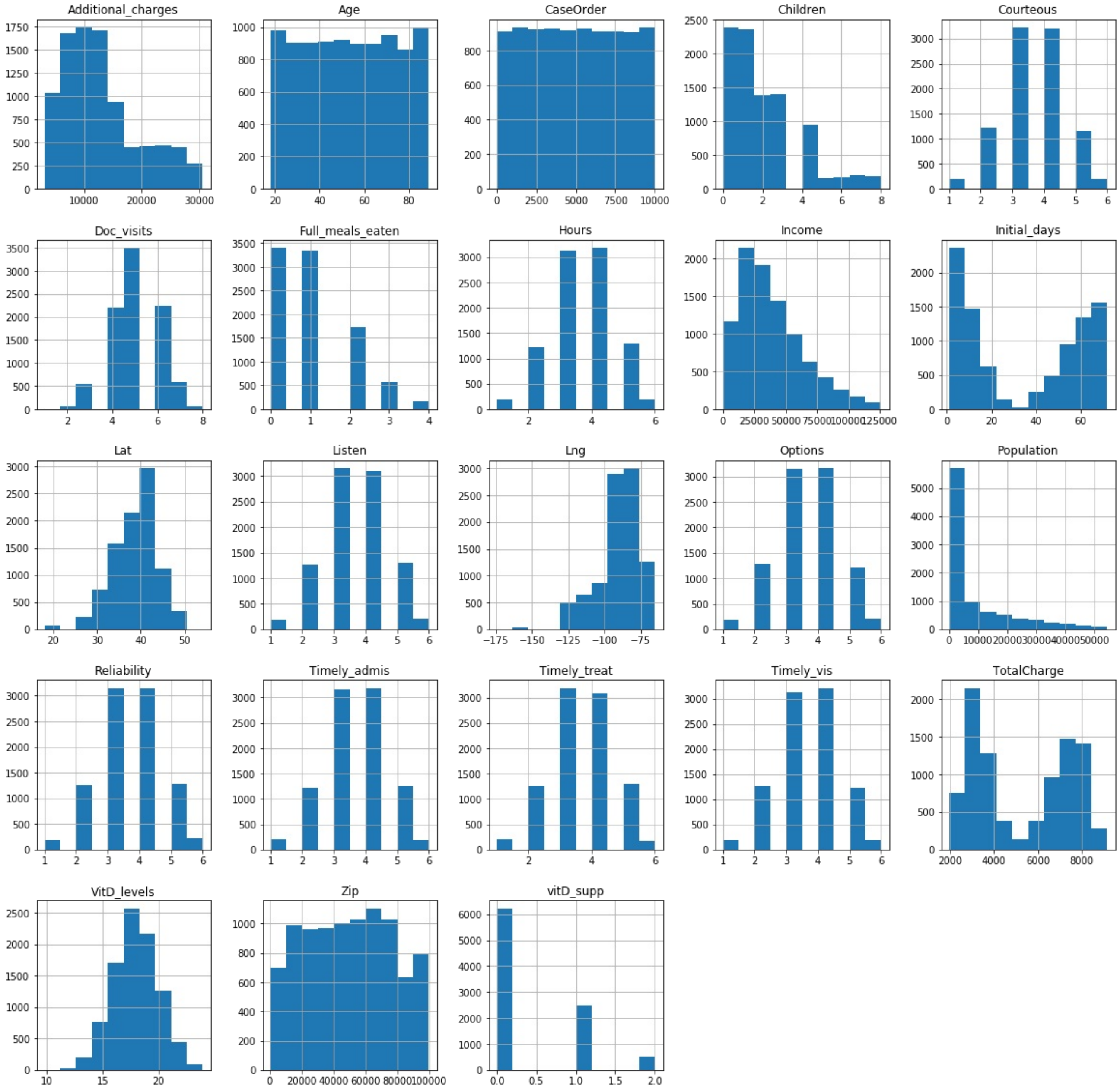
Feature Distribution and Normality

In [35]:

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

```
Out[35]:

array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA64AD488>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA65FE208>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA6636D48>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA666FE48>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA66A6F88>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA66E4048>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA671D148>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7A661C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7A6DD88>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7AA3F88>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7B114C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7B49608>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7B80708>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7BB9808>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7BF2948>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7C2A9C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7C62AC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7C9CC08>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7CD2D08>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7D0EE08>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7D48EC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7D80FC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7DBC108>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7DF4248>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000024EA7E2D408>]],
      dtype=object)
```



Additional_charges

Age

CaseOrder

Children

Courteous

Doc_visits

Full_meals_eaten

Hours

Income

Initial_days

Lat

Listen

Lng

Options

Population

Reliability

Timely_admis

Timely_treat

Timely_vis

TotalCharge

VitD_levels

Zip

vitD_supp

Bivariate Visualization

Scatterplots with

X-Axis = Additional_charges

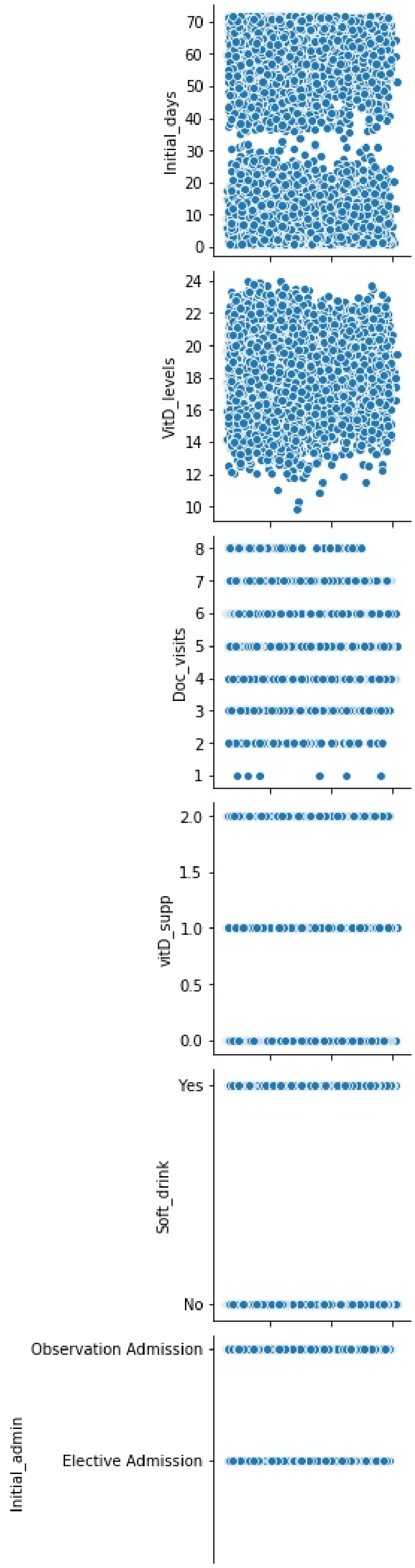
Y-Axis = Independent feature

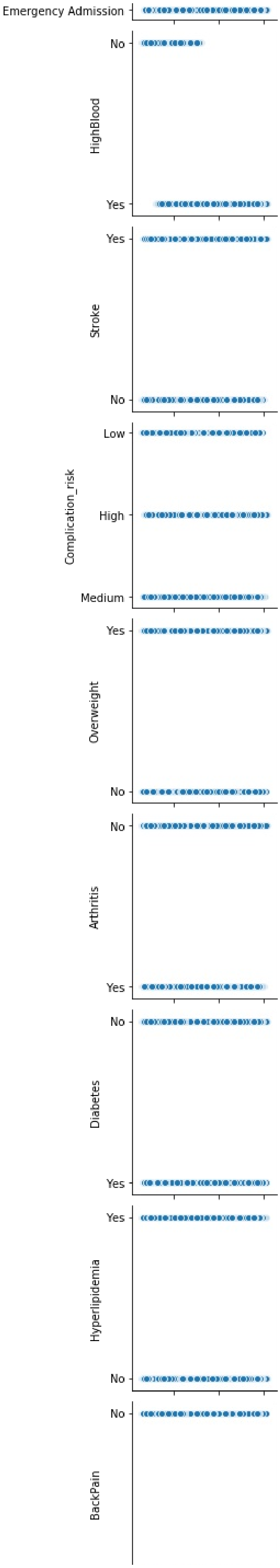
In [36]:

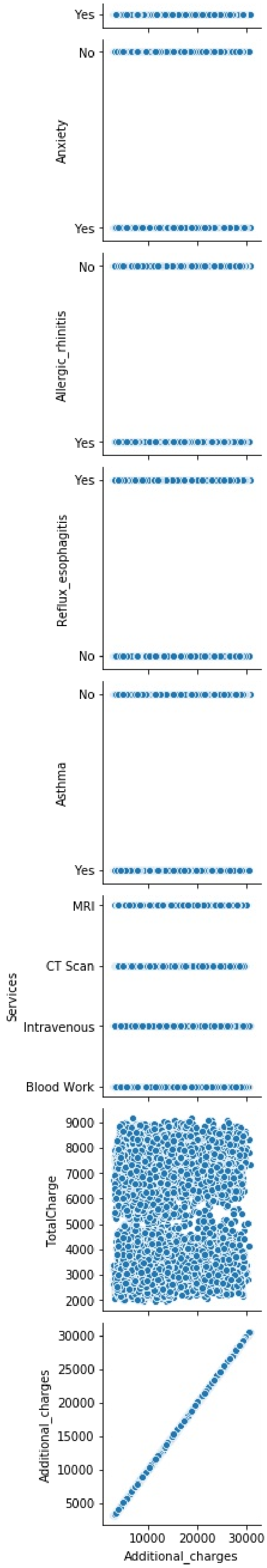
```python
sns.pairplot(df, x_vars=['Additional_charges'], y_vars=['Initial_days','VitD_levels','Doc_visits','vitD_supp','Soft_drink','Initial_admin','HighBlood','Stroke','Complication_risk','Overweight','Arthritis','Diabetes','Hyperlipidemia','BackPain','Anxiety','Allergic_rhinitis','Reflux_esophagitis','Asthma','Services','TotalCharge','Additional_charges'])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

Out[36]:

```
<seaborn.axisgrid.PairGrid at 0x24ea7fa7b48>
```




Emergency Admission
No
HighBlood
Yes
Yes
Stroke
No
Low
Complication_risk
High
Medium
Yes
Overweight
No
No
Arthritis
Yes
No
Diabetes
Yes
Yes
Hyperlipidemia
No
No
BackPain

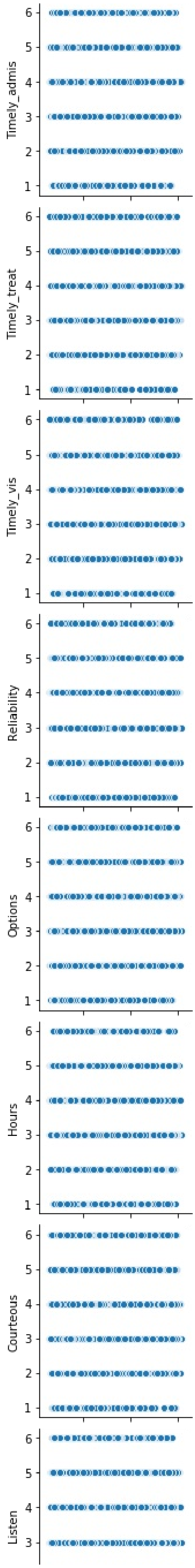


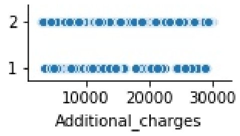
In [37]:

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

```
Out[37]:

<seaborn.axisgrid.PairGrid at 0x24eaaae9cc8>
```


Timely_admis
6
5
4
3
2
1
Timely_treat
6
5
4
3
2
1
Timely_vis
6
5
4
3
2
1
Reliability
6
5
4
3
2
1
Options
6
5
4
3
2
1
Hours
6
5
4
3
2
1
Courteous
6
5
4
3
2
1
Listen
6
5
4
3



In [ ]:

Dummy variables

Drop_first parameter set to True,

ensuring k-1 features to avoid multicollinearity issues

Rename necessary variables

Code Reference (Pandas.get_dummies, n.d.)

In [38]:

```
#Get dummies for categorical features,
#scroll to show drop_first=True at end in Panopto
df = pd.get_dummies(df, columns=['Area','Marital','Gender','Doc_visits','vitD_supp','ReAdmis','Soft_drink','Initial_admin','HighBlood','Stroke','Complication_risk','Overweight','Arthritis','Diabetes','Hyperlipidemia','BackPain','Anxiety','Allergic_rhinitis','Reflux_esophagitis','Asthma','Services'], drop_first=True)
```

In [39]:

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

In [40]:

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

Look at data set size and

Variable correlation

In [41]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9206 entries, 0 to 9999
Data columns (total 98 columns):
 #   Column                               Non-Null Count  Dtype  
---  ------                               --------------  -----  
 0   CaseOrder                            9206 non-null   int64  
 1   Zip                                  9206 non-null   int64  
 2   Lat                                  9206 non-null   float64
 3   Lng                                  9206 non-null   float64
 4   Population                           9206 non-null   int64  
 5   Children                             9206 non-null   int64  
 6   Age                                  9206 non-null   int64  
 7   Income                               9206 non-null   float64
 8   VitD_levels                          9206 non-null   float64
 9   Full_meals_eaten                     9206 non-null   int64  
 10  Initial_days                         9206 non-null   float64
 11  TotalCharge                          9206 non-null   float64
 12  Additional_charges                   9206 non-null   float64
 13  Customer_id                          9206 non-null   object 
 14  Interaction                          9206 non-null   object 
 15  UID                                  9206 non-null   object 
 16  City                                 9206 non-null   object 
 17  State                                9206 non-null   object 
 18  County                               9206 non-null   object 
 19  TimeZone                             9206 non-null   object 
 20  Job                                  9206 non-null   object 
 21  Area_Suburban                        9206 non-null   uint8  
 22  Area_Urban                           9206 non-null   uint8
 23  Marital_Married                      9206 non-null   uint8
 24  Marital_Never_Married                9206 non-null   uint8
 25  Marital_Separated                    9206 non-null   uint8
 26  Marital_Widowed                      9206 non-null   uint8
 27  Gender_Male                          9206 non-null   uint8
 28  Gender_Nonbinary                     9206 non-null   uint8
 29  Doc_visits_2                         9206 non-null   uint8
 30  Doc_visits_3                         9206 non-null   uint8
 31  Doc_visits_4                         9206 non-null   uint8
 32  Doc_visits_5                         9206 non-null   uint8
 33  Doc_visits_6                         9206 non-null   uint8
 34  Doc_visits_7                         9206 non-null   uint8
 35  Doc_visits_8                         9206 non-null   uint8
 36  vitD_supp_1                          9206 non-null   uint8
 37  vitD_supp_2                          9206 non-null   uint8
 38  ReAdmis_Yes                          9206 non-null   uint8
 39  Soft_drink_Yes                       9206 non-null   uint8
 40  Initial_admin_Emergency_Admission    9206 non-null   uint8
 41  Initial_admin_Observation_Admission  9206 non-null   uint8
 42  HighBlood_Yes                        9206 non-null   uint8
 43  Stroke_Yes                           9206 non-null   uint8
 44  Complication_risk_Low                9206 non-null   uint8
 45  Complication_risk_Medium             9206 non-null   uint8
 46  Overweight_Yes                       9206 non-null   uint8
 47  Arthritis_Yes                        9206 non-null   uint8
 48  Diabetes_Yes                         9206 non-null   uint8
 49  Hyperlipidemia_Yes                   9206 non-null   uint8
 50  BackPain_Yes                         9206 non-null   uint8
 51  Anxiety_Yes                          9206 non-null   uint8
 52  Allergic_rhinitis_Yes                9206 non-null   uint8
 53  Reflux_esophagitis_Yes               9206 non-null   uint8
 54  Asthma_Yes                           9206 non-null   uint8
 55  Services_CT Scan                     9206 non-null   uint8
 56  Services_Intravenous                 9206 non-null   uint8
 57  Services_MRI                         9206 non-null   uint8
 58  Timely_admis_2                       9206 non-null   uint8
 59  Timely_admis_3                       9206 non-null   uint8
 60  Timely_admis_4                       9206 non-null   uint8
 61  Timely_admis_5                       9206 non-null   uint8
 62  Timely_admis_6                       9206 non-null   uint8
 63  Timely_treat_2                       9206 non-null   uint8
 64  Timely_treat_3                       9206 non-null   uint8
 65  Timely_treat_4                       9206 non-null   uint8
 66  Timely_treat_5                       9206 non-null   uint8
 67  Timely_treat_6                       9206 non-null   uint8
 68  Timely_vis_2                         9206 non-null   uint8
 69  Timely_vis_3                         9206 non-null   uint8
 70  Timely_vis_4                         9206 non-null   uint8
 71  Timely_vis_5                         9206 non-null   uint8
 72  Timely_vis_6                         9206 non-null   uint8
 73  Reliability_2                        9206 non-null   uint8
 74  Reliability_3                        9206 non-null   uint8
 75  Reliability_4                        9206 non-null   uint8
 76  Reliability_5                        9206 non-null   uint8
 77  Reliability_6                        9206 non-null   uint8
 78  Options_2                            9206 non-null   uint8
 79  Options_3                            9206 non-null   uint8
 80  Options_4                            9206 non-null   uint8
 81  Options_5                            9206 non-null   uint8
 82  Options_6                            9206 non-null   uint8
 83  Hours_2                              9206 non-null   uint8
 84  Hours_3                              9206 non-null   uint8
 85  Hours_4                              9206 non-null   uint8
 86  Hours_5                              9206 non-null   uint8
 87  Hours_6                              9206 non-null   uint8
 88  Courteous_2                          9206 non-null   uint8
 89  Courteous_3                          9206 non-null   uint8
 90  Courteous_4                          9206 non-null   uint8
 91  Courteous_5                          9206 non-null   uint8
 92  Courteous_6                          9206 non-null   uint8
 93  Listen_2                             9206 non-null   uint8
 94  Listen_3                             9206 non-null   uint8
 95  Listen_4                             9206 non-null   uint8
 96  Listen_5                             9206 non-null   uint8
 97  Listen_6                             9206 non-null   uint8
dtypes: float64(7), int64(6), object(8), uint8(77)
memory usage: 2.2+ MB
```

In [42]:

```
df.corr()
```

Out[42]:

	CaseOrder	Zip	Lat	Lng	Population	Children	Age	Income	VitD_levels	Full_meals_eaten	...	Cou
CaseOrder	1.000000	0.010465	-0.012946	-0.012081	0.001489	0.017027	-0.003011	-0.012265	-0.015026	-0.020805	...	
Zip	0.010465	1.000000	-0.084258	-0.913573	0.012947	0.014307	-0.003327	0.010507	-0.010747	0.013077	...	
Lat	-0.012946	-0.084258	1.000000	0.001062	-0.187334	0.005874	-0.000132	-0.015414	-0.005158	-0.001353	...	
Lng	-0.012081	-0.913573	0.001062	1.000000	-0.018263	-0.014141	0.002780	-0.008175	0.000931	-0.013120	...	
Population	0.001489	0.012947	-0.187334	-0.018263	1.000000	0.007810	-0.018884	0.002162	0.004719	-0.025711	...	
...	...	...	...	...	...	...	...	...	...	...	...	
Listen_2	0.006526	0.009945	-0.006435	-0.003020	0.013885	0.008641	0.002051	-0.012318	0.014576	0.014706	...	
Listen_3	0.002207	-0.003263	0.008912	0.005719	-0.006724	0.006598	-0.006864	-0.014727	0.010725	0.003840	...	
Listen_4	-0.001205	0.016042	-0.012675	-0.021459	0.003848	-0.022668	-0.003206	0.028375	0.001107	0.000904	...	
Listen_5	-0.003449	-0.021017	0.001714	0.017698	-0.001610	0.008179	0.006968	0.005117	-0.021362	-0.015708	...	
Listen_6	-0.014306	-0.005841	0.018726	0.007310	-0.009906	-0.000516	0.004917	-0.011113	-0.016527	-0.012259	...	

90 rows × 90 columns

Heatmaps for correlation visualization

Code reference (seaborn.heatmap, n.d.)

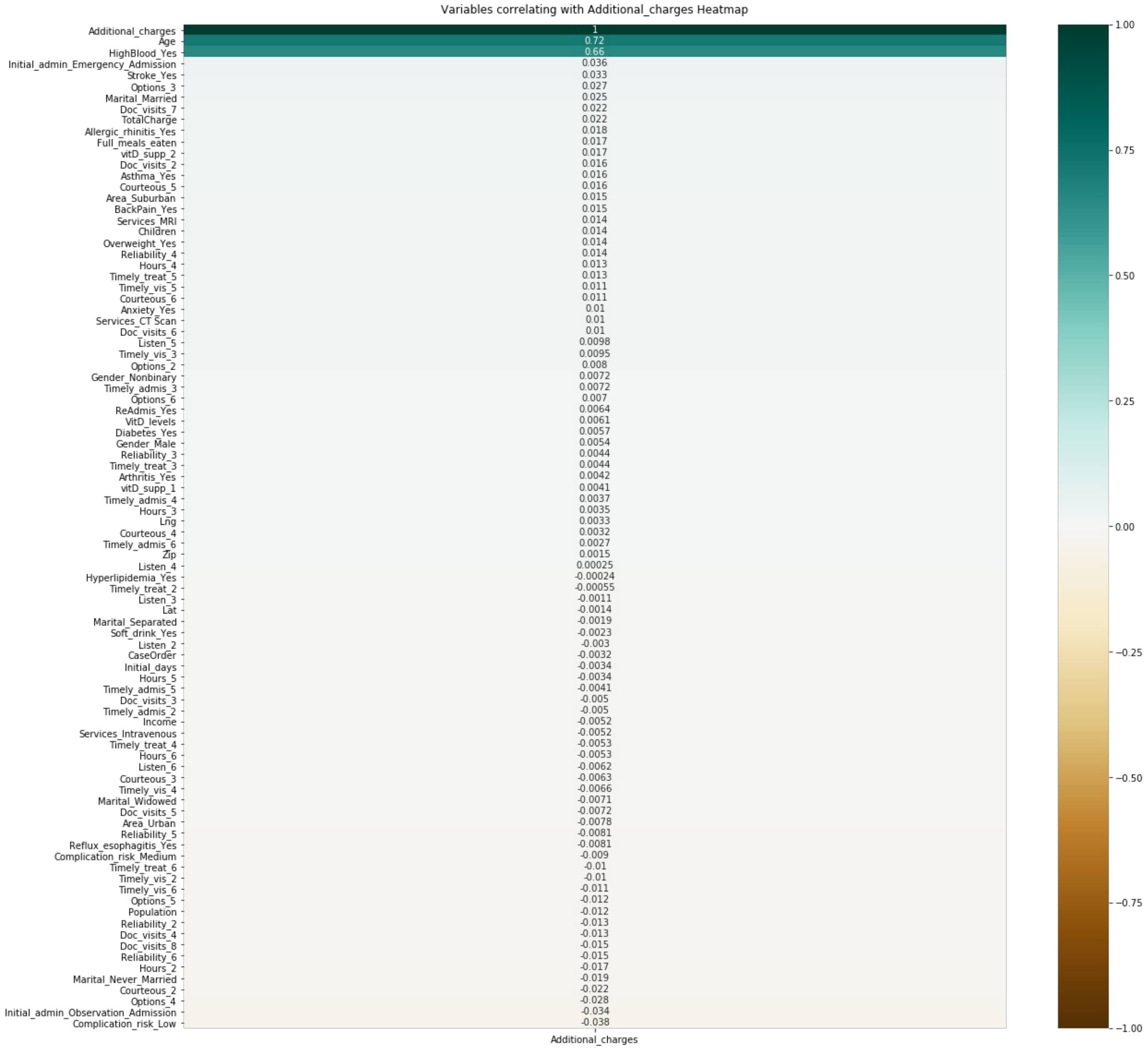
In [43]:

```
import matplotlib
matplotlib.pyplot.figure(figsize=(20,20))
heatmap = sns.heatmap(df.corr()[['Additional_charges']].sort_values(by='Additional_charges', ascending=False), vm
in=-1, vmax=1, annot=True, cmap='BrBG')
heatmap.set_title('Variables correlating with Additional_charges Heatmap',pad=12)

abs(df.corr()['Additional_charges'])
```

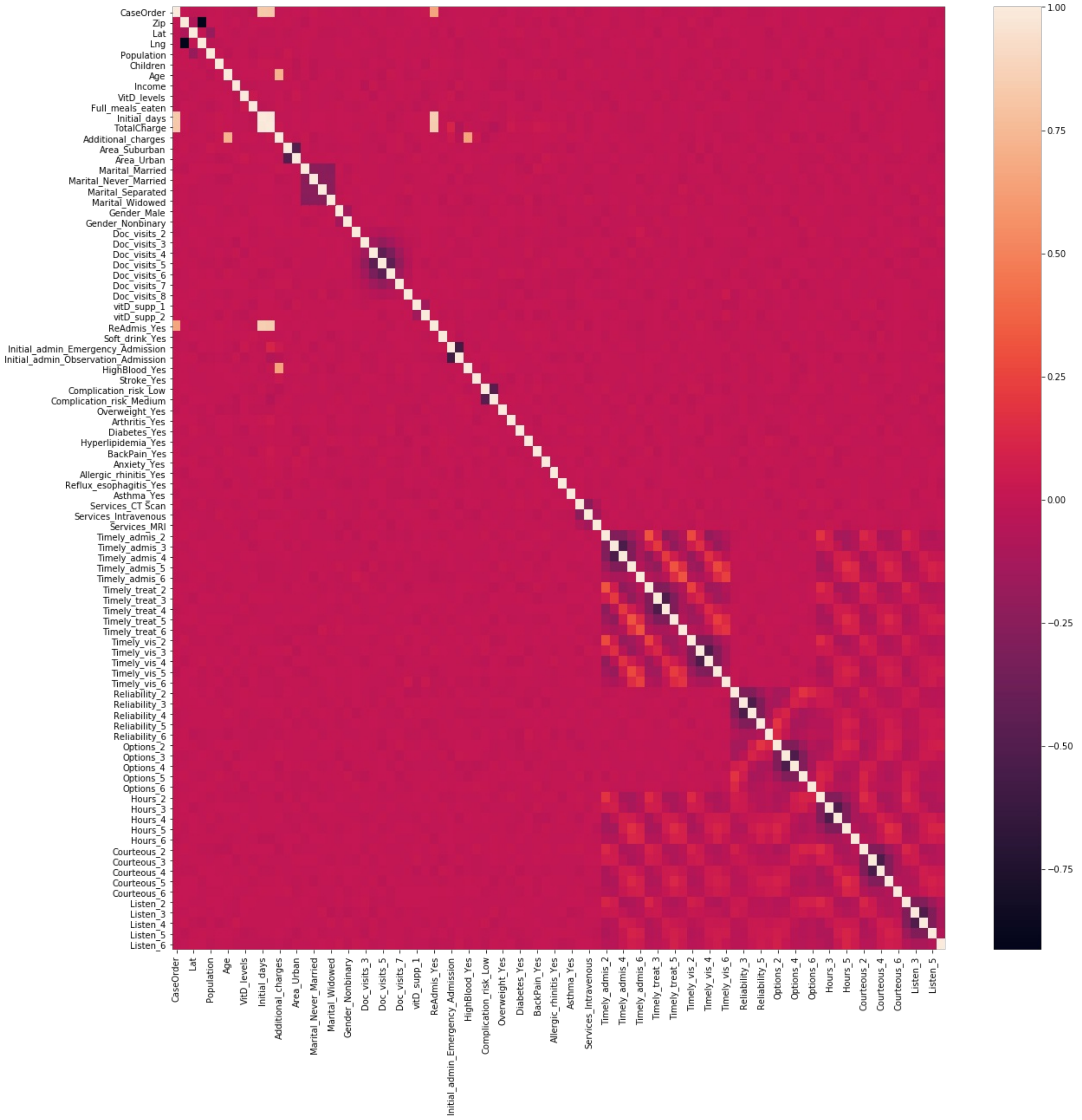
Out[43]:

```
CaseOrder     0.003178
Zip           0.001545
Lat           0.001433
Lng           0.003290
Population    0.011835
                ...
Listen_2      0.002972
Listen_3      0.001072
Listen_4      0.000247
Listen_5      0.009820
Listen_6      0.006197
Name: Additional_charges, Length: 90, dtype: float64
```



In [44]:

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



In [ ]:

```

```

C2 Summary Statistics

In [45]:

```
dfc = df[['Additional_charges','Age','TotalCharge','Marital_Married','Doc_visits_7','Initial_admin_Emergency_Admission','Initial_admin_Observation_Admission','HighBlood_Yes','Stroke_Yes','Complication_risk_Low','Options_3','Options_4','Courteous_2']]
```

In [46]:

```
dfc.describe()
```

Out[46]:

	Additional_charges	Age	TotalCharge	Marital_Married	Doc_visits_7	Initial_admin_Emergency_Admission	Initial_admin_Obs
count	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	
mean	12927.980718	53.543124	5306.435876	0.203889	0.063871	0.505323	
std	6540.592828	20.609439	2181.251460	0.402909	0.244537	0.499999	
min	3125.703000	18.000000	1938.312067	0.000000	0.000000	0.000000	
25%	7991.171750	36.000000	3178.291852	0.000000	0.000000	0.000000	
50%	11556.775000	53.000000	5100.260500	0.000000	0.000000	1.000000	
75%	15602.158960	71.000000	7458.542500	0.000000	0.000000	1.000000	
max	30566.070000	89.000000	9180.728000	1.000000	1.000000	1.000000	

In [47]:

```
dfc.corr()
```

Out[47]:

	Additional_charges	Age	TotalCharge	Marital_Married	Doc_visits_7	Initial_admin_Emergency
Additional_charges	1.000000	0.716409	0.022020	0.025245	0.022486	
Age	0.716409	1.000000	0.010785	0.012580	0.005877	
TotalCharge	0.022020	0.010785	1.000000	0.000992	-0.000072	
Marital_Married	0.025245	0.012580	0.000992	1.000000	0.005638	
Doc_visits_7	0.022486	0.005877	-0.000072	0.005638	1.000000	
Initial_admin_Emergency_Admission	0.036228	-0.004498	0.107284	0.016453	0.015878	
Initial_admin_Observation_Admission	-0.034389	-0.010404	-0.069032	-0.012139	0.000431	
HighBlood_Yes	0.655680	0.008265	0.015240	0.022163	0.026361	
Stroke_Yes	0.033301	0.011657	-0.007641	-0.006432	0.012226	
Complication_risk_Low	-0.038131	0.000604	-0.014872	-0.001313	-0.005083	
Options_3	0.026605	0.014939	-0.001122	-0.006154	-0.019437	
Options_4	-0.027679	-0.019079	0.000477	0.013666	0.010899	
Courteous_2	-0.022164	-0.023345	0.006867	-0.008226	-0.006694	

In [48]:

```
dfc.mean()
```

Out[48]:

```
Additional_charges                     12927.980718
Age                                       53.543124
TotalCharge                             5306.435876
Marital_Married                            0.203889
Doc_visits_7                               0.063871
Initial_admin_Emergency_Admission          0.505323
Initial_admin_Observation_Admission        0.244189
HighBlood_Yes                              0.407886
Stroke_Yes                                 0.199001
Complication_risk_Low                      0.212036
Options_3                                  0.341408
Options_4                                  0.344123
Courteous_2                                0.132848
dtype: float64
```

In [49]:

```
dfc.median()
```

Out[49]:

```
Additional_charges                      11556.7750
Age                                        53.0000
TotalCharge                              5100.2605
Marital_Married                             0.0000
Doc_visits_7                                0.0000
Initial_admin_Emergency_Admission           1.0000
Initial_admin_Observation_Admission         0.0000
HighBlood_Yes                               0.0000
Stroke_Yes                                  0.0000
Complication_risk_Low                       0.0000
Options_3                                   0.0000
Options_4                                   0.0000
Courteous_2                                 0.0000
dtype: float64
```

In [ ]:

```

```

In [50]:

```
df.to_excel('C:/Users/ericy/Desktop/D208_all_variables.xlsx', index=False)
```

Statistical Feature Selection

Correlation > .02

Casts a wide net for initial feature selection

In [51]:

```
abs(df.corr()["Additional_charges"][abs(df.corr()["Additional_charges"])>0.02].drop('Additional_charges')).index.
tolist()
```

Out[51]:

```
['Age',
 'TotalCharge',
 'Marital_Married',
 'Doc_visits_7',
 'Initial_admin_Emergency_Admission',
 'Initial_admin_Observation_Admission',
 'HighBlood_Yes',
 'Stroke_Yes',
 'Complication_risk_Low',
 'Options_3',
 'Options_4',
 'Courteous_2']
```

In [ ]:

```

```

In [52]:

```
df.to_excel('C:/Users/ericy/Desktop/D208_clean.xlsx', index=False)
```

Initial Multiple Regression

Using Ordinary Least Squared (OLS) Regression

Code Reference ("Ordinary Least Squares (OLS) using statsmodels", 2022)

In [53]:

```
X=df[['Age','TotalCharge','Marital_Married','Doc_visits_7','Initial_admin_Emergency_Admission','Initial_admin_Obs
ervation_Admission','HighBlood_Yes','Stroke_Yes','Complication_risk_Low','Options_3','Options_4','Courteous_2']]
y=df['Additional_charges']
```

In [54]:

```
import statsmodels.api as sm

X= sm.add_constant(X)
```

In [55]:

```
ols = sm.OLS(y, X).fit()
```

In [56]:

```
print(ols.summary())
```

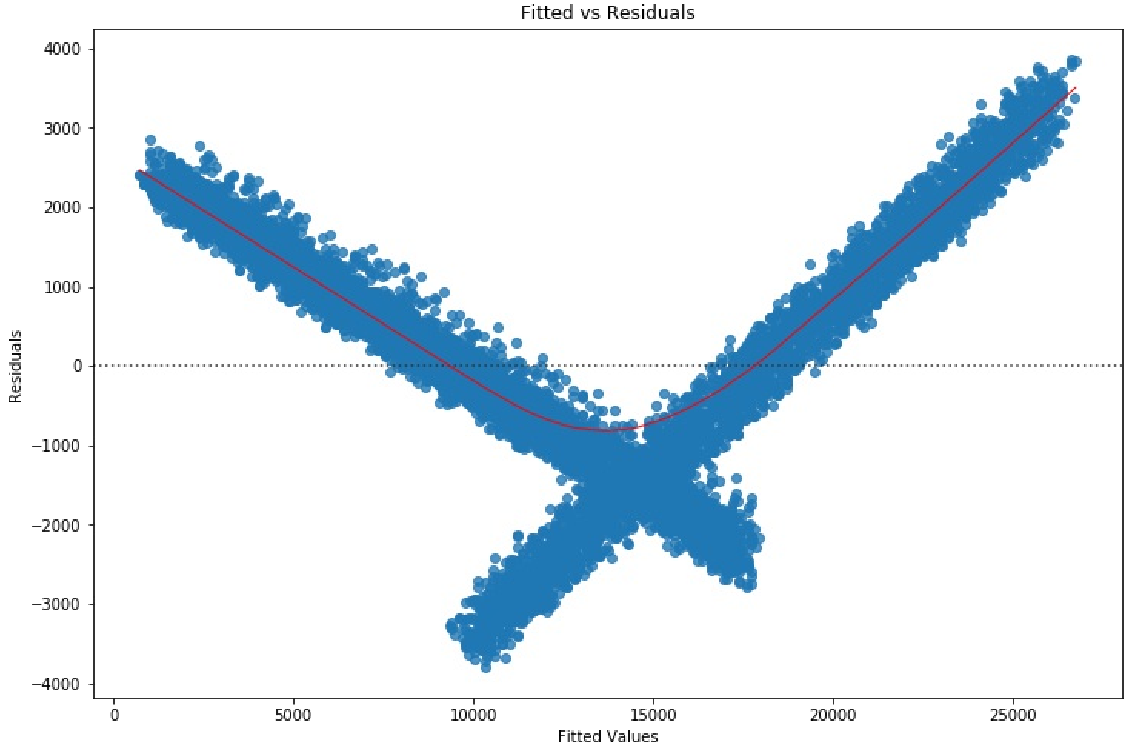
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Additional_charges   R-squared:                       0.938
Model:                            OLS   Adj. R-squared:                  0.938
Method:                 Least Squares   F-statistic:                 1.157e+04
Date:                Thu, 30 Jun 2022   Prob (F-statistic):               0.00
Time:                        18:50:14   Log-Likelihood:                -81153.
No. Observations:                9206   AIC:                         1.623e+05
Df Residuals:                    9193   BIC:                         1.624e+05
Df Model:                          12
Covariance Type:            nonrobust
====================================================================================================
===
                                          coef    std err          t      P>|t|      [0.025      0.9
75]
----------------------------------------------------------------------------------------------------
---
const                               -2927.6667     76.352    -38.344      0.000   -3077.334   -2777.
999
Age                                   225.5595      0.826    273.225      0.000     223.941     227.
178
TotalCharge                             0.0001      0.008      0.017      0.987      -0.015       0.
016
Marital_Married                        22.9864     42.219      0.544      0.586     -59.772     105.
745
Doc_visits_7                           11.2221     69.579      0.161      0.872    -125.169     147.
613
Initial_admin_Emergency_Admission     470.9664     41.702     11.294      0.000     389.222     552.
711
Initial_admin_Observation_Admission  -102.0149     48.366     -2.109      0.035    -196.823      -7.
207
HighBlood_Yes                        8636.1851     34.638    249.325      0.000    8568.286    8704.
084
Stroke_Yes                            353.0603     42.611      8.286      0.000     269.533     436.
587
Complication_risk_Low                -292.2519     41.620     -7.022      0.000    -373.836    -210.
668
Options_3                             100.4664     42.089      2.387      0.017      17.963     182.
969
Options_4                             -14.0420     41.966     -0.335      0.738     -96.304      68.
221
Courteous_2                            -7.7094     50.265     -0.153      0.878    -106.240      90.
822
==============================================================================
Omnibus:                     1309.636   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              316.742
Skew:                          -0.020   Prob(JB):                     1.66e-69
Kurtosis:                       2.092   Cond. No.                     2.84e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.84e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Additional_charges = -2927.6667 +Age(225.5595)+TotalCharge(.0001)+Marital_Married(22.9864)+Initial_admin_Emergency_Ad Initial_admin_Observation_Admission(102.0149)+HighBlood_Yes(8636.1851)+Stroke_Yes(35 Complication_risk_Low(292.2519)+Options_3(100.4664)-Options_4(14.0420)- Courteous_2(7.7094)

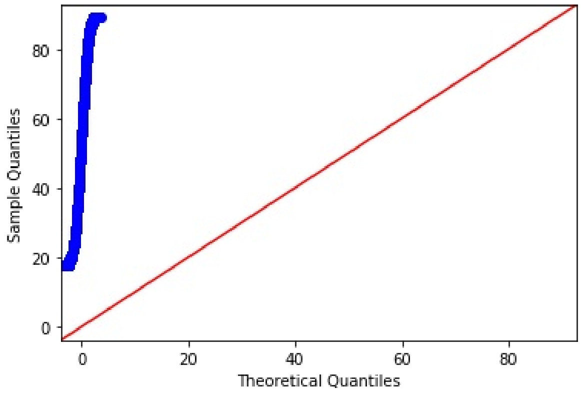
In [57]:

```
plt.figure(figsize=(12,8))
plt.title('Fitted vs Residuals')
sns.residplot(ols.fittedvalues,ols.resid,lowess=True,line_kws={'color':'r','lw':1})
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.show()
```



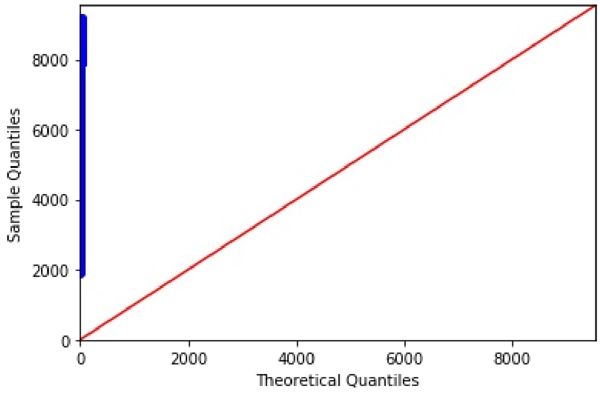
In [58]:

```
fig = sm.qqplot(df['Age'], line='45')
plt.show()
```



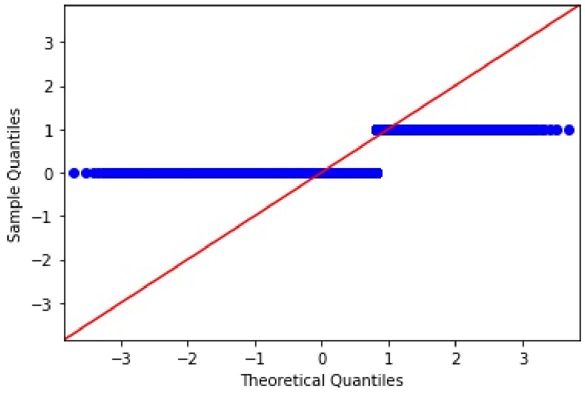
In [59]:

```
fig = sm.qqplot(df['TotalCharge'], line='45')
plt.show()
```



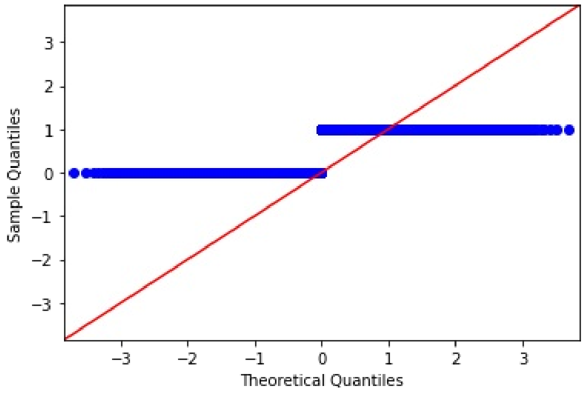
In [60]:

```
fig = sm.qqplot(df['Marital_Married'], line='45')
plt.show()
```



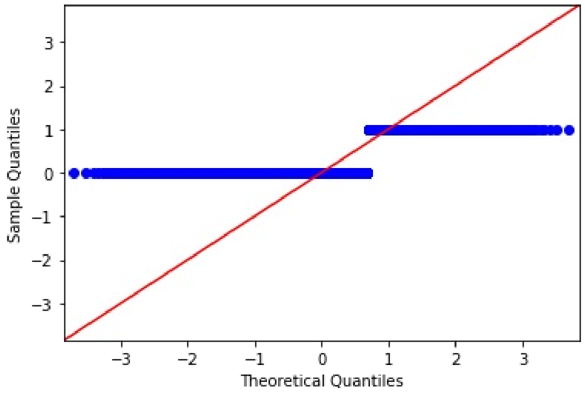
In [61]:

```
fig = sm.qqplot(df['Initial_admin_Emergency_Admission'], line='45')
plt.show()
```



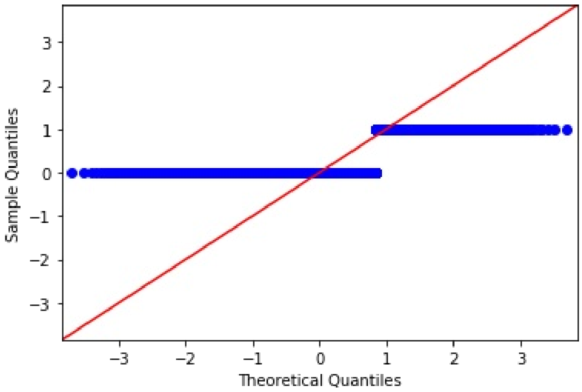
In [62]:

```
fig = sm.qqplot(df['Initial_admin_Observation_Admission'], line='45')
plt.show()
```



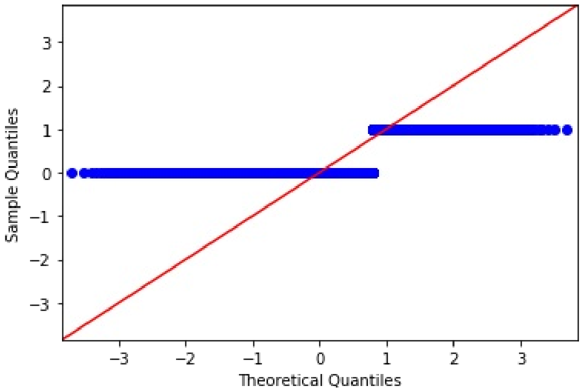
In [63]:

```
fig = sm.qqplot(df['Stroke_Yes'], line='45')
plt.show()
```



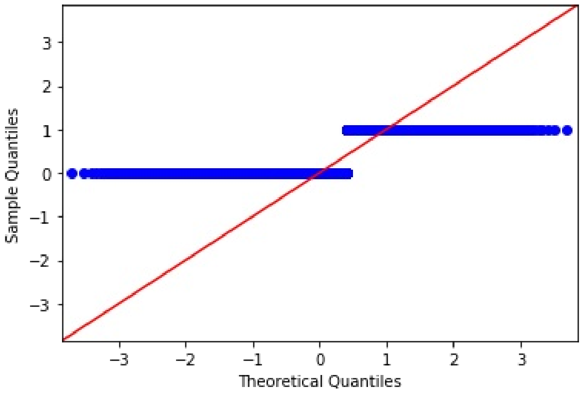
In [64]:

```
fig = sm.qqplot(df['Complication_risk_Low'], line='45')
plt.show()
```



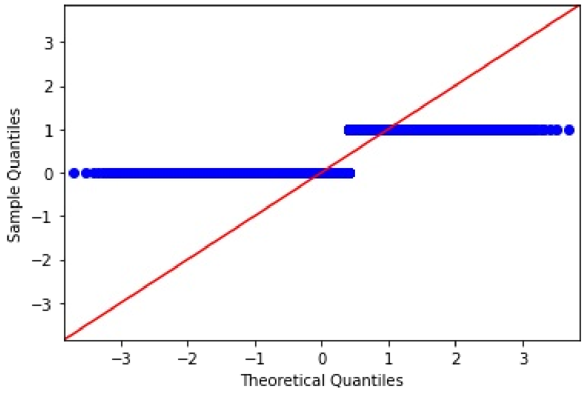
In [65]:

```
fig = sm.qqplot(df['Options_3'], line='45')
plt.show()
```



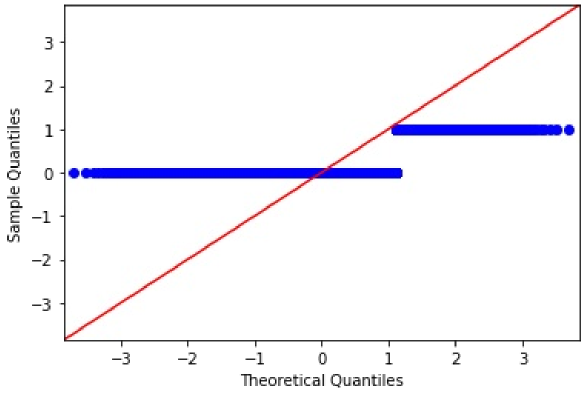
In [66]:

```
fig = sm.qqplot(df['Options_4'], line='45')
plt.show()
```



In [67]:

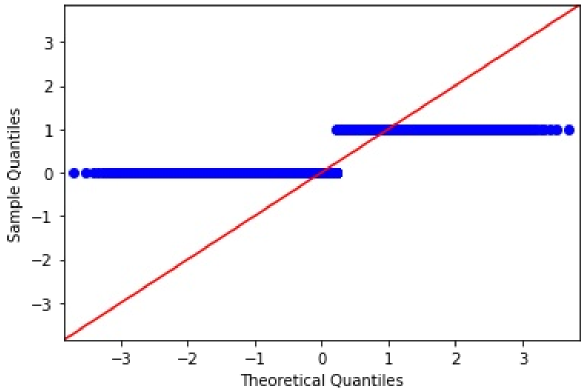
```
fig = sm.qqplot(df['Courteous_2'], line='45')
plt.show()
```



In [ ]:

In [68]:

```
fig = sm.qqplot(df['HighBlood_Yes'], line='45')
plt.show()
```



Multiple Regression Feature Reduction

Variance Inflation Factor (VIF)

Look for Multicollinearity between Independent variables

Code Reference (Zach, 2020)

In [69]:

```
from patsy import dmatrices
from statsmodels.stats.outliers_influence import variance_inflation_factor

y, X = dmatrices('Additional_charges ~ Age+TotalCharge+Marital_Married+Initial_admin_Emergency_Admission+Initial_
admin_Observation_Admission+HighBlood_Yes+Stroke_Yes+Complication_risk_Low+Options_3+Options_4+Courteous_2', data
=df, return_type='dataframe')
```

In [70]:

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

In [71]:

```
vif
```

Out[71]:

	VIF	variable
0	20.132471	Intercept
1	1.001660	Age
2	1.012415	TotalCharge
3	1.001234	Marital_Married
4	1.503821	Initial_admin_Emergency_Admission
5	1.493896	Initial_admin_Observation_Admission
6	1.002135	HighBlood_Yes
7	1.001448	Stroke_Yes
8	1.001558	Complication_risk_Low
9	1.378054	Options_3
10	1.375611	Options_4
11	1.007220	Courteous_2

Feature selection, differing levels of correlation

and corresponding R_Squared & Root Mean Square Error (RMSE)

In [72]:

```
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 Nearast Neighbors & correlation for feature selection

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

In [73]:

```
X=df[['Age','Marital_Married','Initial_admin_Emergency_Admission','Initial_admin_Observation_Admission','HighBloo
d_Yes','Stroke_Yes','Complication_risk_Low','Options_3','Options_4','Courteous_2']]
y = df.Additional_charges
```

In [74]:

```
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("R_squared: " + str(round(r2_score(y,y_pred),2)))
```

```
RMSE: 1395.79
R_squared: 0.95
```

In [75]:

```
vals = [0.02,0.05,0.08,0.1,0.2]
for val in vals:
    features = abs(df.corr()["Additional_charges"][abs(df.corr()["Additional_charges"])>val].drop('Additional_cha
rges')).index.tolist()

    X = df.drop(columns='Additional_charges')
    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)))
    print("R_squared: " + str(round(r2_score(y,y_pred),2)))
```

```
['Age', 'TotalCharge', 'Marital_Married', 'Doc_visits_7', 'Initial_admin_Emergency_Admission', 'Init
ial_admin_Observation_Admission', 'HighBlood_Yes', 'Stroke_Yes', 'Complication_risk_Low', 'Options_3
', 'Options_4', 'Courteous_2']
RMSE: 1838.19
R_squared: 0.92
['Age', 'HighBlood_Yes']
RMSE: 408.78
R_squared: 1.0
['Age', 'HighBlood_Yes']
RMSE: 408.78
R_squared: 1.0
['Age', 'HighBlood_Yes']
RMSE: 408.78
R_squared: 1.0
['Age', 'HighBlood_Yes']
RMSE: 408.78
R_squared: 1.0
```

OLS, Reduced Model With

Age (continuous independent feature)

HighBlood_Yes (catagorical independent feature)

In [76]:

```
X=df[['Age','HighBlood_Yes']]
y=df['Additional_charges']
```

In [77]:

```
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import OLSInfluence
X= sm.add_constant(X)
```

In [78]:

```
ols = sm.OLS(y, X).fit()
```

In [79]:

```
print(ols.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Additional_charges   R-squared:                       0.935
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                 6.669e+04
Date:                Thu, 30 Jun 2022   Prob (F-statistic):               0.00
Time:                        18:50:59   Log-Likelihood:                -81330.
No. Observations:                9206   AIC:                         1.627e+05
Df Residuals:                    9203   BIC:                         1.627e+05
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const         -2681.5619     50.211    -53.406      0.000   -2779.986   -2583.137
Age             225.6544      0.840    268.483      0.000     224.007     227.302
HighBlood_Yes  8647.7532     35.245    245.361      0.000    8578.665    8716.841
==============================================================================
Omnibus:                      931.630   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              269.728
Skew:                          -0.021   Prob(JB):                     2.69e-59
Kurtosis:                       2.163   Cond. No.                         172.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [ ]:

```

```

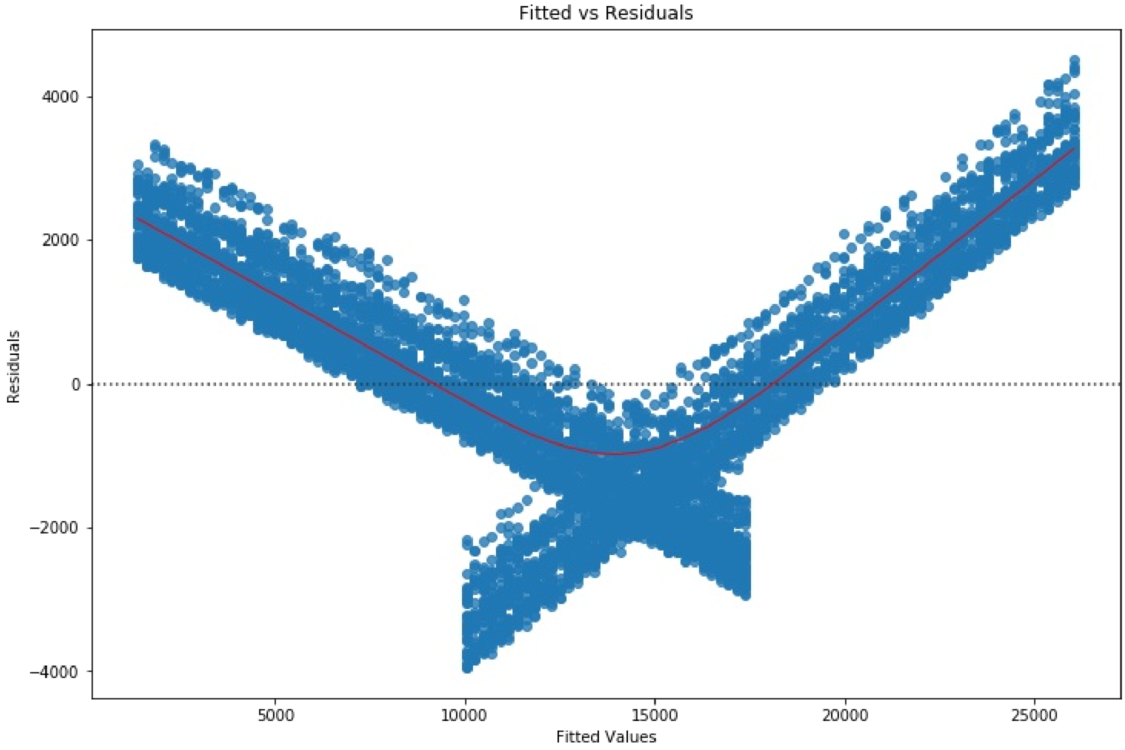
Additional_charges = -2681.5619+Age(225.6544)+HighBlood_Yes(8647.7532)

Fitted vs Residuals of model

Code Reference (seaborn.residplot, n.d.)

In [80]:

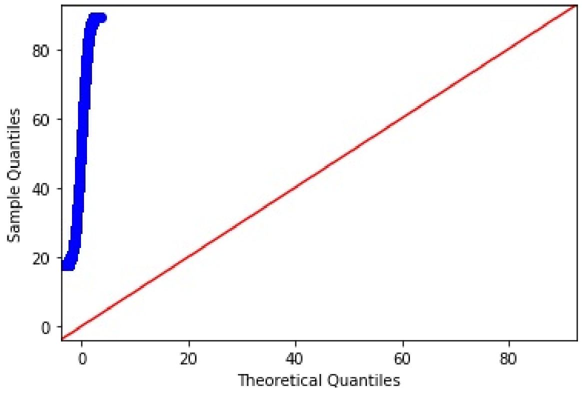
```
plt.figure(figsize=(12,8))
plt.title('Fitted vs Residuals')
sns.residplot(ols.fittedvalues,ols.resid,lowess=True,line_kws={'color':'r','lw':1})
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.show()
```



QQ Plots, Independent Features

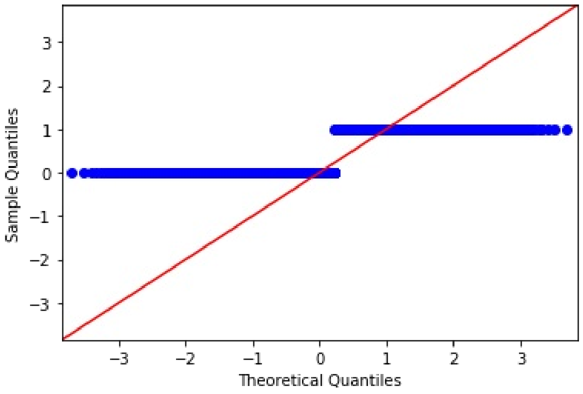
In [81]:

```
fig = sm.qqplot(df['Age'], line='45')
plt.show()
```



In [82]:

```
fig = sm.qqplot(df['HighBlood_Yes'], line='45')
plt.show()
```



In [ ]:

In [ ]:

In [ ]: