WGU D208

January 10, 2024

[1]:
```
%%html
<style>
.toc-item > li {
    list-style-type: upper-alpha;
}
</style>
```

```
<IPython.core.display.HTML object>
```

0.1 Kamal Shaham

0.2 D209: Predictive Modeling

0.3 Instructor: Dr. Eric Straw

```
<h2>Table of Contents</h2>
<ul class="toc-item">
    <li><a href="#question">Research Question</a>
        <ul>
```

Goals of data analysis

```
        </ul>
    </li>
    <li><a href="#assumptions">Method Justification</a>
        <ul>
            <li><a href="#assumptions">Four assumptions of linear regression model</a></li>
            <li><a href="#benefits">Two benefits of using python</a></li>
            <li><a href="#justification2">Justification for using multiple linear regression</
        </ul>
    </li>
    <li><a href="#cleaning-goals">Data Preparation</a>
    <ul>
            <li><a href="#cleaning-goals">Data cleaning goals</a></li>
            <li><a href="#variables">Variable statistical summaries</a></li>
            <li><a href="#visualizations">Univariate/Bivariate visualizations</a></li>
            <li><a href="#transform-goals">Data transformation goals</a></li>
            <li><a href="#csv">Prepared data CSV</a></li>
        </ul>
    </li>
    <li><a href="#initial-model">Model Comparison and Analysis</a>
    <ul>
            <li><a href="#initial-model">Initial multiple linear regression model</a></li>
            <li><a href="#evaluation">Model evaluation metric</a></li>
            <li><a href="#reduced-model">Reduced linear regression model</a></li>
        </ul>
    </li>
    <li><a href="#compare-models">Reduced Model Analysis</a>
    <ul>
            <li><a href="#compare-models">Compare initial and reduced models</a></li>
            <li><a href="#outputs">Output and calculations of analysis</a></li>
            <li><a href="#linear-code">Linear regression code</a></li>
        </ul>
    </li>
    <li><a href="#results">Data Summary and Implications</a>
    <ul>
            <li><a href="#results">Results of data analysis</a></li>
            <li><a href="#action">Reccomended course of action</a></li>
        </ul>
    </li>
    <li><a href="#video">Panopto video</a></li>
    <li><a href="#thirdparty">Third-party code references</a></li>
    <li><a href="#references">References</a></li>
</ul>
```

0.4 A. Research Question

According to a study by Baek et al. (2018), shorter hospital stays have been linked to lower risks of opportunistic infections and side effects from medications. They also reduce the financial burden of hospital fees and aid in increasing bed turnover rates, thereby enhancing hospital profits. Hospitals continuously grapple with the issue of readmissions, and any factors contributing to their reduction are of significant value. As defined in the medical data dictionary (D208 Datasets), a patient readmitted to the hospital within a month of discharge is categorized as a readmission. This leads to an important research question: Is there a correlation between patient observations and the initial length of their hospital stay? Utilizing patient medical data, we aim to determine if certain variables might impact the duration of a patient's initial hospital stay.

0.4.1 A2. Goals of data analysis

The goal of this data analysis is to determine if any variables influence a patient's hospital readmission rate. Through the application of analytical models like multiple linear regression, we aim to identify the best fit. As emphasized in Dr. Middleton's webinars, the dependent variable in this context will be continuous. By examining factors that affect the duration of a patient's initial hospital stay, we can provide hospitals with valuable data that can be instrumental to enhancing treatments and potentially reducing readmission rates.

0.5 B. Method Justification

Assumptions are made when using a multiple linear regression model to determine if there is a good fit. According to (Zach, 2021) these include: - Linear relationships - there needs to be a linear relationship between the independent variable x, and the dependent variable, y. Linear relationships can be found by using visualizations (scatterplots) or statistical tests. - Error distribution - errors need to be normally distributed. To check if the residuals of a model follow a normal distribution a histogram or Q-Q plot can be used to verify normality. - No multicollinearity - two or more of the predictors do not correlate strongly with each other. This can be checked via generating a matrix of the tolerances and variance inflation factor of each independent variable. - Homoscedasticity - the variance of the residuals needs to be the same for all values of x, this case being the independent variable. Can be checked with a scatterplot of residuals vs predicted values.

0.5.1 B2. Two benefits of using Python

Python will be used to perform this data analysis. Python has several statistical packages such as Matplotlib, SciPy, and Statsmodels. Tools in Python allow for intuitive visualizations of statistical observations. For example, multiple linear regression can be utilized by importing the sklearn and statsmodels packages. The Seaborn package is used for visualizations throughout this analysis.

0.5.2 B3. Justification for using multiple linear regression

Multiple linear regression, in the context of our research question, will involve analyzing each factor in the dataset to determine if any variables correlate with the 'Initial_days' variable. Several factors within the dataset are continuous variables, which may potentially influence 'Initial_days'. By utilizing this data, hospitals can offer improved treatment options and potentially reduce readmissions.

## C. Data Preparation The data will need to be prepared prior to running the data analysis model. Dealing with missing or null values will need to be addressed. Missing values can potentially be filled with zeros or populated with the average of the respective column. Duplicated data will also need to be removed from the dataset. Outliers will need to be identified and potentially addressed. In order to run linear regression on categorical variables, they will need to be converted to numerical values. When dealing with categorical variables that have more than two levels and can't be sorted ordinally, one-hot encoding will need to be utilized. In addition to one-hot encoding the variables, we'll need to drop one variable when adding them to our regression model to mitigate multicollinearity. Patient location/job demographics (state, city, job, area, etc.) will not provide any benefit to our analysis and thus can be removed.

[12]:

```python
%matplotlib inline
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from scipy import stats
data=pd.read_csv('medical_clean.csv')
print(data.head())
print(data.columns)
```

```
#check for missing/null values
print(data.isnull().sum())

#check for duplicate values of any rows
print(data.duplicated().any())

# Check for duplicate values based on customer_id unique key
print(data.duplicated('Customer_id').any())

# remove unused columns
data.drop(['CaseOrder', 'Customer_id','Interaction', 'UID', 'City', 'Marital',␣
 ↪'State', 'County', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'TimeZone',␣
 ↪'Job', 'Gender'], axis=1, inplace=True)
print(data.head())
print(data.columns)
print(data.info())

# rename unclear survey response columns
survey_col_names = {
    'Item1': 'Timely_admis',
    'Item2': 'Timely_treat',
    'Item3': 'Timely_visits',
    'Item4': 'Reliability',
    'Item5': 'Options',
    'Item6': 'Hours_treat',
    'Item7': 'Courteous_staff',
    'Item8': 'Active_listening'
}

data = data.rename(columns=survey_col_names)

categorical_cols = ['ReAdmis', 'Complication_risk', 'Initial_admin', 'Services',
                    'Overweight', 'Anxiety', 'Arthritis', 'Asthma',␣
 ↪'Soft_drink',
                    'Diabetes', 'Allergic_rhinitis', 'BackPain', 'Stroke',␣
 ↪'HighBlood',
                    'Hyperlipidemia', 'Reflux_esophagitis']

# Generate dummy variables
data = pd.get_dummies(data, columns=categorical_cols, drop_first=True)

print(data.info())

print(data.head())

# Set the option to display all the columns
pd.set_option('display.max_columns', None)
```

```python
# Set the option to display all the rows
pd.set_option('display.max_rows', None)
print(data.describe(include='all'))
print(data.columns)
```

```
   CaseOrder Customer_id                           Interaction  \
0          1     C412403  8cd49b13-f45a-4b47-a2bd-173ffa932c2f   
1          2     Z919181  d2450b70-0337-4406-bdbb-bc1037f1734c   
2          3     F995323  a2057123-abf5-4a2c-abad-8ffe33512562   
3          4     A879973  1dec528d-eb34-4079-adce-0d7a40e82205   
4          5     C544523  5885f56b-d6da-43a3-8760-83583af94266   

                                UID          City State        County    Zip  \
0  3a83ddb66e2ae73798bdf1d705dc0932           Eva    AL        Morgan  35621   
1  176354c5eef714957d486009feabf195      Marianna    FL       Jackson  32446   
2  e19a0fa00aeda885b8a436757e889bc9   Sioux Falls    SD     Minnehaha  57110   
3  cd17d7b6d152cb6f23957346d11c3f07  New Richland    MN        Waseca  56072   
4  d2f0425877b10ed6bb381f3e2579424a    West Point    VA  King William  23181   

        Lat       Lng  Population      Area          TimeZone  \
0  34.34960 -86.72508        2951  Suburban   America/Chicago   
1  30.84513 -85.22907       11303     Urban   America/Chicago   
2  43.54321 -96.63772       17125  Suburban   America/Chicago   
3  43.89744 -93.51479        2162  Suburban   America/Chicago   
4  37.59894 -76.88958        5287     Rural  America/New_York   

                                Job  Children  Age    Income   Marital  \
0  Psychologist, sport and exercise         1   53  86575.93  Divorced   
1      Community development worker         3   51  46805.99   Married   
2           Chief Executive Officer         3   53  14370.14   Widowed   
3               Early years teacher         0   78  39741.49   Married   
4       Health promotion specialist         1   22   1209.56   Widowed   

   Gender ReAdmis  VitD_levels  Doc_visits  Full_meals_eaten  vitD_supp  \
0    Male      No    19.141466           6                 0          0   
1  Female      No    18.940352           4                 2          1   
2  Female      No    18.057507           4                 1          0   
3    Male      No    16.576858           4                 1          0   
4  Female      No    17.439069           5                 0          2   

  Soft_drink        Initial_admin HighBlood Stroke Complication_risk  \
0         No  Emergency Admission       Yes     No            Medium   
1         No  Emergency Admission       Yes     No              High   
2         No   Elective Admission       Yes     No            Medium   
3         No   Elective Admission        No    Yes            Medium   
4        Yes   Elective Admission        No     No               Low   
```

```
  Overweight Arthritis Diabetes Hyperlipidemia BackPain Anxiety  \
0         No       Yes      Yes             No      Yes     Yes
1        Yes        No       No             No       No      No
2        Yes        No      Yes             No       No      No
3         No       Yes       No             No       No      No
4         No        No       No            Yes       No      No

  Allergic_rhinitis Reflux_esophagitis Asthma     Services  Initial_days  \
0               Yes                 No    Yes   Blood Work     10.585770
1                No                Yes     No  Intravenous     15.129562
2                No                 No     No   Blood Work      4.772177
3                No                Yes    Yes   Blood Work      1.714879
4               Yes                 No     No      CT Scan      1.254807

   TotalCharge  Additional_charges  Item1  Item2  Item3  Item4  Item5  Item6  \
0  3726.702860        17939.403420      3      3      2      2      4      3
1  4193.190458        17612.998120      3      4      3      4      4      4
2  2434.234222        17505.192460      2      4      4      4      3      4
3  2127.830423        12993.437350      3      5      5      3      4      5
4  2113.073274         3716.525786      2      1      3      3      5      3

   Item7  Item8
0      3      4
1      3      3
2      3      3
3      5      5
4      4      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')
CaseOrder             0
Customer_id           0
Interaction           0
UID                   0
City                  0
State                 0
County                0
Zip                   0
Lat                   0
Lng                   0
```

```
Population            0
Area                  0
TimeZone              0
Job                   0
Children              0
Age                   0
Income                0
Marital               0
Gender                0
ReAdmis               0
VitD_levels           0
Doc_visits            0
Full_meals_eaten      0
vitD_supp             0
Soft_drink            0
Initial_admin         0
HighBlood             0
Stroke                0
Complication_risk     0
Overweight            0
Arthritis             0
Diabetes              0
Hyperlipidemia        0
BackPain              0
Anxiety               0
Allergic_rhinitis     0
Reflux_esophagitis    0
Asthma                0
Services              0
Initial_days          0
TotalCharge           0
Additional_charges    0
Item1                 0
Item2                 0
Item3                 0
Item4                 0
Item5                 0
Item6                 0
Item7                 0
Item8                 0
dtype: int64
False
False
   Children  Age    Income ReAdmis  VitD_levels  Doc_visits  Full_meals_eaten  \
0         1   53  86575.93      No    19.141466           6                 0
1         3   51  46805.99      No    18.940352           4                 2
2         3   53  14370.14      No    18.057507           4                 1
3         0   78  39741.49      No    16.576858           4                 1
```

```
4         1   22   1209.56      No    17.439069          5                 0

   vitD_supp Soft_drink        Initial_admin HighBlood Stroke  \
0          0         No  Emergency Admission       Yes     No
1          1         No  Emergency Admission       Yes     No
2          0         No   Elective Admission       Yes     No
3          0         No   Elective Admission        No    Yes
4          2        Yes   Elective Admission        No     No

  Complication_risk Overweight Arthritis Diabetes Hyperlipidemia BackPain  \
0            Medium         No       Yes      Yes             No      Yes
1              High        Yes        No       No             No       No
2            Medium        Yes        No      Yes             No       No
3            Medium         No       Yes       No             No       No
4               Low         No        No       No            Yes       No

  Anxiety Allergic_rhinitis Reflux_esophagitis Asthma     Services  \
0     Yes               Yes                 No    Yes   Blood Work
1      No                No                Yes     No  Intravenous
2      No                No                 No     No   Blood Work
3      No                No                Yes    Yes   Blood Work
4      No               Yes                 No     No      CT Scan

   Initial_days  TotalCharge  Additional_charges  Item1  Item2  Item3  Item4  \
0     10.585770  3726.702860        17939.403420      3      3      2      2
1     15.129562  4193.190458        17612.998120      3      4      3      4
2      4.772177  2434.234222        17505.192460      2      4      4      4
3      1.714879  2127.830423        12993.437350      3      5      5      3
4      1.254807  2113.073274         3716.525786      2      1      3      3

   Item5  Item6  Item7  Item8
0      4      3      3      4
1      4      4      3      3
2      3      4      3      3
3      4      5      5      5
4      5      3      4      3
Index(['Children', 'Age', 'Income', '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')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 34 columns):
 #   Column              Non-Null Count  Dtype
```

```
---  ------              --------------  -----
 0   Children            10000 non-null  int64
 1   Age                 10000 non-null  int64
 2   Income              10000 non-null  float64
 3   ReAdmis             10000 non-null  object
 4   VitD_levels         10000 non-null  float64
 5   Doc_visits          10000 non-null  int64
 6   Full_meals_eaten    10000 non-null  int64
 7   vitD_supp           10000 non-null  int64
 8   Soft_drink          10000 non-null  object
 9   Initial_admin       10000 non-null  object
 10  HighBlood           10000 non-null  object
 11  Stroke              10000 non-null  object
 12  Complication_risk   10000 non-null  object
 13  Overweight          10000 non-null  object
 14  Arthritis           10000 non-null  object
 15  Diabetes            10000 non-null  object
 16  Hyperlipidemia      10000 non-null  object
 17  BackPain            10000 non-null  object
 18  Anxiety             10000 non-null  object
 19  Allergic_rhinitis   10000 non-null  object
 20  Reflux_esophagitis  10000 non-null  object
 21  Asthma              10000 non-null  object
 22  Services            10000 non-null  object
 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
dtypes: float64(5), int64(13), object(16)
memory usage: 2.6+ MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 38 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Children                              10000 non-null  int64
 1   Age                                   10000 non-null  int64
 2   Income                                10000 non-null  float64
 3   VitD_levels                           10000 non-null  float64
 4   Doc_visits                            10000 non-null  int64
 5   Full_meals_eaten                     10000 non-null  int64
 6   vitD_supp                            10000 non-null  int64
 7   Initial_days                         10000 non-null  float64
 8   TotalCharge                          10000 non-null  float64
 9   Additional_charges                   10000 non-null  float64
 10  Timely_admis                         10000 non-null  int64
 11  Timely_treat                         10000 non-null  int64
 12  Timely_visits                        10000 non-null  int64
 13  Reliability                          10000 non-null  int64
 14  Options                              10000 non-null  int64
 15  Hours_treat                          10000 non-null  int64
 16  Courteous_staff                      10000 non-null  int64
 17  Active_listening                     10000 non-null  int64
 18  ReAdmis_Yes                          10000 non-null  uint8
 19  Complication_risk_Low                10000 non-null  uint8
 20  Complication_risk_Medium             10000 non-null  uint8
 21  Initial_admin_Emergency Admission    10000 non-null  uint8
 22  Initial_admin_Observation Admission  10000 non-null  uint8
 23  Services_CT Scan                     10000 non-null  uint8
 24  Services_Intravenous                 10000 non-null  uint8
 25  Services_MRI                         10000 non-null  uint8
 26  Overweight_Yes                       10000 non-null  uint8
 27  Anxiety_Yes                          10000 non-null  uint8
 28  Arthritis_Yes                        10000 non-null  uint8
 29  Asthma_Yes                           10000 non-null  uint8
 30  Soft_drink_Yes                       10000 non-null  uint8
 31  Diabetes_Yes                         10000 non-null  uint8
 32  Allergic_rhinitis_Yes                10000 non-null  uint8
 33  BackPain_Yes                         10000 non-null  uint8
 34  Stroke_Yes                           10000 non-null  uint8
 35  HighBlood_Yes                        10000 non-null  uint8
 36  Hyperlipidemia_Yes                   10000 non-null  uint8
 37  Reflux_esophagitis_Yes               10000 non-null  uint8
dtypes: float64(5), int64(13), uint8(20)
memory usage: 1.6 MB
None
   Children  Age    Income  VitD_levels  Doc_visits  Full_meals_eaten  \
0         1   53  86575.93    19.141466           6                 0
1         3   51  46805.99    18.940352           4                 2
2         3   53  14370.14    18.057507           4                 1
3         0   78  39741.49    16.576858           4                 1
4         1   22   1209.56    17.439069           5                 0

   vitD_supp  Initial_days  TotalCharge  Additional_charges  Timely_admis  \
0          0     10.585770  3726.702860        17939.403420             3
1          1     15.129562  4193.190458        17612.998120             3
2          0      4.772177  2434.234222        17505.192460             2
3          0      1.714879  2127.830423        12993.437350             3
```

```
4          2      1.254807  2113.073274         3716.525786             2

   Timely_treat  Timely_visits  Reliability  Options  Hours_treat  \
0             3              2            2        4            3
1             4              3            4        4            4
2             4              4            4        3            4
3             5              5            3        4            5
4             1              3            3        5            3

   Courteous_staff  Active_listening  ReAdmis_Yes  Complication_risk_Low  \
0                3                 4            0                      0
1                3                 3            0                      0
2                3                 3            0                      0
3                5                 5            0                      0
4                4                 3            0                      1

   Complication_risk_Medium  Initial_admin_Emergency Admission  \
0                         1                                  1
1                         0                                  1
2                         1                                  0
3                         1                                  0
4                         0                                  0

   Initial_admin_Observation Admission  Services_CT Scan  \
0                                    0                 0
1                                    0                 0
2                                    0                 0
3                                    0                 0
4                                    0                 1

   Services_Intravenous  Services_MRI  Overweight_Yes  Anxiety_Yes  \
0                     0             0               0            1
1                     1             0               1            0
2                     0             0               1            0
3                     0             0               0            0
4                     0             0               0            0

   Arthritis_Yes  Asthma_Yes  Soft_drink_Yes  Diabetes_Yes  \
0              1           1               0             1
1              0           0               0             0
2              0           0               0             1
3              1           1               0             0
4              0           0               1             0

   Allergic_rhinitis_Yes  BackPain_Yes  Stroke_Yes  HighBlood_Yes  \
0                      1             1           0              1
1                      0             0           0              1
2                      0             0           0              1
```

```
3                      0             0           1              0
4                      1             0           0              0

   Hyperlipidemia_Yes  Reflux_esophagitis_Yes
0                   0                       0
1                   0                       1
2                   0                       0
3                   0                       1
4                   1                       0
           Children           Age         Income   VitD_levels    Doc_visits  \
count  10000.000000  10000.000000   10000.000000  10000.000000  10000.000000
mean       2.097200     53.511700   40490.495160     17.964262      5.012200
std        2.163659     20.638538   28521.153293      2.017231      1.045734
min        0.000000     18.000000     154.080000      9.806483      1.000000
25%        0.000000     36.000000   19598.775000     16.626439      4.000000
50%        1.000000     53.000000   33768.420000     17.951122      5.000000
75%        3.000000     71.000000   54296.402500     19.347963      6.000000
max       10.000000     89.000000  207249.100000     26.394449      9.000000

       Full_meals_eaten     vitD_supp  Initial_days   TotalCharge  \
count      10000.000000  10000.000000  10000.000000  10000.000000
mean           1.001400      0.398900     34.455299   5312.172769
std            1.008117      0.628505     26.309341   2180.393838
min            0.000000      0.000000      1.001981   1938.312067
25%            0.000000      0.000000      7.896215   3179.374015
50%            1.000000      0.000000     35.836244   5213.952000
75%            2.000000      1.000000     61.161020   7459.699750
max            7.000000      5.000000     71.981490   9180.728000

       Additional_charges  Timely_admis  Timely_treat  Timely_visits  \
count        10000.000000  10000.000000  10000.000000   10000.000000
mean         12934.528587      3.518800      3.506700       3.511100
std           6542.601544      1.031966      1.034825       1.032755
min           3125.703000      1.000000      1.000000       1.000000
25%           7986.487755      3.000000      3.000000       3.000000
50%          11573.977735      4.000000      3.000000       4.000000
75%          15626.490000      4.000000      4.000000       4.000000
max          30566.070000      8.000000      7.000000       8.000000

        Reliability       Options   Hours_treat  Courteous_staff  \
count  10000.000000  10000.000000  10000.000000     10000.000000
mean       3.515100      3.496900      3.522500         3.494000
std        1.036282      1.030192      1.032376         1.021405
min        1.000000      1.000000      1.000000         1.000000
25%        3.000000      3.000000      3.000000         3.000000
50%        4.000000      3.000000      4.000000         3.000000
75%        4.000000      4.000000      4.000000         4.000000
max        7.000000      7.000000      7.000000         7.000000
```

```
       Active_listening   ReAdmis_Yes  Complication_risk_Low  \
count      10000.000000  10000.000000           10000.000000
mean           3.509700      0.366900               0.212500
std            1.042312      0.481983               0.409097
min            1.000000      0.000000               0.000000
25%            3.000000      0.000000               0.000000
50%            3.000000      0.000000               0.000000
75%            4.000000      1.000000               0.000000
max            7.000000      1.000000               1.000000

       Complication_risk_Medium  Initial_admin_Emergency Admission  \
count              10000.000000                       10000.000000
mean                   0.451700                           0.506000
std                    0.497687                           0.499989
min                    0.000000                           0.000000
25%                    0.000000                           0.000000
50%                    0.000000                           1.000000
75%                    1.000000                           1.000000
max                    1.000000                           1.000000

       Initial_admin_Observation Admission  Services_CT Scan  \
count                         10000.000000      10000.000000
mean                              0.243600          0.122500
std                               0.429276          0.327879
min                               0.000000          0.000000
25%                               0.000000          0.000000
50%                               0.000000          0.000000
75%                               0.000000          0.000000
max                               1.000000          1.000000

       Services_Intravenous  Services_MRI  Overweight_Yes   Anxiety_Yes  \
count          10000.000000  10000.000000    10000.000000  10000.000000
mean               0.313000      0.038000        0.709400      0.321500
std                0.463738      0.191206        0.454062      0.467076
min                0.000000      0.000000        0.000000      0.000000
25%                0.000000      0.000000        0.000000      0.000000
50%                0.000000      0.000000        1.000000      0.000000
75%                1.000000      0.000000        1.000000      1.000000
max                1.000000      1.000000        1.000000      1.000000

       Arthritis_Yes   Asthma_Yes  Soft_drink_Yes  Diabetes_Yes  \
count   10000.000000  10000.00000    10000.000000   10000.00000
mean        0.357400      0.28930        0.257500       0.27380
std         0.479258      0.45346        0.437279       0.44593
min         0.000000      0.00000        0.000000       0.00000
25%         0.000000      0.00000        0.000000       0.00000
50%         0.000000      0.00000        0.000000       0.00000
```

```
75%         1.000000      1.00000        1.000000       1.00000
max         1.000000      1.00000        1.000000       1.00000

       Allergic_rhinitis_Yes  BackPain_Yes    Stroke_Yes  HighBlood_Yes  \
count           10000.000000  10000.000000  10000.000000   10000.000000
mean                0.394100      0.411400      0.199300       0.409000
std                 0.488681      0.492112      0.399494       0.491674
min                 0.000000      0.000000      0.000000       0.000000
25%                 0.000000      0.000000      0.000000       0.000000
50%                 0.000000      0.000000      0.000000       0.000000
75%                 1.000000      1.000000      0.000000       1.000000
max                 1.000000      1.000000      1.000000       1.000000

       Hyperlipidemia_Yes  Reflux_esophagitis_Yes
count        10000.000000            10000.000000
mean             0.337200                0.413500
std              0.472777                0.492486
min              0.000000                0.000000
25%              0.000000                0.000000
50%              0.000000                0.000000
75%              1.000000                1.000000
max              1.000000                1.000000
Index(['Children', 'Age', 'Income', 'VitD_levels', 'Doc_visits',
       'Full_meals_eaten', 'vitD_supp', 'Initial_days', 'TotalCharge',
       'Additional_charges', 'Timely_admis', 'Timely_treat', 'Timely_visits',
       'Reliability', 'Options', 'Hours_treat', 'Courteous_staff',
       'Active_listening', 'ReAdmis_Yes', 'Complication_risk_Low',
       'Complication_risk_Medium', 'Initial_admin_Emergency Admission',
       'Initial_admin_Observation Admission', 'Services_CT Scan',
       'Services_Intravenous', 'Services_MRI', 'Overweight_Yes', 'Anxiety_Yes',
       'Arthritis_Yes', 'Asthma_Yes', 'Soft_drink_Yes', 'Diabetes_Yes',
       'Allergic_rhinitis_Yes', 'BackPain_Yes', 'Stroke_Yes', 'HighBlood_Yes',
       'Hyperlipidemia_Yes', 'Reflux_esophagitis_Yes'],
      dtype='object')
```

### C2. Variable Statistical Summaries In order to use multiple linear regression to answer this research question, summary statistics will need to be generated for every variable used. Geographic variables of the patient, such as population, city, and state, will not provide benefit to our analysis, so they will not be included. There is potential for different modeling to be used on these columns for further analysis. The tables below contain the independent variables (with 'Initial_days' being our dependent variable), their data types, their categorical/continuous classification, and sample data from each column. A summary statistics table is generated, detailing each variable's standard deviation, interquartile ranges, mean, and median (noted as 50% value in the output).

The categorical variables in the data were converted to numerical types to perform regression analysis. Histograms and box plots were generated for each variable to check for distribution. Based on these histograms, it can be seen that 'Income', 'Children', 'vitD_supp', and 'Full_meals_eaten' are not normally distributed.

- Age: Integer, Example: 53
- ReAdmis: Character (binary categorical), Example: No
- VitD_levels: Numeric, Example: 19.141466
- Doc_visits: Integer, Example: 6
- Full_meals_eaten: Integer, Example: 0
- vitD_supp: Integer, Example: 0
- Soft_drink: Character (binary categorical), Example: No
- Initial_admin: Character (nominal categorical), Example: Emergency Admission
- HighBlood: Character (binary categorical), Example: Yes
- Stroke: Character (binary categorical), Example: No
- Complication_risk: Character (ordinal categorical), Example: Medium
- Overweight: Character (binary categorical), Example: No
- Arthritis: Character (binary categorical), Example: Yes
- Diabetes: Character (binary categorical), Example: Yes
- Hyperlipidemia: Character (binary categorical), Example: No
- BackPain: Character (binary categorical), Example: Yes
- Anxiety: Character (binary categorical), Example: Yes
- Allergic_rhinitis: Character (binary categorical), Example: Yes
- Reflux_esophagitis: Character (binary categorical), Example: No
- Asthma: Character (binary categorical), Example: Yes
- Services: Character (Nominal categorical), Example: Blood Work
- Initial_days: Numeric, Example: 10.585770
- TotalCharge: Numeric, Example: 3726.702860
- Additional_charges: Numeric, Example: 17939.403420
- Item1 (Timely_admis): Integer, Example: 3
- Item2 (Timely_treat): Integer, Example: 3
- Item3 (Timely_visits): Integer, Example: 2
- Item4 (Reliability): Integer, Example: 2
- Item5 (Options): Integer, Example: 4
- Item6 (Hours_treat): Integer, Example: 3
- Item7 (Courteous_staff): Integer, Example: 3
- Item8 (Active_listening): Integer, Example: 4

[108]:

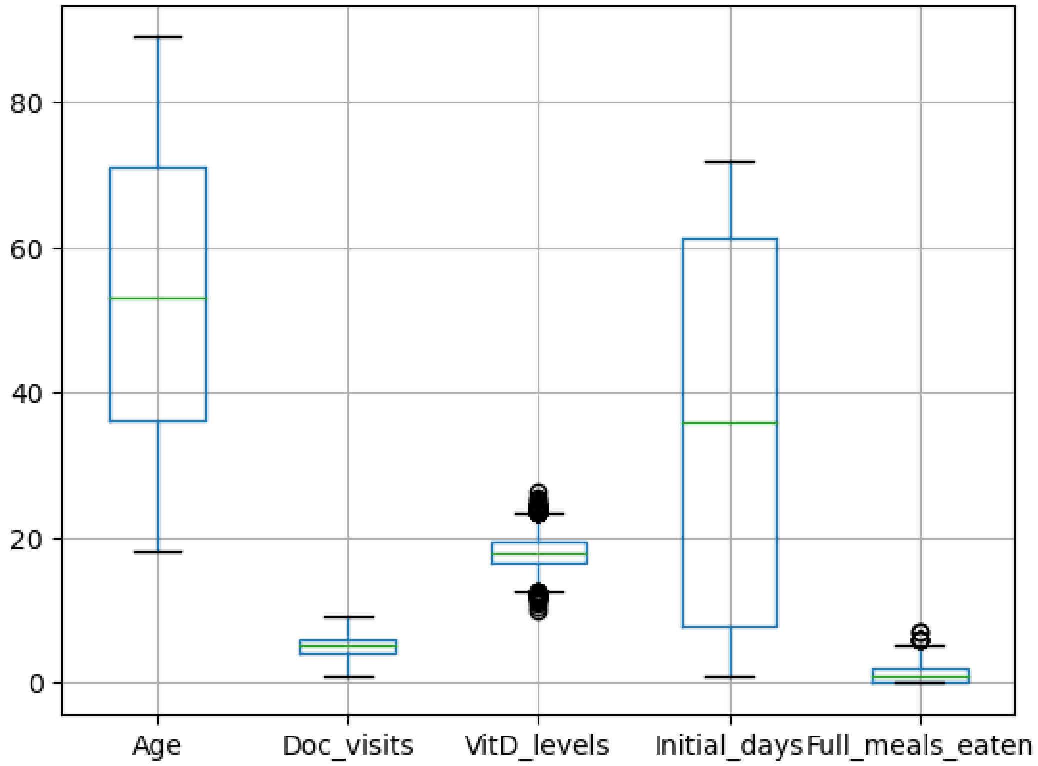
```python
print(data.info())
print(data.head())
print(data.describe(include='all'))
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 36 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Age                             10000 non-null  int64
 1   VitD_levels                     10000 non-null  float64
 2   Doc_visits                      10000 non-null  int64
 3   Full_meals_eaten                10000 non-null  int64
 4   vitD_supp                       10000 non-null  int64
 5   Initial_days                    10000 non-null  float64
 6   TotalCharge                         10000 non-null  float64
 7   Additional_charges                  10000 non-null  float64
 8   Timely_admis                        10000 non-null  int64
 9   Timely_treat                        10000 non-null  int64
 10  Timely_visits                       10000 non-null  int64
 11  Reliability                         10000 non-null  int64
 12  Options                             10000 non-null  int64
 13  Hours_treat                         10000 non-null  int64
 14  Courteous_staff                     10000 non-null  int64
 15  Active_listening                    10000 non-null  int64
 16  ReAdmis_Yes                         10000 non-null  uint8
 17  Complication_risk_Low               10000 non-null  uint8
 18  Complication_risk_Medium            10000 non-null  uint8
 19  Initial_admin_Emergency Admission   10000 non-null  uint8
 20  Initial_admin_Observation Admission  10000 non-null  uint8
 21  Services_CT Scan                    10000 non-null  uint8
 22  Services_Intravenous                10000 non-null  uint8
 23  Services_MRI                        10000 non-null  uint8
 24  Overweight_Yes                      10000 non-null  uint8
 25  Anxiety_Yes                         10000 non-null  uint8
 26  Arthritis_Yes                       10000 non-null  uint8
 27  Asthma_Yes                          10000 non-null  uint8
 28  Soft_drink_Yes                      10000 non-null  uint8
 29  Diabetes_Yes                        10000 non-null  uint8
 30  Allergic_rhinitis_Yes               10000 non-null  uint8
 31  BackPain_Yes                        10000 non-null  uint8
 32  Stroke_Yes                          10000 non-null  uint8
 33  HighBlood_Yes                       10000 non-null  uint8
 34  Hyperlipidemia_Yes                  10000 non-null  uint8
 35  Reflux_esophagitis_Yes              10000 non-null  uint8
dtypes: float64(4), int64(12), uint8(20)
memory usage: 1.4 MB
None
   Age  VitD_levels  Doc_visits  Full_meals_eaten  vitD_supp  Initial_days  \
0   53    19.141466           6                 0          0     10.585770
1   51    18.940352           4                 2          1     15.129562
2   53    18.057507           4                 1          0      4.772177
3   78    16.576858           4                 1          0      1.714879
4   22    17.439069           5                 0          2      1.254807

   TotalCharge  Additional_charges  Timely_admis  Timely_treat  Timely_visits  \
0  3726.702860        17939.403420             3             3              2
1  4193.190458        17612.998120             3             4              3
2  2434.234222        17505.192460             2             4              4
3  2127.830423        12993.437350             3             5              5
4  2113.073274         3716.525786             2             1              3

   Reliability  Options  Hours_treat  Courteous_staff  Active_listening  \
0            2        4            3                3                 4
1            4        4            4                3                 3
2            4        3            4                3                 3
3            3        4            5                5                 5
4            3        5            3                4                 3

   ReAdmis_Yes  Complication_risk_Low  Complication_risk_Medium  \
0            0                      0                         1
1            0                      0                         0
2            0                      0                         1
3            0                      0                         1
4            0                      1                         0

   Initial_admin_Emergency Admission  Initial_admin_Observation Admission  \
0                                  1                                    0
1                                  1                                    0
2                                  0                                    0
3                                  0                                    0
4                                  0                                    0

   Services_CT Scan  Services_Intravenous  Services_MRI  Overweight_Yes  \
0                 0                     0             0               0
1                 0                     1             0               1
2                 0                     0             0               1
3                 0                     0             0               0
4                 1                     0             0               0

   Anxiety_Yes  Arthritis_Yes  Asthma_Yes  Soft_drink_Yes  Diabetes_Yes  \
0            1              1           1               0             1
1            0              0           0               0             0
2            0              0           0               0             1
3            0              1           1               0             0
4            0              0           0               1             0

   Allergic_rhinitis_Yes  BackPain_Yes  Stroke_Yes  HighBlood_Yes  \
0                      1             1           0              1
1                      0             0           0              1
2                      0             0           0              1
3                      0             0           1              0
4                      1             0           0              0

   Hyperlipidemia_Yes  Reflux_esophagitis_Yes
0                   0                       0
1                   0                       1
2                   0                       0
3                   0                       1
4                   1                       0
                Age   VitD_levels    Doc_visits  Full_meals_eaten  \
count  10000.000000  10000.000000  10000.000000       10000.000000
mean      53.511700     17.964262      5.012200           1.001400
std       20.638538      2.017231      1.045734           1.008117
min       18.000000      9.806483      1.000000           0.000000
25%       36.000000     16.626439      4.000000           0.000000
50%       53.000000     17.951122      5.000000           1.000000
75%       71.000000     19.347963      6.000000           2.000000
max       89.000000     26.394449      9.000000           7.000000

          vitD_supp  Initial_days   TotalCharge  Additional_charges  \
count  10000.000000  10000.000000  10000.000000        10000.000000
mean       0.398900     34.455299   5312.172769        12934.528587
std        0.628505     26.309341   2180.393838         6542.601544
min        0.000000      1.001981   1938.312067         3125.703000
25%        0.000000      7.896215   3179.374015         7986.487755
50%        0.000000     35.836244   5213.952000        11573.977735
75%        1.000000     61.161020   7459.699750        15626.490000
max        5.000000     71.981490   9180.728000        30566.070000

        Timely_admis  Timely_treat  Timely_visits   Reliability       Options  \
count  10000.000000  10000.000000   10000.000000  10000.000000  10000.000000
mean       3.518800      3.506700       3.511100      3.515100      3.496900
std        1.031966      1.034825       1.032755      1.036282      1.030192
min        1.000000      1.000000       1.000000      1.000000      1.000000
25%        3.000000      3.000000       3.000000      3.000000      3.000000
50%        4.000000      3.000000       4.000000      4.000000      3.000000
75%        4.000000      4.000000       4.000000      4.000000      4.000000
max        8.000000      7.000000       8.000000      7.000000      7.000000

        Hours_treat  Courteous_staff  Active_listening   ReAdmis_Yes  \
count  10000.000000     10000.000000      10000.000000  10000.000000
mean       3.522500         3.494000          3.509700      0.366900
std        1.032376         1.021405          1.042312      0.481983
min        1.000000         1.000000          1.000000      0.000000
25%        3.000000         3.000000          3.000000      0.000000
50%        4.000000         3.000000          3.000000      0.000000
75%        4.000000         4.000000          4.000000      1.000000
max        7.000000         7.000000          7.000000      1.000000

       Complication_risk_Low  Complication_risk_Medium  \
count           10000.000000              10000.000000
mean                0.212500                  0.451700
std                 0.409097                  0.497687
min                 0.000000                  0.000000
25%                 0.000000                  0.000000
50%                 0.000000                  0.000000
75%                 0.000000                  1.000000
max                 1.000000                  1.000000
```

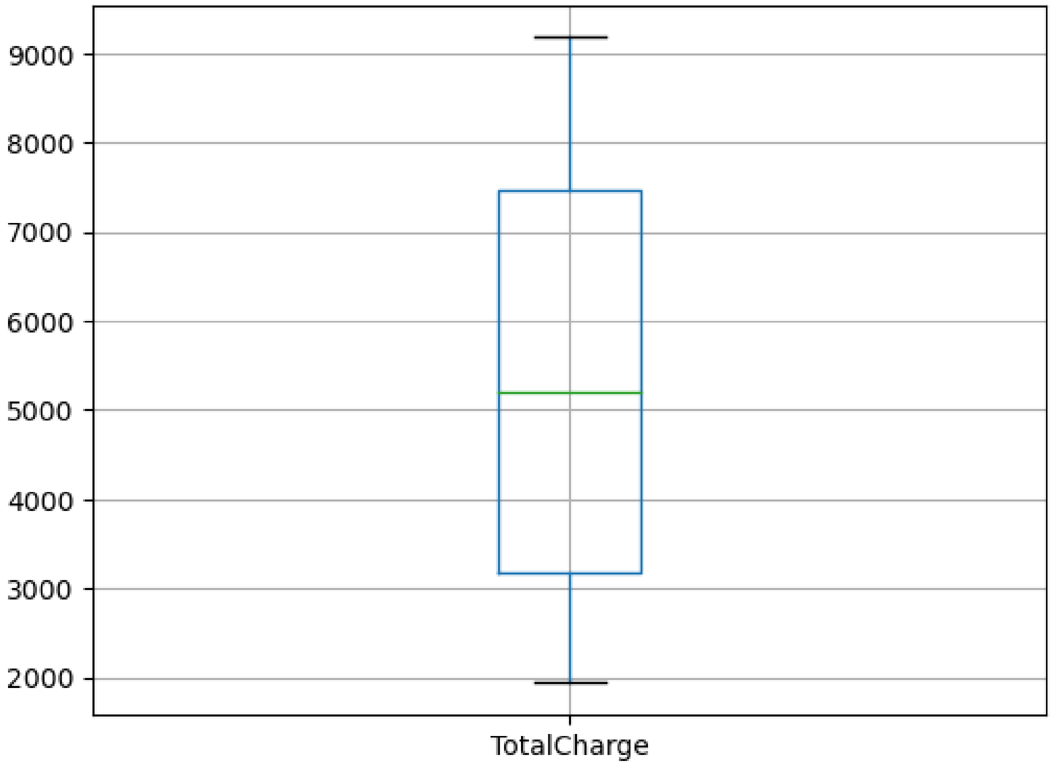
```
       Initial_admin_Emergency Admission  Initial_admin_Observation Admission  \
count                       10000.000000                         10000.000000
mean                            0.506000                             0.243600
std                             0.499989                             0.429276
min                             0.000000                             0.000000
25%                             0.000000                             0.000000
50%                             1.000000                             0.000000
75%                             1.000000                             0.000000
max                             1.000000                             1.000000

       Services_CT Scan  Services_Intravenous  Services_MRI  Overweight_Yes  \
count      10000.000000          10000.000000  10000.000000    10000.000000
mean           0.122500              0.313000      0.038000        0.709400
std            0.327879              0.463738      0.191206        0.454062
min            0.000000              0.000000      0.000000        0.000000
25%            0.000000              0.000000      0.000000        0.000000
50%            0.000000              0.000000      0.000000        1.000000
75%            0.000000              1.000000      0.000000        1.000000
max            1.000000              1.000000      1.000000        1.000000

        Anxiety_Yes  Arthritis_Yes   Asthma_Yes  Soft_drink_Yes  Diabetes_Yes  \
count  10000.000000   10000.000000  10000.00000    10000.000000   10000.00000
mean       0.321500       0.357400      0.28930        0.257500       0.27380
std        0.467076       0.479258      0.45346        0.437279       0.44593
min        0.000000       0.000000      0.00000        0.000000       0.00000
25%        0.000000       0.000000      0.00000        0.000000       0.00000
50%        0.000000       0.000000      0.00000        0.000000       0.00000
75%        1.000000       1.000000      1.00000        1.000000       1.00000
max        1.000000       1.000000      1.00000        1.000000       1.00000

       Allergic_rhinitis_Yes  BackPain_Yes    Stroke_Yes  HighBlood_Yes  \
count           10000.000000  10000.000000  10000.000000   10000.000000
mean                0.394100      0.411400      0.199300       0.409000
std                 0.488681      0.492112      0.399494       0.491674
min                 0.000000      0.000000      0.000000       0.000000
25%                 0.000000      0.000000      0.000000       0.000000
50%                 0.000000      0.000000      0.000000       0.000000
75%                 1.000000      1.000000      0.000000       1.000000
max                 1.000000      1.000000      1.000000       1.000000

       Hyperlipidemia_Yes  Reflux_esophagitis_Yes
count        10000.000000            10000.000000
mean             0.337200                0.413500
std              0.472777                0.492486
min              0.000000                0.000000
25%              0.000000                0.000000
50%              0.000000                0.000000
```

```
75%               1.000000                  1.000000
max               1.000000                  1.000000
```

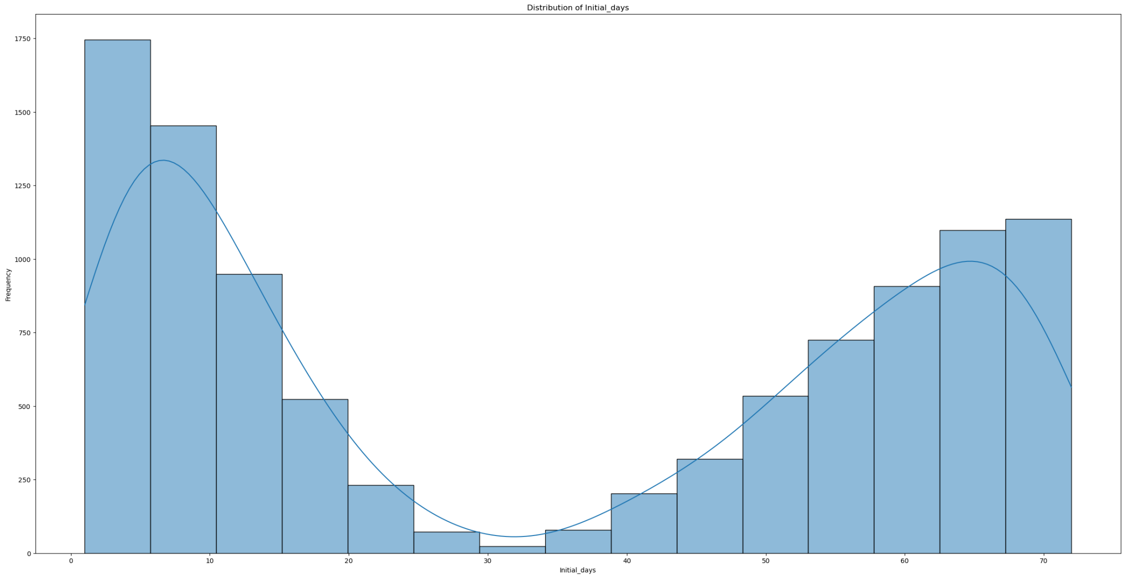
[3]:
```
# check for outliers with smaller group of variables
data.boxplot(column=['Age', 'Doc_visits', 'VitD_levels', 'Initial_days',␣
 ↪'Full_meals_eaten'])
plt.show()
```



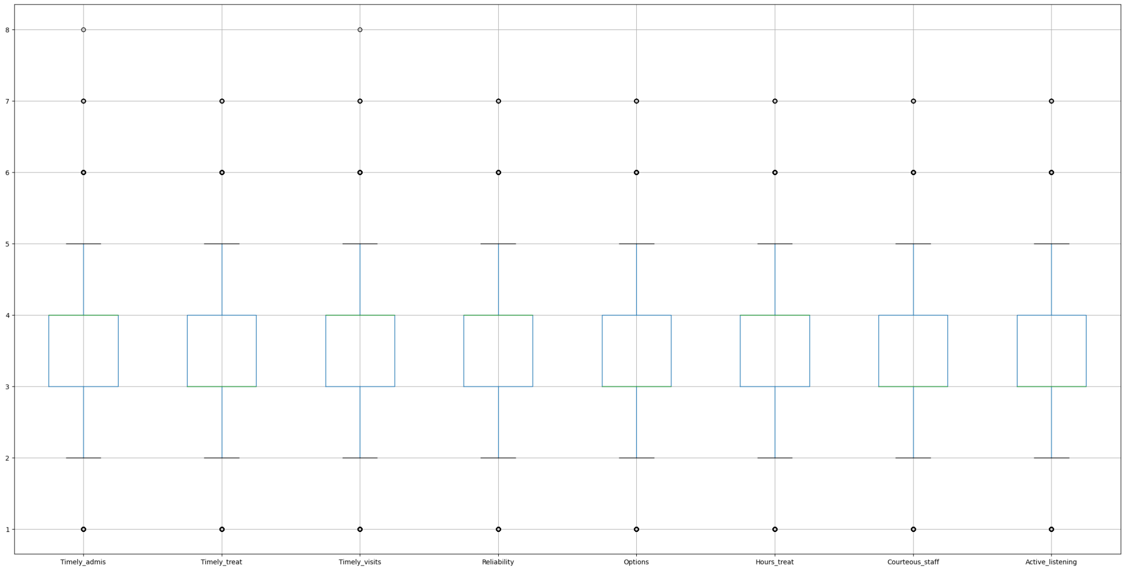
[4]:
```
# check for outliers with smaller group of variables
data.boxplot(column=['TotalCharge'])
plt.show()
```



[5]:
```
plt.subplots(figsize=(30,15))

# Create a histogram of the frequency of initial_days
sns.histplot(data['Initial_days'], kde=True)

plt.title('Distribution of Initial_days')
plt.xlabel('Initial_days')
plt.ylabel('Frequency')
plt.show()
```



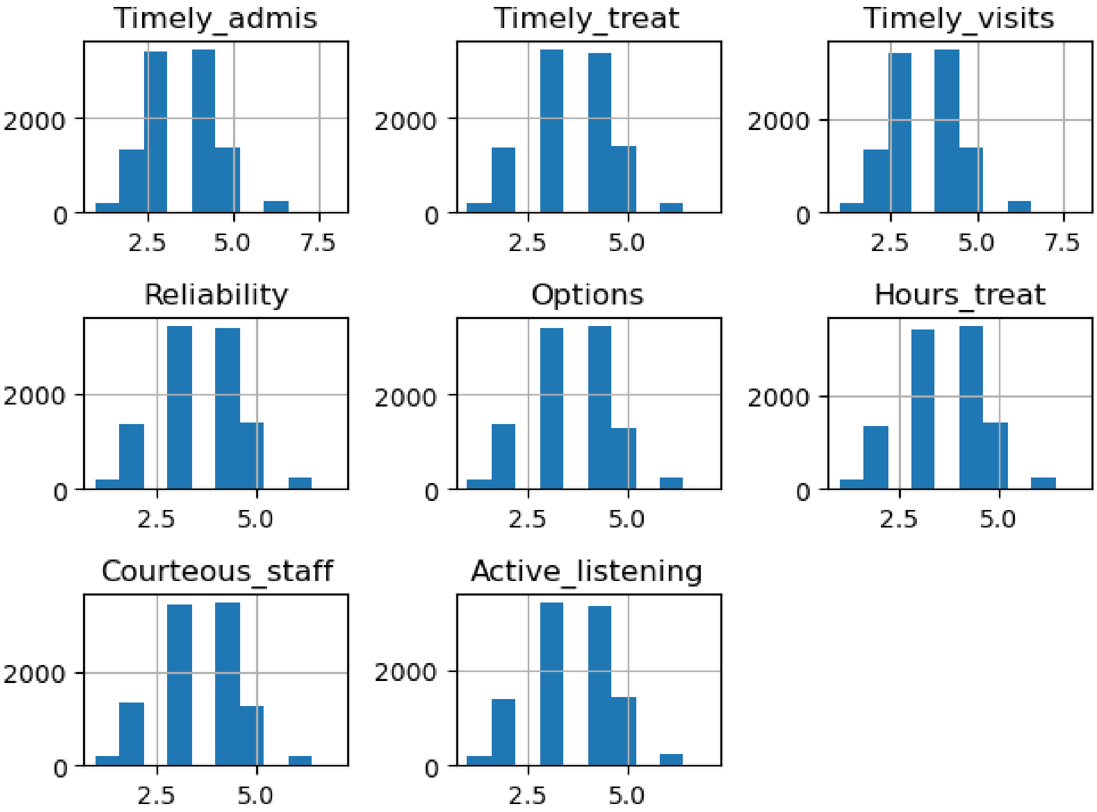
[6]:
```
plt.subplots(figsize=(30,15))

data.boxplot(column=['Timely_admis', 'Timely_treat', 'Timely_visits',␣
  ↪'Reliability', 'Options', 'Hours_treat', 'Courteous_staff',␣
  ↪'Active_listening'])
plt.show()
```



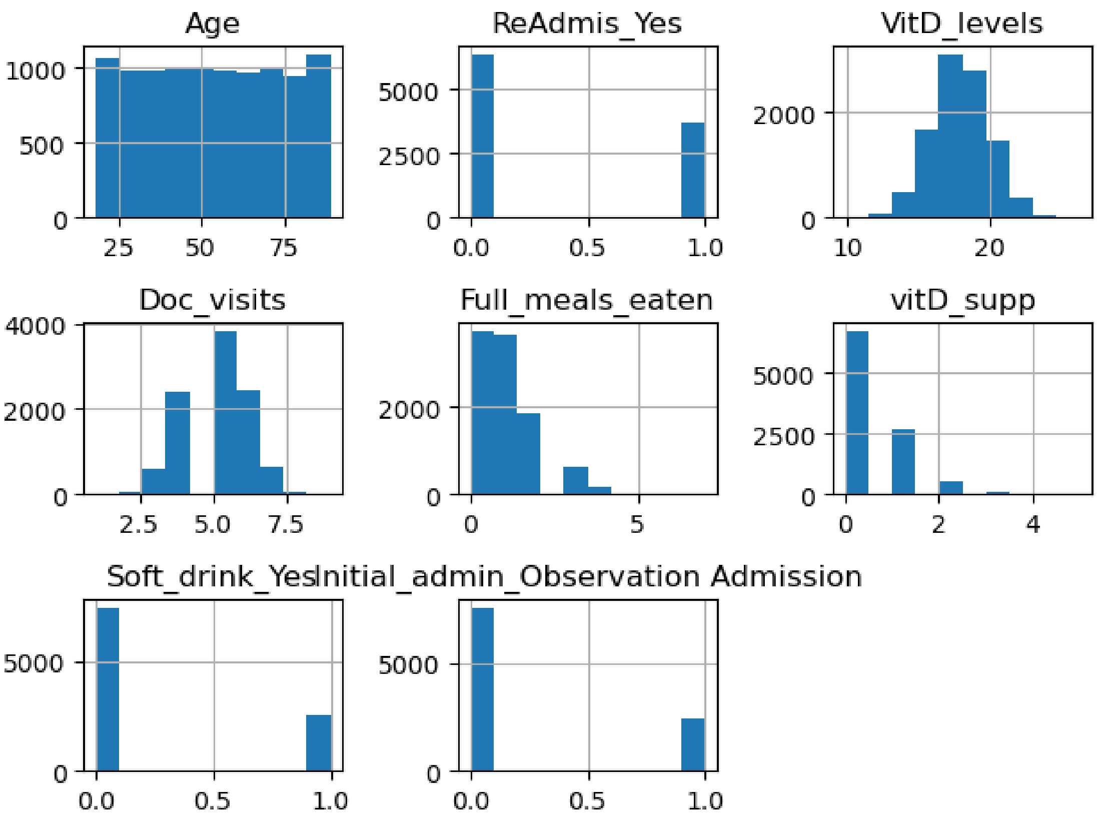
### C3. Univariate/Bivariate Visualizations

Below we generate both univariate and bivariate visualizations for the independent and dependent variables.

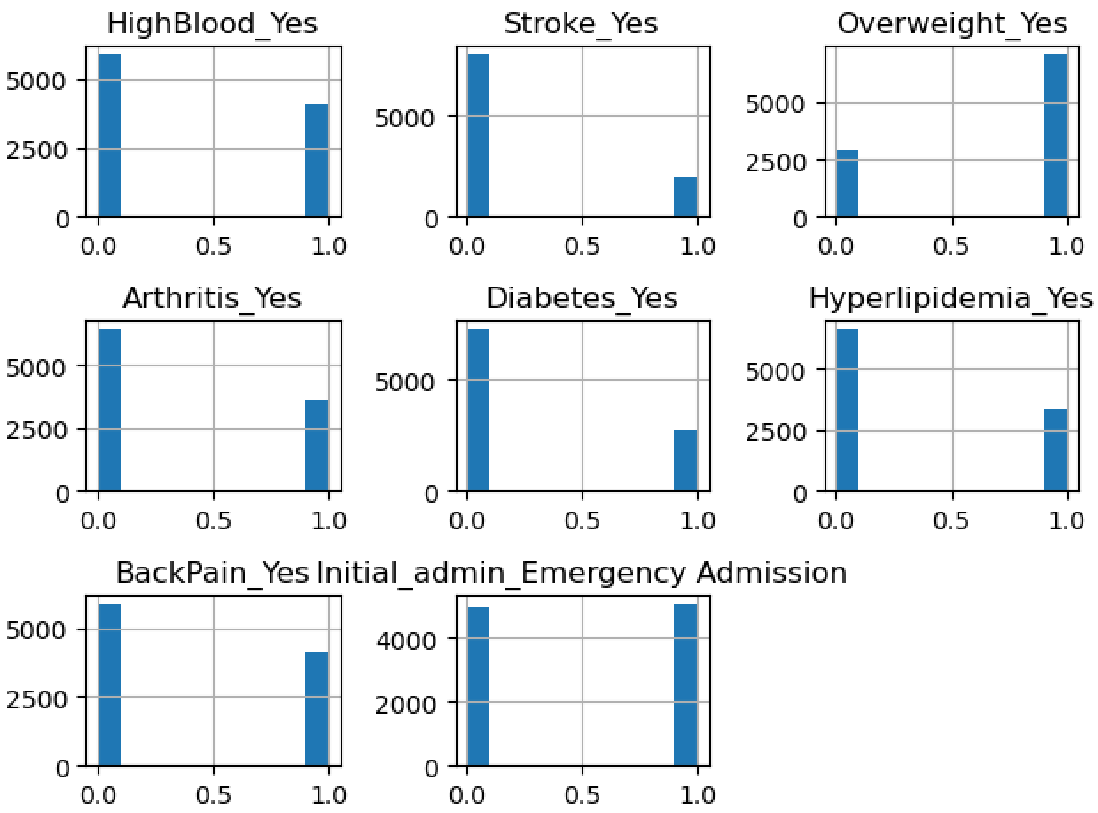
```
[7]: data[['Timely_admis', 'Timely_treat', 'Timely_visits', 'Reliability',␣
      ↪'Options', 'Hours_treat', 'Courteous_staff', 'Active_listening']].hist()
     plt.tight_layout()
     plt.show()
```



```
[52]: data[['Age', 'ReAdmis_Yes', 'VitD_levels',
              'Doc_visits', 'Full_meals_eaten', 'vitD_supp', 'Soft_drink_Yes',␣
       ↪'Initial_admin_Observation Admission']].hist()
      plt.tight_layout()
      plt.show()
```

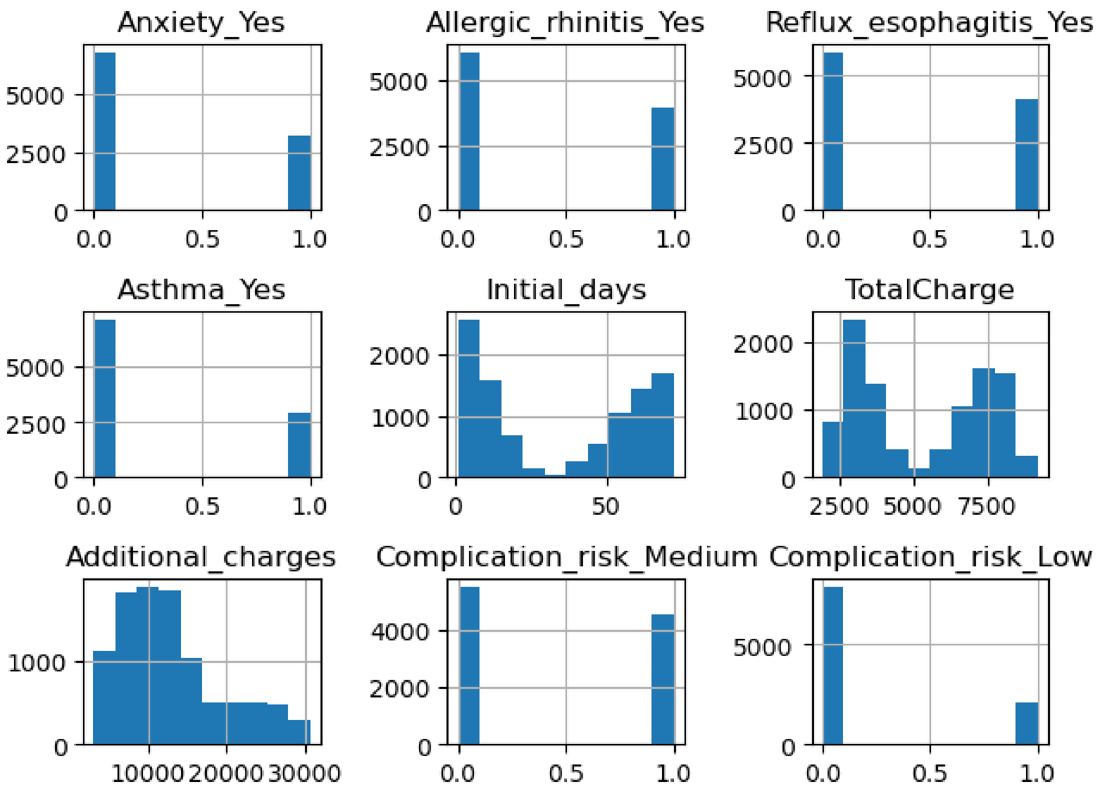


[8]:
```
data[['HighBlood_Yes', 'Stroke_Yes', 'Overweight_Yes', 'Arthritis_Yes',␣
 ↪'Diabetes_Yes',
       'Hyperlipidemia_Yes', 'BackPain_Yes', 'Initial_admin_Emergency␣
 ↪Admission']].hist()
plt.tight_layout()
plt.show()
```



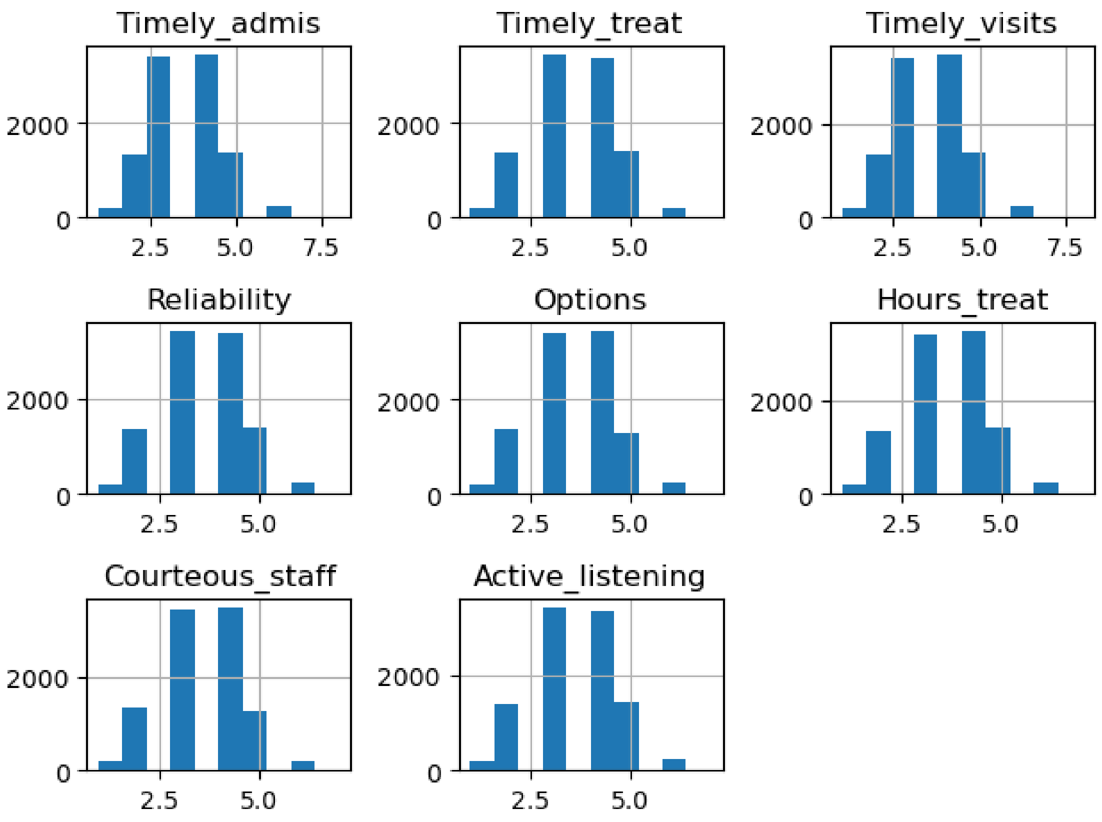
[45]:

```
data[['Anxiety_Yes', 'Allergic_rhinitis_Yes',
       'Reflux_esophagitis_Yes', 'Asthma_Yes', 'Initial_days', 'TotalCharge',
       'Additional_charges', 'Complication_risk_Medium',␣
 ↪'Complication_risk_Low']].hist()
plt.tight_layout()
plt.show()
```

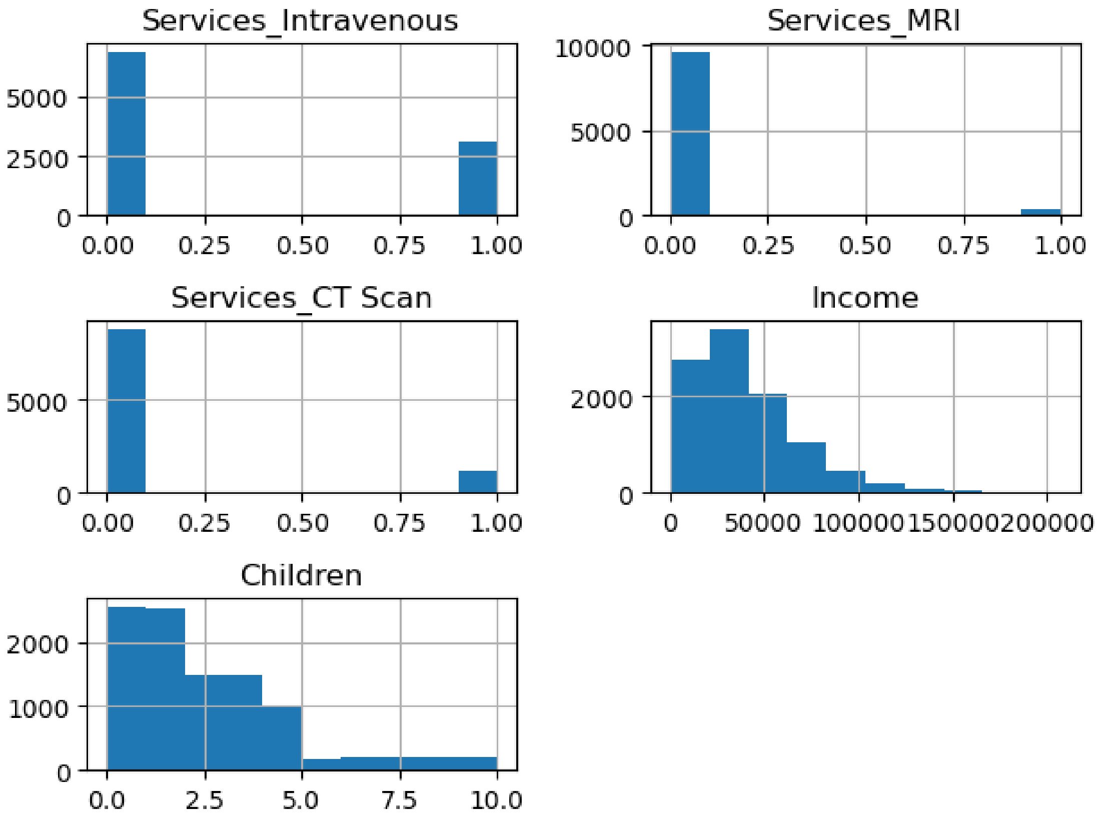


[9]:

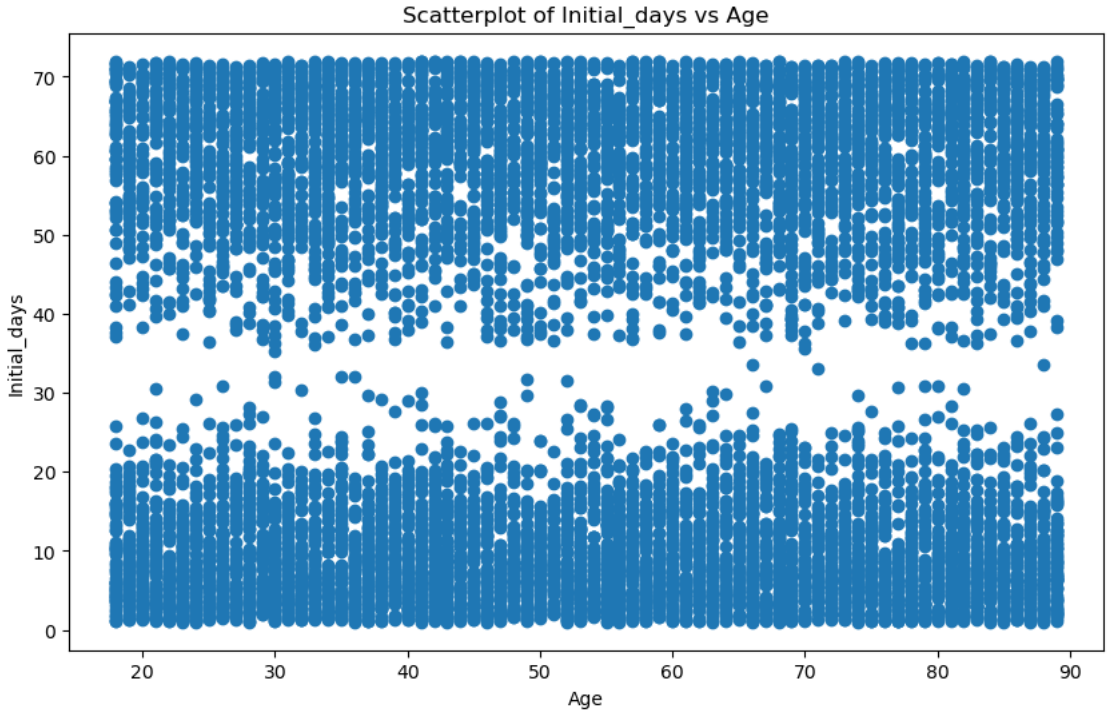
```
data[['Timely_admis', 'Timely_treat', 'Timely_visits',
       'Reliability', 'Options', 'Hours_treat', 'Courteous_staff',
       'Active_listening']].hist()
plt.tight_layout()
plt.show()
```

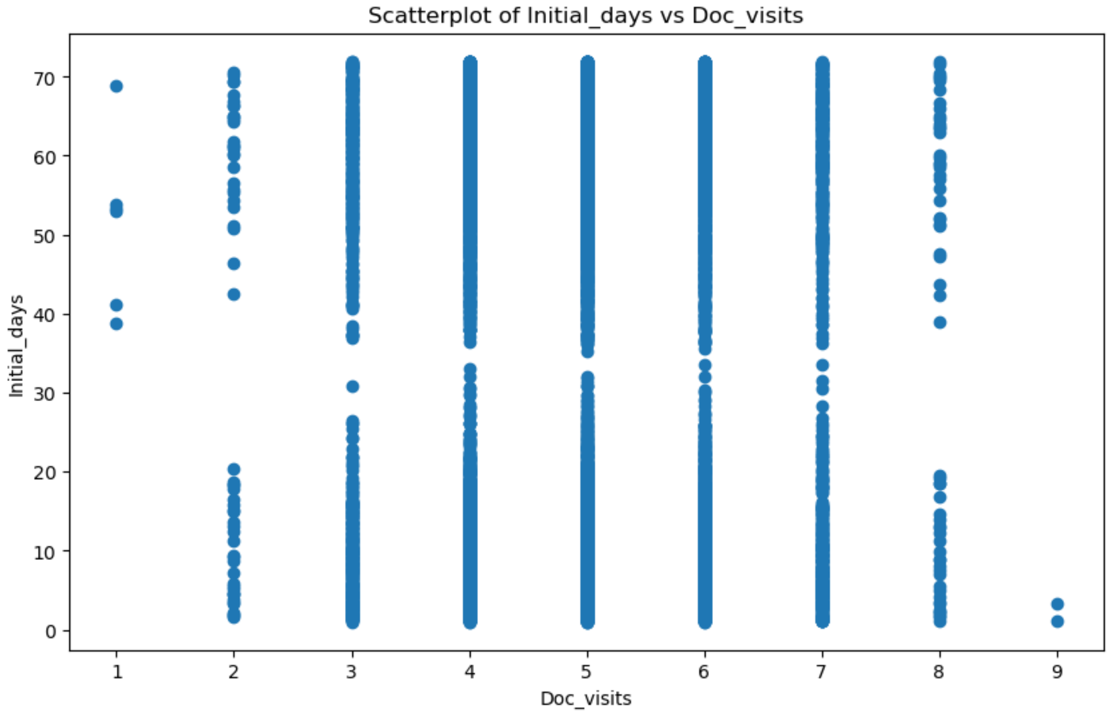


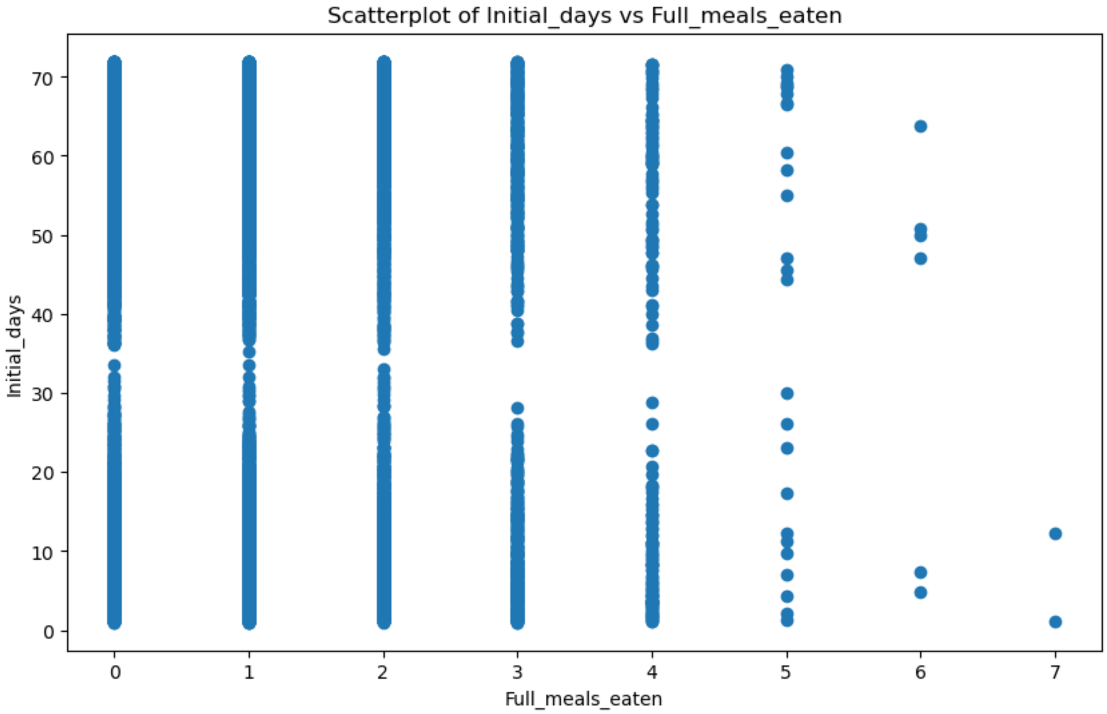
```
[10]: data[['Services_Intravenous', 'Services_MRI', 'Services_CT Scan', 'Income',␣
       ↪'Children']].hist()
      plt.tight_layout()
      plt.show()
```

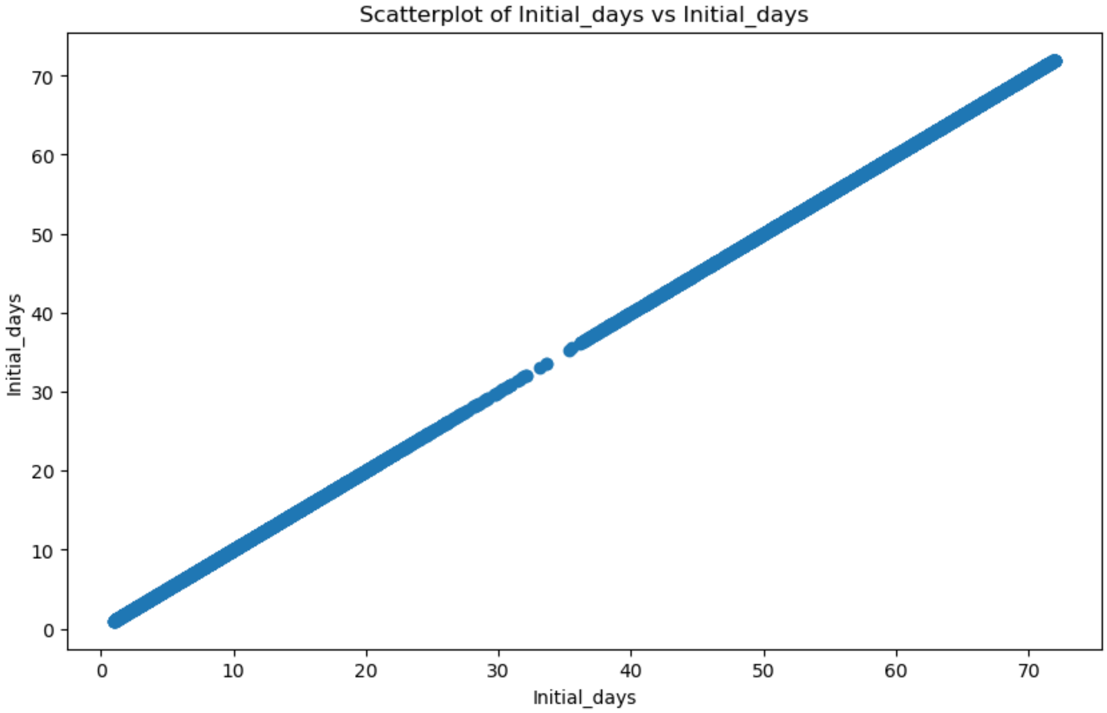


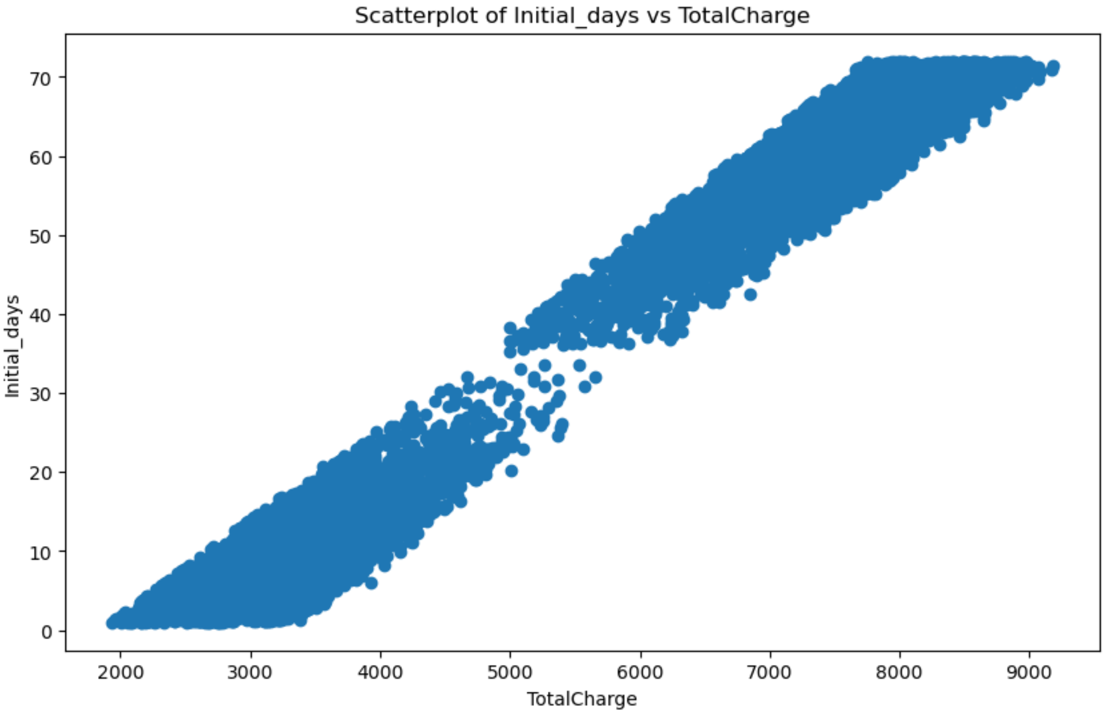
[11]:
```
contData = ['Age', 'Doc_visits', 'Full_meals_eaten',
        'Initial_days', 'TotalCharge', 'Additional_charges', 'Timely_admis',
        'Timely_treat', 'Timely_visits', 'Reliability', 'Options',
        'Hours_treat', 'Courteous_staff', 'Active_listening', 'ReAdmis_Yes',
        'Complication_risk_Low', 'Complication_risk_Medium',
        'Initial_admin_Emergency Admission',
        'Initial_admin_Observation Admission', 'Services_CT Scan',
        'Services_Intravenous', 'Services_MRI', 'Overweight_Yes', 'Anxiety_Yes',
        'Arthritis_Yes', 'Asthma_Yes', 'Diabetes_Yes',
        'Allergic_rhinitis_Yes', 'BackPain_Yes', 'Stroke_Yes', 'HighBlood_Yes',
        'Hyperlipidemia_Yes', 'Reflux_esophagitis_Yes', 'Income', 'Children']
for column in contData:
    plt.figure(figsize=(10, 6))
    plt.scatter(data[column], data['Initial_days'])
    plt.title(f'Scatterplot of Initial_days vs {column}')
    plt.xlabel(column)
    plt.ylabel('Initial_days')
    plt.show()
```

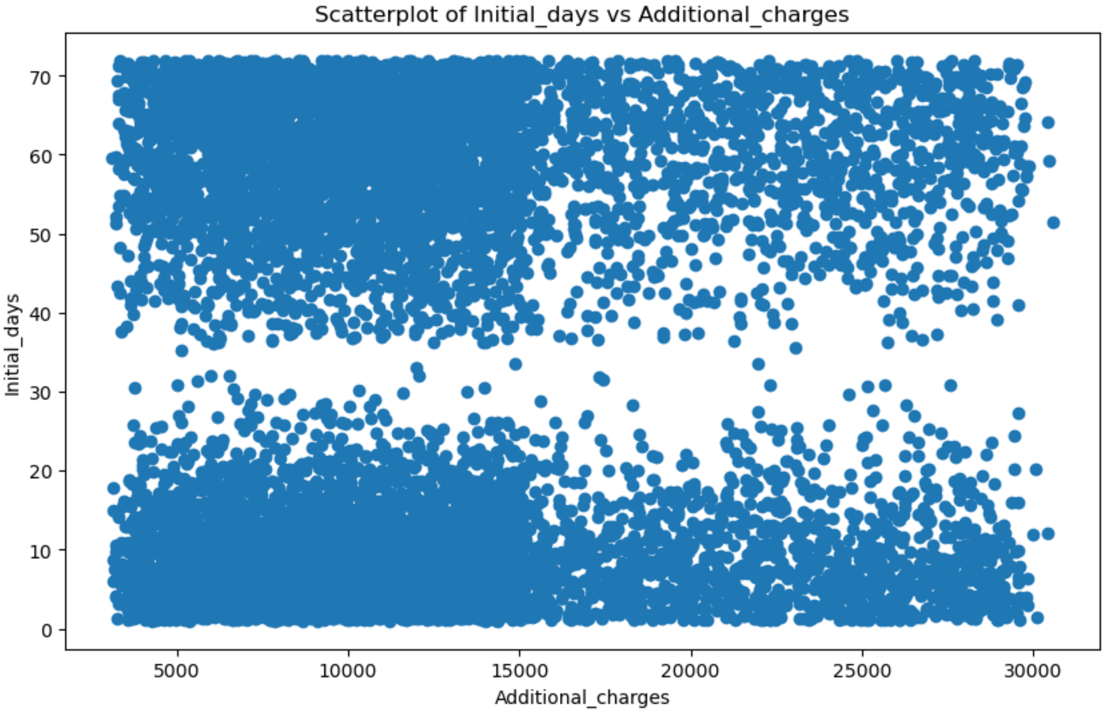

Scatterplot of Initial_days vs Age
Initial_days
0
10
20
30
40
50
60
70
20
30
40
50
60
70
80
90
Age

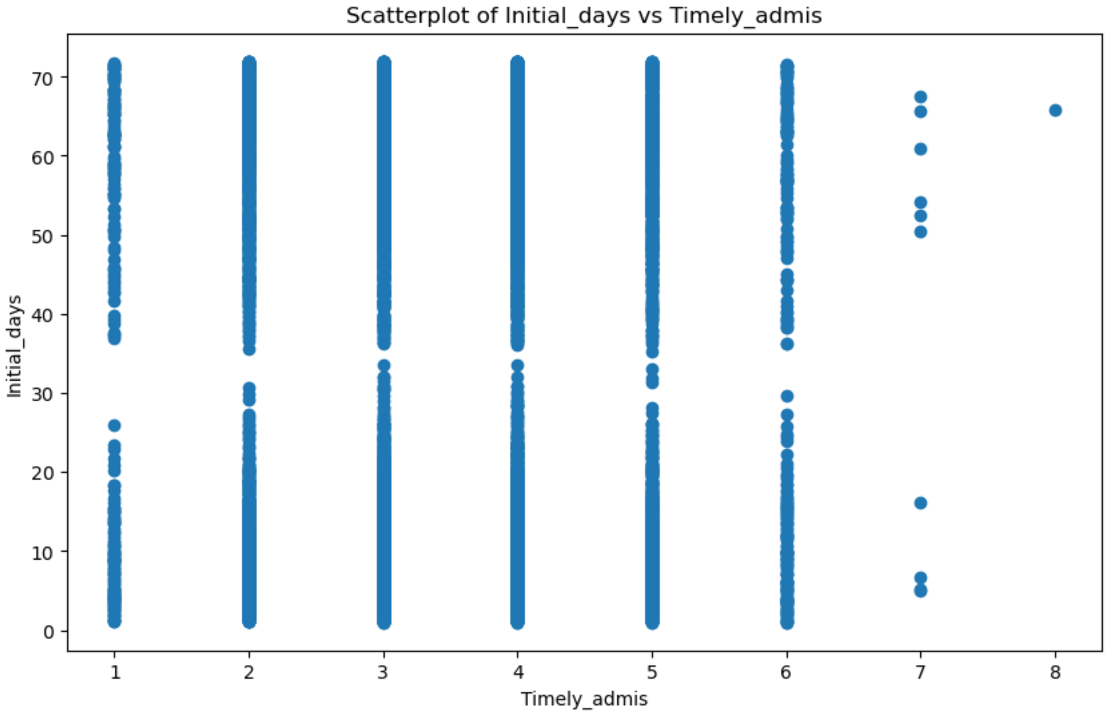

Scatterplot of Initial_days vs Doc_visits
Initial_days
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
8
9
Doc_visits

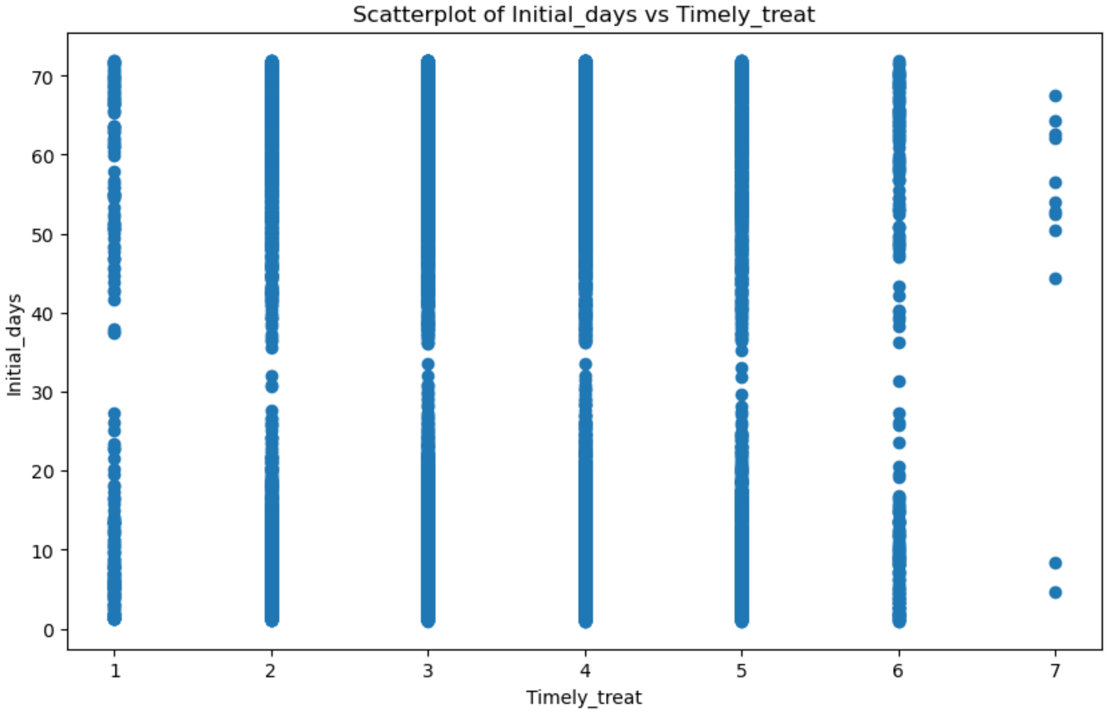

Scatterplot of Initial_days vs Full_meals_eaten
Initial_days
0
10
20
30
40
50
60
70
Full_meals_eaten
0
1
2
3
4
5
6
7

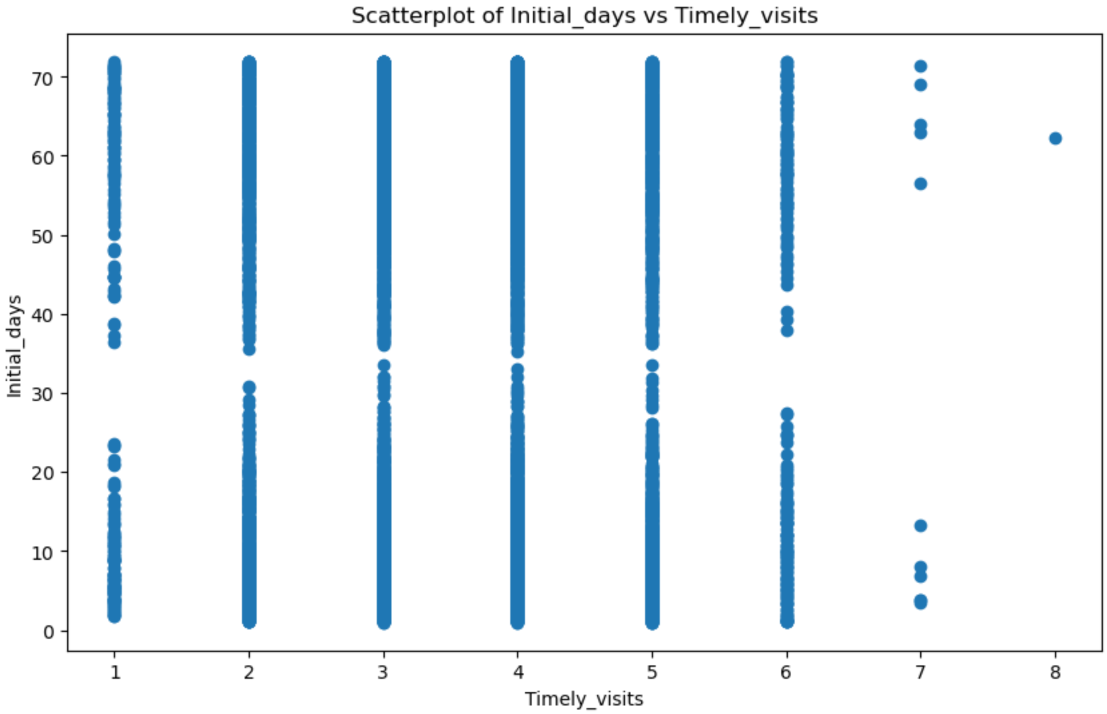

Scatterplot of Initial_days vs Initial_days
Initial_days
0
10
20
30
40
50
60
70
Initial_days
0
10
20
30
40
50
60
70

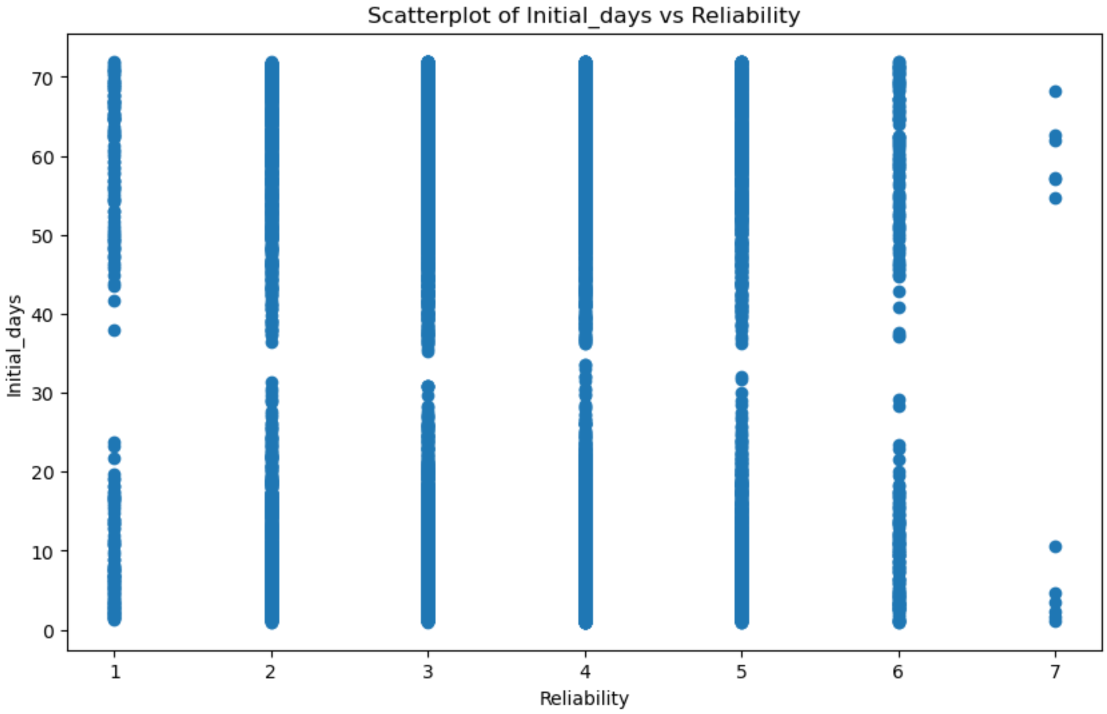

Scatterplot of Initial_days vs TotalCharge
Initial_days
0
10
20
30
40
50
60
70
TotalCharge
2000
3000
4000
5000
6000
7000
8000
9000

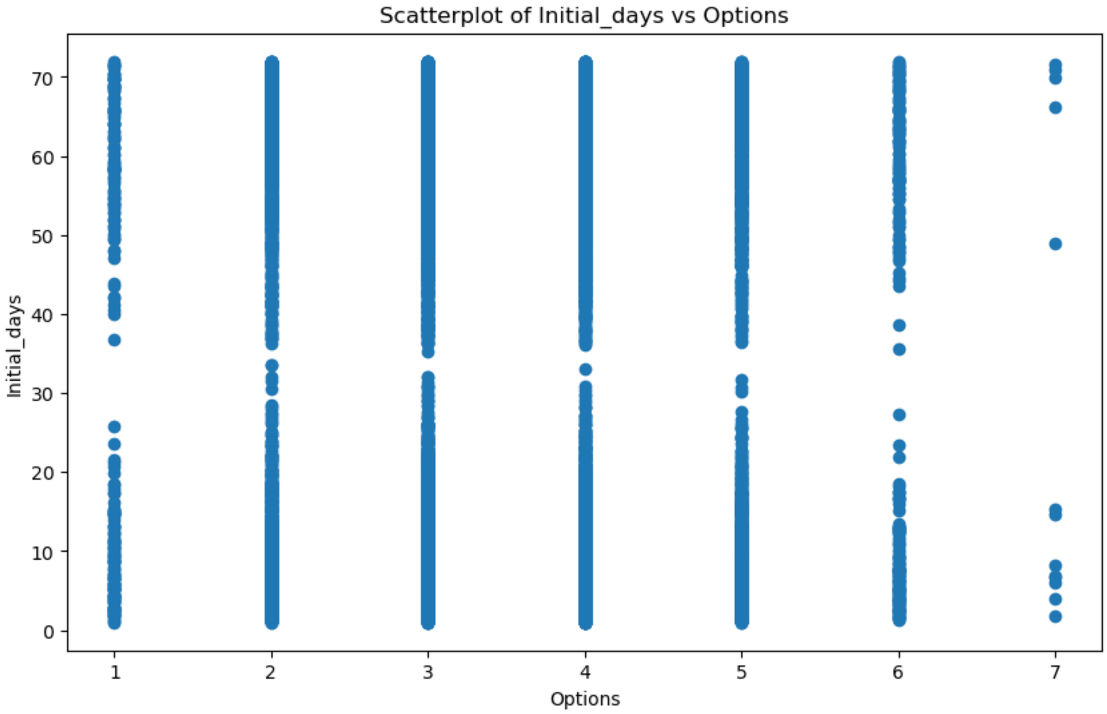

Scatterplot of Initial_days vs Additional_charges
Initial_days
0
10
20
30
40
50
60
70
Additional_charges
5000
10000
15000
20000
25000
30000

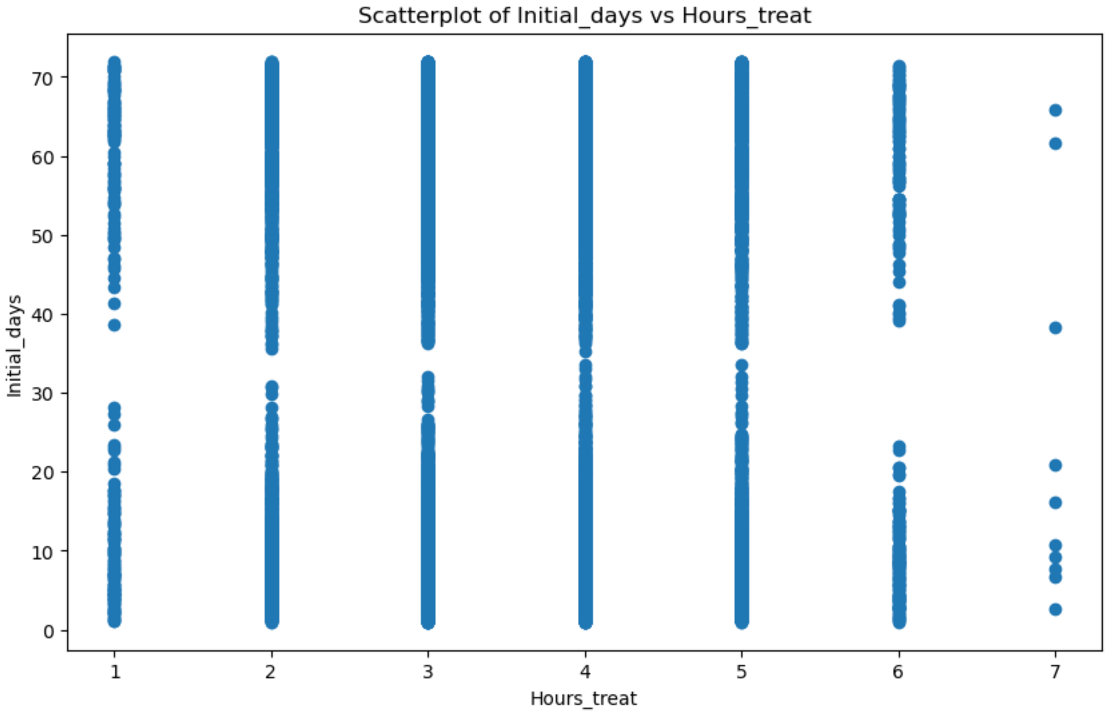

Scatterplot of Initial_days vs Timely_admis
Initial_days
Timely_admis
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
8

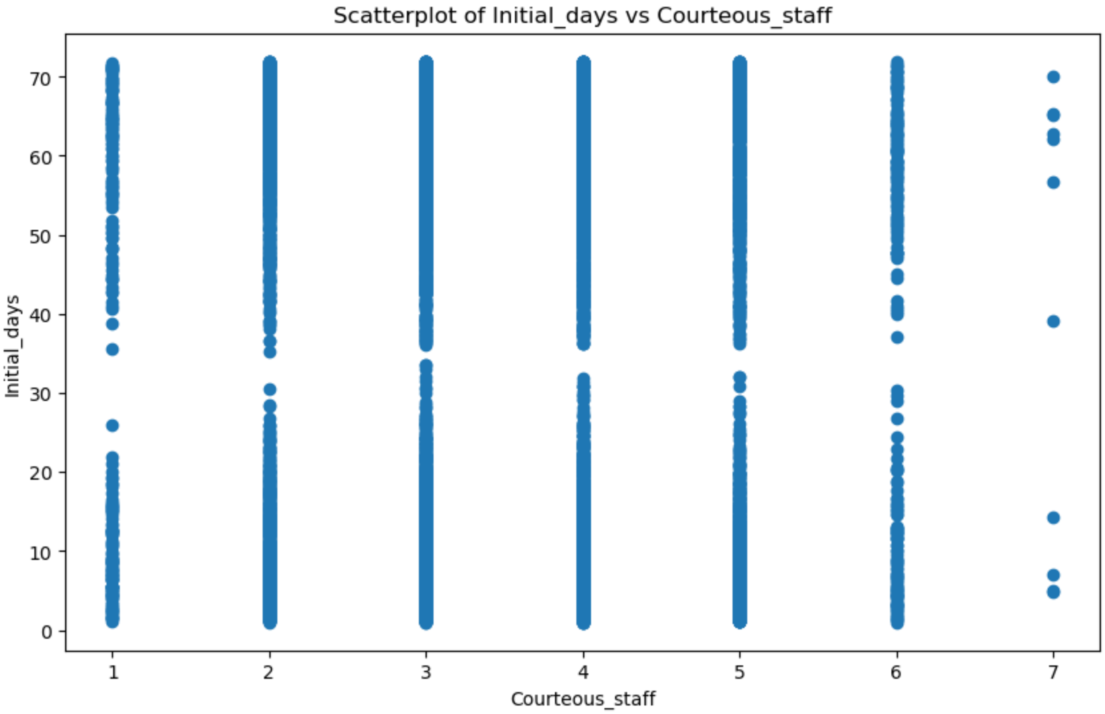

Scatterplot of Initial_days vs Timely_treat
Initial_days
Timely_treat
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7

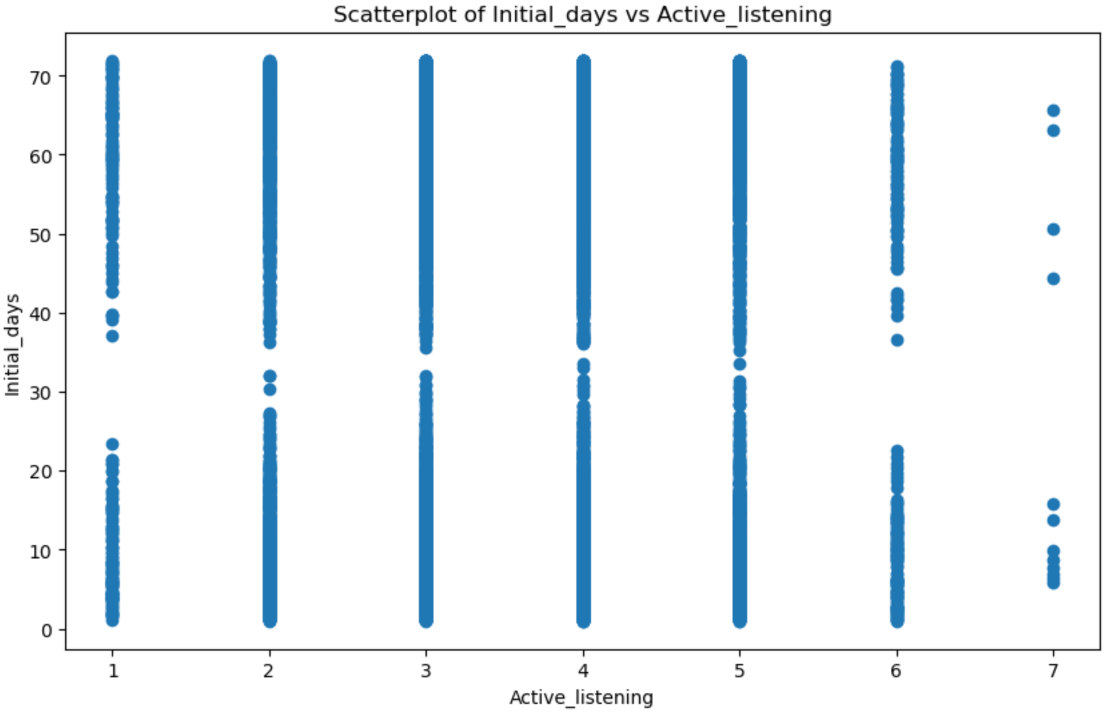

Scatterplot of Initial_days vs Timely_visits
Initial_days
0
10
20
30
40
50
60
70
Timely_visits
1
2
3
4
5
6
7
8

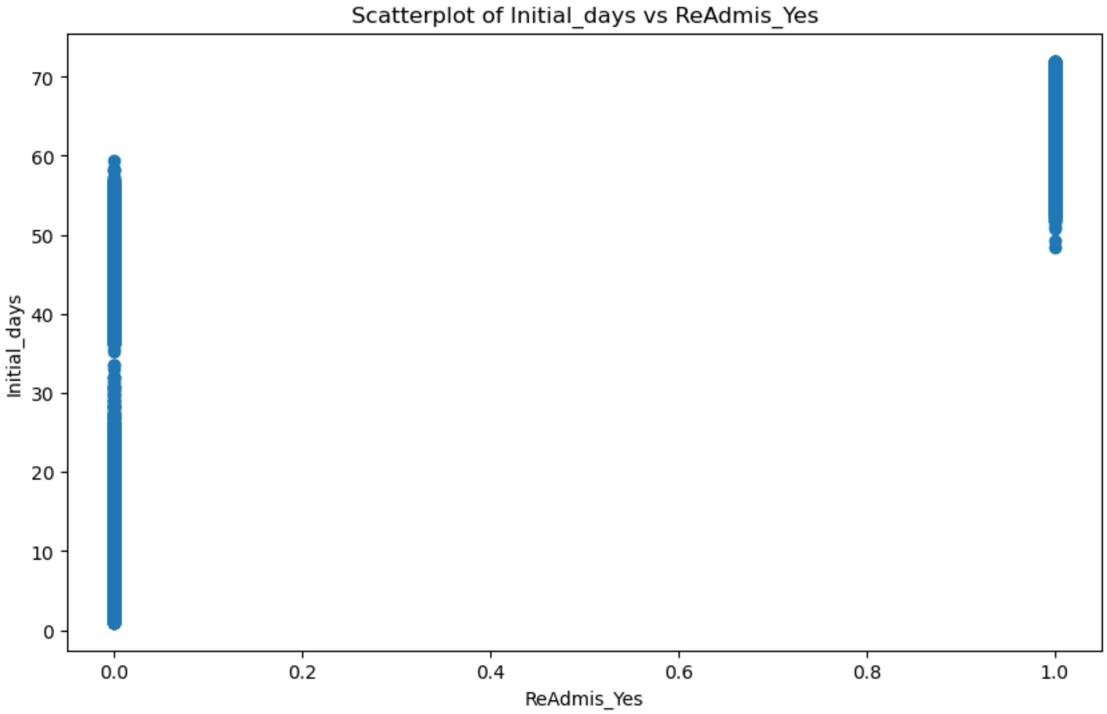

Scatterplot of Initial_days vs Reliability
Initial_days
0
10
20
30
40
50
60
70
Reliability
1
2
3
4
5
6
7

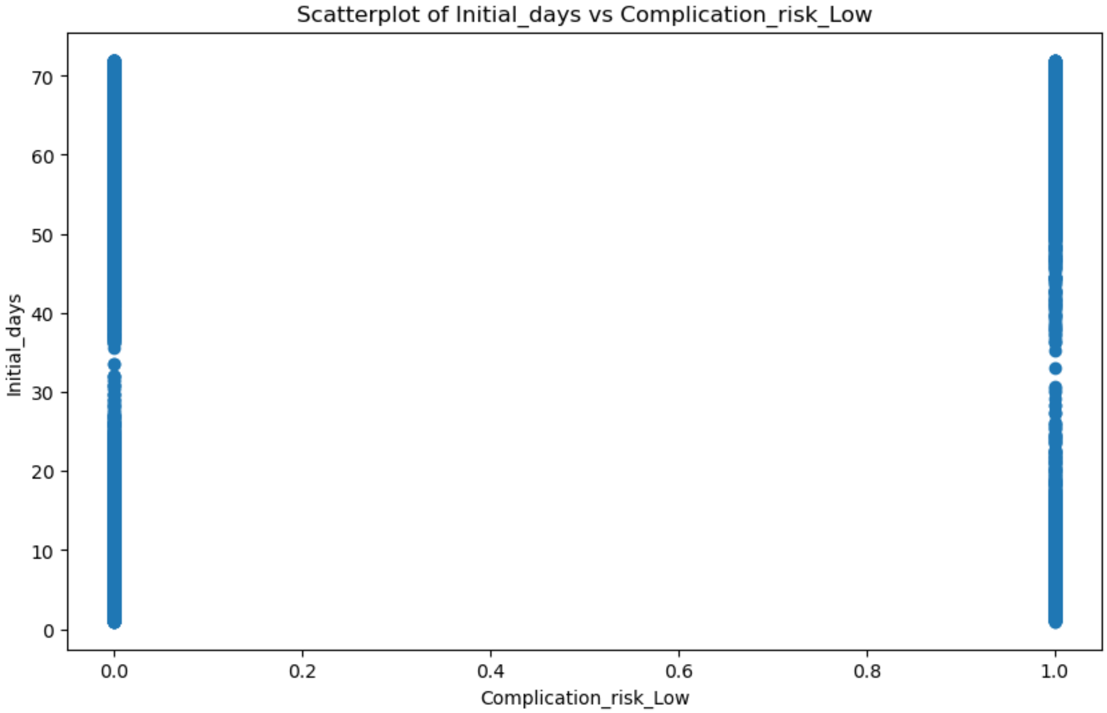

Scatterplot of Initial_days vs Options
Initial_days
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
Options

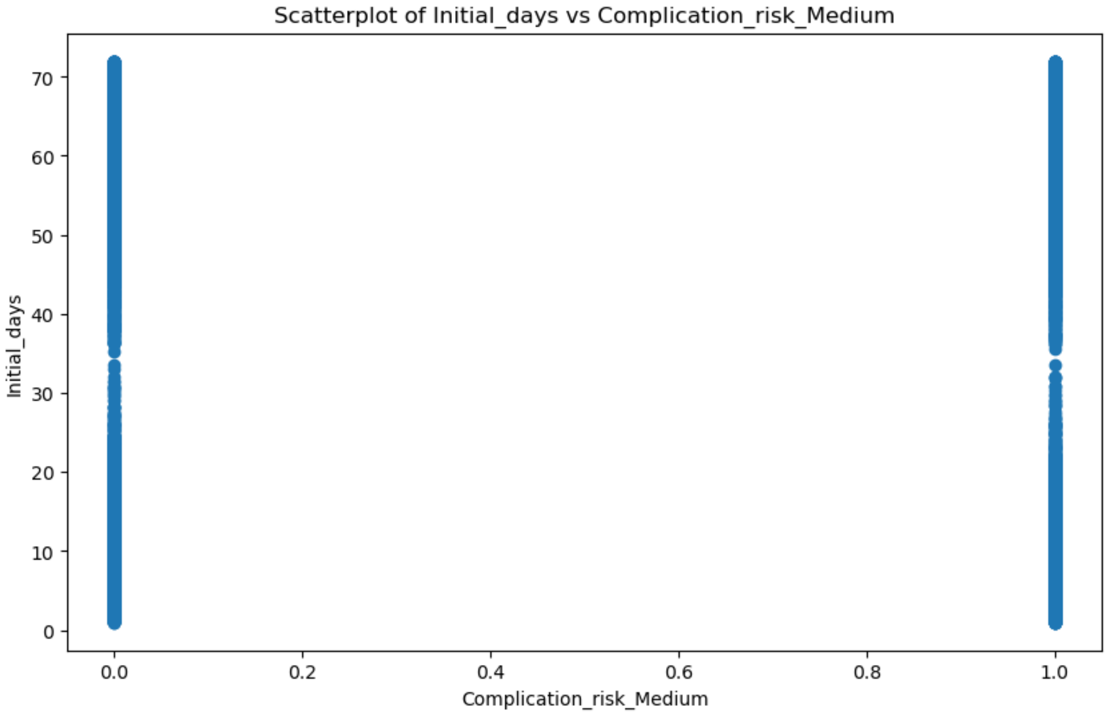

Scatterplot of Initial_days vs Hours_treat
Initial_days
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
Hours_treat

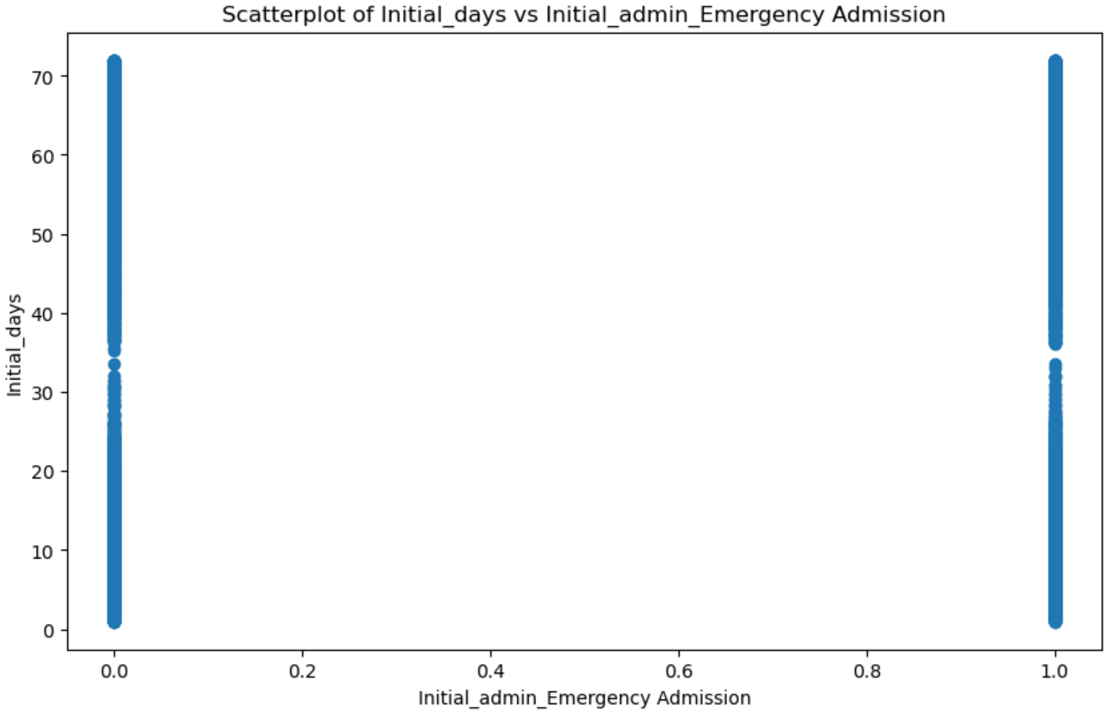

Scatterplot of Initial_days vs Courteous_staff
Initial_days
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
Courteous_staff

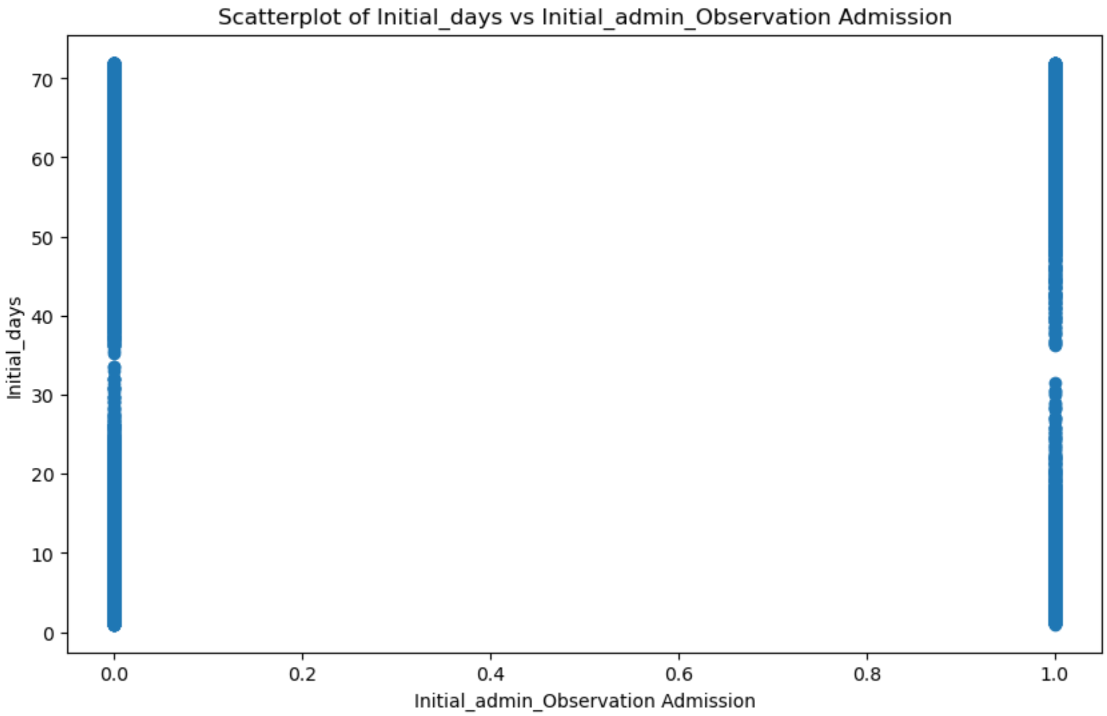

Scatterplot of Initial_days vs Active_listening
Initial_days
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
Active_listening

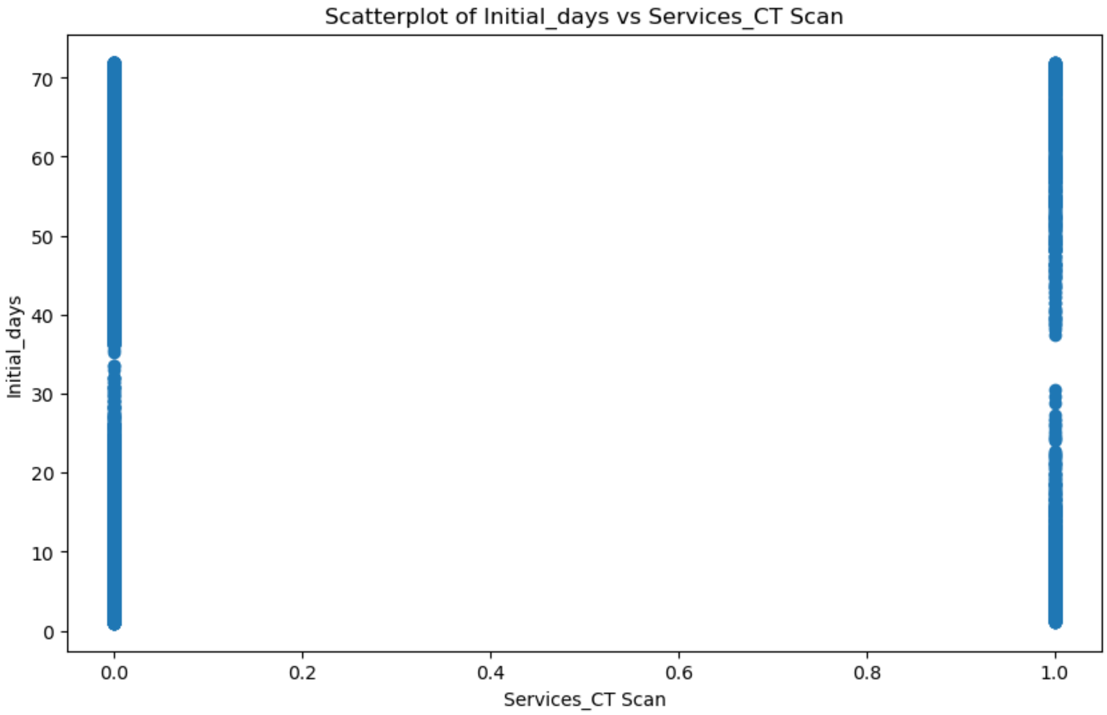

Scatterplot of Initial_days vs ReAdmis_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
ReAdmis_Yes

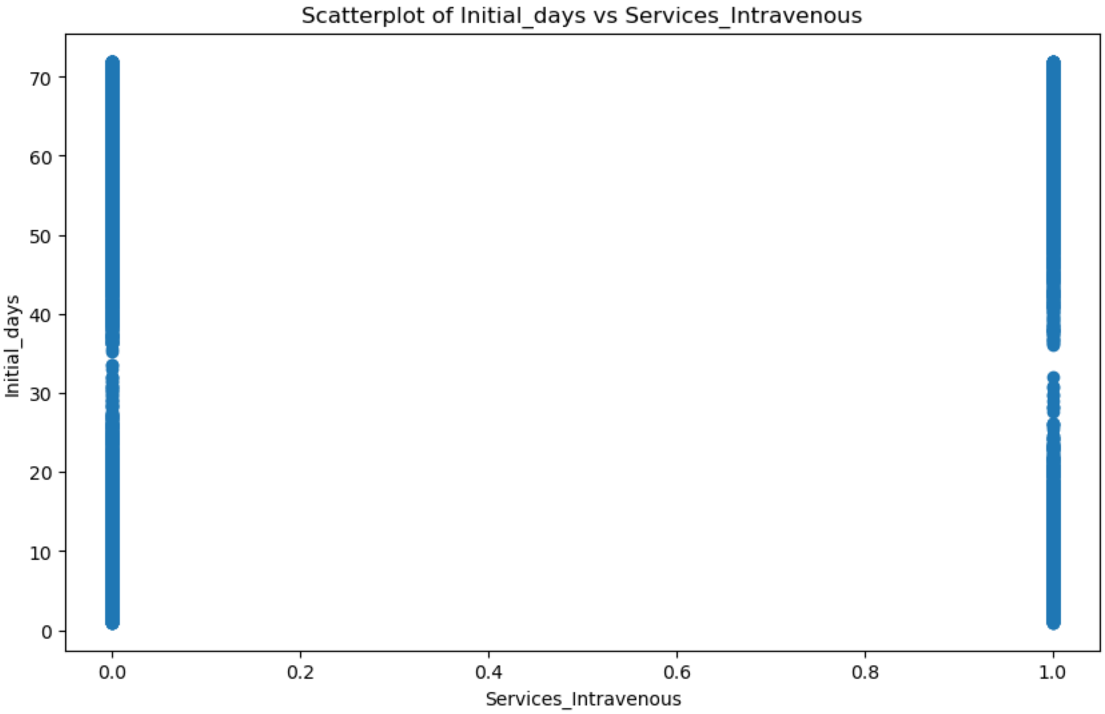

Scatterplot of Initial_days vs Complication_risk_Low
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Complication_risk_Low

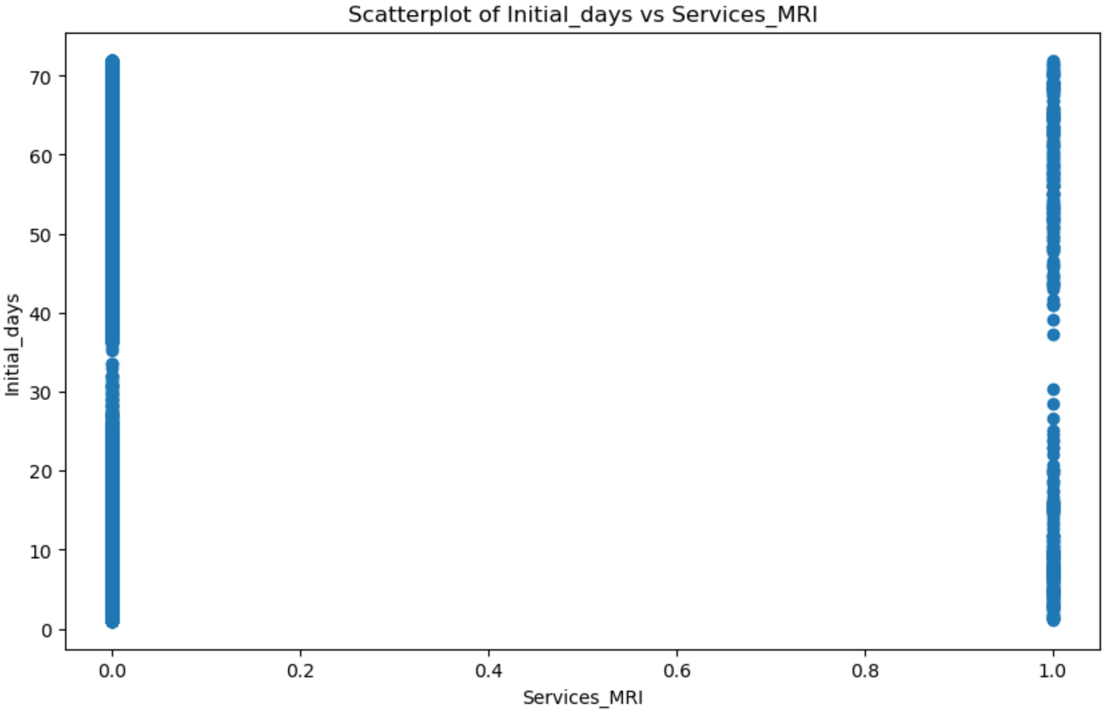

Scatterplot of Initial_days vs Complication_risk_Medium
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Complication_risk_Medium

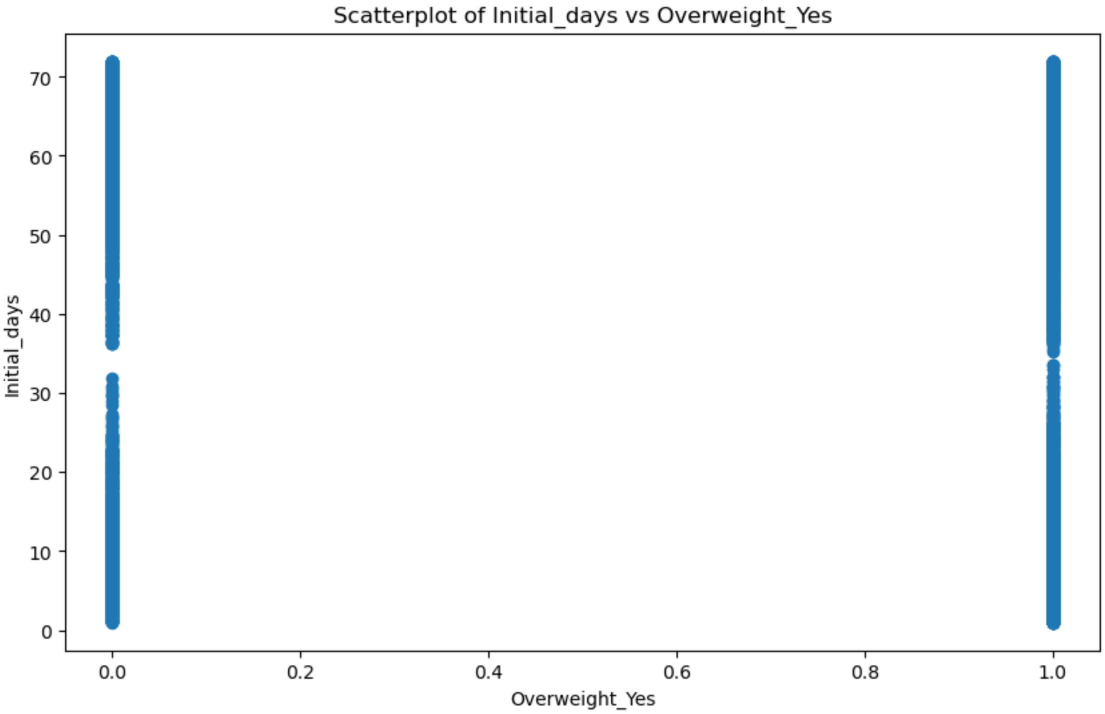

Scatterplot of Initial_days vs Initial_admin_Emergency Admission
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Initial_admin_Emergency Admission

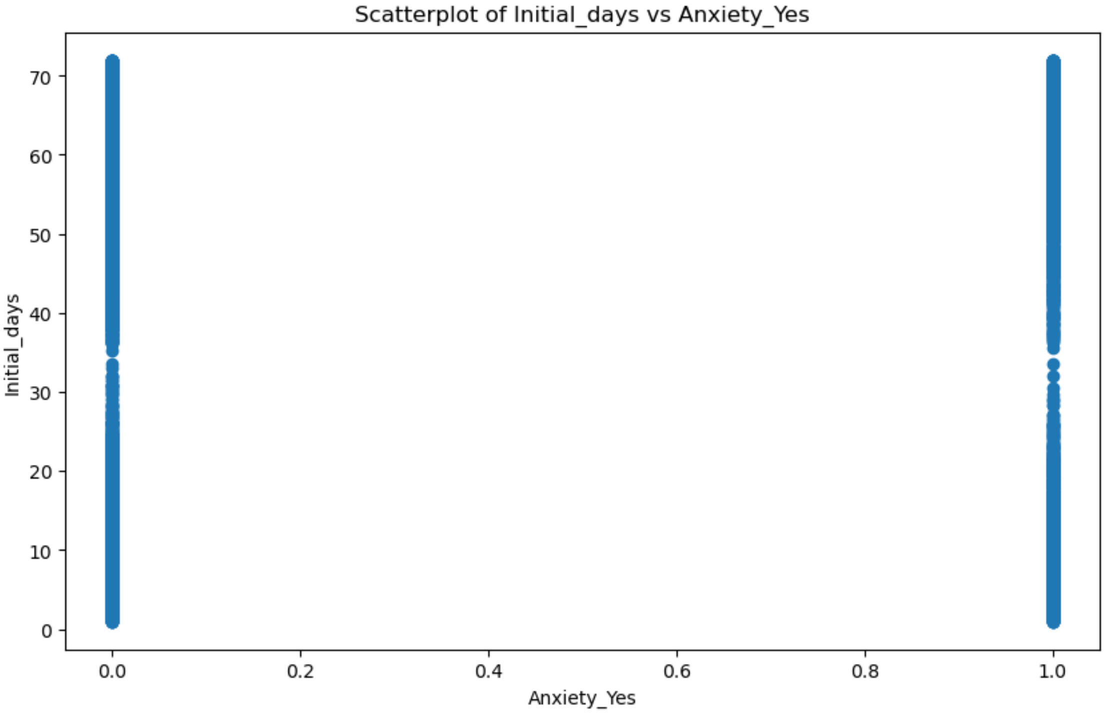

Scatterplot of Initial_days vs Initial_admin_Observation Admission
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Initial_admin_Observation Admission

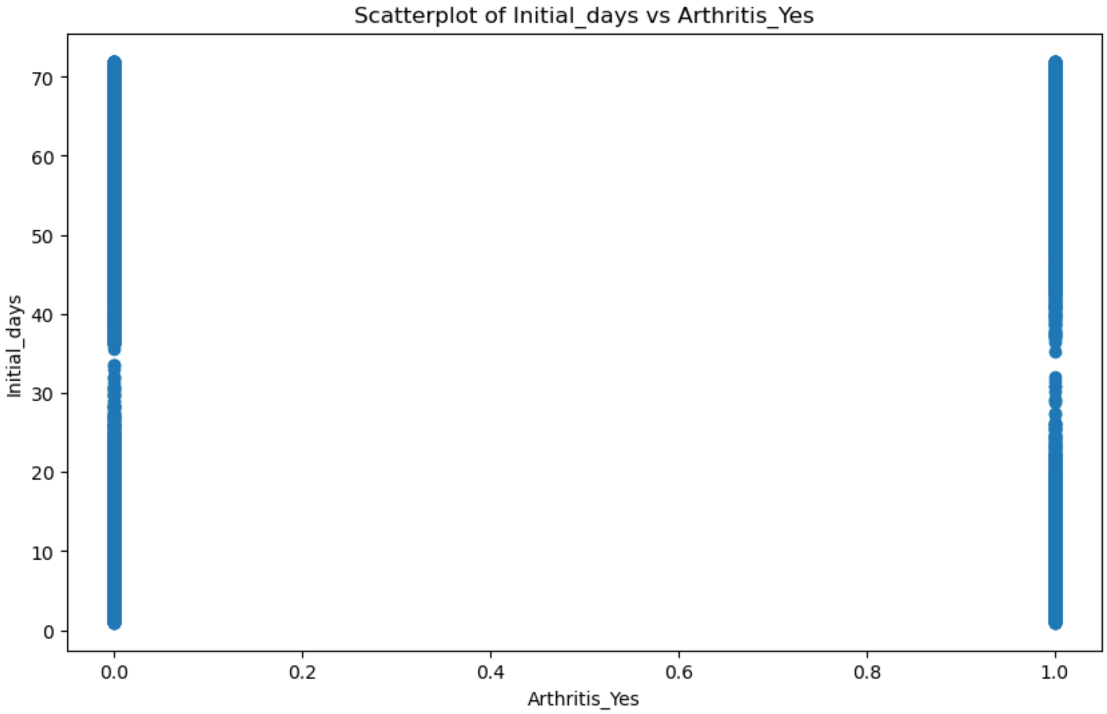

Scatterplot of Initial_days vs Services_CT Scan
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Services_CT Scan

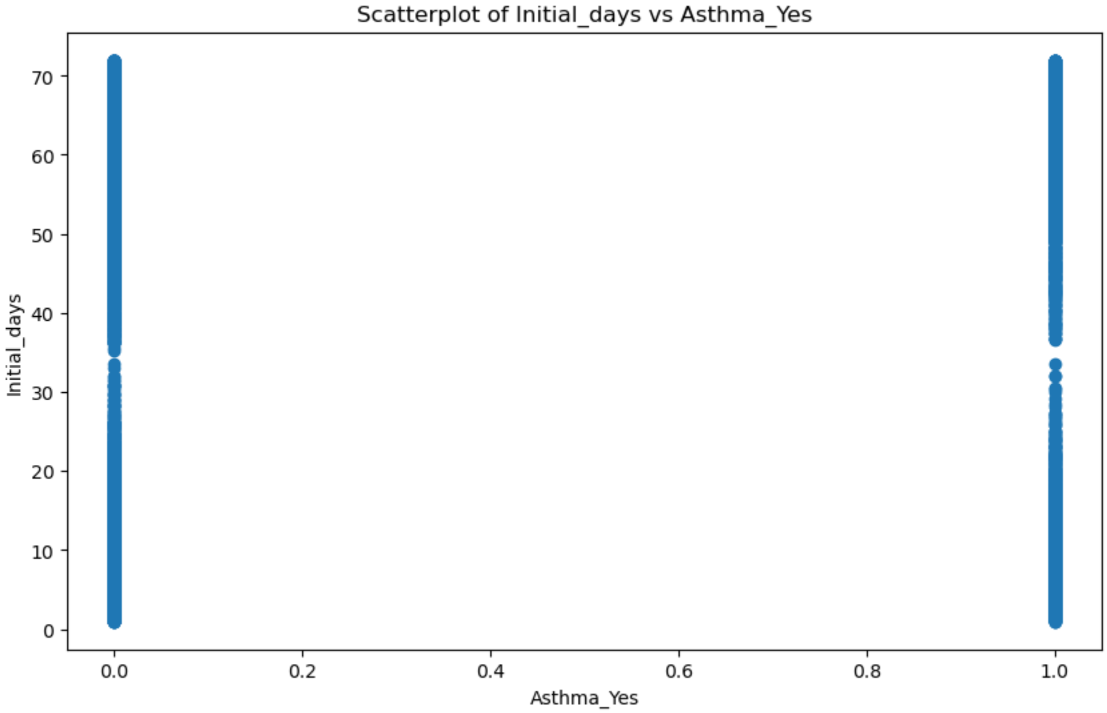

Scatterplot of Initial_days vs Services_Intravenous
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Services_Intravenous

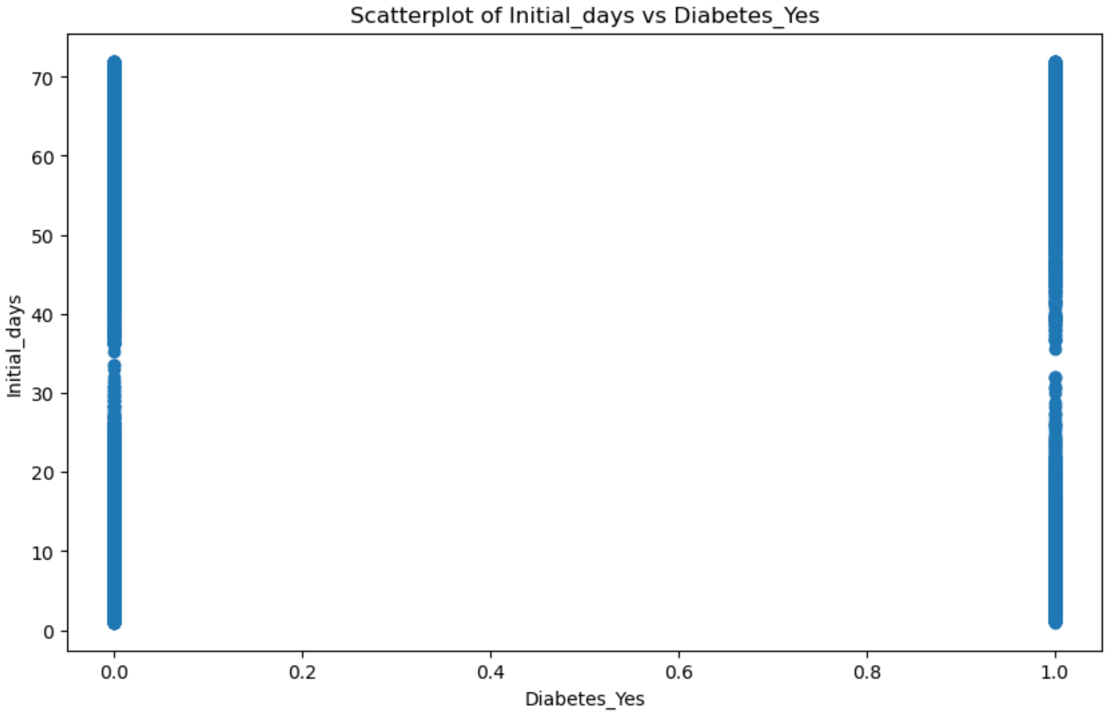

Scatterplot of Initial_days vs Services_MRI
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Services_MRI

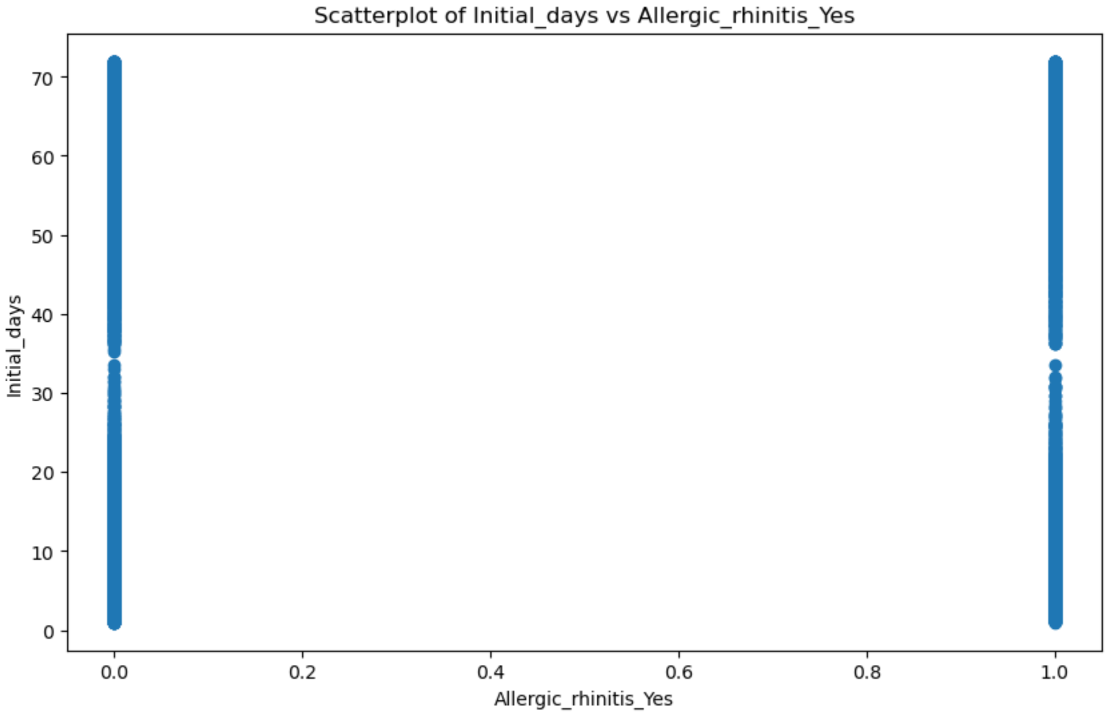

Scatterplot of Initial_days vs Overweight_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Overweight_Yes

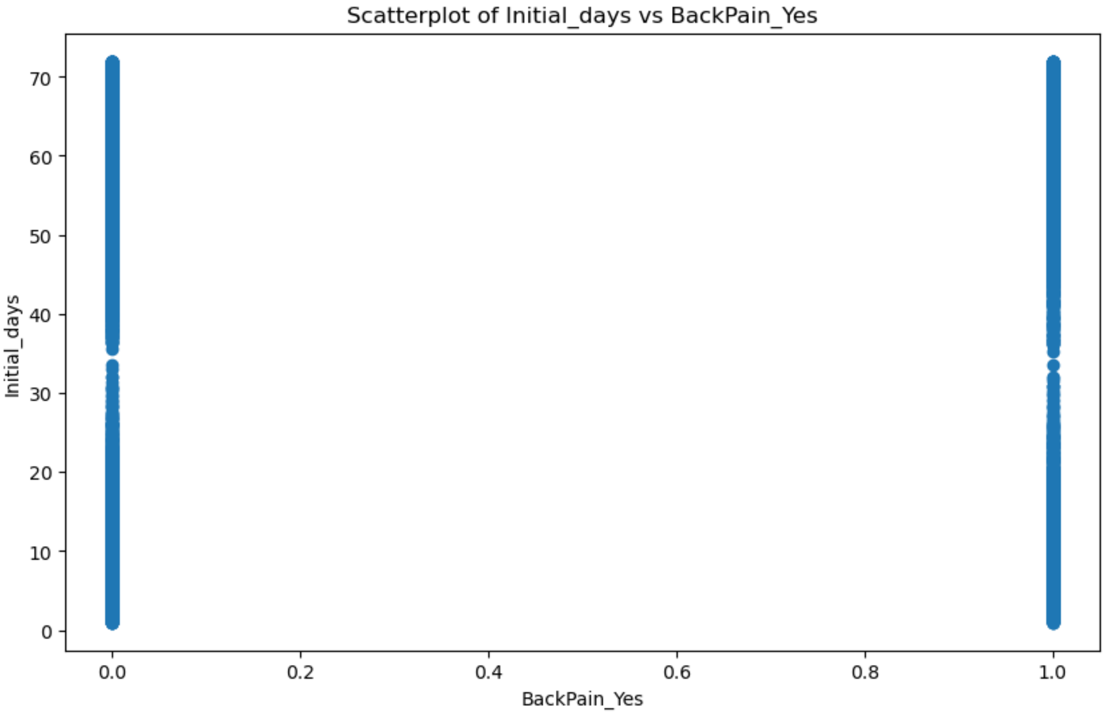

Scatterplot of Initial_days vs Anxiety_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Anxiety_Yes

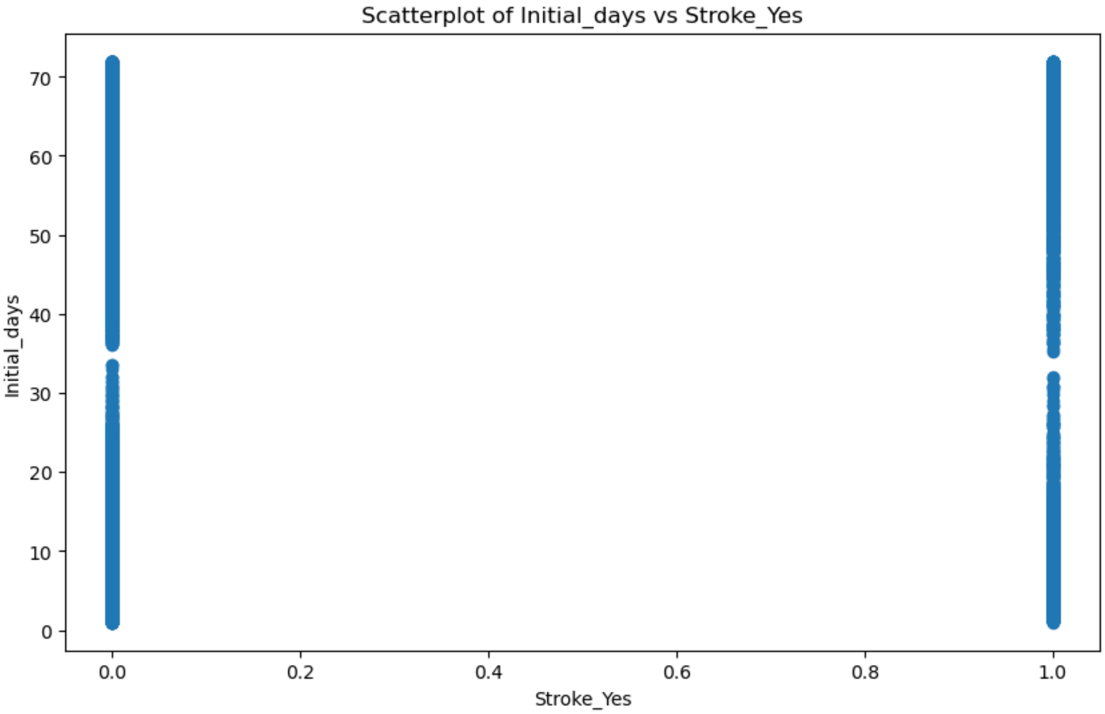

Scatterplot of Initial_days vs Arthritis_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Arthritis_Yes

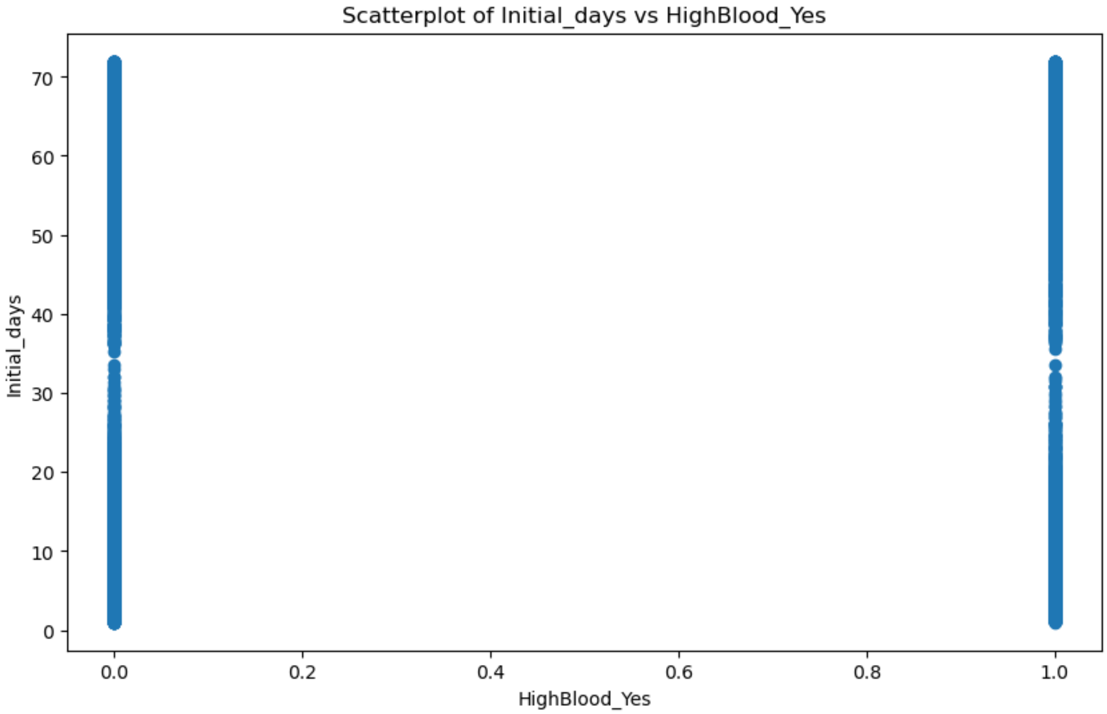

Scatterplot of Initial_days vs Asthma_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Asthma_Yes

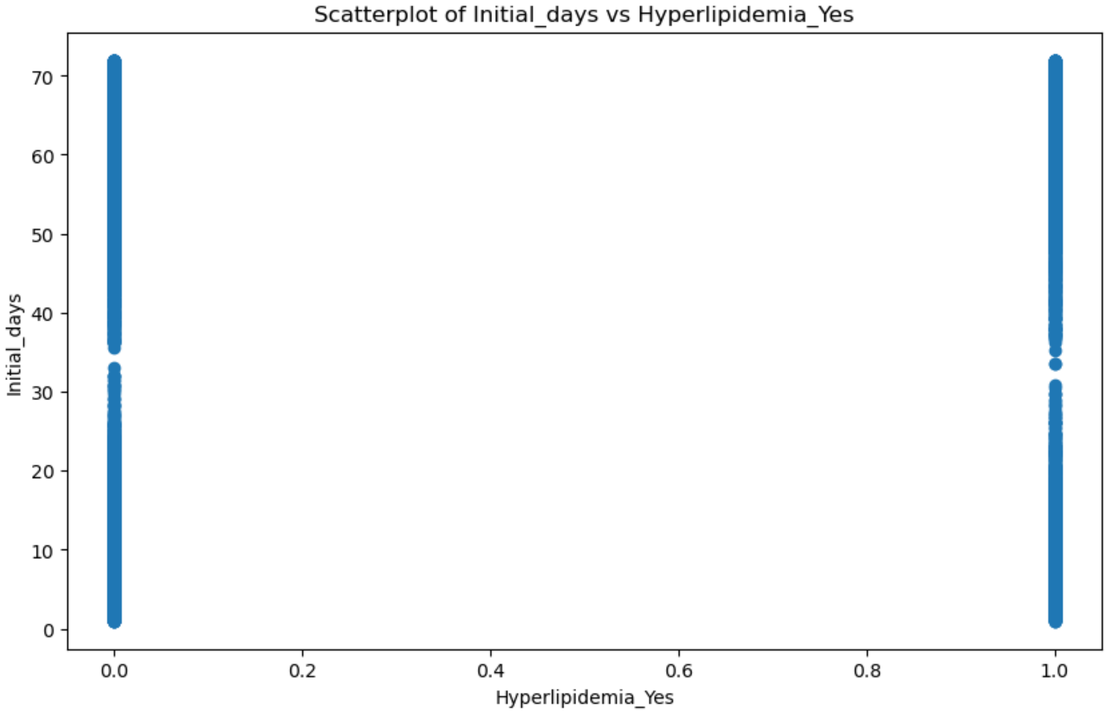

Scatterplot of Initial_days vs Diabetes_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Diabetes_Yes

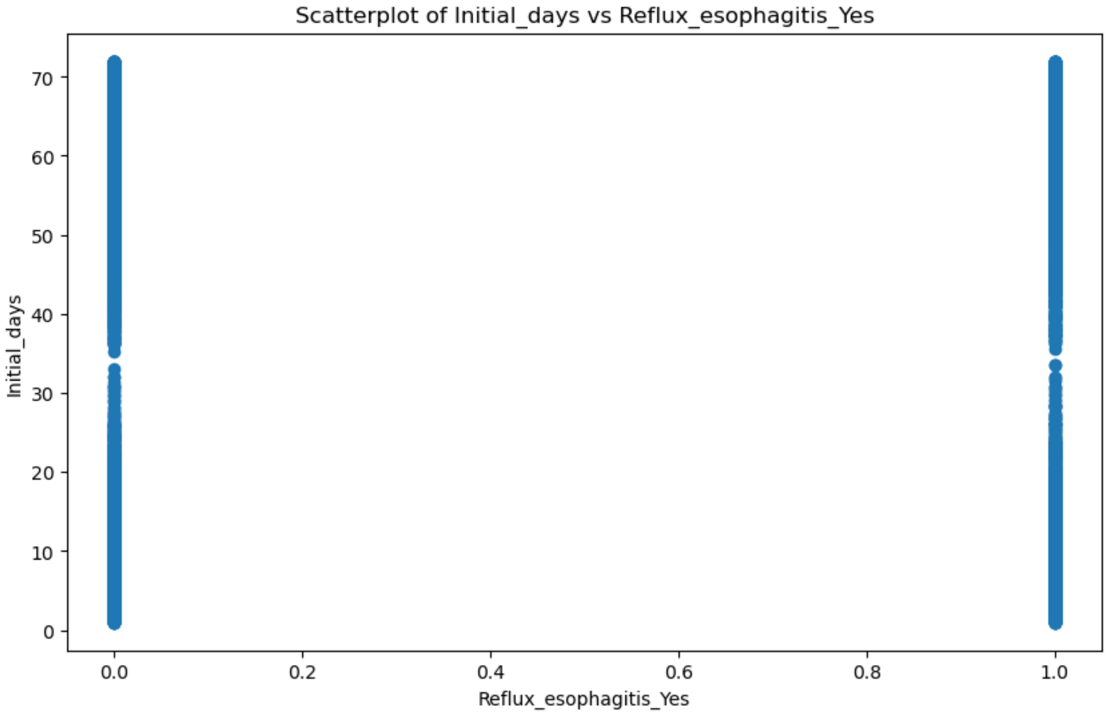

Scatterplot of Initial_days vs Allergic_rhinitis_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Allergic_rhinitis_Yes

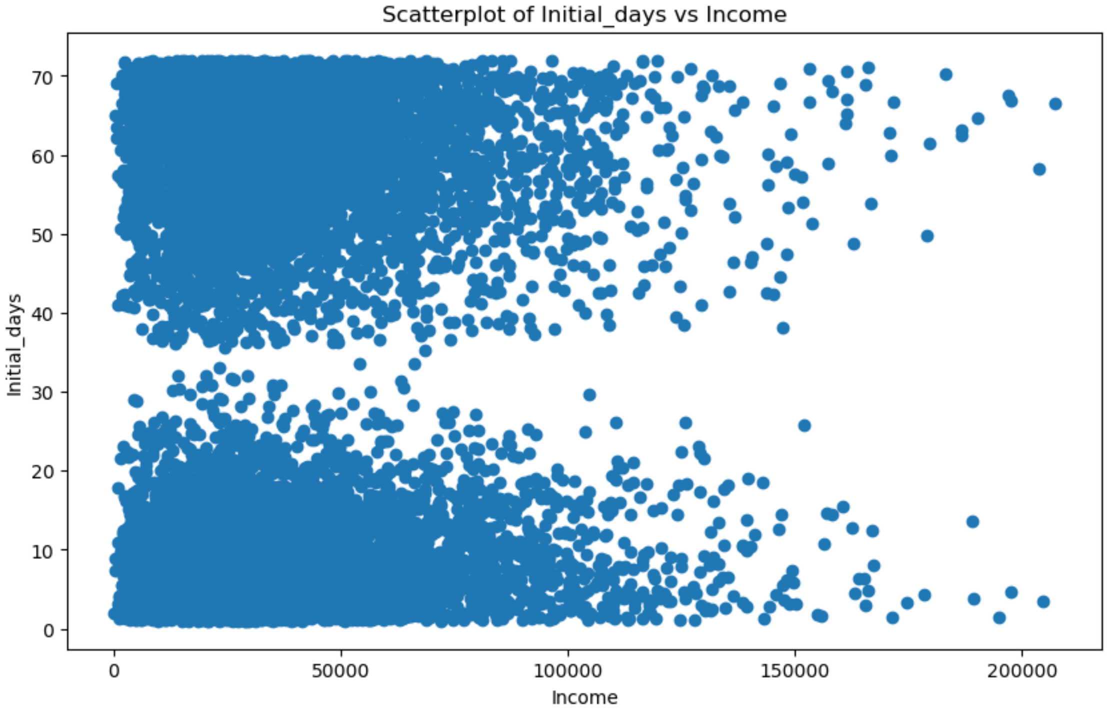

Scatterplot of Initial_days vs BackPain_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
BackPain_Yes

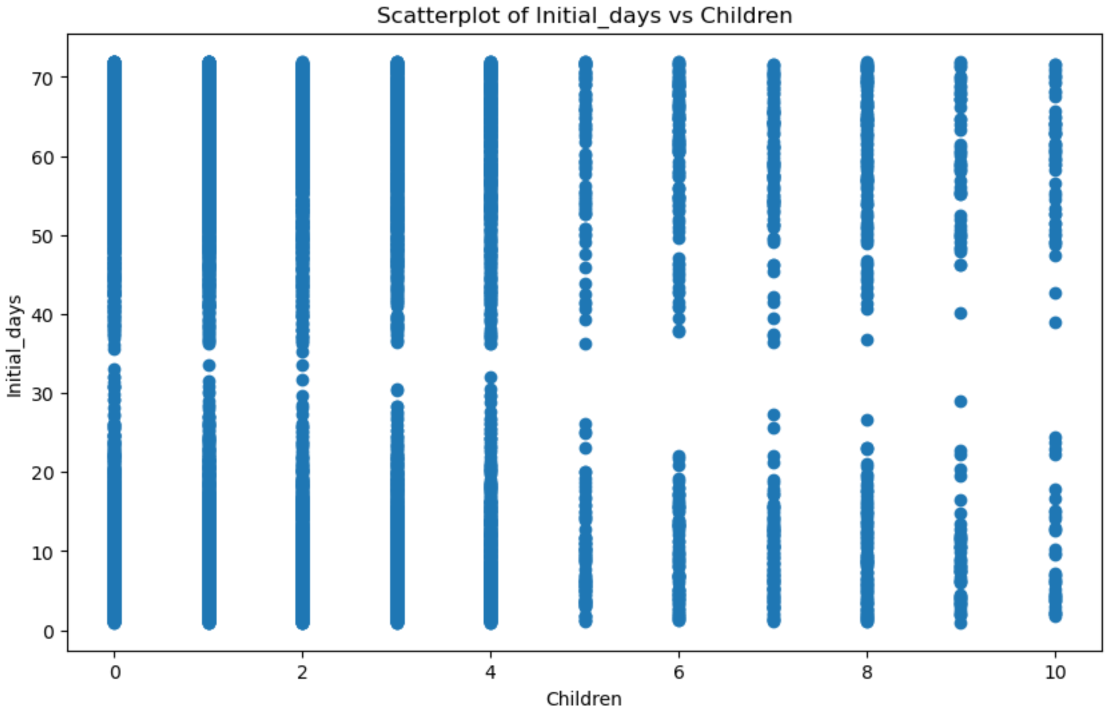

Scatterplot of Initial_days vs Stroke_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Stroke_Yes


Scatterplot of Initial_days vs HighBlood_Yes
Initial_days
0
10
20
30
40
50
60
70
HighBlood_Yes
0.0
0.2
0.4
0.6
0.8
1.0


Scatterplot of Initial_days vs Hyperlipidemia_Yes
Initial_days
0
10
20
30
40
50
60
70
Hyperlipidemia_Yes
0.0
0.2
0.4
0.6
0.8
1.0


Scatterplot of Initial_days vs Reflux_esophagitis_Yes
Initial_days
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Reflux_esophagitis_Yes


Scatterplot of Initial_days vs Income
Initial_days
0
10
20
30
40
50
60
70
0
50000
100000
150000
200000
Income



### C4. Data Transformation Goals

Before statistical analysis can take place, the data will need a review. The first step is to check for missing/null values and verifying none exist. We also check if any duplicated data exists and handle it accordingly. The majority of the columns identified as not providing any benefit to this particular analysis were patient location based and thus needed to be dropped. Several column names such as Item1 through Item8 will be renamed to provide clarity during analysis. Lastly, we need to convert categorical variables to numerical types. Any columns with Yes/No values or Low, Medium, and High, will be converted to numerical along with using the drop-one method:

[71]:
```
%matplotlib inline
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
data=pd.read_csv('medical_clean.csv')
print(data.head())
print(data.columns)

#check for missing/null values
print(data.isnull().sum())

#check for duplicate values of any rows
print(data.duplicated().any())
```

```
# Check for duplicate values based on customer_id unique key
print(data.duplicated('Customer_id').any())

# remove unused columns
data.drop(['CaseOrder', 'Customer_id','Interaction', 'UID', 'City', 'Children',␣
 ↪'Income', 'Marital', 'State', 'County', 'Zip', 'Lat', 'Lng', 'Population',␣
 ↪'Area', 'TimeZone', 'Job', 'Gender'], axis=1, inplace=True)
print(data.head())
print(data.columns)
print(data.info())

# rename unclear survey response columns
survey_col_names = {
    'Item1': 'Timely_admis',
    'Item2': 'Timely_treat',
    'Item3': 'Timely_visits',
    'Item4': 'Reliability',
    'Item5': 'Options',
    'Item6': 'Hours_treat',
    'Item7': 'Courteous_staff',
    'Item8': 'Active_listening'
}

data = data.rename(columns=survey_col_names)

categorical_cols = ['ReAdmis', 'Complication_risk', 'Initial_admin', 'Services',
                    'Overweight', 'Anxiety', 'Arthritis', 'Asthma',␣
 ↪'Soft_drink',
                    'Diabetes', 'Allergic_rhinitis', 'BackPain', 'Stroke',␣
 ↪'HighBlood',
                    'Hyperlipidemia', 'Reflux_esophagitis']

# dummy variables
data = pd.get_dummies(data, columns=categorical_cols, drop_first=True)

print(data.info())

print(data.head())

# display all columns
pd.set_option('display.max_columns', None)
# display all the rows
pd.set_option('display.max_rows', None)

print(data.describe(include='all'))
print(data.columns)
```

```
   CaseOrder Customer_id                           Interaction  \
0          1     C412403  8cd49b13-f45a-4b47-a2bd-173ffa932c2f
1          2     Z919181  d2450b70-0337-4406-bdbb-bc1037f1734c
2          3     F995323  a2057123-abf5-4a2c-abad-8ffe33512562
3          4     A879973  1dec528d-eb34-4079-adce-0d7a40e82205
4          5     C544523  5885f56b-d6da-43a3-8760-83583af94266

                                UID          City State        County    Zip  \
0  3a83ddb66e2ae73798bdf1d705dc0932           Eva    AL        Morgan  35621
1  176354c5eef714957d486009feabf195      Marianna    FL       Jackson  32446
2  e19a0fa00aeda885b8a436757e889bc9   Sioux Falls    SD     Minnehaha  57110
3  cd17d7b6d152cb6f23957346d11c3f07  New Richland    MN        Waseca  56072
4  d2f0425877b10ed6bb381f3e2579424a    West Point    VA  King William  23181

       Lat       Lng  Population      Area          TimeZone  \
0  34.34960 -86.72508        2951  Suburban   America/Chicago
1  30.84513 -85.22907       11303     Urban   America/Chicago
2  43.54321 -96.63772       17125  Suburban   America/Chicago
3  43.89744 -93.51479        2162  Suburban   America/Chicago
4  37.59894 -76.88958        5287     Rural  America/New_York

                                Job  Children  Age    Income   Marital  \
0  Psychologist, sport and exercise         1   53  86575.93  Divorced
1      Community development worker         3   51  46805.99   Married
2           Chief Executive Officer         3   53  14370.14   Widowed
3               Early years teacher         0   78  39741.49   Married
4       Health promotion specialist         1   22   1209.56   Widowed

   Gender ReAdmis  VitD_levels  Doc_visits  Full_meals_eaten  vitD_supp  \
0    Male      No    19.141466           6                 0          0
1  Female      No    18.940352           4                 2          1
2  Female      No    18.057507           4                 1          0
3    Male      No    16.576858           4                 1          0
4  Female      No    17.439069           5                 0          2

  Soft_drink        Initial_admin HighBlood Stroke Complication_risk  \
0         No  Emergency Admission       Yes     No            Medium
1         No  Emergency Admission       Yes     No              High
2         No   Elective Admission       Yes     No            Medium
3         No   Elective Admission        No    Yes            Medium
4        Yes   Elective Admission        No     No               Low

  Overweight Arthritis Diabetes Hyperlipidemia BackPain Anxiety  \
0         No       Yes      Yes             No      Yes     Yes
1        Yes        No       No             No       No      No
2        Yes        No      Yes             No       No      No
3         No       Yes       No             No       No      No
4         No        No       No            Yes       No      No
```

```
  Allergic_rhinitis Reflux_esophagitis Asthma     Services  Initial_days  \
0               Yes                 No    Yes   Blood Work     10.585770
1                No                Yes     No  Intravenous     15.129562
2                No                 No     No   Blood Work      4.772177
3                No                Yes    Yes   Blood Work      1.714879
4               Yes                 No     No      CT Scan      1.254807

   TotalCharge  Additional_charges  Item1  Item2  Item3  Item4  Item5  Item6  \
0  3726.702860        17939.403420      3      3      2      2      4      3
1  4193.190458        17612.998120      3      4      3      4      4      4
2  2434.234222        17505.192460      2      4      4      4      3      4
3  2127.830423        12993.437350      3      5      5      3      4      5
4  2113.073274         3716.525786      2      1      3      3      5      3

   Item7  Item8
0      3      4
1      3      3
2      3      3
3      5      5
4      4      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')
CaseOrder             0
Customer_id           0
Interaction           0
UID                   0
City                  0
State                 0
County                0
Zip                   0
Lat                   0
Lng                   0
Population            0
Area                  0
TimeZone              0
Job                   0
Children              0
Age                   0
```

```
Income                0
Marital               0
Gender                0
ReAdmis               0
VitD_levels           0
Doc_visits            0
Full_meals_eaten      0
vitD_supp             0
Soft_drink            0
Initial_admin         0
HighBlood             0
Stroke                0
Complication_risk     0
Overweight            0
Arthritis             0
Diabetes              0
Hyperlipidemia        0
BackPain              0
Anxiety               0
Allergic_rhinitis     0
Reflux_esophagitis    0
Asthma                0
Services              0
Initial_days          0
TotalCharge           0
Additional_charges    0
Item1                 0
Item2                 0
Item3                 0
Item4                 0
Item5                 0
Item6                 0
Item7                 0
Item8                 0
dtype: int64
False
False
   Age ReAdmis  VitD_levels  Doc_visits  Full_meals_eaten  vitD_supp  \
0   53      No    19.141466           6                 0          0
1   51      No    18.940352           4                 2          1
2   53      No    18.057507           4                 1          0
3   78      No    16.576858           4                 1          0
4   22      No    17.439069           5                 0          2

  Soft_drink        Initial_admin HighBlood Stroke Complication_risk  \
0         No  Emergency Admission       Yes     No            Medium
1         No  Emergency Admission       Yes     No              High
2         No   Elective Admission       Yes     No            Medium
```

```
3         No   Elective Admission        No    Yes            Medium
4        Yes   Elective Admission        No     No               Low

  Overweight Arthritis Diabetes Hyperlipidemia BackPain Anxiety  \
0         No       Yes      Yes             No      Yes     Yes
1        Yes        No       No             No       No      No
2        Yes        No      Yes             No       No      No
3         No       Yes       No             No       No      No
4         No        No       No            Yes       No      No

  Allergic_rhinitis Reflux_esophagitis Asthma     Services  Initial_days  \
0               Yes                 No    Yes   Blood Work     10.585770
1                No                Yes     No  Intravenous     15.129562
2                No                 No     No   Blood Work      4.772177
3                No                Yes    Yes   Blood Work      1.714879
4               Yes                 No     No      CT Scan      1.254807

   TotalCharge  Additional_charges  Item1  Item2  Item3  Item4  Item5  Item6  \
0  3726.702860        17939.403420      3      3      2      2      4      3
1  4193.190458        17612.998120      3      4      3      4      4      4
2  2434.234222        17505.192460      2      4      4      4      3      4
3  2127.830423        12993.437350      3      5      5      3      4      5
4  2113.073274         3716.525786      2      1      3      3      5      3

   Item7  Item8
0      3      4
1      3      3
2      3      3
3      5      5
4      4      3
Index(['Age', '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')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 32 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 10000 non-null  int64
 1   ReAdmis             10000 non-null  object
 2   VitD_levels         10000 non-null  float64
 3   Doc_visits          10000 non-null  int64
 4   Full_meals_eaten    10000 non-null  int64
 5   vitD_supp           10000 non-null  int64  
 6   Soft_drink          10000 non-null  object 
 7   Initial_admin       10000 non-null  object 
 8   HighBlood           10000 non-null  object 
 9   Stroke              10000 non-null  object 
 10  Complication_risk   10000 non-null  object 
 11  Overweight          10000 non-null  object 
 12  Arthritis           10000 non-null  object 
 13  Diabetes            10000 non-null  object 
 14  Hyperlipidemia      10000 non-null  object 
 15  BackPain            10000 non-null  object 
 16  Anxiety             10000 non-null  object 
 17  Allergic_rhinitis   10000 non-null  object 
 18  Reflux_esophagitis  10000 non-null  object 
 19  Asthma              10000 non-null  object 
 20  Services            10000 non-null  object 
 21  Initial_days        10000 non-null  float64
 22  TotalCharge         10000 non-null  float64
 23  Additional_charges  10000 non-null  float64
 24  Item1               10000 non-null  int64  
 25  Item2               10000 non-null  int64  
 26  Item3               10000 non-null  int64  
 27  Item4               10000 non-null  int64  
 28  Item5               10000 non-null  int64  
 29  Item6               10000 non-null  int64  
 30  Item7               10000 non-null  int64  
 31  Item8               10000 non-null  int64  
dtypes: float64(4), int64(12), object(16)
memory usage: 2.4+ MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 36 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Age                                   10000 non-null  int64  
 1   VitD_levels                           10000 non-null  float64
 2   Doc_visits                            10000 non-null  int64  
 3   Full_meals_eaten                      10000 non-null  int64  
 4   vitD_supp                             10000 non-null  int64  
 5   Initial_days                          10000 non-null  float64
 6   TotalCharge                           10000 non-null  float64
 7   Additional_charges                    10000 non-null  float64
 8   Timely_admis                          10000 non-null  int64  
 9   Timely_treat                          10000 non-null  int64  
 10  Timely_visits                         10000 non-null  int64  
 11  Reliability                           10000 non-null  int64  
 12  Options                               10000 non-null  int64  
 13  Hours_treat                         10000 non-null  int64
 14  Courteous_staff                     10000 non-null  int64
 15  Active_listening                    10000 non-null  int64
 16  ReAdmis_Yes                         10000 non-null  uint8
 17  Complication_risk_Low               10000 non-null  uint8
 18  Complication_risk_Medium            10000 non-null  uint8
 19  Initial_admin_Emergency Admission   10000 non-null  uint8
 20  Initial_admin_Observation Admission  10000 non-null  uint8
 21  Services_CT Scan                    10000 non-null  uint8
 22  Services_Intravenous                10000 non-null  uint8
 23  Services_MRI                        10000 non-null  uint8
 24  Overweight_Yes                      10000 non-null  uint8
 25  Anxiety_Yes                         10000 non-null  uint8
 26  Arthritis_Yes                       10000 non-null  uint8
 27  Asthma_Yes                          10000 non-null  uint8
 28  Soft_drink_Yes                      10000 non-null  uint8
 29  Diabetes_Yes                        10000 non-null  uint8
 30  Allergic_rhinitis_Yes               10000 non-null  uint8
 31  BackPain_Yes                        10000 non-null  uint8
 32  Stroke_Yes                          10000 non-null  uint8
 33  HighBlood_Yes                       10000 non-null  uint8
 34  Hyperlipidemia_Yes                  10000 non-null  uint8
 35  Reflux_esophagitis_Yes              10000 non-null  uint8
dtypes: float64(4), int64(12), uint8(20)
memory usage: 1.4 MB
None
   Age  VitD_levels  Doc_visits  Full_meals_eaten  vitD_supp  Initial_days  \
0   53    19.141466           6                 0          0     10.585770
1   51    18.940352           4                 2          1     15.129562
2   53    18.057507           4                 1          0      4.772177
3   78    16.576858           4                 1          0      1.714879
4   22    17.439069           5                 0          2      1.254807

   TotalCharge  Additional_charges  Timely_admis  Timely_treat  Timely_visits  \
0  3726.702860        17939.403420             3             3              2
1  4193.190458        17612.998120             3             4              3
2  2434.234222        17505.192460             2             4              4
3  2127.830423        12993.437350             3             5              5
4  2113.073274         3716.525786             2             1              3

   Reliability  Options  Hours_treat  Courteous_staff  Active_listening  \
0            2        4            3                3                 4
1            4        4            4                3                 3
2            4        3            4                3                 3
3            3        4            5                5                 5
4            3        5            3                4                 3

   ReAdmis_Yes  Complication_risk_Low  Complication_risk_Medium  \
0            0                      0                         1
1            0                      0                         0
2            0                      0                         1
3            0                      0                         1
4            0                      1                         0

   Initial_admin_Emergency Admission  Initial_admin_Observation Admission  \
0                                  1                                    0
1                                  1                                    0
2                                  0                                    0
3                                  0                                    0
4                                  0                                    0

   Services_CT Scan  Services_Intravenous  Services_MRI  Overweight_Yes  \
0                 0                     0             0               0
1                 0                     1             0               1
2                 0                     0             0               1
3                 0                     0             0               0
4                 1                     0             0               0

   Anxiety_Yes  Arthritis_Yes  Asthma_Yes  Soft_drink_Yes  Diabetes_Yes  \
0            1              1           1               0             1
1            0              0           0               0             0
2            0              0           0               0             1
3            0              1           1               0             0
4            0              0           0               1             0

   Allergic_rhinitis_Yes  BackPain_Yes  Stroke_Yes  HighBlood_Yes  \
0                      1             1           0              1
1                      0             0           0              1
2                      0             0           0              1
3                      0             0           1              0
4                      1             0           0              0

   Hyperlipidemia_Yes  Reflux_esophagitis_Yes
0                   0                       0
1                   0                       1
2                   0                       0
3                   0                       1
4                   1                       0
              Age   VitD_levels    Doc_visits  Full_meals_eaten  \
count  10000.000000  10000.000000  10000.000000      10000.000000
mean      53.511700     17.964262      5.012200          1.001400
std       20.638538      2.017231      1.045734          1.008117
min       18.000000      9.806483      1.000000          0.000000
25%       36.000000     16.626439      4.000000          0.000000
50%       53.000000     17.951122      5.000000          1.000000
75%       71.000000     19.347963      6.000000          2.000000
```

```
max        89.000000    26.394449     9.000000          7.000000

         vitD_supp  Initial_days   TotalCharge  Additional_charges  \
count  10000.000000  10000.000000  10000.000000        10000.000000
mean       0.398900     34.455299   5312.172769        12934.528587
std        0.628505     26.309341   2180.393838         6542.601544
min        0.000000      1.001981   1938.312067         3125.703000
25%        0.000000      7.896215   3179.374015         7986.487755
50%        0.000000     35.836244   5213.952000        11573.977735
75%        1.000000     61.161020   7459.699750        15626.490000
max        5.000000     71.981490   9180.728000        30566.070000

       Timely_admis  Timely_treat  Timely_visits   Reliability       Options  \
count  10000.000000  10000.000000   10000.000000  10000.000000  10000.000000
mean       3.518800      3.506700       3.511100      3.515100      3.496900
std        1.031966      1.034825       1.032755      1.036282      1.030192
min        1.000000      1.000000       1.000000      1.000000      1.000000
25%        3.000000      3.000000       3.000000      3.000000      3.000000
50%        4.000000      3.000000       4.000000      4.000000      3.000000
75%        4.000000      4.000000       4.000000      4.000000      4.000000
max        8.000000      7.000000       8.000000      7.000000      7.000000

        Hours_treat  Courteous_staff  Active_listening   ReAdmis_Yes  \
count  10000.000000     10000.000000      10000.000000  10000.000000
mean       3.522500         3.494000          3.509700      0.366900
std        1.032376         1.021405          1.042312      0.481983
min        1.000000         1.000000          1.000000      0.000000
25%        3.000000         3.000000          3.000000      0.000000
50%        4.000000         3.000000          3.000000      0.000000
75%        4.000000         4.000000          4.000000      1.000000
max        7.000000         7.000000          7.000000      1.000000

       Complication_risk_Low  Complication_risk_Medium  \
count           10000.000000              10000.000000
mean                0.212500                  0.451700
std                 0.409097                  0.497687
min                 0.000000                  0.000000
25%                 0.000000                  0.000000
50%                 0.000000                  0.000000
75%                 0.000000                  1.000000
max                 1.000000                  1.000000

       Initial_admin_Emergency Admission  Initial_admin_Observation Admission  \
count                       10000.000000                         10000.000000
mean                            0.506000                             0.243600
std                             0.499989                             0.429276
min                             0.000000                             0.000000
25%                             0.000000                             0.000000
```

```
50%                             1.000000                             0.000000
75%                             1.000000                             0.000000
max                             1.000000                             1.000000

       Services_CT Scan  Services_Intravenous  Services_MRI  Overweight_Yes  \
count      10000.000000          10000.000000  10000.000000    10000.000000
mean           0.122500              0.313000      0.038000        0.709400
std            0.327879              0.463738      0.191206        0.454062
min            0.000000              0.000000      0.000000        0.000000
25%            0.000000              0.000000      0.000000        0.000000
50%            0.000000              0.000000      0.000000        1.000000
75%            0.000000              1.000000      0.000000        1.000000
max            1.000000              1.000000      1.000000        1.000000

        Anxiety_Yes  Arthritis_Yes   Asthma_Yes  Soft_drink_Yes  Diabetes_Yes  \
count  10000.000000   10000.000000  10000.00000    10000.000000   10000.00000
mean       0.321500       0.357400      0.28930        0.257500       0.27380
std        0.467076       0.479258      0.45346        0.437279       0.44593
min        0.000000       0.000000      0.00000        0.000000       0.00000
25%        0.000000       0.000000      0.00000        0.000000       0.00000
50%        0.000000       0.000000      0.00000        0.000000       0.00000
75%        1.000000       1.000000      1.00000        1.000000       1.00000
max        1.000000       1.000000      1.00000        1.000000       1.00000

       Allergic_rhinitis_Yes  BackPain_Yes    Stroke_Yes  HighBlood_Yes  \
count           10000.000000  10000.000000  10000.000000   10000.000000
mean                0.394100      0.411400      0.199300       0.409000
std                 0.488681      0.492112      0.399494       0.491674
min                 0.000000      0.000000      0.000000       0.000000
25%                 0.000000      0.000000      0.000000       0.000000
50%                 0.000000      0.000000      0.000000       0.000000
75%                 1.000000      1.000000      0.000000       1.000000
max                 1.000000      1.000000      1.000000       1.000000

       Hyperlipidemia_Yes  Reflux_esophagitis_Yes
count        10000.000000            10000.000000
mean             0.337200                0.413500
std              0.472777                0.492486
min              0.000000                0.000000
25%              0.000000                0.000000
50%              0.000000                0.000000
75%              1.000000                1.000000
max              1.000000                1.000000
Index(['Age', 'VitD_levels', 'Doc_visits', 'Full_meals_eaten', 'vitD_supp',
       'Initial_days', 'TotalCharge', 'Additional_charges', 'Timely_admis',
       'Timely_treat', 'Timely_visits', 'Reliability', 'Options',
       'Hours_treat', 'Courteous_staff', 'Active_listening', 'ReAdmis_Yes',
       'Complication_risk_Low', 'Complication_risk_Medium',
       'Initial_admin_Emergency Admission',
       'Initial_admin_Observation Admission', 'Services_CT Scan',
       'Services_Intravenous', 'Services_MRI', 'Overweight_Yes', 'Anxiety_Yes',
       'Arthritis_Yes', 'Asthma_Yes', 'Soft_drink_Yes', 'Diabetes_Yes',
       'Allergic_rhinitis_Yes', 'BackPain_Yes', 'Stroke_Yes', 'HighBlood_Yes',
       'Hyperlipidemia_Yes', 'Reflux_esophagitis_Yes'],
      dtype='object')
```

### C5. Prepared Data CSV Attached prepared data csv as: prepared-data.csv

[4]:

```
data.to_csv('prepared-data.csv')
```

## D. Model Comparison and Analysis

The initial regression model will be run on the predictor variables mentioned above, and compared with the Initial_days variable as the target. Observations removed from the initial model were minor patient demographics (Marital, Gender) and some small medical observations such as VitD_levels and vitD_supp. The OLS Regression results can be found in the codeblock below with further analysis continued:

[5]:

```
import statsmodels.api as sm

columns = ['Age', 'Doc_visits', 'Full_meals_eaten',
       'Initial_days', 'TotalCharge', 'Additional_charges', 'Timely_admis',
       'Timely_treat', 'Timely_visits', 'Reliability', 'Options',
       'Hours_treat', 'Courteous_staff', 'Active_listening', 'ReAdmis_Yes',
       'Complication_risk_Low', 'Complication_risk_Medium',
       'Initial_admin_Emergency Admission',
       'Initial_admin_Observation Admission', 'Services_CT Scan',
       'Services_Intravenous', 'Services_MRI', 'Overweight_Yes', 'Anxiety_Yes',
       'Arthritis_Yes', 'Asthma_Yes', 'Diabetes_Yes',
       'Allergic_rhinitis_Yes', 'BackPain_Yes', 'Stroke_Yes', 'HighBlood_Yes',
       'Hyperlipidemia_Yes', 'Reflux_esophagitis_Yes']

data = data[columns]

X = data.drop('Initial_days', axis=1)
X = sm.add_constant(X)

# our dependent variable
y = data['Initial_days']

# initial ols model
model = sm.OLS(y, X).fit()

# summary of the initial model
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 2.116e+16
Date:                Tue, 09 Jan 2024   Prob (F-statistic):               0.00
Time:                        23:27:30   Log-Likelihood:             1.1236e+05
No. Observations:               10000   AIC:                        -2.247e+05
Df Residuals:                    9967   BIC:                        -2.244e+05
Df Model:                          32
Covariance Type:            nonrobust
================================================================================
=======================
                                          coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
-----------------------
const                                 -27.6939   3.74e-07  -7.41e+07      0.000
-27.694     -27.694
Age                                  6.291e-09   4.71e-09      1.335      0.182
-2.94e-09    1.55e-08
Doc_visits                           1.232e-08   3.06e-08      0.402      0.688
-4.77e-08    7.24e-08
Full_meals_eaten                    -4.278e-08   3.18e-08     -1.347      0.178
-1.05e-07    1.95e-08
TotalCharge                             0.0122   2.84e-11    4.3e+08      0.000
0.012       0.012
Additional_charges                  -3.336e-11   1.97e-11     -1.692      0.091
-7.2e-11    5.28e-12
Timely_admis                        -3.225e-09   4.61e-08     -0.070      0.944
-9.36e-08    8.71e-08
Timely_treat                        -6.709e-08   4.25e-08     -1.578      0.115
-1.5e-07    1.63e-08
Timely_visits                         5.85e-08   3.93e-08      1.490      0.136
-1.84e-08    1.35e-07
Reliability                         -2.642e-09    3.5e-08     -0.076      0.940
-7.12e-08    6.59e-08
Options                              3.107e-08   3.68e-08      0.844      0.399
-4.11e-08    1.03e-07
Hours_treat                          1.482e-08    3.8e-08      0.390      0.697
-5.97e-08    8.93e-08
Courteous_staff                     -1.456e-08   3.58e-08     -0.407      0.684
-8.48e-08    5.56e-08
Active_listening                     5.151e-09   3.37e-08      0.153      0.879
-6.09e-08    7.12e-08
ReAdmis_Yes                          1.777e-07   1.27e-07      1.402      0.161
-7.08e-08    4.26e-07
Complication_risk_Low                   5.0464   8.99e-08   5.61e+07      0.000
5.046       5.046
Complication_risk_Medium                5.0464   7.44e-08   6.78e+07      0.000
```

```
5.046       5.046
Initial_admin_Emergency Admission      -6.2526   7.95e-08  -7.87e+07      0.000
-6.253      -6.253
Initial_admin_Observation Admission  1.235e-08   9.12e-08      0.135      0.892
-1.66e-07    1.91e-07
Services_CT Scan                     2.916e-08   1.02e-07      0.287      0.774
-1.7e-07    2.28e-07
Services_Intravenous                 5.913e-08   7.23e-08      0.818      0.413
-8.25e-08    2.01e-07
Services_MRI                         1.424e-07    1.7e-07      0.838      0.402
-1.91e-07    4.76e-07
Overweight_Yes                       4.336e-09   7.05e-08      0.061      0.951
-1.34e-07    1.43e-07
Anxiety_Yes                            -1.0510   6.86e-08  -1.53e+07      0.000
-1.051      -1.051
Arthritis_Yes                          -0.8781   6.69e-08  -1.31e+07      0.000
-0.878      -0.878
Asthma_Yes                          -4.627e-08   7.06e-08     -0.655      0.512
-1.85e-07    9.22e-08
Diabetes_Yes                           -0.9178   7.19e-08  -1.28e+07      0.000
-0.918      -0.918
Allergic_rhinitis_Yes                  -0.7394   6.56e-08  -1.13e+07      0.000
-0.739      -0.739
BackPain_Yes                           -1.0392   6.52e-08  -1.59e+07      0.000
-1.039      -1.039
Stroke_Yes                           3.639e-08   8.04e-08      0.452      0.651
-1.21e-07    1.94e-07
HighBlood_Yes                          -1.3708   1.82e-07  -7.52e+06      0.000
-1.371      -1.371
Hyperlipidemia_Yes                     -1.1471   6.78e-08  -1.69e+07      0.000
-1.147      -1.147
Reflux_esophagitis_Yes                 -0.7284   6.51e-08  -1.12e+07      0.000
-0.728      -0.728
==============================================================================
Omnibus:                      405.752   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1309.867
Skew:                           0.013   Prob(JB):                    3.68e-285
Kurtosis:                       4.773   Cond. No.                     1.81e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.81e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Based on the above OLS regression output we find the initial model equation:

$\hat{Y}$ = -27.6939 + (6.291e-9 * Age) + (1.232e-8 * Doc_visits) - (4.278e-8 * Full_meals_eaten)

+ 0.0122 * TotalCharge - (3.336e-11 * Additional_charges) - (3.225e-9 * Timely_admis) - (6.709e-8 * Timely_treat) + (5.85e-8 * Timely_visits) - (2.642e-9 * Reliability) + (3.107e-8 * Options) + (1.482e-8 * Hours_treat) - (1.456e-8 * Courteous_staff) + (5.151e-9 * Active_listening) + (1.777e-7 * ReAdmis_Yes) + 5.0464 * Complication_risk_Low + 5.0464 * Complication_risk_Medium - 6.2526 * Initial_admin_Emergency_Admission + (1.235e-8 * Initial_admin_Observation_Admission) + (2.916e-8 * Services_CT_Scan) + (5.913e-8 * Services_Intravenous) + (1.424e-7 * Services_MRI) + (4.336e-9 * Overweight_Yes) - 1.0510 * Anxiety_Yes - 0.8781 * Arthritis_Yes - (4.627e-8 * Asthma_Yes) - 0.9178 * Diabetes_Yes - 0.7394 * Allergic_rhinitis_Yes - 1.0392 * BackPain_Yes + (3.639e-8 * Stroke_Yes) - 1.3708 * HighBlood_Yes - 1.1471 * Hyperlipidemia_Yes - 0.7284 * Reflux_esophagitis_Yes

The initial model had an R-squared value of 1.00 which means 100% of the variation can be explained by this model. As mentioned in Dr.Middleton's Webinar, the Prob (F-statistic) value being 0.00 typically indicates that the regression model is statistically significant. The model above also tells us the condition number is large at 1.81e+05, this may indicate strong multicollinearity. To find where there may be multicollinearity, a correlation matrix and heatmap will be used. By using these tools, we can further narrow our model down.

### D2. Model Evaluation Metric

We will first generate a correlation matrix of our independent variables. Then we will create a heatmap of these values along with our dependent variable. Lastly for each independent variable we will get there variance inflation factors and their p-values. These strategies will help with narrowing our model down further:

[9]:

```python
from statsmodels.stats.outliers_influence import variance_inflation_factor
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.api as sm

data1 = pd.read_csv('prepared-data.csv')
pd.set_option('display.max_columns', None)

columns = ['Age', 'Doc_visits', 'Full_meals_eaten', 'Initial_days',
           'TotalCharge', 'Additional_charges', 'Timely_admis', 'Timely_treat',
           'Timely_visits', 'Reliability', 'Options', 'Hours_treat',
           'Courteous_staff', 'Active_listening', 'ReAdmis_Yes',
           'Complication_risk_Low', 'Complication_risk_Medium',
           'Initial_admin_Emergency Admission', 'Initial_admin_Observation␣
 ↪Admission',
           'Services_CT Scan', 'Services_Intravenous', 'Services_MRI',
           'Overweight_Yes', 'Anxiety_Yes', 'Arthritis_Yes', 'Asthma_Yes',
           'Diabetes_Yes', 'Allergic_rhinitis_Yes', 'BackPain_Yes',␣
 ↪'Stroke_Yes',
           'HighBlood_Yes', 'Hyperlipidemia_Yes', 'Reflux_esophagitis_Yes']

data1 = data1[columns]
```

```python
X = data1.drop('Initial_days', axis=1)

X_with_target = data1

correlation_matrix = X.corr()
print(correlation_matrix)
```

```
                                          Age  Doc_visits  Full_meals_eaten  \
Age                                  1.000000    0.006898          0.008555   
Doc_visits                           0.006898    1.000000         -0.002767   
Full_meals_eaten                     0.008555   -0.002767          1.000000   
TotalCharge                          0.016876   -0.005043         -0.014306   
Additional_charges                   0.716854    0.008072          0.018763   
Timely_admis                         0.005552    0.003680          0.003724   
Timely_treat                         0.003967    0.006024         -0.002022   
Timely_visits                        0.004709   -0.002718          0.008246   
Reliability                          0.003377   -0.006538         -0.009019   
Options                             -0.008827   -0.009434          0.009538   
Hours_treat                         -0.002087    0.012530          0.004294   
Courteous_staff                      0.009423    0.008589          0.004087   
Active_listening                    -0.003367    0.004571         -0.018382   
ReAdmis_Yes                          0.015810    0.000246         -0.012172   
Complication_risk_Low                0.001085   -0.006061         -0.012119   
Complication_risk_Medium            -0.006021   -0.008091         -0.003254   
Initial_admin_Emergency Admission   -0.004538    0.003686          0.006333   
Initial_admin_Observation Admission -0.008336    0.015658          0.004527   
Services_CT Scan                     0.009506    0.014600         -0.002939   
Services_Intravenous                 0.004142   -0.008700          0.016177   
Services_MRI                         0.008529   -0.012822         -0.018954   
Overweight_Yes                      -0.008292    0.011890         -0.008287   
Anxiety_Yes                          0.006130   -0.001684          0.008602   
Arthritis_Yes                        0.007110   -0.000719          0.011591   
Asthma_Yes                           0.009229   -0.017989          0.012459   
Diabetes_Yes                         0.003694    0.012781          0.009603   
Allergic_rhinitis_Yes                0.012092    0.002920          0.015120   
BackPain_Yes                         0.021081    0.008514         -0.015676   
Stroke_Yes                           0.012035   -0.002230          0.002784   
HighBlood_Yes                        0.007147    0.008967          0.014784   
Hyperlipidemia_Yes                   0.003736   -0.026730          0.000688   
Reflux_esophagitis_Yes              -0.019609   -0.005330         -0.000562   

                                     TotalCharge  Additional_charges  \
Age                                     0.016876            0.716854   
Doc_visits                             -0.005043            0.008072   
Full_meals_eaten                       -0.014306            0.018763   
TotalCharge                             1.000000            0.029256   
Additional_charges                      0.029256            1.000000   
Timely_admis                           -0.019706            0.002423   
```

```
Timely_treat                           -0.006055            0.002815
Timely_visits                          -0.009051           -0.004422
Reliability                            -0.010318           -0.000771
Options                                 0.003532           -0.014323
Hours_treat                            -0.010480           -0.000448
Courteous_staff                         0.004556            0.015209
Active_listening                       -0.008250           -0.000467
ReAdmis_Yes                             0.843726            0.013620
Complication_risk_Low                  -0.013344           -0.035234
Complication_risk_Medium               -0.068781           -0.009418
Initial_admin_Emergency Admission       0.106985            0.034762
Initial_admin_Observation Admission    -0.066870           -0.029231
Services_CT Scan                        0.010561            0.013137
Services_Intravenous                   -0.016170           -0.001095
Services_MRI                            0.007341            0.010134
Overweight_Yes                         -0.012782            0.012771
Anxiety_Yes                             0.031199            0.011666
Arthritis_Yes                           0.032932            0.004788
Asthma_Yes                             -0.014290            0.014083
Diabetes_Yes                            0.011524            0.002450
Allergic_rhinitis_Yes                   0.018919            0.016154
BackPain_Yes                            0.035828            0.014245
Stroke_Yes                             -0.003694            0.035140
HighBlood_Yes                           0.019910            0.654316
Hyperlipidemia_Yes                      0.017565           -0.002475
Reflux_esophagitis_Yes                  0.026284           -0.011405

                                     Timely_admis  Timely_treat  \
Age                                      0.005552      0.003967
Doc_visits                               0.003680      0.006024
Full_meals_eaten                         0.003724     -0.002022
TotalCharge                             -0.019706     -0.006055
Additional_charges                       0.002423      0.002815
Timely_admis                             1.000000      0.655578
Timely_treat                             0.655578      1.000000
Timely_visits                            0.579585      0.521728
Reliability                             -0.004614      0.003077
Options                                 -0.000368     -0.010018
Hours_treat                              0.421233      0.366075
Courteous_staff                          0.332855      0.291039
Active_listening                         0.278067      0.242962
ReAdmis_Yes                             -0.016785     -0.002423
Complication_risk_Low                   -0.011477     -0.000883
Complication_risk_Medium                 0.000697      0.001405
Initial_admin_Emergency Admission        0.011605      0.015289
Initial_admin_Observation Admission     -0.015080     -0.012455
Services_CT Scan                        -0.012275     -0.005809
Services_Intravenous                    -0.012715     -0.013123
```

```
Services_MRI                             0.002968      0.006800
Overweight_Yes                           0.002056     -0.001177
Anxiety_Yes                             -0.007458     -0.009733
Arthritis_Yes                           -0.008532     -0.012492
Asthma_Yes                              -0.011303     -0.007648
Diabetes_Yes                             0.013806      0.005994
Allergic_rhinitis_Yes                    0.009402      0.014654
BackPain_Yes                            -0.011687     -0.005413
Stroke_Yes                               0.001948     -0.007706
HighBlood_Yes                           -0.011017     -0.007745
Hyperlipidemia_Yes                       0.019393      0.000697
Reflux_esophagitis_Yes                   0.011367      0.017425

                                     Timely_visits  Reliability   Options  \
Age                                       0.004709     0.003377 -0.008827
Doc_visits                               -0.002718    -0.006538 -0.009434
Full_meals_eaten                          0.008246    -0.009019  0.009538
TotalCharge                              -0.009051    -0.010318  0.003532
Additional_charges                       -0.004422    -0.000771 -0.014323
Timely_admis                              0.579585    -0.004614 -0.000368
Timely_treat                              0.521728     0.003077 -0.010018
Timely_visits                             1.000000    -0.006324 -0.010496
Reliability                              -0.006324     1.000000 -0.447372
Options                                  -0.010496    -0.447372  1.000000
Hours_treat                               0.312874     0.235444 -0.310154
Courteous_staff                           0.252302     0.192223 -0.268186
Active_listening                          0.209498     0.161528 -0.229557
ReAdmis_Yes                              -0.011699    -0.001983  0.005614
Complication_risk_Low                    -0.023929    -0.010047  0.009038
Complication_risk_Medium                  0.008826    -0.001688  0.011217
Initial_admin_Emergency Admission         0.012944    -0.002974  0.015472
Initial_admin_Observation Admission      -0.017605     0.002297 -0.024299
Services_CT Scan                         -0.018045    -0.011773  0.017853
Services_Intravenous                     -0.019993     0.002651 -0.012204
Services_MRI                              0.012045    -0.004915  0.014307
Overweight_Yes                           -0.002505    -0.003001  0.005771
Anxiety_Yes                              -0.004807     0.005363 -0.016946
Arthritis_Yes                            -0.020948     0.001416  0.002447
Asthma_Yes                                0.000296    -0.004721  0.019796
Diabetes_Yes                             -0.007469    -0.005052  0.017958
Allergic_rhinitis_Yes                     0.001140    -0.007704 -0.004029
BackPain_Yes                             -0.015480     0.012723  0.006659
Stroke_Yes                                0.001304    -0.013430  0.005025
HighBlood_Yes                            -0.015244    -0.004075 -0.013292
Hyperlipidemia_Yes                        0.018551     0.020022 -0.009352
Reflux_esophagitis_Yes                   -0.005584     0.016277 -0.011764

                                     Hours_treat  Courteous_staff  \
Age                                    -0.002087         0.009423
Doc_visits                              0.012530         0.008589
Full_meals_eaten                        0.004294         0.004087
TotalCharge                            -0.010480         0.004556
Additional_charges                     -0.000448         0.015209
Timely_admis                            0.421233         0.332855
Timely_treat                            0.366075         0.291039
Timely_visits                           0.312874         0.252302
Reliability                             0.235444         0.192223
Options                                -0.310154        -0.268186
Hours_treat                             1.000000         0.377368
Courteous_staff                         0.377368         1.000000
Active_listening                        0.319886         0.274499
ReAdmis_Yes                            -0.016894        -0.004974
Complication_risk_Low                  -0.004100         0.002932
Complication_risk_Medium                0.008344        -0.020342
Initial_admin_Emergency Admission       0.017467         0.013583
Initial_admin_Observation Admission    -0.008758        -0.014229
Services_CT Scan                       -0.025723        -0.018858
Services_Intravenous                    0.008058        -0.009126
Services_MRI                           -0.010412        -0.006002
Overweight_Yes                         -0.003118        -0.009151
Anxiety_Yes                            -0.002248         0.003520
Arthritis_Yes                          -0.018074        -0.008286
Asthma_Yes                             -0.009740        -0.013202
Diabetes_Yes                           -0.004259        -0.004737
Allergic_rhinitis_Yes                  -0.012721         0.008445
BackPain_Yes                           -0.016056        -0.015781
Stroke_Yes                              0.004282        -0.005280
HighBlood_Yes                          -0.002369         0.007277
Hyperlipidemia_Yes                      0.017648         0.016202
Reflux_esophagitis_Yes                  0.009729        -0.013657

                                     Active_listening  ReAdmis_Yes  \
Age                                         -0.003367     0.015810
Doc_visits                                   0.004571     0.000246
Full_meals_eaten                            -0.018382    -0.012172
TotalCharge                                 -0.008250     0.843726
Additional_charges                          -0.000467     0.013620
Timely_admis                                 0.278067    -0.016785
Timely_treat                                 0.242962    -0.002423
Timely_visits                                0.209498    -0.011699
Reliability                                  0.161528    -0.001983
Options                                     -0.229557     0.005614
Hours_treat                                  0.319886    -0.016894
Courteous_staff                              0.274499    -0.004974
Active_listening                             1.000000    -0.016740
ReAdmis_Yes                                 -0.016740     1.000000
```

```
Complication_risk_Low                       -0.007297     0.001186
Complication_risk_Medium                     0.000518     0.002799
Initial_admin_Emergency Admission           -0.008268     0.019707
Initial_admin_Observation Admission          0.013270    -0.011972
Services_CT Scan                            -0.011525     0.024395
Services_Intravenous                         0.005305    -0.020313
Services_MRI                                -0.006868     0.009309
Overweight_Yes                               0.009549    -0.008586
Anxiety_Yes                                  0.014650     0.002406
Arthritis_Yes                                0.002869     0.007663
Asthma_Yes                                   0.002209    -0.017133
Diabetes_Yes                                -0.014752    -0.003058
Allergic_rhinitis_Yes                        0.005355    -0.004651
BackPain_Yes                                 0.000213     0.013313
Stroke_Yes                                   0.000040     0.000918
HighBlood_Yes                                0.002601     0.002270
Hyperlipidemia_Yes                          -0.001970     0.004307
Reflux_esophagitis_Yes                      -0.003236     0.005422

                                     Complication_risk_Low  \
Age                                               0.001085
Doc_visits                                       -0.006061
Full_meals_eaten                                 -0.012119
TotalCharge                                      -0.013344
Additional_charges                               -0.035234
Timely_admis                                     -0.011477
Timely_treat                                     -0.000883
Timely_visits                                    -0.023929
Reliability                                      -0.010047
Options                                           0.009038
Hours_treat                                      -0.004100
Courteous_staff                                   0.002932
Active_listening                                 -0.007297
ReAdmis_Yes                                       0.001186
Complication_risk_Low                             1.000000
Complication_risk_Medium                         -0.471487
Initial_admin_Emergency Admission                 0.009168
Initial_admin_Observation Admission              -0.001509
Services_CT Scan                                 -0.015145
Services_Intravenous                              0.002570
Services_MRI                                      0.004155
Overweight_Yes                                   -0.009947
Anxiety_Yes                                      -0.002192
Arthritis_Yes                                     0.002818
Asthma_Yes                                        0.003902
Diabetes_Yes                                      0.006675
Allergic_rhinitis_Yes                             0.013276
BackPain_Yes                                      0.019759
```

```
Stroke_Yes                                       -0.001537
HighBlood_Yes                                    -0.027906
Hyperlipidemia_Yes                               -0.005972
Reflux_esophagitis_Yes                            0.000652

                                     Complication_risk_Medium  \
Age                                                 -0.006021
Doc_visits                                          -0.008091
Full_meals_eaten                                    -0.003254
TotalCharge                                         -0.068781
Additional_charges                                  -0.009418
Timely_admis                                         0.000697
Timely_treat                                         0.001405
Timely_visits                                        0.008826
Reliability                                         -0.001688
Options                                              0.011217
Hours_treat                                          0.008344
Courteous_staff                                     -0.020342
Active_listening                                     0.000518
ReAdmis_Yes                                          0.002799
Complication_risk_Low                               -0.471487
Complication_risk_Medium                             1.000000
Initial_admin_Emergency Admission                   -0.010290
Initial_admin_Observation Admission                  0.023246
Services_CT Scan                                     0.002861
Services_Intravenous                                -0.005122
Services_MRI                                         0.011933
Overweight_Yes                                       0.018871
Anxiety_Yes                                          0.004640
Arthritis_Yes                                        0.017453
Asthma_Yes                                           0.006750
Diabetes_Yes                                        -0.001241
Allergic_rhinitis_Yes                               -0.017744
BackPain_Yes                                        -0.009920
Stroke_Yes                                           0.000886
HighBlood_Yes                                        0.014528
Hyperlipidemia_Yes                                   0.010995
Reflux_esophagitis_Yes                              -0.005622

                                     Initial_admin_Emergency Admission  \
Age                                                          -0.004538
Doc_visits                                                    0.003686
Full_meals_eaten                                              0.006333
TotalCharge                                                   0.106985
Additional_charges                                            0.034762
Timely_admis                                                  0.011605
Timely_treat                                                  0.015289
Timely_visits                                                 0.012944
```

```
Reliability                                                   -0.002974
Options                                                        0.015472
Hours_treat                                                    0.017467
Courteous_staff                                                0.013583
Active_listening                                              -0.008268
ReAdmis_Yes                                                    0.019707
Complication_risk_Low                                          0.009168
Complication_risk_Medium                                      -0.010290
Initial_admin_Emergency Admission                              1.000000
Initial_admin_Observation Admission                           -0.574347
Services_CT Scan                                               0.007412
Services_Intravenous                                          -0.003787
Services_MRI                                                   0.009122
Overweight_Yes                                                -0.009940
Anxiety_Yes                                                    0.008655
Arthritis_Yes                                                 -0.000603
Asthma_Yes                                                    -0.005672
Diabetes_Yes                                                  -0.008266
Allergic_rhinitis_Yes                                          0.006080
BackPain_Yes                                                   0.000535
Stroke_Yes                                                    -0.009743
HighBlood_Yes                                                 -0.001440
Hyperlipidemia_Yes                                             0.018941
Reflux_esophagitis_Yes                                        -0.000126

                                     Initial_admin_Observation Admission  \
Age                                                             -0.008336
Doc_visits                                                       0.015658
Full_meals_eaten                                                 0.004527
TotalCharge                                                     -0.066870
Additional_charges                                              -0.029231
Timely_admis                                                    -0.015080
Timely_treat                                                    -0.012455
Timely_visits                                                   -0.017605
Reliability                                                      0.002297
Options                                                         -0.024299
Hours_treat                                                     -0.008758
Courteous_staff                                                 -0.014229
Active_listening                                                 0.013270
ReAdmis_Yes                                                     -0.011972
Complication_risk_Low                                           -0.001509
Complication_risk_Medium                                         0.023246
Initial_admin_Emergency Admission                               -0.574347
Initial_admin_Observation Admission                              1.000000
Services_CT Scan                                                -0.019476
Services_Intravenous                                             0.010817
Services_MRI                                                    -0.010440
Overweight_Yes                                                   0.009698
```

```
Anxiety_Yes                                                    -0.004077
Arthritis_Yes                                                   0.000182
Asthma_Yes                                                      0.001678
Diabetes_Yes                                                    0.000535
Allergic_rhinitis_Yes                                          -0.025280
BackPain_Yes                                                    0.009861
Stroke_Yes                                                      0.005543
HighBlood_Yes                                                   0.006006
Hyperlipidemia_Yes                                             -0.009077
Reflux_esophagitis_Yes                                         -0.008650

                                     Services_CT Scan  Services_Intravenous  \
Age                                          0.009506              0.004142
Doc_visits                                   0.014600             -0.008700
Full_meals_eaten                            -0.002939              0.016177
TotalCharge                                  0.010561             -0.016170
Additional_charges                           0.013137             -0.001095
Timely_admis                                -0.012275             -0.012715
Timely_treat                                -0.005809             -0.013123
Timely_visits                               -0.018045             -0.019993
Reliability                                 -0.011773              0.002651
Options                                      0.017853             -0.012204
Hours_treat                                 -0.025723              0.008058
Courteous_staff                             -0.018858             -0.009126
Active_listening                            -0.011525              0.005305
ReAdmis_Yes                                  0.024395             -0.020313
Complication_risk_Low                       -0.015145              0.002570
Complication_risk_Medium                     0.002861             -0.005122
Initial_admin_Emergency Admission            0.007412             -0.003787
Initial_admin_Observation Admission         -0.019476              0.010817
Services_CT Scan                             1.000000             -0.252196
Services_Intravenous                        -0.252196              1.000000
Services_MRI                                -0.074259             -0.134152
Overweight_Yes                               0.002005              0.004074
Anxiety_Yes                                 -0.005771              0.007251
Arthritis_Yes                                0.000754             -0.001198
Asthma_Yes                                   0.013862             -0.013559
Diabetes_Yes                                 0.014087              0.001937
Allergic_rhinitis_Yes                        0.007008             -0.003766
BackPain_Yes                                 0.016757             -0.013446
Stroke_Yes                                   0.013635             -0.019871
HighBlood_Yes                                0.011772             -0.008408
Hyperlipidemia_Yes                           0.000600             -0.003848
Reflux_esophagitis_Yes                       0.017628             -0.022007

                                     Services_MRI  Overweight_Yes  \
Age                                      0.008529       -0.008292
Doc_visits                              -0.012822        0.011890
```

```
Full_meals_eaten                        -0.018954       -0.008287
TotalCharge                              0.007341       -0.012782
Additional_charges                       0.010134        0.012771
Timely_admis                             0.002968        0.002056
Timely_treat                             0.006800       -0.001177
Timely_visits                            0.012045       -0.002505
Reliability                             -0.004915       -0.003001
Options                                  0.014307        0.005771
Hours_treat                             -0.010412       -0.003118
Courteous_staff                         -0.006002       -0.009151
Active_listening                        -0.006868        0.009549
ReAdmis_Yes                              0.009309       -0.008586
Complication_risk_Low                    0.004155       -0.009947
Complication_risk_Medium                 0.011933        0.018871
Initial_admin_Emergency Admission        0.009122       -0.009940
Initial_admin_Observation Admission     -0.010440        0.009698
Services_CT Scan                        -0.074259        0.002005
Services_Intravenous                    -0.134152        0.004074
Services_MRI                             1.000000       -0.002963
Overweight_Yes                          -0.002963        1.000000
Anxiety_Yes                             -0.006909       -0.011186
Arthritis_Yes                           -0.004160        0.003954
Asthma_Yes                              -0.001077        0.013943
Diabetes_Yes                             0.018715       -0.007575
Allergic_rhinitis_Yes                    0.000259        0.002819
BackPain_Yes                            -0.000353        0.010083
Stroke_Yes                              -0.003580       -0.001011
HighBlood_Yes                            0.001681        0.026231
Hyperlipidemia_Yes                      -0.002363       -0.006102
Reflux_esophagitis_Yes                   0.001986       -0.012240

                                     Anxiety_Yes  Arthritis_Yes  Asthma_Yes  \
Age                                     0.006130       0.007110    0.009229
Doc_visits                             -0.001684      -0.000719   -0.017989
Full_meals_eaten                        0.008602       0.011591    0.012459
TotalCharge                             0.031199       0.032932   -0.014290
Additional_charges                      0.011666       0.004788    0.014083
Timely_admis                           -0.007458      -0.008532   -0.011303
Timely_treat                           -0.009733      -0.012492   -0.007648
Timely_visits                          -0.004807      -0.020948    0.000296
Reliability                             0.005363       0.001416   -0.004721
Options                                -0.016946       0.002447    0.019796
Hours_treat                            -0.002248      -0.018074   -0.009740
Courteous_staff                         0.003520      -0.008286   -0.013202
Active_listening                        0.014650       0.002869    0.002209
ReAdmis_Yes                             0.002406       0.007663   -0.017133
Complication_risk_Low                  -0.002192       0.002818    0.003902
Complication_risk_Medium                0.004640       0.017453    0.006750
```

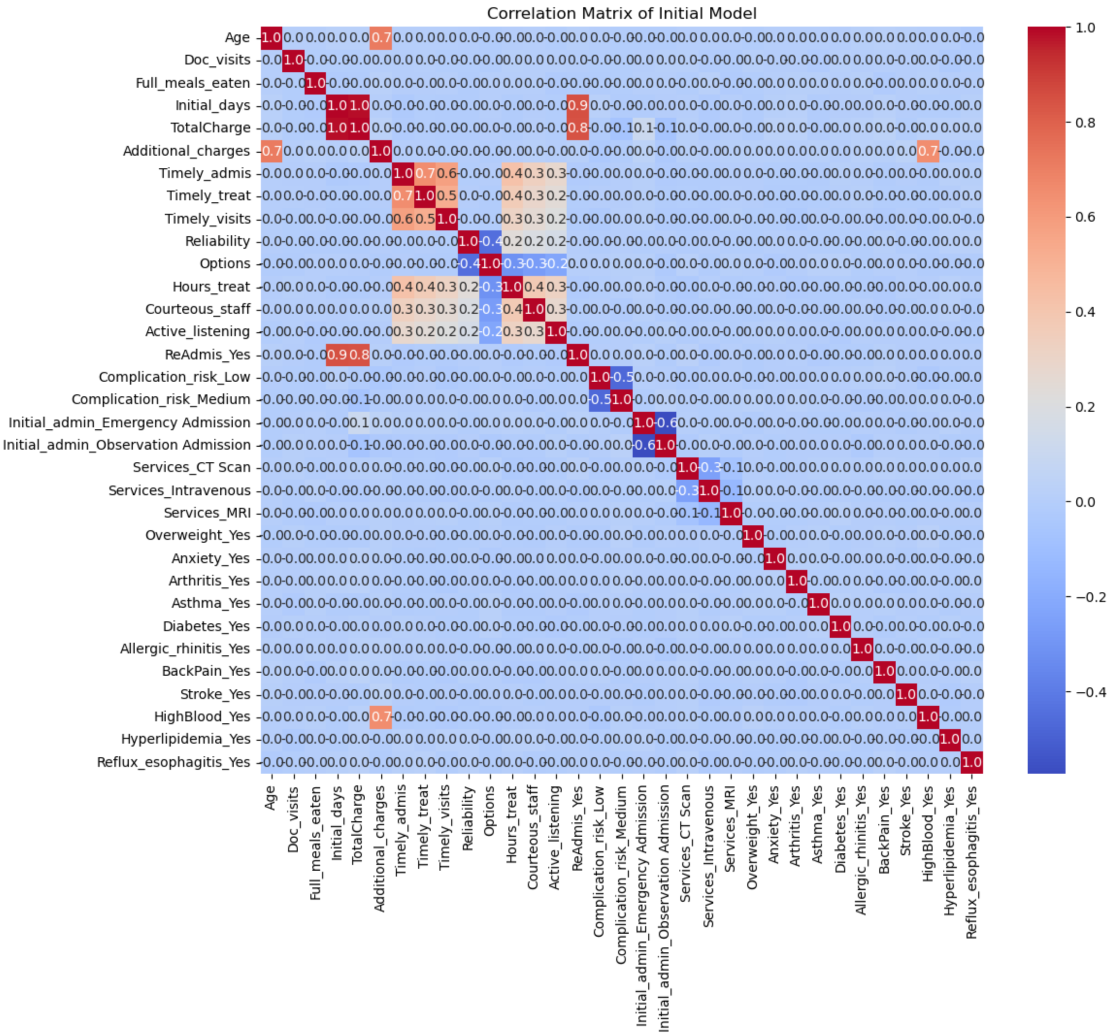
```
Initial_admin_Emergency Admission       0.008655      -0.000603   -0.005672
Initial_admin_Observation Admission    -0.004077       0.000182    0.001678
Services_CT Scan                       -0.005771       0.000754    0.013862
Services_Intravenous                    0.007251      -0.001198   -0.013559
Services_MRI                           -0.006909      -0.004160   -0.001077
Overweight_Yes                         -0.011186       0.003954    0.013943
Anxiety_Yes                             1.000000       0.012045    0.011758
Arthritis_Yes                           0.012045       1.000000   -0.006423
Asthma_Yes                              0.011758      -0.006423    1.000000
Diabetes_Yes                           -0.002529       0.009097    0.016765
Allergic_rhinitis_Yes                   0.004368       0.008748    0.004454
BackPain_Yes                            0.009289      -0.018804    0.014261
Stroke_Yes                             -0.013801      -0.018438    0.002443
HighBlood_Yes                           0.008303       0.007314    0.006174
Hyperlipidemia_Yes                     -0.013178      -0.007130   -0.009106
Reflux_esophagitis_Yes                 -0.007566       0.014894   -0.001458

                                     Diabetes_Yes  Allergic_rhinitis_Yes  \
Age                                      0.003694               0.012092
Doc_visits                               0.012781               0.002920
Full_meals_eaten                         0.009603               0.015120
TotalCharge                              0.011524               0.018919
Additional_charges                       0.002450               0.016154
Timely_admis                             0.013806               0.009402
Timely_treat                             0.005994               0.014654
Timely_visits                           -0.007469               0.001140
Reliability                             -0.005052              -0.007704
Options                                  0.017958              -0.004029
Hours_treat                             -0.004259              -0.012721
Courteous_staff                         -0.004737               0.008445
Active_listening                        -0.014752               0.005355
ReAdmis_Yes                             -0.003058              -0.004651
Complication_risk_Low                    0.006675               0.013276
Complication_risk_Medium                -0.001241              -0.017744
Initial_admin_Emergency Admission       -0.008266               0.006080
Initial_admin_Observation Admission      0.000535              -0.025280
Services_CT Scan                         0.014087               0.007008
Services_Intravenous                     0.001937              -0.003766
Services_MRI                             0.018715               0.000259
Overweight_Yes                          -0.007575               0.002819
Anxiety_Yes                             -0.002529               0.004368
Arthritis_Yes                            0.009097               0.008748
Asthma_Yes                               0.016765               0.004454
Diabetes_Yes                             1.000000               0.005486
Allergic_rhinitis_Yes                    0.005486               1.000000
BackPain_Yes                            -0.013405               0.004023
Stroke_Yes                               0.005792              -0.004837
HighBlood_Yes                           -0.005858               0.011709
```

```
Hyperlipidemia_Yes                       0.011739              -0.009049
Reflux_esophagitis_Yes                  -0.007816              -0.007731

                                     BackPain_Yes  Stroke_Yes  HighBlood_Yes  \
Age                                      0.021081    0.012035       0.007147
Doc_visits                               0.008514   -0.002230       0.008967
Full_meals_eaten                        -0.015676    0.002784       0.014784
TotalCharge                              0.035828   -0.003694       0.019910
Additional_charges                       0.014245    0.035140       0.654316
Timely_admis                            -0.011687    0.001948      -0.011017
Timely_treat                            -0.005413   -0.007706      -0.007745
Timely_visits                           -0.015480    0.001304      -0.015244
Reliability                              0.012723   -0.013430      -0.004075
Options                                  0.006659    0.005025      -0.013292
Hours_treat                             -0.016056    0.004282      -0.002369
Courteous_staff                         -0.015781   -0.005280       0.007277
Active_listening                         0.000213    0.000040       0.002601
ReAdmis_Yes                              0.013313    0.000918       0.002270
Complication_risk_Low                    0.019759   -0.001537      -0.027906
Complication_risk_Medium                -0.009920    0.000886       0.014528
Initial_admin_Emergency Admission        0.000535   -0.009743      -0.001440
Initial_admin_Observation Admission      0.009861    0.005543       0.006006
Services_CT Scan                         0.016757    0.013635       0.011772
Services_Intravenous                    -0.013446   -0.019871      -0.008408
Services_MRI                            -0.000353   -0.003580       0.001681
Overweight_Yes                           0.010083   -0.001011       0.026231
Anxiety_Yes                              0.009289   -0.013801       0.008303
Arthritis_Yes                           -0.018804   -0.018438       0.007314
Asthma_Yes                               0.014261    0.002443       0.006174
Diabetes_Yes                            -0.013405    0.005792      -0.005858
Allergic_rhinitis_Yes                    0.004023   -0.004837       0.011709
BackPain_Yes                             1.000000    0.003602       0.003048
Stroke_Yes                               0.003602    1.000000       0.007568
HighBlood_Yes                            0.003048    0.007568       1.000000
Hyperlipidemia_Yes                      -0.000963   -0.014847      -0.009529
Reflux_esophagitis_Yes                   0.016036   -0.000054       0.001150

                                     Hyperlipidemia_Yes  \
Age                                            0.003736
Doc_visits                                    -0.026730
Full_meals_eaten                               0.000688
TotalCharge                                    0.017565
Additional_charges                            -0.002475
Timely_admis                                   0.019393
Timely_treat                                   0.000697
Timely_visits                                  0.018551
Reliability                                    0.020022
Options                                       -0.009352
```

```
Hours_treat                                     0.017648
Courteous_staff                                 0.016202
Active_listening                               -0.001970
ReAdmis_Yes                                     0.004307
Complication_risk_Low                          -0.005972
Complication_risk_Medium                        0.010995
Initial_admin_Emergency Admission               0.018941
Initial_admin_Observation Admission            -0.009077
Services_CT Scan                                0.000600
Services_Intravenous                           -0.003848
Services_MRI                                   -0.002363
Overweight_Yes                                 -0.006102
Anxiety_Yes                                    -0.013178
Arthritis_Yes                                  -0.007130
Asthma_Yes                                     -0.009106
Diabetes_Yes                                    0.011739
Allergic_rhinitis_Yes                          -0.009049
BackPain_Yes                                   -0.000963
Stroke_Yes                                     -0.014847
HighBlood_Yes                                  -0.009529
Hyperlipidemia_Yes                              1.000000
Reflux_esophagitis_Yes                          0.001580

                                     Reflux_esophagitis_Yes
Age                                               -0.019609
Doc_visits                                        -0.005330
Full_meals_eaten                                  -0.000562
TotalCharge                                        0.026284
Additional_charges                                -0.011405
Timely_admis                                       0.011367
Timely_treat                                       0.017425
Timely_visits                                     -0.005584
Reliability                                        0.016277
Options                                           -0.011764
Hours_treat                                        0.009729
Courteous_staff                                   -0.013657
Active_listening                                  -0.003236
ReAdmis_Yes                                        0.005422
Complication_risk_Low                              0.000652
Complication_risk_Medium                          -0.005622
Initial_admin_Emergency Admission                 -0.000126
Initial_admin_Observation Admission               -0.008650
Services_CT Scan                                   0.017628
Services_Intravenous                              -0.022007
Services_MRI                                       0.001986
Overweight_Yes                                    -0.012240
Anxiety_Yes                                       -0.007566
Arthritis_Yes                                      0.014894
```

```
Asthma_Yes                                     -0.001458
Diabetes_Yes                                   -0.007816
Allergic_rhinitis_Yes                          -0.007731
BackPain_Yes                                    0.016036
Stroke_Yes                                     -0.000054
HighBlood_Yes                                   0.001150
Hyperlipidemia_Yes                              0.001580
Reflux_esophagitis_Yes                          1.000000
```

[10]:
```python
# plotting the initial_model correlation matrix
correlation_matrix = X_with_target.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".1f")
plt.title("Correlation Matrix of Initial Model")
plt.show()
```



```
[8]: import numpy as np

     vif_data = pd.DataFrame()
     vif_data["Variable"] = X.columns

     vif_values = []
     for i in range(len(X.columns)):
         vif = variance_inflation_factor(X.values, i)
         vif_values.append(vif)

     vif_data["VIF"] = vif_values

     # pulling p-values from the model
     p_values = model.pvalues

     # add p-values to the vif dataframe
     vif_data['P-Value'] = vif_data['Variable'].apply(lambda var: p_values.get(var,␣
      ↪np.nan))

     print(vif_data)
```

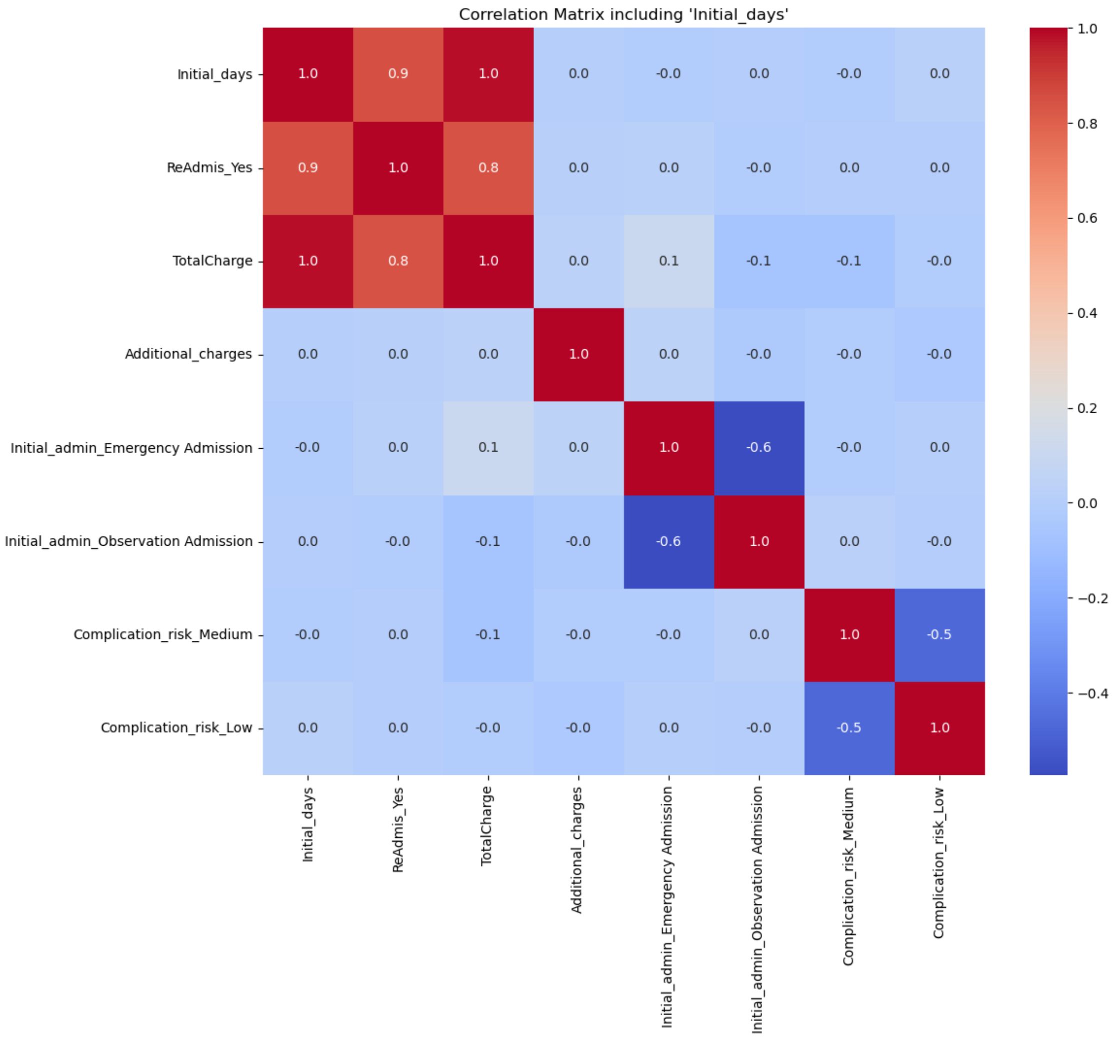
```
                                 Variable        VIF   P-Value
0                                     Age  68.480567  0.181855
1                              Doc_visits  19.941977  0.687606
2                        Full_meals_eaten   1.978869  0.178170
3                             TotalCharge  23.757244  0.000000
4                      Additional_charges  78.307975  0.090644
5                            Timely_admis  27.940825  0.944225
6                            Timely_treat  23.599967  0.114698
7                           Timely_visits  19.957174  0.136150
8                             Reliability  13.129180  0.939779
9                                 Options  11.956790  0.398910
10                            Hours_treat  18.638020  0.696698
11                        Courteous_staff  16.070136  0.684208
12                       Active_listening  14.358288  0.878522
13                            ReAdmis_Yes   5.512167  0.161024
14                  Complication_risk_Low   1.658864  0.000000
15               Complication_risk_Medium   2.400300  0.000000
16      Initial_admin_Emergency Admission   3.095585  0.000000
17    Initial_admin_Observation Admission   1.951168  0.892261
18                       Services_CT Scan   1.233743  0.774173
19                   Services_Intravenous   1.585907  0.413132
20                           Services_MRI   1.073312  0.402218
21                         Overweight_Yes   3.391201  0.950970
22                            Anxiety_Yes   1.477685  0.000000
23                          Arthritis_Yes   1.562574  0.000000
24                             Asthma_Yes   1.408097  0.512425
25                           Diabetes_Yes   1.381342  0.000000
```

```
26                Allergic_rhinitis_Yes   1.652144  0.000000
27                         BackPain_Yes   1.707121  0.000000
28                           Stroke_Yes   1.256522  0.650942
29                         HighBlood_Yes  12.980122  0.000000
30                    Hyperlipidemia_Yes   1.510806  0.000000
31                Reflux_esophagitis_Yes   1.705367  0.000000
```

As seen in the above heatmap, we can use this to find any variables with correlation to the Initial_days variable. We see that the ReAdmis and TotalCharge variables are strong Initial_days predictors, we will include Additional_charges as well. From the table of variance inflation factor's and P-values we see that the dummy variables for Initial_admin and Complication_risk had low P-values and low variance inflation factor numbers and thus will be kept. We use the variables mentioned to generate a new heatmap, and a new variance inflation factors table with P-values:

[12]:

```
# reduced_model columns
selected_columns = ['Initial_days', 'ReAdmis_Yes', 'TotalCharge',␣
 ↪'Additional_charges',
                    'Initial_admin_Emergency Admission',␣
 ↪'Initial_admin_Observation Admission','Complication_risk_Medium',␣
 ↪'Complication_risk_Low']
filtered_data = X[selected_columns]

# plotting the new correlation matrix
correlation_matrix = filtered_data.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".1f")
plt.title("Correlation Matrix of reduced model")
plt.show()
```



[36]:
```
import numpy as np

selected_columns = ['ReAdmis_Yes', 'TotalCharge', 'Additional_charges',
                    'Initial_admin_Emergency Admission',␣
 ↪'Complication_risk_Medium', 'Complication_risk_Low']
vif_cols = X[selected_columns]


vif_data = pd.DataFrame()
vif_data["Variable"] = vif_cols.columns

vif_values = []
for i in range(len(vif_cols.columns)):
    vif = variance_inflation_factor(X.values, i)
    vif_values.append(vif)
```

```
vif_data["VIF"] = vif_values

# pulling p-values from the model
p_values = model.pvalues

# Add p-values to vif dataframe
vif_data['P-Value'] = vif_data['Variable'].apply(lambda var: p_values.get(var,␣
 ↪np.nan))

print(vif_data)
```

```
                               Variable        VIF   P-Value
0                           ReAdmis_Yes  23.438315  0.161024
1                           TotalCharge   3.609938  0.000000
2                    Additional_charges   3.678486  0.090644
3  Initial_admin_Emergency Admission   2.072822  0.000000
4                Complication_risk_Medium   1.032847  0.000000
5                   Complication_risk_Low   1.323609  0.000000
```

As we can see above, the only P-value that was of issue was the ReAdmis variable but checking our heatmap and correlation matrix we can see their appears to be a linear relationship between the number of days a patient was initially admitted and their readmission. It also shows a relationship between initial_days and the patient's total charges. The new multiple linear regression model will be run with these now-reduced set of variables.

### D3. Reduced Linear Regression Model

A reduced linear regression model can be run with the narrowed down values mentioned earlier. The output of which can be seen along with further analysis continued below:

[37]:
```
import pandas as pd

data1 = pd.read_csv('prepared-data.csv')
pd.set_option('display.max_columns', None)

selected_columns = ['Initial_days', 'ReAdmis_Yes', 'TotalCharge',␣
 ↪'Additional_charges',
                    'Initial_admin_Emergency Admission',␣
 ↪'Complication_risk_Medium', 'Complication_risk_Low', 'Age']

data2 = data1[selected_columns]

X = data2.drop('Initial_days', axis=1)
X = sm.add_constant(X)

# our dependent variable
y = data2['Initial_days']
```

```python
# initial ols model
reduced_model = sm.OLS(y, X).fit()

# summary of the initial model
print(reduced_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 6.845e+05
Date:                Tue, 09 Jan 2024   Prob (F-statistic):               0.00
Time:                        20:08:42   Log-Likelihood:                -16014.
No. Observations:               10000   AIC:                         3.204e+04
Df Residuals:                    9992   BIC:                         3.210e+04
Df Model:                           7
Covariance Type:            nonrobust
================================================================================
=====================
                                        coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
---------------------
const                               -29.8795      0.058   -514.049      0.000
-29.993     -29.766
ReAdmis_Yes                           0.4313      0.047      9.113      0.000
0.339       0.524
TotalCharge                           0.0121   1.06e-05   1144.217      0.000
0.012       0.012
Additional_charges                   -0.0001   2.64e-06    -52.558      0.000
-0.000      -0.000
Initial_admin_Emergency Admission    -6.1609      0.024   -252.440      0.000
-6.209      -6.113
Complication_risk_Medium              4.9182      0.028    177.195      0.000
4.864       4.973
Complication_risk_Low                 4.8927      0.034    145.946      0.000
4.827       4.958
Age                                   0.0305      0.001     36.501      0.000
0.029       0.032
==============================================================================
Omnibus:                      121.461   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              120.583
Skew:                          -0.249   Prob(JB):                     6.54e-27
Kurtosis:                       2.798   Cond. No.                     8.65e+04
==============================================================================

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

```
specified.
[2] The condition number is large, 8.65e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Based on the reduced model above, our R-squared is now 0.998 which means the model still accounts for 99.8% of the variance. We now have a new reduced model equation as follows:

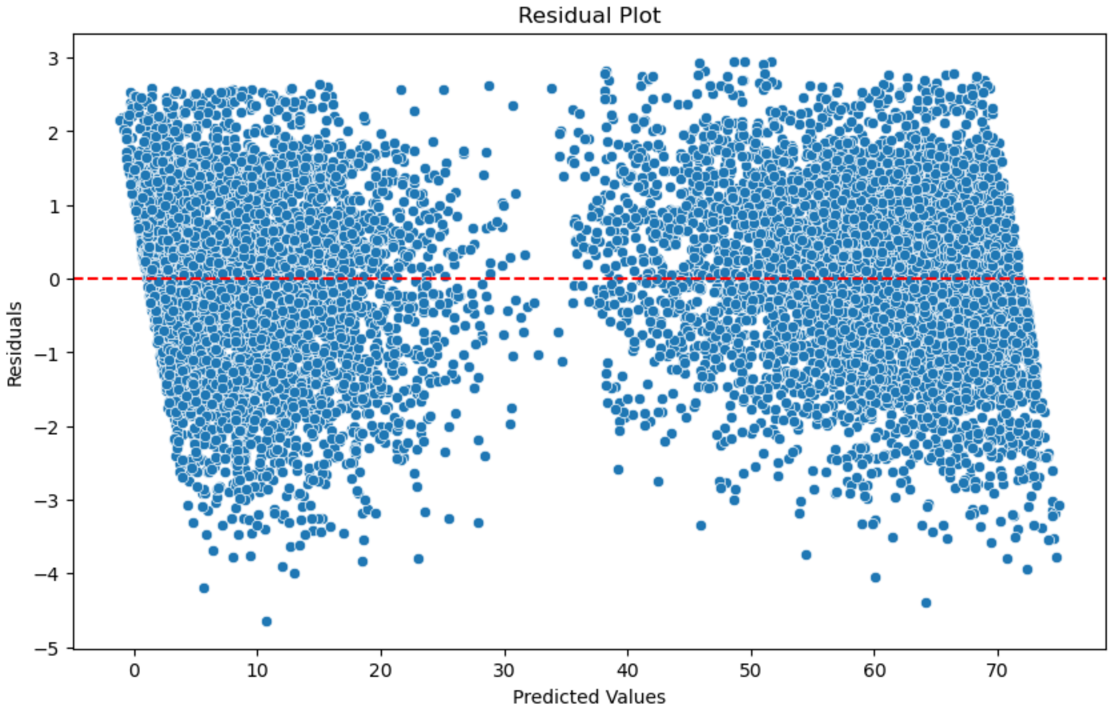
Ŷ = -29.8795 - 6.1609 * Initial_admin_Emergency Admission + 0.4313 * ReAdmis_Yes + 0.0121 * TotalCharge - 0.0001 * Additional_charges + 4.9182 * Complication_risk_Medium + 4.8927 * Complication_risk_Low + 0.0305 * Age

## E. Reduced Model Analysis

The strategy for the variable selection was an analysis of the correlation matrix, heatmap, variance inflation factors, and P-values. These tools helped to identify the variables that had a correlation with the initial_days variable. A new regression model was ran with the R-squared results shown above along with the new reduced model equation. New variance inflation factors and p-values were also generated. Lastly we will need to generate a residual plot and calculate the residual standard error:

[31]:

```python
import seaborn as sns
import matplotlib.pyplot as plt

# calculating residuals
residuals = y - reduced_model.predict(X)

# plotting residual plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x=reduced_model.predict(X), y=residuals)
plt.axhline(y=0, color='red', linestyle='--')
plt.xlabel('Predicted Values')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.show()

# calculating residual standard error
RSE = (sum(residuals ** 2) / (len(y) - X.shape[1])) ** 0.5
print("Residual Standard Error:", RSE)
```



```
Residual Standard Error: 1.20061821423433
```

### E2. Output and Calculations of Analysis

All output and calculation analysis are noted in the above sections and visualizations.

### E3. Linear Regression Code

The code used to create the multiple regression models, both the initial and reduced versions can be found below:

[32]:
```
import statsmodels.api as sm

columns = ['Age', 'Doc_visits', 'Full_meals_eaten',
       'Initial_days', 'TotalCharge', 'Additional_charges', 'Timely_admis',
       'Timely_treat', 'Timely_visits', 'Reliability', 'Options',
       'Hours_treat', 'Courteous_staff', 'Active_listening', 'ReAdmis_Yes',
       'Complication_risk_Low', 'Complication_risk_Medium',
       'Initial_admin_Emergency Admission',
       'Initial_admin_Observation Admission', 'Services_CT Scan',
       'Services_Intravenous', 'Services_MRI', 'Overweight_Yes', 'Anxiety_Yes',
       'Arthritis_Yes', 'Asthma_Yes', 'Diabetes_Yes',
       'Allergic_rhinitis_Yes', 'BackPain_Yes', 'Stroke_Yes', 'HighBlood_Yes',
       'Hyperlipidemia_Yes', 'Reflux_esophagitis_Yes']

data = data[columns]
```

```python
X = data.drop('Initial_days', axis=1)
X = sm.add_constant(X)

# our dependent variable
y = data['Initial_days']

# initial ols model
model = sm.OLS(y, X).fit()

# summary of the initial model
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 2.116e+16
Date:                Tue, 09 Jan 2024   Prob (F-statistic):               0.00
Time:                        19:08:41   Log-Likelihood:             1.1236e+05
No. Observations:               10000   AIC:                        -2.247e+05
Df Residuals:                    9967   BIC:                        -2.244e+05
Df Model:                          32
Covariance Type:            nonrobust
================================================================================
=======================
                                          coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
-----------------------
const                                 -27.6939   3.74e-07  -7.41e+07      0.000
-27.694     -27.694
Age                                  6.291e-09   4.71e-09      1.335      0.182
-2.94e-09    1.55e-08
Doc_visits                           1.232e-08   3.06e-08      0.402      0.688
-4.77e-08    7.24e-08
Full_meals_eaten                    -4.278e-08   3.18e-08     -1.347      0.178
-1.05e-07    1.95e-08
TotalCharge                             0.0122   2.84e-11    4.3e+08      0.000
0.012       0.012
Additional_charges                  -3.336e-11   1.97e-11     -1.692      0.091
-7.2e-11    5.28e-12
Timely_admis                        -3.225e-09   4.61e-08     -0.070      0.944
-9.36e-08    8.71e-08
Timely_treat                        -6.709e-08   4.25e-08     -1.578      0.115
-1.5e-07    1.63e-08
Timely_visits                         5.85e-08   3.93e-08      1.490      0.136
-1.84e-08    1.35e-07
```

```
Reliability                         -2.642e-09    3.5e-08     -0.076      0.940
-7.12e-08    6.59e-08
Options                              3.107e-08   3.68e-08      0.844      0.399
-4.11e-08    1.03e-07
Hours_treat                          1.482e-08    3.8e-08      0.390      0.697
-5.97e-08    8.93e-08
Courteous_staff                     -1.456e-08   3.58e-08     -0.407      0.684
-8.48e-08    5.56e-08
Active_listening                     5.151e-09   3.37e-08      0.153      0.879
-6.09e-08    7.12e-08
ReAdmis_Yes                          1.777e-07   1.27e-07      1.402      0.161
-7.08e-08    4.26e-07
Complication_risk_Low                   5.0464   8.99e-08   5.61e+07      0.000
5.046       5.046
Complication_risk_Medium                5.0464   7.44e-08   6.78e+07      0.000
5.046       5.046
Initial_admin_Emergency Admission      -6.2526   7.95e-08  -7.87e+07      0.000
-6.253      -6.253
Initial_admin_Observation Admission  1.235e-08   9.12e-08      0.135      0.892
-1.66e-07    1.91e-07
Services_CT Scan                     2.916e-08   1.02e-07      0.287      0.774
-1.7e-07    2.28e-07
Services_Intravenous                 5.913e-08   7.23e-08      0.818      0.413
-8.25e-08    2.01e-07
Services_MRI                         1.424e-07    1.7e-07      0.838      0.402
-1.91e-07    4.76e-07
Overweight_Yes                       4.336e-09   7.05e-08      0.061      0.951
-1.34e-07    1.43e-07
Anxiety_Yes                            -1.0510   6.86e-08  -1.53e+07      0.000
-1.051      -1.051
Arthritis_Yes                          -0.8781   6.69e-08  -1.31e+07      0.000
-0.878      -0.878
Asthma_Yes                          -4.627e-08   7.06e-08     -0.655      0.512
-1.85e-07    9.22e-08
Diabetes_Yes                           -0.9178   7.19e-08  -1.28e+07      0.000
-0.918      -0.918
Allergic_rhinitis_Yes                  -0.7394   6.56e-08  -1.13e+07      0.000
-0.739      -0.739
BackPain_Yes                           -1.0392   6.52e-08  -1.59e+07      0.000
-1.039      -1.039
Stroke_Yes                           3.639e-08   8.04e-08      0.452      0.651
-1.21e-07    1.94e-07
HighBlood_Yes                          -1.3708   1.82e-07  -7.52e+06      0.000
-1.371      -1.371
Hyperlipidemia_Yes                     -1.1471   6.78e-08  -1.69e+07      0.000
-1.147      -1.147
Reflux_esophagitis_Yes                 -0.7284   6.51e-08  -1.12e+07      0.000
-0.728      -0.728
```

```
==============================================================================
Omnibus:                      405.752   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1309.867
Skew:                           0.013   Prob(JB):                    3.68e-285
Kurtosis:                       4.773   Cond. No.                     1.81e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.81e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[58]:

```python
import pandas as pd

pd.set_option('display.max_columns', None)

selected_columns = ['Initial_days', 'ReAdmis_Yes', 'TotalCharge',␣
 ↪'Additional_charges',
                    'Initial_admin_Emergency Admission',␣
 ↪'Complication_risk_Medium', 'Complication_risk_Low', 'Age']

data2 = data[selected_columns]

X = data2.drop('Initial_days', axis=1)
X = sm.add_constant(X)

# our dependent variable
y = data2['Initial_days']

# initial ols model
reduced_model = sm.OLS(y, X).fit()

# summary of the initial model
print(reduced_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 6.845e+05
Date:                Tue, 09 Jan 2024   Prob (F-statistic):               0.00
Time:                        21:05:07   Log-Likelihood:                -16014.
No. Observations:               10000   AIC:                         3.204e+04
Df Residuals:                    9992   BIC:                         3.210e+04
Df Model:                           7
Covariance Type:            nonrobust
================================================================================
=====================
                                        coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
---------------------
const                               -29.8795      0.058   -514.049      0.000
-29.993     -29.766
ReAdmis_Yes                           0.4313      0.047      9.113      0.000
0.339       0.524
TotalCharge                           0.0121   1.06e-05   1144.217      0.000
0.012       0.012
Additional_charges                   -0.0001   2.64e-06    -52.558      0.000
-0.000      -0.000
Initial_admin_Emergency Admission    -6.1609      0.024   -252.440      0.000
-6.209      -6.113
Complication_risk_Medium              4.9182      0.028    177.195      0.000
4.864       4.973
Complication_risk_Low                 4.8927      0.034    145.946      0.000
4.827       4.958
Age                                   0.0305      0.001     36.501      0.000
0.029       0.032
==============================================================================
Omnibus:                      121.461   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              120.583
Skew:                          -0.249   Prob(JB):                     6.54e-27
Kurtosis:                       2.798   Cond. No.                     8.65e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 8.65e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

## F. Data Summary and Implications

The final regression model can be found below along with the coefficients interpretations:

Ŷ = -29.8795 - 6.1609 * Initial_admin_Emergency Admission + 0.4313 * ReAdmis_Yes + 0.0121 * TotalCharge - 0.0001 * Additional_charges + 4.9182 * Complication_risk_Medium + 4.8927 * Complication_risk_Low + 0.0305 * Age

Using the equation, the coefficients seem to show a high correlation with the initial_days variable. Each coefficient was found to have affected the Initial_days:

- Age: increase of Initial_days by 0.0305 units
- Complication_risk_Low: increase of Initial_days by 4.8927 units
- Complication_risk_Medium: increase of Initial_days by 4.9182 units
- Initial_admin_Emergency Admission: decrease of Initial_days by 6.1609 units
- Additional_charges: decrease of Initial_days by 0.0001 units

- TotalCharge: increase of Initial_days by 0.0121 units
- ReAdmis_Yes: increase of Initial_days by 0.4313 units

The P-values in all of the variables above also indicated significance aside from the ReAdmis & Additional_charges columns. It indicates that some of these variables impact the Initial_days variable, it may also indicate that Initial_days may change depending on if these variables increase or decrese. The model found that a patient's total charges and being readmitted has a correlation with the Initial_days variable. In a practical sense, these variables identified can be monitored to help predict patient readmissions. If a patient's total charges are high and they've been readmitted, the model is showing that there is a good chance their Initial_days will be high too. The same logic can be applied if a patient is being readmitted as well (high Initial_days and total charges).

The analysis does have limitations, the majority of our P-values were 0.000 but not all of them. We used Additional_charges as it may relate to TotalCharge in some way and the patient's readmission status. This analysis is meant to represent only this subset of variables. The independent variables may need to change if a different dependent variable is selected. Using only one model on a subset of variables may not lead to an entireley accurate answer to other research questions. If there are other variables not recorded by the hospital these may also change the results of the analysis.

### F2. Recommended course of action

Based on the regression analysis conducted above, hospitals should work to reduce the likelihood of a patient being readmitted to the hospital, and their total charges. The longer the patient remains in the hospital, the more of a burden it becomes on the patient. Financially, it is in the best interest of the patient for hospital staff to diagnose and treat patients as soon as they can. Other factors such as the Complication_risk and Initial_admin may also play a role in the time a patient spends in the hospital. Ensuring the patient is classified under the correct complication risk and initial admission would help. Other multiple regression models should be run using different independent variables for a more accurate analysis. Hospitals can develop programs for patients that are at higher-risk of readmission based on these models and can treat them accordingly.

## G. Panopto Video https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=a5c35e1a-4c54-4cc9-bb2b-b0f30049b98e

## H. Third-party Code References

cosine. (2020, August 14). Detecting Multicollinearity with VIF - Python. *GeeksforGeeks.*
Retrieved from www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

isitapol. (2022, February 17). How to Create a Residual Plot in Python. *GeeksforGeeks.*
Retrieved from www.geeksforgeeks.org/how-to-create-a-residual-plot-in-python/

(2020, November 10). Seaborn Heatmap - a Comprehensive Guide. *GeeksforGeeks.*
Retrieved from www.geeksforgeeks.org/seaborn-heatmap-a-comprehensive-guide/

Einblick Content Team. (2023, January 20). Three Multicollinearity Tests in Python. *Einblick.*
Retrieved from www.einblick.ai/python-code-examples/multicollinearity-test

Zach. (2021, November 22). How to Perform Bivariate Analysis in Python (with Examples). *Statology.*
Retrieved from www.statology.org/bivariate-analysis-in-python/

## I. References

Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., & Yoo, S. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLOS ONE*, 13(4), e0195901.
Retrieved from https://doi.org/10.1371/journal.pone.0195901

D208 Datasets. (n.d.). *WGU Performance Assessment*. Tasks.wgu.edu. Retrieved from https://tasks.wgu.edu/student/004659020/course/29780017/task/3784/overview

Middleton, Dr. K. (2021, October 21).
Retrieved from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=09b8fdbb-a374-452b-ba53-af39001ff3f3

Zach. (2021, November 16). The Five Assumptions of Multiple Linear Regression. *Statology*.
Retrieved from https://www.statology.org/multiple-linear-regression-assumptions/

[ ]:

```

```