Lora Milam

Western Governors University

D208 Data Cleaning

25 February 2022

D208 Performance Assessment Task 1

1 Research Question

Lora Milam Masters Data Analytics (2/14/2022) Program Mentor:d208@wgu.edu

1.1 Question

Is there a correlation between how many days a patient has in their initial stay and other characteristics of a patient?

1.2 Objectives

Hospitals could benefit from further analysis of these fields and researching further into the correlation between these fields and the amount of days for the patient's initial visit. Future findings could provide valuable information on how to increase productivity and decrease the amount of days associated with a patient's initial visit.

2 Data Analysis

2.1 Summary of Assumptions

The assumptions of multiple regression are:

- The dependant variables and the independent variables have a linear relationship
- Independant variables do not have strong relationships with each other
- y_i observations are chosen independently and randomly from the population
- Normally, residuals are distributed with a mean of zero

2.2 Justification of Tools

I will be utilizing Python's many capabilities to better analyze the database of medical patient records. Python3 is the latest iteration of the programming language Python, as provided within Jupyter Notebooks. Python is a high level, general purpose language that utilizes a variety of packages to tailor data.

Even though there are other methods that can be used to address this problem, I find Python3 and Jupyter Notebooks to be a convenient and intuitive way to visualize and draw conclusions from databases.

2.3 Appropriate Technique

Multiple regression analysis is an appropriate technique for analyzing the research question because the number of days associated with a patient's initial visit is a continuous variable. Determining which explanatory variables that have the most precedent on how long a patient will be initially admitted will allow hospitals and medical personnel to focus time and resources more on the appropriate areas. By adding and removing independent variables from the multiple regression model, it can be discerned which variables have a higher correlation to the dependent variable.

3 Data Preparation

3.1 Data Preparation Goals

The approach for this analysis includes:

1. Read in dataset: read_csv()
2. Better understand input data by evaluating data structure
3. Name dataset "medical_df" and subdatasets as "df"
4. Evaluate possible misspellings, ambiguous variable names, and absent data
5. Utilize histograms to locate outliers correlated to statistical significance, whether they hide this or create it.
6. If necessary, use imputation to replace missing data with meaningful information related to central tendency; mean,median,mode. Possibly remove outliers that are several deviations above the mean.
7. Convert character object values to numeric values/separate into separate variables using dummy variables where applicable (Himamsh)

The data set is 10,000 raw medical patient records. The target variable is the number of days of the initial visit. The title of the column is "Initial_days".

The predictor variables that are provided in the dataset may have a correlation with the probability of the patient being readmitted due to previous ailments or problems from past admissions. These predictor variables include patient medical conditions (high blood pressure, stroke, obesity, arthritis, diabetes, etc.), patient information (service while hospitalized, days in hospital, type of initial admission, etc.), patient demographics (gender,age, job, education level, etc.). These original predictor variables can be seen in the table below:

Variable	Description
CaseOrder:	- integer index - correlated to original order of raw data
Customer_id:	- character string object - unique to patient
Interaction:	- character string object - unique to patient transactions, procedures and admissions
UID:	- character string object - unique to the transactions, procedures and admissions of a patient
City:	- character string object - the city of residence of the patient
State:	- character string object - the state of residence of the patient
County:	- character string object - the county of residence of the patient
Zip:	- integer - the zip code corresponding to the residence of the patient's
Lat:	- continuous numeric (floating numeric) - GPS coordinates indicating the latitude corresponding to the patient's residence
Lng:	- continuous numeric (floating numeric) - GPS coordinates indicating the longitude corresponding to the patient's residence

Population:	- integer - the population that is within a mile radius of patient's resident

Area:	- nominal categorical
	- character string object
	- the area type corresponding to the patient's residence
	- based on unofficial census data
	- the unique values are ['Emergency Admission', 'Elective Admission', 'Observation Admission']
	`In [12]: df['Area'].unique()`
	`Out[12]: array(['Suburban', 'Urban', 'Rural'], dtype=object)`

Timezone:	- nominal categorical
	- character string object
	- the time zone corresponding to the patient's residence
	- the unique values are ['America/Chicago', 'America/New_York', 'America/Los_Angeles', 'America/Indiana/Indianapolis', 'America/Detroit', 'America/Denver', 'America/Nome', 'America/Anchorage', 'America/Phoenix', 'America/Boise', 'America/Puerto_Rico', 'America/Yakutat', 'Pacific/Honolulu', 'America/Menominee', 'America/Kentucky/Louisville', 'America/Indiana/Vincennes', 'America/Toronto', 'America/Indiana/Marengo', 'America/Indiana/Winamac', 'America/Indiana/Tell_City', 'America/Sitka', 'America/Indiana/Knox', 'America/North_Dakota/New_Salem', 'America/Indiana/Vevay', 'America/Adak', 'America/North_Dakota/Beulah']
	`In [13]: df['Timezone'].unique()`
	`Out[13]: array(['America/Chicago', 'America/New_York', 'America/Los_Angeles', 'America/Indiana/Indianapolis', 'America/Detroit', 'America/Denver', 'America/Nome', 'America/Anchorage', 'America/Phoenix', 'America/Boise', 'America/Puerto_Rico', 'America/Yakutat', 'Pacific/Honolulu', 'America/Menominee', 'America/Kentucky/Louisville', 'America/Indiana/Vincennes', 'America/Toronto', 'America/Indiana/Marengo', 'America/Indiana/Winamac', 'America/Indiana/Tell_City', 'America/Sitka', 'America/Indiana/Knox', 'America/North_Dakota/New_Salem', 'America/Indiana/Vevay', 'America/Adak', 'America/North_Dakota/Beulah'], dtype=object)`

Job:	- nominal categorical

	- character string object - the occupation of the patient or insurance holder
Children:	- integer - the amount of children within patient's household
Age:	- integer - the patient's age
Income:	- numeric value - the patient's or insurance holder's annual income
Marital:	- nominal categorical - character string object - the marital status of the patient or insurance holder - the unique values are ['Divorced', 'Married', 'Widowed', 'Never Married', 'Separated'] In [17]: `df['Marital'].unique()` Out[17]: `array(['Divorced', 'Married', 'Widowed', 'Never Married', 'Separated'], dtype=object)`
Gender:	- nominal categorical - character string object - the gender of the patient - the unique values are ['Male', 'Female', 'Prefer not to answer'] In [18]: `df['Gender'].unique()` Out[18]: `array(['Male', 'Female', 'Prefer not to answer'], dtype=object)`
ReAdmis	- binary categorical - character string object - whether or not, within a month of release, each customer has been readmitted to the hospital

VitD_levels:	- continuous numeric (floating numeric) - value of the vitamin D levels of the patient - measured in ng/mL
Doc_visits:	- integer - the number of times during the initial hospitalization that the primary physician visited the patient
Full_meals_eaten:	- integer - number of full meals eaten Note: It counts as zero if the patient only eats a partial meal
VitD_supp:	- integer - the number of times that vitamin D supplements were administered to patient
Soft_drink:	- binary categorical - character string object - whether or not a patient on a daily basis drinks three or more sodas - the unique values are [Yes, No]
Initial_admin:	- nominal categorical - character string object - the reason why the patient was initially admitted into the hospital In [19]: `df['Initial_admin'].unique()` Out[19]: `array(['Emergency Admission', 'Elective Admission', 'Observation Admission'], dtype=object)`
HighBlood:	- binary categorical - character string object - whether or not the patient has high blood pressure - the unique values are [Yes, No]

Stroke:	- binary categorical - character string object - whether or not the patient has had a stroke - the unique values are [Yes, No]
Complication_risk:	- ordinal categorical - character string object - the level of complication risk - the unique values are [High, Medium, Low]
Overweight:	- binary categorical - integer - whether or not the patient is overweight, as determined by their BMI elements: age, gender, and height - the unique values are [1,0]
Arthritis:	- binary categorical - character string object - whether or not the patient has arthritis - the unique values are [Yes, No]
Diabetes:	- binary categorical - character string object - whether or not the patient has diabetes - the unique values are [Yes, No]
Hyperlipidemia:	- binary categorical - character string object - whether or not the patient has hyperlipidemia

	- the unique values are [Yes, No]
BackPain:	- binary categorical - character string object - whether or not the patient has chronic backpain - the unique values are [Yes, No]
Anxiety:	- binary categorical - integer - whether or not the patient has an anxiety disorder - the unique values are [1,0]
Allergic_rhinitis:	- binary categorical - character string object - whether or not the patient has allergic rhinitis - the unique values are [Yes, No]
Reflux_esophagitis:	- binary categorical - character string object - whether or not the patient has reflux esophagitis - the unique values are [Yes, No]
Asthma:	- binary categorical - character string object - whether or not the patient has asthma - the unique values are [Yes, No]
Services:	- nominal categorical - character string object

	- the primary service the patient received while hospitalized
	- the unique values are ['Blood Work', 'Intravenous', 'CT Scan', 'MRI']
	`In [20]: df['Services'].unique()`
	`Out[20]: array(['Blood Work', 'Intravenous', 'CT Scan', 'MRI'], dtype=object)`

Initial_days:	- numeric value - number of days,during the initial visit, the patient stayed in the hospital - **target variable**
TotalCharge:	- numeric value - patient's average daily charges,during the initial visit, for typical (not specialized) treatments and services
Additional_charges:	- numeric value - patient's average charges,during the initial visit, for additional treatments and services
Item1:	- integer value - from most important to least important, the level of importance of timely admission - the unique values are [1:8]
Item2:	- integer value - from most important to least important, the level of importance of timely treatment - the unique values are [1:8]
Item3:	- integer value - from most important to least important, the level of importance of timely visits - the unique values are [1:8]

Item4:	- integer value - from most important to least important, the level of importance of reliability - the unique values are [1:8]
Item5:	- integer value - from most important to least important, the level of importance of options
Item6:	- integer value - from most important to least important, the level of importance of hours of treatment - the unique values are [1:8]
Item7:	- integer value - from most important to least important, the level of importance of courteous staff - the unique values are [1:8]
Item8:	- integer value - from doctor from most important to least important, the level of importance of evidence of active listening - the unique values are [1:8]

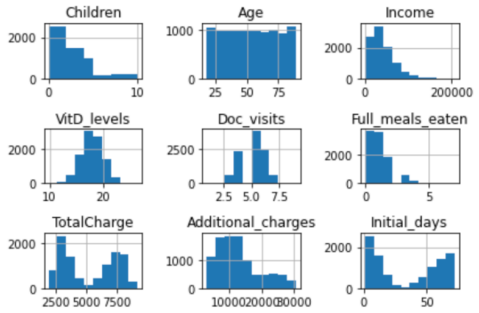
3.2 Summary Statistics

For this analysis, the less relevant variables were removed ('CaseOrder', 'Customer_id', 'Interaction', 'UID', ‘City’, ‘State’, ‘County’, ‘Zip’, ‘Lat’, ‘Lng’, ‘Population’, ‘Area’, ‘Timezone’, 'Job', 'Marital'). Binomial “Yes”/”No” or “Male”/”Female” variables were imputed to a corresponding numeric variable, 1/0. The resulting data frame will consist of 35 columns, including the target variable. The dataset had been preemptively cleaned, resulting in no NULL, NAs or missing values.

Measures of central tendency through histograms and boxplots showed normal distributions for “TotalCharge” and “Doc_visits”. The cleaned dataset no longer had any unjustifiable outliers, meaning that any remaining outliers were pertinent to the analysis. The histograms for “TotalCharge”, and “Inital_days” both had a binomial distribution. In the scatter plots between

the dependent variable and the independent variables, there was a direct linear relationship between “Inital_days” and “TotalCharge”.

In [13]:
```
# Histogram of continuous variables
medical_df[['Children', 'Age', 'Income', 'VitD_levels', 'Doc_visits',
            'Full_meals_eaten', 'TotalCharge', 'Additional_charges', 'Initial_days']].hist()
plt.savefig('medical_pyplot.jpg')
plt.tight_layout()
```



In [69]:
```
# Reduced multiple regression model
medical_df['intercept'] = 1
lm_ReAdmis_reduced = sm.OLS(medical_df['Initial_days'], medical_df[['TotalCharge', 'Survey_TimelyAdmin',
                                                                    'Survey_TimelyTreatment', 'Survey_TimelyVisits',
                                                                    'intercept']]).fit()
print(lm_ReAdmis_reduced.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.975
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                 9.925e+04
Date:                Sat, 26 Feb 2022   Prob (F-statistic):               0.00
Time:                        01:07:48   Log-Likelihood:                -28355.
No. Observations:               10000   AIC:                         5.672e+04
Df Residuals:                    9995   BIC:                         5.676e+04
Df Model:                           4
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
TotalCharge                  0.0119   1.89e-05    629.883      0.000       0.012       0.012
Survey_TimelyAdmin          -0.0689      0.057     -1.209      0.227      -0.180       0.043
Survey_TimelyTreatment       0.0062      0.054      0.114      0.910      -0.100       0.112
Survey_TimelyVisits         -0.0112      0.050     -0.223      0.824      -0.110       0.088
intercept                  -28.5873      0.199   -143.886      0.000     -28.977     -28.198
==============================================================================
Omnibus:                      740.927   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              272.868
Skew:                          -0.135   Prob(JB):                     5.59e-60
Kurtosis:                       2.237   Cond. No.                     2.78e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.78e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

3.3 Data Preparation Steps

To properly analyze the data given, first the data will be cleaned by using Python3 and data cleaning methods. Python3 is the latest iteration of the programming language Python, as provided within Jupyter Notebooks. The plan for cleaning the data will be done through the following steps:

1. Import the dataset to a Python dataframe
2. Relabel survey columns to something more identifiable

3. Describe dataframe, structure, and data types
4. Summary statistics
5. Remove any redundant, irrelevant, or misleading fields
6. Impute missing data with measures of central tendency (mean, median, mode) or remove outliers that are several standard deviations above the mean.
7. Convert character object values to numeric values/separate into separate variables using dummy variables where applicable (Himamsh)
8. Univariate and bivariate visualizations
9. Place 'Initial_days' at the end of the dataframe
10. Prepared dataset will be exported as "medical_prepared.csv"
11. No data will be deleted or changed unless it is an obvious mistake. Note: Adjusting data too much will skew the model towards a conclusion that may not generalize due to the data being overly-manipulated.

In [1]:
```
# Standard imports
import numpy as np
import pandas as pd
from pandas import Series, DataFrame

# Visualization libraries
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

# Statistics packages
import pylab
from pylab import rcParams
import statsmodels.api as sm
import statistics
from scipy import stats

# Scikit-learn
import sklearn
from sklearn import preprocessing
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import classification_report

# Import chisquare from SciPy.stats
from scipy.stats import chisquare
from scipy.stats import chi2_contingency
```

In [2]:
```
# Load dataset into Pandas dataframe
medical_df = pd.read_csv('medical_clean.csv')
```

In [3]:
```
# Rename survey columns to more identifiable names
medical_df.rename(columns =
    {'Item1':'Survey_TimelyAdmin',
     'Item2':'Survey_TimelyTreatment',
     'Item3':'Survey_TimelyVisits',
     'Item4':'Survey_Reliability',
     'Item5':'Survey_Options',
     'Item6':'Survey_HoursTreatment',
     'Item7':'Survey_CourteousStaff',
     'Item8':'Survey_ActiveListening'}, inplace=True)

# Display updated Medical dataframe
medical_df
```

Out[3]:

	CaseOrder	Customer_id	Interaction	UID	City	State	County	Zip	Lat	Lng	...	TotalCharge
0	1	C412403	8cd49b13-f45a-4b47-a2bd-173ffa932c2f	3a83ddb66e2ae73798bdf1d705dc0932	Eva	AL	Morgan	35621	34.34960	-86.72508	...	3726.702860
1	2	Z919181	d2450b70-0337-4406-bdbb-bc1037f1734c	176354c5eef714957d486009feabf195	Marianna	FL	Jackson	32446	30.84513	-85.22907	...	4193.190458
2	3	F995323	a2057123-abf5-4a2c-abad-8ffe33512562	e19a0fa00aeda885b8a436757e889bc9	Sioux Falls	SD	Minnehaha	57110	43.54321	-96.63772	...	2434.234222
3	4	A879973	1dec528d-eb34-4079-adce-0d7a40e82205	cd17d7b6d152cb6f23957346d11c3f07	New Richland	MN	Waseca	56072	43.89744	-93.51479	...	2127.830423
4	5	C544523	5885f56b-d6da-43a3-8760-83583af94266	d2f0425877b10ed6bb381f3e2579424a	West Point	VA	King William	23181	37.59894	-76.88958	...	2113.073274
...	...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	B863060	a25b594d-0328-486f-a9b9-0567eb0f9723	39184dc28cc038871912ccc4500049e5	Norlina	NC	Warren	27563	36.42886	-78.23716	...	6850.942000
9996	9997	P712040	70711574-f7b1-4a17-b15f-48c54564b70f	3cd124ccd43147404292e883bf9ec55c	Milmay	NJ	Atlantic	8340	39.43609	-74.87302	...	7741.690000
9997	9998	R778890	1d79569d-8e0f-4180-a207-d67ee4527d26	41b770aeee97a5b9e7f69c906a8119d7	Southside	TN	Montgomery	37171	36.36655	-87.29988	...	8276.481000
9998	9999	E344109	f5a68e69-2a60-409b-a92f-ac0847b27db0	2bb491ef5b1beb1fed758cc6885c167a	Quinn	SD	Pennington	57775	44.10354	-102.01590	...	7644.483000
9999	10000	I569847	bc482c02-f8c9-4423-99de-3db5e62a18d5	95663a202338000abdf7e09311c2a8a1	Coraopolis	PA	Allegheny	15108	40.49998	-80.19959	...	7887.553000

10000 rows × 50 columns

In [4]:
```
# List dataframe columns
df = medical_df.columns
print(df)
```

```
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', 'Survey_TimelyAdmin',
       'Survey_TimelyTreatment', 'Survey_TimelyVisits', 'Survey_Reliability',
       'Survey_Options', 'Survey_HoursTreatment', 'Survey_CourteousStaff',
       'Survey_ActiveListening'],
      dtype='object')
```

In [5]:
```
# Number of records and columns of dataframe
medical_df.shape
```

Out[5]:
```
(10000, 50)
```

In [6]:
```
# Describe Medical dataset stats
medical_df.describe()
```

Out[6]:

	CaseOrder	Zip	Lat	Lng	Population	Children	Age	Income	VitD_levels	Doc_visits	...	Tot
count	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	...	1000
mean	5000.50000	50159.323900	38.751099	-91.243080	9965.253800	2.097200	53.511700	40490.495160	17.964262	5.012200	...	53
std	2886.89568	27469.588208	5.403085	15.205998	14824.758614	2.163659	20.638538	28521.153293	2.017231	1.045734	...	21
min	1.00000	610.000000	17.967190	-174.209700	0.000000	0.000000	18.000000	154.080000	9.806483	1.000000	...	19
25%	2500.75000	27592.000000	35.255120	-97.352982	694.750000	0.000000	36.000000	19598.775000	16.626439	4.000000	...	31
50%	5000.50000	50207.000000	39.419355	-88.397230	2769.000000	1.000000	53.000000	33768.420000	17.951122	5.000000	...	52
75%	7500.25000	72411.750000	42.044175	-80.438050	13945.000000	3.000000	71.000000	54296.402500	19.347963	6.000000	...	74
max	10000.00000	99929.000000	70.560990	-65.290170	122814.000000	10.000000	89.000000	207249.100000	26.394449	9.000000	...	91

8 rows × 23 columns

In [7]:
```
# Remove less relevant fields from stats description
medical_df = medical_df.drop(columns=['CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State',
                                      'County', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'TimeZone', 'Job',
                                      'Marital'])
medical_df.describe()
medical_df.shape
```

Out[7]:
```
(10000, 35)
```

In [8]:
```
# Count of missing values by column
data_nulls = medical_df.isnull().sum()
print(data_nulls)
```

```
Children                0
Age                     0
Income                  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
Survey_TimelyAdmin      0
Survey_TimelyTreatment  0
Survey_TimelyVisits     0
Survey_Reliability      0
Survey_Options          0
Survey_HoursTreatment   0
Survey_CourteousStaff   0
Survey_ActiveListening  0
dtype: int64
```

In [9]:
```
# Impute qualitive data fields by creating binary dummy columns then drop pre-existing column
# Exclude redundant values: ex. gender is categorized as male or female, the value 'prefer not to answer' can be identi
dmy = pd.get_dummies(medical_df['Gender'])
dmy = dmy.iloc[:,:-1]
medical_df = pd.concat([medical_df,dmy],axis=1)
medical_df = medical_df.drop(columns = 'Gender')

dmy = pd.get_dummies(medical_df['Initial_admin'])
medical_df = pd.concat([medical_df,dmy],axis=1)
medical_df = medical_df.drop(columns = 'Initial_admin')

dmy = pd.get_dummies(medical_df['Complication_risk'])
medical_df = pd.concat([medical_df,dmy],axis=1)
medical_df = medical_df.drop(columns = 'Complication_risk')

dmy = pd.get_dummies(medical_df['Services'])
medical_df = pd.concat([medical_df,dmy],axis=1)
medical_df = medical_df.drop(columns = 'Services')

medical_df.describe()
```

Out[9]:

	Children	Age	Income	VitD_levels	Doc_visits	Full_meals_eaten	vitD_supp	Initial_days	TotalCharge	Additional_charges
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	2.097200	53.511700	40490.495160	17.964262	5.012200	1.001400	0.398900	34.455299	5312.172769	12934.528587
std	2.163659	20.638538	28521.153293	2.017231	1.045734	1.008117	0.628505	26.309341	2180.393838	6542.601544
min	0.000000	18.000000	154.080000	9.806483	1.000000	0.000000	0.000000	1.001981	1938.312067	3125.703000
25%	0.000000	36.000000	19598.775000	16.626439	4.000000	0.000000	0.000000	7.896215	3179.374015	7986.487755
50%	1.000000	53.000000	33768.420000	17.951122	5.000000	1.000000	0.000000	35.836244	5213.952000	11573.977735
75%	3.000000	71.000000	54296.402500	19.347963	6.000000	2.000000	1.000000	61.161020	7459.699750	15626.490000
max	10.000000	89.000000	207249.100000	26.394449	9.000000	7.000000	5.000000	71.981490	9180.728000	30566.070000

8 rows × 27 columns

In [10]:
```
df = medical_df.columns
print(df)
```

```
Index(['Children', 'Age', 'Income', 'ReAdmis', 'VitD_levels', 'Doc_visits',
       'Full_meals_eaten', 'vitD_supp', 'Soft_drink', 'HighBlood', 'Stroke',
       'Complication_risk', 'Overweight', 'Arthritis', 'Diabetes',
       'Hyperlipidemia', 'BackPain', 'Anxiety', 'Allergic_rhinitis',
       'Reflux_esophagitis', 'Asthma', 'Initial_days', 'TotalCharge',
       'Additional_charges', 'Survey_TimelyAdmin', 'Survey_TimelyTreatment',
       'Survey_TimelyVisits', 'Survey_Reliability', 'Survey_Options',
       'Survey_HoursTreatment', 'Survey_CourteousStaff',
       'Survey_ActiveListening', 'Female', 'Male', 'Elective Admission',
       'Emergency Admission', 'Observation Admission', 'Blood Work', 'CT Scan',
       'Intravenous', 'MRI'],
      dtype='object')
```

In [11]:
```
# Impute binomial categoricals as numerics
data = medical_df['ReAdmis']
readmis_dict = {
    "No":0,"Yes":1
}
for k,v in readmis_dict.items():
    data = data.replace(k,v)

medical_df['ReAdmis'] = data

data = medical_df['Soft_drink']
for k,v in readmis_dict.items():
    data = data.replace(k,v)

medical_df['Soft_drink'] = data

data = medical_df['HighBlood']
for k,v in readmis_dict.items():
    data = data.replace(k,v)

medical_df['HighBlood'] = data

data = medical_df['Stroke']
for k,v in readmis_dict.items():
    data = data.replace(k,v)
medical_df['Stroke'] = data

data = medical_df['Arthritis']
for k,v in readmis_dict.items():
    data = data.replace(k,v)
medical_df['Arthritis'] = data

data = medical_df['Diabetes']
for k,v in readmis_dict.items():
    data = data.replace(k,v)
medical_df['Diabetes'] = data

data = medical_df['Hyperlipidemia']
for k,v in readmis_dict.items():
    data = data.replace(k,v)
medical_df['Hyperlipidemia'] = data

data = medical_df['BackPain']
for k,v in readmis_dict.items():
    data = data.replace(k,v)
medical_df['BackPain'] = data

data = medical_df['Allergic_rhinitis']
for k,v in readmis_dict.items():
    data = data.replace(k,v)
medical_df['Allergic_rhinitis'] = data

data = medical_df['Reflux_esophagitis']
for k,v in readmis_dict.items():
    data = data.replace(k,v)
medical_df['Reflux_esophagitis'] = data

data = medical_df['Asthma']
for k,v in readmis_dict.items():
    data = data.replace(k,v)
medical_df['Asthma'] = data

df = medical_df.columns
print(df)
```

```
Index(['Children', 'Age', 'Income', 'ReAdmis', 'VitD_levels', 'Doc_visits',
       'Full_meals_eaten', 'vitD_supp', 'Soft_drink', 'HighBlood', 'Stroke',
       'Complication_risk', 'Overweight', 'Arthritis', 'Diabetes',
       'Hyperlipidemia', 'BackPain', 'Anxiety', 'Allergic_rhinitis',
       'Reflux_esophagitis', 'Asthma', 'Initial_days', 'TotalCharge',
       'Additional_charges', 'Survey_TimelyAdmin', 'Survey_TimelyTreatment',
       'Survey_TimelyVisits', 'Survey_Reliability', 'Survey_Options',
       'Survey_HoursTreatment', 'Survey_CourteousStaff',
       'Survey_ActiveListening', 'Female', 'Male', 'Elective Admission',
       'Emergency Admission', 'Observation Admission', 'Blood Work', 'CT Scan',
       'Intravenous', 'MRI'],
      dtype='object')
```

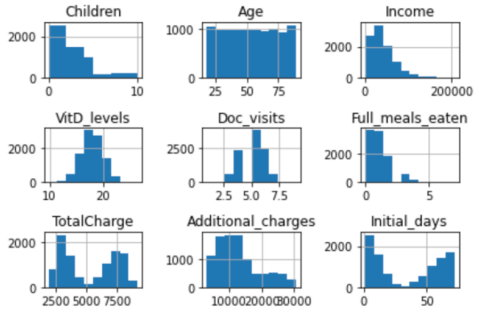
In [12]:
```
# Move Initial_days to the end of the dataframe
ReAdmis = medical_df[['Initial_days']]
medical_df.pop("Initial_days")
medical_df['Initial_days'] = ReAdmis

df = medical_df.columns
print(df)
medical_df.head()
```

```
Index(['Children', 'Age', 'Income', 'ReAdmis', 'VitD_levels', 'Doc_visits',
       'Full_meals_eaten', 'vitD_supp', 'Soft_drink', 'HighBlood', 'Stroke',
       'Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain',
       'Anxiety', 'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma',
       'TotalCharge', 'Additional_charges', 'Survey_TimelyAdmin',
       'Survey_TimelyTreatment', 'Survey_TimelyVisits', 'Survey_Reliability',
       'Survey_Options', 'Survey_HoursTreatment', 'Survey_CourteousStaff',
       'Survey_ActiveListening', 'Female', 'Male', 'Elective Admission',
       'Emergency Admission', 'Observation Admission', 'High', 'Low', 'Medium',
       'Blood Work', 'CT Scan', 'Intravenous', 'MRI', 'Initial_days'],
      dtype='object')
```

3.4 Visualizations

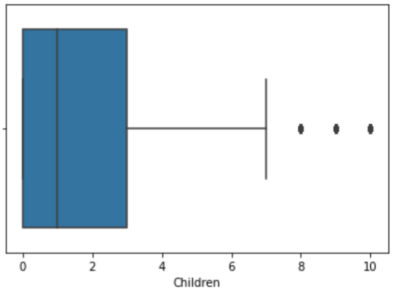
3.4.1 Univariate Statistics

In [13]:
```
# Histogram of continuous variables
medical_df[['Children', 'Age', 'Income', 'VitD_levels', 'Doc_visits',
            'Full_meals_eaten', 'TotalCharge', 'Additional_charges', 'Initial_days']].hist()
plt.savefig('medical_pyplot.jpg')
plt.tight_layout()
```



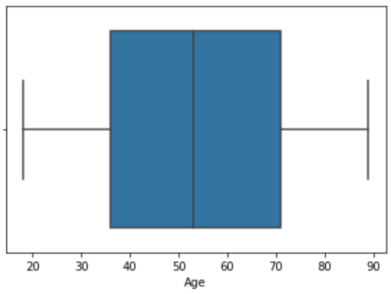
In [14]:
```
# Seaborn boxplots for continuous variables
sns.boxplot('Children', data = medical_df)
plt.show()
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



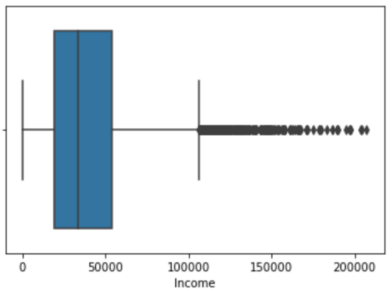
In [15]:
```
sns.boxplot('Age', data = medical_df)
plt.show()
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



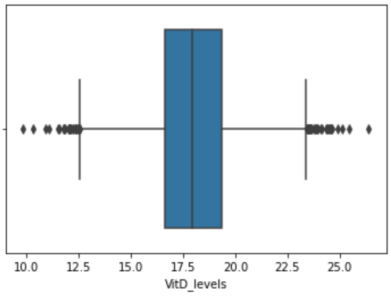
In [16]:
```
sns.boxplot('Income', data = medical_df)
plt.show()
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



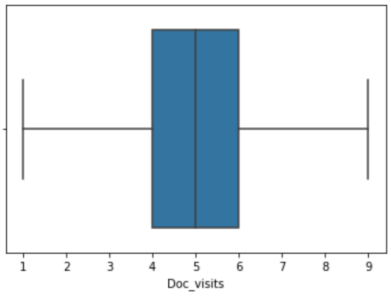
In [17]:
```
sns.boxplot('VitD_levels', data = medical_df)
plt.show()
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



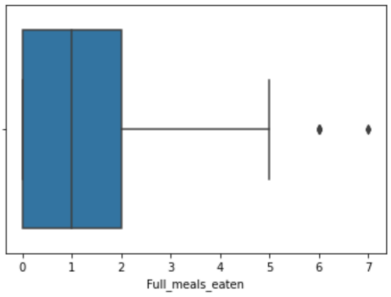
In [18]:
```
sns.boxplot('Doc_visits', data = medical_df)
plt.show()
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



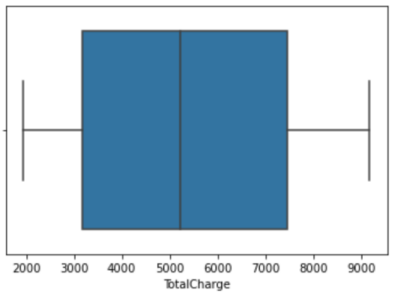
In [19]:
```
sns.boxplot('Full_meals_eaten', data = medical_df)
plt.show()
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



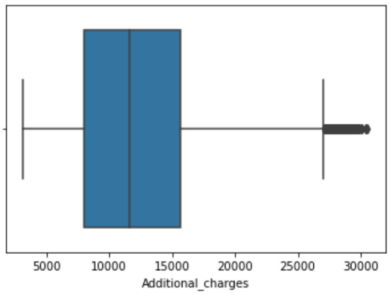
In [20]:
```
sns.boxplot('TotalCharge', data = medical_df)
plt.show()
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



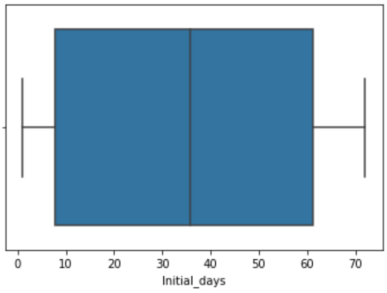
In [21]:
```
sns.boxplot('Additional_charges', data = medical_df)
plt.show()
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



In [22]:
```
sns.boxplot('Initial_days', data = medical_df)
plt.show()
```

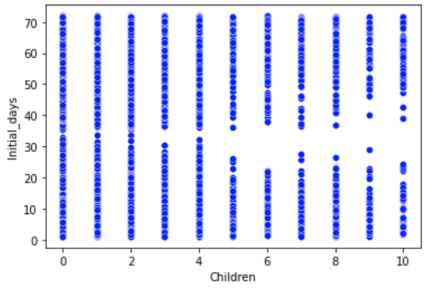
```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments wit
hout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



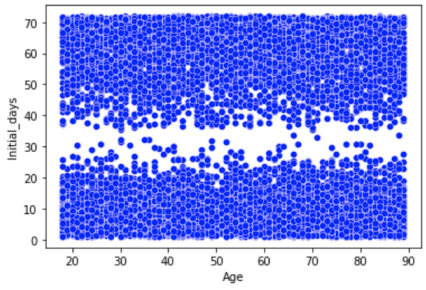
In [23]:
```
# The remaining outliers are justifiable and do not need to be removed
```

3.4.2 Bivariate Statistics

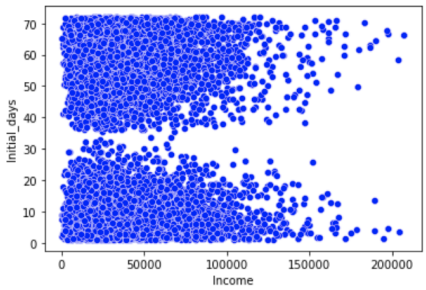
In [24]:
```
# Scatterplots to show relationships between target and independant variables
sns.scatterplot(x=medical_df['Children'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



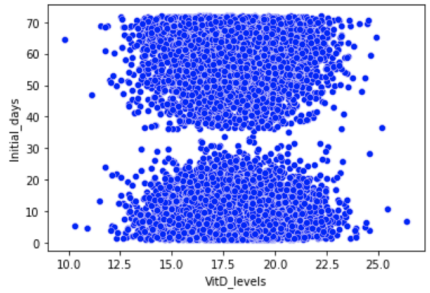
In [25]:
```
sns.scatterplot(x=medical_df['Age'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



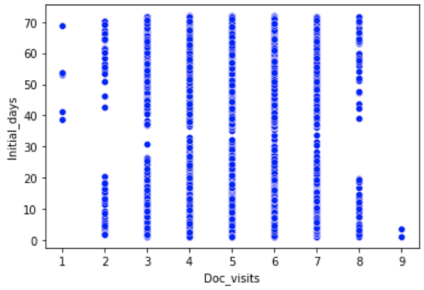
In [26]:
```
sns.scatterplot(x=medical_df['Income'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



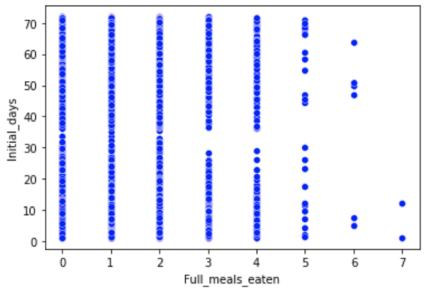
In [27]:
```
sns.scatterplot(x=medical_df['VitD_levels'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



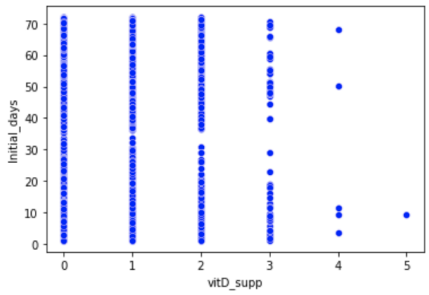
In [28]:
```
sns.scatterplot(x=medical_df['Doc_visits'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



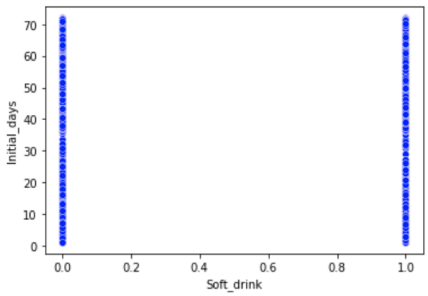
In [29]:
```
sns.scatterplot(x=medical_df['Full_meals_eaten'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



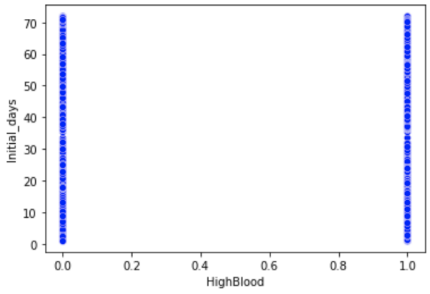
In [30]:
```
sns.scatterplot(x=medical_df['vitD_supp'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



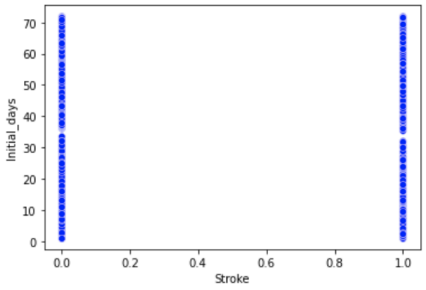
In [31]:
```
sns.scatterplot(x=medical_df['Soft_drink'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



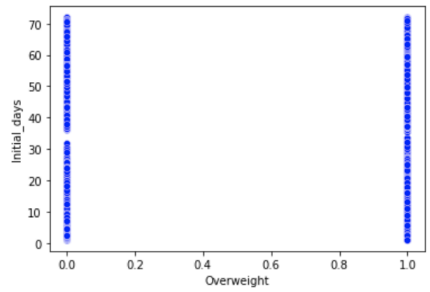
In [32]:
```
sns.scatterplot(x=medical_df['HighBlood'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



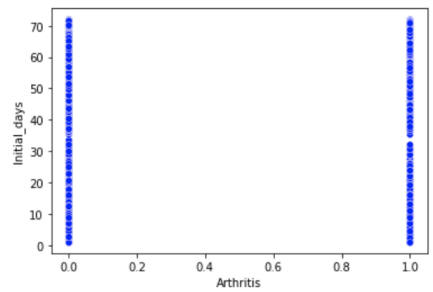
In [33]:
```
sns.scatterplot(x=medical_df['Stroke'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



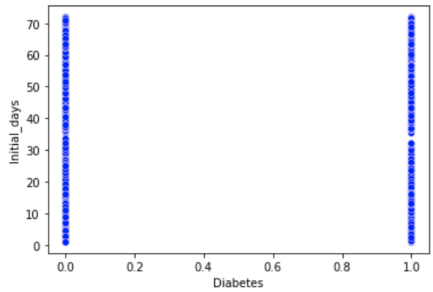
In [34]:
```
sns.scatterplot(x=medical_df['Overweight'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



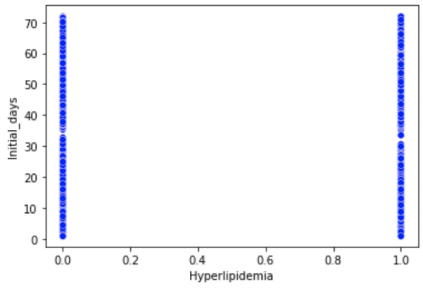
In [35]:
```
sns.scatterplot(x=medical_df['Arthritis'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



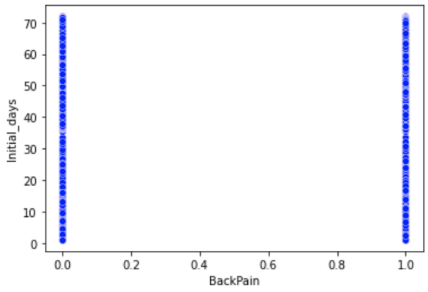
In [36]:
```
sns.scatterplot(x=medical_df['Diabetes'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



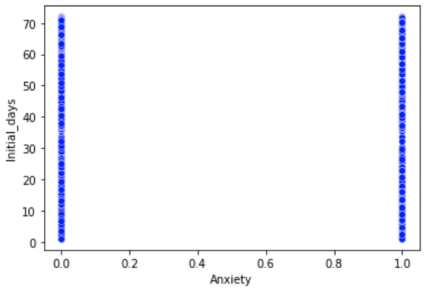
In [37]:
```
sns.scatterplot(x=medical_df['Hyperlipidemia'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



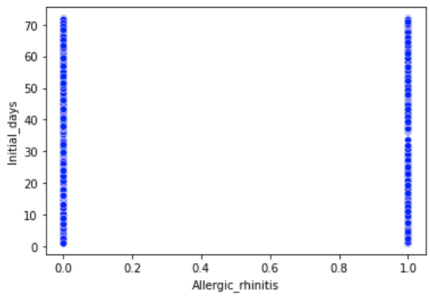
In [38]:
```
sns.scatterplot(x=medical_df['BackPain'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



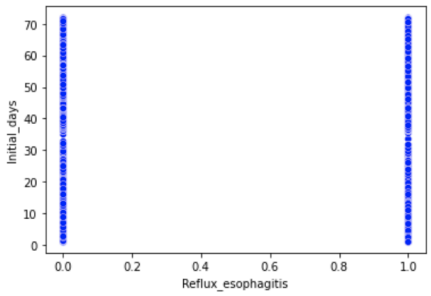
In [39]:
```
sns.scatterplot(x=medical_df['Anxiety'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



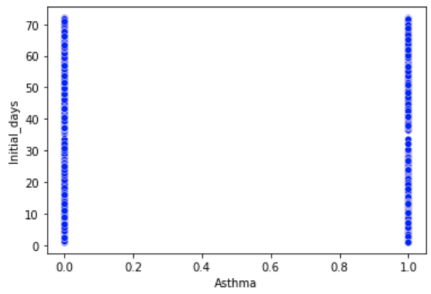
In [40]:
```
sns.scatterplot(x=medical_df['Allergic_rhinitis'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



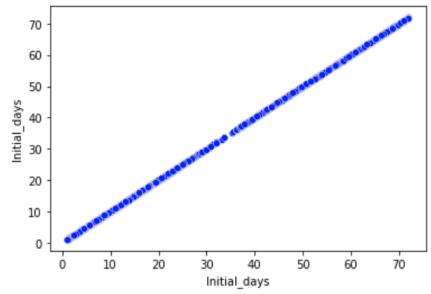
In [41]:
```
sns.scatterplot(x=medical_df['Reflux_esophagitis'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



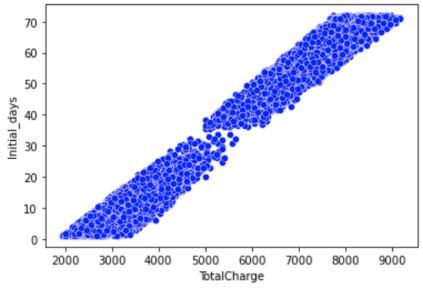
In [42]:
```
sns.scatterplot(x=medical_df['Asthma'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



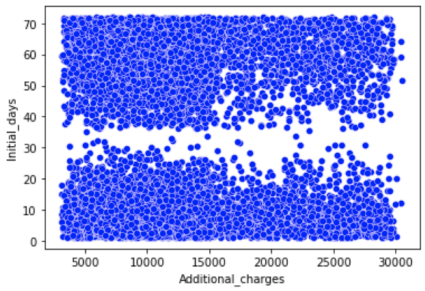
In [43]:
```
sns.scatterplot(x=medical_df['Initial_days'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



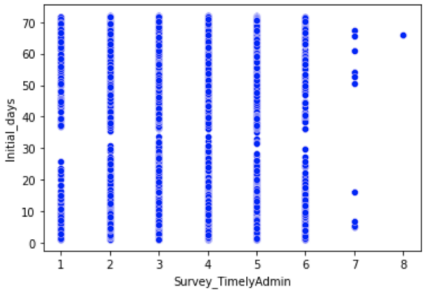
In [44]:
```
sns.scatterplot(x=medical_df['TotalCharge'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



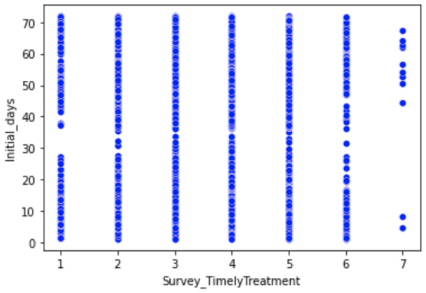
In [45]:
```
sns.scatterplot(x=medical_df['Additional_charges'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



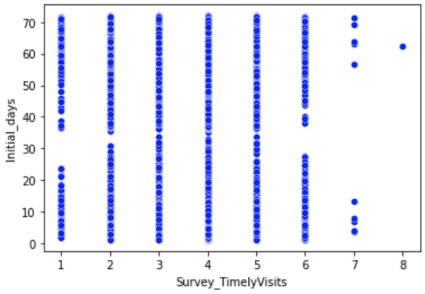
In [46]:
```
sns.scatterplot(x=medical_df['Survey_TimelyAdmin'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



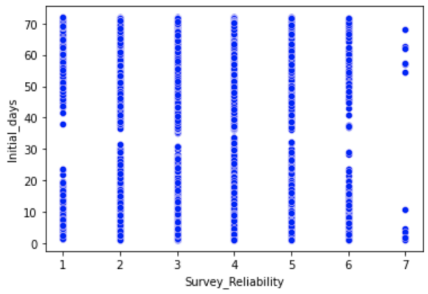
In [47]:
```
sns.scatterplot(x=medical_df['Survey_TimelyTreatment'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



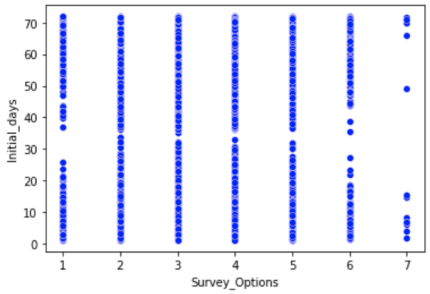
In [48]:
```
sns.scatterplot(x=medical_df['Survey_TimelyVisits'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



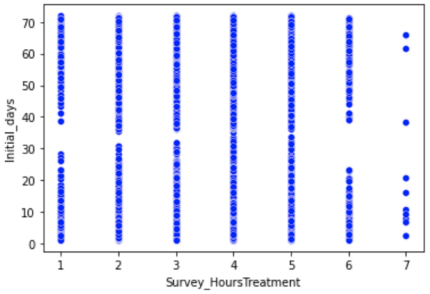
In [49]:
```
sns.scatterplot(x=medical_df['Survey_Reliability'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



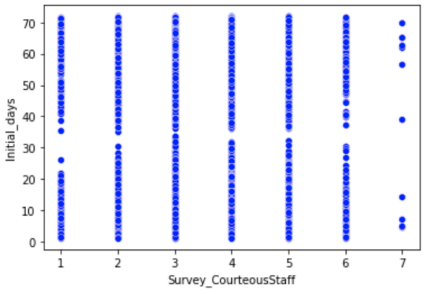
In [50]:
```
sns.scatterplot(x=medical_df['Survey_Options'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



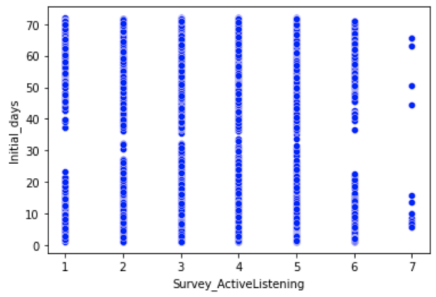
In [51]:
```
sns.scatterplot(x=medical_df['Survey_HoursTreatment'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



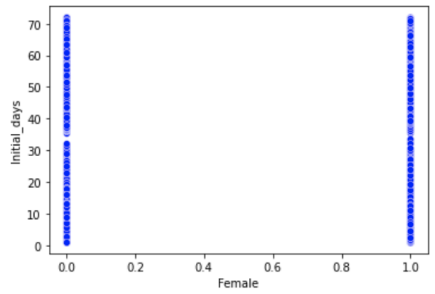
In [52]:
```
sns.scatterplot(x=medical_df['Survey_CourteousStaff'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



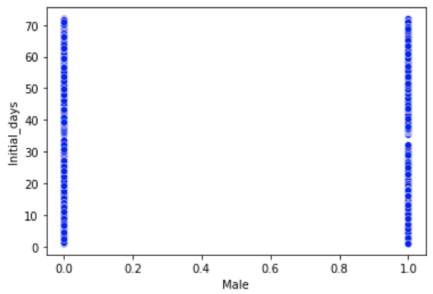
In [53]:
```
sns.scatterplot(x=medical_df['Survey_ActiveListening'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



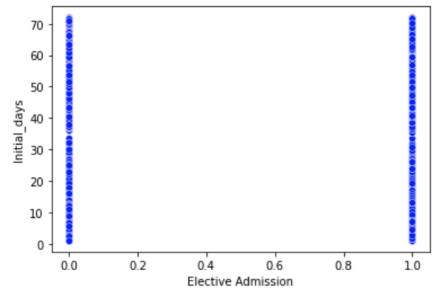
In [54]:
```
sns.scatterplot(x=medical_df['Female'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



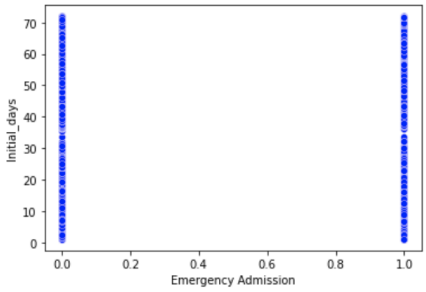
In [55]:
```
sns.scatterplot(x=medical_df['Male'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



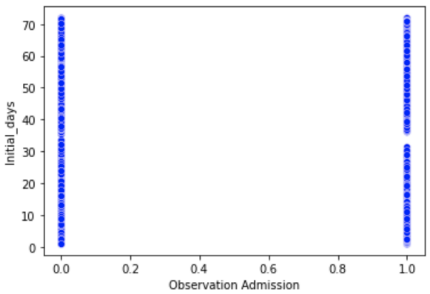
In [56]:
```
sns.scatterplot(x=medical_df['Elective Admission'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



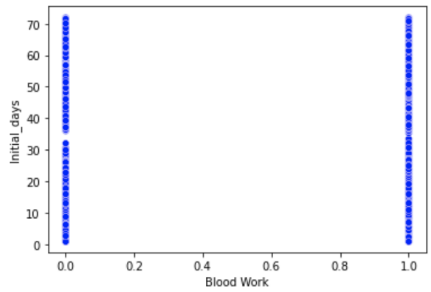
In [57]:
```
sns.scatterplot(x=medical_df['Emergency Admission'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



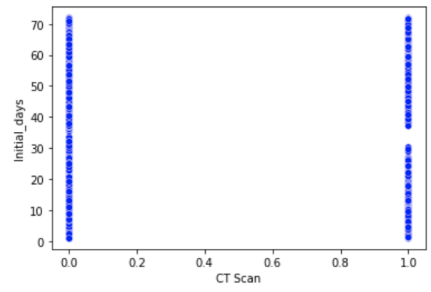
In [58]:
```
sns.scatterplot(x=medical_df['Observation Admission'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



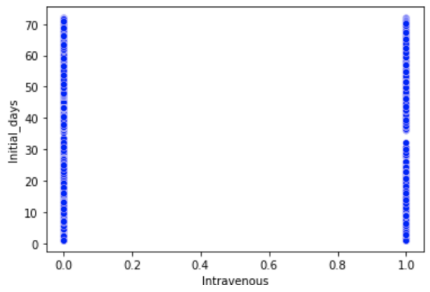
In [59]:
```
sns.scatterplot(x=medical_df['Blood Work'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



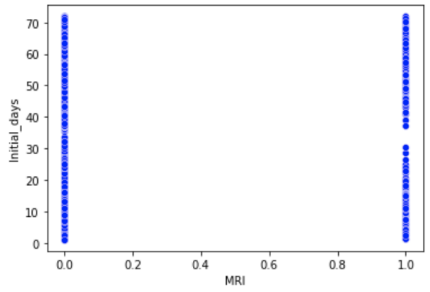
In [60]:
```
sns.scatterplot(x=medical_df['CT Scan'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



In [61]:
```
sns.scatterplot(x=medical_df['Intravenous'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



In [62]:
```
sns.scatterplot(x=medical_df['MRI'], y=medical_df['Initial_days'], color='blue')
plt.show();
```



3.5 Prepared Dataset

In [64]:
```
# Extract Clean dataset
medical_df.to_csv('medical_prepared.csv')
```

This can be found in the attached file named “medical_prepared.csv”.

4 Model Comparison

4.1 Initial Multiple Regression Models

In [64]:
```
# Initial estimated regression equation that could be used to predict the Initial_days, without dummy variables
medical_df['intercept'] = 1
lm_ReAdmis = sm.OLS(medical_df['Initial_days'], medical_df[['Children', 'Age', 'Income', 'VitD_levels', 'Doc_visits',
          'Full_meals_eaten', 'vitD_supp', 'TotalCharge', 'Additional_charges', 'Survey_TimelyAdmin',
        'Survey_TimelyTreatment', 'Survey_TimelyVisits', 'Survey_Reliability', 'Survey_Options', 'Survey_HoursTreatment
        'Survey_CourteousStaff', 'Survey_ActiveListening', 'intercept']]).fit()
print(lm_ReAdmis.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.977
Model:                            OLS   Adj. R-squared:                  0.977
Method:                 Least Squares   F-statistic:                 2.458e+04
Date:                Sat, 26 Feb 2022   Prob (F-statistic):               0.00
Time:                        01:07:47   Log-Likelihood:                -28097.
No. Observations:               10000   AIC:                         5.623e+04
Df Residuals:                    9982   BIC:                         5.636e+04
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Children                   -0.0117      0.019     -0.630      0.529      -0.048       0.025
Age                         0.0451      0.003     16.124      0.000       0.040       0.051
Income                  1.454e-06   1.41e-06      1.030      0.303   -1.31e-06    4.22e-06
VitD_levels                -0.0274      0.020     -1.373      0.170      -0.067       0.012
Doc_visits                 -0.0409      0.038     -1.064      0.287      -0.116       0.034
Full_meals_eaten           -0.0627      0.040     -1.569      0.117      -0.141       0.016
vitD_supp                  -0.0271      0.064     -0.424      0.672      -0.153       0.098
TotalCharge                 0.0119   1.85e-05    645.655      0.000       0.012       0.012
Additional_charges         -0.0002   8.82e-06    -22.660      0.000      -0.000      -0.000
Survey_TimelyAdmin         -0.0613      0.058     -1.057      0.290      -0.175       0.052
Survey_TimelyTreatment      0.0134      0.053      0.250      0.802      -0.091       0.118
Survey_TimelyVisits        -0.0265      0.049     -0.538      0.591      -0.123       0.070
Survey_Reliability         -0.0510      0.044     -1.160      0.246      -0.137       0.035
Survey_Options             -0.0189      0.046     -0.409      0.683      -0.110       0.072
Survey_HoursTreatment       0.0120      0.048      0.250      0.802      -0.082       0.106
Survey_CourteousStaff      -0.0588      0.045     -1.306      0.192      -0.147       0.029
Survey_ActiveListening      0.0371      0.042      0.874      0.382      -0.046       0.120
intercept                 -27.4525      0.565    -48.620      0.000     -28.559     -26.346
==============================================================================
Omnibus:                      831.505   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              291.078
Skew:                          -0.139   Prob(JB):                     6.21e-64
Kurtosis:                       2.212   Cond. No.                     7.16e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.16e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [65]:
```
medical_df_dummies = medical_df.columns
print(medical_df_dummies)
```

```
Index(['Children', 'Age', 'Income', 'ReAdmis', 'VitD_levels', 'Doc_visits',
       'Full_meals_eaten', 'vitD_supp', 'Soft_drink', 'HighBlood', 'Stroke',
       'Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain',
       'Anxiety', 'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma',
       'TotalCharge', 'Additional_charges', 'Survey_TimelyAdmin',
       'Survey_TimelyTreatment', 'Survey_TimelyVisits', 'Survey_Reliability',
       'Survey_Options', 'Survey_HoursTreatment', 'Survey_CourteousStaff',
       'Survey_ActiveListening', 'Female', 'Male', 'Elective Admission',
       'Emergency Admission', 'Observation Admission', 'High', 'Low', 'Medium',
       'Blood Work', 'CT Scan', 'Intravenous', 'MRI', 'Initial_days',
       'intercept'],
      dtype='object')
```

In [66]:
```
# Model including dummy variables
medical_df['intercept'] = 1
lm_ReAdmis = sm.OLS(medical_df['Initial_days'], medical_df[['Children', 'Age', 'Income', 'VitD_levels', 'Doc_visits',
        'Full_meals_eaten', 'vitD_supp', 'TotalCharge', 'Additional_charges', 'Survey_TimelyAdmin',
       'Survey_TimelyTreatment', 'Survey_TimelyVisits', 'Survey_Reliability', 'Survey_Options', 'Survey_HoursTreatment',
       'Survey_CourteousStaff', 'Survey_ActiveListening','Female', 'Male', 'Elective Admission', 'Emergency Admission',
       'Observation Admission', 'Blood Work', 'CT Scan', 'Intravenous', 'MRI','High', 'Low', 'Medium', 'intercept']]).
print(lm_ReAdmis.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 1.829e+05
Date:                Sat, 26 Feb 2022   Prob (F-statistic):               0.00
Time:                        01:07:47   Log-Likelihood:                -16043.
No. Observations:               10000   AIC:                         3.214e+04
Df Residuals:                    9973   BIC:                         3.233e+04
Df Model:                          26
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Children                9.001e-05      0.006      0.016      0.987      -0.011       0.011
Age                        0.0306      0.001     36.489      0.000       0.029       0.032
Income                 -3.322e-07   4.23e-07     -0.785      0.432   -1.16e-06    4.97e-07
VitD_levels                0.0110      0.006      1.831      0.067      -0.001       0.023
Doc_visits                 0.0025      0.012      0.217      0.828      -0.020       0.025
Full_meals_eaten          -0.0126      0.012     -1.054      0.292      -0.036       0.011
vitD_supp                  0.0077      0.019      0.400      0.689      -0.030       0.045
TotalCharge                0.0122   5.59e-06   2176.867      0.000       0.012       0.012
Additional_charges        -0.0001   2.65e-06    -52.477      0.000      -0.000      -0.000
Survey_TimelyAdmin        -0.0351      0.017     -2.018      0.044      -0.069      -0.001
Survey_TimelyTreatment     0.0018      0.016      0.114      0.909      -0.030       0.033
Survey_TimelyVisits        0.0247      0.015      1.669      0.095      -0.004       0.054
Survey_Reliability        -0.0258      0.013     -1.961      0.050      -0.052   -6.43e-06
Survey_Options             0.0073      0.014      0.523      0.601      -0.020       0.034
Survey_HoursTreatment      0.0219      0.014      1.531      0.126      -0.006       0.050
Survey_CourteousStaff      0.0114      0.013      0.847      0.397      -0.015       0.038
Survey_ActiveListening    -0.0028      0.013     -0.218      0.827      -0.028       0.022
Female                     0.1128      0.084      1.339      0.180      -0.052       0.278
Male                       0.1446      0.084      1.716      0.086      -0.021       0.310
Elective Admission        -3.0290      0.038    -80.105      0.000      -3.103      -2.955
Emergency Admission       -9.2182      0.037   -249.273      0.000      -9.291      -9.146
Observation Admission     -3.0221      0.038    -79.820      0.000      -3.096      -2.948
Blood Work                -3.8098      0.032   -118.821      0.000      -3.873      -3.747
CT Scan                   -3.8708      0.039    -98.254      0.000      -3.948      -3.794
Intravenous               -3.7869      0.034   -112.732      0.000      -3.853      -3.721
MRI                       -3.8017      0.055    -68.864      0.000      -3.910      -3.693
High                      -8.3796      0.037   -223.741      0.000      -8.453      -8.306
Low                       -3.4619      0.038    -89.941      0.000      -3.537      -3.386
Medium                    -3.4277      0.036    -94.075      0.000      -3.499      -3.356
intercept                -15.2692      0.099   -154.734      0.000     -15.463     -15.076
==============================================================================
Omnibus:                      123.339   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              122.436
Skew:                          -0.251   Prob(JB):                     2.59e-27
Kurtosis:                       2.797   Cond. No.                     1.25e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.66e-19. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Current multiple regression equation:

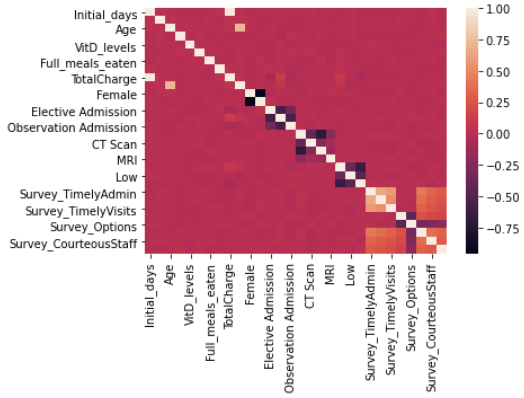
y = 9.001e-05*Children + .0306*Age - 3.322e-07*Income + .0110*VitD_levels + .0025*Doc_visits - .0126*Full_meals_eaten + .0077*vitD_supp + .0122*TotalCharge - .0001*Additional_charges - .0351*Survey_TimelyAdmin + .0018*Survey_TimelyTreatment + .0247*Survey_TimelyVisits - .0258*Survey_Reliability +.0073*Survey_Options + .0219*Survey_HoursTreatment + .0114*Survey_CourteousStaff -.0028*Survey_ActiveListening + .1128*Female + .1446*Male - 3.0290*Elective Admission - 9.2182*Emergency Admission - .3.0221*Observation Admission - 3.8098*Blood Work - 3.8708*CT Scan -

3.7869*Intravenous - 3.8017*MRI - 8.3796*High - 3.4277*Medium - 3.4619*Low -15.2692

The R^2 value = 0.998. So, 99% of the variation is explained by this model. The condition number is large which might suggest multicollinearity. All of these variables are not needed to explain the variance. Next, a heatmap for bivariate analysis & a principal component analysis will be created in order to reduce variables.¶

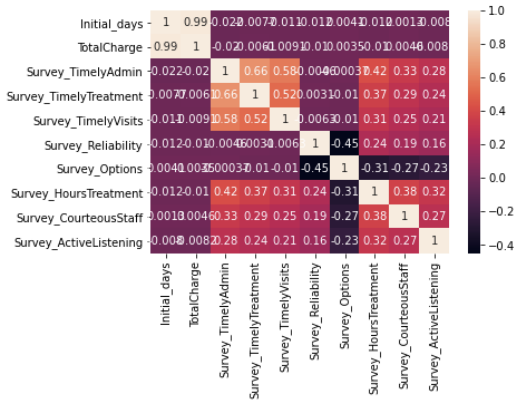
4.2 Justification

In [67]:
```
# Dataframe for heatmap bivariate analysis
medical_bivariate = medical_df[['Initial_days', 'Children', 'Age', 'Income', 'VitD_levels', 'Doc_visits',
          'Full_meals_eaten', 'vitD_supp', 'TotalCharge', 'Additional_charges','Female', 'Male',
        'Elective Admission', 'Emergency Admission', 'Observation Admission', 'Blood Work', 'CT Scan', 'Intravenous',
        'MRI','High', 'Low', 'Medium', 'Survey_TimelyAdmin', 'Survey_TimelyTreatment', 'Survey_TimelyVisits',
        'Survey_Reliability', 'Survey_Options', 'Survey_HoursTreatment','Survey_CourteousStaff',
        'Survey_ActiveListening']]
sns.heatmap(medical_bivariate.corr(), annot=False)
plt.show()
```



According to the heatmap, the strongest correlation for the target variable Initial_days is the TotalCharge variable, as well as some of the survey answers. So, the other variables will be removed except the correlating variables. So, the variables ['Initial_days', 'TotalCharge', 'Survey_TimelyAdmin', 'Survey_TimelyTreatment', 'Survey_TimelyVisits', 'Survey_Reliability', 'Survey_HoursTreatment','Survey_CourteousStaff', 'Survey_ActiveListening'] will stay since they seem to be the most relevant.

```
In [68]: bivariate = medical_df[['Initial_days','TotalCharge','Survey_TimelyAdmin', 'Survey_TimelyTreatment', 'Survey_TimelyVisit
         'Survey_Reliability', 'Survey_Options', 'Survey_HoursTreatment','Survey_CourteousStaff',
         'Survey_ActiveListening']]
         map(medical_bivariate.corr(), annot=True)
         )
```



The above heatmap indicates that for Initial_days, TotalCharge is the predictor for most of the variance. There is clearly a strong linear relationship between how many days a patient stays during their initial visit and how much the total charge a patient has. Next, a multiple linear regression will be run on the variables with .5 or above
The reduced regression equation will include the independent variables [TotalCharge, Survey_TimelyAdmin, Survey_TimelyTreatment, Survey_TimelyVisits]

4.3 Reduced Multiple Regression Model

```
In [69]: # Reduced multiple regression model
         medical_df['intercept'] = 1
         lm_ReAdmis_reduced = sm.OLS(medical_df['Initial_days'], medical_df[['TotalCharge', 'Survey_TimelyAdmin',
                                                                             'Survey_TimelyTreatment', 'Survey_TimelyVisits',
                                                                             'intercept']]).fit()
         print(lm_ReAdmis_reduced.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.975
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                 9.925e+04
Date:                Sat, 26 Feb 2022   Prob (F-statistic):               0.00
Time:                        01:07:48   Log-Likelihood:                -28355.
No. Observations:               10000   AIC:                         5.672e+04
Df Residuals:                    9995   BIC:                         5.676e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
TotalCharge    0.0119   1.89e-05    629.883      0.000       0.012       0.012
Survey_TimelyAdmin    -0.0689      0.057     -1.209      0.227      -0.180       0.043
Survey_TimelyTreatment     0.0062      0.054      0.114      0.910      -0.100       0.112
Survey_TimelyVisits    -0.0112      0.050     -0.223      0.824      -0.110       0.088
intercept    -28.5873      0.199   -143.886      0.000     -28.977     -28.198
==============================================================================
Omnibus:                      740.927   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              272.868
Skew:                          -0.135   Prob(JB):                     5.59e-60
Kurtosis:                       2.237   Cond. No.                     2.78e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.78e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

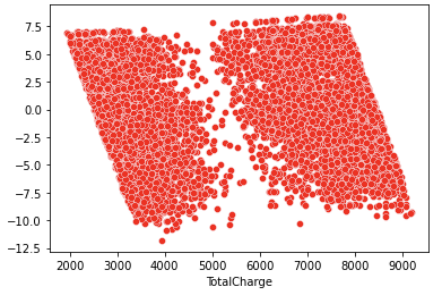
After removing the thos variables, the model still explains the variance at 72.4%.
Current multiple regression equation:

y = .0119*TotalCharge - .0689*Survey_TimelyAdmin + .0062*Survey_TimelyTreatment - .0112*Survey_TimelyVisits - 28.5873

5 Model Analysis

5.1 Models Comparison

```
In [70]: # Residual plot
medical_df = pd.read_csv('medical_prepared.csv')
medical_df['intercept'] = 1
residuals = medical_df['Initial_days'] - lm_ReAdmis_reduced.predict(medical_df[['TotalCharge',
                                                                                'Survey_TimelyAdmin', 'Survey_TimelyTreatmen
                                                                                'Survey_TimelyVisits', 'intercept']])

sns.scatterplot(x=medical_df['TotalCharge'],y=residuals,color='red')
plt.show();
```



5.2 Output

Output and calculations can be found above.

5.3 Code

All code for analysis can be found above.

6 Summary

6.1 Results of Analysis

- The final multiple regression equation has **5** independent variables:
  y = .0119*TotalCharge - .0689*Survey_TimelyAdmin + .0062*Survey_TimelyTreatment - .0112*Survey_TimelyVisits - 28.5873
- The coefficients suggest that for every 1 unit of:
  - TotalCharge - the initial days will increase by .0119
  - Survey_TimelyAdmin - the initial days will decrease by .0689
  - Survey_TimelyTreatment - the initial days will increase by .0062
  - Survey_TimelyVisits - the initial days will decrease by .0112

- P-values for TotalCharge, 0.000, indicate that it are statistically significant. While Survey_TimelyAdmin,Survey_TimelyTreatment, and Survey_TimelyVisits are not statistically significant given that their p-values are .227, .910, and .824 respectively.

6.2 Limitations of Analysis

Generally, the limitation of multiple regression is that correlation between variables does not always mean causation. This means that the relationships found are not conclusive to whether or not the independent variables have a direct effect on the dependable variable or if there are other unknown variables affecting both. It can also be noted that it is inconclusive to whether or not the dependent variable is in fact affecting the independent variables instead. Another limitation that can be found with this analysis is the dataset is small and could benefit from more data collection.

6.3 Recommended Course of Action

From this analysis, there is a significant linear relationship between the initial days a patient is admitted and the total amount charged for the initial visit. It could be assumed that it would be in the best interest of medical personnel to increase productivity of a hospital so as to decrease the initial days of a patient. Intuitively however, there are most likely some unknown variables affecting the relationship, as well as the knowledge that this analysis does not indicate causation. In this scenario, an unknown variable could be a pre-existing illness. A pre-existing illness would increase the days of an initial visit, increase the total amount charged to the patient due to the need of a specialist and most likely need followup appointments. Overall, there would need to be more data collected and further analysis done before any definitive conclusions can be drawn.

7 Supporting Documents

7.1 Video

This can be found within the attached file ‘Panopto Recording’.

7.2 Sources for Third-Party Code

Himamsh, Viveka. “How to Create Dummy Variables in Python with Pandas?” *GeeksforGeeks*, 8 Oct. 2021, https://www.geeksforgeeks.org/how-to-create-dummy-variables-in-python-with-pandas/.

“Python List Pop() Method.” Python List Pop() Method, https://www.w3schools.com/python/ref_list_pop.asp.

7.3 Sources

Gagner, David. (2022). D207 Exploratory Data Analysis . Salt Lake City ; Western Governors University.

Western Governors University. (n.d.). D207 D208 D209 Medical Data Considerations and Dictionary. Salt Lake City.