Multiple Regression – Medical Data

Benjamin Vickers

Western Governor's University

Abstract

This paper will provide a multiple regression model of medical raw data. A business question will be identified that would be a real-world situation or issue an organization may face. Multiple regression will be conducted to answer the business question, using Python. Visualizations will be provided to help the reader understand the regression and predictions. The code used for the regression and predictions will be provided. Lastly, the regression equation, statistical and practical significance, limitations, and a recommended course of action will be provided.

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:

1. **Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using multiple regression.**

One research question that would be relevant to a real-world organization, based on the dataset provided, would be "How do patient observations and initial days correlate?". The author aims to answer this question using multiple linear regression.

2. **Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.**

The goal of the data analysis is to predict the probability of readmission of a patient. This will be accomplished by determining which independent variables in the dataset are a good indicator and fit. Once the appropriate independent variables are determined, they will be used in the multiple linear regression to determine the probability of readmission. A multiple regression equation will be provided as well as the statistics for the variables identified.

Part II: Method Justification

B. Describe multiple regression methods by doing the following:

1. **Summarize the assumptions of a multiple regression model.**

There are several assumptions that should be reviewed and met for a multiple regression model to be a good fit. According to Statistics Solutions (2021) these include:

- Linear relationship – the relationship between the independent and dependent variables needs to be linear. This can be checked using scatterplots It predicts the probability of an outcome occurring.

- Errors should be normally distributed – errors between observed and predicted values should be normally distributed. This can be checked using a histogram or a Q-Q-Plot Explains the relationship between one dependent binary variable and one or more independent nominal variables.
- No multicollinearity – the independent variables should not be too highly correlated with each other. This can be checked using a correlation matrix or a variance inflation factor.
- Homoscedasticity – there should be no clear pattern in the distribution of data in a scatterplot of residuals versus predicted values.

2. Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.

Python will be the tool used for the multiple regression of the medical data. Python has several benefits. Python has several packages that allow for statistical analyses, such as Pandas, Scipy and Statsmodels. The syntax is not difficult to understand. Python allows for easy-to-understand visualizations of variables and the observations in them. It allows you to parse data easily into separate groups without changing the original dataset. Python also has commands specifically for multiple regression, such as linear_regression.fit, sm.OLS, and model.coef_. These make it simpler to run the regression analyses as well as visualize them.

3. Explain why multiple regression is an appropriate technique to analyze the research question summarized in Part I.

Multiple regression is an appropriate technique to analyze the previously noted research question. This analysis technique allows for each patient observation variable and determine the variables that correlate with initial days. Many of the variables are numeric independent

variables. These have the potential to impact the number of initial days a patient is admitted. The information provided from the analysis would allow the hospital to reduce the initial days a patient is admitted. Using multiple regression, the specific patient observations can be identified that correlate with initial days

Part III: Data Preparation

C. Summarize the data preparation process for multiple regression analysis by doing the following:

1. **Describe your data preparation goals and the data manipulations that will be used to achieve the goals.**

Before conducting the multiple regression, the data needs to be prepared. First any null or missing values need to be mitigated. This will include determining the best method of changing the values, whether changing null values to zeros or populating based on average values for the variable(s). Categorical variables will be converted to numerical variables so a linear regression can be conducted. Variables that are demographic data can be removed as they identify static data on the patients and cannot be changed by the hospital. These unnecessary variables will be removed. Any missing data will be identified and mitigated, and duplicate data will be removed if it exists.

2. **Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.**

In order for the multiple regression to answer the research question, summary statistics need to be identified. P-values for the independent variables need to be identified. Coefficients for the independent variables need to be identified as well. This will indicate which independent variables will impact the target variable. The target variable is the Initial_days variable. The predictor variables are patient observations, such as Services, Complication_risk, and Doc_visits.

In the tables below the numeric variables are shown with their summary statistics. We see in the table the standard deviations of each numerical variable as well as the dispersion in the interquartile ranges. We can see that Children, Full_meals_eaten, Income, and vitD_supp are not normally distributed variables, while Age, VitD_Levels, Doc_visits, and Initial_days are normally distributed.

	Children	Age	Income	Marital	Gender	ReAdmis	VitD_levels	Doc_visits	Full_meals_eaten	vitD_supp
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	0.202300	1.544600	0.366900	17.964262	5.012200	1.001400	0.398900
std	2.163659	20.638538	28521.153293	0.401735	0.539296	0.481983	2.017231	1.045734	1.008117	0.628505
min	0.000000	18.000000	154.080000	0.000000	1.000000	0.000000	9.806483	1.000000	0.000000	0.000000
25%	0.000000	36.000000	19598.775000	0.000000	1.000000	0.000000	16.626439	4.000000	0.000000	0.000000
50%	1.000000	53.000000	33768.420000	0.000000	2.000000	0.000000	17.951122	5.000000	1.000000	0.000000
75%	3.000000	71.000000	54296.402500	0.000000	2.000000	1.000000	19.347963	6.000000	2.000000	1.000000
max	10.000000	89.000000	207249.100000	1.000000	3.000000	1.000000	26.394449	9.000000	7.000000	5.000000

Soft_drink	Initial_admin	HighBlood	Stroke	Complication_risk	Overweight	Arthritis	Diabetes	Hyperlipidemia	BackPain
10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000
0.257500	2.255600	0.409000	0.199300	2.123300	0.709400	0.357400	0.27380	0.337200	0.411400
0.437279	0.831347	0.491674	0.399494	0.730172	0.454062	0.479258	0.44593	0.472777	0.492112
0.000000	1.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.00000	0.000000	0.000000
0.000000	1.000000	0.000000	0.000000	2.000000	0.000000	0.000000	0.00000	0.000000	0.000000
0.000000	3.000000	0.000000	0.000000	2.000000	1.000000	0.000000	0.00000	0.000000	0.000000
1.000000	3.000000	1.000000	0.000000	3.000000	1.000000	1.000000	1.00000	1.000000	1.000000
1.000000	3.000000	1.000000	1.000000	3.000000	1.000000	1.000000	1.00000	1.000000	1.000000

Anxiety	Allergic_rhinitis	Reflux_esophagitis	Asthma	Services	Initial_days	TotalCharge	Additional_charges	Timely_admis	Timely_treat
10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
0.321500	0.394100	0.413500	0.28930	1.862500	34.455299	5312.172769	12934.528587	3.518800	3.506700
0.467076	0.488681	0.492486	0.45346	0.986251	26.309341	2180.393838	6542.601544	1.031966	1.034825
0.000000	0.000000	0.000000	0.00000	1.000000	1.001981	1938.312067	3125.703000	1.000000	1.000000
0.000000	0.000000	0.000000	0.00000	1.000000	7.896215	3179.374015	7986.487755	3.000000	3.000000
0.000000	0.000000	0.000000	0.00000	1.000000	35.836244	5213.952000	11573.977735	4.000000	3.000000
1.000000	1.000000	1.000000	1.00000	3.000000	61.161020	7459.699750	15626.490000	4.000000	4.000000
1.000000	1.000000	1.000000	1.00000	4.000000	71.981490	9180.728000	30566.070000	8.000000	7.000000

Timely_visits	Reliability	Options	Hrs_treat	Courteous	Active_listen
10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
3.511100	3.515100	3.496900	3.522500	3.494000	3.509700
1.032755	1.036282	1.030192	1.032376	1.021405	1.042312
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
3.000000	3.000000	3.000000	3.000000	3.000000	3.000000
4.000000	4.000000	3.000000	4.000000	3.000000	3.000000
4.000000	4.000000	4.000000	4.000000	4.000000	4.000000
8.000000	7.000000	7.000000	7.000000	7.000000	7.000000

The categorical variables were also converted to numerical variables to that regression analysis can be conducted and statistical data is compared correctly. Visualizations of the variables can be seen below in Section 4. A summary of the statistics for these variables indicate that Age, Reflux_esophagitis, BackPain, and Gender are normally distrubuted, outside of the Nonbinary response in Gender. Responses for Marital, Soft_drink, Initial_admin, Stroke, Complication_risk, and Diabetes are separated by large margins. Additionally, we can see that the number of responses equal to No to ReAdmis are more prevalent than those with a Yes response. In terms of the bivariate analyses, one can see that Doc_visits and Initial_admin are normally distributed when compared against Initial_days, as is Complication_risk, Stroke and Diabetes. ReAdmi, Timely_admis, and Timely_visits are not normally distributed when compared against Initial_days.

3. Explain the steps used to prepare the data for the analysis, including the annotated code.

Before conducting the analysis, the data needs to be prepared. The first step is to ensure there is no missing data in any of the columns. Next, we will want to check to make sure that none of the data in the columns is duplicated. We will also want to make sure that none of the columns or rows are duplicated, so we will want to check that and ensure the result is a “False”. There are several columns in the dataset that were deemed irrelevant to the logistic analysis, such as patient

demographics that cannot be changed and are related to the interaction and location of the patient, so those should be dropped from the dataset. This makes working with the data easier. The categorical variables need to be converted to numerical, so any “yes/no” or other categorical options need to be changed to number values. Additionally, the survey columns need to be renamed to provide for more clear understanding and determination of applicable variables. Next is to identify the target variable and move to the first column of the dataset for easier visual cues.

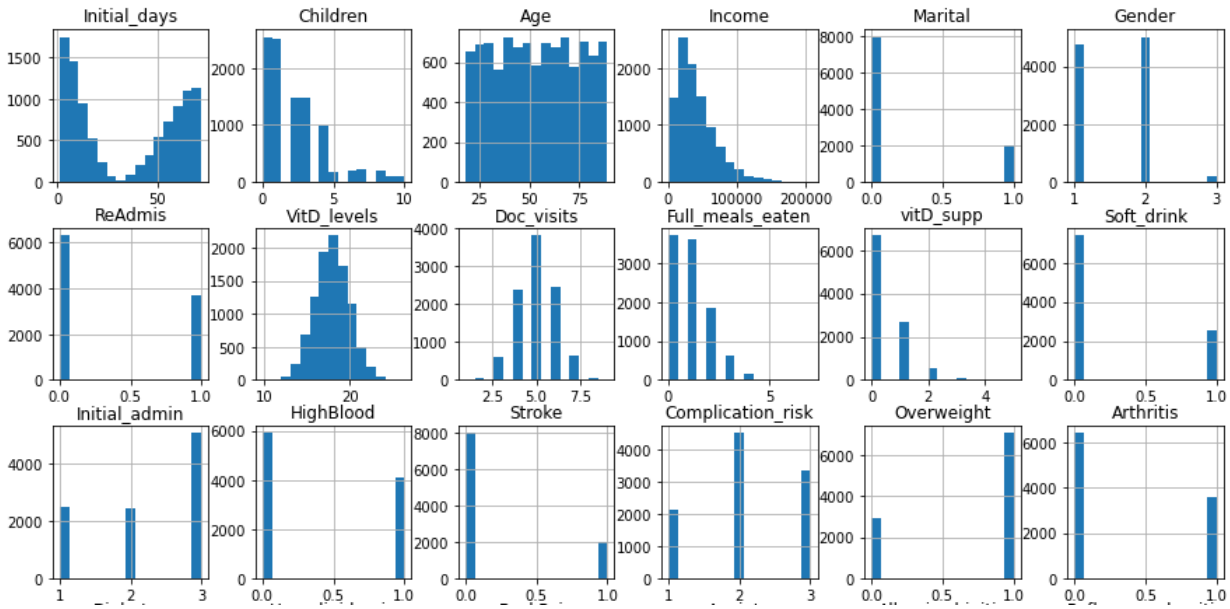
Code

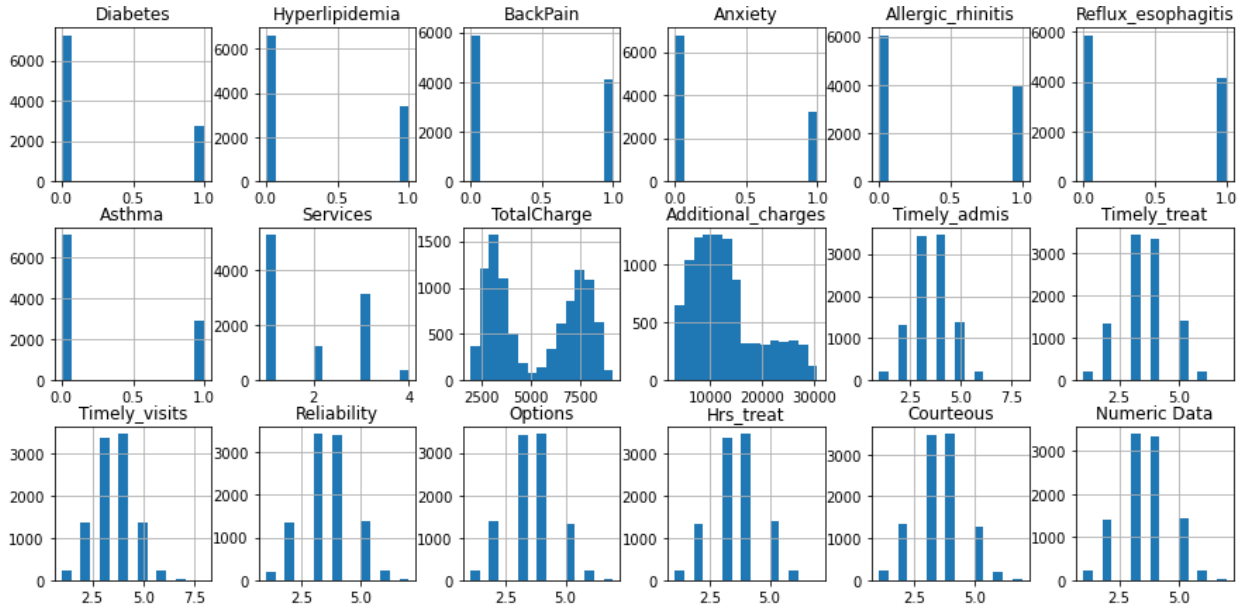
```
import numpy as np
import pandas as pd
from sklearn import linear_model
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
pd.set_option('display.max_columns', None)
import pylab
from pylab import rcParams
import statsmodels.api as sm
import statistics
from scipy import stats
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
from scipy.stats import chisquare
from scipy.stats import chi2_contingency
df = pd.read_csv (r'C:\Users\benvi\Desktop\medical_clean.csv')
df.rename(columns={'Item1':'Timely_admis','Item2':'Timely_treat',
                   'Item3':'Timely_visits','Item4':'Reliability',
                   'Item5':'Options','Item6':'Hrs_treat',
                   'Item7':'Courteous','Item8':'Active_listen'},inplace=True)
df.head()
df.info()
#check for missing data
```

```
df.isna().any()
#check for duplicate data in columns
df[df.duplicated()]
# check if any cols are duplicated - Looking for False
df.columns.duplicated().any()
# check if any rows are duplicated - looking for False
df.duplicated().any()
# drop demographic data
df =
df.drop(['CaseOrder','Customer_id','Interaction','UID','City','State','County','Zip','Lat','Lng','Popul
ation','Area','TimeZone','Job'], axis=1)
# verify columns were dropped
df.head()
#change yes/no to 1/0
df = df.replace(to_replace = ['Yes','No'],value = [1,0])
df['Gender'] = df['Gender'].replace(['Male','Female','Nonbinary'],[1,2,3])
#change Marital to "Married/Not Married", then change to integer 1/0
df['Marital'] = df['Marital'].replace(['Divorced','Widowed','Separated','Never Married'],'Not
Married')
df['Marital'] = df['Marital'].replace(['Married','Not Married'],[1,0])
#convert Initial_Admin, Complication_risk, and Services to integers
df['Initial_admin'] = df['Initial_admin'].replace(['Elective Admission','Observation
Admission','Emergency Admission'],[1,2,3])
df['Complication_risk'] = df['Complication_risk'].replace(['Low','Medium','High'],[1,2,3])
df['Services'] = df['Services'].replace(['Blood Work','CT Scan','Intravenous','MRI'],[1,2,3,4])
df.info()
df.describe()
my_list = df.columns.values.tolist()
print(my_list)
#Move target variable to beginning of columns
df=df[['Initial_days','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', 'TotalCharge', 'Additional_charges',
'Timely_admis', 'Timely_treat', 'Timely_visits', 'Reliability', 'Options', 'Hrs_treat', 'Courteous',
'Active_listen']]
#Verify target variable was moved
my_list = df.columns.values.tolist()
print(my_list)
#export prepared dataset
df.to_csv(r'C:\Users\benvi\Desktop\medical_PREPARED-TASK1.csv', index = False)
```

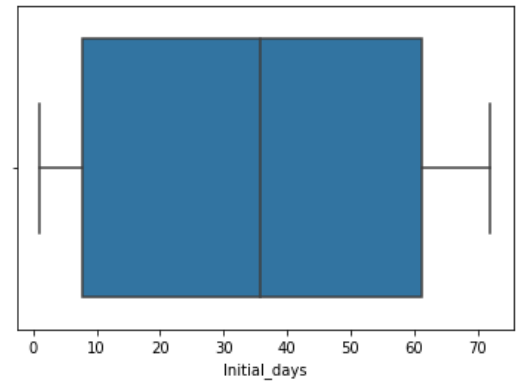
4. Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.

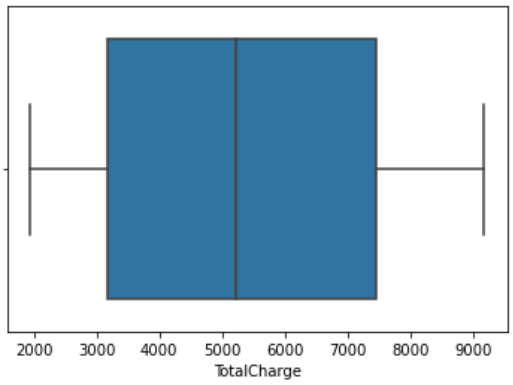
Visualizations of the variables that have been prepared and cleaned will help make the determination of which variables should be used in the linear regression model. Below are the histograms of the newly defined numerical data. These can be used to visualize the data and see if it even distributed.

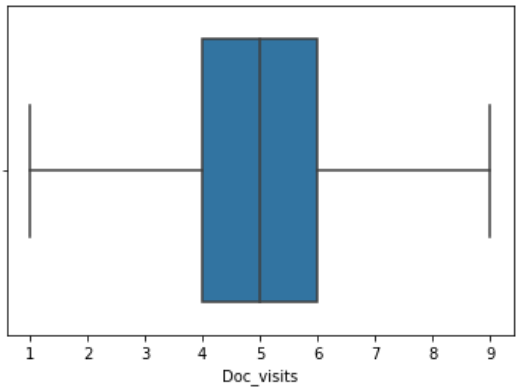




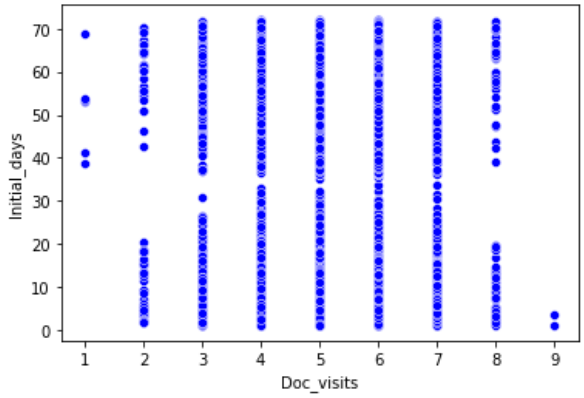
We can see that while there are several variables that are not evenly distributed. However, we can see that anomalies and outliers have been removed in the box plots below:

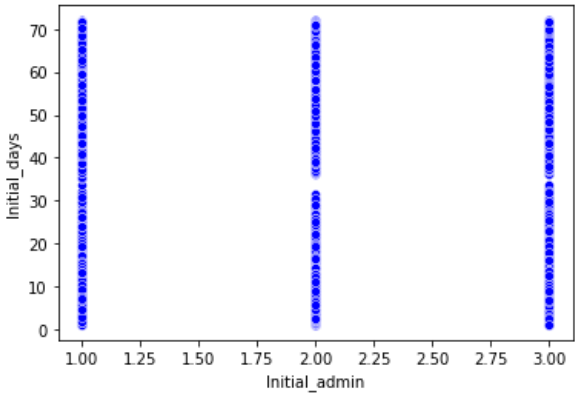


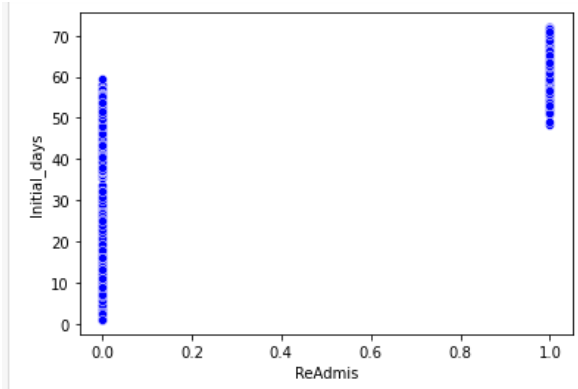


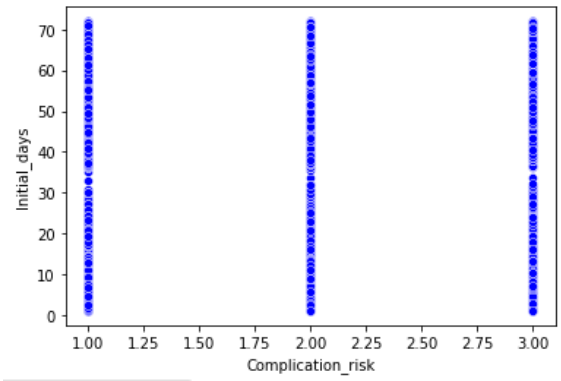


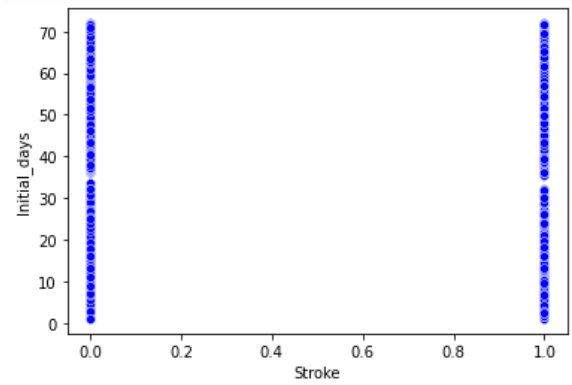
Bivariate analyses can be visualized as well. Below the target Initial_days is compared with potential predictor in scatterplots.

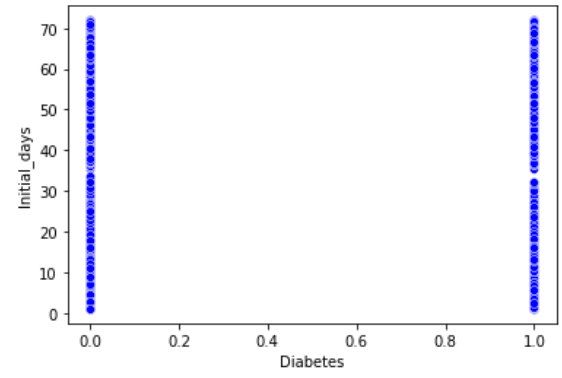


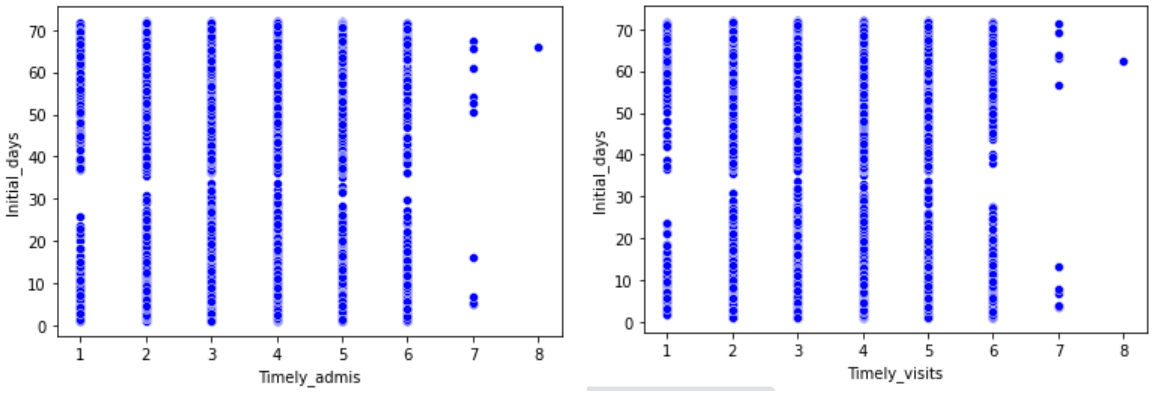












5. **Provide a copy of the prepared data set.**

A copy of the prepared dataset has been uploaded in the task evaluation.

Part IV: Model Comparison and Analysis

D. Compare an initial and a reduced multiple regression model by doing the following:

1. **Construct an initial multiple regression model from *all* predictors that were identified in Part C2.**

An initial regression will be run on potential predictor variables. These are compared against the target of Initial_days. The OLS Regression Results can be seen below:

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.994
Model:	OLS	Adj. R-squared:	0.994
Method:	Least Squares	F-statistic:	9.209e+04
Date:	Sat, 27 Nov 2021	Prob (F-statistic):	0.00
Time:	17:53:58	Log-Likelihood:	-21027.
No. Observations:	10000	AIC:	4.209e+04
Df Residuals:	9980	BIC:	4.224e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Age	0.0197	0.003	6.780	0.000	0.014	0.025
ReAdmis	0.9935	0.078	12.755	0.000	0.841	1.146
Doc_visits	0.0197	0.019	1.037	0.300	-0.018	0.057
Initial_admin	-3.3450	0.024	-137.279	0.000	-3.393	-3.297
HighBlood	-0.4696	0.112	-4.175	0.000	-0.690	-0.249
Stroke	0.0806	0.050	1.617	0.106	-0.017	0.178
Complication_risk	-2.7188	0.028	-98.534	0.000	-2.773	-2.665
Diabetes	-0.9084	0.045	-20.395	0.000	-0.996	-0.821
Anxiety	-1.0138	0.043	-23.827	0.000	-1.097	-0.930
Allergic_rhinitis	-0.8316	0.041	-20.459	0.000	-0.911	-0.752
Reflux_esophagitis	-0.7705	0.040	-19.103	0.000	-0.850	-0.691
Asthma	0.0236	0.044	0.540	0.589	-0.062	0.109
Services	0.0096	0.020	0.475	0.635	-0.030	0.049
TotalCharge	0.0119	1.73e-05	687.854	0.000	0.012	0.012
Additional_charges	-9.473e-05	1.22e-05	-7.786	0.000	-0.000	-7.09e-05
Timely_admis	-0.0339	0.027	-1.279	0.201	-0.086	0.018
Timely_treat	0.0089	0.026	0.347	0.729	-0.041	0.059
Hrs_treat	0.0138	0.022	0.629	0.529	-0.029	0.057
Active_listen	0.0236	0.020	1.158	0.247	-0.016	0.064
intercept	-14.5799	0.186	-78.567	0.000	-14.944	-14.216

Omnibus:	284.824	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	223.140
Skew:	0.283	Prob(JB):	3.51e-49
Kurtosis:	2.535	Cond. No.	1.47e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.47e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Variables removed from the initial model included minor observations/variables such as VitD_levels, Income, Marital, Gender, Full_meals_eaten, vitD_supp, and Soft_drink. The initial model equation is as follows:

$$\hat{Y} = -14.5799 + 0.0197(Age) + 0.9935(ReAdmis) + 0.0197(Doc\_visits) - 3.3450(Initial\_admin) - 0.4696(HighBlood) + 0.0806(Stroke) - 2.7188(Complication\_risk) - 0.9084(Diabetes) - 1.0138(Anxiety) - 0.8316(Allergic\_rhinitis) - 0.7705(Reflux\_esophagitis) + 0.0236(Asthma) + 0.0096(Services) + 0.0119(TotalCharge) - 9.473(Additional\_charges) - 0.0339(Timely\_admis) + 0.0089(Timely\_treat) + 0.0138(Hrs\_treat) + 0.0236(Active\_listen).$$

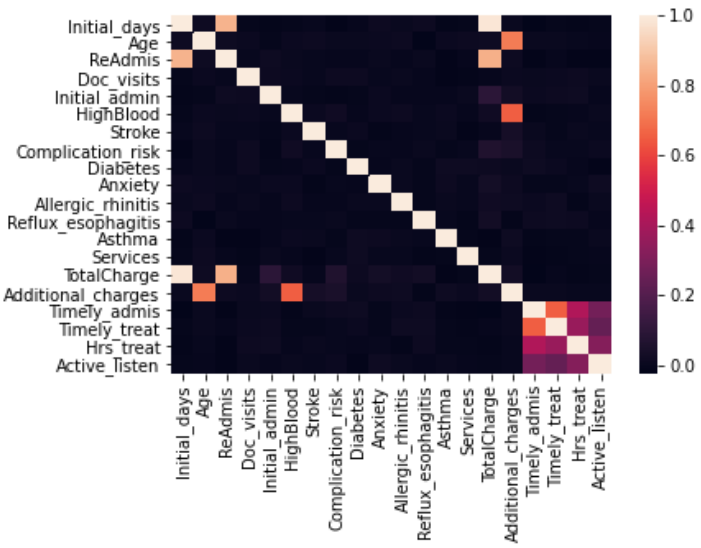
This initial model has an R-squared value of 0.99, meaning 99% of the variation can be explained by this model. The condition number is large, suggesting strong multicollinearity, meaning all variables may not be needed. A heatmap and correlation matrix can be used to visualize where there may be multicollinearity and start to narrow down which variables to use in the reduced model.

2. **Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.**

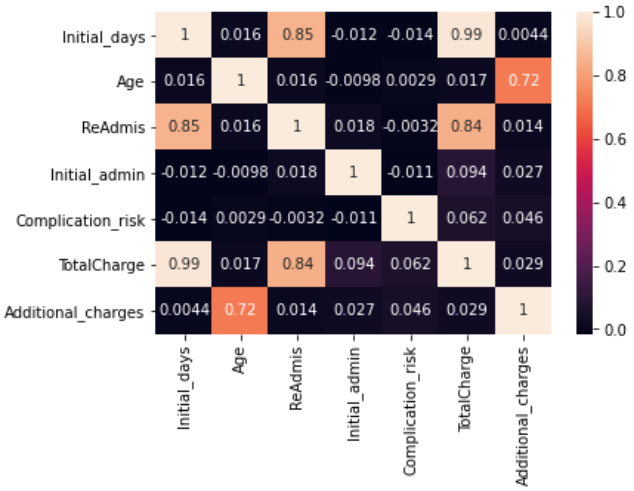
A correlation matrix can be used to help determine the best variables for a reduced regression model. A heatmap will help to visualize this data as well. These can be seen below:

	Age	ReAdmis	Doc_visits	Initial_admin	HighBlood	Stroke	Complication_risk	Diabetes	Anxiety	Allergic_rhinitis	I
Age	1.000000	0.015810	0.006898	-0.009763	0.007147	0.012035	0.002887	0.003694	0.006130	0.012092	
ReAdmis	0.015810	1.000000	0.000246	0.017522	0.002270	0.000918	-0.003236	-0.003058	0.002406	-0.004651	
Doc_visits	0.006898	0.000246	1.000000	0.012518	0.008967	-0.002230	0.012306	0.012781	-0.001684	0.002920	
Initial_admin	-0.009763	0.017522	0.012518	1.000000	0.001369	-0.008856	-0.011229	-0.009667	0.008305	-0.005741	
HighBlood	0.007147	0.002270	0.008967	0.001369	1.000000	0.007568	0.021368	-0.005858	0.008303	0.011709	
Stroke	0.012035	0.000918	-0.002230	-0.008856	0.007568	1.000000	0.001119	0.005792	-0.013801	-0.004837	
Complication_risk	0.002887	-0.003236	0.012306	-0.011229	0.021368	0.001119	1.000000	-0.006633	-0.000707	-0.002782	
Diabetes	0.003694	-0.003058	0.012781	-0.009667	-0.005858	0.005792	-0.006633	1.000000	-0.002529	0.005486	
Anxiety	0.006130	0.002406	-0.001684	0.008305	0.008303	-0.013801	-0.000707	-0.002529	1.000000	0.004368	
Allergic_rhinitis	0.012092	-0.004651	0.002920	-0.005741	0.011709	-0.004837	-0.002782	0.005486	0.004368	1.000000	
Reflux_esophagitis	-0.019609	0.005422	-0.005330	-0.004618	0.001150	-0.000054	0.003102	-0.007816	-0.007566	-0.007731	
Asthma	0.009229	-0.017133	-0.017989	-0.005956	0.006174	0.002443	-0.008973	0.016765	0.011758	0.004454	
Services	0.012016	-0.005578	-0.010785	0.003836	-0.003016	-0.016236	-0.001870	0.017390	0.000882	-0.001061	
TotalCharge	0.016876	0.843726	-0.005043	0.094157	0.019910	-0.003694	0.061834	0.011524	0.031199	0.018919	
Additional_charges	0.716854	0.013620	0.008072	0.026720	0.654316	0.035140	0.045901	0.002450	0.011666	0.016154	
Timely_admis	0.005552	-0.016785	0.003680	0.006172	-0.011017	0.001948	0.012386	0.013806	-0.007458	0.009402	
Timely_treat	0.003967	-0.002423	0.006024	0.011959	-0.007745	-0.007706	0.000032	0.005994	-0.009733	0.014654	
Hrs_treat	-0.002087	-0.016894	0.012530	0.016487	-0.002369	0.004282	-0.001094	-0.004259	-0.002248	-0.012721	
Active_listen	-0.003367	-0.016740	0.004571	-0.003092	0.002601	0.000040	0.007824	-0.014752	0.014650	0.005355	

Reflux_esophagitis	Asthma	Services	TotalCharge	Additional_charges	Timely_admis	Timely_treat	Hrs_treat	Active_listen
-0.019609	0.009229	0.012016	0.016876	0.716854	0.005552	0.003967	-0.002087	-0.003367
0.005422	-0.017133	-0.005578	0.843726	0.013620	-0.016785	-0.002423	-0.016894	-0.016740
-0.005330	-0.017989	-0.010785	-0.005043	0.008072	0.003680	0.006024	0.012530	0.004571
-0.004618	-0.005956	0.003836	0.094157	0.026720	0.006172	0.011959	0.016487	-0.003092
0.001150	0.006174	-0.003016	0.019910	0.654316	-0.011017	-0.007745	-0.002369	0.002601
-0.000054	0.002443	-0.016236	-0.003694	0.035140	0.001948	-0.007706	0.004282	0.000040
0.003102	-0.008973	-0.001870	0.061834	0.045901	0.012386	0.000032	-0.001094	0.007824
-0.007816	0.016765	0.017390	0.011524	0.002450	0.013806	0.005994	-0.004259	-0.014752
-0.007566	0.011758	0.000882	0.031199	0.011666	-0.007458	-0.009733	-0.002248	0.014650
-0.007731	0.004454	-0.001061	0.018919	0.016154	0.009402	0.014654	-0.012721	0.005355
1.000000	-0.001458	-0.013680	0.026284	-0.011405	0.011367	0.017425	0.009729	-0.003236
-0.001458	1.000000	-0.008769	-0.014290	0.014083	-0.011303	-0.007648	-0.009740	0.002209
-0.013680	-0.008769	1.000000	-0.007425	0.009232	-0.014312	-0.010317	-0.007029	-0.002837
0.026284	-0.014290	-0.007425	1.000000	0.029256	-0.019706	-0.006055	-0.010480	-0.008250
-0.011405	0.014083	0.009232	0.029256	1.000000	0.002423	0.002815	-0.000448	-0.000467
0.011367	-0.011303	-0.014312	-0.019706	0.002423	1.000000	0.655578	0.421233	0.278067
0.017425	-0.007648	-0.010317	-0.006055	0.002815	0.655578	1.000000	0.366075	0.242962
0.009729	-0.009740	-0.007029	-0.010480	-0.000448	0.421233	0.366075	1.000000	0.319886
-0.003236	0.002209	-0.002837	-0.008250	-0.000467	0.278067	0.242962	0.319886	1.000000



This heatmap can be used to find variables that correlate with Initial_days. We can note by this heatmap that there are several variables that can be removed. The correlation matrix and heatmap help to identify variables that may not be strong variables for predictors. Based on the heatmap, it appears that ReAdmis and TotalCharge are strong predictors. To narrow the list to other potential variables, diagnosis and survey variables will be removed. Age and Complication_risk are left as they societally tend to lead to longer admission days. The new heatmap can be seen below:



ReAdmis and TotalCharge are still very strong predictors for most of the variance. There appears to be a linear relationship between the number of initial days a patient was admitted and their potential readmission and high total charges. A multiple linear regression model will be run on these reduced variables. These include both categorical and continuous variables.

3. **Provide a reduced multiple regression model that includes *both* categorical and continuous variables.**

Note: The output should include a screenshot of each model.

A reduced multiple regression model can be run using the identified variables above. Below is the reduced OLS Regression Results as identified in the correlation matrix and the heat map:

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.993
Model:	OLS	Adj. R-squared:	0.993
Method:	Least Squares	F-statistic:	2.478e+05
Date:	Sat, 27 Nov 2021	Prob (F-statistic):	0.00
Time:	19:34:52	Log-Likelihood:	-21843.
No. Observations:	10000	AIC:	4.370e+04
Df Residuals:	9993	BIC:	4.375e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Age	0.0308	0.001	20.537	0.000	0.028	0.034
ReAdmis	1.2333	0.084	14.650	0.000	1.068	1.398
Initial_admin	-3.3141	0.026	-126.496	0.000	-3.365	-3.263
Complication_risk	-2.6898	0.030	-90.460	0.000	-2.748	-2.632
TotalCharge	0.0119	1.87e-05	633.739	0.000	0.012	0.012
Additional_charges	-0.0001	4.73e-06	-30.383	0.000	-0.000	-0.000
intercept	-15.6501	0.120	-130.377	0.000	-15.885	-15.415

Omnibus:	138.342	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	112.748
Skew:	0.189	Prob(JB):	3.29e-25
Kurtosis:	2.642	Cond. No.	8.98e+04

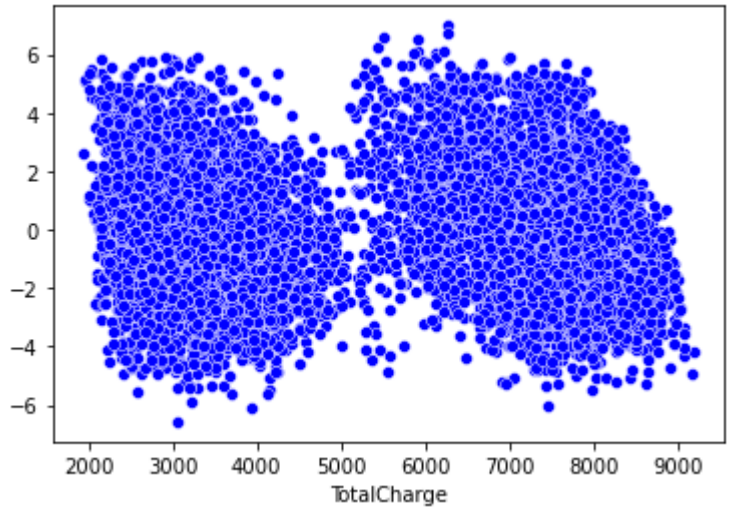
As you can see, the reduced model still accounts for 99.3% of the variance. The multiple linear regression model equation is as follows:

$\hat{Y}$= -15.6501 + 0.0308(Age) + 1.2333(ReAdmis) – 3.3141(Initial_admin) – 2.6898(Complication_risk + 0.0119(TotalCharge) – 0.0001(Additional_charges).

E. Analyze the data set using your reduced multiple regression model by doing the following:

1. **Explain your data analysis process by comparing the initial and reduced multiple regression models, including the following elements:**
   - **the logic of the variable selection technique**
   - **the model evaluation metric**
   - **a residual plot**

The technique for the variable selection was based on the results of the correlation matrix and mapping the variables using a heatmap. This identified the variables that correlated the most with the Initial_days variable. The model evaluation metric can be seen in the regression results listed above with the model equation and analysis, including the R-squared values. Below is the residual plot for the model:



2. **Provide the output and *any* calculations of the analysis you performed, including the model's residual error.**

Note: The output should include the predictions from the refined model you used to perform the analysis.

The output of the calculations as well as the model's residual error is noted above in the tables and visualizations.

3. **Provide the code used to support the implementation of the multiple regression models.**

The code to support the implementation of the multiple regression models can be found in Exhibit A below, after the Citations.

Part V: Data Summary and Implications

F. Summarize your findings and assumptions by doing the following:

1. **Discuss the results of your data analysis, including the following elements:**

- **a regression equation for the reduced model**

The final multiple linear regression equation for the reduced model is as follows:

$$\hat{Y}=-15.6501 + 0.0308(\text{Age}) + 1.2333(\text{ReAdmis}) - 3.3141(\text{Initial\_admin}) - 2.6898(\text{Complication\_risk} + 0.0119(\text{TotalCharge}) - 0.0001(\text{Additional\_charges}).$$

- **an interpretation of coefficients of the statistically significant variables of the model**

The coefficients of the statistically significant variables from the reduced model show a high correlation with the Initial_days. The coefficients suggest that for every 1 unit of:

- Age – Initial_days will increase by 0.0308 units
- ReAdmis – Initial_days will increase by 1.2333 units
- Initial_admin – Initial_days will decrease by 3.3141 units
- Complication_risk – Initial_days will decrease by 2.6898 units
- TotalCharge – Initial_days will increase by 0.0119 units
- Additional_charges – Initial_days will decrease by 0.0001 units

The p-values for the above listed variables are statistically significant at 0.000.

- **the statistical and practical significance of the model**

The multiple regression model has both statistical and practical significance. It indicates that there are variables that impact the initial days a patient is admitted, and that as these variables increase or decrease, the initial days will also increase or decrease. It identifies there is a strong correlation between a patient being readmitted and high total charges with the Initial days. From a practical standpoint, it identified predictor variables for the hospital to monitor in order to predict if a patient will be readmitted. If a patient has high total charges associated with their admittance and they are readmitted, the initial days will likely be high as well. It also indicates that if a patient is readmitted, there is a high likelihood of the total charges and initial days being high.

- **the limitations of the data analysis**

This analysis does have limitations. It looks at six specific variables that have a p-value of 0.000, meaning they are statistically significant. Other variables showed a low p-value as well and were not considered in this analysis. The values, independent variables and coefficients may change based on the dependent variable that is chose by the analyst. The results could be different based on the change in the dependent variable. In other words, only one multiple regression model was run, and the results could differ based on a different model. The results are only predictions, and while there is a strong confidence, there may be other variables that are not captured by the hospital that could impact and skew the results. There could be other variables that are not tracked by the hospital that may be better variables to use to conduct the analysis, so

this is limited to only the datapoints the hospital has deemed important to capture. The analysis only provides a prediction, not a definitive answer.

2. **Recommend a course of action based on your results.**

Based on the analysis conducted, the hospital should work to reduce the potential of a patient being readmitted and total charges, which are strong predictors related to the initial days. The longer a patient is admitted, the higher the total charges to the hospital, and the higher the likelihood of the patient being readmitted. Focusing efforts of the doctors and nurses to make more timely and accurate diagnoses can reduce this variable. Additionally, ensuring the appropriate tests are conducted in a timely fashion can impact this variable. An additional multiple regression should be conducted on other variables to determine what other factors play a part in a patient being readmitted, increasing the initial days, and increasing the total charges. There is a strong likelihood that other conditions impact the initial days, such as being the complication risk and age of the patient. These variables are outside the control of the hospital and patient. Ensuring that a patient is initially admitted under the correct classification may help reduce the initial days as well.

Citations

Unknown. (2021, August 11). *Assumptions of multiple linear regression*. Statistics Solutions. Retrieved November 25, 2021, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-multiple-linear-regression/.

Exhibit A – Full Code Used

```
import numpy as np
import pandas as pd
from sklearn import linear_model
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
pd.set_option('display.max_columns', None)
import pylab
from pylab import rcParams
import statsmodels.api as sm
import statistics
from scipy import stats
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
from scipy.stats import chisquare
from scipy.stats import chi2_contingency
df = pd.read_csv (r'C:\Users\benvi\Desktop\medical_clean.csv')
df.rename(columns={'Item1':'Timely_admis','Item2':'Timely_treat',
                   'Item3':'Timely_visits','Item4':'Reliability',
                   'Item5':'Options','Item6':'Hrs_treat',
                   'Item7':'Courteous','Item8':'Active_listen'},inplace=True)
df.head()
df.info()
#check for missing data
df.isna().any()
#check for duplicate data in columns
df[df.duplicated()]
# check if any cols are duplicated - Looking for False
df.columns.duplicated().any()
# check if any rows are duplicated - looking for False
df.duplicated().any()
# drop demographic data
df =
        df.drop(['CaseOrder','Customer_id','Interaction','UID','City','State','County','Zip','Lat','Lng','
        Population','Area','TimeZone','Job'], axis=1)
# verify columns were dropped
```

```
df.head()
#change yes/no to 1/0
df = df.replace(to_replace = ['Yes','No'],value = [1,0])
df
#Convert genders to number
df['Gender'] = df['Gender'].replace(['Male','Female','Nonbinary'],[1,2,3])
#change Marital to "Married/Not Married", then change to integer 1/0
df['Marital'] = df['Marital'].replace(['Divorced','Widowed','Separated','Never Married'],'Not
    Married')
df['Marital'] = df['Marital'].replace(['Married','Not Married'],[1,0])
#convert Initial_Admin, Complication_risk, and Services to integers
df['Initial_admin'] = df['Initial_admin'].replace(['Elective Admission','Observation
    Admission','Emergency Admission'],[1,2,3])
df['Complication_risk'] = df['Complication_risk'].replace(['Low','Medium','High'],[1,2,3])
df['Services'] = df['Services'].replace(['Blood Work','CT Scan','Intravenous','MRI'],[1,2,3,4])
df.info()
df.describe()
my_list = df.columns.values.tolist()
print(my_list)
#Move target variable to beginning of columns
df=df[['Initial_days','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',
    'TotalCharge', 'Additional_charges', 'Timely_admis', 'Timely_treat', 'Timely_visits',
    'Reliability', 'Options', 'Hrs_treat', 'Courteous', 'Active_listen']]
#Verify target variable was moved
my_list = df.columns.values.tolist()
print(my_list)
#export prepared dataset
df.to_csv(r'C:\Users\benvi\Desktop\medical_PREPARED-TASK1.csv', index = False)
# Columns for numerical data
NumericalData = df.select_dtypes(include = "number").columns
print (NumericalData)
# histogram plot numeric data
fig = plt.figure(figsize=(10, 20))
ax = df[NumericalData].hist(bins = 15, figsize=(15,15))
plt.title('Numeric Data')
fig.tight_layout(h_pad=5, w_pad=5)
plt.show()
# Create boxplots for continuous variables
sns.boxplot('Initial_days', data = df)
plt.show()
```

```
sns.boxplot('TotalCharge', data = df)
plt.show()
sns.boxplot('Doc_visits', data = df)
plt.show()
# Create scatterplots to show relationships between target variable and potential predictor
    variables
sns.scatterplot(x=df['Doc_visits'],y=df['Initial_days'],color='blue')
plt.show();
sns.scatterplot(x=df['Initial_admin'],y=df['Initial_days'],color='blue')
plt.show();
sns.scatterplot(x=df['ReAdmis'],y=df['Initial_days'],color='blue')
plt.show();
sns.scatterplot(x=df['Complication_risk'],y=df['Initial_days'],color='blue')
plt.show();
sns.scatterplot(x=df['Stroke'],y=df['Initial_days'],color='blue')
plt.show();
sns.scatterplot(x=df['Diabetes'],y=df['Initial_days'],color='blue')
plt.show();
sns.scatterplot(x=df['Timely_admis'],y=df['Initial_days'],color='blue')
plt.show();
sns.scatterplot(x=df['Timely_visits'],y=df['Initial_days'],color='blue')
plt.show();
df['intercept'] = 1
lm_initialdays = sm.OLS(df['Initial_days'],df[['Age',
    'ReAdmis', 'Doc_visits', 'Initial_admin', 'HighBlood', 'Stroke',
    'Complication_risk', 'Diabetes','Anxiety', 'Allergic_rhinitis',
    'Reflux_esophagitis', 'Asthma', 'Services', 'TotalCharge',
    'Additional_charges', 'Timely_admis', 'Timely_treat',
    'Hrs_treat','Active_listen','intercept']]).fit()

print(lm_initialdays.summary())
#heatmap and correlatin matrix dataframe creation
medical_heatmap = df[['Initial_days','Age',
    'ReAdmis', 'Doc_visits', 'Initial_admin', 'HighBlood', 'Stroke',
    'Complication_risk', 'Diabetes','Anxiety', 'Allergic_rhinitis',
    'Reflux_esophagitis', 'Asthma', 'Services', 'TotalCharge',
    'Additional_charges', 'Timely_admis', 'Timely_treat',
    'Hrs_treat','Active_listen']]
#Initial model heatmap
sns.heatmap(medical_heatmap.corr(), annot=False)
plt.show
medical_heatmap.corr()
#Narrow results, removing diagnosis and survey variables
```

```
medical_heatmap = df[['Initial_days','Age',
      'ReAdmis', 'Initial_admin','Complication_risk',
      'TotalCharge','Additional_charges']]
#Reduced Initial model heatmap
sns.heatmap(medical_heatmap.corr(), annot=True)
plt.show
#Reduced multiple regression model
df['intercept'] = 1
lm_initialdays_reduced = sm.OLS(df['Initial_days'],df[['Age',
      'ReAdmis','Initial_admin',
      'Complication_risk','TotalCharge',
      'Additional_charges','intercept']]).fit()

print(lm_initialdays_reduced.summary())
#load cleansed data for residual plot
med_df = pd.read_csv (r'C:\Users\benvi\Desktop\medical_PREPARED-TASK1.csv')
#Create residual plot
med_df['intercept'] = 1
residuals = med_df['Initial_days'] -
        lm_initialdays_reduced.predict(med_df[['Age','ReAdmis','Initial_admin','Complication_ris
        k',
            'TotalCharge','Additional_charges','intercept']])
sns.scatterplot(x=med_df['TotalCharge'],y=residuals,color='blue')
plt.show();
```