Predictive Modeling – D208
Task 1

Western Governor's University
Performance Assessment
Matthew Morgan
Student ID: 010471280
11/21/2022

Part I: Research Question
A1. What patient demographics and medical conditions lead to longer initial hospital stays?

A2. Ultimately, we want to identify the variables that cause patients to be readmitted. However, when running a bivariate analysis on patient readmissions nothing shows a relationship except initial days. By looking at other variables that might have a relationship with initial days we can find a trend or pattern in patients who end up back in the hospital then we can start looking for ways to reduce readmissions and ultimately reduce the chances of receiving a fine.

Part II: Method Justification
B1. Multiple linear regression makes several assumptions. The first being that there must be a linear relationship between the outcome variable and independent variables. It also assumes that the residuals are normally distributed and that the independent variables are not highly correlated with each other.
B2. I have decided to use Python for the entirety of the project. I am familiar with it, was provided with the necessary education on datacamp, and it's versatile enough to produce all the output needed for this assessment.
B3. Multiple linear regression is used to estimate the relationship between a dependent variable and two or more independent variables using a straight line. It allows us to visualize the relationship between many different variables and determine possible relationships.

Part III: Data Preparation
C1. To prepare and manipulate data for this project first we check for missing and duplicate values. None were found. Then we try and convert categorical variables to numerical variables so they can be used for analysis. For this I used pd.get_dummies to save time and ensure we mitigate multicollinearity.
C2. The target variable I chose was initial_days, as it is a continuous variable and can be used for Multiple Linear Regression. The predictor variables I chose were the following; vitD_supp, Children, Income, Full_Meals_Eaten, Additional_charges, TotalCharge, VitD_levels, Age, Doc_Visits, HighBlood, Stroke, Arthritis, Diabetes, Hyperlipidemia, BackPain, Allergic_rhinitis, Reflux_esophagitis, Asthma, Marital, Services, Gender, Initial_admin, and Complication_risk.

I included screenshots of summary statistics below:

Summary statistics

	Count	Missing	Unique	Dtype	Numeric	Mean
Population	10000	0	5951	int64	True	9965.2538
vitD_supp	10000	0	6	int64	True	0.3989
Children	10000	0	11	int64	True	2.0972
Income	10000	0	9993	float64	True	40490.49516
Full_meals_eaten	10000	0	8	int64	True	1.0014
Additional_charges	10000	0	9418	float64	True	12934.528587
Initial_days	10000	0	9997	float64	True	34.455299
TotalCharge	10000	0	9997	float64	True	5312.172769
VitD_levels	10000	0	9976	float64	True	17.964262
Zip	10000	0	8612	int64	True	50159.3239
Age	10000	0	72	int64	True	53.5117
Doc_visits	10000	0	9	int64	True	5.0122
City	10000	0	6072	object	False	-
County	10000	0	1607	object	False	-
Job	10000	0	639	object	False	-
State	10000	0	52	object	False	-
TimeZone	10000	0	26	object	False	-
Marital	10000	0	5	object	False	-
Services	10000	0	4	object	False	-
Area	10000	0	3	object	False	-
Gender	10000	0	3	object	False	-
Initial_admin	10000	0	3	object	False	-
Complication_risk	10000	0	3	object	False	-
ReAdmis	10000	0	2	object	False	-
Soft_drink	10000	0	2	object	False	-
HighBlood	10000	0	2	object	False	-
Stroke	10000	0	2	object	False	-
Overweight	10000	0	2	object	False	-
Arthritis	10000	0	2	object	False	-
Diabetes	10000	0	2	object	False	-
Hyperlipidemia	10000	0	2	object	False	-
BackPain	10000	0	2	object	False	-
Anxiety	10000	0	2	object	False	-
Allergic_rhinitis	10000	0	2	object	False	-
Reflux_esophagitis	10000	0	2	object	False	-
Asthma	10000	0	2	object	False	-

	Mode	Min	Median	Max	\
Population	0	0	2769.0	122814	
vitD_supp	0	0	0.0	5	
Children	0	0	1.0	10	
Income	14572.4	154.08	33768.42	207249.1	
Full_meals_eaten	0	0	1.0	7	
Additional_charges	3883.66416	3125.703	11573.977735	30566.07	
Initial_days	63.54432	1.001981	35.836244	71.98149	
TotalCharge	7555.452	1938.312067	5213.952	9180.728	
VitD_levels	15.26009	9.806483	17.951122	26.394449	
Zip	24136	610	50207.0	99929	
Age	47	18	53.0	89	
Doc_visits	5	1	5.0	9	
City	-	-	-	-	
County	-	-	-	-	
Job	-	-	-	-	
State	-	-	-	-	
TimeZone	-	-	-	-	
Marital	-	-	-	-	
Services	-	-	-	-	
Area	-	-	-	-	
Gender	-	-	-	-	
Initial_admin	-	-	-	-	
Complication_risk	-	-	-	-	
ReAdmis	-	-	-	-	
Soft_drink	-	-	-	-	
HighBlood	-	-	-	-	
Stroke	-	-	-	-	
Overweight	-	-	-	-	
Arthritis	-	-	-	-	

	Std	Skew	Kurt
Population	14824.758614	2.229959	5.880913
vitD_supp	0.628505	1.550205	2.330763
Children	2.163659	1.448013	2.076321
Income	28521.153293	1.405899	2.74569
Full_meals_eaten	1.008117	1.009461	1.042727
Additional_charges	6542.601544	0.831842	-0.142684
Initial_days	26.309341	0.070286	-1.754525
TotalCharge	2180.393838	0.069661	-1.668267
VitD_levels	2.017231	0.032435	-0.022112
Zip	27469.588208	0.022813	-1.065366
Age	20.638538	0.005117	-1.189527
Doc_visits	1.045734	-0.018563	0.025999
City	-	-	-
Marital	-	-	-
Services	-	-	-
Area	-	-	-
Gender	-	-	-
Initial_admin	-	-	-
Complication_risk	-	-	-
ReAdmis	-	-	-
Soft_drink	-	-	-
HighBlood	-	-	-
Stroke	-	-	-
Overweight	-	-	-
Arthritis	-	-	-
Diabetes	-	-	-
Hyperlipidemia	-	-	-
BackPain	-	-	-
Anxiety	-	-	-
Allergic_rhinitis	-	-	-
Reflux_esophagitis	-	-	-
Asthma	-	-	-

The summary stats show us that the dataset has many continuous variables, due to the mean/max/mode being outside of 0 and 1. This also gives us a plan for starting to re-express our categorical variables (variables that are simply yes/no, or have a defined group) in the next step so begin building models to test our dependent variable.

The summary statistics overall show us that our average patient lives in an area with a population of 9,965 people (with a standard deviation of 14,824) within a mile radius, has 2 children (with a standard deviation of 2.16), has an income of $40k/yr (with a standard deviation of $28,521), eats 1 full meal a day while in the hospital (with a standard deviation of 1), receives $12,934 in additional charges (with a standard deviation of $6,542), spends 34 days on their initial stay in the hospital (with a standard deviation of 26 days), receives $5,312 in total charges (with a standard deviation of $2180), has Vitamin D levels of 17.96 ng/mL upon admission (with a standard deviation of 2), is 53 years old (with a standard deviation of 21) , and is visited by their doctor 5 times during their stay (with a standard deviation of 1).

C3. To prepare the data for analysis I re-expressed some categorical variables on my own, and used pd.get_dummies to automate one hot encoding of others. Code snippets are below.

Re-expression of categorical variables

```
#Data Wrangling; turn categorical values into quantitative data

df['ReAdmis_numeric'] = df['ReAdmis']

dict_ReAdmis = {"ReAdmis_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_ReAdmis, inplace=True)


df['Soft_drink_numeric'] = df['Soft_drink']

dict_Soft_drink = {"Soft_drink_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_Soft_drink, inplace=True)


df['HighBlood_numeric'] = df['HighBlood']

dict_HighBlood = {"HighBlood_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_HighBlood, inplace=True)


df['Stroke_numeric'] = df['Stroke']

dict_stroke = {"Stroke_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_stroke, inplace=True)


df['Arthritis_numeric'] = df['Arthritis']

dict_arthritis = {"Arthritis_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_arthritis, inplace=True)


df['Diabetes_numeric'] = df['Diabetes']

dict_diabetes = {"Diabetes_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_diabetes, inplace=True)
```

```
df['Hyperlipidemia_numeric'] = df['Hyperlipidemia']

dict_hyperlipidemia = {"Hyperlipidemia_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_hyperlipidemia, inplace=True)



df['BackPain_numeric'] = df['BackPain']

dict_backpain = {"BackPain_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_backpain, inplace=True)



df['Allergic_rhinitis_numeric'] = df['Allergic_rhinitis']

dict_allergies = {"Allergic_rhinitis_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_allergies, inplace=True)



df['Reflux_esophagitis_numeric'] = df['Reflux_esophagitis']

dict_reflux = {"Reflux_esophagitis_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_reflux, inplace=True)



df['Asthma_numeric'] = df['Asthma']

dict_asthma = {"Asthma_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_asthma, inplace=True)
```

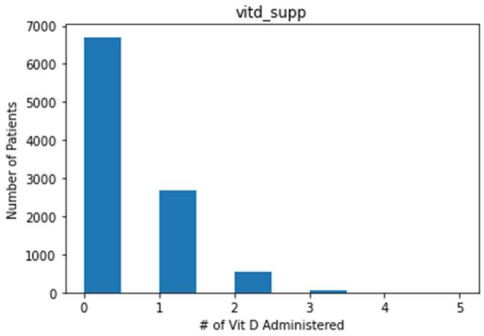
Pd.get_dummies one-hot encoding

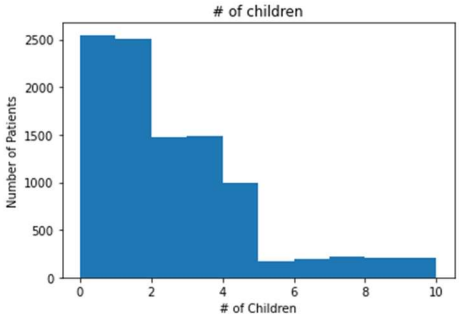
```
df = pd.get_dummies(df, columns=["Marital", "Services", "Gender", "Initial_admin",
"Complication_risk"])
```

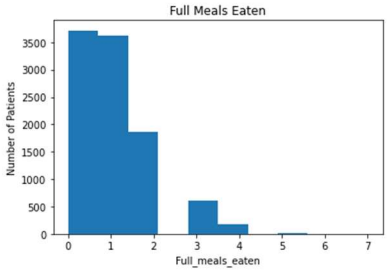
I used both methods for the practice and to get experience with both.

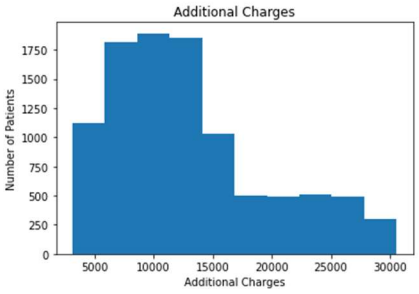
C4. Univariate and Bivariate visualizations are below:

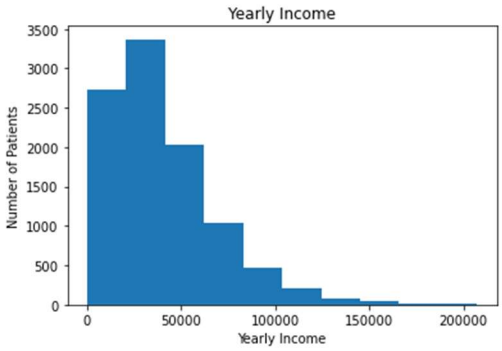
Univariate Visualizations

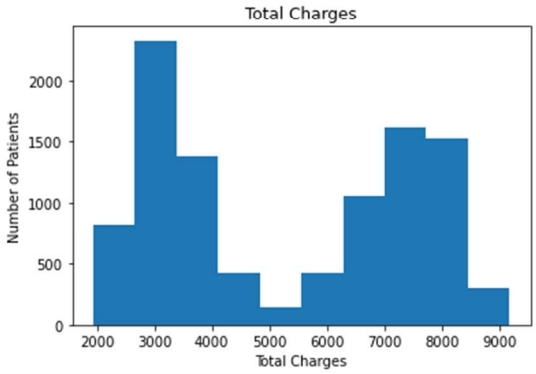

vitd_supp
Number of Patients
# of Vit D Administered

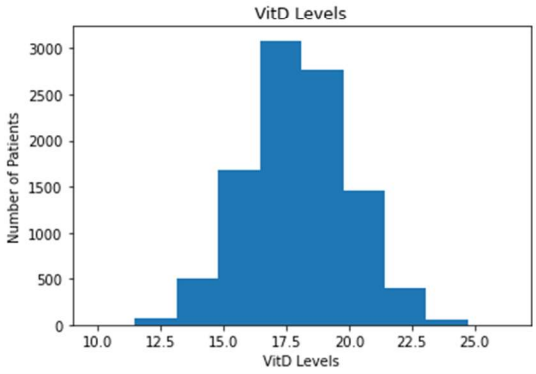

# of children
Number of Patients
# of Children

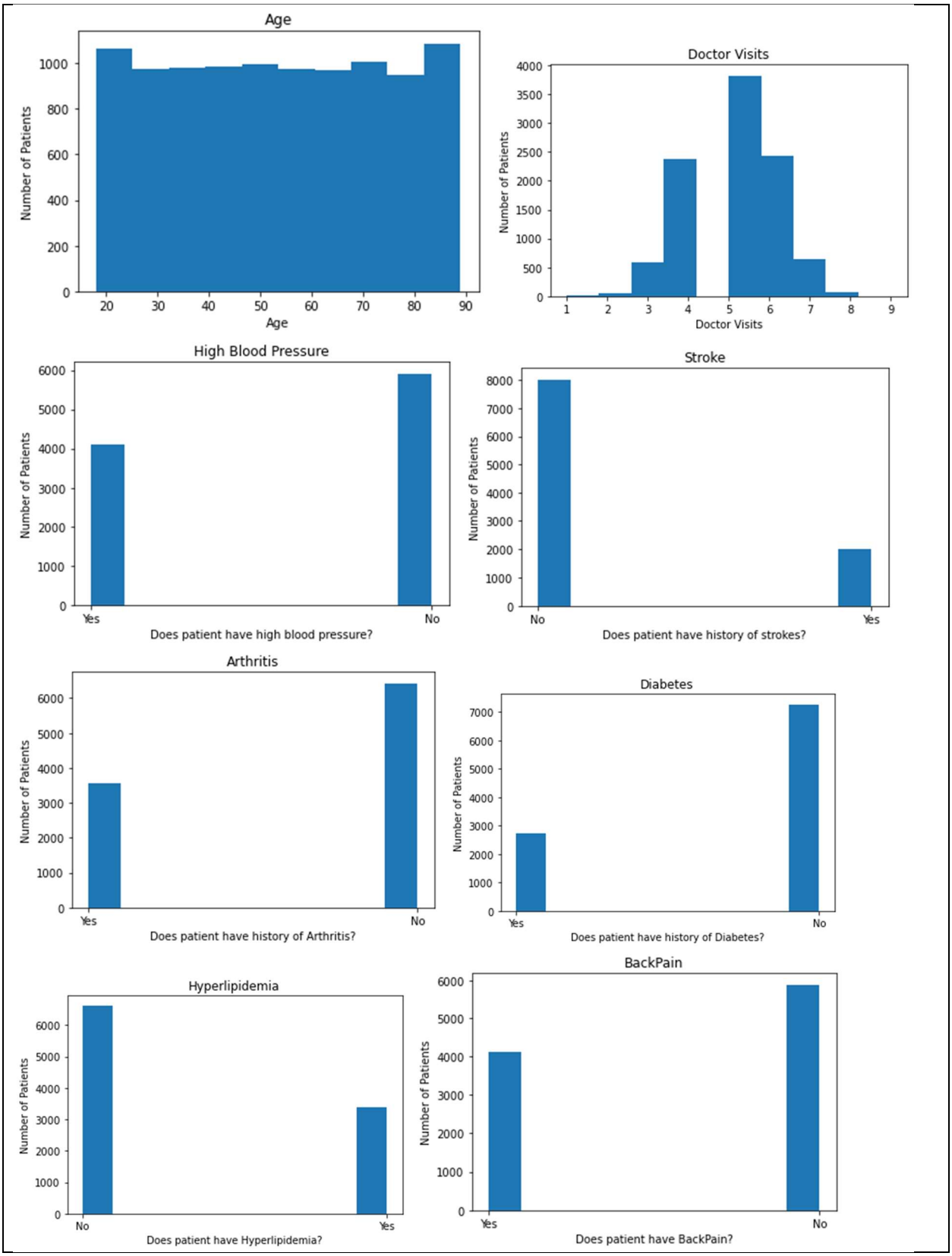

Full Meals Eaten
Number of Patients
Full_meals_eaten

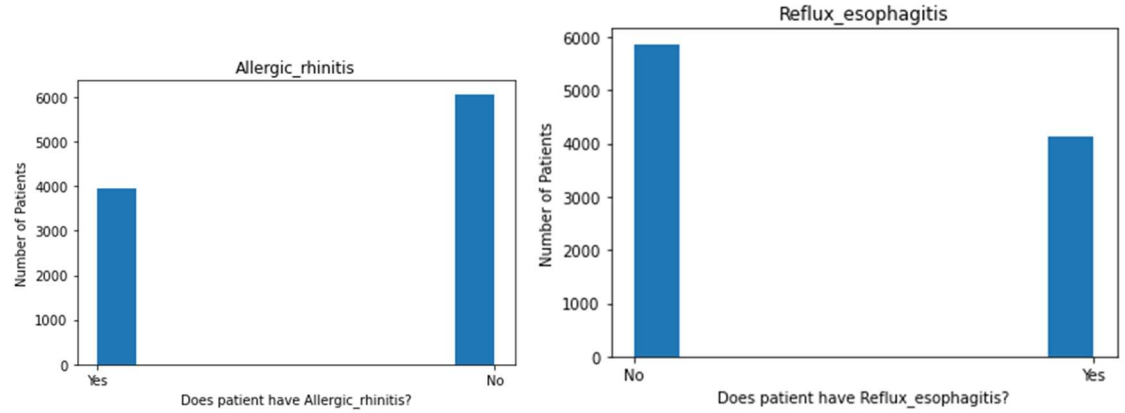

Additional Charges
Number of Patients
Additional Charges


Yearly Income
Number of Patients
Yearly Income

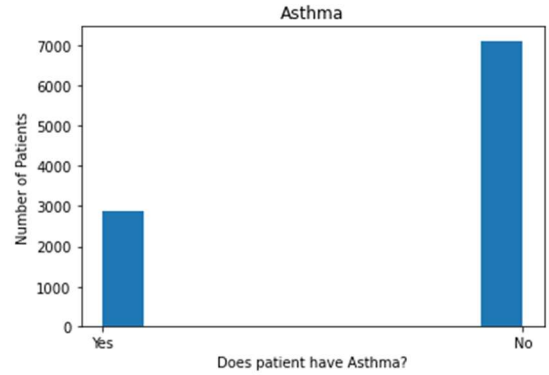

Total Charges
Number of Patients
Total Charges

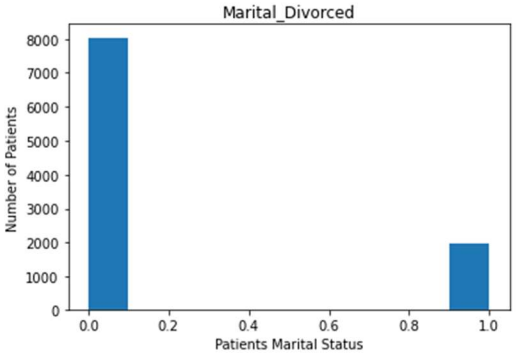

VitD Levels
Number of Patients
VitD Levels

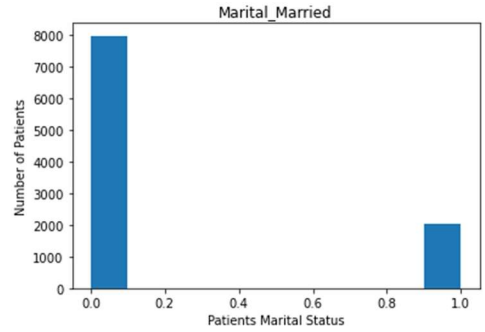

Age
Number of Patients
Age
Doctor Visits
Number of Patients
Doctor Visits
High Blood Pressure
Number of Patients
Yes
No
Does patient have high blood pressure?
Stroke
Number of Patients
No
Yes
Does patient have history of strokes?
Arthritis
Number of Patients
Yes
No
Does patient have history of Arthritis?
Diabetes
Number of Patients
Yes
No
Does patient have history of Diabetes?
Hyperlipidemia
Number of Patients
No
Yes
Does patient have Hyperlipidemia?
BackPain
Number of Patients
Yes
No
Does patient have BackPain?

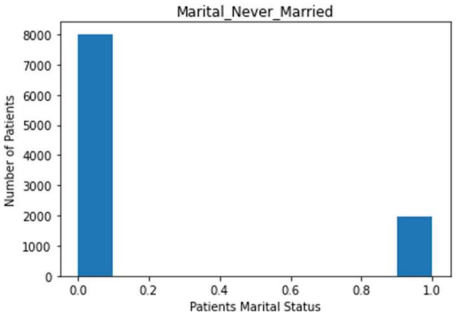

Allergic_rhinitis
Number of Patients
Yes
No
Does patient have Allergic_rhinitis?

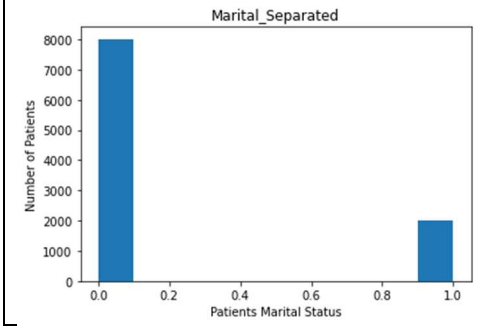
Reflux_esophagitis
Number of Patients
No
Yes
Does patient have Reflux_esophagitis?

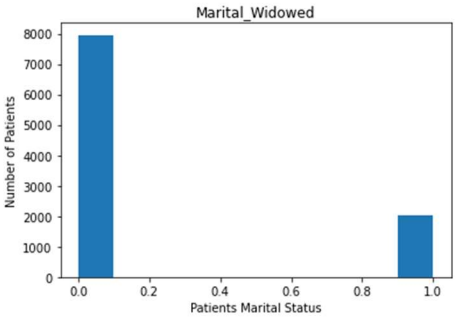

Asthma
Number of Patients
Yes
No
Does patient have Asthma?

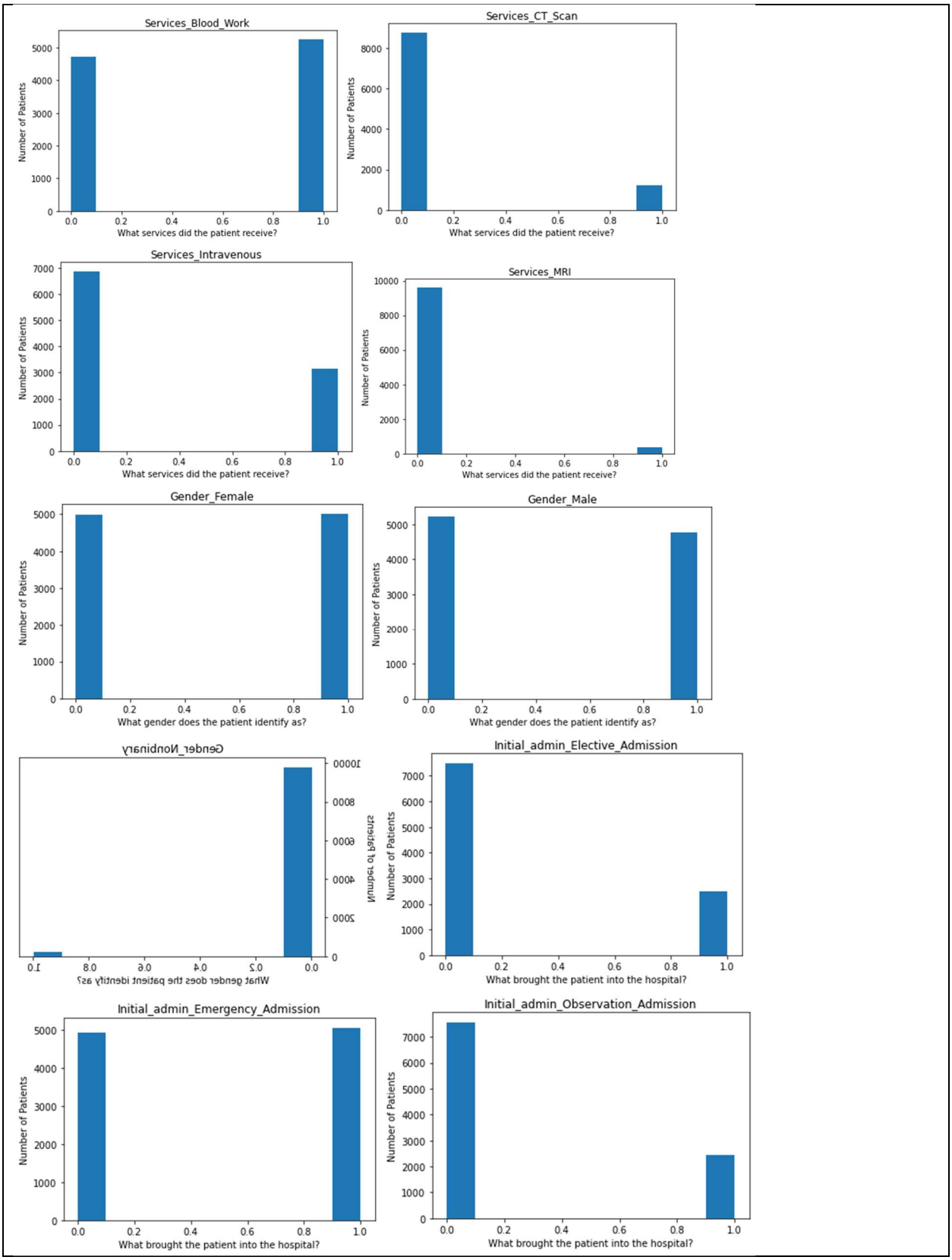

Marital_Divorced
Number of Patients
Patients Marital Status

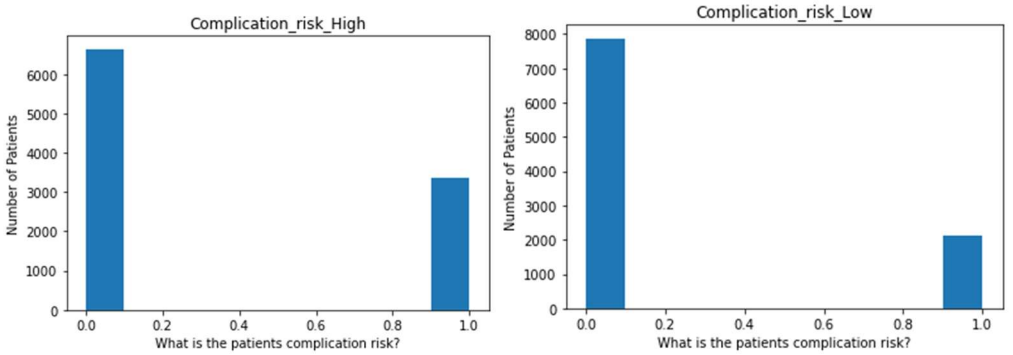

Marital_Married
Number of Patients
Patients Marital Status

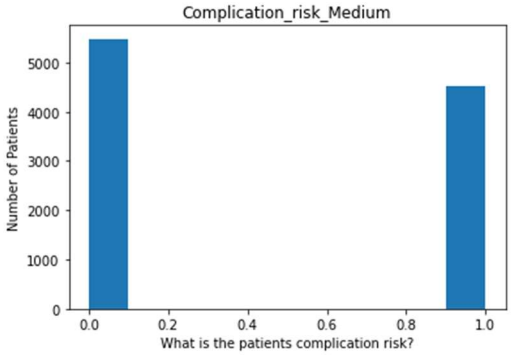

Marital_Never_Married
Number of Patients
Patients Marital Status


Marital_Separated
Number of Patients
Patients Marital Status

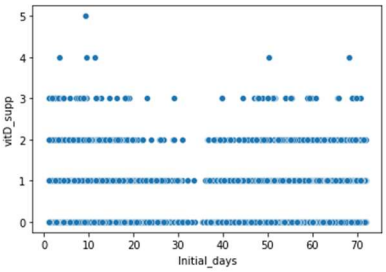

Marital_Widowed
Number of Patients
Patients Marital Status

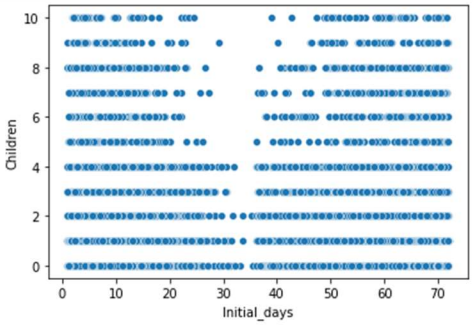

Services_Blood_Work
Number of Patients
What services did the patient receive?
Services_CT_Scan
Number of Patients
What services did the patient receive?
Services_Intravenous
Number of Patients
What services did the patient receive?
Services_MRI
Number of Patients
What services did the patient receive?
Gender_Female
Number of Patients
What gender does the patient identify as?
Gender_Male
Number of Patients
What gender does the patient identify as?
Initial_admin_Elective_Admission
Number of Patients
What brought the patient into the hospital?
Initial_admin_Emergency_Admission
Number of Patients
What brought the patient into the hospital?
Initial_admin_Observation_Admission
Number of Patients
What brought the patient into the hospital?

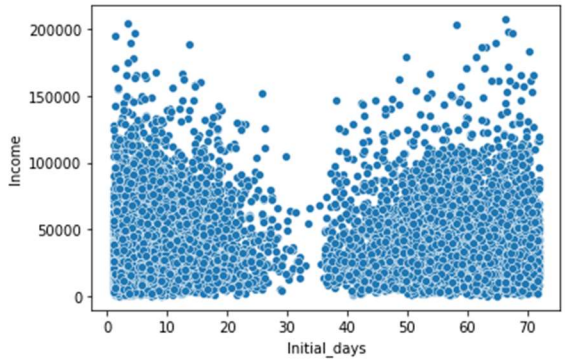


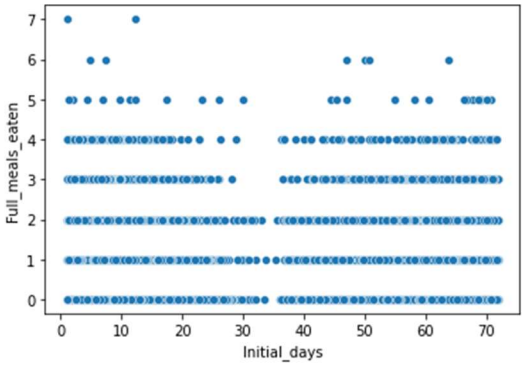


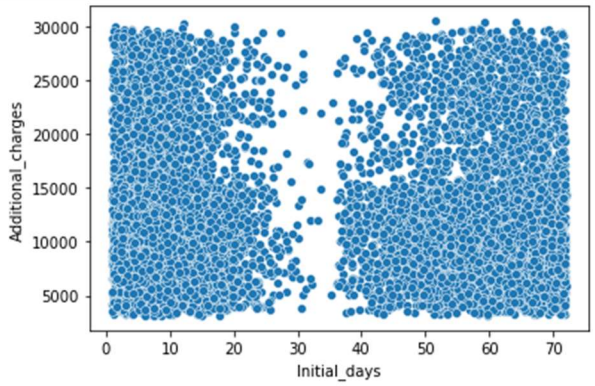
Bivariate Visualizations

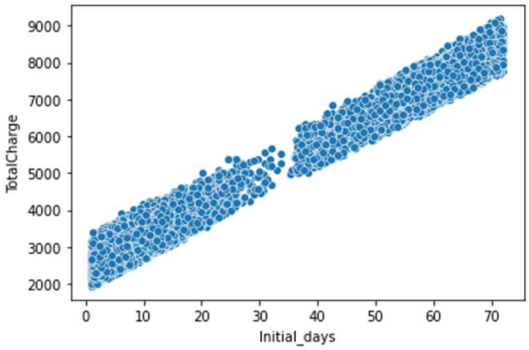


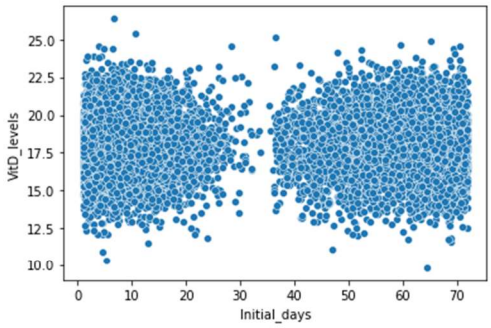


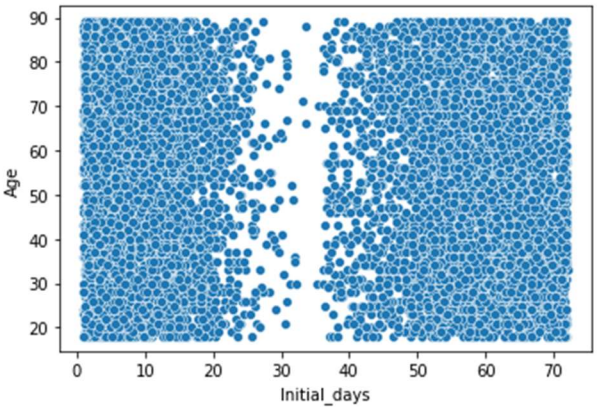


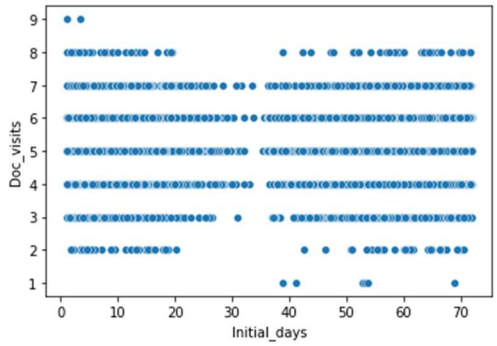


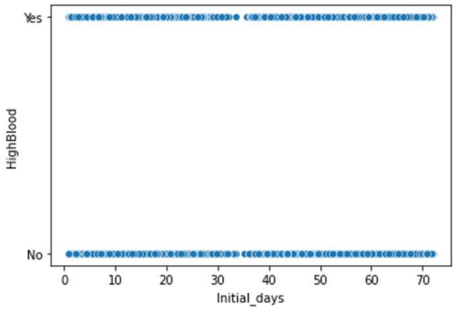

Additional_charges
Initial_days

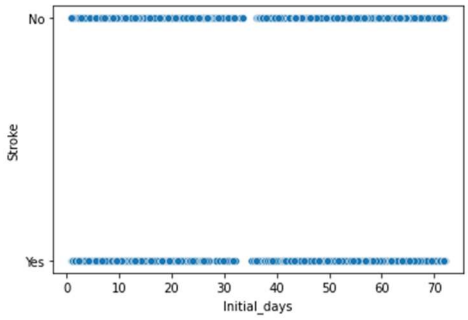

TotalCharge
Initial_days

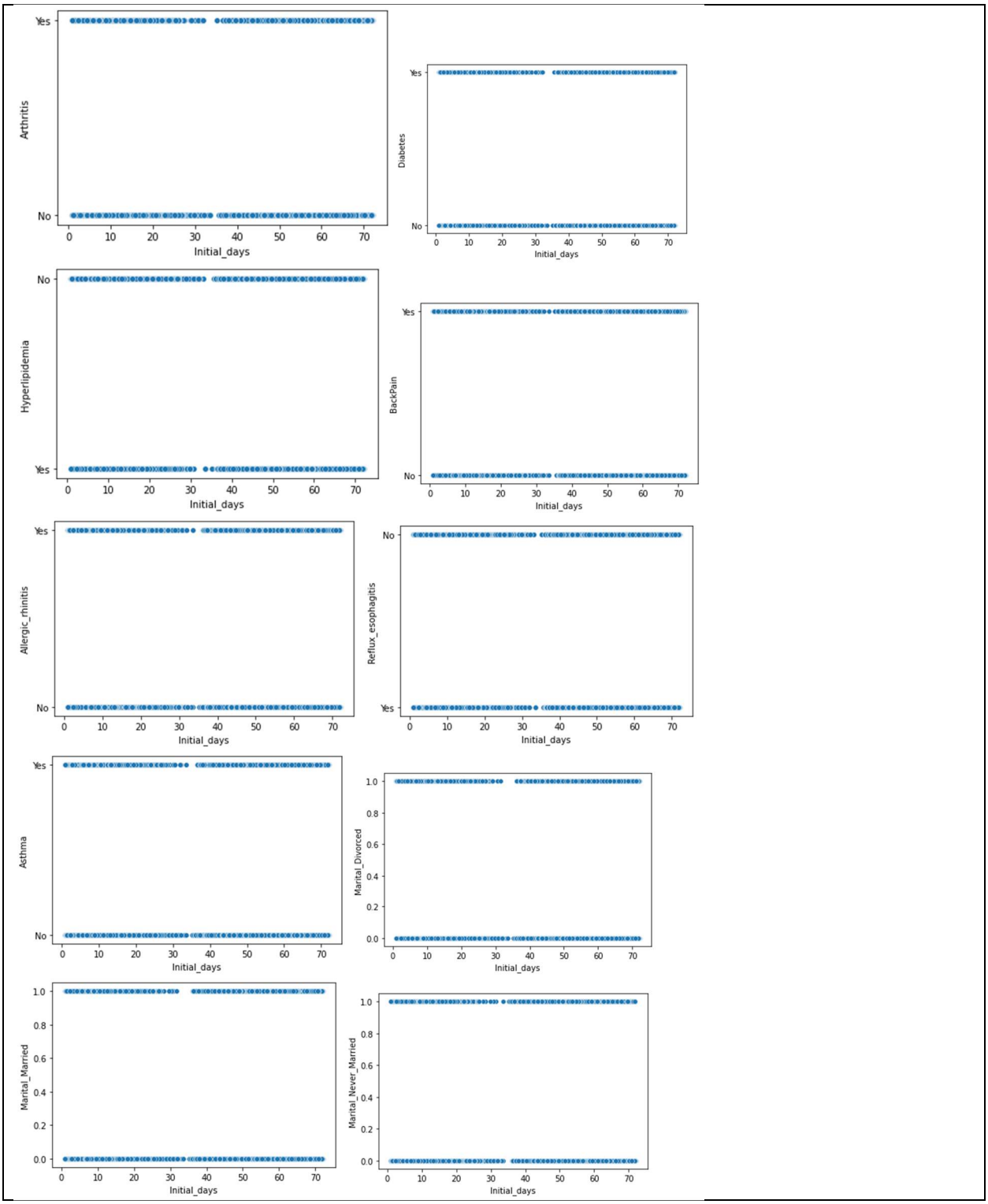

VitD_levels
Initial_days

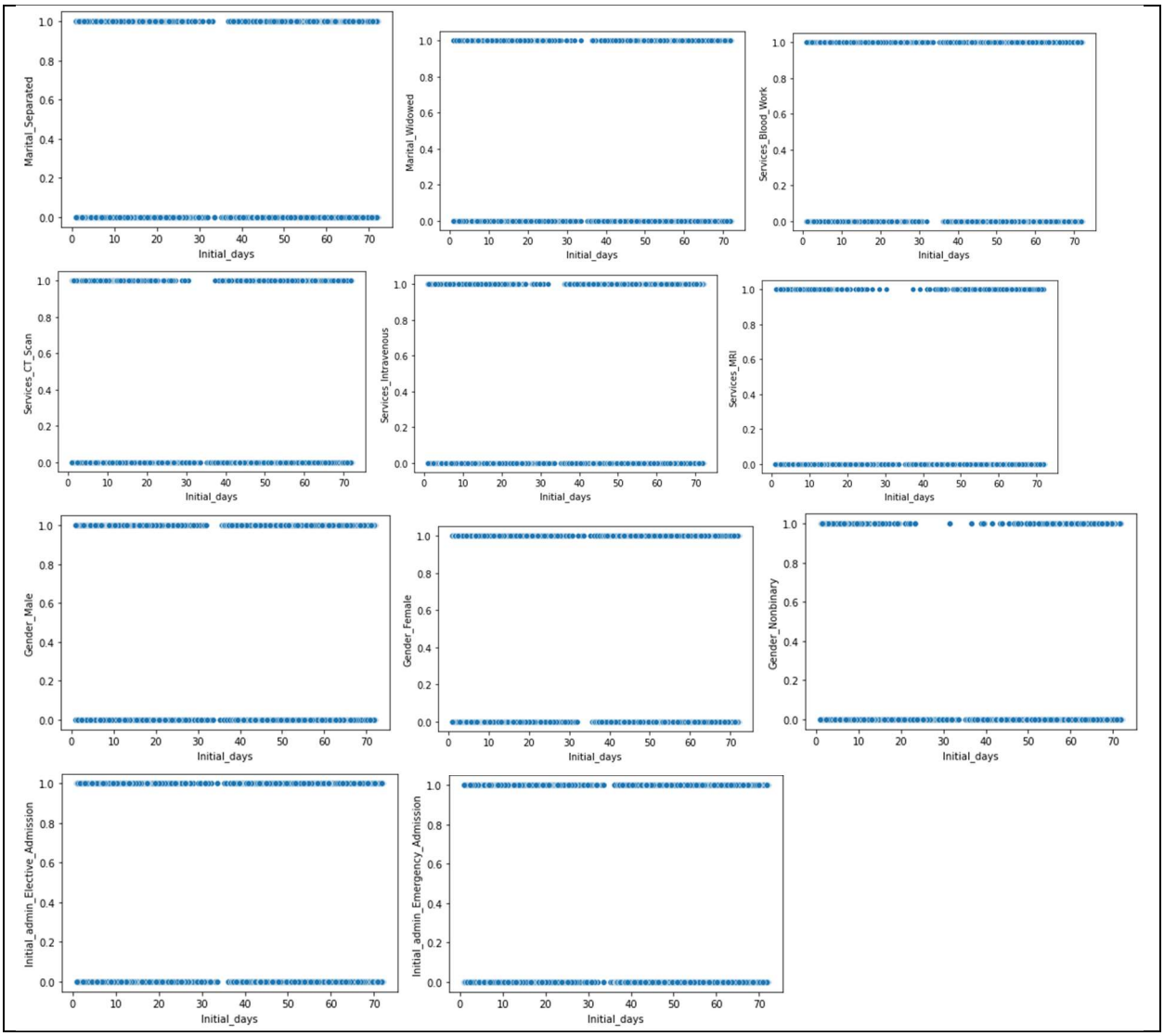

Age
Initial_days

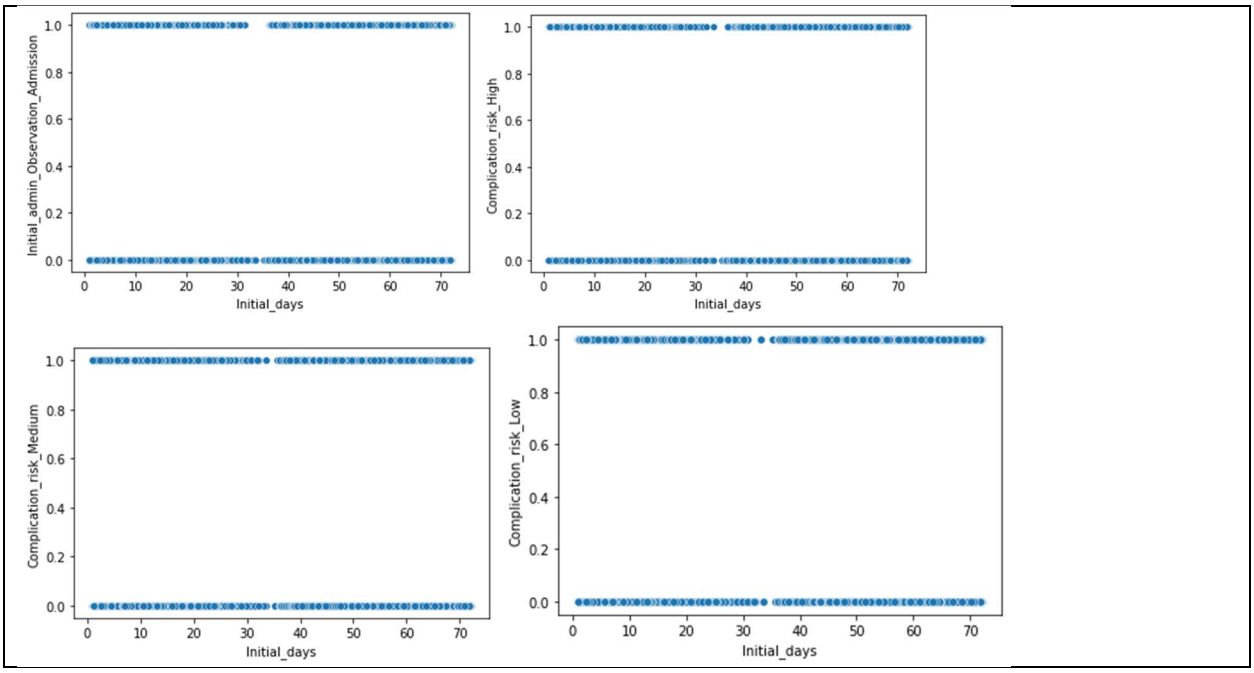

Doc_visits
Initial_days


HighBlood
Yes
No
Initial_days


Stroke
No
Yes
Initial_days


Arthritis
Yes
No
Initial_days
Diabetes
Yes
No
Initial_days
Hyperlipidemia
No
Yes
Initial_days
BackPain
Yes
No
Initial_days
Allergic_rhinitis
Yes
No
Initial_days
Reflux_esophagitis
No
Yes
Initial_days
Asthma
Yes
No
Initial_days
Marital_Divorced
Initial_days
Marital_Married
Initial_days
Marital_Never_Married
Initial_days


Marital_Separated
Marital_Widowed
Services_Blood_Work
Services_CT_Scan
Services_Intravenous
Services_MRI
Gender_Male
Gender_Female
Gender_Nonbinary
Initial_admin_Elective_Admission
Initial_admin_Emergency_Admission
Initial_days



C5. Prepared data set uploaded as past of submission.

Part IV: Model Comparison and Analysis

D1. Initial regression model with variables identified in C2

OLS Regression Results

Dep. Variable:	Initial_days	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	9.258e+05
Date:	Sun, 20 Nov 2022	**Prob (F-statistic):**	0.00
Time:	14:50:51	**Log-Likelihood:**	-7060.1
No. Observations:	10000	**AIC:**	1.418e+04
Df Residuals:	9968	**BIC:**	1.441e+04
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-10.8883	0.023	-471.390	0.000	-10.934	-10.843
vitD_supp	0.0001	0.008	0.019	0.985	-0.015	0.015
Children	-0.0016	0.002	-0.710	0.478	-0.006	0.003
Income	-9.605e-09	1.72e-07	-0.056	0.956	-3.48e-07	3.28e-07
Full_meals_eaten	-0.0042	0.005	-0.861	0.389	-0.014	0.005
Additional_charges	-2.634e-06	3.03e-06	-0.869	0.385	-8.57e-06	3.31e-06
TotalCharge	0.0122	2.28e-06	5345.969	0.000	0.012	0.012
VitD_levels	-0.0017	0.002	-0.702	0.482	-0.006	0.003
Age	0.0005	0.001	0.662	0.508	-0.001	0.002
Doc_visits	0.0009	0.005	0.191	0.848	-0.008	0.010
HighBlood_numeric	-1.3551	0.028	-48.387	0.000	-1.410	-1.300
Stroke_numeric	0.0181	0.012	1.463	0.144	-0.006	0.042
Arthritis_numeric	-0.8892	0.010	-86.563	0.000	-0.909	-0.869
Diabetes_numeric	-0.9146	0.011	-82.879	0.000	-0.936	-0.893
Hyperlipidemia_numeric	-1.1332	0.010	-108.907	0.000	-1.154	-1.113
BackPain_numeric	-1.0474	0.010	-104.629	0.000	-1.067	-1.028
Allergic_rhinitis_numeric	-0.7427	0.010	-73.806	0.000	-0.762	-0.723
Reflux_esophagitis_numeric	-0.7201	0.010	-72.092	0.000	-0.740	-0.701
Asthma_numeric	-0.0128	0.011	-1.179	0.238	-0.034	0.008
Marital_Divorced	-2.1805	0.011	-199.011	0.000	-2.202	-2.159
Marital_Married	-2.1820	0.011	-200.749	0.000	-2.203	-2.161
Marital_Never_Married	-2.1642	0.011	-199.295	0.000	-2.186	-2.143
Marital_Separated	-2.1874	0.011	-201.002	0.000	-2.209	-2.166
Marital_Widowed	-2.1742	0.011	-201.888	0.000	-2.195	-2.153
Services_Blood_Work	-2.7271	0.010	-264.686	0.000	-2.747	-2.707
Services_CT_Scan	-2.7189	0.014	-195.632	0.000	-2.746	-2.692
Services_Intravenous	-2.7327	0.011	-246.551	0.000	-2.754	-2.711
Services_MRI	-2.7095	0.021	-130.861	0.000	-2.750	-2.669
Gender_Female	-3.6287	0.013	-271.876	0.000	-3.655	-3.603
Gender_Male	-3.6210	0.013	-272.503	0.000	-3.647	-3.595
Gender_Nonbinary	-3.6386	0.025	-143.538	0.000	-3.688	-3.589
Initial_admin_Elective_Admission	-1.5435	0.011	-141.986	0.000	-1.565	-1.522
Initial_admin_Emergency_Admission	-7.8003	0.010	-769.676	0.000	-7.820	-7.780
Initial_admin_Observation_Admission	-1.5445	0.011	-139.863	0.000	-1.566	-1.523
Complication_risk_High	-6.9899	0.011	-663.987	0.000	-7.011	-6.969
Complication_risk_Low	-1.9465	0.011	-172.527	0.000	-1.969	-1.924
Complication_risk_Medium	-1.9519	0.010	-194.182	0.000	-1.972	-1.932

Omnibus:	90391.781	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1792.347
Skew:	-0.761	**Prob(JB):**	0.00
Kurtosis:	1.592	**Cond. No.**	2.91e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.06e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

D2. I get an r-squared value of 1. I also have a multicollinearity issue. This means my model does explain the dependent variable, however, I should look at reducing the number of variables to make the model less complex and reduce multicollinearities. I can use VIF to look at variables that are producing multicollinearity.

Results of VIF below:

	feature	VIF
0	Initial_days	2880.163153
1	vitD_supp	1.003676
2	Children	1.003506
3	Income	1.002683
4	Full_meals_eaten	1.004107
5	Additional_charges	16.303881
6	TotalCharge	2944.078834
7	VitD_levels	1.003914
8	Age	9.273563
9	Doc_visits	1.003377
10	HighBlood_numeric	9.711378
11	Stroke_numeric	1.010014
12	Arthritis_numeric	1.760819
13	Diabetes_numeric	1.696512
14	Hyperlipidemia_numeric	2.198017
15	BackPain_numeric	2.112077
16	Allergic_rhinitis_numeric	1.551215
17	Reflux_esophagitis_numeric	1.527188
18	Asthma_numeric	1.003104
19	Marital_Divorced	inf
20	Marital_Married	inf
21	Marital_Never_Married	inf
22	Marital_Separated	inf
23	Marital_Widowed	inf
24	Services_Blood_Work	inf
25	Services_CT_Scan	inf
26	Services_Intravenous	inf
27	Services_MRI	inf
28	Gender_Female	inf
29	Gender_Male	inf
30	Gender_Nonbinary	inf
31	Initial_admin_Elective_Admission	inf
32	Initial_admin_Emergency_Admission	inf
33	Initial_admin_Observation_Admission	inf
34	Complication_risk_High	inf
35	Complication_risk_Low	inf
36	Complication_risk_Medium	inf

Results of new model with variables of a VIF = infinity removed:

D3. Reducing the number of variables using VIF to produce an even smaller model with continuous and categorical variables. VIF results are below:

OLS Regression Results

Dep. Variable:	Initial_days	**R-squared:**	0.979
Model:	OLS	**Adj. R-squared:**	0.979
Method:	Least Squares	**F-statistic:**	2.531e+04
Date:	Sun, 20 Nov 2022	**Prob (F-statistic):**	0.00
Time:	17:14:46	**Log-Likelihood:**	-27675.
No. Observations:	10000	**AIC:**	5.539e+04
Df Residuals:	9981	**BIC:**	5.552e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-26.8453	0.433	-62.064	0.000	-27.693	-25.997
vitD_supp	-0.0345	0.061	-0.562	0.574	-0.155	0.086
Children	-0.0093	0.018	-0.523	0.601	-0.044	0.026
Income	1.843e-06	1.35e-06	1.362	0.173	-8.09e-07	4.49e-06
Full_meals_eaten	-0.0540	0.038	-1.411	0.158	-0.129	0.021
Additional_charges	-0.0005	2.33e-05	-19.913	0.000	-0.001	-0.000
TotalCharge	0.0120	1.77e-05	674.076	0.000	0.012	0.012
VitD_levels	-0.0443	0.019	-2.317	0.021	-0.082	-0.007
Age	0.1051	0.006	18.841	0.000	0.094	0.116
Doc_visits	-0.0523	0.037	-1.416	0.157	-0.125	0.020
HighBlood_numeric	2.6166	0.216	12.121	0.000	2.193	3.040
Stroke_numeric	0.2519	0.097	2.599	0.009	0.062	0.442
Arthritis_numeric	-0.7721	0.081	-9.583	0.000	-0.930	-0.614
Diabetes_numeric	-0.7978	0.087	-9.214	0.000	-0.968	-0.628
Hyperlipidemia_numeric	-1.2028	0.082	-14.733	0.000	-1.363	-1.043
BackPain_numeric	-1.0001	0.078	-12.742	0.000	-1.154	-0.846
Allergic_rhinitis_numeric	-0.7978	0.079	-10.104	0.000	-0.953	-0.643
Reflux_esophagitis_numeric	-0.7123	0.078	-9.088	0.000	-0.866	-0.559
Asthma_numeric	0.0840	0.085	0.987	0.324	-0.083	0.251

Omnibus:	1193.071	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	341.251
Skew:	-0.130	**Prob(JB):**	7.91e-75
Kurtosis:	2.133	**Cond. No.**	5.74e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.74e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Continuous variables in this model include; Children, Income, vitD_supp, Full_meals_eaten, Additional_charges, TotalCharge, VitD_levels, Age and Doc_visits.
Categorical variables include all of the numeric columns.

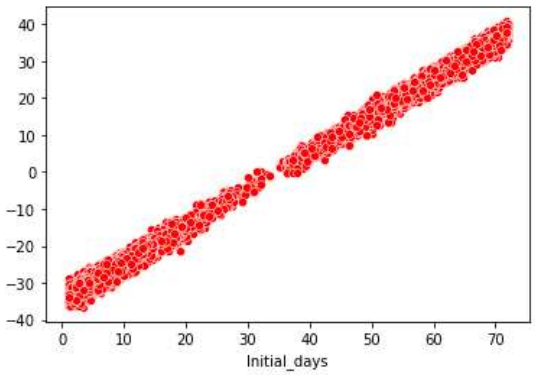
E. Analyze the data set.

E1. When building the model I chose to select patient demographics and patient health conditions to try and find a specific cohort of patients that were at risk of spending long initial days in the hospital. I chose Initial_days as the dependent variable because we are supposed to choose a continuous variable for this task and I saw in previous assignments that Initial_days has a very strong correlation for patient readmissions.

Both models I built had high R-squared and Adj. R-Squared scores. However, the F-Statistic is very high and P-value of the individual variables vary drastically. With that said, the models I built are very complex and need to be cleaned up more. More variables could also be chosen, though when I played with this data set I couldn't find anything that had a large correlation to Initial_days. I think this dataset just isn't a great fit for answering this question.

Below is a screenshot of my Residuals Plot:

```
In [79]: #Residual plot
         df['intercept'] = 1
         residuals = df['Initial_days'] - mdl_initial_vs_variables.predi
         sns.scatterplot(x=df['Initial_days'], y=residuals, color='red')
Out[79]: <AxesSubplot:xlabel='Initial_days'>
```



E2. Residual Standard Error:

```
In [78]: #Residual Standard Error
         np.sqrt(mdl_initial_vs_variables.scale)
Out[78]: 26.295731213305718
```

(Techhelpnotes, 2022)

E3.

Initial Model
mdl_initial_vs_variables = ols("Initial_days ~ vitD_supp + Children + Income + Full_meals_eaten + Additional_charges + TotalCharge + VitD_levels + Age + Doc_visits + HighBlood_numeric + Stroke_numeric + Arthritis_numeric + Diabetes_numeric + Hyperlipidemia_numeric + BackPain_numeric + Allergic_rhinitis_numeric + Reflux_esophagitis_numeric + Asthma_numeric + Marital_Divorced + Marital_Married + Marital_Never_Married + Marital_Separated +

```
Marital_Widowed + Services_Blood_Work + Services_CT_Scan + Services_Intravenous +
Services_MRI + Gender_Female + Gender_Male + Gender_Nonbinary +
Initial_admin_Elective_Admission + Initial_admin_Emergency_Admission +
Initial_admin_Observation_Admission + Complication_risk_High + Complication_risk_Low +
Complication_risk_Medium", data=df).fit()



mdl_initial_vs_variables.summary()
```

VIF to reduce model

```
# Checking for the VIF values of the variables.

from statsmodels.stats.outliers_influence import variance_inflation_factor



X = df[['Initial_days', 'vitD_supp', 'Children', 'Income', 'Full_meals_eaten', 'Additional_charges',
'TotalCharge', 'VitD_levels', 'Age', 'Doc_visits', 'HighBlood_numeric', 'Stroke_numeric',
'Arthritis_numeric', 'Diabetes_numeric', 'Hyperlipidemia_numeric', 'BackPain_numeric',
'Allergic_rhinitis_numeric', 'Reflux_esophagitis_numeric', 'Asthma_numeric', 'Marital_Divorced',
'Marital_Married', 'Marital_Never_Married', 'Marital_Separated', 'Marital_Widowed',
'Services_Blood_Work', 'Services_CT_Scan', 'Services_Intravenous', 'Services_MRI',
'Gender_Female', 'Gender_Male', 'Gender_Nonbinary', 'Initial_admin_Elective_Admission',
'Initial_admin_Emergency_Admission', 'Initial_admin_Observation_Admission',
'Complication_risk_High', 'Complication_risk_Low', 'Complication_risk_Medium']]





# VIF dataframe

vif_data = pd.DataFrame()

vif_data["feature"] = X.columns



# calculating VIF for each feature

vif_data["VIF"] = [variance_inflation_factor(X.values, i)

                   for i in range(len(X.columns))]



print(vif_data)
```

(GeeksforGeeks, 2019)

Reduced Model
#Running new model based on VIF results, removing variables with VIF = infinity mdl_initial_vs_variables = ols("Initial_days ~ vitD_supp + Children + Income + Full_meals_eaten + Additional_charges + TotalCharge + VitD_levels + Age + Doc_visits + HighBlood_numeric + Stroke_numeric + Arthritis_numeric + Diabetes_numeric + Hyperlipidemia_numeric + BackPain_numeric + Allergic_rhinitis_numeric + Reflux_esophagitis_numeric + Asthma_numeric", data=df).fit() mdl_initial_vs_variables.summary()

Part V: Data Summary and Implications
F1. My linear regression equation:

Y = -26.85 + -0.03 (vitD_supp) + -0.009 (Children) + 1.843x10^6 (Income) + -0.05 (Full_meals_eaten) + -0.0005 (Additional_charges) + 0.012 (TotalCharge) + -0.0443 (VitD_levels) + 0.1051 (Age) + -0.05 (Doc_visits) + 2.6166 (HighBlood) + 0.252 (Stroke) + -0.7721 (Arthritis) + -0.7978 (Diabetes) + -1.2028 (Hyperlipidemia) + -1.0001 (BackPain) + -0.7978 (Allergic_rhinitis) + -0.7123 (Reflux_esophagitis) + 0.0840 (Asthma)
This line means for every 1 unit of:

vitD_supp, Initial_days will decrease 0.03 units
Children, Initial_days will decrease 0.009 units
Income, Initial_days will increase 1.843x10^6 units
Full_meals_eaten, Initial_days will decrease 0.05 units
Additional_charges, Initial_days will decrease 0.0005 units
TotalCharge, Initial_days will increase 0.012 units
VitD_levels, Initial_days will decrease -0.0443 units
Age, Initial_days will increase 0.1051 units
Doc_visits, Initial_days will decrease 0.0523 units
HighBlood_numeric, Initial_days will increase 2.6166 units
Stroke_numeric, Initial_days will increase 0.2519 units
Arthritis_numeric, Initial_days will decrease 0.7721 units
Diabetes_numeric, Initial_days will decrease 0.7978 units
Hyperlipidemia_numeric, Initial_days will decrease 1.2028 units
BackPain_numeric, Initial_days will decrease 1.0001 units
Allergic_rhinitis_numeric, Initial_days will decrease .7978 units
Reflux_esophagitis_numeric, Initial_days will decrease 0.7123 units
Asthma_numeric, Initial_days will increase 0.0840 units

My model is statistically significant based on the R-squared value and p-values of the many different variables. However, this result is likely due to chance due to the complexity of the model and number of variables. Meaning that overall, this model carries no practical significance as well. This model will not produce repeatable results for there to be any real world impact that our hospital can act on.

This model is very limited. The task asked for us to choose a continuous variable, I chose Initial_days, though the main problem we are trying to solve for is ReAdmis, which is a categorical variable. Because I was limited with choosing my dependent variable, the analysis was limited by the scope of the dataset itself. I am not sure that this dataset is robust enough to answer the main question at this point with my current level of knowledge. I have to imagine that I will have to apply a machine learning algorithm to this dataset to figure out what the best model is in the future.

F2. Based on my results, there really isn't a course of action to be recommended. The model is too robust and complex and no predictions can really be made. The course of action would be to start back at square 1 and pick a different dependent variable, reducing our model with a different method, and maybe evaluating what data we capture moving forward.

Part VI: Demonstration

G. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=a459c80e-2e2e-4a35-bbf4-af5401196114

H.

Techhelpnotes, 2022 https://techhelpnotes.com/residual-standard-error-of-a-regression-in-python/

GeeksforGeeks, 2019 https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

I. No sources used.