Western Governors University

D208 – Predictive Modeling – Task 1

By: Krista Moik

Table of Contents

Part I: Research Question

A1: Research Question

Using the provided medical_clean CSV, my research question is: **Which variables affect the length of a patient's initial hospital admission?**

A2: Goals

Knowing which variables affect the length of an initial hospital admission is important for stakeholders. This knowledge could allow hospitals to predict hospital bed availability, staffing needs, and could also allow for techniques that reduce the length of stay thus also reducing costs. My goal is to develop a multiple linear regression model that will allow me to find which variables in the dataset have the greatest impact on the length of admission for patients. In this analysis, the Initial_admin column is my dependent variable and the other variables in the dataset are my independent variables that I will be testing.

Part II: Method Justification

B1: Summary of Assumptions

A multiple linear regression model is based on several assumptions. First, multiple linear regression assumes there is a linear relationship between the dependent and independent variables. This model also assumes that the independent variables have little to no multicollinearity. Additionally, it assumes that observations are selected at random and independently from the population. Finally, this model assumes that residuals have a normal distribution with a mean of zero (Western Governors University, n.d.).

B2: Tool Benefits

Python and R are both capable of supporting the various phases of this analysis. While R is a programming language that has many of the statistical abilities needed for this analysis already built in, I will be using Python as it was recommended to not change programming languages for this project. That being said, Python has many libraries that can be imported that are just as statistically capable as R. Additionally, Python is also beneficial to use in terms of its quickness, readability, mathematical abilities, and visualizations (Western Governors University Information Technology, 2024). As always, I will import pandas and numpy for their database and basic mathematical abilities. Additionally, I will also be importing libraries from scipy.stats, statsmodels, and sklearn to complete the statistical portions of this analysis. I will also be importing matplotlib and seaborn to complete the visualization portions of this analysis. While I am familiar with most of these packages, the statistical models were suggested in WGU Course Materials (Western Governors University, n.d.).

B3: Appropriate Technique

Multiple Linear Regression is an appropriate technique to develop a model to determine the relationship between a continuous dependent variable with multiple independent explanatory variables, which include those that are qualitative or quantitative in nature (Statology, 2020). My dependent variable is Initial_days, which is a numeric and continuous variable. My independent variables include those that are both quantitative and qualitative in nature. Thus, my research question and chosen variables align with the capabilities and requirements of multiple linear regression. Using multiple linear regression will allow me to discover whether any of the independent variables have a relationship with the dependent variable that could be used to predict the number of days a patient will be admitted for.

Part III: Data Preparation

C1: Data Cleaning

First, even though the data states it is cleaned, I will check for duplicate and null values using the print(df.duplicated().value_counts()) and df.isnull().sum() functions. I then plan to check for outliers in my chosen variables using the describe() function, histograms (using matplotlib plt.hist() function), and boxplots (using seaborn- sns.boxplot()), and will perform imputation as needed using fillna(). Once the data is cleaned, I will check the dataset again using boxplots to see if there are any changes. Finally, I will review my variables for categorical values and use ordinal encoding or one hot encoding as needed using functions like creating a dictionary, replace(), drop(), and onehot_encoder() as appropriate. To confirm the data has been cleaned in alignment with my research question, I will visualize my data again to confirm the effects. See the attached document in pdf and ipynb format titled KMoikD208Code1 to see the actual code used per column. By verifying the data is cleaned and ready for further analysis, I am preventing errors and unclean data from affecting my analysis.

C2: Summary Statistics

The dependent variable I am using is Initial_days. This is a continuous variable. By using the describe() function, I obtained the below descriptive statistical information regarding this variable. The minimum days spent in the hospital is slightly over 1 day and the maximum is almost 72 days. The one day minimum is affected by the dataset only including stays that were at least 1 day. The average number of days a patient was initially admitted was almost 34.5 days.

```
#statistically describe dependent variable - Initial_days
df.Initial_days.describe()
```

```
count    10000.000000
mean        34.455299
std         26.309341
min          1.001981
25%          7.896215
50%         35.836244
75%         61.161020
max         71.981490
Name: Initial_days, dtype: float64
```

For the independent variables, I will be using the columns Children, Age, Income, Doc_visits, Overweight, Diabetes, Anxiety, HighBlood, Stroke, Gender, Marital, Complication_risk, and Initial_admin. I chose these independent variables as I felt that they would be more likely to have an impact on how many days a patient was admitted for rather than some of the other variables present in the dataset such as those that

counted how many meals a patient ate while in the hospital or whether or not the patient received a Vitamin D supplement.

The statistics for my chosen variables are below:

1. Children, which is a quantitative and discrete variable. The variable counts the number of children each patient has. As you can see from the below statistics, the minimum number is 0 children and the most any patient has is 10. The average number of children patients have is approximately 2.

```
#Statistically describe independent variable - Children
df.Children.describe()
```

```
count    10000.000000
mean         2.097200
std          2.163659
min          0.000000
25%          0.000000
50%          1.000000
75%          3.000000
max         10.000000
Name: Children, dtype: float64
```

2. Age, which is a quantitative and continuous variable. Age is the age of the patient in years. As you can see by the minimum age of 18, data is not available for patients under the age of 18. The highest age is 89.0 years old. The average age of patients are about 53.5 years old.

```
#Statistically describe independent variable - Age
df.Age.describe()
```

```
count    10000.000000
mean        53.511700
std         20.638538
min         18.000000
25%         36.000000
50%         53.000000
75%         71.000000
max         89.000000
Name: Age, dtype: float64
```

3. Income, which is a quantitative and continuous variable. Income is the reported income of each patient, or the primary insurance holder. The minimum is $154.08, which seems low, but would not be unreasonable for an 18-year-old patient who just got their first job. The highest is over $200,000 a year, which is also not unreasonable for an older professional who has progressed their career over the years. The average salary amongst patients is approximately $40,490.

```
#Statistically describe independent variable - Income
df.Income.describe()
```

```
count     10000.000000
mean      40490.495160
std       28521.153293
min         154.080000
25%       19598.775000
50%       33768.420000
75%       54296.402500
max      207249.100000
Name: Income, dtype: float64
```

4. Marital, which is a qualitative and nominal variable. Marital provides the marital status of the patient, or primary insurance holder. Somewhat surprisingly, widowed is the most common response out of 5 response options of never married, separated, married, divorced, and widowed.

```
#Statistically describe independent variable - Marital
df.Marital.describe()
```

```
count       10000
unique          5
top       Widowed
freq         2045
Name: Marital, dtype: object
```

Looking deeper into this data using value_counts() we can see that both widowed and married are close in number to each other. The other three options are also relatively similar to each other.

```
df.Marital.value_counts()
```

```
Marital
Widowed          2045
Married          2023
Separated        1987
Never Married    1984
Divorced         1961
Name: count, dtype: int64
```

5. Gender, which is a qualitative and nominal variable. This variable is the self-identification of the patient as male, female, or nonbinary. From the statistics below, we can see that there were more female patients than male patients or nonbinary patients.

```
#Statistically describe independent variable - Gender
df.Gender.describe()
```

```
count      10000
unique         3
top       Female
freq        5018
Name: Gender, dtype: object
```

Looking closer at the data using value_counts(), we can see that the number of male patients is only slightly smaller than female, but that the number of patients reporting as nonbinary is much smaller than those who reported as male or female.

```
df.Gender.value_counts()
```

```
Gender
Female       5018
Male         4768
Nonbinary     214
Name: count, dtype: int64
```

6. Doc_visits, which is a quantitative and discrete variable. This is a count of the number of times the primary physician visited the patient during the initial hospitalization. The average number of doctor visits during initial admission was a little over 5 days. No physician visited more than 9 times, and every patient had at least 1 visit by the doctor.

```
#Statistically describe independent variable - Doc_visits
df.Doc_visits.describe()
```

```
count    10000.000000
mean         5.012200
std          1.045734
min          1.000000
25%          4.000000
50%          5.000000
75%          6.000000
max          9.000000
Name: Doc_visits, dtype: float64
```

7. Initial_admin, which is a qualitative and nominal variable. This indicates whether a patient was initially admitted as an emergency, elective, or observation. A little over half of all patients were admitted as an Emergency.

```
#Statistically describe independent variable - Initial_admin
df.Initial_admin.describe()
```

```
count                   10000
unique                      3
top       Emergency Admission
freq                     5060
Name: Initial_admin, dtype: object
```

Looking closer at the data using value.counts() we can see that patients admitted as elective and observation are almost equal.

```
df.Initial_admin.value_counts()
```

```
Initial_admin
Emergency Admission      5060
Elective Admission       2504
Observation Admission    2436
Name: count, dtype: int64
```

8. HighBlood, which is a qualitative and nominal variable. This indicates as yes or no whether the patient has high blood pressure. The below statistics show that a little over half of all patients did not have high blood pressure.

```
#Statistically describe independent variable - HighBlood
df.HighBlood.describe()
```

```
count     10000
unique        2
top          No
freq       5910
Name: HighBlood, dtype: object
```

9. Stroke, which is a qualitative and nominal variable. This indicates as yes or no whether the patient has had a stroke. The below statistics shows that the majority of patients did not have a stroke.

```
#Statistically describe independent variable - Stroke
df.Stroke.describe()
```

```
count     10000
unique        2
top          No
freq       8007
Name: Stroke, dtype: object
```

10. Complication_risk, which is a qualitative and ordinal variable. This variable uses the classifiers high, medium, and low to indicate the level of complication risk for the patient. The below statistics show that almost half of all patients were considered medium for complication risks.

```
#Statistically describe independent variable - Complication_risk
df.Complication_risk.describe()
```

```
count      10000
unique         3
top       Medium
freq        4517
Name: Complication_risk, dtype: object
```

Using value_counts() we can see that patients determined at high risk were the next highest, and that patients determined to be at low risk accounted for the least amount of patients.

```
df.Complication_risk.value_counts()
```

```
Complication_risk
Medium    4517
High      3358
Low       2125
Name: count, dtype: int64
```

11. Overweight, which is a qualitative and nominal variable. This variable indicates as either yes or no whether the patient is overweight. The below statistics show that more than half of all patients are overweight.

```
#Statistically describe independent variable - Overweight
df.Overweight.describe()
```

```
count     10000
unique        2
top         Yes
freq       7094
Name: Overweight, dtype: object
```

12. Diabetes, which is a qualitative and nominal variable. This variable indicates as either yes or no whether the patient has diabetes. The below statistics show that more than half of patients do not have diabetes.

```
#Statistically describe independent variable - Diabetes
df.Diabetes.describe()
```

```
count     10000
unique        2
top          No
freq       7262
Name: Diabetes, dtype: object
```

13. Anxiety, which is a qualitative and nominal variable. This variable indicates as either yes or no whether the patient has anxiety. The below statistics show that more than half of patients did not have anxiety.

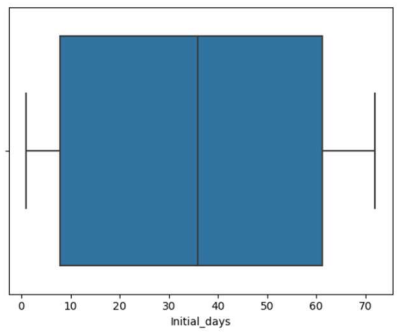
```
: #Statistically describe independent variable - Anxiety
  df.Anxiety.describe()
```

```
: count     10000
  unique        2
  top          No
  freq       6785
  Name: Anxiety, dtype: object
```

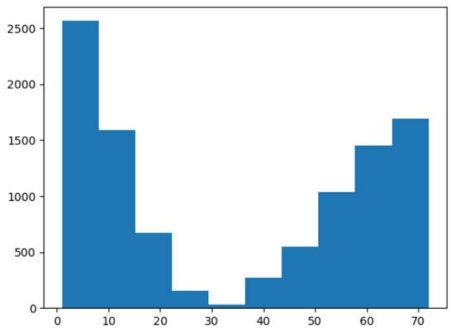
C3: Visualizations

Please see the below univariate visualization of the dependent variable Initial_days using the sns.boxplot() and plt.hist() functions. These visualizations show a bimodal skew without any apparent outliers.

```
: sns.boxplot(df, x='Initial_days')
: <Axes: xlabel='Initial_days'>
```

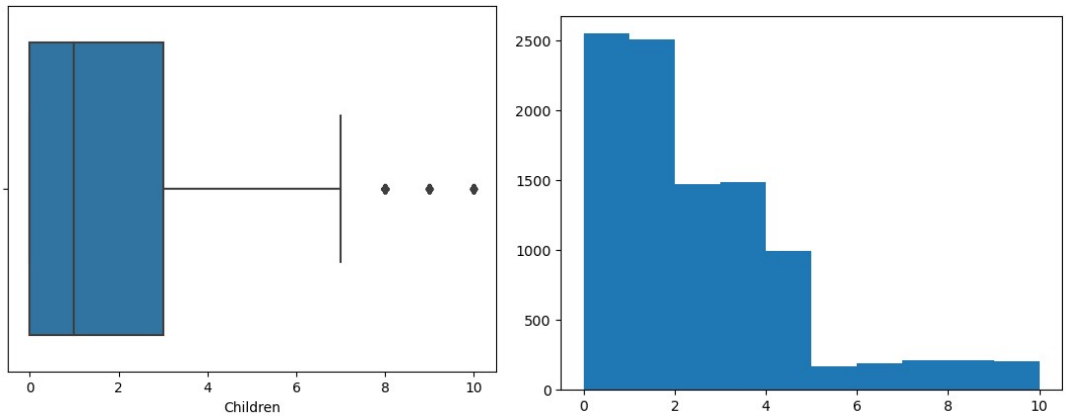



```
plt.hist(df['Initial_days'])
plt.show()
```

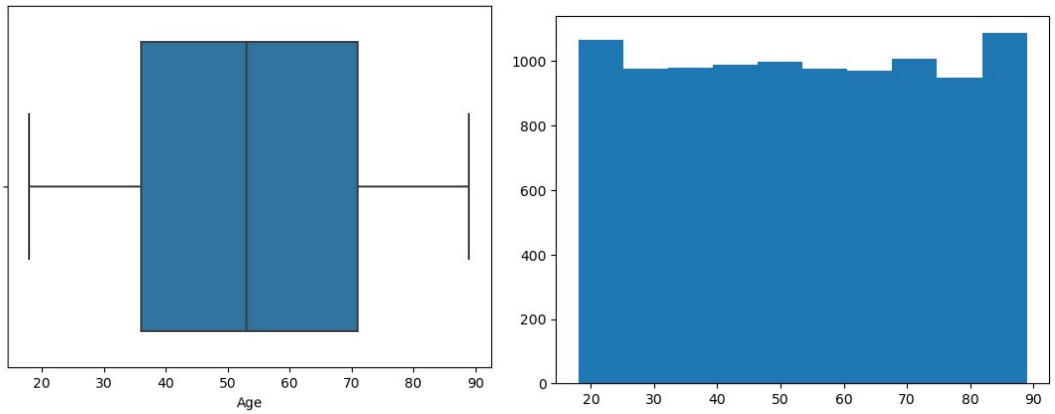


Univariate visualizations of my independent variables:

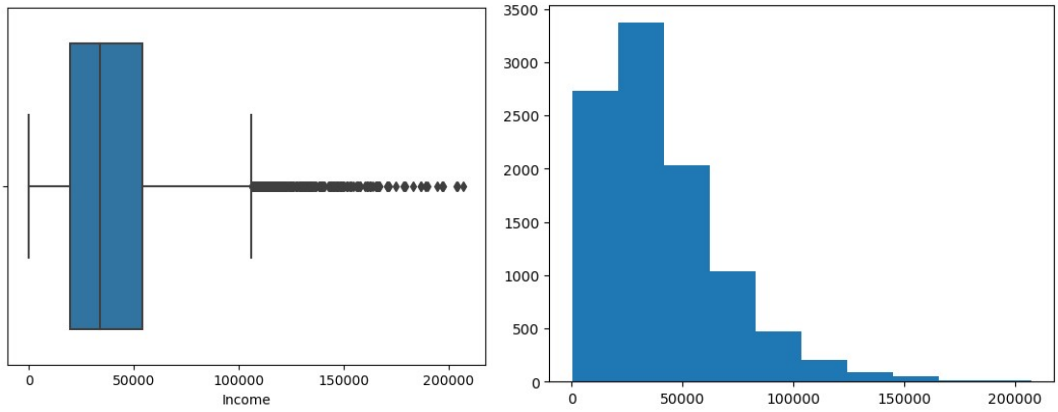
1. The Children variable is shown below as a boxplot and histogram. The boxplot clearly shows outliers, and the histogram indicates the data is positively skewed.



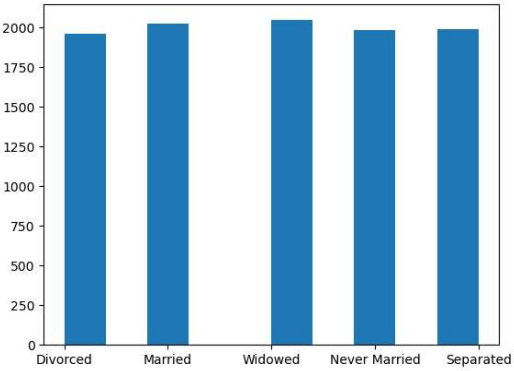
2. The Age variable is shown below as a boxplot and histogram. Per the boxplot, there do not appear to be any outliers. The histogram shows a uniform distribution.



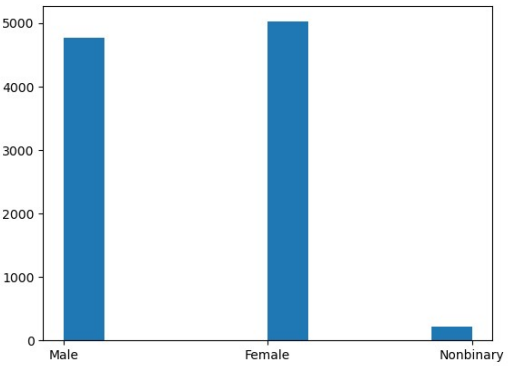
3. The Income variable is shown below as a boxplot and histogram. The boxplot shows there are many outliers. The histogram shows a positive skew.



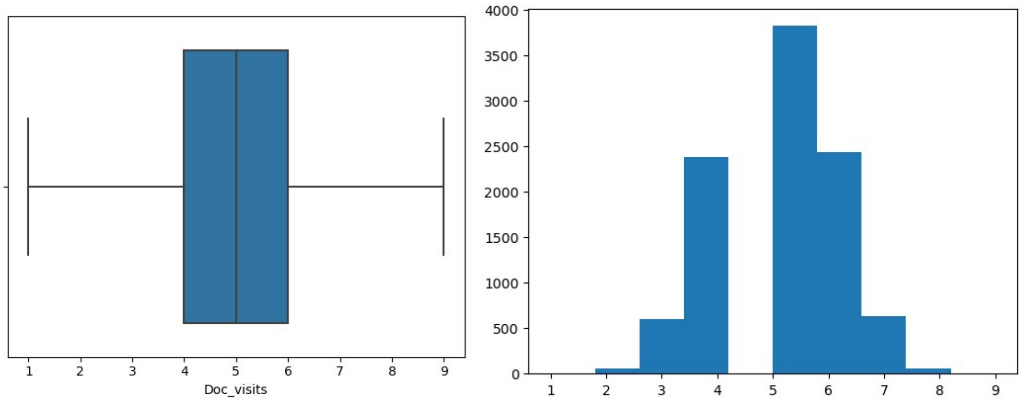
4. The marital variable is shown below as a histogram and seems to be a somewhat uniform distribution.



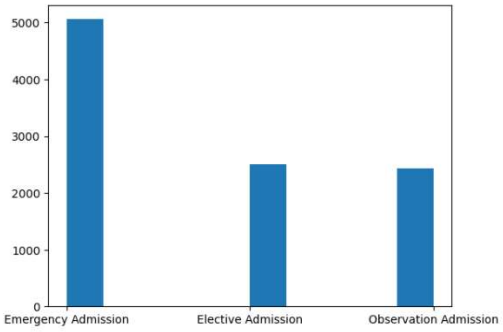
5. The gender variable is shown below as a histogram. The values are in line with what our describe() and value.counts() functions showed us.



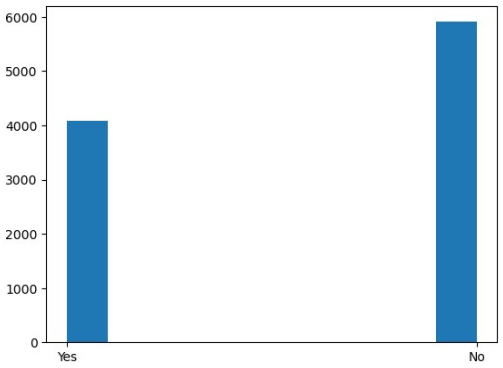
6. The Doc_visits variable is shown below as a boxplot and histogram. The boxplot does not seem to show any outliers. This histogram shows a bimodal distribution.



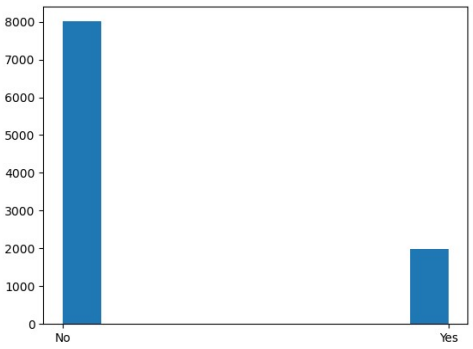
7. The Initial_admin variable is shown below as a histogram. This visualization is in line with what our describe() and value.counts() functions showed us.



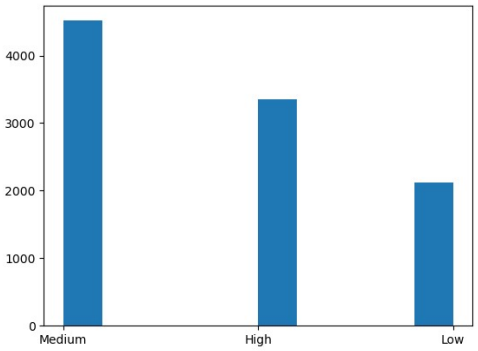
8. The HighBlood variable is shown below as a histogram. This visualization is in line with what our describe() function showed us.



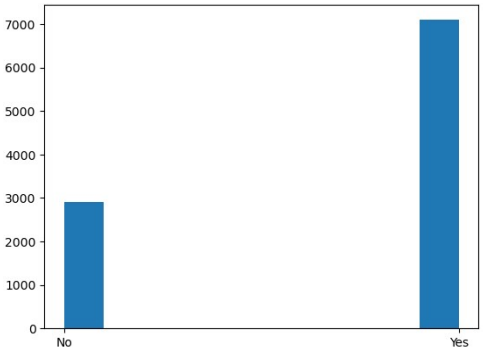
9. The Stroke variable is shown below as a histogram. This visualization is in line with what our describe() function showed us.



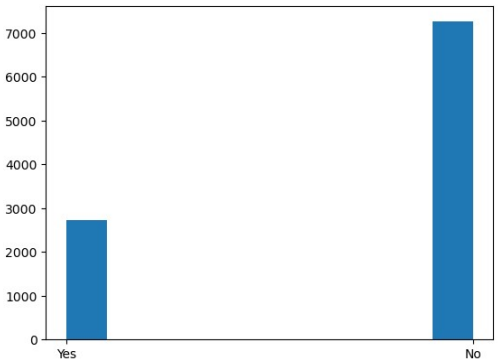
10. The Complication_risk variable is shown below as a histogram. This visualization is in line with what our describe() and value.counts() functions showed us.



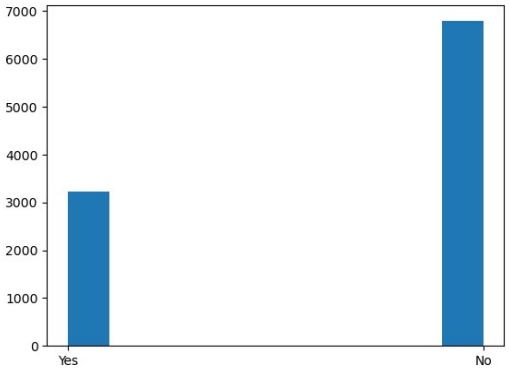
11. The Overweight variable is shown below as a histogram. This visualization is in line with what our describe() function showed us.



12. The Diabetes variable is shown below as a histogram. This visualization is in line with what our describe() function showed us.

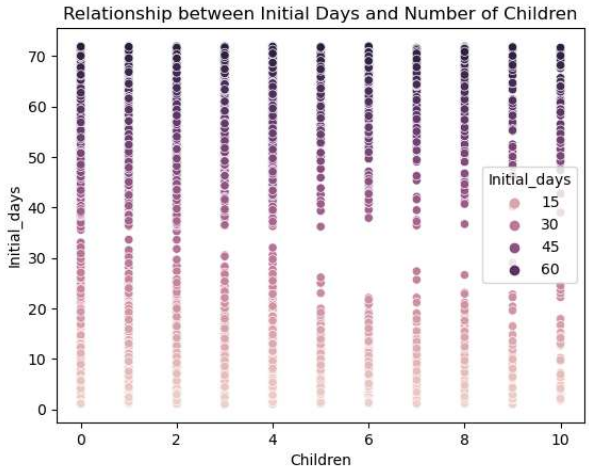


13. The Anxiety variable is shown below as a histogram. This visualization is in line with what our describe() function showed us.

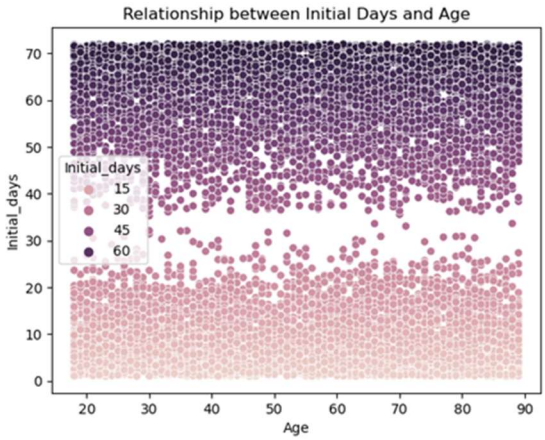


Bivariate visualizations of my dependent and independent variables using the sns.scatterplot() function are below:

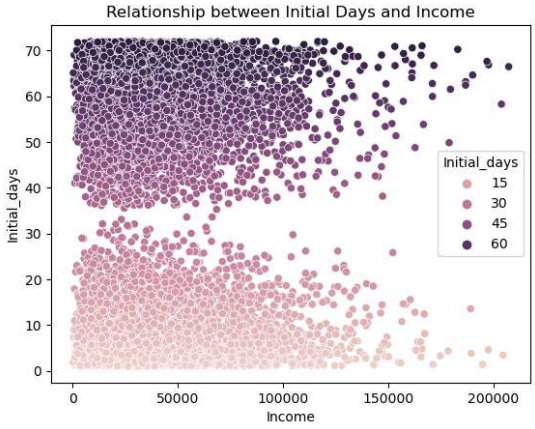
1. The Initial_days and Children scatterplot is below. The sparsity that is present around 30-40 initial days is interesting.



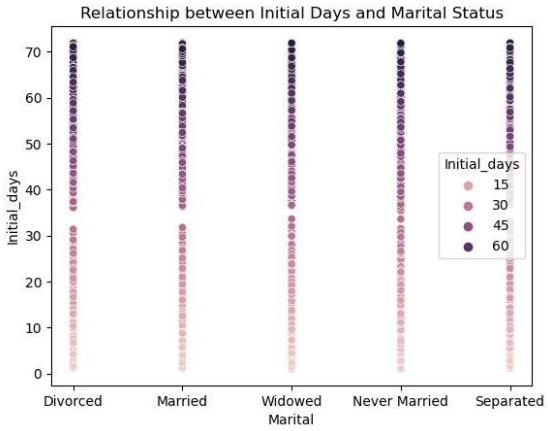
2. The Initial_days and Age scatterplot is below. It appears pretty well distributed, and the sparsity around 30-40 initial days is present.



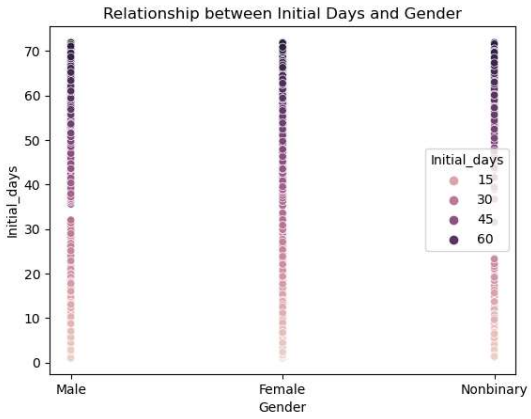
3. The Initial_days and Income scatterplot is below. It is reasonable that the lower income figures are more represented as they make up most of the income variable. It is interesting to see the continued sparsity around 30-40 days.



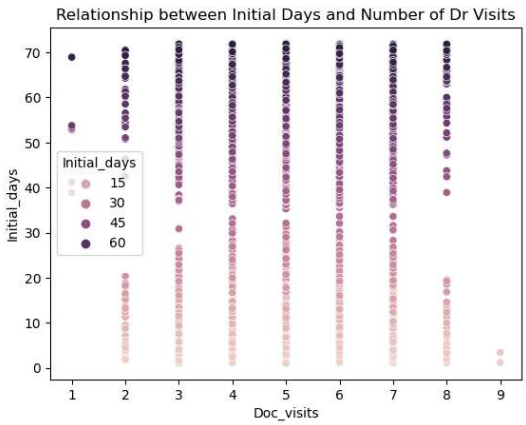
4. The Initial_days and Marital scatterplot is below. All marital statuses appear equally represented, which is expected per the statistics we looked at previously. Again, there is a noted sparsity around 30-40 initial days.



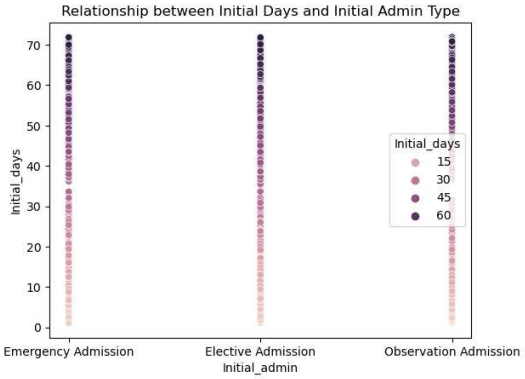
5. The Initial_days and Gender scatterplot is below.



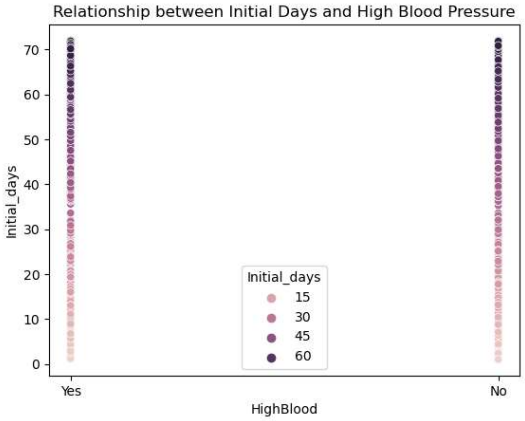
6. The Initial_days and Doc_visits scatterplot is below. The plot appears in line with the statistics we looked at for Doc_visits earlier in the analysis.



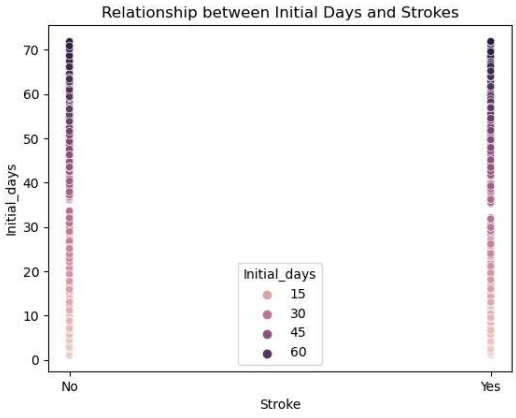
7. The Initial_days and Initial_admin scatterplot is below.



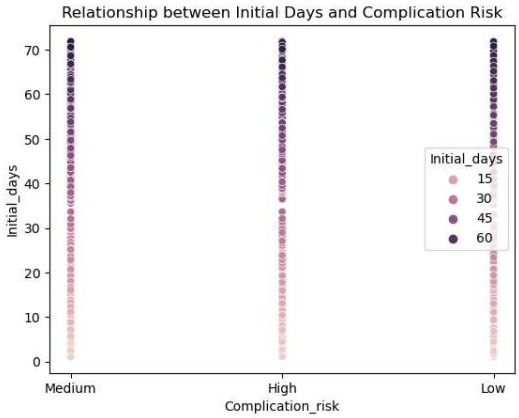
8. The Initial_days and HighBlood scatterplot is below.



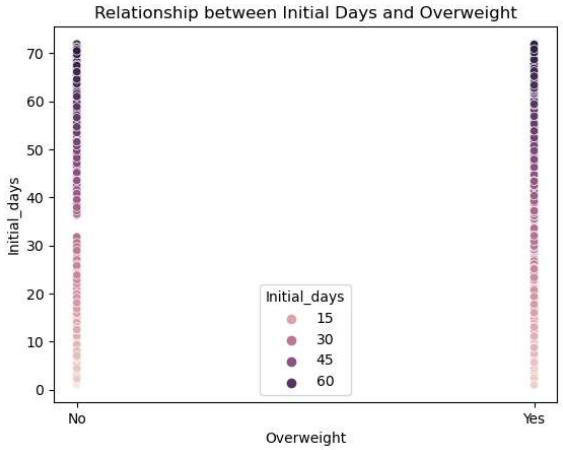
9. The Initial_days and Stroke scatterplot is below. It is interesting to see that sparsity around 30-40 days again.



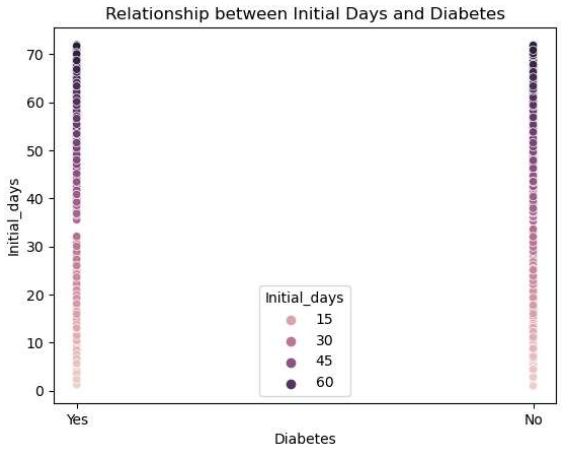
10. The Initial_days and Complication_risk scatterplot is below.



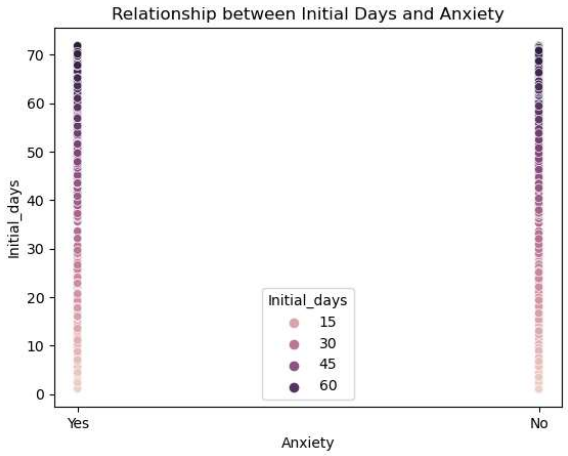
11. The Initial_days and Overweight scatterplot is below.



12. The Initial_days and Diabetes scatterplot is below.



13. The Initial_days and Anxiety scatterplot is below.



C4: Data Transformation

My data transformation goals include re-expressing my qualitative variables to numeric values so I can use them in my multiple linear regression model. For the variables that have yes/no responses, I will use ordinal encoding to set the yes responses as 1s and the no responses as 0s. For the categorical values that have 3 or more response types, I will use one hot encoding as explained by Practical Business Python (Moffitt, 2017). Additionally, I will use the df.drop() function to drop all columns I am not using in my multiple linear regression analysis. This will provide me with a dataset of my dependent and independent variables which are all now numeric values so I can create my regression model.

Here is an example of the ordinal encoding code I used to change Yes and No responses to 1s and 0s:

```
#re-expressing Overweight
df['Overweight_numeric']=df['Overweight']
```

```
#set up dictionary
dict_overweight={'Overweight_numeric' : {'No':0, 'Yes':1}}
```

```
#replace variable's values
df.replace(dict_overweight, inplace=True)
```

```
# drop original column
df=df.drop(columns=['Overweight'])
```

Below is an example of using the onehot encoder to re-express categorical variables with more than two response types:

```
from sklearn.preprocessing import OneHotEncoder

onehot_encoder = OneHotEncoder(sparse=False)
```

```
onehot_encoded=onehot_encoder.fit_transform(df[['Gender', 'Marital', 'Complication_risk', 'Initial_admin']])
```

```
C:\Users\Kmoik WGU\anaconda3\Lib\site-packages\sklearn\preprocessing\_encoders.py:972: FutureWarning: `sparse` was renamed to `
sparse_output` in version 1.2 and will be removed in 1.4. `sparse_output` is ignored unless you leave `sparse` to its default v
alue.
  warnings.warn(
```

```
df_encoded=pd.DataFrame(onehot_encoded, columns=onehot_encoder.get_feature_names_out(['Gender', 'Marital', 'Complication_risk', '
```

```
#drop first columns to reduce multicollinearity
df_encoded.drop(['Gender_Female', 'Marital_Divorced', 'Complication_risk_High', 'Initial_admin_Elective Admission'], axis=1, inpl
```

```
#merging encoded columns to df
df=pd.concat([df, df_encoded], axis=1)
```

After the encoding was completed, I then used the drop() function to drop the columns I was not including in my analysis as well as the first column of the re-expressed variables to reduce instances of multicollinearity.

Please see the attached pdf and ipynb files titled KMoikD208Code1 for the full code used.

C5: Prepared Data Set

Please see the attached CSV file titled: KMoikclean_medical.csv showing my prepared dataset. The columns in the clean data set are below:

```
#confirm data set has been updated appropriately
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 20 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Children                            10000 non-null  int64
 1   Age                                 10000 non-null  int64
 2   Income                              10000 non-null  float64
 3   Doc_visits                          10000 non-null  int64
 4   Initial_days                        10000 non-null  float64
 5   Overweight_numeric                  10000 non-null  int64
 6   Diabetes_numeric                    10000 non-null  int64
 7   Anxiety_numeric                     10000 non-null  int64
 8   HighBlood_numeric                   10000 non-null  int64
 9   Stroke_numeric                      10000 non-null  int64
 10  Gender_Male                         10000 non-null  float64
 11  Gender_Nonbinary                    10000 non-null  float64
 12  Marital_Married                     10000 non-null  float64
 13  Marital_Never Married               10000 non-null  float64
 14  Marital_Separated                   10000 non-null  float64
 15  Marital_Widowed                     10000 non-null  float64
 16  Complication_risk_Low               10000 non-null  float64
 17  Complication_risk_Medium            10000 non-null  float64
 18  Initial_admin_Emergency Admission   10000 non-null  float64
 19  Initial_admin_Observation Admission 10000 non-null  float64
dtypes: float64(12), int64(8)
memory usage: 1.5 MB
```

No duplicate or missing values were located in the dataset. The only outliers in the data were those that were previously explored in D206. I chose to still retain these outliers as I found them to be reasonable and justifiable to keep in the dataset.

After re-expressing my variables and dropping unused and unneeded columns, my resulting data set has 20 columns which includes my 1 dependent variable and 19 independent variables. The columns that changed were the re-expressed columns. I added _numeric to the end of the columns where the Yes and No responses were re-encoded as 1s and 0s. For the columns with more than 2 variables that needed to be re-expressed, a new column was made for every response, with the first column being dropped, and 1s and 0s added. So, the marital column with 5 possible responses became: Marital_Married, Marital_Never Married, Marital_Separated, and Marital_Widowed. Marital_Divorced was dropped to reduce multicollinearity.

Part IV: Model Comparisons and Analysis

D1: Initial Model

I chose to use the Ordinary Least Squares (OLS) method to minimize residuals and find the best-fitting line. Using Ordinary Least Squares (OLS) code from Statology, my initial model with all of my independent variables appeared like this (Statology, 2020):

```
#view linear regression model - OLS
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.334
Date:                Tue, 30 Jan 2024   Prob (F-statistic):              0.150
Time:                        11:47:11   Log-Likelihood:                -46875.
No. Observations:               10000   AIC:                         9.379e+04
Df Residuals:                    9980   BIC:                         9.394e+04
Df Model:                          19
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            33.0968      1.823     18.154      0.000      29.523      36.670
Children                          0.2789      0.122      2.292      0.022       0.040       0.517
Age                               0.0204      0.013      1.598      0.110      -0.005       0.045
Income                        -1.173e-05   9.23e-06     -1.271      0.204   -2.98e-05    6.37e-06
Doc_visits                       -0.1448      0.252     -0.575      0.565      -0.638       0.349
Overweight_numeric               -0.6277      0.580     -1.082      0.279      -1.765       0.509
Diabetes_numeric                 -0.1920      0.590     -0.325      0.745      -1.349       0.965
Anxiety_numeric                   0.6631      0.563      1.177      0.239      -0.441       1.767
HighBlood_numeric                -0.2903      0.536     -0.542      0.588      -1.340       0.760
Stroke_numeric                   -0.1263      0.659     -0.192      0.848      -1.417       1.165
Gender_Male                       0.4112      0.532      0.773      0.440      -0.632       1.455
Gender_Nonbinary                  1.0921      1.837      0.594      0.552      -2.509       4.693
Marital_Married                   1.2619      0.834      1.513      0.130      -0.373       2.897
Marital_Never Married             1.8263      0.838      2.178      0.029       0.183       3.470
Marital_Separated                 1.8174      0.838      2.170      0.030       0.176       3.459
Marital_Widowed                   1.6711      0.832      2.009      0.045       0.040       3.302
Complication_risk_Low             1.1827      0.730      1.621      0.105      -0.248       2.613
Complication_risk_Medium         -0.0685      0.600     -0.114      0.909      -1.244       1.107
Initial_admin_Emergency Admission -0.7608     0.643     -1.183      0.237      -2.022       0.500
Initial_admin_Observation Admission -0.2571   0.749     -0.343      0.732      -1.726       1.212
==============================================================================
Omnibus:                    41464.123   Durbin-Watson:                   0.164
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1277.480
Skew:                           0.070   Prob(JB):                     3.97e-278
Kurtosis:                       1.255   Cond. No.                      3.68e+05
==============================================================================
```

This model gives us the linear regression equation of:

Initial_days = 33.1 + 0.28*Children + 0.02*Age – 1.173e-05*Income – 0.14*Doc_visits – 0.63*Overweight_numeric - 0.19*Diabetes_numeric + 0.66*Anxiety_numeric – 0.29*HighBlood_numeric – 0.13*Stroke_numeric + 0.41*Gender_male + 1.09*Gender_Nonbinary + 1.26*Marital_Married + 1.83*Marital_Never Married + 1.82*Marital_Separated + 1.67*Marital_Widowed + 1.18*Complication_risk_Low – 0.07*Complication_risk_Medium – 0.76*Initial_admin_Emergency Admission – 0.26*Initial_admin_Observation Admission

The R-Squared value of 0.003 in this model is very small. Ideally, we want this value to be close to 1 to indicate the values are close to the actual regression line. This indicates that the predictor variables most likely do not affect the variance of the dependent variable. Additionally, the Prob F-statistic at 0.15 is greater than our alpha value of 0.05 indicating this model as a whole is not statistically significant in its current state.

D2: Justification of Model Reduction

My initial model using OLS shows that several of the variables have p-values greater than our alpha value of 0.05. As such, we will use variance inflation factors (VIF) and backward stepwise elimination to remove the variables that have high levels of multicollinearity and are not statistically relevant to our data. The VIF value represents multicollinearity, and we want to remove any variables that have a value of 10 or greater. The backward stepwise elimination process will then be performed until all variables remaining have a p-value smaller than our alpha value, indicating they are statistically significant. The result of these processes will be variables that have low multicollinearity and are statistically significant to our dependent variable.

D3: Reduced Linear Regression Model

First, I calculated the VIF using code obtained from WGU Course Materials (Western Governors University, n.d.):

```
X=df[['Children', 'Age', 'Income', 'Doc_visits', 'Overweight_numeric', 'Diabetes_numeric', 'Anxiety_numeric', 'HighBlo
```

```
#VIF dataframe
vif_df=pd.DataFrame()
vif_df["feature"]=X.columns
```

```
#calculate VIF for all independent variables
vif_df["VIF"]=[variance_inflation_factor(X.values, i) for i in range(len(X.columns))]
```

```
print(vif_df)
```

```
                            feature        VIF
0                          Children   1.904355
1                               Age   6.635519
2                            Income   2.889753
3                        Doc_visits  12.735734
4                 Overweight_numeric   3.284719
5                   Diabetes_numeric   1.369569
6                     Anxiety_numeric   1.460602
7                  HighBlood_numeric   1.677480
8                     Stroke_numeric   1.243162
9                         Gender_Male   1.904398
10                  Gender_Nonbinary   1.041556
11                   Marital_Married   1.921127
12              Marital_Never Married   1.886120
13                  Marital_Separated   1.896776
14                    Marital_Widowed   1.924777
15              Complication_risk_Low   1.591768
16           Complication_risk_Medium   2.259976
17   Initial_admin_Emergency Admission   2.862566
18  Initial_admin_Observation Admission   1.907937
```

Once I calculated the VIF values for each independent variable, I proceeded to remove Doc_visits from my model as the VIF was greater than 10 at 12.74. After removing Doc_visits, I checked the updated VIF model:

```
#reducing model - removing Doc_visits which has the highest VIF of 12.7 indicating high multicollinearity
X=df[['Children', 'Age', 'Income', 'Overweight_numeric', 'Diabetes_numeric', 'Anxiety_numeric', 'HighBlood_numeric', 'Stroke_nume
```

```
#VIF dataframe
vif_df=pd.DataFrame()
vif_df["feature"]=X.columns
```

```
#calculating VIF for independent variables
vif_df["VIF"]=[variance_inflation_factor(X.values, i) for i in range(len(X.columns))]
```

```
print(vif_df)
```

```
                            feature       VIF
0                          Children  1.871955
1                               Age  5.607090
2                            Income  2.749537
3                 Overweight_numeric  3.113494
4                   Diabetes_numeric  1.356744
5                    Anxiety_numeric  1.448349
6                  HighBlood_numeric  1.656069
7                     Stroke_numeric  1.237705
8                        Gender_Male  1.865492
9                    Gender_Nonbinary  1.039609
10                    Marital_Married  1.831676
11              Marital_Never Married  1.790556
12                  Marital_Separated  1.803622
13                    Marital_Widowed  1.830069
14              Complication_risk_Low  1.560728
15           Complication_risk_Medium  2.196086
16   Initial_admin_Emergency Admission  2.686578
17  Initial_admin_Observation Admission  1.823480
```

Removing Doc_visits resulted in all the remaining variables to have a VIF of 5.6 or less. Instead of removing Age, which has the next highest VIF of 5.6, I will keep it for now as, thinking logically, Age could be an important variable in my model and is below 10 for the VIF.

Now that my reduction based on VIF is complete, I will create a new OLS model with the remaining variables:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.390
Date:                Tue, 30 Jan 2024   Prob (F-statistic):              0.125
Time:                        11:51:16   Log-Likelihood:                -46876.
No. Observations:               10000   AIC:                         9.379e+04
Df Residuals:                    9981   BIC:                         9.393e+04
Df Model:                          18
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            32.3779      1.328     24.387      0.000      29.775      34.980
Children                          0.2791      0.122      2.294      0.022       0.041       0.518
Age                               0.0203      0.013      1.594      0.111      -0.005       0.045
Income                        -1.18e-05   9.23e-06     -1.279      0.201   -2.99e-05    6.29e-06
Overweight_numeric               -0.6319      0.580     -1.090      0.276      -1.769       0.505
Diabetes_numeric                 -0.1965      0.590     -0.333      0.739      -1.353       0.960
Anxiety_numeric                   0.6637      0.563      1.178      0.239      -0.441       1.768
HighBlood_numeric                -0.2928      0.536     -0.547      0.585      -1.343       0.757
Stroke_numeric                   -0.1252      0.659     -0.190      0.849      -1.416       1.166
Gender_Male                       0.4129      0.532      0.776      0.438      -0.630       1.456
Gender_Nonbinary                  1.0940      1.837      0.596      0.551      -2.507       4.695
Marital_Married                   1.2666      0.834      1.518      0.129      -0.369       2.902
Marital_Never Married             1.8339      0.838      2.187      0.029       0.191       3.477
Marital_Separated                 1.8221      0.838      2.176      0.030       0.180       3.464
Marital_Widowed                   1.6760      0.832      2.015      0.044       0.045       3.306
Complication_risk_Low             1.1875      0.730      1.628      0.104      -0.243       2.618
Complication_risk_Medium         -0.0639      0.600     -0.107      0.915      -1.239       1.112
Initial_admin_Emergency Admission -0.7666     0.643     -1.192      0.233      -2.027       0.494
Initial_admin_Observation Admission -0.2665   0.749     -0.356      0.722      -1.735       1.202
==============================================================================
Omnibus:                    41456.246   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1277.757
Skew:                           0.070   Prob(JB):                    3.46e-278
Kurtosis:                       1.254   Cond. No.                     3.48e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The R-Squared value is still low at the same 0.003 and there are quite a few p-values that are greater than the 0.05 alpha value, indicating those variables are not statistically significant. Also, the model notes suggest that there is most likely strong multicollinearity among the variables. Per WGU Course Materials, I will now use backward stepwise elimination process to create updated OLS models by systematically removing the variables with the highest p-values until no variables with p-values greater than the alpha value of 0.05 exist in the model (Western Governors University, n.d.). I will start by removing the variable with the highest p-value, which is Complication_risk_Medium and has a p-value of 0.915:

```
: #view linear regression model - OLS - with Complication_risk_Medium removed
  print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                          OLS   Adj. R-squared:                  0.001
Method:               Least Squares   F-statistic:                     1.471
Date:              Tue, 30 Jan 2024   Prob (F-statistic):             0.0948
Time:                      11:53:38   Log-Likelihood:                -46876.
No. Observations:             10000   AIC:                         9.379e+04
Df Residuals:                  9982   BIC:                         9.392e+04
Df Model:                        17
Covariance Type:           nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            32.3418      1.284     25.192      0.000      29.825      34.858
Children                          0.2792      0.122      2.294      0.022       0.041       0.518
Age                               0.0203      0.013      1.594      0.111      -0.005       0.045
Income                        -1.18e-05   9.23e-06     -1.278      0.201   -2.99e-05    6.29e-06
Overweight_numeric               -0.6328      0.580     -1.091      0.275      -1.769       0.504
Diabetes_numeric                 -0.1967      0.590     -0.333      0.739      -1.354       0.960
Anxiety_numeric                   0.6634      0.563      1.178      0.239      -0.441       1.768
HighBlood_numeric                -0.2929      0.536     -0.547      0.585      -1.343       0.757
Stroke_numeric                   -0.1253      0.659     -0.190      0.849      -1.416       1.166
Gender_Male                       0.4125      0.532      0.775      0.438      -0.631       1.456
Gender_Nonbinary                  1.0952      1.837      0.596      0.551      -2.505       4.696
Marital_Married                   1.2675      0.834      1.520      0.129      -0.368       2.903
Marital_Never Married             1.8348      0.838      2.189      0.029       0.192       3.478
Marital_Separated                 1.8221      0.837      2.176      0.030       0.180       3.464
Marital_Widowed                   1.6764      0.832      2.015      0.044       0.046       3.307
Complication_risk_Low             1.2242      0.643      1.903      0.057      -0.037       2.485
Initial_admin_Emergency Admission -0.7673     0.643     -1.193      0.233      -2.028       0.493
Initial_admin_Observation Admission -0.2686   0.749     -0.359      0.720      -1.737       1.199
==============================================================================
Omnibus:                    41455.614   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1277.772
Skew:                           0.070   Prob(JB):                    3.43e-278
Kurtosis:                       1.254   Cond. No.                     3.48e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

My R-Squared value has been reduced to 0.002 and I still have large p-values. I will continue with my backward stepwise elimination process. The next variable I will eliminate is Stroke_numeric which has a p-value of 0.8:

```
#view linear regression model - OLS - without Stroke_numeric
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.561
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0705
Time:                        11:53:58   Log-Likelihood:                -46876.
No. Observations:               10000   AIC:                         9.379e+04
Df Residuals:                    9983   BIC:                         9.391e+04
Df Model:                          16
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            32.3168      1.277     25.308      0.000      29.814      34.820
Children                          0.2791      0.122      2.294      0.022       0.041       0.518
Age                               0.0203      0.013      1.592      0.111      -0.005       0.045
Income                        -1.18e-05   9.23e-06     -1.279      0.201   -2.99e-05    6.29e-06
Overweight_numeric               -0.6327      0.580     -1.091      0.275      -1.769       0.504
Diabetes_numeric                 -0.1972      0.590     -0.334      0.738      -1.354       0.960
Anxiety_numeric                   0.6649      0.563      1.181      0.238      -0.439       1.769
HighBlood_numeric                -0.2936      0.536     -0.548      0.584      -1.344       0.756
Gender_Male                       0.4130      0.532      0.776      0.438      -0.630       1.456
Gender_Nonbinary                  1.0941      1.837      0.596      0.551      -2.506       4.695
Marital_Married                   1.2692      0.834      1.522      0.128      -0.366       2.904
Marital_Never Married             1.8364      0.838      2.191      0.028       0.193       3.479
Marital_Separated                 1.8226      0.837      2.176      0.030       0.181       3.464
Marital_Widowed                   1.6788      0.832      2.019      0.044       0.049       3.309
Complication_risk_Low             1.2244      0.643      1.903      0.057      -0.037       2.485
Initial_admin_Emergency Admission -0.7663     0.643     -1.192      0.233      -2.027       0.494
Initial_admin_Observation Admission -0.2686   0.749     -0.359      0.720      -1.737       1.199
==============================================================================
Omnibus:                    41456.088   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1277.772
Skew:                           0.070   Prob(JB):                    3.43e-278
Kurtosis:                       1.254   Cond. No.                     3.48e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

My R-Squared value remains the same, as does the note regarding strong multicollinearity. I will next remove the variable Diabetes_numeric which has a p-value of 0.73:

```
: #view linear regression model - OLS - without Diabetes_numeric
  print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.658
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0520
Time:                        11:54:19   Log-Likelihood:                -46876.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9984   BIC:                         9.390e+04
Df Model:                          15
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           32.2601      1.266     25.490      0.000      29.779      34.741
Children                         0.2781      0.122      2.287      0.022       0.040       0.517
Age                              0.0203      0.013      1.591      0.112      -0.005       0.045
Income                       -1.177e-05   9.23e-06     -1.275      0.202   -2.99e-05    6.32e-06
Overweight_numeric               -0.6314      0.580     -1.089      0.276      -1.768       0.505
Anxiety_numeric                  0.6654      0.563      1.181      0.237      -0.439       1.769
HighBlood_numeric                -0.2927      0.536     -0.546      0.585      -1.343       0.757
Gender_Male                      0.4133      0.532      0.776      0.437      -0.630       1.456
Gender_Nonbinary                 1.0958      1.837      0.597      0.551      -2.504       4.696
Marital_Married                  1.2727      0.834      1.526      0.127      -0.362       2.907
Marital_Never Married            1.8372      0.838      2.192      0.028       0.194       3.480
Marital_Separated                1.8213      0.837      2.175      0.030       0.180       3.463
Marital_Widowed                  1.6819      0.832      2.023      0.043       0.052       3.312
Complication_risk_Low            1.2230      0.643      1.901      0.057      -0.038       2.484
Initial_admin_Emergency Admission   -0.7642      0.643     -1.189      0.235      -2.025       0.496
Initial_admin_Observation Admission   -0.2674      0.749     -0.357      0.721      -1.735       1.200
==============================================================================
Omnibus:                    41454.652   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1277.833
Skew:                           0.070   Prob(JB):                    3.33e-278
Kurtosis:                       1.254   Cond. No.                     3.48e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The next variable with the highest p-value is Initial_admin_Observation_Admission with a p-value of 0.71. I will now remove that variable from the OLS model:

```
#view linear regression model - OLS - without Initial_admin_Observation Admission
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.707
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0527
Time:                        11:54:37   Log-Likelihood:                -46877.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9986   BIC:                         9.388e+04
Df Model:                          13
Covariance Type:            nonrobust
=====================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                33.2288      0.990     33.551      0.000      31.287      35.170
Children                              0.2800      0.122      2.302      0.021       0.042       0.518
Income                           -1.199e-05   9.23e-06     -1.299      0.194   -3.01e-05    6.1e-06
Overweight_numeric                   -0.6402      0.580     -1.104      0.269      -1.777       0.496
Anxiety_numeric                       0.6703      0.563      1.190      0.234      -0.434       1.774
HighBlood_numeric                    -0.2880      0.536     -0.538      0.591      -1.338       0.762
Gender_Male                           0.4044      0.532      0.760      0.447      -0.638       1.447
Gender_Nonbinary                      1.1024      1.837      0.600      0.548      -2.498       4.703
Marital_Married                       1.2840      0.834      1.540      0.124      -0.351       2.919
Marital_Never Married                 1.8233      0.838      2.176      0.030       0.180       3.466
Marital_Separated                     1.8189      0.837      2.172      0.030       0.177       3.460
Marital_Widowed                       1.6832      0.832      2.024      0.043       0.053       3.313
Complication_risk_Low                 1.2232      0.643      1.901      0.057      -0.038       2.484
Initial_admin_Emergency Admission    -0.6371      0.526     -1.210      0.226      -1.669       0.395
==============================================================================
Omnibus:                    41412.477   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1279.339
Skew:                           0.070   Prob(JB):                    1.57e-278
Kurtosis:                       1.253   Cond. No.                     3.47e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.47e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The R-Squared value remains at 0.002. I will now remove HighBlood_numeric as it has the highest p-value left of 0.59:

```
#view linear regression model - OLS - without HighBlood_numeric
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.825
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0388
Time:                        11:55:01   Log-Likelihood:                -46877.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9987   BIC:                         9.387e+04
Df Model:                          12
Covariance Type:            nonrobust
=====================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                33.1138      0.967     34.243      0.000      31.218      35.009
Children                              0.2799      0.122      2.301      0.021       0.041       0.518
Income                           -1.198e-05   9.23e-06     -1.299      0.194   -3.01e-05    6.1e-06
Overweight_numeric                   -0.6484      0.580     -1.119      0.263      -1.784       0.488
Anxiety_numeric                       0.6678      0.563      1.186      0.236      -0.436       1.772
Gender_Male                           0.4026      0.532      0.757      0.449      -0.640       1.445
Gender_Nonbinary                      1.0873      1.836      0.592      0.554      -2.512       4.687
Marital_Married                       1.2800      0.834      1.535      0.125      -0.355       2.915
Marital_Never Married                 1.8299      0.838      2.184      0.029       0.187       3.472
Marital_Separated                     1.8229      0.837      2.177      0.030       0.182       3.464
Marital_Widowed                       1.6906      0.831      2.033      0.042       0.061       3.320
Complication_risk_Low                 1.2327      0.643      1.917      0.055      -0.028       2.493
Initial_admin_Emergency Admission    -0.6367      0.526     -1.210      0.226      -1.668       0.395
==============================================================================
Omnibus:                    41410.649   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1279.372
Skew:                           0.070   Prob(JB):                    1.54e-278
Kurtosis:                       1.253   Cond. No.                     3.47e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.47e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The next variable I will remove is Gender_Nonbinary that has a p-value of 0.55:

```
#view linear regression model - OLS - without Gender_Nonbinary
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.008
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0199
Time:                        11:55:14   Log-Likelihood:                -46876.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9987   BIC:                         9.387e+04
Df Model:                          12
Covariance Type:            nonrobust
=====================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                32.0606      1.186     27.022      0.000      29.735      34.386
Children                              0.2779      0.122      2.285      0.022       0.040       0.516
Age                                   0.0203      0.013      1.593      0.111      -0.005       0.045
Income                           -1.179e-05   9.23e-06     -1.278      0.201   -2.99e-05     6.3e-06
Overweight_numeric                   -0.6391      0.579     -1.103      0.270      -1.775       0.497
Anxiety_numeric                       0.6639      0.563      1.179      0.238      -0.440       1.768
Gender_Male                           0.3720      0.527      0.706      0.480      -0.660       1.404
Marital_Married                       1.2579      0.834      1.509      0.131      -0.376       2.892
Marital_Never Married                 1.8344      0.838      2.190      0.029       0.192       3.477
Marital_Separated                     1.8184      0.837      2.172      0.030       0.177       3.459
Marital_Widowed                       1.6835      0.831      2.025      0.043       0.054       3.313
Complication_risk_Low                 1.2285      0.643      1.911      0.056      -0.032       2.489
Initial_admin_Emergency Admission    -0.6326      0.526     -1.202      0.229      -1.664       0.399
==============================================================================
Omnibus:                    41442.045   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1278.250
Skew:                           0.070   Prob(JB):                    2.70e-278
Kurtosis:                       1.254   Cond. No.                     2.88e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.88e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The next variable I will remove from the model is Gender_male, which has a p-value of 0.48:

```
#view linear regression model - OLS without Gender_Male
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.145
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0146
Time:                        11:55:34   Log-Likelihood:                -46876.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9988   BIC:                         9.386e+04
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            32.2480      1.156     27.887      0.000      29.981      34.515
Children                          0.2776      0.122      2.282      0.022       0.039       0.516
Age                               0.0202      0.013      1.582      0.114      -0.005       0.045
Income                        -1.177e-05   9.23e-06     -1.276      0.202   -2.99e-05    6.31e-06
Overweight_numeric               -0.6405      0.579     -1.105      0.269      -1.776       0.495
Anxiety_numeric                   0.6610      0.563      1.174      0.241      -0.443       1.765
Marital_Married                   1.2666      0.833      1.520      0.129      -0.367       2.900
Marital_Never Married             1.8387      0.838      2.195      0.028       0.197       3.481
Marital_Separated                 1.8170      0.837      2.170      0.030       0.176       3.458
Marital_Widowed                   1.6842      0.831      2.026      0.043       0.055       3.314
Complication_risk_Low             1.2264      0.643      1.908      0.056      -0.034       2.487
Initial_admin_Emergency Admission -0.6367      0.526     -1.210      0.226      -1.668       0.395
==============================================================================
Omnibus:                    41430.012   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1278.688
Skew:                           0.070   Prob(JB):                    2.17e-278
Kurtosis:                       1.254   Cond. No.                     2.86e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.86e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The R-squared value remains at 0.002 with high p-values and indications of strong multicollinearity. The next variable I will remove from the model is Overweight_numeric with a p-value of 0.26:

```
#view linear regression model - OLS without Overweight_numeric
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.237
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0134
Time:                        11:55:53   Log-Likelihood:                -46877.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9989   BIC:                         9.386e+04
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            31.7780      1.075     29.550      0.000      29.670      33.886
Children                          0.2793      0.122      2.297      0.022       0.041       0.518
Age                               0.0203      0.013      1.591      0.112      -0.005       0.045
Income                        -1.158e-05   9.22e-06     -1.256      0.209   -2.97e-05     6.5e-06
Anxiety_numeric                   0.6677      0.563      1.186      0.236      -0.436       1.771
Marital_Married                   1.2562      0.833      1.507      0.132      -0.377       2.890
Marital_Never Married             1.8238      0.838      2.177      0.029       0.182       3.466
Marital_Separated                 1.8122      0.837      2.165      0.030       0.171       3.453
Marital_Widowed                   1.6716      0.831      2.011      0.044       0.042       3.301
Complication_risk_Low             1.2332      0.643      1.918      0.055      -0.027       2.493
Initial_admin_Emergency Admission -0.6308      0.526     -1.199      0.231      -1.662       0.401
==============================================================================
Omnibus:                    41403.749   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1279.625
Skew:                           0.070   Prob(JB):                    1.36e-278
Kurtosis:                       1.253   Cond. No.                     2.82e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.82e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The next variable I will remove from the model is Anxiety_numeric with a p-value of 0.236:

```
#view linear regression model - OLS without Anxiety_numeric
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.329
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0129
Time:                        11:56:10   Log-Likelihood:                -46878.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9990   BIC:                         9.385e+04
Df Model:                           9
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            31.9844      1.061     30.139      0.000      29.904      34.065
Children                          0.2805      0.122      2.306      0.021       0.042       0.519
Age                               0.0204      0.013      1.598      0.110      -0.005       0.045
Income                        -1.158e-05   9.22e-06     -1.256      0.209   -2.97e-05     6.5e-06
Marital_Married                   1.2584      0.833      1.510      0.131      -0.375       2.892
Marital_Never Married             1.8142      0.838      2.166      0.030       0.172       3.456
Marital_Separated                 1.8173      0.837      2.171      0.030       0.176       3.458
Marital_Widowed                   1.6683      0.831      2.007      0.045       0.039       3.298
Complication_risk_Low             1.2312      0.643      1.915      0.056      -0.029       2.491
Initial_admin_Emergency Admission    -0.6255      0.526     -1.189      0.235      -1.657       0.406
==============================================================================
Omnibus:                    41386.297   Durbin-Watson:                   0.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1280.283
Skew:                           0.070   Prob(JB):                     9.78e-279
Kurtosis:                       1.253   Cond. No.                      2.81e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.81e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The next variable I will remove from the model is Initial_admin_Emergency admission with a p-value of 0.235:

```
#view linear regression model - OLS without Initial_admin_Emergency Admission
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.444
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0122
Time:                        11:56:26   Log-Likelihood:                -46878.
No. Observations:               10000   AIC:                         9.377e+04
Df Residuals:                    9991   BIC:                         9.384e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         31.6601      1.026     30.870      0.000      29.650      33.671
Children                       0.2792      0.122      2.296      0.022       0.041       0.517
Age                            0.0204      0.013      1.604      0.109      -0.005       0.045
Income                     -1.135e-05   9.22e-06     -1.230      0.219   -2.94e-05    6.73e-06
Marital_Married                1.2482      0.833      1.498      0.134      -0.385       2.882
Marital_Never Married          1.8218      0.838      2.175      0.030       0.180       3.464
Marital_Separated              1.8178      0.837      2.171      0.030       0.177       3.459
Marital_Widowed                1.6612      0.831      1.999      0.046       0.032       3.291
Complication_risk_Low          1.2242      0.643      1.904      0.057      -0.036       2.484
==============================================================================
Omnibus:                    41360.321   Durbin-Watson:                   0.162
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1281.173
Skew:                           0.070   Prob(JB):                     6.26e-279
Kurtosis:                       1.252   Cond. No.                      2.80e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.8e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The next variable I will remove from the model is Income with a p-value of 0.21:

```
#view linear regression model - OLS without Income
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.577
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0119
Time:                        11:56:40   Log-Likelihood:                -46879.
No. Observations:               10000   AIC:                         9.377e+04
Df Residuals:                    9992   BIC:                         9.383e+04
Df Model:                           7
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                      31.1853      0.950     32.818      0.000      29.323      33.048
Children                    0.2780      0.122      2.287      0.022       0.040       0.516
Age                         0.0206      0.013      1.619      0.105      -0.004       0.046
Marital_Married             1.2644      0.833      1.517      0.129      -0.369       2.898
Marital_Never Married       1.8225      0.838      2.176      0.030       0.181       3.464
Marital_Separated           1.8294      0.837      2.185      0.029       0.188       3.470
Marital_Widowed             1.6753      0.831      2.016      0.044       0.046       3.305
Complication_risk_Low       1.2196      0.643      1.897      0.058      -0.041       2.480
==============================================================================
Omnibus:                    41353.973   Durbin-Watson:                   0.162
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1281.450
Skew:                           0.070   Prob(JB):                   5.45e-279
Kurtosis:                       1.252   Cond. No.                         321.
==============================================================================

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

The model notes now indicate that there no longer is strong multicollinearity among the variables. However, there are still variables remaining that have a p-value greater than the alpha value of 0.05, which indicates a lack of statistic relevance to our model. I will continue to remove these high p-value variables. The next variable I will remove from the model is Marital_Married with a p-value of 0.12:

```
#view linear regression model - OLS without Marital_Married
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.622
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0153
Time:                        11:56:53   Log-Likelihood:                -46880.
No. Observations:               10000   AIC:                         9.377e+04
Df Residuals:                    9993   BIC:                         9.382e+04
Df Model:                           6
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                      31.8187      0.854     37.272      0.000      30.145      33.492
Children                    0.2785      0.122      2.290      0.022       0.040       0.517
Age                         0.0208      0.013      1.631      0.103      -0.004       0.046
Marital_Never Married       1.1807      0.723      1.633      0.102      -0.236       2.598
Marital_Separated           1.1875      0.722      1.644      0.100      -0.229       2.604
Marital_Widowed             1.0334      0.715      1.444      0.149      -0.369       2.436
Complication_risk_Low       1.2185      0.643      1.895      0.058      -0.042       2.479
==============================================================================
Omnibus:                    41326.496   Durbin-Watson:                   0.162
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1282.545
Skew:                           0.071   Prob(JB):                   3.15e-279
Kurtosis:                       1.251   Cond. No.                         231.
==============================================================================

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

Removing Marital_Married caused some of the remaining variables' p-values to increase. The next variable I will remove from the model is Marital_Widowed which now has a p-value of 0.14:

```
#view linear regression model - OLS without Marital_Widowed
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.748
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0267
Time:                        11:58:15   Log-Likelihood:                -46883.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9995   BIC:                         9.381e+04
Df Model:                           4
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       33.2894      0.448     74.311      0.000      32.411      34.168
Children                     0.2774      0.122      2.281      0.023       0.039       0.516
Marital_Never Married        0.8081      0.681      1.187      0.235      -0.527       2.143
Marital_Separated            0.8298      0.680      1.219      0.223      -0.504       2.164
Complication_risk_Low        1.2186      0.643      1.895      0.058      -0.042       2.479
==============================================================================
Omnibus:                    41263.743   Durbin-Watson:                   0.161
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1284.867
Skew:                           0.071   Prob(JB):                    9.88e-280
Kurtosis:                       1.250   Cond. No.                         9.49
==============================================================================

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

The R-squared value is now 0.001 but there are still variables in the model that are not statistically relevant and their p-values have actually increased with the removal of other variables with higher p-values. The next variable I will remove is Marital_Never Married with a p-value of 0.23:

```
#view linear regression model - OLS without Marital_Never Married
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     3.194
Date:                Tue, 30 Jan 2024   Prob (F-statistic):             0.0225
Time:                        11:58:49   Log-Likelihood:                -46883.
No. Observations:               10000   AIC:                         9.377e+04
Df Residuals:                    9996   BIC:                         9.380e+04
Df Model:                           3
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       33.4967      0.413     81.203      0.000      32.688      34.305
Children                     0.2745      0.122      2.258      0.024       0.036       0.513
Marital_Separated            0.6296      0.659      0.955      0.340      -0.663       1.922
Complication_risk_Low        1.2133      0.643      1.887      0.059      -0.047       2.474
==============================================================================
Omnibus:                    41235.073   Durbin-Watson:                   0.160
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1285.912
Skew:                           0.071   Prob(JB):                    5.86e-280
Kurtosis:                       1.249   Cond. No.                         8.10
==============================================================================

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

The next variable I will remove is Marital_Separated with a p-value of 0.34:

```
#view linear regression model - OLS without Marital_Separated
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.001
Model:                          OLS   Adj. R-squared:                  0.001
Method:               Least Squares   F-statistic:                     4.336
Date:              Tue, 30 Jan 2024   Prob (F-statistic):             0.0131
Time:                      11:59:09   Log-Likelihood:                -46884.
No. Observations:             10000   AIC:                         9.377e+04
Df Residuals:                  9997   BIC:                         9.380e+04
Df Model:                         2
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       33.6225      0.391     86.010      0.000      32.856      34.389
Children                     0.2731      0.122      2.247      0.025       0.035       0.511
Complication_risk_Low        1.2234      0.643      1.903      0.057      -0.037       2.484
==============================================================================
Omnibus:                    41218.705   Durbin-Watson:                   0.160
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1286.475
Skew:                           0.071   Prob(JB):                     4.42e-280
Kurtosis:                       1.249   Cond. No.                         7.82
==============================================================================

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

The last variable I will remove from the model is Complication_risk_Low. This variable is close to being statistically relevant with a p-value of 0.057, but it is greater than the alpha value of 0.05:

```
#view linear regression model - OLS without Complication_risk_Low
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.001
Model:                          OLS   Adj. R-squared:                  0.000
Method:               Least Squares   F-statistic:                     5.049
Date:              Tue, 30 Jan 2024   Prob (F-statistic):             0.0247
Time:                      12:22:21   Log-Likelihood:                -46886.
No. Observations:             10000   AIC:                         9.378e+04
Df Residuals:                  9998   BIC:                         9.379e+04
Df Model:                         1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.8824      0.366     92.490      0.000      33.164      34.600
Children       0.2732      0.122      2.247      0.025       0.035       0.512
==============================================================================
Omnibus:                    41168.684   Durbin-Watson:                   0.159
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1288.297
Skew:                           0.070   Prob(JB):                     1.78e-280
Kurtosis:                       1.247   Cond. No.                         4.43
==============================================================================

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

My final linear regression model using OLS shows that only one of my independent variables, Children, is statistically relevant to our dependent variable Initial_days. The R-squared value is 0.001, which indicates that 0.1% of the variance of Initial_days can be explained using the Children variable. Logically, this makes sense that the number of children a patient has generally would not be believed to affect how long a patient is initially admitted to a hospital for. However, the Prob (F-statistic) of 0.0247 indicates that our model as a whole is statistically significant. As this value is less than 0.05, this means that Children has a significant association with the number of Initial_days a patient is admitted for (Statology, 2020).

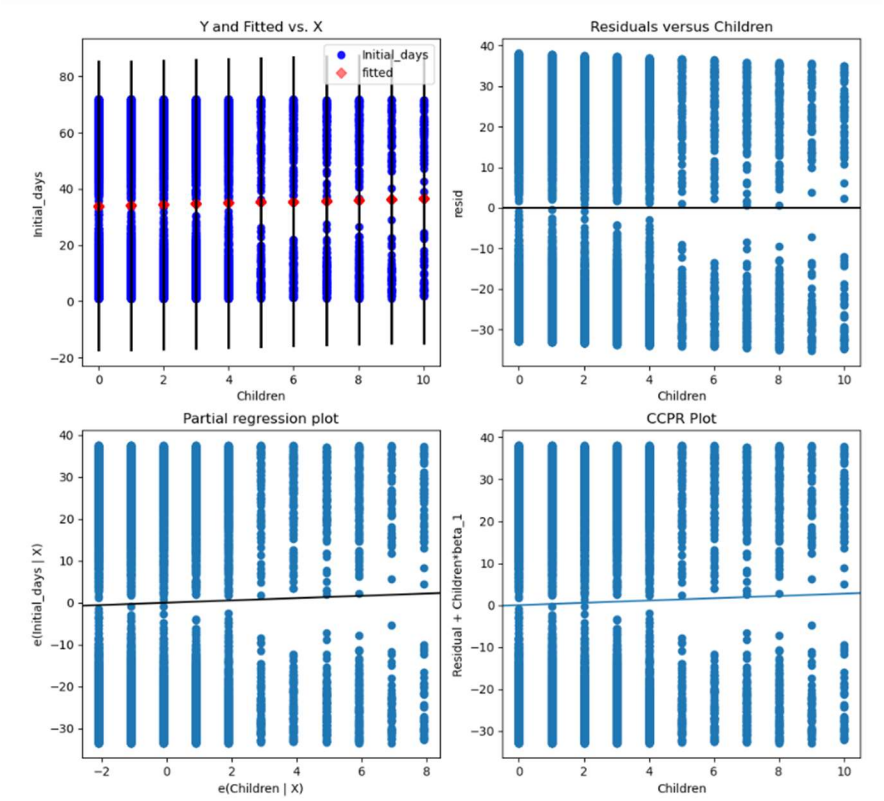
E1: Model Comparison

My initial multiple linear regression model had many variables in it, not all of which were statistically important to the model itself. One variable was removed due to a VIF greater than 12, indicating a high probability of multicollinearity. An additional 17 variables were removed from the model using the backward stepwise elimination method. These variables were removed as their p-value was greater than the 0.05 alpha value. Variables with p-values greater than the alpha value generally are not considered statistically significant, and therefore should be removed from the model. This process was completed one variable at a time starting with the variable with the highest p-value and then reviewing the updated model to see how the removal of that variable affected the other variables left including their p-value. Ultimately, I ended up with only one variable that had p-value less than the alpha value and thus was found to be statistically significant.

My initial model had a very low R-squared value; however, my final model had an even smaller, but similar, R-squared value. This indicates that my model does not explain the variance between my independent and dependent variables. However, my final model did have a Prob F-statistic value less than the alpha value, indicating that the final model as a whole may be statistically relevant. This would seem to indicate that the final model is a better model than the initial model.

E2: Output and Calculations

Using code from Statology, I was able to produce a residual plot of my final model (Statology, 2020):



Using additional code from Statology, I was also able to calculate the standardized residuals (Statology, 2020):

```
print(standardized_residuals)
```

```
[-0.89610922 -0.74412663 -1.13790654 ...  1.34786755  1.08944013
  1.32289383]
```

The residuals are all generally small values below an absolute value of 3, which indicates the model fits the data well and there are no outliers (Statology, 2020).

E3: Code

Please see the attached document titled KMoikD208Code1 in both pdf and ipynb format.

Part V: Data Summary and Implications

F1: Results

My initial equation was: Initial_days = 33.1 + 0.28*Children + 0.02*Age – 1.173e-05*Income – 0.14*Doc_visits – 0.63*Overweight_numeric - 0.19*Diabetes_numeric + 0.66*Anxiety_numeric – 0.29*HighBlood_numeric – 0.13*Stroke_numeric + 0.41*Gender_male + 1.09*Gender_Nonbinary + 1.26*Marital_Married + 1.83*Marital_Never Married + 1.82*Marital_Separated + 1.67*Marital_Widowed

+ 1.18*Complication_risk_Low – 0.07*Complication_risk_Medium – 0.76*Initial_admin_Emergency Admission – 0.26*Initial_admin_Observation Admission

After performing multiple linear regression, my final equation was: Initial_days = 33.88 + 0.27*Children. In this equation, 33.88 is the y-intercept, which is the predicted value of Initial_days when all other variables are equal to 0. The coefficient 0.27*Children means that an increase of 1 child would also cause the number of initial days to increase by 0.27.

While the model does not necessarily explain the variance the best, it is statistically significant as whole per the Prob F-statistic and per the low p-value of the variable Children.

Some of the limitations of my analysis include the type of data provided, or not provided as the case may be. This also includes the fact that there is no data provided for any patients under the age of 18 or for those who were admitted for less than 1 day. Additionally, other limitations of my analysis include the assumptions of linear regression in general. It was assumed there was a linear relationship between the number of Initial_days and my independent variables. Linear regression can also result in errors when there is strong correlation between variables or too much data (Vidyashri, 2024).

F2: Recommendations

Unfortunately, I do not believe my model is significant in a practical sense as, while Children may be statistically significant per its p-value, logically, I do not believe it is the most significant variable that would affect Initial_days. My recommendations would involve obtaining more and complete data that would also include more variables. Additional variables I would be interested in and that I think would provide more impactful insight would include more information on the reason for hospitalization, perhaps a diagnosis or type of diagnosis such as ones like cancer, concussion, etc. I would then rerun my analysis using additional variables. If my model remained the same or seemingly not very practical still, I would then consider a different type of model and modify my statistical analysis 1to compare the results and see if it fit the data better.

G: Panopto Demonstration

My Panopto Video can be viewed at this URL:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=584cd179-6aea-4bb1-bb12-b1090149a83e

H: Sources of Third-Party Code

For C4: Moffitt, C. (2017). *Guide to Encoding Categorical Values in Python.* Practical Business Python. https://pbpython.com/categorical-encoding.html

For D1: Statology. (2020). *A Complete Guide to Linear Regression in Python*. Statology. https://www.statology.org/linear-regression-python/

For D3: Western Governors University. (n.d.). *D208 Predictive Modeling Episode 1* [Video]. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=567da34c-96e3-44c7-a160-ae3100f9433d

For D3: Western Governors University. (n.d.). *D208 Predictive Modeling Episode 5* [Video]. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=15852089-76a2-4828-8a42-ad3100e8460c

For E2: Statology. (2020). *How to Create a Residual Plot in Python*. Statology. https://www.statology.org/residual-plot-python/

For E2: Statology. (2020). *How to Calculate Standardized Residuals in Python*. Statology. https://www.statology.org/standardized-residuals-python/

I: Sources

For B1: Western Governors University. (n.d.) *D208 Predictive Modeling Episode 2* [Video]. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b7ead95b-c392-4973-aa9c-ad1901031ab1

For B2: Western Governors University Information Technology. (2024). *R or Python.* Western Governors University. https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html

For B2: Western Governors University. (n.d.). *D208 Predictive Modeling Episode 5* [Video]. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=15852089-76a2-4828-8a42-ad3100e8460c

For B3 & D3: Statology. (2020). *A Complete Guide to Linear Regression in Python*. Statology. https://www.statology.org/linear-regression-python/

For D3: Western Governors University. (n.d.) *Getting Started with D208 Part I* [Video]. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=15e09c73-c5aa-439d-852f-af47001b8970

For E2: Statology. (2020). *How to Create a Residual Plot in Python*. Statology. https://www.statology.org/residual-plot-python/

For F1: Vidyashri, M. H. (2024). *Advantages And Disadvantages Of Regression Model.* VTUPulse. https://vtupulse.com/machine-learning/advantages-and-disadvantages-of-regression-model/

J: Professional Communication