Linear Regression

Matthew E. Heino

D 208 Predictive Modeling

Linear Regression

Introduction

This paper seeks to cover the concepts of multiple regression using a suitable algorithm for the given dataset. This paper will be composed of a series of sections. Each of these sections will address a specific point in the assessment and the rubric. The format will follow the layout as given in the assessment documentation.

All relevant code for the programming part of the assessment can be found in the included Jupyter Notebook. Code will be included where required or it makes sense to do so.

Background

The goal of this paper is to perform appropriate data preparation, develop a suitable research question, describe the assumptions of a linear regression model, provide justification for using Python as the preferred programming language, complete a model comparison along with analysis of the model, provide a summary of the findings and limitations of the regression model. Each of these will be discussed in depth in the appropriate section of the paper.

The dataset that this analysis will rely upon is the medical dataset. This dataset is composed of 10,000 observations and 50 feature columns. The columns are composed of both categorical and numerical data. The data in these columns model the data that is commonly associated with patients.

For example, there is personal information about the patient. There is information about the charges that the patient has incurred while staying at the hospital. The features that were selected for this assessment will be discussed in a subsequent section.

Part I: Research Question

In this section, there will be a discussion about the question that will be beneficial to the organization. The research question will be relevant to the organization and will be considered a bona fide question of interest as well as defining the goals of the data analysis.

A1: The Organizational Research Question. The question that should be researched makes best use of the data that is found in the dataset. The question needs to be something that the organization is interested in and will be something that adds business value to the organization or its patients. The question that the research seeks to understand is the following: "What factors lead to the initial length of stay in the hospital?"

The reason why this question is of value to the organization is it may be beneficial to see what features are common in the length of stay for the patient. Knowing the features that may play a role may lead to more insight into what can make the patient stay shorter and therefore make the time in the hospital more productive for the staff and less taxing on the patient.

Additional insight may lead to the organization's ability to create cost-cutting measures that will save the medical organization money and resources. Saving in these components means that the organization can use these to pursue other projects and ventures that may be used to provide better services to the patients and better experiences for the patients.

A2: Goals of the Analysis. The goals of the analysis are the following:

1. There will be a valid way to predict the length of stay of the patient based on certain features within the dataset.
2. Provide a way to ascertain what features play an intricate role in the length of stay of the patient.

3. After the analysis is there any course of action that could be implemented to help either reduce the length of stay?
4. If a course of action can be inferred from the research how can it be used to help the organization be more responsive to the treatment of the patients under the organization's care?

Ideally when the research question is answered we want to see a venue that can be pursued to decrease the hospital stay of the patient. This will brought to fruition by looking at the features that are the most contributing to the hospital stay. After examining these features, there will be a need to ask questions about what can be done in the services provided to the patient to help them spend less time in the hospital.

Part II: Justification

In this section, there will be a discussion of assumptions that are made about the multiple linear regression model. There will be a description of using Python in support of the data analysis of the dataset. There will be an explanation why the multiple linear regression for analyzing the research question that was discussed in **Section A1**.

B1: Summary of Assumptions. The four assumptions that must be adhered to. If these assumptions are not observed in the data then any model that is based on the data may be erroneous or misleading. This could lead to undesirable predictions that could affect the decisions that are based on the models and the data that the model is based on. However, some assumptions can be used to ascertain if the given model is good. The multiple linear model must meet the following four assumptions.

The first assumption is that there exists a linear relationship between the variables. There must be a relationship between the independent variable (x) and the dependent variable (y). This assumption can be easy to assess by creating a scatter plot of the data points and seeing if there is a linear relationship between the independent (x values or explanatory variables) and the dependent variables (y values or the response variables). If these points fall into a relatively straight line then it can be assumed that these variables exhibit a linear relationship and they satisfy this assumption (Zach, 2020).

The second assumption is that independent variables are not correlated with one another. The reason why this assumption needs to be met is that it will bring to fruition the concept of multicollinearity. If this assumption is violated then it will be difficult to deduce which one of the independent variables is contributing to the variance within the model (Taylor, n.d.). To look at this assumption another way, when you have an independent or predictor variable there should be no other way to arrive at this variable. It should be distinct in the way it contributes to the model. If other variables compose the model and they can be used to derive the independent variable there is no way to ascertain how much influence the variable has on the model. You will not be able to calculate its effect on the variance.

The third assumption is the residuals of the variance of the residuals is constant over the course of the dataset. In essence, there is a normal distribution of the residuals. The scenario displayed in this assumption is referred to as homoscedasticity (Taylor, n.d.). This assumption means that the regression line will be a straight line that passes "evenly" through the plotted data. If the regression is anything other than this it violates this assumption and another method may need to be used to model the data. Depending on the data that is used it might need to be modeled if the data is not continuous. For example, if the data is composed of categorical data then you would be able to employ a logistic regression to model the data more realistically.

The fourth assumption is the observations that are present in the dataset are their residuals are independent. The values are independent. This will also yield a straight line for the regression (Taylor, n.d.).

B2: Benefits of Python. The benefit of using Python is this language provides a wealth of packages that are available to clean the data and perform the required data analysis. This is illustrated in the table given below.

The first benefit of using Python is that it has packages like **statsmodels** that provide the required modeling class to create multiple linear regression. The other benefit is that it integrates well into a Jupyter Notebook. This integration allows code to be run in a more user-friendly manner. Users can execute the cells and step through the code on a cell-by-cell basis. This ability is often overlooked but it can be a great advantage as it allows the user to step through the code without any knowledge of coding (the user does not need to know how to program).

The Python programming language comes equipped with various packages and libraries that offer extensive facilities to process data and perform the multiple linear regression that is the focus of this document. These packages will be very useful during the creation of the multiple linear regression model. The table below lists some of the packages that will be used in the data preparation and the creation of the regression model.

Python Package	Description
matplotlib	Plotting function for the graphs that are used to graph the features of the dataset.
missingno	Used to graph the existence of any missing values within the dataset.
numpy	Mathematics library that provides math functions like square root.
pandas	Provides methods that can be used to create a dataframe and manipulate the content within it.
seaborn	Used to plot various types of quality graphs,e.g., histogram, violin plots, etc.
statsmodels	Used to create the model for the multiple linear regression.
sklearn	Used to create testing sets for testing the model that was created.

Table 1: Python Packages

B3: Why is Multiple Linear Regression Appropriate? The reason for using multiple linear regression stems from the nature of the data and the question that needs to be answered. Using a simple linear regression will only allow the analysis to focus on two variables at a time. A simple linear regression will be composed of one independent and one dependent variable. This does not always afford the ability to capture all the variance that could be contained within the dataset.

With multiple linear regression, there will be an ability to look at multiple independent features (the predictors) and how they can be used to predict the dependent variable (the response). This type of modeling allows for the analysis of multiple independent variables and seeing how they influence the outcome or the features that are chosen may not have any influence on the outcome.

Using a regression model it may become apparent that only a handful of variables play an integral role in determining the outcome of the response variable. If this is the case we can reduce the model to only the bare minimum variables that can be used to predict the outcome.

Using a simple linear regression we are not able to capture the full variance of the dataset. The analyst can only look at them in a piecemeal fashion. The analyst only gets to see how two variables interact with one another and there is no way to see if there is a relationship that can be used to answer the question proposed in **Section A**.

It is often helpful to look at the cause and the outcome in a way that is composed of many constituent parts. An outcome is not usually composed of one item, it is usually an amalgamation of different events that lead to the outcome. Using multiple linear regression the analyst can look at various features from the dataset and see if they have any role in the outcome. If the features do they can be kept; if the features do not they can be "dropped" from the model.

Part III: Data Preparation

This section there will be a description of the goals of preparing the data for analysis. The section will describe the variables that will be needed to conduct the research and will compose the multiple linear regression model. For each of these variables, there will be univariate and bivariate visualizations.

C1: Goals of Cleaning. The goal of the cleaning stage of data analysis is to make sure that all of the data that is proposed in **Section A** is in the right state. The right state means that all features are in a form that is conducive to analysis. This could mean changing the values that are string to numeric. This will allow for the use of models like multiple linear regression to be used. The multiple linear regression model relies on the fact that any inputs are in a numeric form.

The steps that will be followed are those that were discussed in a previous assessment. The first step will be to look for columns that will not be needed to answer the question. There will be a visual look at the spreadsheet and a review of the data dictionary. These files were provided by the organization. Using this it will be easy to see what variables will be needed and what variables will not be needed. Any variables that are not needed will not be included in the initial reading of the data into the data frame.

These columns or features will be "dropped" from the dataframe as they are not needed for the creation of the multiple linear regression that will developed in subsequent sections of this assessment. This stage will also look for duplicated data. Using duplicates can affect the accuracy of the model.

The next step would be to look for any null values that may be in the columns or rows of the dataset. If values are missing an appropriate strategy will need to be employed to remedy the missing values. While actual imputation will be handled after looking at the outliers.

The next stage will be to look for outliers that may affect the model by changing the equation that will be used to predict any future values. The outliers may have a dramatic effect on the regression line for the model. The outliers will affect the magnitudes of the coefficients.

For example, for variables that are categorical and missing values, it would be prudent to impute using the mode of the variable. If the variable exhibits a normal distribution then the mean would be used to impute the missing variables.

If the column is of the wrong type there will be a need to transform that column's data into a suitable data type that can be used for the model. This will be the case for the categorical data types that may be used for the model.

During this stage, there will be a check for multicollinearity. This check is done so the initial model will not have to undergo this check when the initial regression model is created. The reason that this check is done is that the variables need to be distinct in their effect on the dependent variable. It must be easy to ascertain the effect of each of the included independent variables on the dependent variable.

This will be undertaken by looking at the VIF of each of the independent variables, and any variable that has a score of 10 or higher will be removed from the candidates for the model and the cleaned dataset. The code to accomplish this task is included in the notebook. This task will be completed at the end of Section C. This is done as part of the data cleaning process as a means to make sure that there are no variables that will have an impact on the creation of the initial multiple linear regression model. This concept was discussed in one of Dr. Sewell's lectures. It has been included as the last step here. This is to keep it in line with the assessment tasks and the rubric.

Note: The code for this section can be found in the Jupyter Notebook that is included with this submission. The appropriate code will be found in **Section C** of the notebook. There will be some additional information or notes included in the notebook. It made more sense to include this with relevant code and not include it in this document.

The file name is:

Heino D208 Predictive Modeling Task 1.ipynb.

C2: Variables Required. The variables that are required are a composition of both categorical and continuous variables. It is important to note what each of these variables are in terms of whether they are categorical or numeric. This distinction may have a bearing on how they are treated. For instance, a categorical variable may need to be encoded into a numeric form using a method like **get_dummies()** from the pandas package.

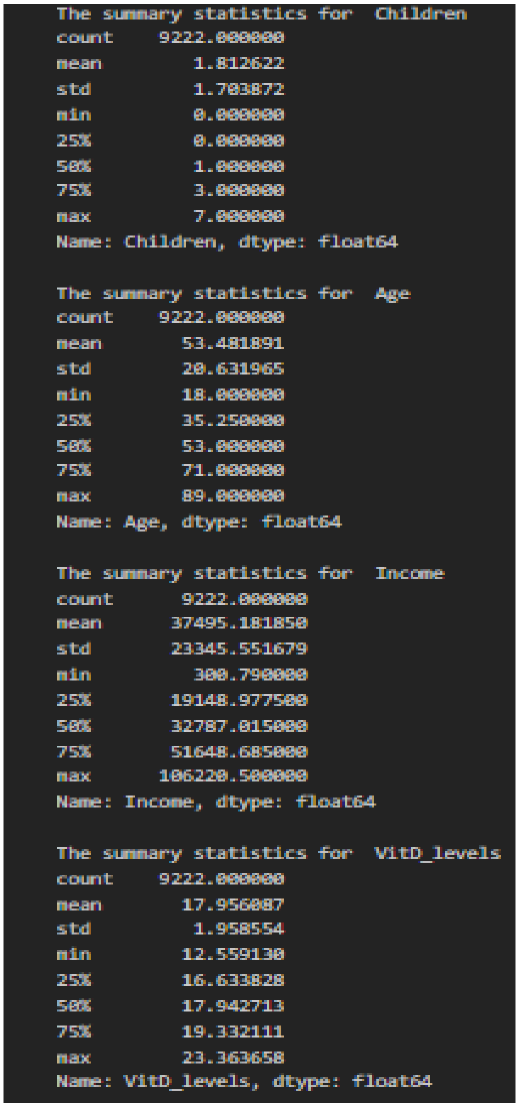
The table below shows the variables that are required to pursue the research question.

	Column Name	Data Type	Description and Notes
1	Children	Continuous	Predictor. Will need to be scaled.
2	Age	Continuous	Predictor. Will need to be scaled.
3	Income	Continuous	Predictor. Will need to be scaled.
4	Gender	Categorical	Predictor. Will need to be transformed
5	VitD_levels	Continuous	Predictor. Will need to be scaled.
6	Doc_visits	Continuous	Predictor. Will need to be scaled.
7	Initial_admin	Categorical	Predictor. Will need to be

			transformed.
8	Complication_risk	Categorical	Predictor. Will need to be transformed.
9	Arthritis	Categorical	Predictor. Will need to be transformed.
10	Diabetes	Categorical	Predictor. Will need to be transformed.
11	BackPain	Categorical	Predictor. Will need to be transformed.
12	TotalCharge	Continuous	Predictor. Will need to be scaled.
13	Initial_days	Continuous	Target. Will need to be scaled.

Table 2: List of Required Variables.

Statistics of the variables. In this section, there will be screenshots of the summary statistics of the continuous variables that will be used in the model are shown below:

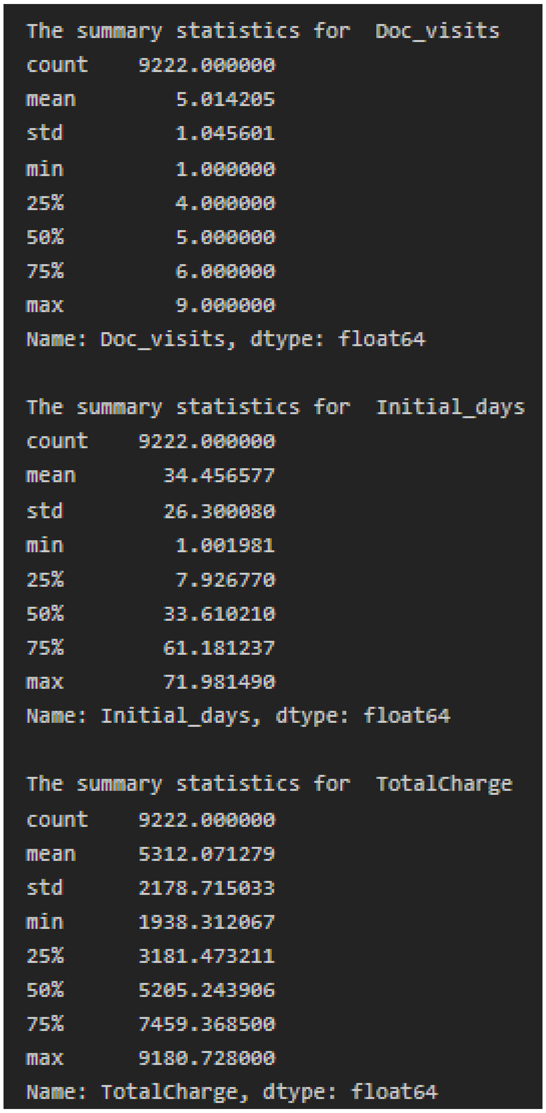


```
The summary statistics for  Children
count    9222.000000
mean        1.812622
std         1.703872
min         0.000000
25%         0.000000
50%         1.000000
75%         3.000000
max         7.000000
Name: Children, dtype: float64

The summary statistics for  Age
count    9222.000000
mean       53.481891
std        20.631965
min        18.000000
25%        35.250000
50%        53.000000
75%        71.000000
max        89.000000
Name: Age, dtype: float64

The summary statistics for  Income
count      9222.000000
mean      37495.181850
std       23345.551679
min         300.790000
25%       19148.977500
50%       32787.015000
75%       51648.685000
max      106220.500000
Name: Income, dtype: float64

The summary statistics for  VitD_levels
count    9222.000000
mean       17.956087
std         1.958554
min        12.559130
25%        16.633828
50%        17.942713
75%        19.332111
max        23.363658
Name: VitD_levels, dtype: float64
```



```
The summary statistics for  Doc_visits
count    9222.000000
mean        5.014205
std         1.045601
min         1.000000
25%         4.000000
50%         5.000000
75%         6.000000
max         9.000000
Name: Doc_visits, dtype: float64

The summary statistics for  Initial_days
count    9222.000000
mean       34.456577
std        26.300080
min         1.001981
25%         7.926770
50%        33.610210
75%        61.181237
max        71.981490
Name: Initial_days, dtype: float64

The summary statistics for  TotalCharge
count    9222.000000
mean     5312.071279
std      2178.715033
min      1938.312067
25%      3181.473211
50%      5205.243906
75%      7459.368500
max      9180.728000
Name: TotalCharge, dtype: float64
```

In reviewing the statistics a few items stand out. The Children variable mean seems to be close to the expected value of 1.94, it is calculated as 1.82 (Average Children per Family U.S. 2022 | Statista, 2023).

If you look at the mean for Age, you will observe that the mean age is roughly 53. With the outliers removed and based on this value the patients that are utilizing medical services are considered middle-aged. It will be interesting if this will have any bearing on the model when it is developed later on. It should be noted that the minimum (min) age for the patients is 18. This could affect how the model will handle ages that are below this value. If these lower age values are used to test the model.

Income has a mean that seems to be a little lower than the average given by FRED. The income reported by FRED is $40,480 while the statistics from the calculation are $37,495.18 (Real Median Personal Income in the United States, 2023b). This may not prove consequential in the model as they seem to be relatively close in approximation.

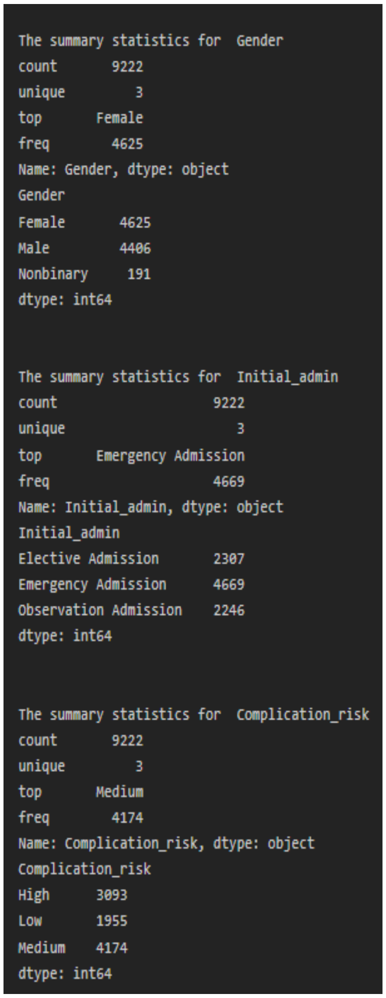
VitD_levels there is difficulty in ascertaining anything of note. There would be a need to investigate as to whether the numbers recorded here are high relative to other known data for this type of measurement.

Doc_visits the mean or the average number of times that the primary physician visited the patient was 5. Which seems good, but it needs to be seen in the appropriate light. If a patient was in the hospital for an extended period it could be interpreted as the physician not visiting enough to keep tabs on the condition of the patient. This could have a bearing on the length of stay the patient has undergone.

Initial_days the mean stay in the hospital for a patient is 53 days. If there is an examination of the percentiles. It is reasonable to deduce that patients are staying a long time in the hospital. Based on the 75 percentile most stay less than 61 days, but that does seem like an excessive amount.

The TotalCharge does seem to be in line with what is expected. The data column is an average charge that is based on the number of days in the hospital.   Based on standard deviation most of the values will fall within the desired two standard deviations.

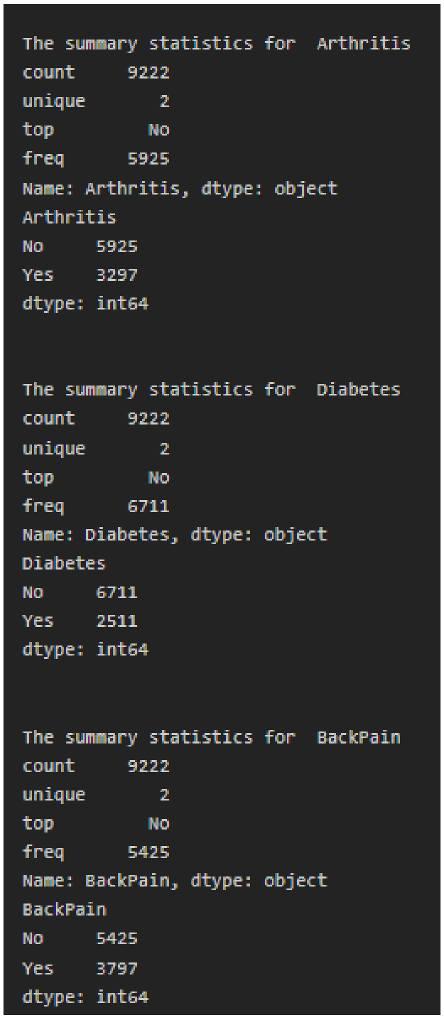
The screenshots below are the summary statistics for the categorical variables that will be used in the model.



```
The summary statistics for  Gender
count       9222
unique         3
top       Female
freq        4625
Name: Gender, dtype: object
Gender
Female       4625
Male         4406
Nonbinary     191
dtype: int64


The summary statistics for  Initial_admin
count                    9222
unique                      3
top       Emergency Admission
freq                     4669
Name: Initial_admin, dtype: object
Initial_admin
Elective Admission       2307
Emergency Admission      4669
Observation Admission    2246
dtype: int64


The summary statistics for  Complication_risk
count       9222
unique         3
top       Medium
freq        4174
Name: Complication_risk, dtype: object
Complication_risk
High      3093
Low       1955
Medium    4174
dtype: int64
```



```
The summary statistics for  Arthritis
count     9222
unique       2
top         No
freq      5925
Name: Arthritis, dtype: object
Arthritis
No     5925
Yes    3297
dtype: int64


The summary statistics for  Diabetes
count     9222
unique       2
top         No
freq      6711
Name: Diabetes, dtype: object
Diabetes
No     6711
Yes    2511
dtype: int64


The summary statistics for  BackPain
count     9222
unique       2
top         No
freq      5425
Name: BackPain, dtype: object
BackPain
No     5425
Yes    3797
dtype: int64
```

Examining these values there are a few things that may be of interest. The gender categorization seems reasonably well distributed with roughly an expected mix of groups.

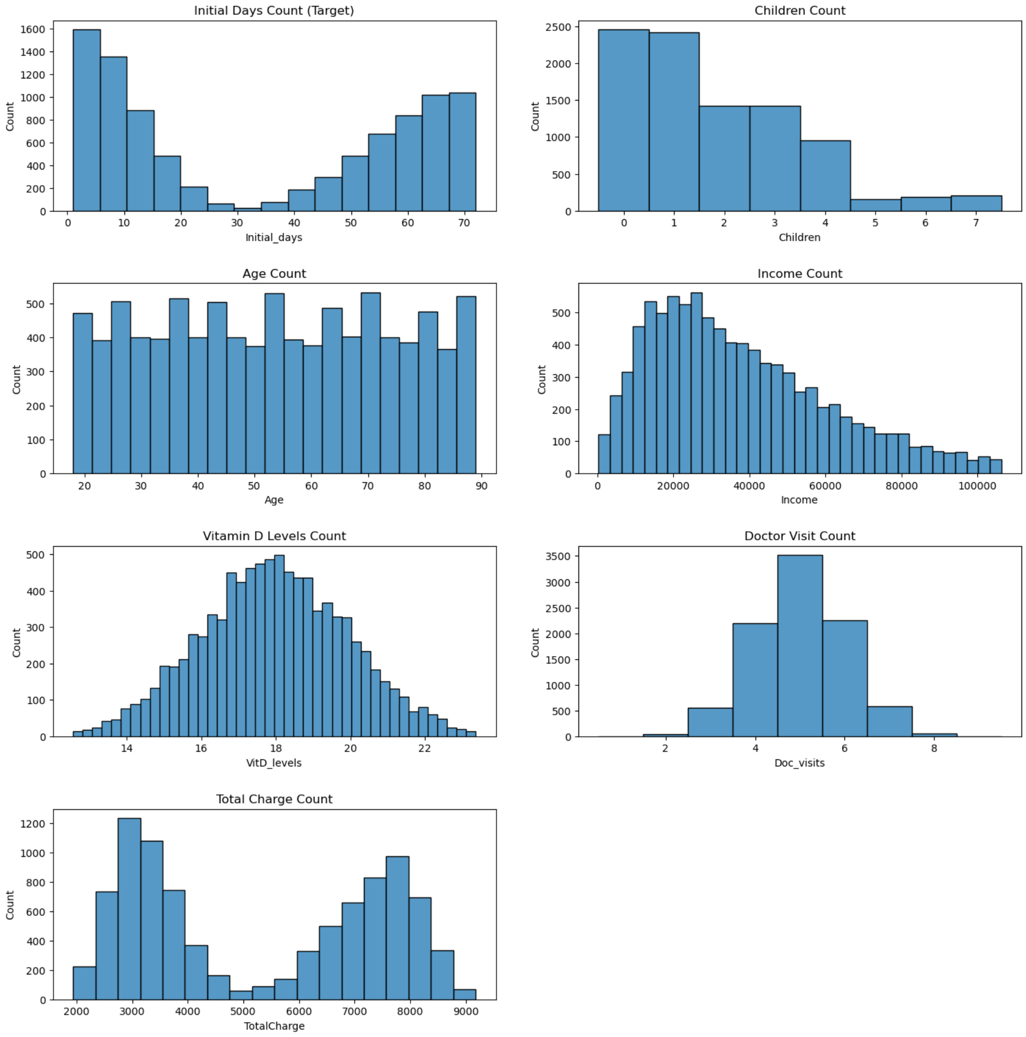
For initial admission, the main reason for the visit is for an emergency. This may influence the duration of the stay as observed in the Initial_days feature. This may be a reason for the long hospital stays of the patient. The Complication_risk seems to indicate that the risk of complication for most of the patients is medium or low as given by the values of 1955 and

4174. This does raise the question of why the hospital stays so long relative to the complication risk. Will this variable play a role in the initial days in the hospital?

The features of Arthritis, Diabetes, and BackPain seem that the vast majority of the patients do not suffer from these afflictions. It will be worth noting if these features will remain in the model after the model is prepared and subsequently reduced in features.

C3: Univariate and Bivariate Visualizations. The chosen variables have graphs that will be associated with their distribution. The summary statistics will not be included but the histogram that is shown below will show how the values in the dataset are distributed.

Univariate Continuous Visualizations

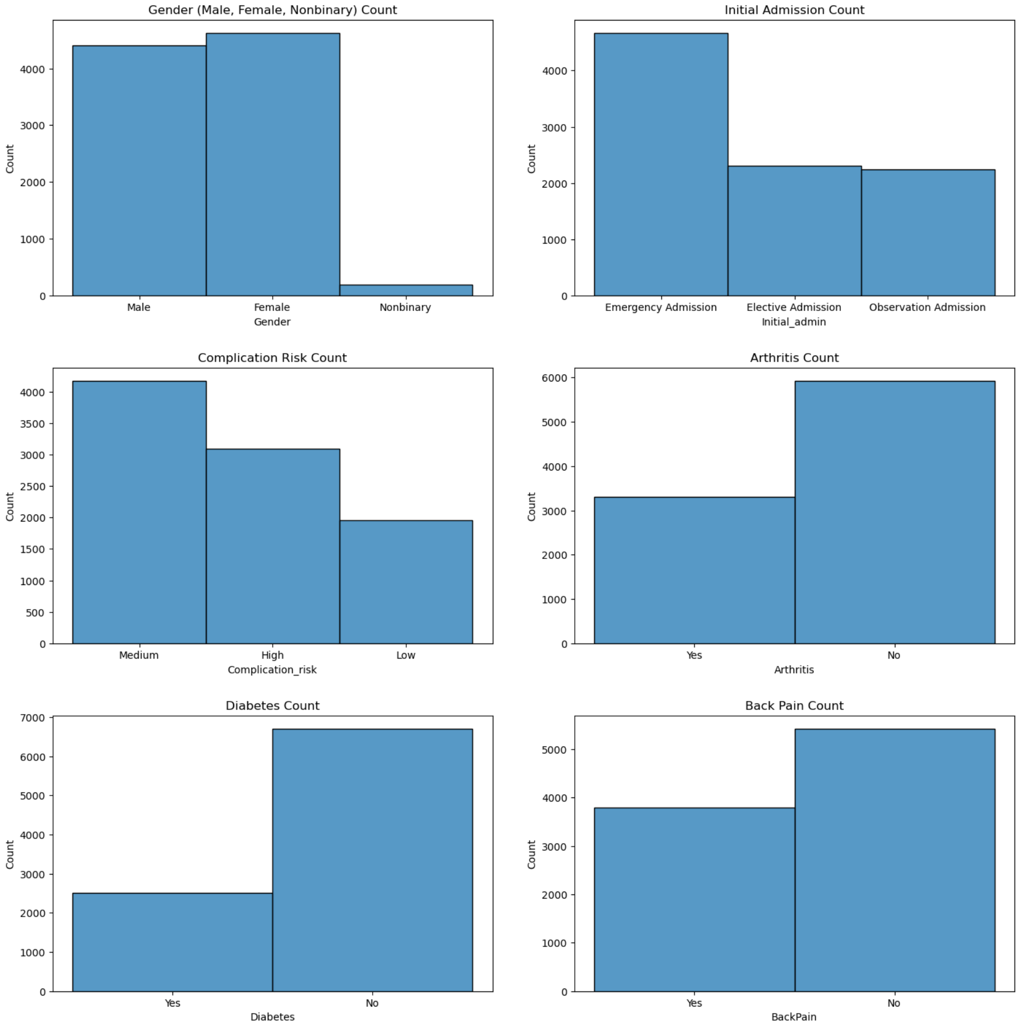


Examining the graphs there are a few things that could be said about them first the VitD_levels and Doc_visits seem to exhibit a very normal distribution of their data. The age column has a distribution that is almost uniform in nature. This does not seem to be a realistic distribution, but this is just a supposition and is not based in fact at this time.

The Initial_days and TotalCharge have a bimodal distribution which may lead to an under-representation of observations in a certain group or subgroup of values. This is evident in the troughs that are depicted in the histogram.

The Children and Income have a right-skewed distribution which is to be expected and nothing seems to be out of the ordinary visually.

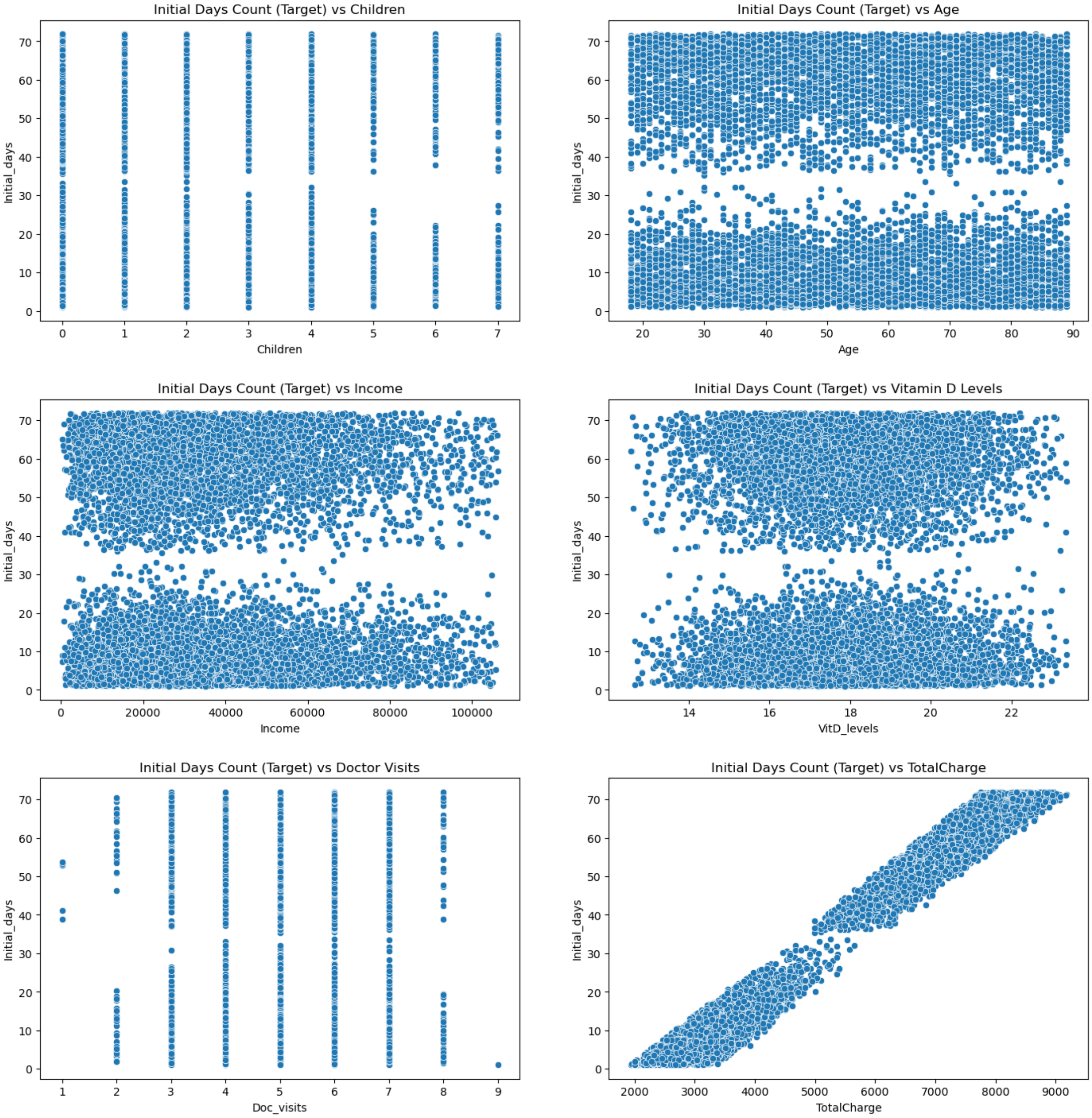
Univariate Categorical Visualizations



Examining these histograms there is nothing out-of-the-ordinarily. The graphs represent the distribution that was observed in statistics that were shown earlier.

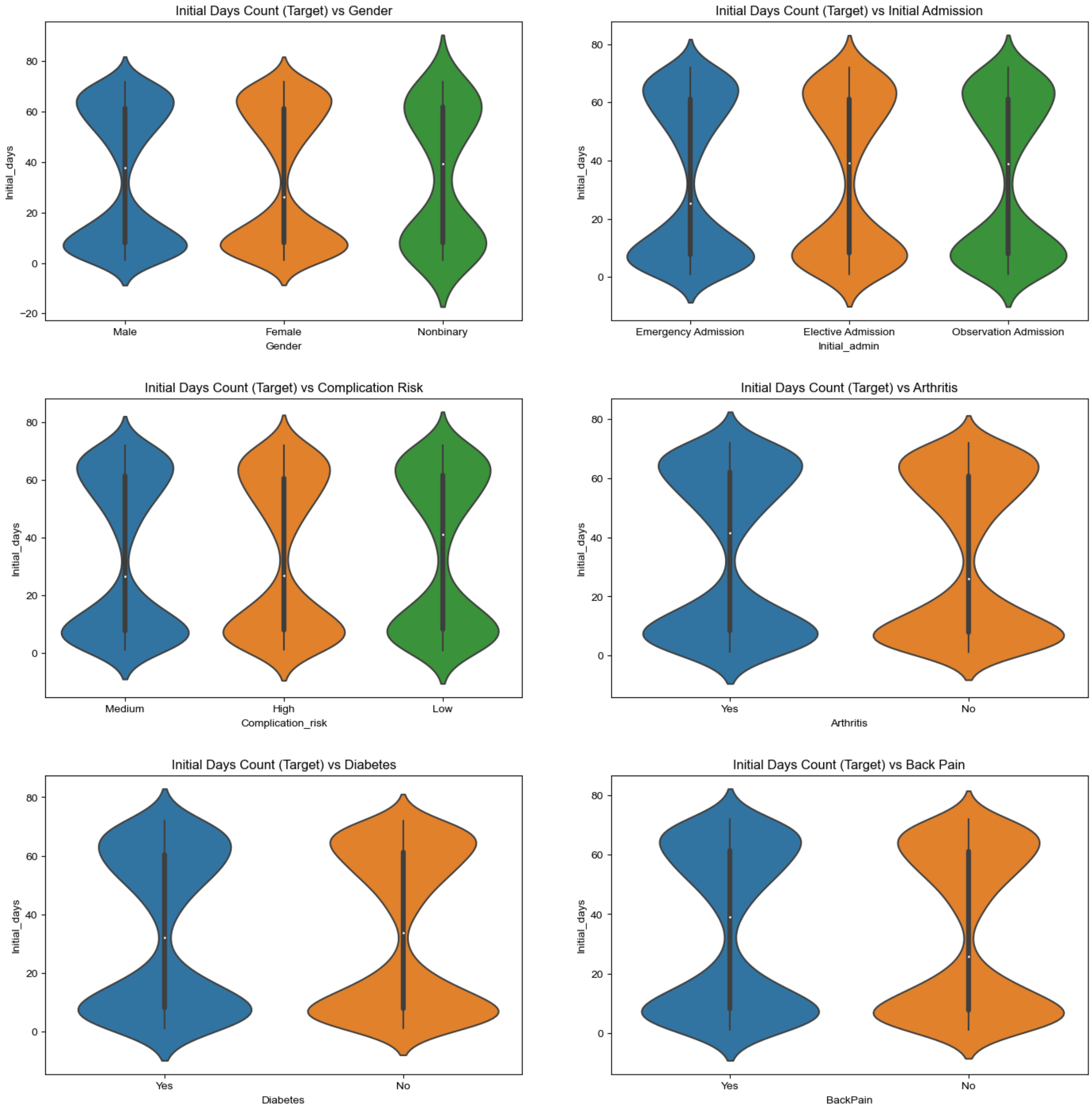
The graphs below show bivariate visualization with each variable plotted against the proposed target variable of Initial_days.

Bivariate Continuous Visualizations



There is something to note with the scatter plots that are shown above. There appear to be two distinct clusters of data in the scattershot for Age, Income, VitD_levels, and TotalCharge. Will these clusters affect the model? It is worthy to note these for possible insight into classifications and/or categorization based on these features. Any further inspection is outside the scope of this paper.

Bivariate Categorical Visualizations



Initial Days Count (Target) vs Gender

Initial Days Count (Target) vs Initial Admission

Initial Days Count (Target) vs Complication Risk

Initial Days Count (Target) vs Arthritis

Initial Days Count (Target) vs Diabetes

Initial Days Count (Target) vs Back Pain

The violin plot was chosen because it was an intuitive way to show the distribution of two variables, the Initial_days (the target) and the categorical candidates. This allowed for inspection of the counts in each of the subgroups within the main group, like Gender or Complication_risk. You will see that there is a depiction of the median as illustrated by the white dot located within the image (Carron, 2021).

The code for the creation of these graphs can be found in the Jupyter Notebook in section C3.

C4: Data Transformation Goals. The goal of the data transformation is to transform the data into a form that will allow a regression model to be used. There are requirements for the creation of the model. Some of these were discussed in a previous section of the paper. As mentioned earlier the multiple linear regression model needs to operate on numerics. The categorical variables that are given in the data file are strings. These categorical variables will need to be converted. This will be handled using the panda's method **get_dummies**.

In this section, there will be an added step. There will be a check of the VIFs of the candidates for the model.

To create a successful initial model there needs to be a check for multicollinearity among the candidate variables. Any variable exhibiting this trait will need to be removed from consideration. The threshold value is a VIF value of 10. If needed the threshold can be adjusted to account for lesser values of correlation. For moderate VIF values, it may be prudent to remove these candidates from consideration. It will depend on how much they affect the standard error (SE) (*Check for Multicollinearity of Model Terms — Check_Collinearity*, n.d.).

C5: Prepared Data Set.

The file that has the prepared dataset is the following:

Heino_reduced_medical_task1.csv

Code for the creation of the CSV file can be found in the accompanying Jupyter Notebook in Section C5.

Part IV: Comparison and Analysis

In this section, there will be a construction of an initial. Then this initial regression will have features removed and this will be discussed in section D2. The final section will illustrate a final linear regression model.

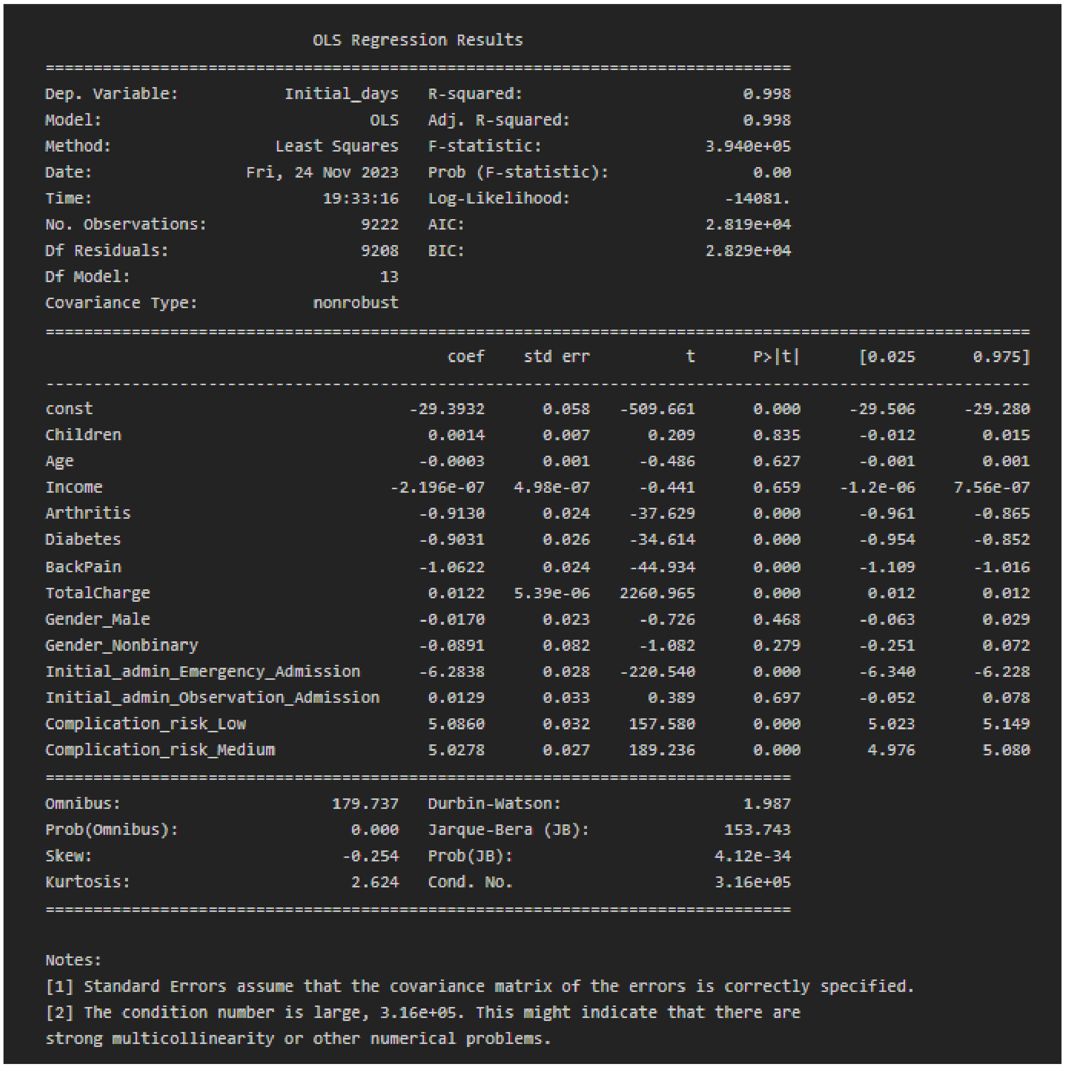
D1: Construction of Initial Multiple Linear Regression Model. After the data has been cleaned and the correct variables have been identified it is now appropriate to start to build a model that can be used to answer the question that was proposed in Section A1.

To revisit an item from the previous section. It was found that VitD_levels and Doc_visits had a VIF score that was above the threshold of 10 that was stated in an earlier section of this paper. These features have been removed from consideration for inclusion into the model. The model will now be composed of the following:

- Children
- Age
- Income
- Arthritis
- Diabetes
- BackPain
- TotalCharge
- Gender_Male
- Gender_Nonbinary
- Initial_admin_Emergency_Admission
- Initial_admin_Observation_Admission
- Complication_risk_Low
- Complication_risk_Medium
- Initial_days

Please note that there are a few extra variables included in the list. This is from using **the get_dummies** method. The code for this can be found in section C4. Data Transformation (cell # 28) of the Jupyter Notebook.

The initial model will be created using the code found in cell 36. The initial regression model summary resulted in the following output.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 3.940e+05
Date:                Fri, 24 Nov 2023   Prob (F-statistic):               0.00
Time:                        19:33:16   Log-Likelihood:                -14081.
No. Observations:                9222   AIC:                         2.819e+04
Df Residuals:                    9208   BIC:                         2.829e+04
Df Model:                          13
Covariance Type:            nonrobust
=======================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
const                                 -29.3932      0.058   -509.661      0.000     -29.506     -29.280
Children                                0.0014      0.007      0.209      0.835      -0.012       0.015
Age                                    -0.0003      0.001     -0.486      0.627      -0.001       0.001
Income                              -2.196e-07   4.98e-07     -0.441      0.659    -1.2e-06    7.56e-07
Arthritis                              -0.9130      0.024    -37.629      0.000      -0.961      -0.865
Diabetes                               -0.9031      0.026    -34.614      0.000      -0.954      -0.852
BackPain                               -1.0622      0.024    -44.934      0.000      -1.109      -1.016
TotalCharge                             0.0122   5.39e-06   2260.965      0.000       0.012       0.012
Gender_Male                            -0.0170      0.023     -0.726      0.468      -0.063       0.029
Gender_Nonbinary                       -0.0891      0.082     -1.082      0.279      -0.251       0.072
Initial_admin_Emergency_Admission      -6.2838      0.028   -220.540      0.000      -6.340      -6.228
Initial_admin_Observation_Admission     0.0129      0.033      0.389      0.697      -0.052       0.078
Complication_risk_Low                   5.0860      0.032    157.580      0.000       5.023       5.149
Complication_risk_Medium                5.0278      0.027    189.236      0.000       4.976       5.080
==============================================================================
Omnibus:                      179.737   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              153.743
Skew:                          -0.254   Prob(JB):                     4.12e-34
Kurtosis:                       2.624   Cond. No.                     3.16e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.16e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 1. Image of Initial Regression Model

Examining the output you will notice the inclusion of the following line of code in the cell: X = sm.add_constant(X). (See cell # 37). This was to add a constant value to remove the following note from the output: "[1] R^2 is computed without centering (uncentered) since the model does not contain a constant." This initial summary used a different method for calculating the R^2 value that does not use a centered mean (Scikit-Learn, n.d.).

D2: Justification of Feature Selection. The features that will be removed from the initial model will be done in a statistically appropriate manner. When creating a model it is advantageous to remove components that do not have a role in the model. The statistic that is often used to determine if a model component is worth keeping in the model is the p-value.

The p-value in the summary is found in the "P>|t" column of the summary output. As discussed in a previous assessment. A value that is greater than 0.05 is a component that is considered not statistically significant. In terms of the feature section, it means that this feature does not need to be included in the features used to create the model. Any feature can be removed from the model and the model should not have been affected by its removal (Bevans, 2023).

This will be employed in a manner that is referred to as Backwards Elimination. This procedure is where you will continue to eliminate features where the p-value is greater than the accepted 0.05. This will eventually result in a list of features that will have a value that is less than 0.05. Meaning that all features that are included are considered statistically significant.

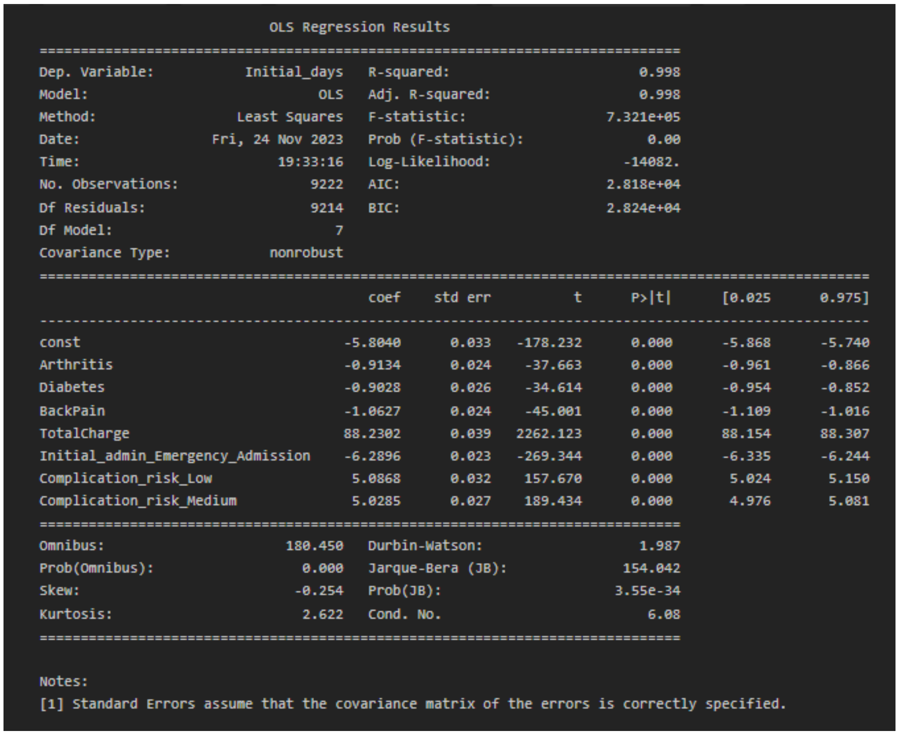
D3: Reduced Linear Regression Model. After performing the feature reduction the model is not composed of the following:

- Children
- Initial_admin_Observation_Admission
- Income
- Age
- Gender_Male
- Gender_Nonbinary

After removing the previous features the model is now composed of the following features:

- Arthritis
- Diabetes
- BackPain
- TotalCharge
- Initial_admin_Emergency_Admission
- Complication_risk_Low
- Complication_risk_Medium

The image below shows the final model with all features that were not statistically significant removed from the model.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 7.321e+05
Date:                Fri, 24 Nov 2023   Prob (F-statistic):               0.00
Time:                        19:33:16   Log-Likelihood:                -14082.
No. Observations:                9222   AIC:                         2.818e+04
Df Residuals:                    9214   BIC:                         2.824e+04
Df Model:                           7
Covariance Type:            nonrobust
=====================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                -5.8040      0.033   -178.232      0.000      -5.868      -5.740
Arthritis                            -0.9134      0.024    -37.663      0.000      -0.961      -0.866
Diabetes                             -0.9028      0.026    -34.614      0.000      -0.954      -0.852
BackPain                             -1.0627      0.024    -45.001      0.000      -1.109      -1.016
TotalCharge                          88.2302      0.039   2262.123      0.000      88.154      88.307
Initial_admin_Emergency_Admission    -6.2896      0.023   -269.344      0.000      -6.335      -6.244
Complication_risk_Low                 5.0868      0.032    157.670      0.000       5.024       5.150
Complication_risk_Medium              5.0285      0.027    189.434      0.000       4.976       5.081
==============================================================================
Omnibus:                      180.450   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              154.042
Skew:                          -0.254   Prob(JB):                     3.55e-34
Kurtosis:                       2.622   Cond. No.                         6.08
==============================================================================

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

Figure 2. Image of Reduced Regression Model

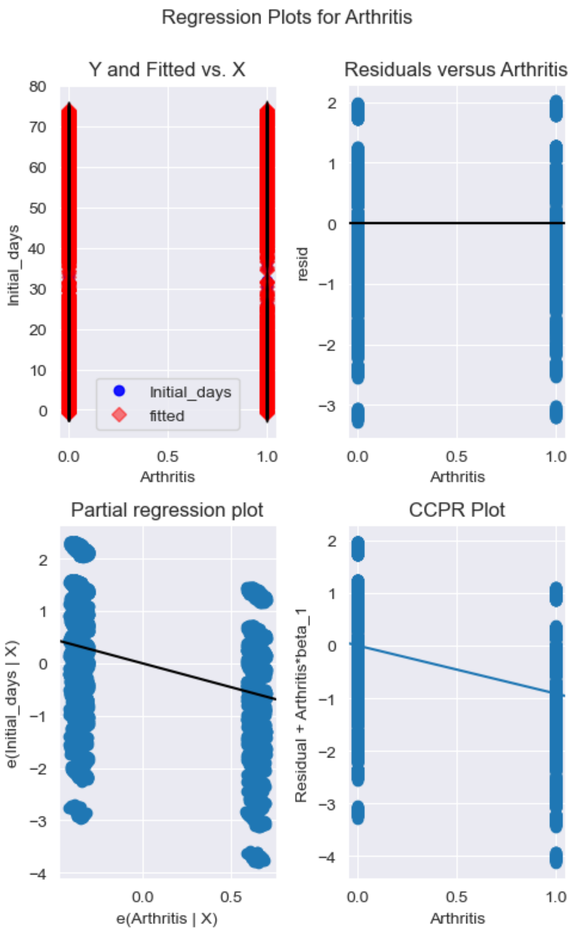
Code for this reduced model can be found in Section D3 Reduced Model and cells 40 – 46. This code was done iteratively and not put into a loop. This is to check to see if the p=values changed between each iteration.

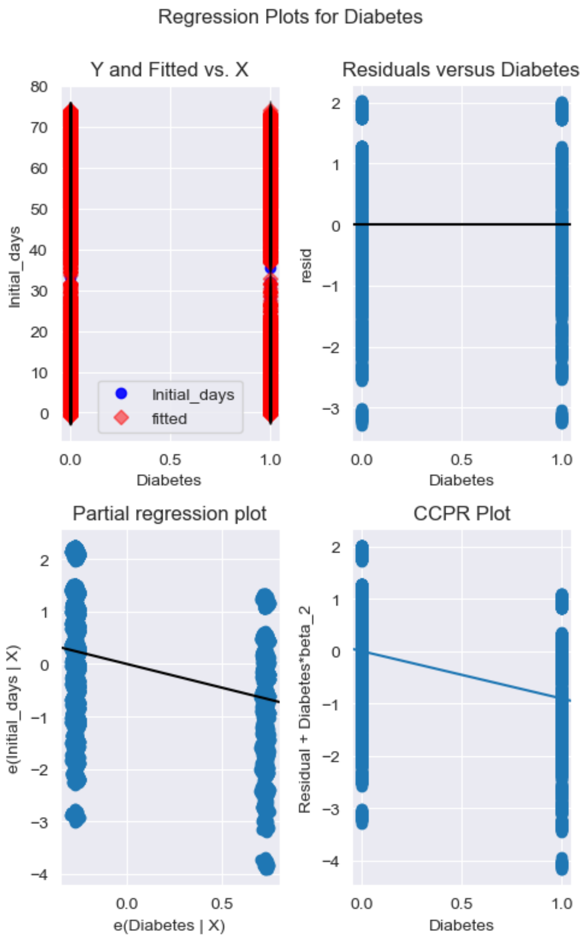
E1: Comparing the Initial and the Reduced Linear Regression Model. The models that have been created both had a stated R2 of .998. The initial regression model had features that were not required and were not beneficial to the model. Also, looking at the standard you will see a slightly higher value for the the included features of the model. In the reduced model the model has a lower standard error which indicates a slightly better model.

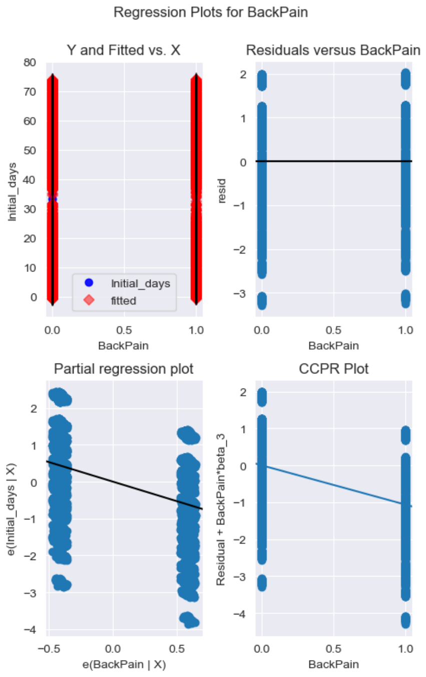
Examining the residual standard error both yielded an RSE of ~1.112. So using this metric it seems that there is no real difference in the models in these metrics. The reduced model does obtain these numbers with fewer components and should be seen as a model that will require less data when comes to processing. This may be preferential in larger datasets. It will reduce the required computing resources.

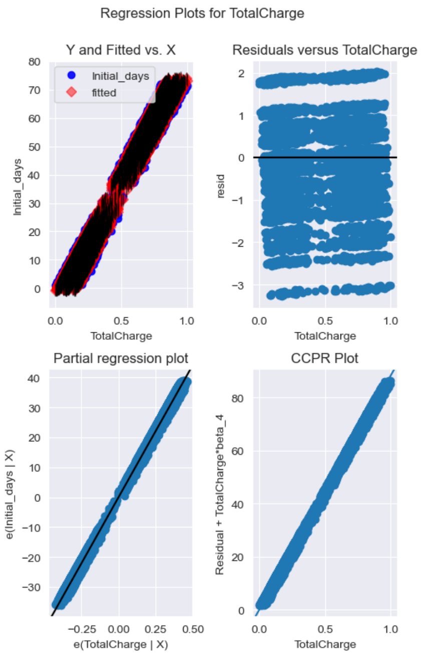
E2: Output and Calculations of Analysis. In this section, there will be a discussion of the calculations and the output that resulted from the reduction of features that yielded to reduced multiple linear regression model.

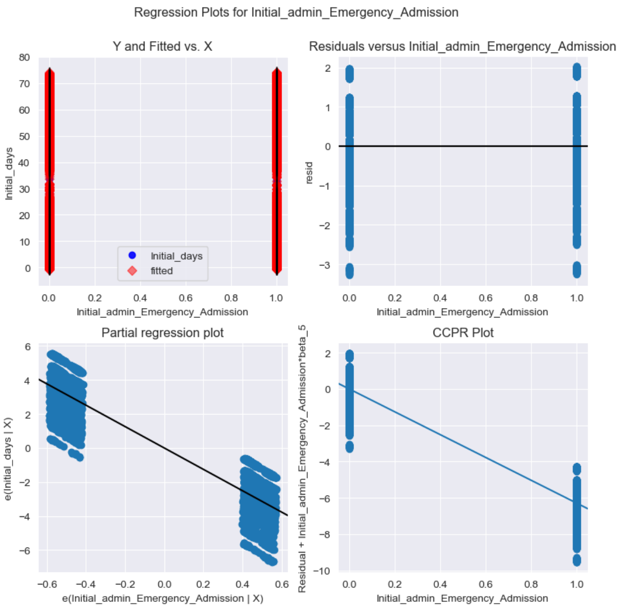
The residual plots are shown on the top right of the included graphs for each of the model features. For a better view of the plots please see the Jupyter Notebook.

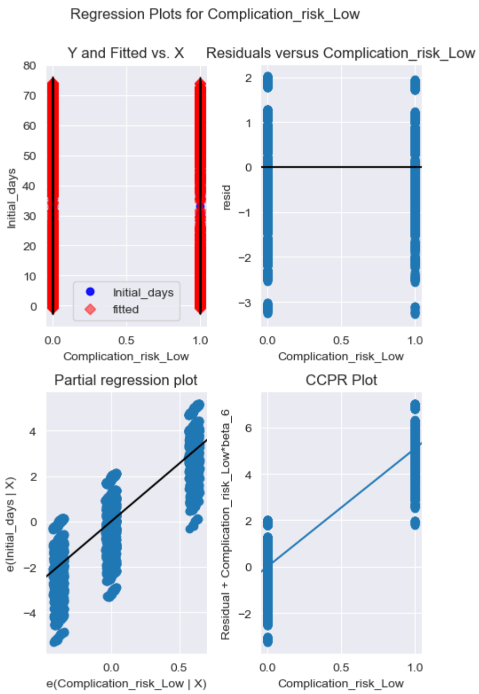

Regression Plots for Arthritis
Y and Fitted vs. X
Initial_days
fitted
Arthritis
Residuals versus Arthritis
resid
Partial regression plot
e(Initial_days | X)
e(Arthritis | X)
CCPR Plot
Residual + Arthritis*beta_1

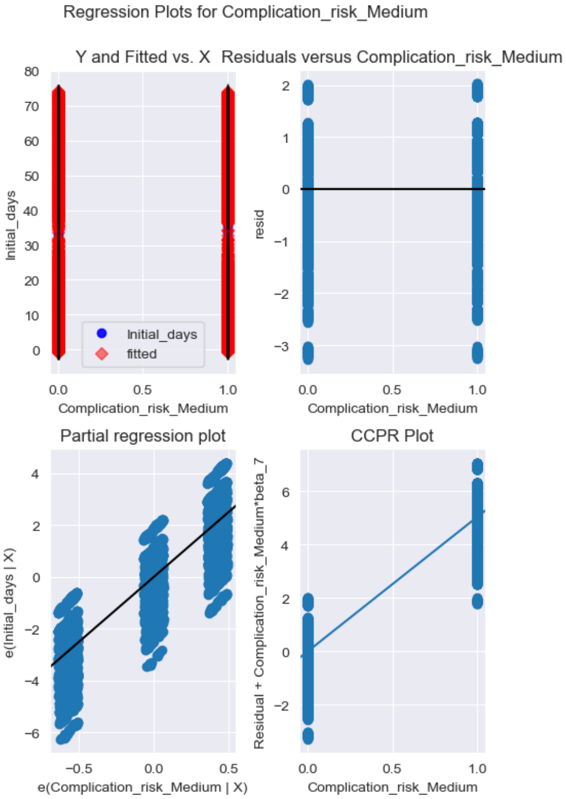

Regression Plots for Diabetes
Y and Fitted vs. X
Initial_days
fitted
Diabetes
Residuals versus Diabetes
resid
Partial regression plot
e(Initial_days | X)
e(Diabetes | X)
CCPR Plot
Residual + Diabetes*beta_2


Regression Plots for BackPain
Y and Fitted vs. X
Initial_days
fitted
BackPain
Residuals versus BackPain
resid
Partial regression plot
e(Initial_days | X)
e(BackPain | X)
CCPR Plot
Residual + BackPain*beta_3


Regression Plots for TotalCharge
Y and Fitted vs. X
Initial_days
fitted
TotalCharge
Residuals versus TotalCharge
resid
Partial regression plot
e(Initial_days | X)
e(TotalCharge | X)
CCPR Plot
Residual + TotalCharge*beta_4


Regression Plots for Initial_admin_Emergency_Admission
Y and Fitted vs. X
Initial_days
Initial_days
fitted
Initial_admin_Emergency_Admission
Residuals versus Initial_admin_Emergency_Admission
resid
Initial_admin_Emergency_Admission
Partial regression plot
e(Initial_days | X)
e(Initial_admin_Emergency_Admission | X)
CCPR Plot
Residual + Initial_admin_Emergency_Admission*beta_5
Initial_admin_Emergency_Admission


Regression Plots for Complication_risk_Low
Y and Fitted vs. X
Initial_days
Initial_days
fitted
Complication_risk_Low
Residuals versus Complication_risk_Low
resid
Complication_risk_Low
Partial regression plot
e(Initial_days | X)
e(Complication_risk_Low | X)
CCPR Plot
Residual + Complication_risk_Low*beta_6
Complication_risk_Low


Regression Plots for Complication_risk_Medium
Y and Fitted vs. X
Initial_days
Initial_days
fitted
Complication_risk_Medium
Residuals versus Complication_risk_Medium
resid
Complication_risk_Medium
Partial regression plot
e(Initial_days | X)
e(Complication_risk_Medium | X)
CCPR Plot
Residual + Complication_risk_Medium*beta_7
Complication_risk_Medium

The model had a RSE of ~1.11. The calculations for this can be found in the section Model's Residual Standard Error (located near the bottom of the Notebook.)

E3: Code. The code that was used to create the multiple linear regression model can be found in the accompanying Jupyter Notebook. The sections in the Notebook follow the layout of this paper. The file name is:

Heino D208 Predictive Modeling Task 1.ipynb

The sections that have code that is relevant to the creation of the multiple linear regression model can be found in Sections D and E.

Part IV: Summary and Implications

This section will provide a summary of the findings as well as their implications. There will be a discussion of the regression of the equation. There will be a section on the recommended course of action based on the results of the model.

F1: Results of Analysis.

The creation of the model yielded an equation of the form:

$$\hat{y} = -5.8076 - 0.8722_{\mathit{Arthritis}} - 0.9068_{\mathit{Diabetes}} - 1.0376_{\text{BackPain}} + 88.2154_{\text{TotalCharge}}$$

$$- 6.2913_{\text{Initial\_admin\_Emergency\_Admission}} + 5.0289_{\text{Complication\_risk\_Low}} + 4.9985_{\text{Complication\_risk\_Medium}}$$

Interpreting the coefficients for the predictor variables these conclusions can be stated:

- Keeping all things constant, patients with arthritis will spend .8722 of a day less in the hospital.
- Keeping all things constant, patients with diabetes will spend .9068 of a day less in the hospital.
- Keeping all things constant, patients with back pain will spend 1.0376 days less in the hospital.

- Keeping all things constant, a one-unit increase in total charge is associated with an 88 increase in days hospitalized.
- Keeping all things constant, patients admitted for an emergency will spend 6.29 days less in the hospital.
- Keeping all things constant, a patient with a low complication risk will spend 5.0289 days more in the hospital.
- Keeping all things constant, a patient with a medium complication risk will spend 5.0289 days more in the hospital.

The models do exhibit some favorable statistics. The probability of the F-statistic is calculated as being 0.00. This statistic assumes that the model is a perfect fit. While theoretically possible it will not occur with other more "realistic" data. This F-statistic value would be indicative of severe overfitting of the data (L, n.d.)[1]. This value is also less than the p-value of 0.05 which means that this is statistically significant. As stated earlier, this model will be overfitting for the data that had been supplied. Yet if this value is taken at face value it means that the regressions are meaningful (Yadav, 2021).

Limitations. Examining the data that was included and observed through analysis there are a few things that stand out. First, look at the histograms and the scatter plots that were present in section C. You can see that there are distinct groupings of the data.

There seems to be a gap in the days range from around 30 days to 40 days. This could result in possible problems with predicting values that fall within this range as there are almost no values in the range. The removal of the outliers did not affect as the 30 to 40 days range is within the bounds of the IQR and would have been included with the prepared data. This is one of the observations that were observed looking at the data depicted in the graphs. Also looking

[1] This was as close to a proper citation as possible with the given information

at the histograms there are a few that approach an almost normal distribution. This does not seem highly likely with data that will be gathered in more realistic settings.

Also excluding some columns may have had an impact on the performance of the model. For example not including the Services column may have affected how much the total charge is. If you undergo more expensive services then there is an increase in the average daily charge that is recorded in the TotalCharge column. This may have some bearing on the model as this relation is not identified in the model but could lead to problems with multicollinearity. If this is true then it could have led to the exclusion of total charge from the model.

In regards to data gathering, it might be beneficial for the inclusion of data that "fills" in the gaps that were discussed earlier. It does not seem that it is possible to have such grouping occurring naturally in the data.

As stated in a previous section the values for the F-statistic are not realistic. It is highly improbable that the model will have a 0.00 F-statistic. This value is because the model includes a lot of parameters to cover as much of the explained variance as possible. It might not be needed to ascertain this high of an R^2 score but it was achieved. It might be possible to remove some of the features and still get the same R^2 value, but this was not tried.

Aside: This may not be relevant now but may be relevant if the model is to be used with other concepts down the line. For example, other concepts that may be encountered. e.g., machine learning. The model identified features that seem to be contrary to what is expected. The steps were followed and do not see how to remedy this discrepancy.

F2: Recommended Course of Action. Based on the findings and analysis the following is the course of action that should be followed. The features that are regarded as significant are features where is no reasonable way to change the values in these features. Most of these features are outside the control of the hospital and any changes will not be possible. Features like diabetes, back pain, and arthritis are outside the scope of things the hospital can change or influence.

Complication risk and total charge are seen as features that increase the stay in the hospital. Looking at these variables there is nothing that the organization can do to change them. Although looking at the total charge it could be the only feature that is within the control of the organization. But it could also be observed that total charge is not mutually exclusive. The meaning here is that it is partially or dependent on other features that may have been excluded from the features that are discussed here.

In regards to data gathering, it might be beneficial for the inclusion of data that "fills" in the gaps that were discussed earlier. It does not seem that it is possible to have such grouping occurring naturally in the data.

Part VI: Demonstration

G. Panopto. The link to the Panopto and the demonstration is the following:

References

H. Web Sources

This assessment submission made use of the resources that were provided by WGU. The resources that are listed below were used to help create the code for various sections of the assessment.

Creating multiple subplots using plt.subplots — Matplotlib 3.8.2 documentation. (n.d.). https://matplotlib.org/stable/gallery/subplots_axes_and_figures/subplots_demo.html

Ebner, J. (2022, March 29). How to use Pandas get dummies in Python - Sharp Sight. Sharp Sight. https://www.sharpsightlabs.com/blog/pandas-get-dummies/

GeeksforGeeks. (2020, October 24). *Reading specific columns of a CSV file using Pandas*. https://www.geeksforgeeks.org/reading-specific-columns-of-a-csv-file-using-pandas/

GeeksforGeeks. (2023a, January 10). Detecting Multicollinearity with VIF Python. https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

GeeksforGeeks. (2023b, March 20). *Pandas GroupBy Count the occurrences of each combination*. https://www.geeksforgeeks.org/pandas-groupby-count-the- occurrences-of-each-combination/

Marques, A. (2022, December 20). *How to show all columns and rows in a Pandas DataFrame*. Built In. https://builtin.com/data-science/pandas-show-all-columns

pandas.DataFrame.select_dtypes — pandas 2.1.3 documentation. (n.d.). Retrieved November 21, 2023, from https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.select_dtypes.html *seaborn.histplot — seaborn 0.13.0 documentation*. (n.d.). Retrieved November 21, 2023, from https://seaborn.pydata.org/generated/seaborn.histplot.html

Seaborn.violinplot() method. (n.d.). Retrieved November 21, 2023, from https://www.tutorialspoint.com/seaborn/seaborn_violinplot_method.htm

statsmodels.regression.linear_model.OLS - statsmodels 0.15.0 (+73). (n.d.). https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

Yadav, J. (2021, December 11). Statistics: How Should I interpret results of OLS? - Jyoti Yadav - Medium. *Medium*. https://jyotiyadav99111.medium.com/statistics-how-should-i-interpret-results-of-ols-3bde1ebeec01

I. In-text Citations

The resources that are found below were used in the creation of the written document and were not used in the creation of the code that is found in the Jupyter Notebook.

Average children per family U.S. 2022 | Statista. (2023, June 2). Statista. Retrieved November 24, 2023, from https://www.statista.com/statistics/718084/average-number-of-own-children-per-family/#:~:text=U.S.%20average%20number%20of%20own,with%20own%20children%201960%2D2022&text=The%20typical%20American%20picture%20of,family%20in%20the%20United%20States.

Bevans, R. (2023, June 22). *Understanding P values | Definition and Examples*. Scribbr. Retrieved November 24, 2023, from https://www.scribbr.com/statistics/p-value/#:~:text=The%20p%20value%20is%20a,to%20reject%20the%20null%20hypothesis.

Carron, J. (2021, December 13). *Violin Plots 101: Visualizing Distribution and Probability Density | Mode*. Retrieved November 24, 2023, from https://mode.com/blog/violin-plot-examples/

Check for multicollinearity of model terms — check_collinearity. (n.d.). Github. Retrieved November 24, 2023, from https://easystats.github.io/performance/reference/check_collinearity.html

L, D. (n.d.). *What does it mean if the F statistic is zero? | Wyzant Ask An Expert*. Wyzant Tutoring. https://www.wyzant.com/resources/answers/694316/what-does-it-mean-if-the-f-statistic-is-zero#:~:text=In%20very%20unusual%20circumstances%2C%20if,severe%20overfitting%20of%20the%20data.

Real median personal income in the United States. (2023b, September 12). Retrieved November 24, 2023, from https://fred.stlouisfed.org/series/MEPAINUSA672N

Scikit-Learn. (n.d.). *LinearRegression produced different R^2 compared to those from `ols` (statsmodels), `stats::lm` (R) and Excel · scikit-learn/scikit-learn · Discussion #21050*. GitHub. Retrieved November 24, 2023, from https://github.com/scikit-learn/scikit-learn/discussions/21050

Taylor, S. (n.d.). *Multiple Linear regression*. Corporate Finance Institute. Retrieved November 14, 2023, from https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/

Zach. (2020, January 8). *The Four Assumptions of Linear Regression*. Statology. Retrieved November 14, 2023, from https://www.statology.org/linear-regression-assumptions/