Part I

A.

1. The research question I am going to be answering is “What factors predict the length of a patient’s hospital stay?”
2. The goal of this analysis is to use information that is generally taken at check in and predict the length of the patient’s stay. This can be used in a practical setting for staffing and supply purposes.

Part II

B.

1. The first assumption of multiple regression is that the relationship between variables is linear. It assumes that the residuals are normally distributed and the independent variables are not highly correlated with each other. Another assumption is that the variance of error terms is normally distributed across all variables. The last assumption is that there are at least two independent variables (Assumptions).
2. I chose to do this project in Python because there are libraries such as SKLearn and Statsmodels that are designed to add functions that data analysts frequently use. I will be using standard scaler, OLS, and mean squared error to start.
3. Multiple regression is an appropriate method to help answer this question because we have one near- continuous dependent variable (Initial\_days) and several independent variables (gender, age, location, etc).

Part III: Data preparation

C.

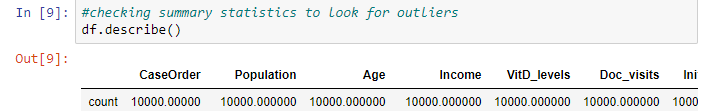
1. My first goal for data preparation is to get rid of the columns I will not be using in my analysis. This allows me to better see the data and get the scope of what I am working with. The next thing I want to do is eliminate duplicate entries. I will do this by looking for duplicate interaction ID’s. Next is checking and changing the data types if need be. After this I will remove outliers for each column and get dummies for categorical variables. I also created some univariate and bivariate distributions to help visualize my dataset and target variable. Finally, I write the prepared data to a csv.
2. The target dependent variable I am focusing on is Initial\_days (length for each hospital stay). The independent variables I will look at for this analysis are the following:

* Soft\_drink\_Yes (Does this person drink three or more sodas each week?)
* Stroke\_Yes (Has this person ever had a stroke?)
* Overweight\_Yes (Is this person overweight?)
* Diabetes\_Yes (Does this person have diabetes?)
* Asthma\_Yes (Does this person have asthma?)
* Age
* TotalCharge (Total charged for this stay)
* VitD\_levels
* HighBlood\_Yes (Does this person have high blood pressure?)
* Arthritis\_Yes (Does this person have arthritis?)
* Lat
* Lng
* Gender
* Doc\_visits
* Complication\_risk

The summary statistics I will need to gather for the analysis are the following:

* The mean for each variable
* The first and third quartiles for each variable
* The interquartile range (IQR) for each variable
* The minimum and maximum for each variable (to remove outliers)
* The residual error

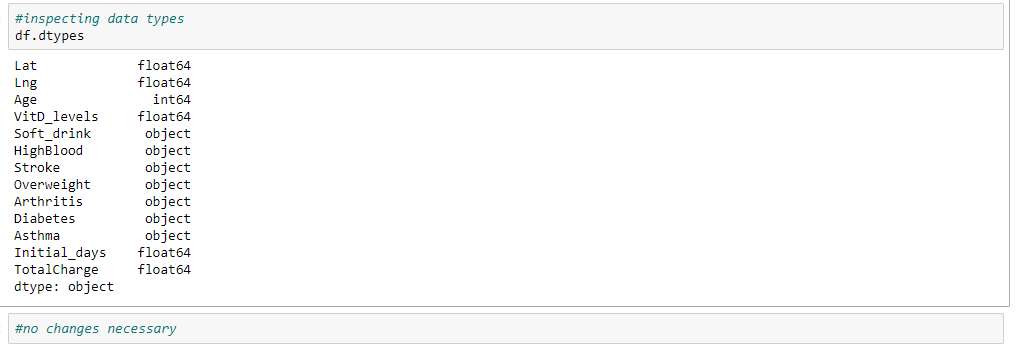
1. The first step I took during preparation was a .describe() statement.



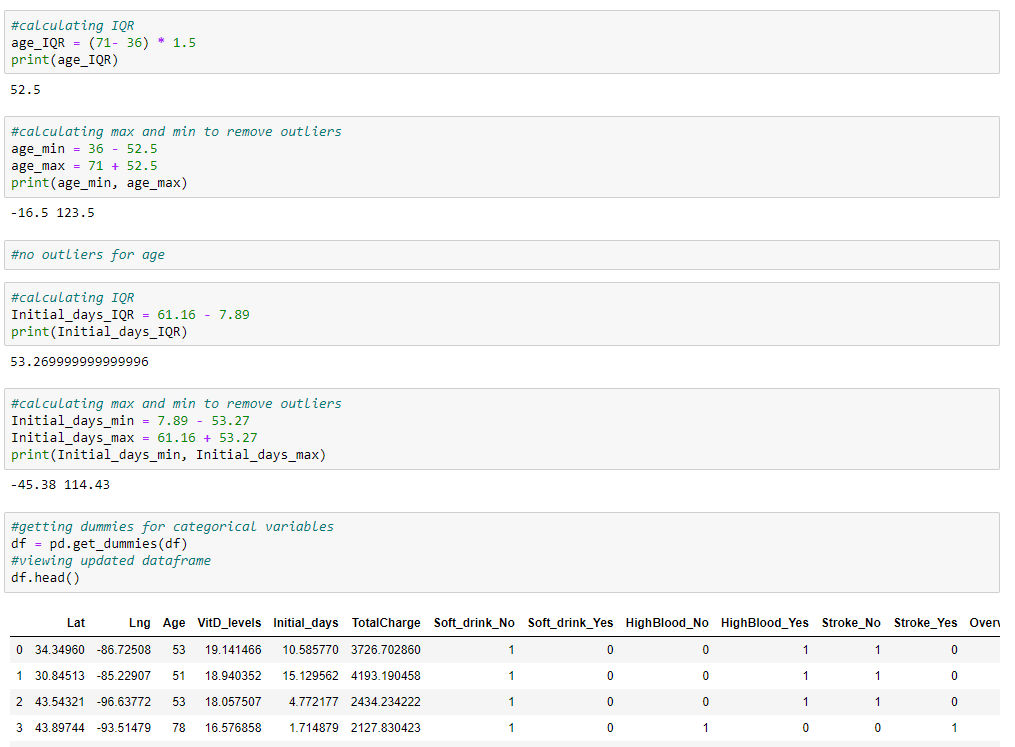
The counts for each column are all 10,000 which implies no missing values. Next, I dropped the columns I will not be using in this analysis.



Then I inspected data types. They all looked appropriate so no changes are necessary.

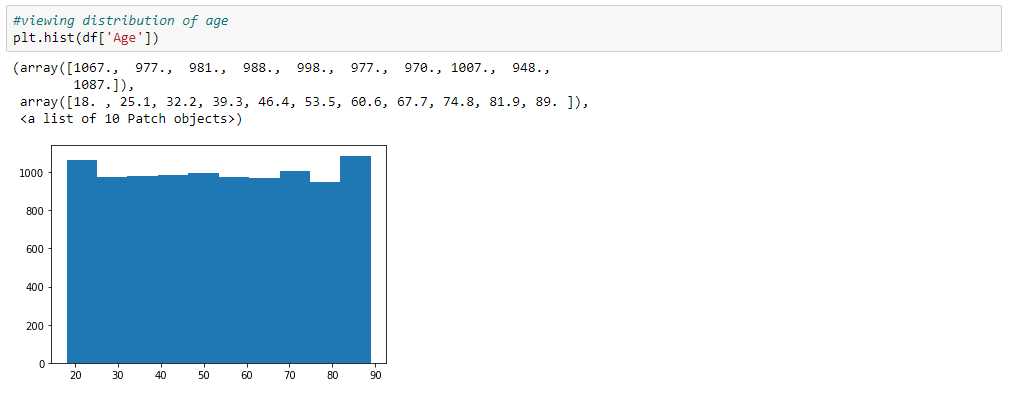


Then I used summary statistics to look for outliers. To do this, I first calculated the interquartile range for each column, which is (the 75th percentile - the 25th percentile) \* 1.5. Next I calculated the maximum value by adding the IQR to the 75th percentile and the minimum value by subtracting the IQR from the 25th percentile (Khan Academy). Next, I wrote statements to exclude the values for each category that are outliers. Afterwards, I check the length of the dataframe to make sure my changes were applied. Finally, I get dummies for the string variables and then view my updated dataframe.

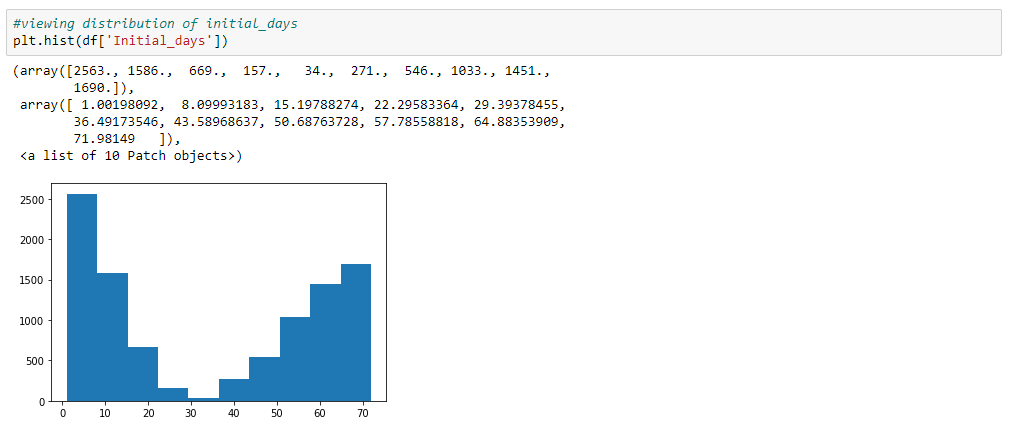


4. The next step in data preparation was to generate univariate and bivariate visualizations of my variables. Those are pictured below.

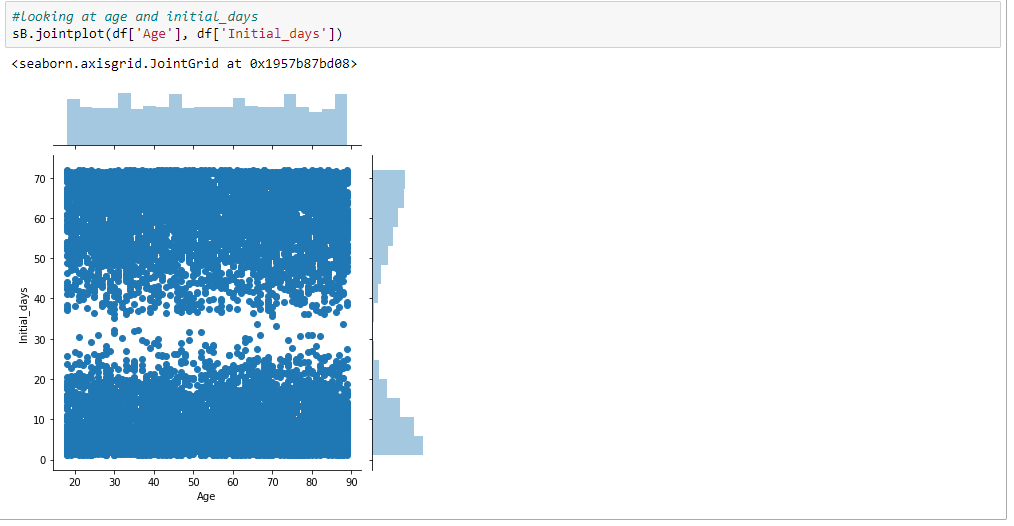
First, I viewed the distribution of ages.



Next I viewed the distribution of initial\_days. This was very interesting to me because there was a clear trend. The pattern is a U-shape, indicating people typically stay less than 20 days or more than 40 .



Then I viewed the relationship between age and initial\_days with a joint plot.



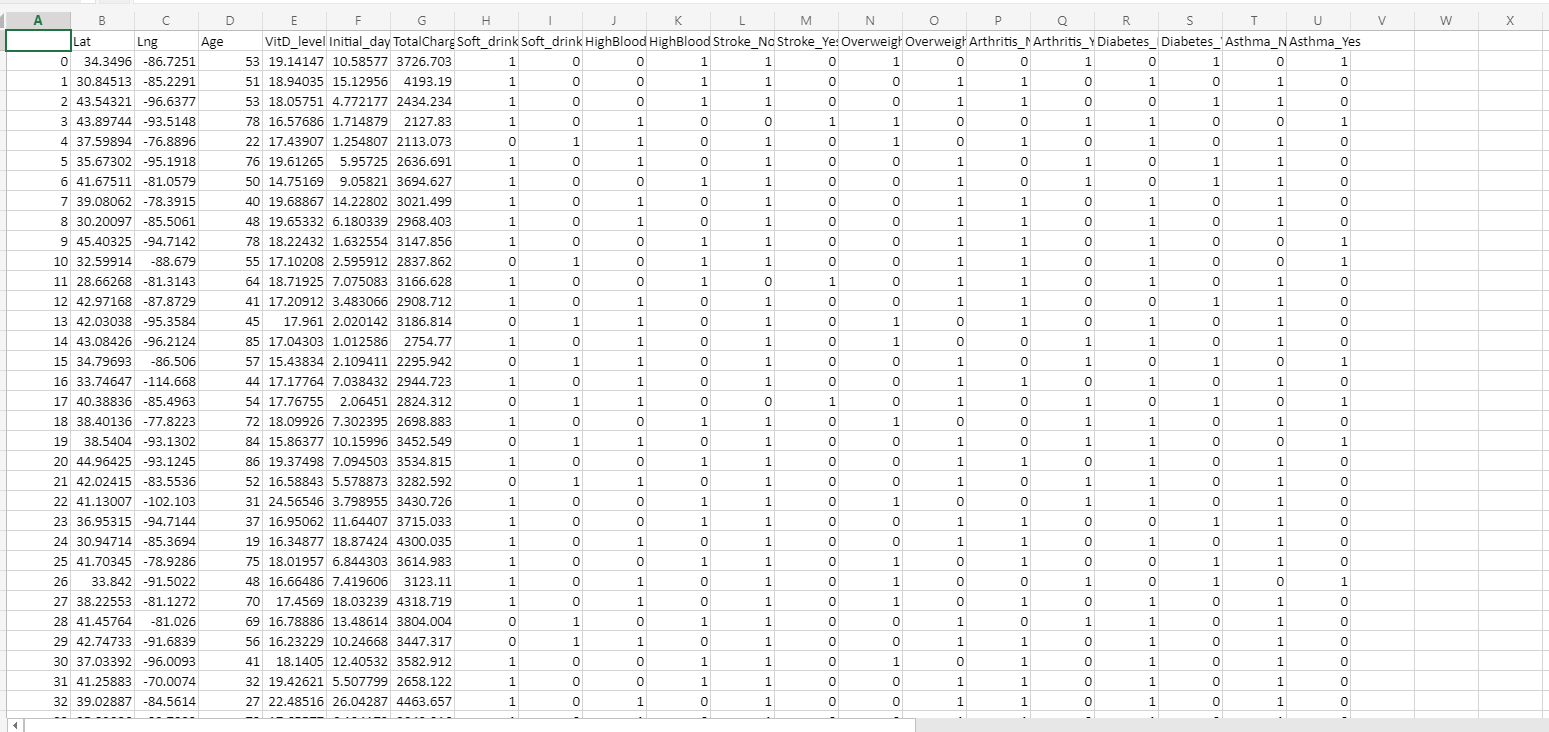
The last visualization is a strip plot to look at HighBlood\_Yes and Initial\_Days. There is no real discernable pattern here which is surprising to me.



Finally, I wrote the prepared data set to a new csv.

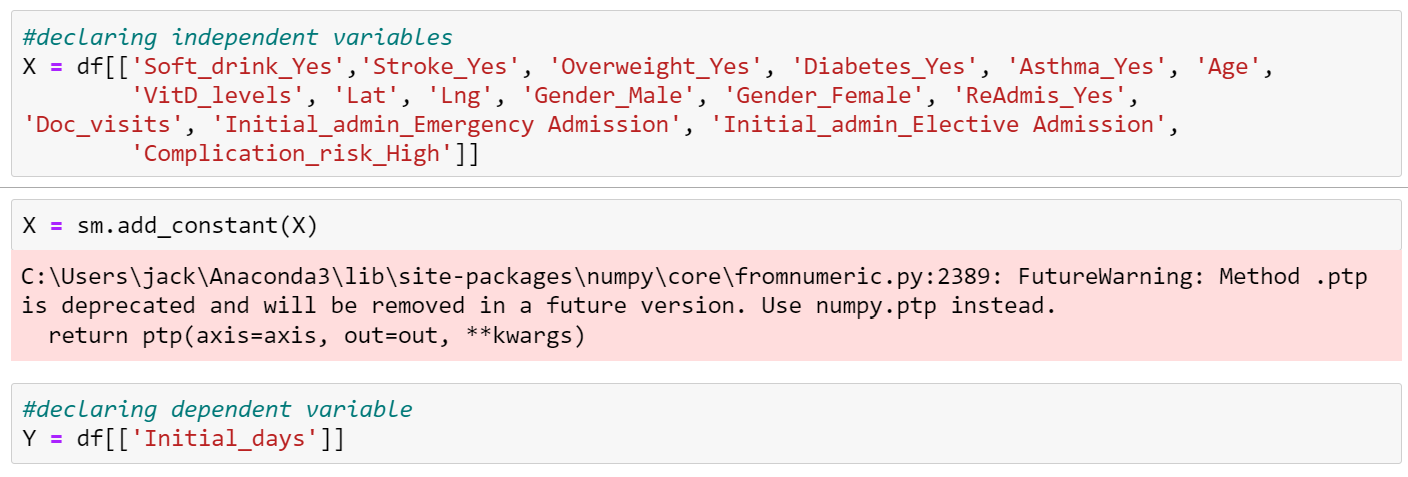


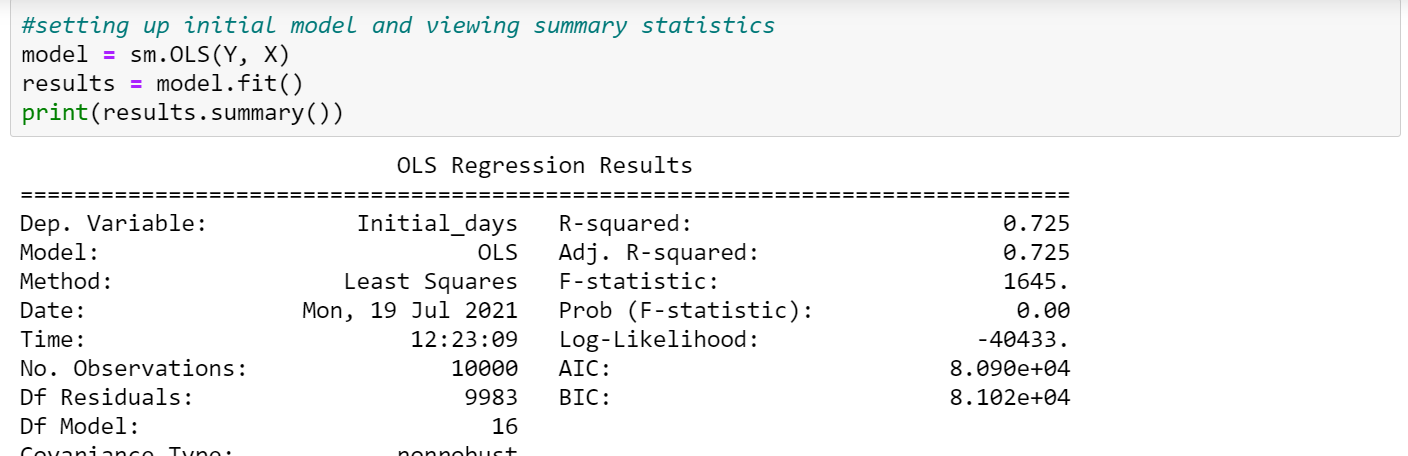
Below is a snippet of the new data set included on the file.



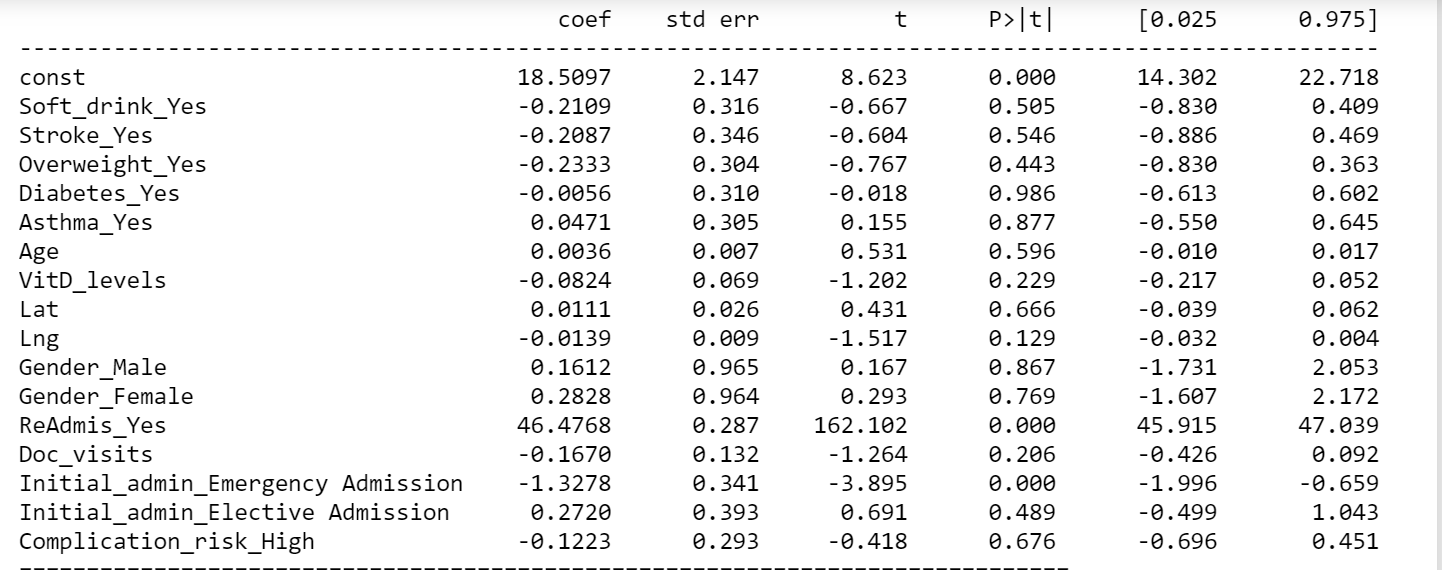
Part IV: Model Comparison and Analysis

D.1. Below is the initial model created with the variables named above.To begin I defined my variables and then added a constant to find the intercept for the model.

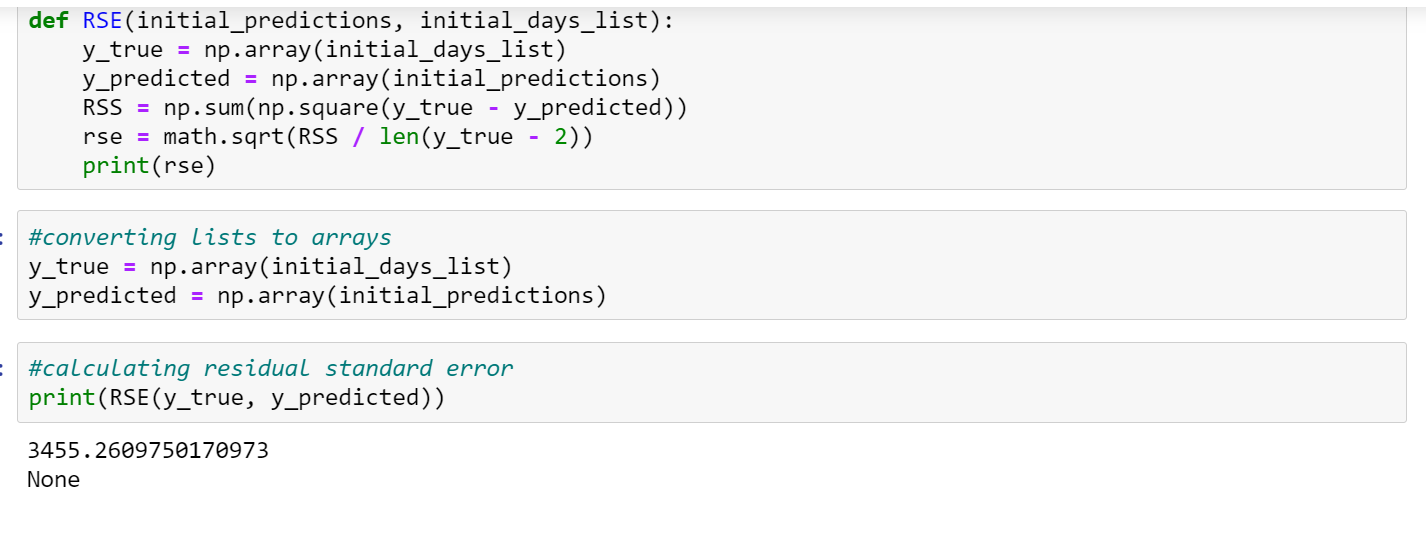




Next I viewed summary statistics for the model. One of my evaluation metrics is r-squared. My result is 0.725 which indicates a semi-efficient model. The coefficients for the model are listed below as well, with the equation for the line beneath the chart.



Besides R Squared, I used residual standard error as another model evaluation metric. To be able to calculate this, I created lists with the expected and actual values and a program that calculates the residual standard error (From Data Preprocessing). For my initial model I got a residual standard error of 3455.26.

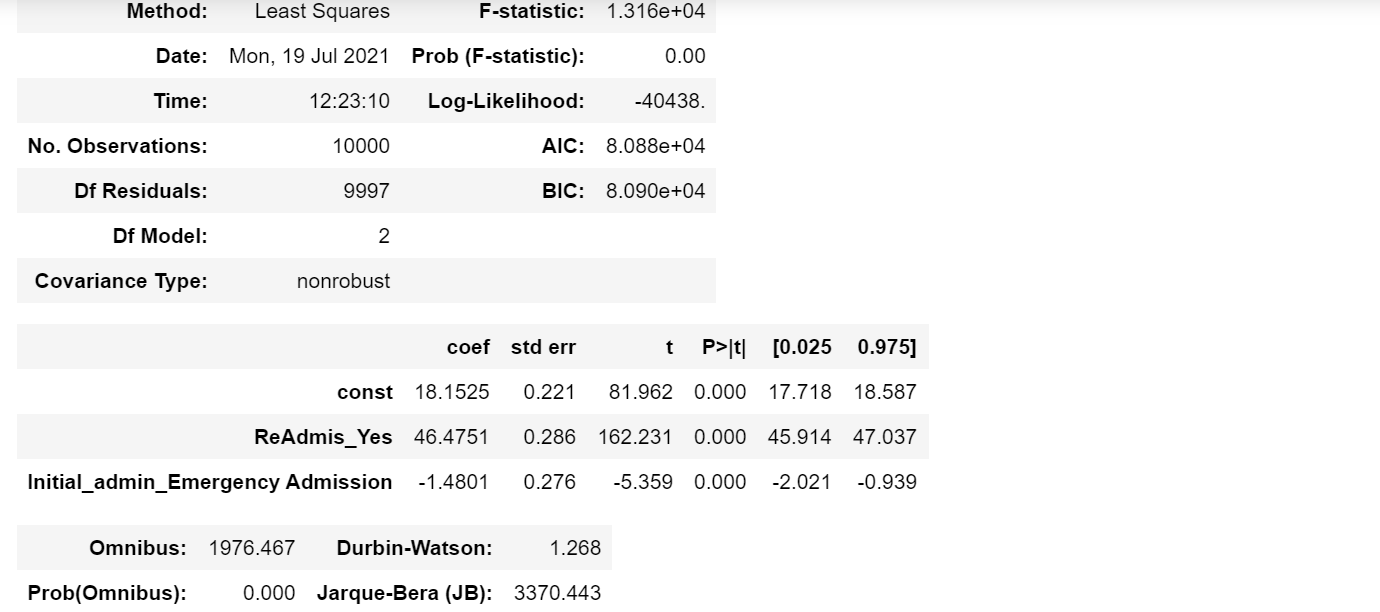


2. The indicator I used for model reduction is p-value. I began with the variables that I thought might have a significant impact on the model. The significance level I chose was .05, so anything with a p-value lower than that was left in. I ended up with ReAdmis\_Yes and Initial\_admin\_Emergency Admission as my independent variables.



3. My reduced model looks at initial\_days as a continuous dependent variable. My categorical independent variables are ReAdmis\_Yes and Initial\_admin\_Emergency Admission.





My equation for the line of best fit is as follows:

Initial\_days = (46.48\*ReAdmis\_yes) + (-1.48\*Initial\_admin\_Emergency Admission) + 18.15

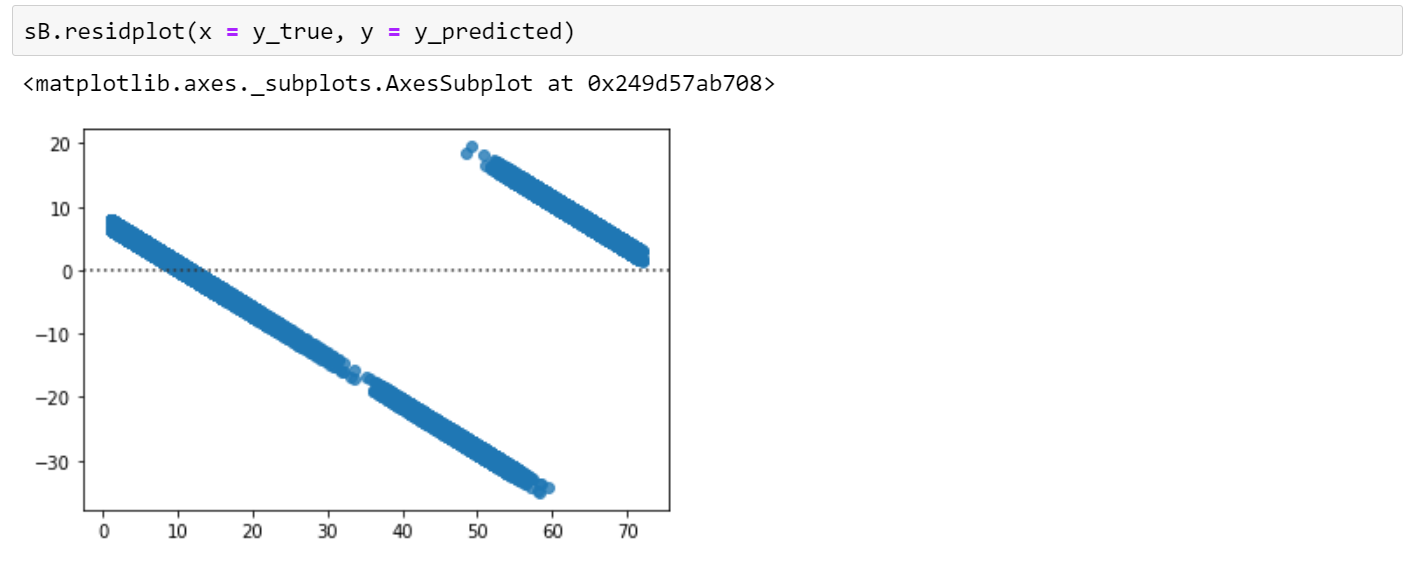
Next I used my model to make predictions and calculate the residual standard error. The RSE for the reduced model is 13.80, compared to my initial model that had a RSE of 3455.26. So the reduced model is quite an improvement in accuracy over the initial.



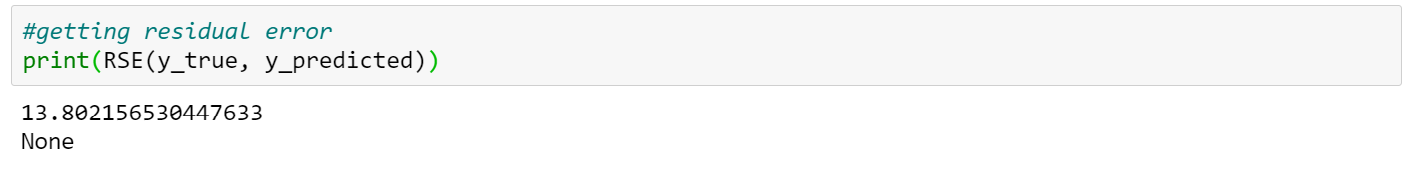
E1. For my original model I selected variables I suspected may be a predictor variable for length of stay. Columns such as interaction ID and city were dropped for a cleaner analysis after using them to check for duplicates. I began with Soft\_drink\_Yes, Stroke\_Yes, Overweight\_Yes, Diabetes\_Yes, Asthma\_Yes, Age, VitD\_levels, Doc\_visits, Gender, ReAdmis\_Yes, Complication\_risk\_high, Initial\_admin\_Emergency Admission, Lat, and Lng. After creating my initial model, I reduced it by selecting only variables with a p-value of 0.05 or less. This left me with ReAdmis\_Yes and Initial\_admin\_Emergency Admission. For my evaluation metric, I calculated the RSE as shown above and got 13.80 for the reduced model compared to 3455.26 for the original model.



Below is my residual plot for the reduced model.

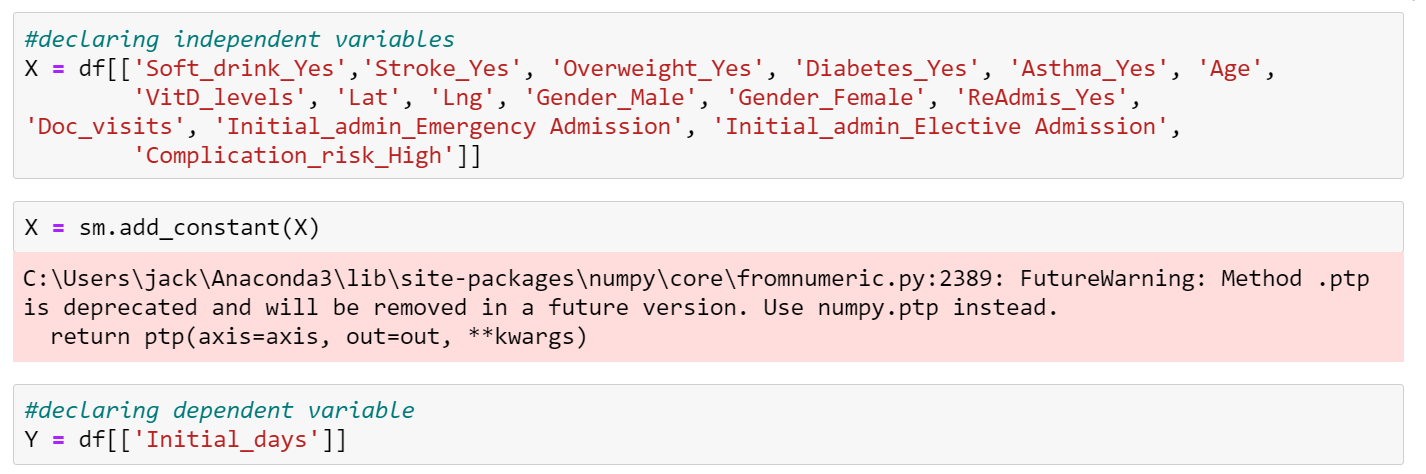


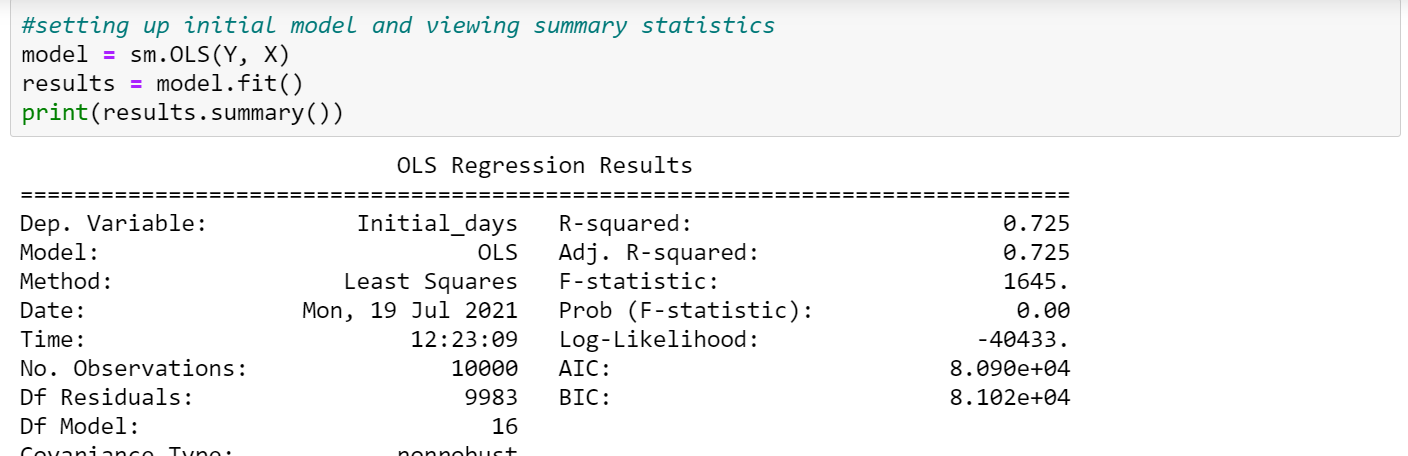
2. Below I calculated the residual standard error for the reduced model.

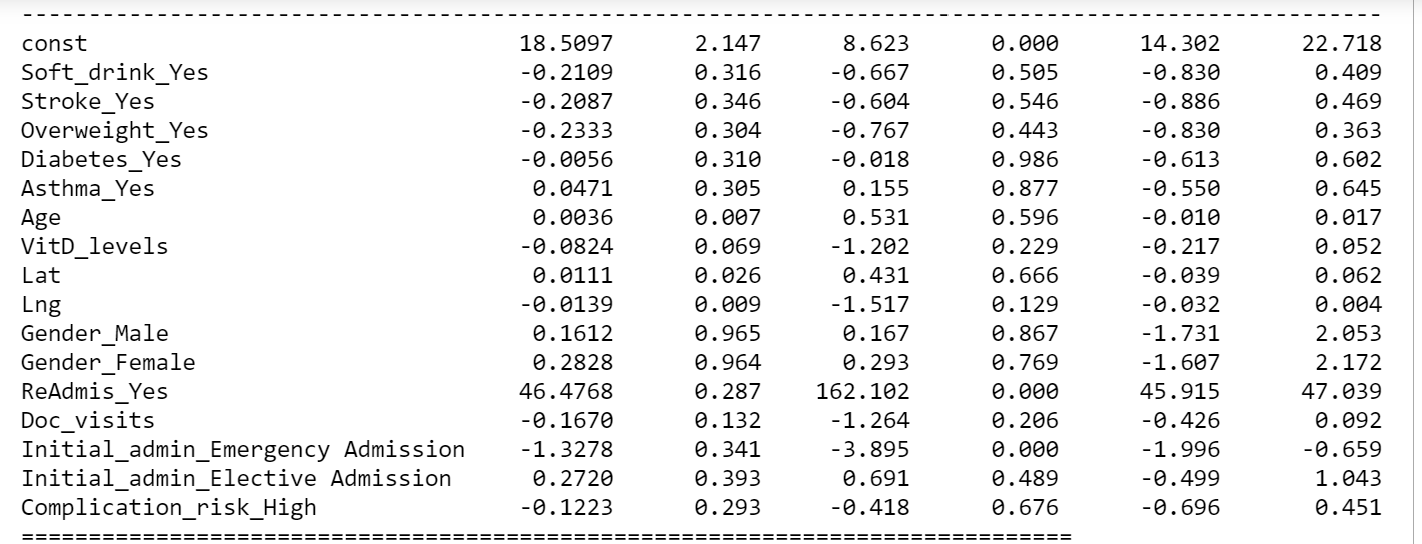


3. Below you will find the code used to create the original and then the reduced model. `

Original:

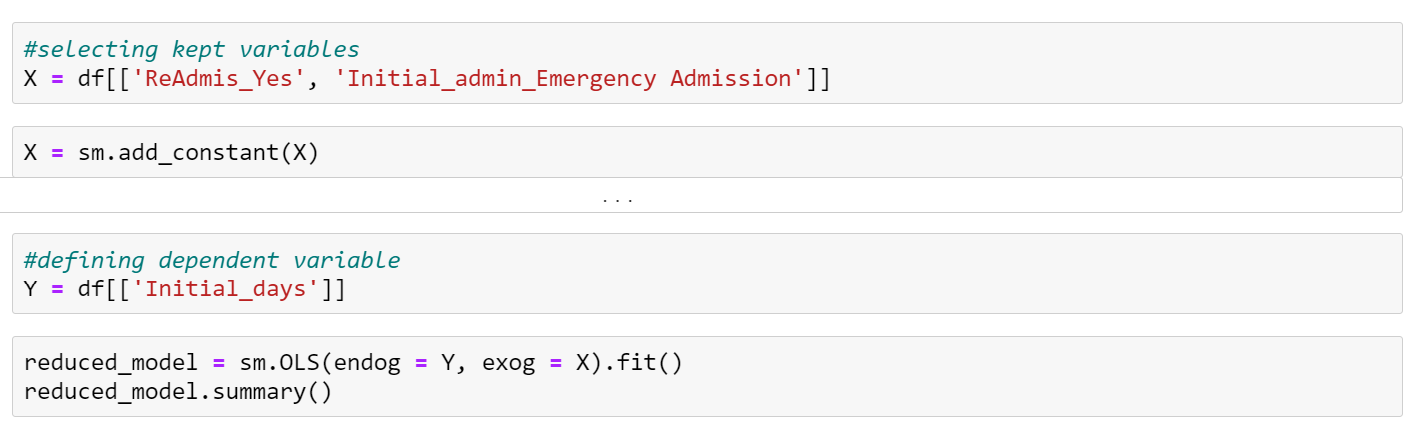


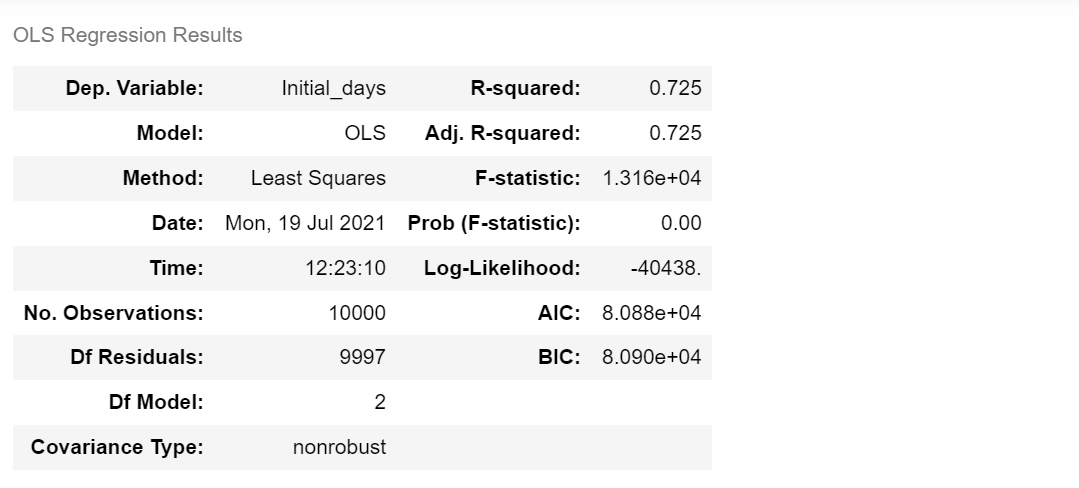






Reduced:









Part V: Data Summary and Implications

F. 1. The result of my analysis is that admission type and readmission status can be used to predict the length of a patient’s stay. This is impractical in real life unless ReAdmis\_Yes also represented having a history of being readmitted. My regression equation for the reduced model is as follows:

Initial\_days = (46.48\*ReAdmis\_yes) + (-1.48\*Initial\_admin\_Emergency Admission) + 18.15

The positive coefficient of 46.48 for ReAdmis\_yes indicates that patients that were readmitted tend to have a longer hospital stay than those that do not. The negative coefficient for Initial\_admin\_Emergency Admission suggests that Emergency admissions tend to have shorter stays than those admitted in other circumstances. The practical application for this model is impossible because you can’t predict how long a person will stay if one of the variables comes after the stay (ReAdmis\_Yes). One limitation here is the completeness of medical history. Having more health data could increase the accuracy of the model.

2. I recommend collecting more detailed medical data. Some things that could potentially be variables are

Heart\_condition - whether the individual has a heart condition

Cancer - whether the individual has cancer or has had cancer

Thyroid - whether or not the individual has a thyroid condition

Autoimmune\_disease - whether or not the individual has an autoimmune disease

Other\_major\_medical - whether or not the individual has any other major medical condition

H. Third party code referenced

*Bronshtein, A. (2019, December 27). Simple and Multiple Linear Regression in Python. Medium. https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9.*

*From Data Pre-processing to Optimizing a Regression Model Performance*. KDnuggets. (n.d.). https://www.kdnuggets.com/2019/07/data-pre-processing-optimizing-regression-model-performance.html.

I. Works cited

Assumptions of Multiple Linear Regression. (2020, March 10). Retrieved May 12, 2021, from <https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/>

Khan Academy. (n.d.). *Identifying outliers with the 1.5xIQR rule* . Khan Academy. https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule#:~:text=A%20commonly%20used%20rule%20says,or%20below%20the%20first%20quartile.