Part I: Research question

A.1. The real world question I’m going to be answering with this analysis is what factors increase a person’s chance at being readmitted within 30 days following a hospital stay?

A.2. The objective of this analysis is to compare the data of those that were readmitted within a month of a hospital stay to those that were not and look for patterns among variables.

Part II: Method Justification

B.1. The first assumption of logistic regression is that the dependent variable is binary. Ordinal logistic regression requires the dependent variable to be ordinal. The observations are assumed to be independent of each other, and there must be little to no multicollinearity among the independent variables. Logistic regression requires a large sample size. The last assumption of logistic regression is linearity of independent variables and log odds.

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.

3. Logistic regression is an appropriate technique for answering this question because the dependent variable is binary (readmitted yes or no) with multiple independent variables.

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. My last goal for data preparation is to search for null and missing entries.

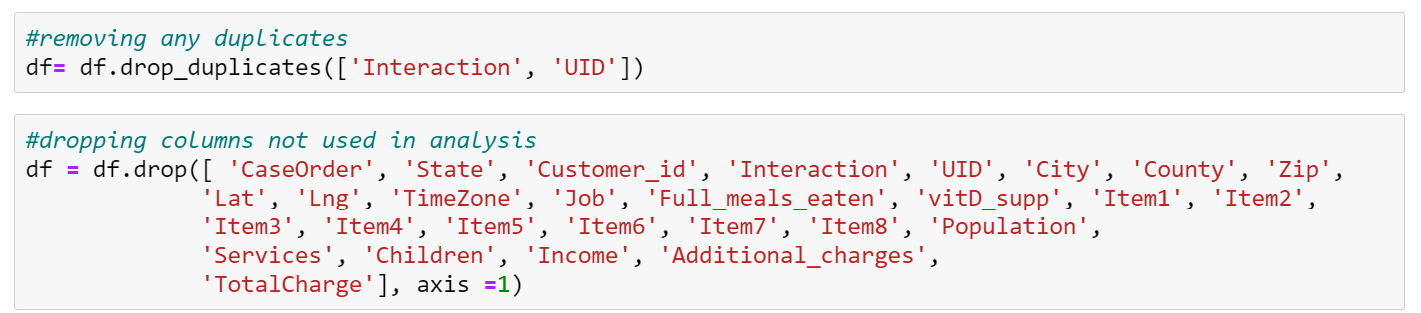
2. The target variable is ReAdmis\_Yes (whether or not a patient was readmitted within 30 days of release). The predictor variables I will be examining are:

* Area
* Population
* Age
* Income
* Gender
* VitD\_levels
* Doc\_visits (The number of doctor’s visits throughout the stay)
* Soft\_drink (Does this person drink 3 or more sodas a day?)
* Initial\_admin (What type was this admission?)
* HighBlood (Do they have high blood pressure?)
* Stroke (Have they had a stroke?)
* Complication\_risk
* Overweight
* Arthritis
* Diabetes
* Hyperlipidemia
* BackPain
* Anxiety
* Allergic rhinitis
* Reflux\_esophagitis
* Asthma
* Services
* Initial\_days
* Complication\_risk
* Children
* Marital

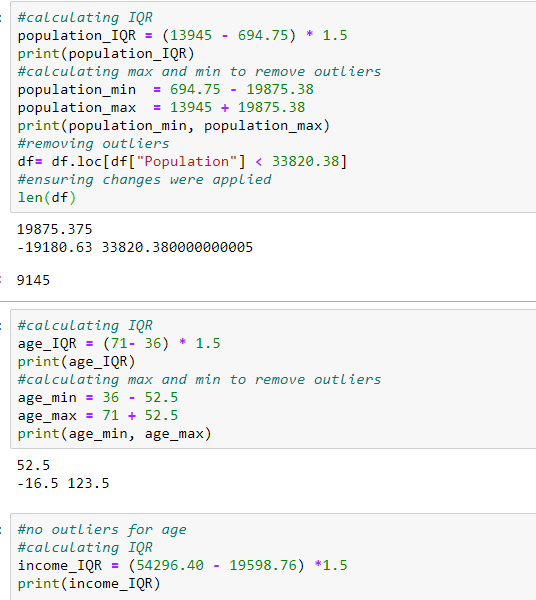
The summary statistics I will need to gather for the analysis are the following:

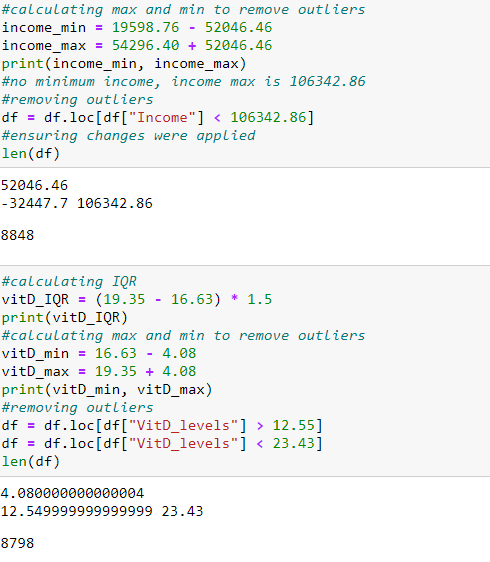
* The correlation amongst numeral variables
* The mean for each variable
* The first and third quartiles for each variable
* The interquartile range (IQR) for each variable
* The minimum and maximum for each variable (to remove outliers)
* The residual error

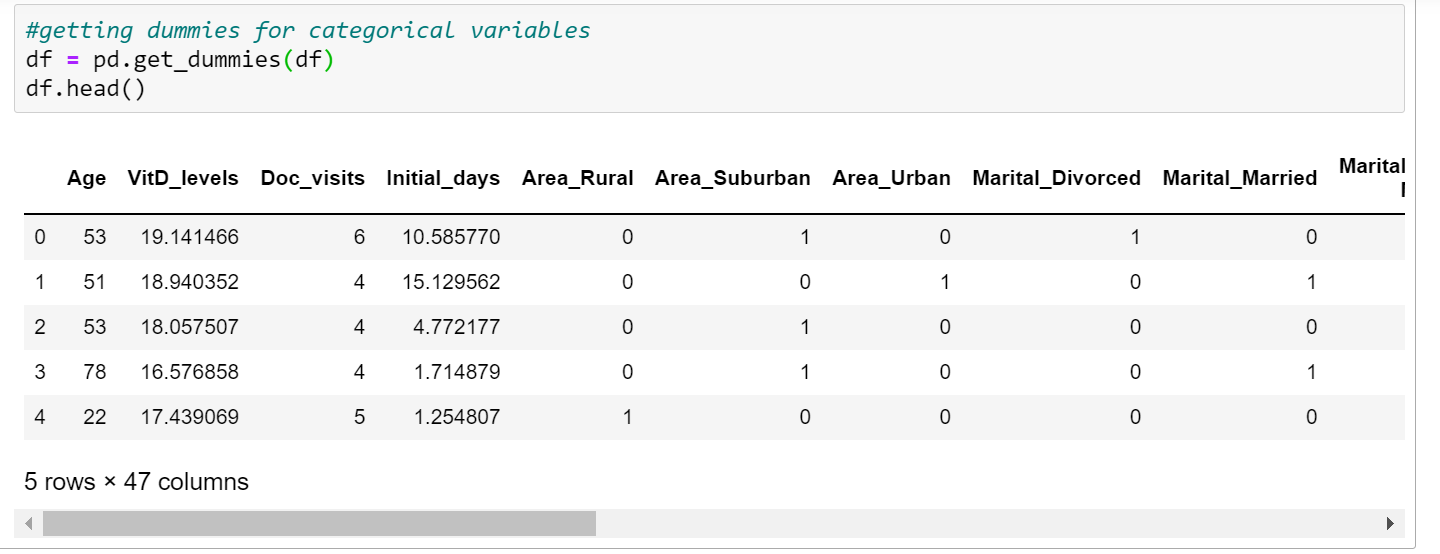
3. The first step I took for data preparation was to drop duplicates by looking for duplicate ID numbers. Next I dropped the columns I will not be using in the analysis.



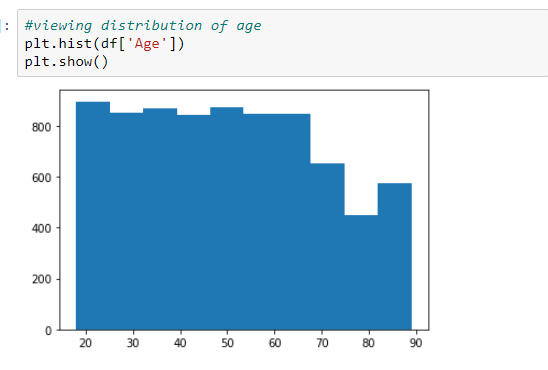
After that I calculated the interquartile range for each variable to look for and remove outliers (Khan Academy).

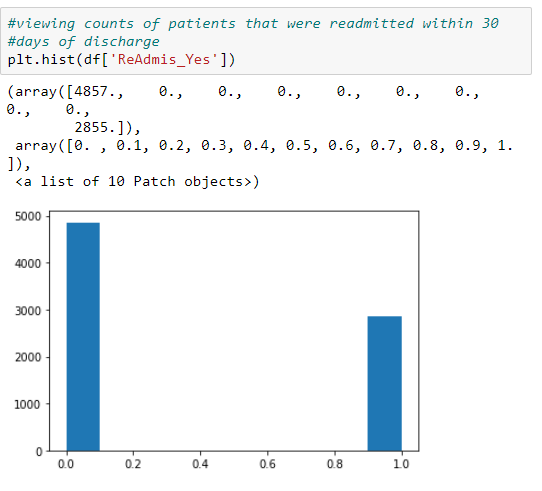
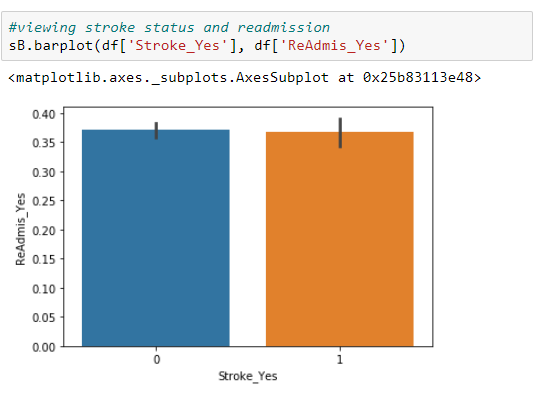


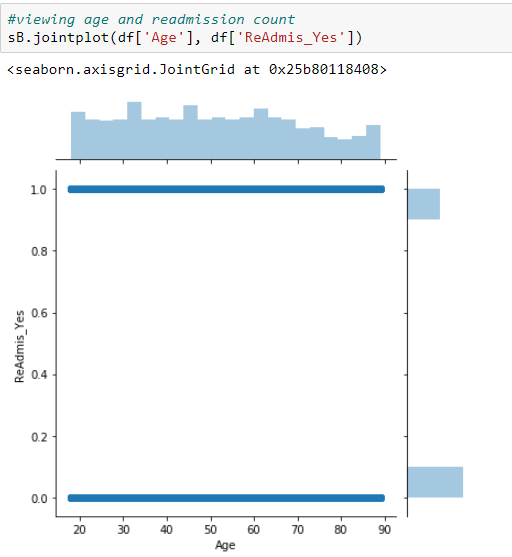
 4. After removing outliers, I got dummies for the categorical variables.

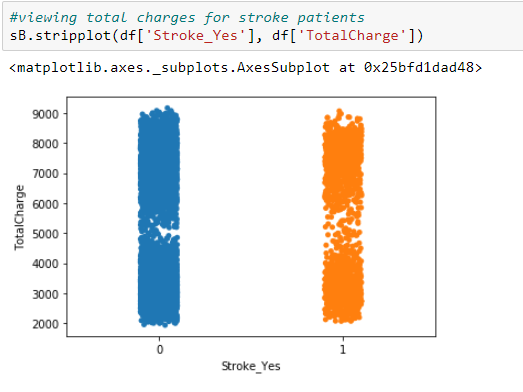


Next, I generated univariate and bivariate visualizations to help explore my data.

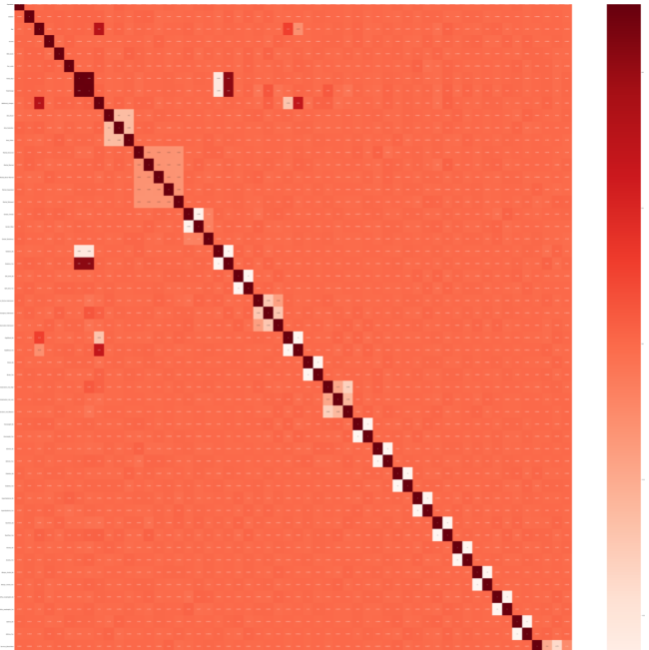




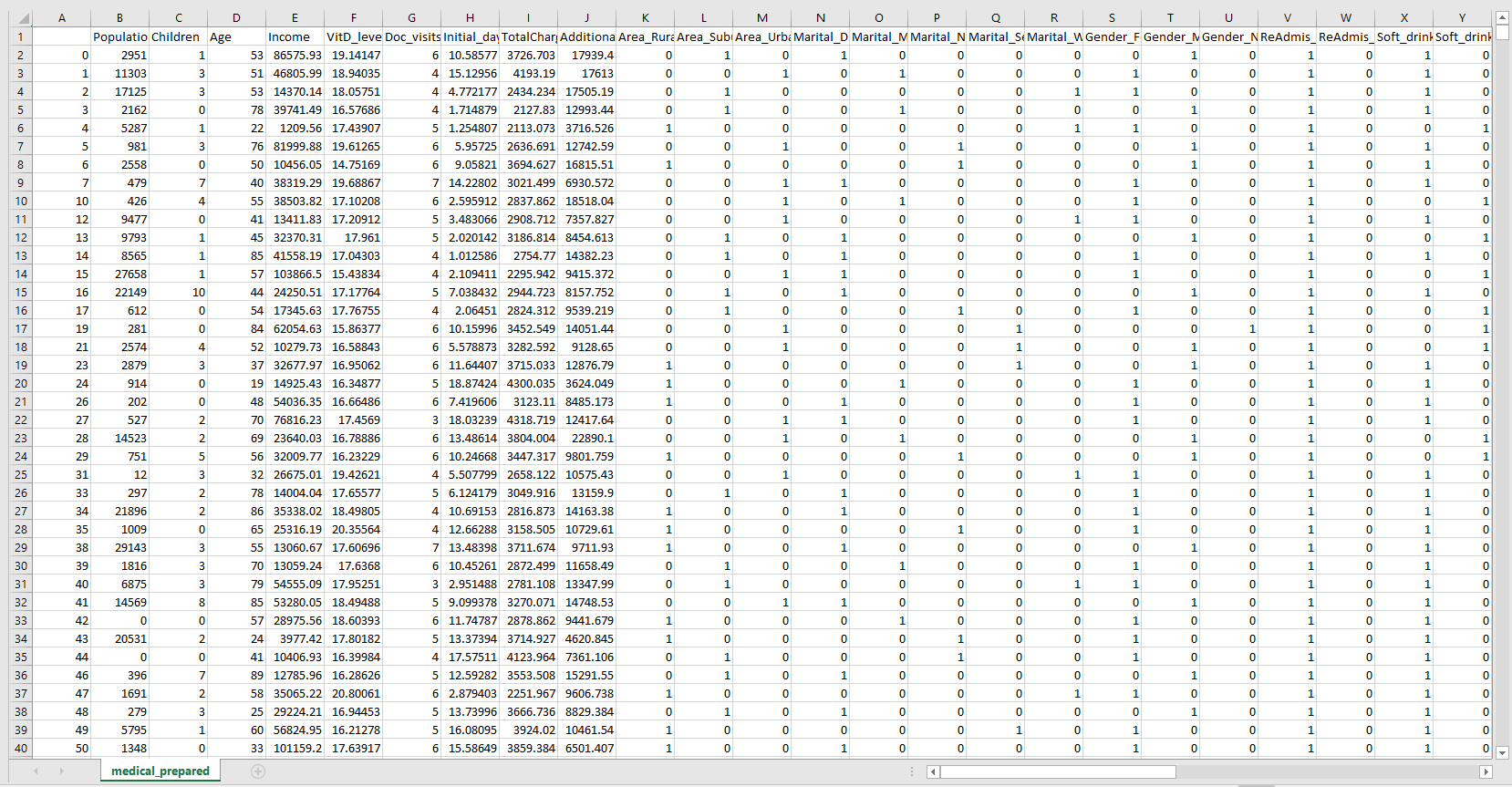




Finally, I visualized the correlations between variables.

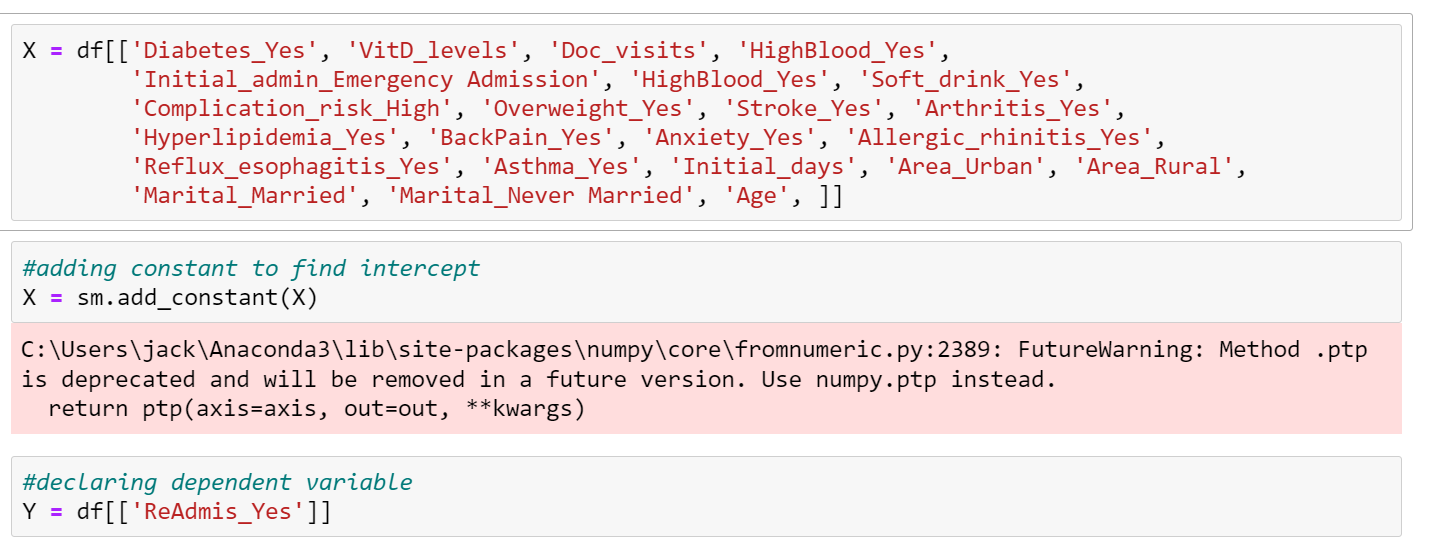


5. Below is a snippet of the cleaned dataset included in this submission.

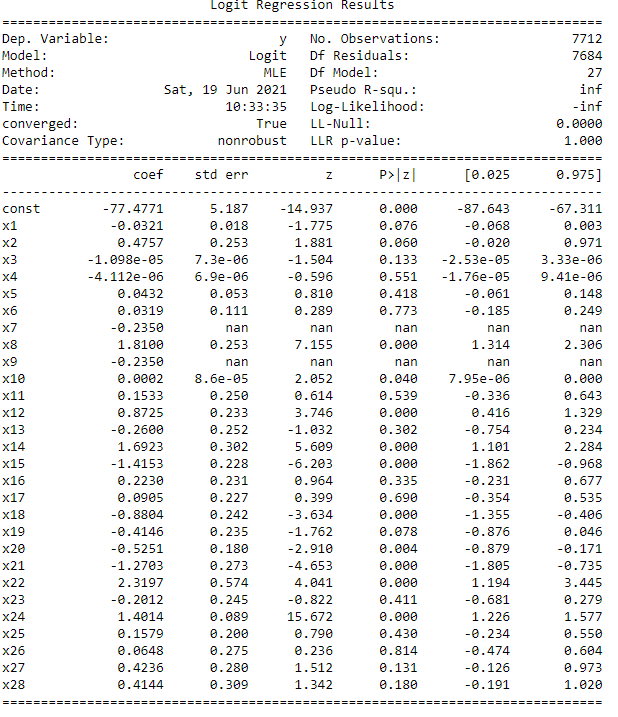


Part IV: Model comparison and analysis

D. 1. Here is the original model using all of the variables identified above (Add\_constant).

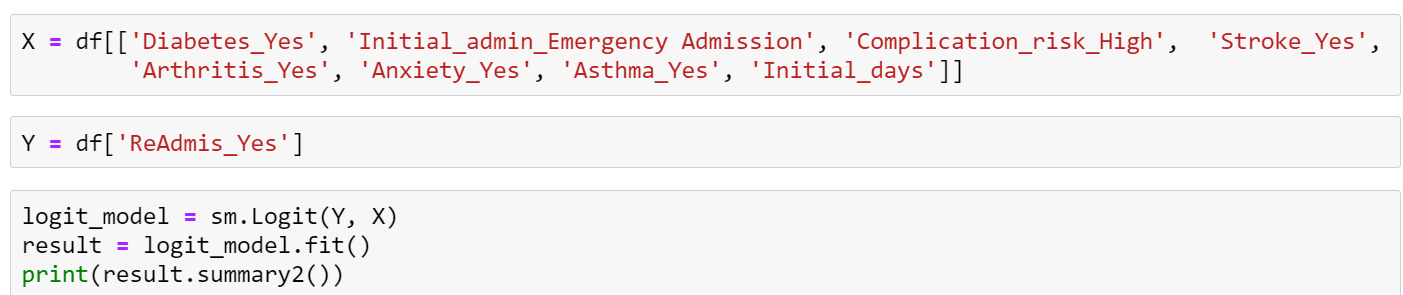


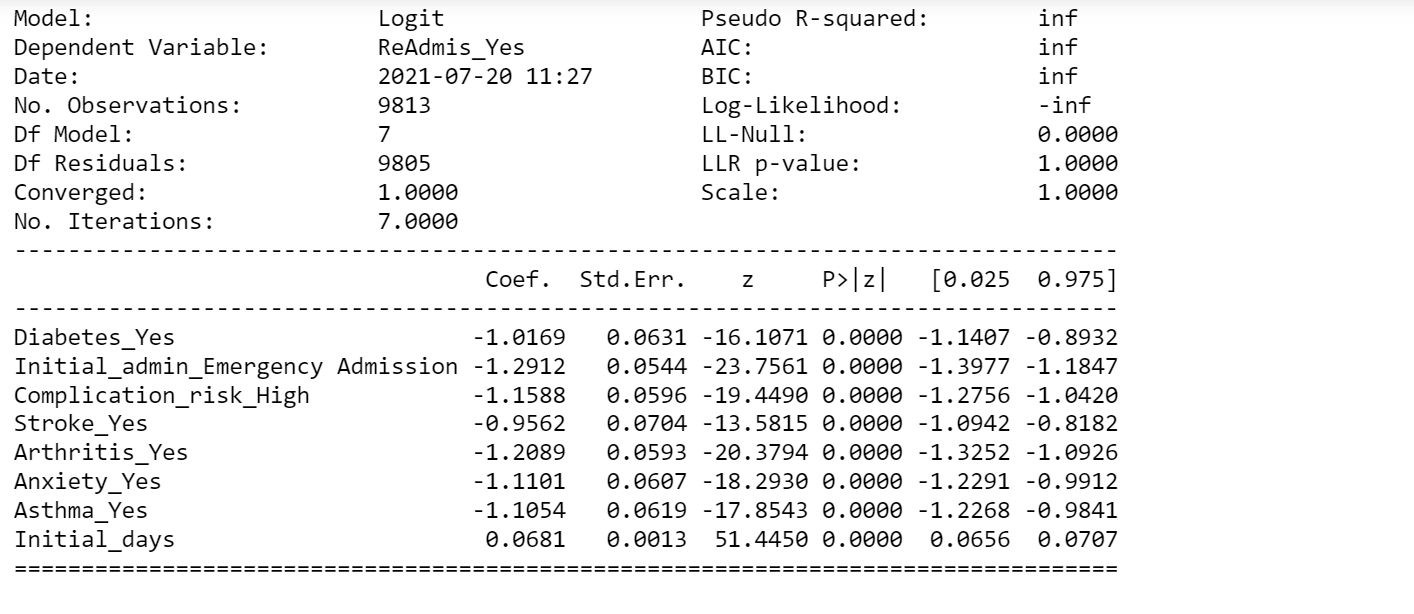
And the result summary:

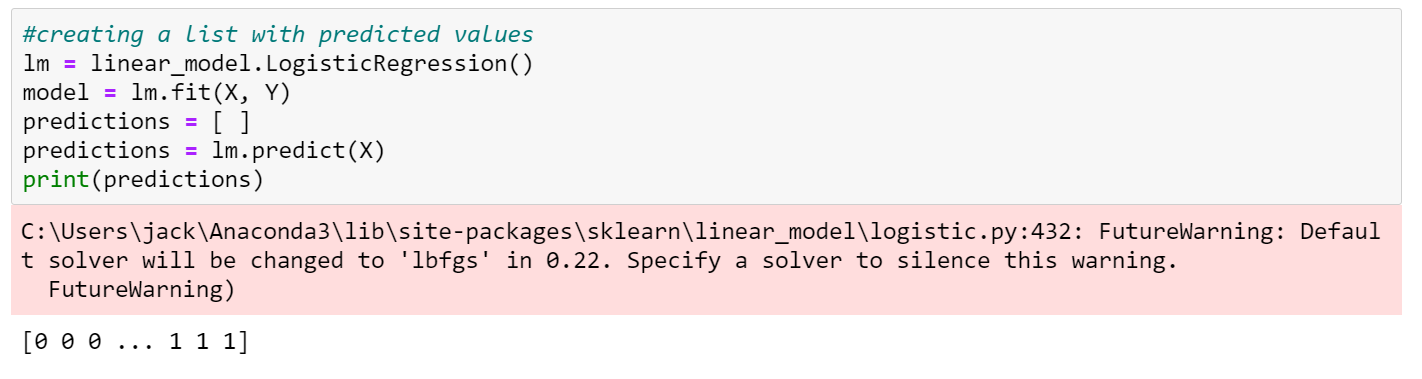


2. The indicator that I used to reduce my model was the p value for each variable. If it was 0.05 or less I kept the variable. The model evaluation metric I used is accuracy score which I found very easily using the SKlearn library. I also calculated the mean squared error for each model.

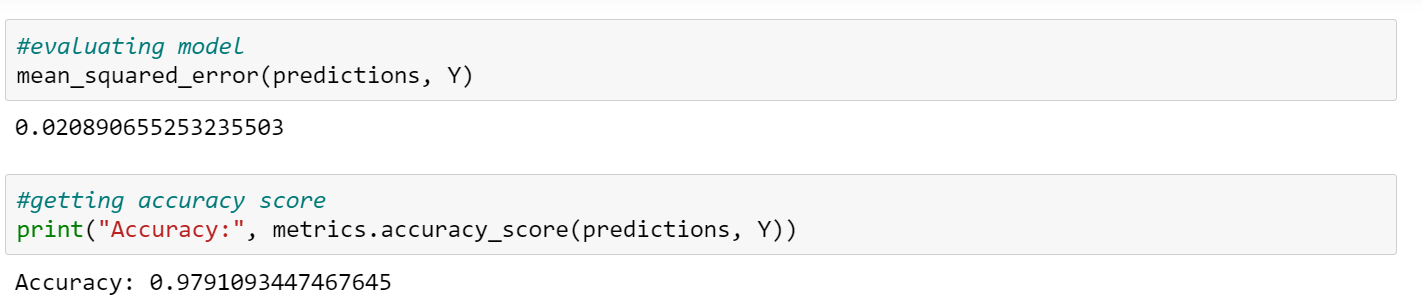
3. Here is the reduced model using the selected features

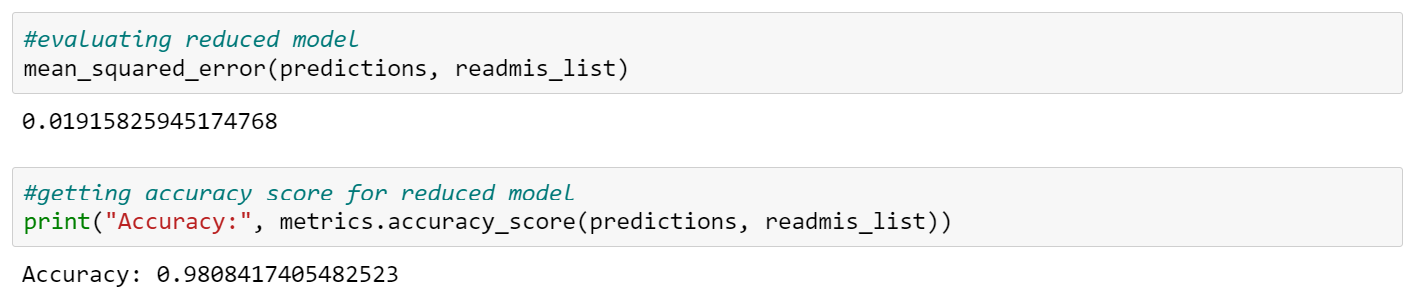






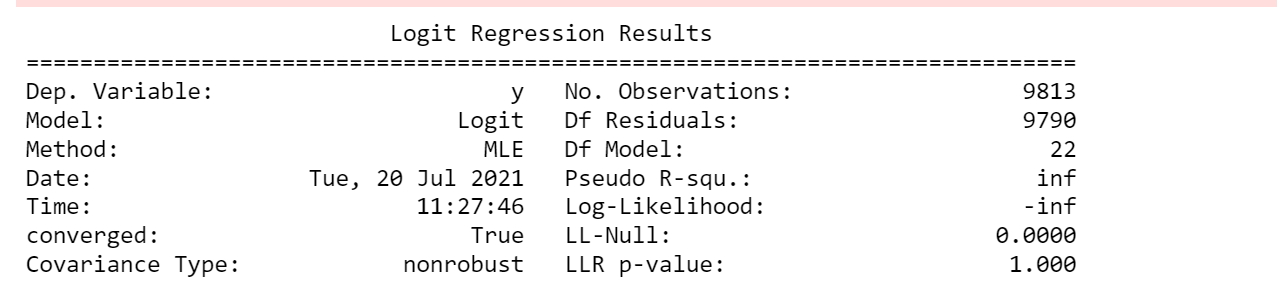
E. 1. My initial model included most of the variables (excluding identifiers), but my reduced model just looks at Initial\_days, Asthma, anxiety, arthritis, stroke, complication risk, admission type and diabetes . To select the variables I retained in my model, I used p values. I set my threshold at 0.05 and was left with the columns above. For model evaluation metrics I chose accuracy score (from SKlearn) and mean squared error. The original model had a MSE of 0.0209, and the reduced model had a MSE of 0.0191. For accuracy scores, the initial model got 0.979 and the reduced model got 0.981. The improvement is very slight but noticeable in the numbers.

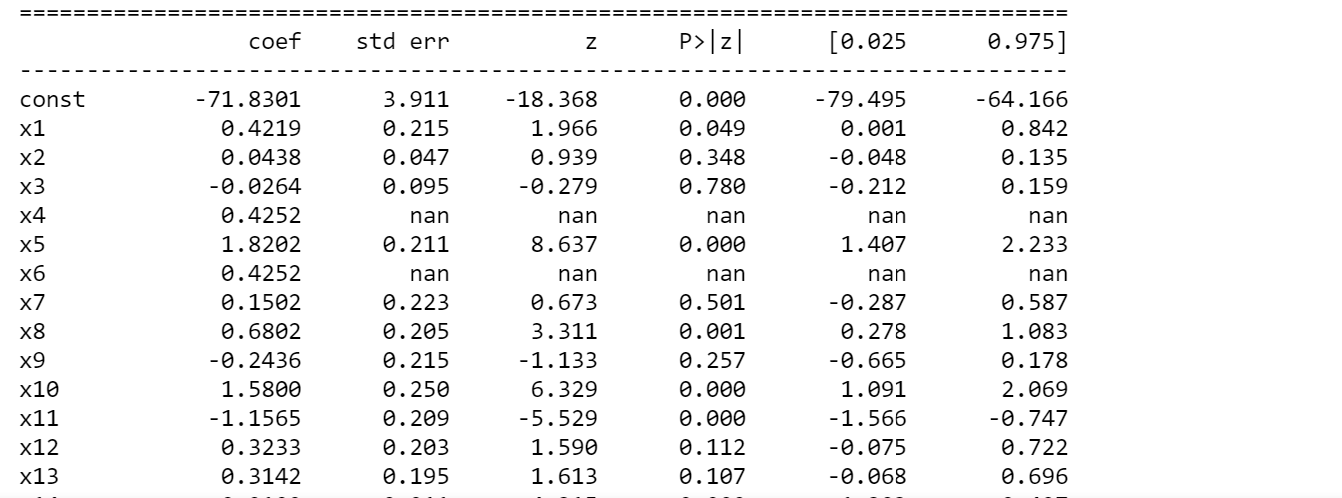


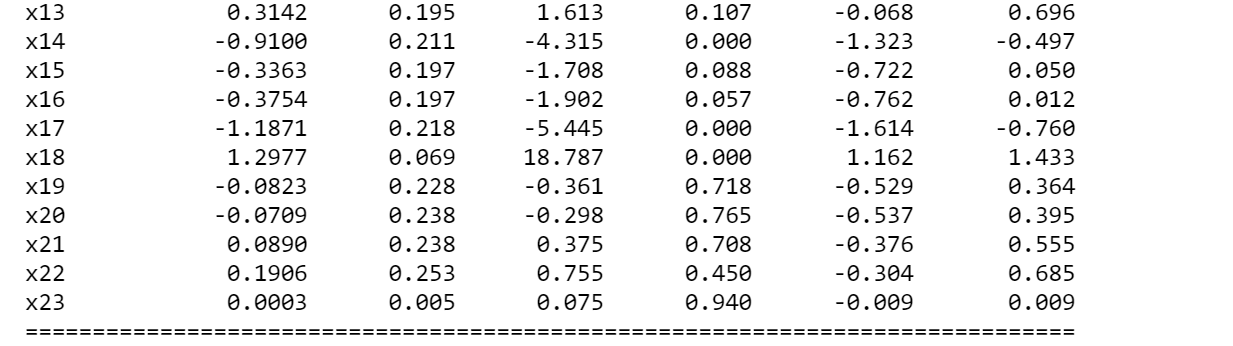


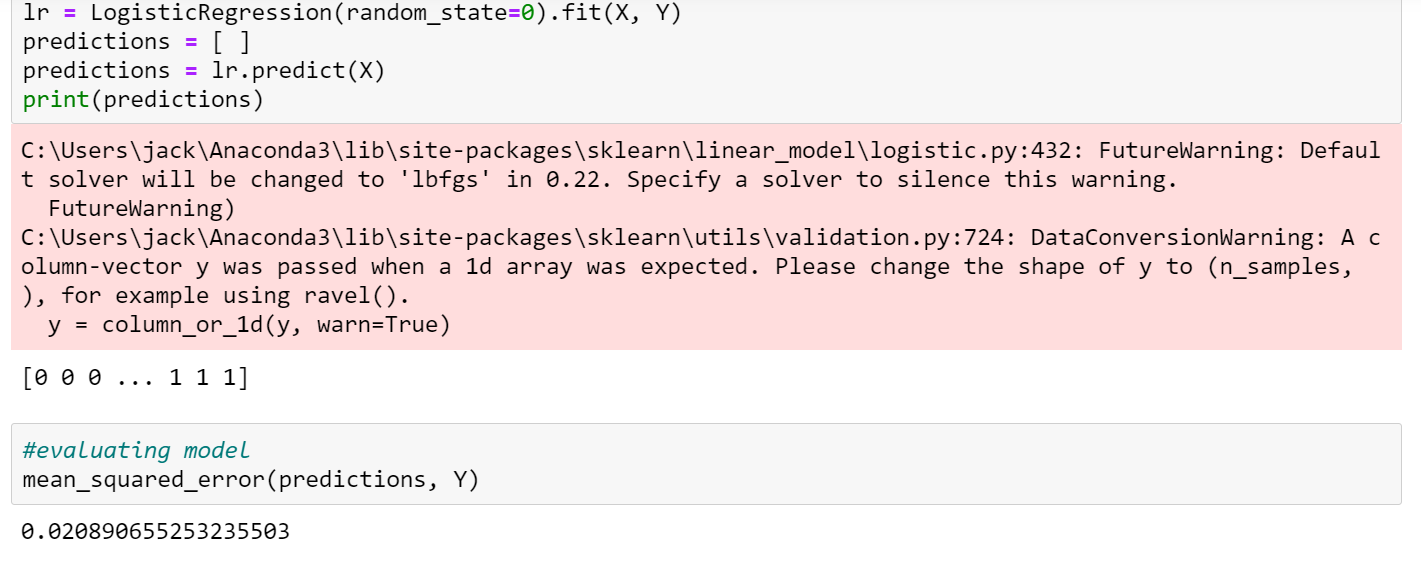
Below is the output for both models.

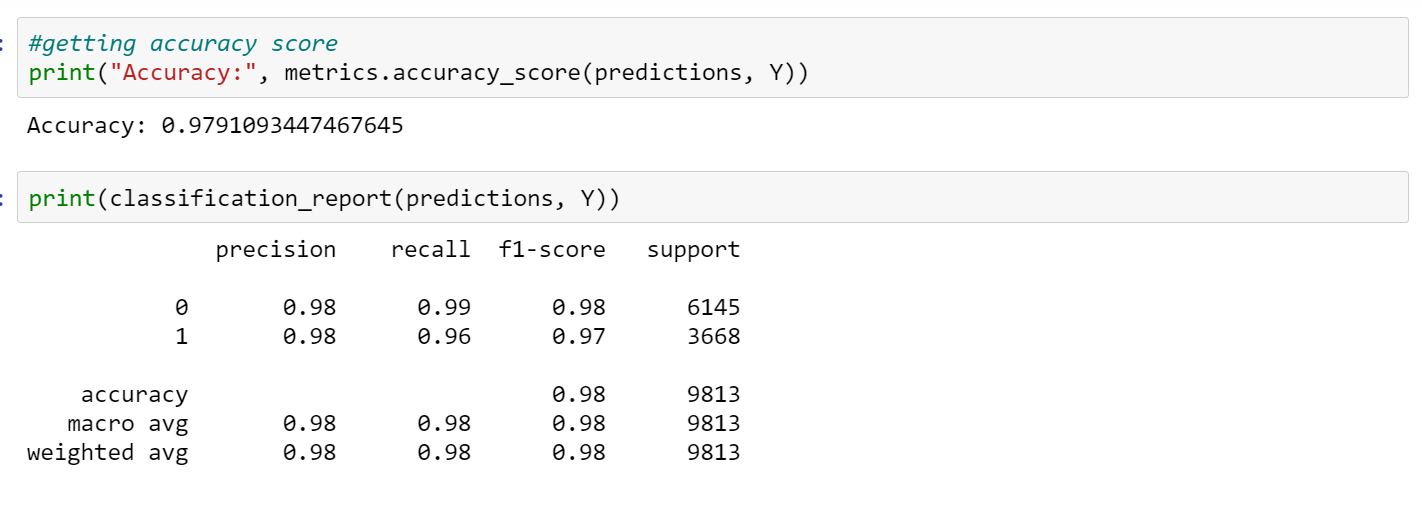
Here is the output for the original model.



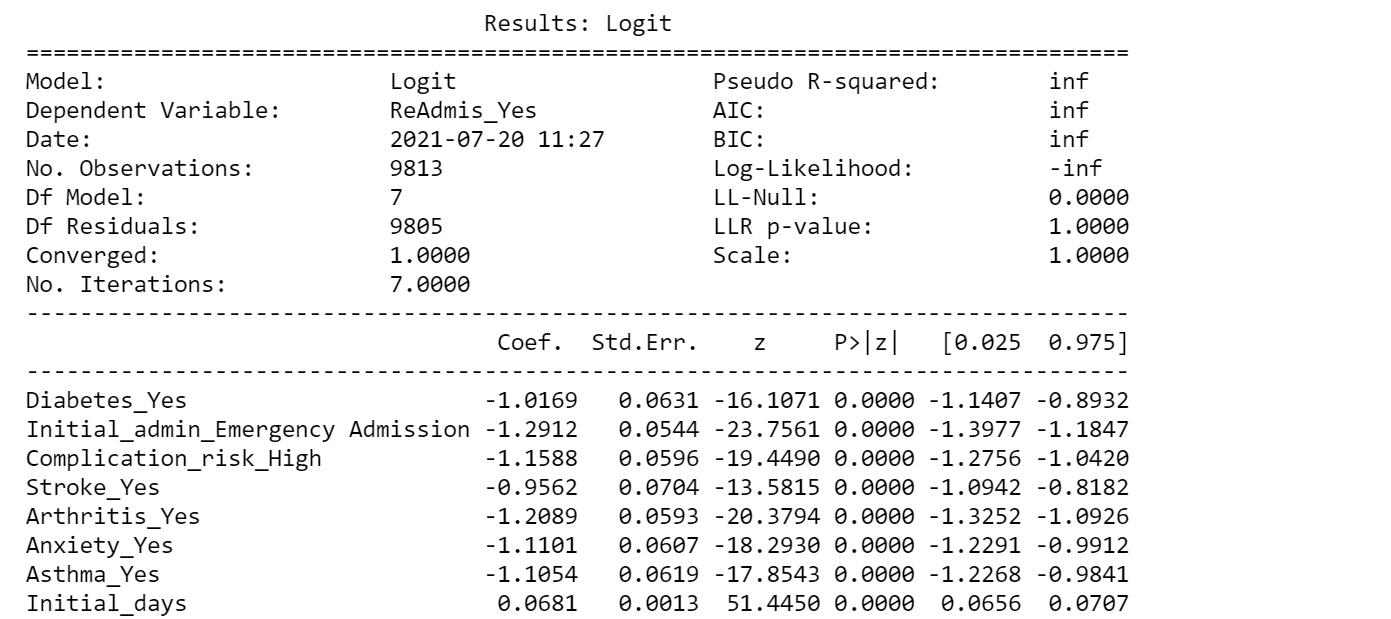


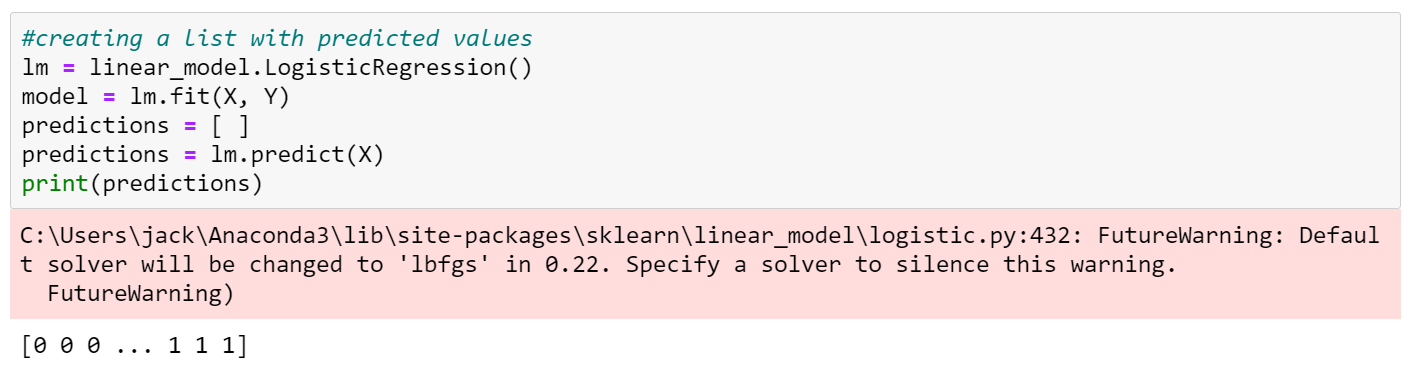


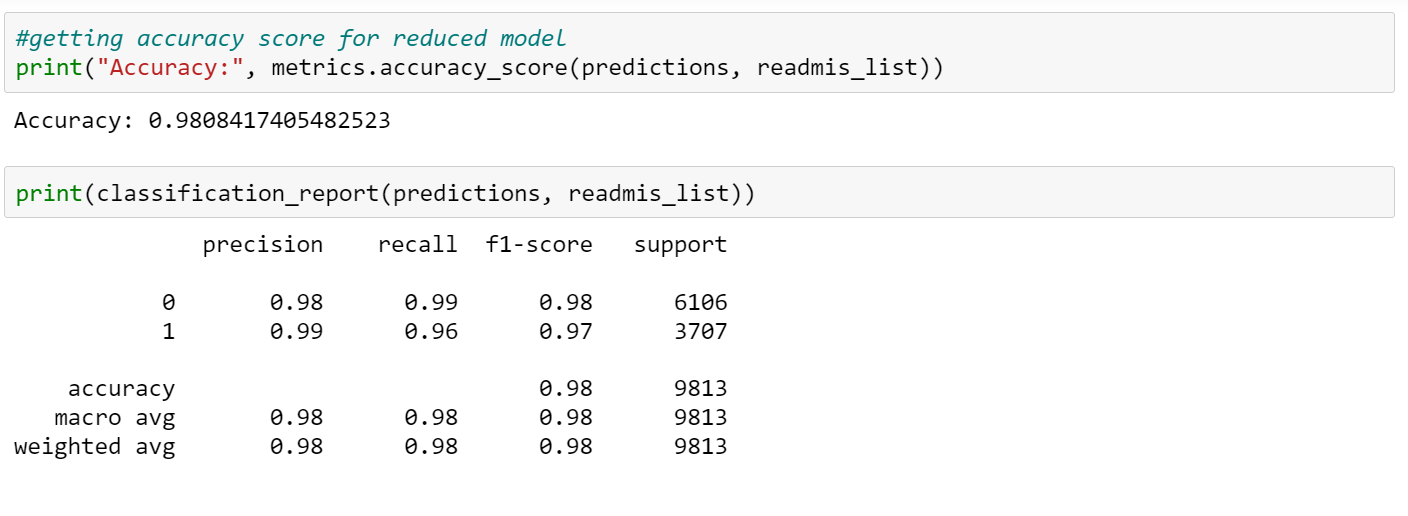




And here is the output for the reduced model.

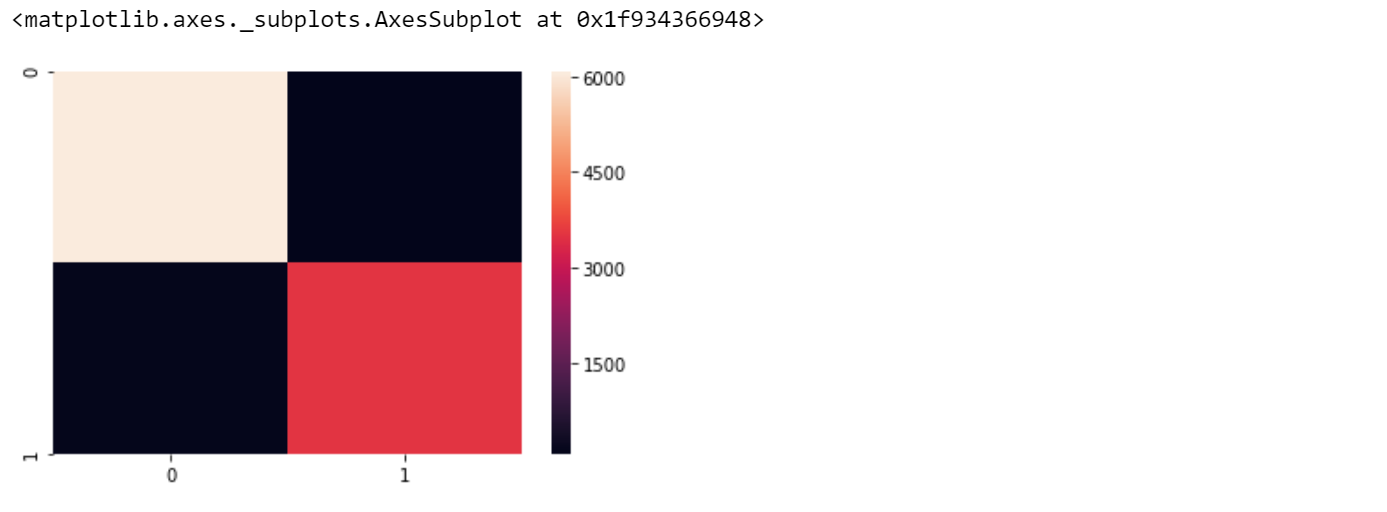




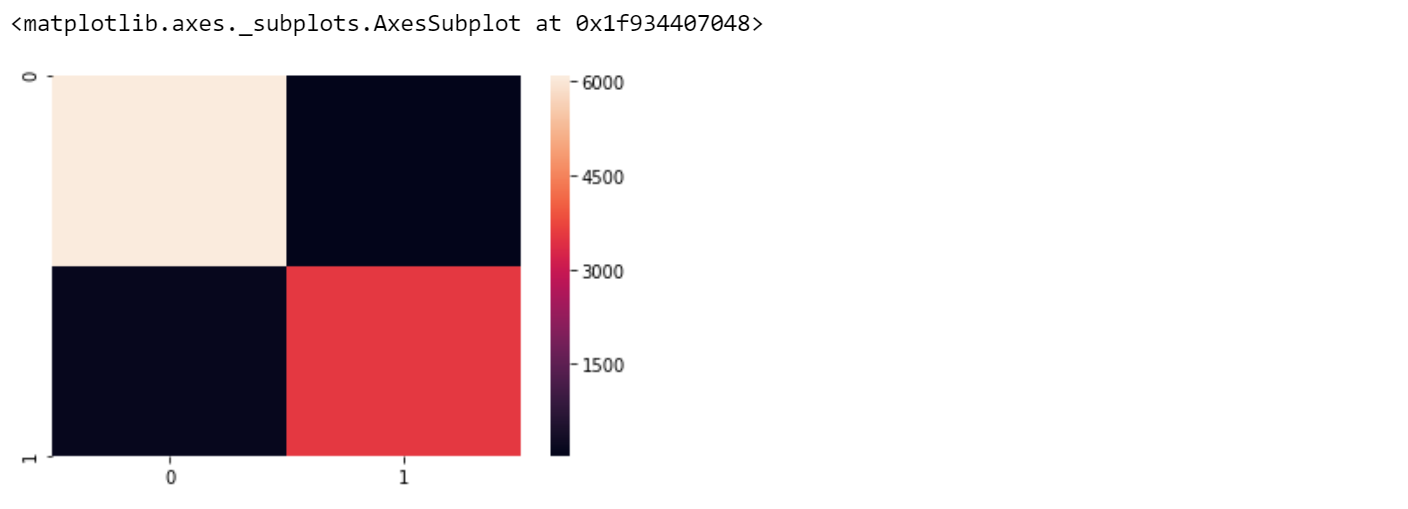


Finally I created a confusion matrix and a heat map to visualize it for each model.

Here is the matrix for the original model.



Here is the matrix for the reduced model.



3. The code used to create these models is included in the .ipynb file

Part V: Data summary and implications

F.1. The result of my analysis is that the reduced model is more accurate than the original model. The equation for that is

ReAdmis\_Yes = (-1.017\*Diabetes\_Yes) + (-1.29\*Initial\_admin\_Emergency Admission) + (-1.16\*Complication\_risk\_High) + (-.956\*Stroke\_Yes) + (-1.209\*Arthritis\_Yes) + (-1.11\*Anxiety\_Yes) + (-1.105\*Asthma\_Yes) + (0.068\*Initial\_days)

All of the coefficients for my variables except initial\_days are negative, meaning that the presence of them tends to reduce the possibility of being admitted. Initial\_days has a positive coefficient which means that those with longer initial hospital stays are more likely to be readmitted. Statistically, this model is accurate and usable to predict whether or not a patient will be readmitted. It could be used in a practical setting to increase follow up care with individuals who are at higher risk of being readmitted. A limitation of the model is the medical history scope it covers. Factors like cancer, autoimmune disease, and thyroid issues are common and not accounted for in this model.

2. I recommend using this model as is to identify those that are at a high risk of being readmitted. Those patients can then be flagged and receive closer follow up care. More data on other major medical conditions could be collected to include more factors in our analysis.

H. Third party code referenced

*Finding coefficients for logistic regression in python*. Stack Overflow. (2019, November 1). https://stackoverflow.com/questions/57924484/finding-coefficients-for-logistic-regression-in-python.

*LogisticRegression*. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html.

1. Works Cited

*Add\_constant. statsmodels. (n.d.). https://www.statsmodels.org/devel/generated/statsmodels.tools.tools.add\_constant.html.*

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.