**D208 Performance Assessment**

**LOGISTIC REGRESSION FOR MEDICAL DATA**

Fahim A. Akbar Student ID 001434895 Masters Data Analytics (January 1, 2021) Program Mentor: Lea Yoakem (877) 435-7948x6422 [fakbar3@wgu.edu](mailto:fakbar3@wgu.edu)

**Part I: Research Question**

**A.  Purpose of using Logistic Regression for Data Analysis**

**A1.  Research Question**

The proposed research question is if to determine which factors most significantly contribute to patients diagnosed with arthritis.

**A2.  Research Goal**

Analyzing the relationships between multiple variables and a patient's arthritis diagnosis can help determine whether there is any discernible link between other healthcare conditions or different patient data and arthritis. To do so, we will test to see which independent variables in the dataset are good indicator and fit. Once we decide on which variables to use, we will be doing a logistic regression analysis and predict the probability of patient readmissions. We will also provide a logistic equation along with the statistics for the identified variables.

**Part II: Method Justification**

**B.  Logistic Regression Methods**

**B1.  Logistic regression model assumptions**

According to Statology, the following assumptions must be met in order for a logistic regression model to be a good fit for the present data: (Zach 2020)

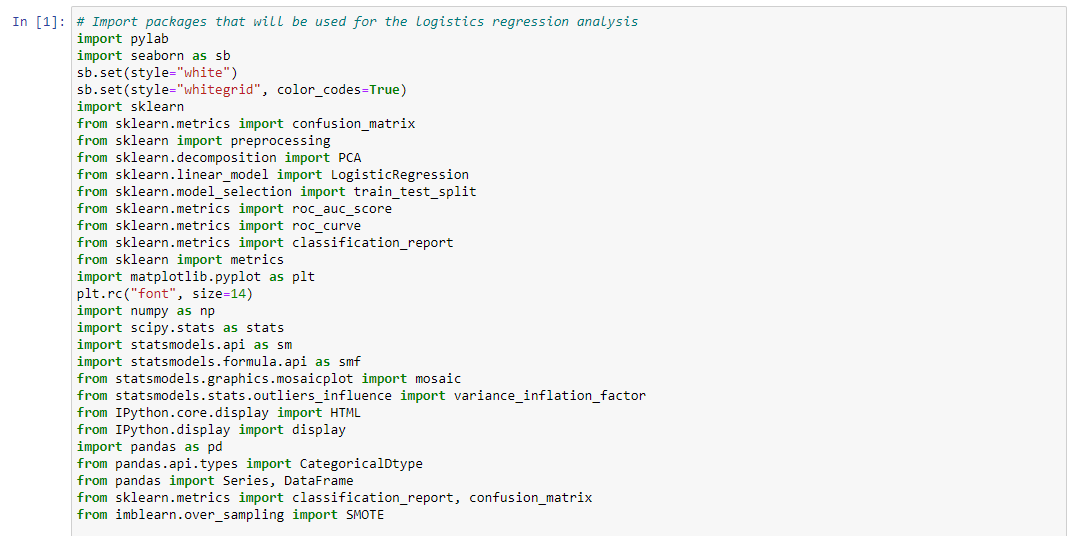
* The dependent variable is binary.
* Observations are independent. Residual plots against time can be used to check if there is a random pattern in the order of observations. If there isn’t, then the observations do not satisfy this assumption.
* No multicollinearity should be present among the variables.
* No extreme outliers should be present.
* Explains the relationship between one dependent binary variable and one or more independent nominal variables.
* A linear relationship should be present between the independent variables and logit of the dependent variable. Logit is defined as Logit(p) = log(p / (1-p)), where p is the probability of a positive outcome.
* The dataset offers a large sample size. Our data set has over 10,000 entries and is therefore sufficient.

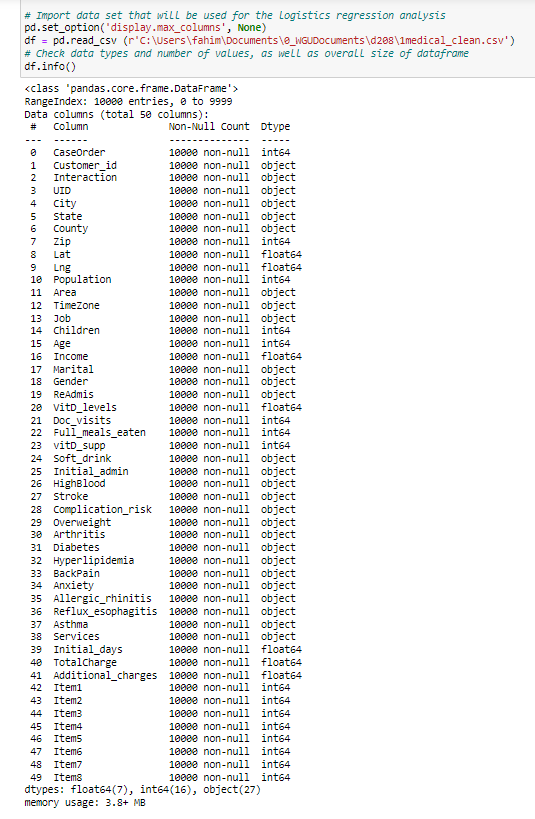
**B2.  Benefits of using Python and chosen packages**

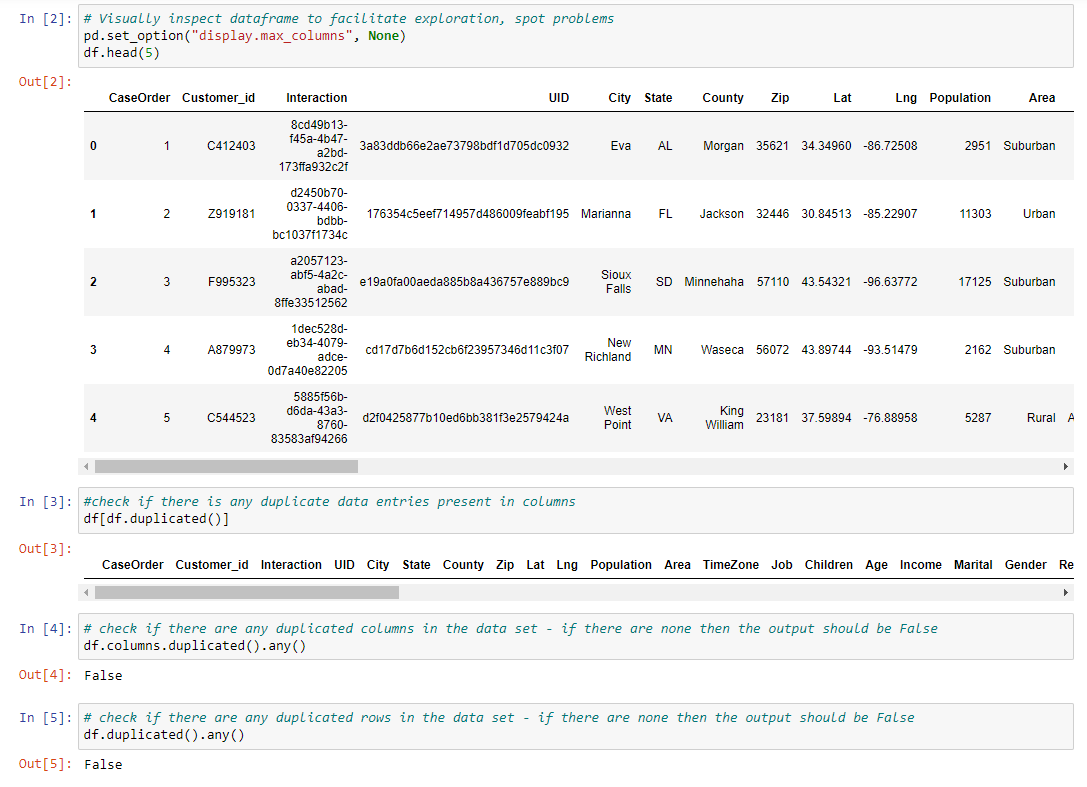
In this assessment, Python will be used for the logistic regression model of the medical data. Python was selected over R as it is the programming language that I’m most comfortable working with. It’s easy to understand syntax is also useful when presenting findings. Python is proficient in supporting data science tasks, especially in utilizing packages specifically designed for logistic regression. Python provides several packages that allow for statistical analyses, such as Pandas, Scipy and Statsmodels. Python enables easy-to-understand visualizations of variables and the observations within them. It allows you to analyze data easily into separate groups without modifying the original dataset. Python also has commands specific to logistic regression, including LogisticRegression, model.fit, and model.coef\_. These commands will help with performing the regression analyses and visualizing them.

The following packages will be imported and used for this analysis:

* Pandas – Standard import for data science projects. Provides methods to read and visualize data and statistical tools to parse and score data.
* Numpy – Standard import for data science projects. Provides methods to read and visualize data and statistical tools to parse and score data.
* Matplotlib – Standard import for visualizations. This package provides more robust tools to visualize reports and data points.
* Seaborn – Provides descriptive and visually intuitive graphs, plots, and matrices
* Scikit-learn/sklearn – Provides methods and arguments for splitting, training, testing, and fitting data. This package also has arguments for predicting and classifying data as well as applying metrics for models







**B3. Why logistic regression is an appropriate technique to analyze the research question**

For this assessment the goal is not to predict specific values. Rather, the goal is to predict the probability of the event on interest, arthritis, occurring. To analyze the probability of patient arthritis, Logistic regression is an appropriate technique. This is because the patient arthritis is a binary predictor to a categorical variable (Arthritis = Yes or No.) The patient observations are the independent variables of the data set that will be tested.

**Part III: Data Preparation**

**C.  Summarize the data preparation process for logistic regression by doing the following:**

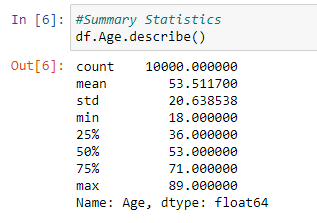
**C1. Relevant data preprocessing goal and data manipulations**

The medical data set that was provided requires preliminary cleaning and preprocessing before use. Logistic regression analysis requires numerical values instead of strings for the categorical and boolean data in this dataset. Booleans can be easily converted from Yes/No to numeric values (1/0). Handling categorical variables depends on their nature: ordinal variables (where the order matters, like "big", "bigger", "biggest") and nominal variables (where order doesn't matter). In this dataset, aside from boolean data, there are other categorical types. For this analysis, nominal categorical columns require the creation of dummy columns. These columns transform categorical data into binary numeric representation. For example, the gender column containing values like Female, Male, and Nonbinary can be converted into two columns conveying the same information. If the first column has a 1, the patient is male; if the second column has a 1, the patient is nonbinary. When both columns show a 0, indicating the absence of a 1, the patient is identified as female. This technique is known as as one-hot encoding, and allows us to represent textual data numerically which the regression analysis can handle. I will be using panda’s get\_dummies function for one-hot encoding.

**C2. Summary Statistics**

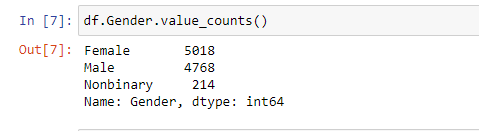
The variables which I intend to examine for this analysis include the dependent variable, Arthritis, and the following independent variables:

**Age**

****

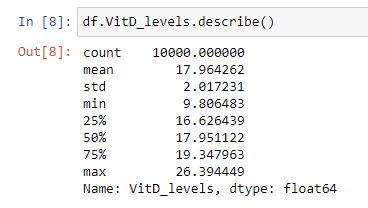
An observation in our data's summary statistics is the absence of patients below the age of 18. For this analysis, this limitation might be less significant because arthritis is uncommon among children unless caused by events like injuries, severe illnesses, accidents, or genetic predispositions. In such cases, the causation likely falls beyond the scope of this analysis.

**Gender**

****

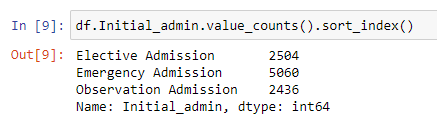
This variable's limitation lies in its categorization of genders into only three options, which oversimplifies the diverse spectrum. However, given the available data, we have no choice but to work with these categories. The data shows that approximately 50% of patients are female, nearly 48% are male, and slightly over 2% identify as nonbinary.

**Vitamin D Levels**

****

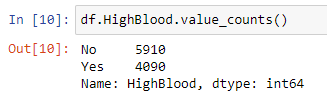
The data indicates a mean close to 18 and a standard deviation slightly above 2. The minimum and maximum values are slightly more than 4 standard deviations away from the mean, and the interquartile range falls between 16.6 and 19.4. This suggests a distribution that is roughly normal in shape.

**Reason for Initial Admission**



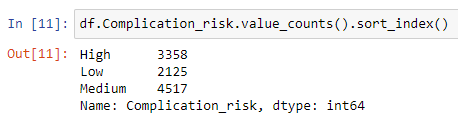
Over half of the patients in this dataset are initially admitted for emergency reasons. This is expected, as hospitals are typically sought out in emergency situations. The remaining half of hospitalized patients are almost evenly divided between those admitted for observation and those undergoing elective procedures.

**High Blood Pressure**

****

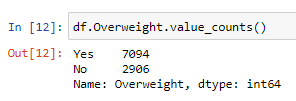
Around 59% of patients in this dataset do not have a high blood pressure diagnosis, whereas 41% have been diagnosed with the condition.

**Complication Risk**



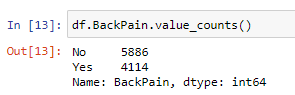
Roughly 45% of patients face a medium risk of complications during their hospitalization. Only 21% are at low risk, with the remaining 34% categorized as high risk for complications.

**Overweight**



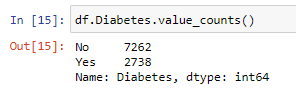
Nearly 71% of the patients in the dataset are classified as overweight, while the remaining 29% fall into the non-overweight category.

**Backpain**

****

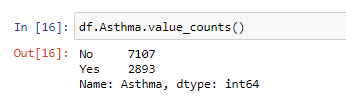
41% of patients from the data set experience back pain, while the remaining 59% do not. I expect backpain and arthritis to be correlated due to both conditions being joint related issues.

**Diabetes**



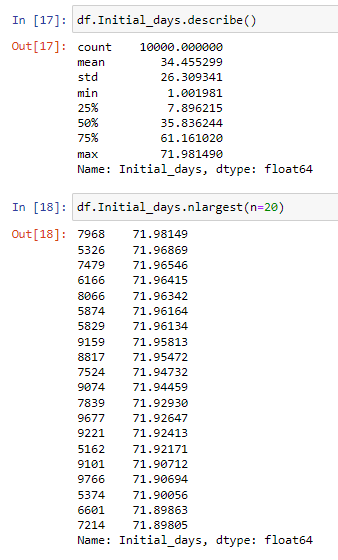
Diabetes has been diagnosed in 27% of the patients within this dataset.

**Asthma**

****

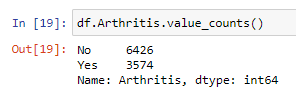
2893, or nearly 29%, of the patients in this dataset have received an asthma diagnosis.

**Length of hospitalization**



In this dataset, the average duration of hospital stay is 34 days, accompanied by a substantial standard deviation of 26 days. The data spans from 1 day to 72 days, placing the minimum at 1.27 standard deviations below the mean, and the maximum at 1.5 standard deviations above the mean. This skewed distribution suggests an imbalance toward one side. Notably, a brief examination of the highest values for this variable confirms that the approximately 72-day maximum is not an isolated outlier.

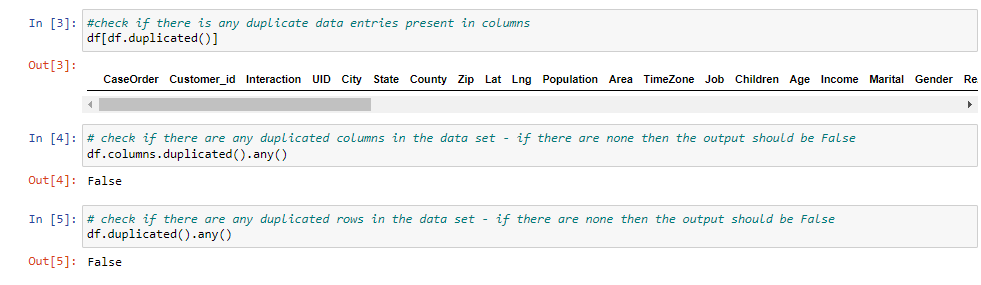
**Arthritis**

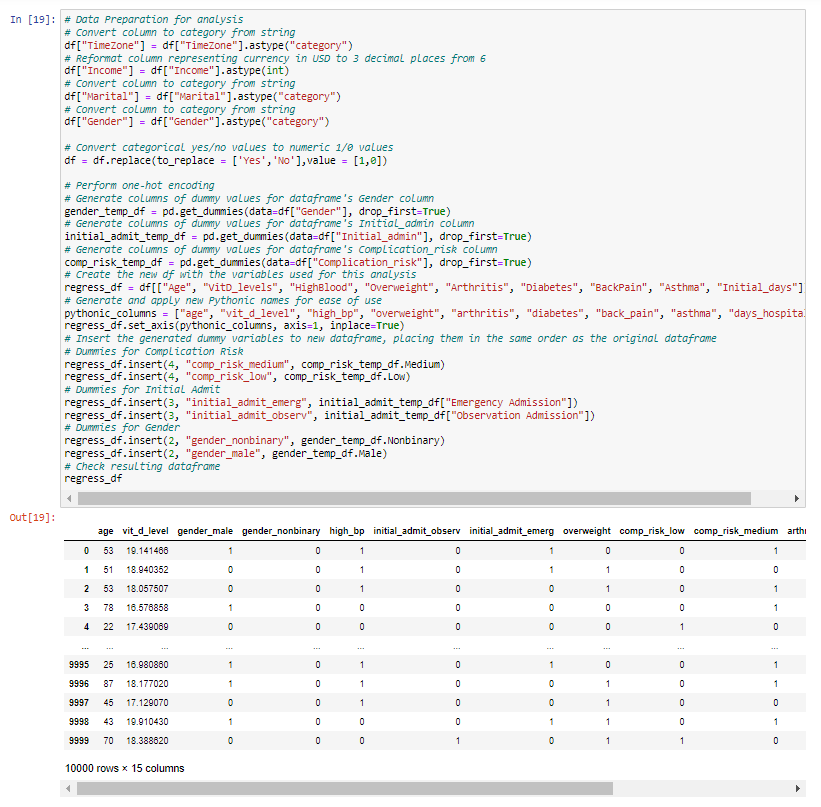
****

All of the previous statistics describe the independent variables, which we are testing for a relationship with diagnoses for arthritis. As such, the summary statistics for arthritis must also be presented here. 3574, or roughly 35%, of hospitalized patients in this dataset have arthritis.

**C3. Data Preparation**

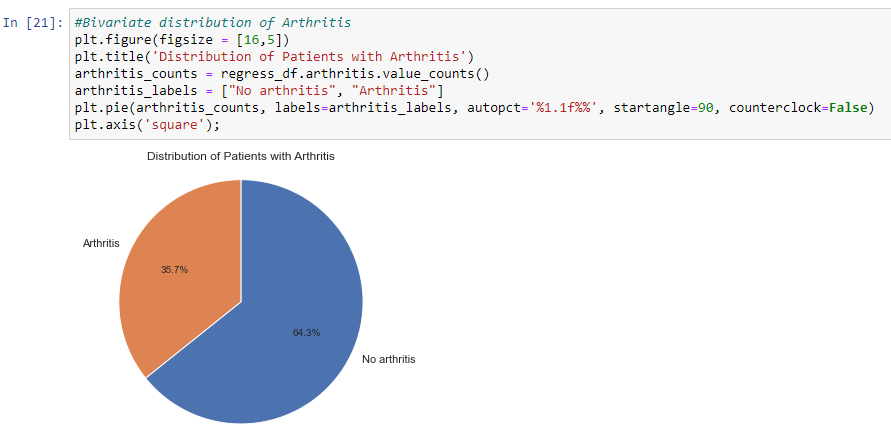
The first step in preparing the data is to make sure that there are no missing data entries in any of the columns. Next, we want to make sure there aren’t any duplicated data entries. We’ll also make sure that there aren’t any duplicated columns or rows to further prevent dealing with repeated entries. Afterwards, the summary statistics for the chosen target variables will be reviewed, as shown in C2. Once that is done, the data will be prepared for the logistic regression analysis. Yes/no responses will be converted into binary numeric values of 0 (No) and 1 (Yes). The categorical datatypes being used for the logistic regression analysis will be "dummied" using one-hot encoding. For ease of use, a new dataframe will be constructed with all of the necessary columns included for this regression, omitting the ~40 columns which will not be tested. Once all the necessary modifications to the data set are made, I will proceed with creating the initial logistics regression model.

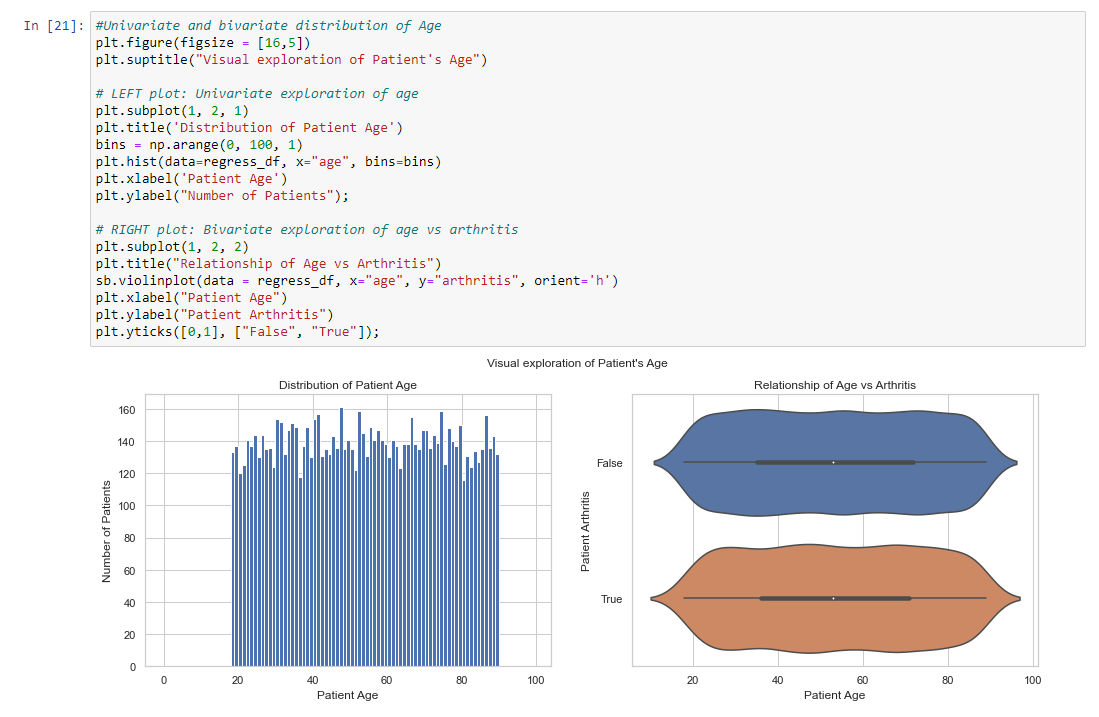




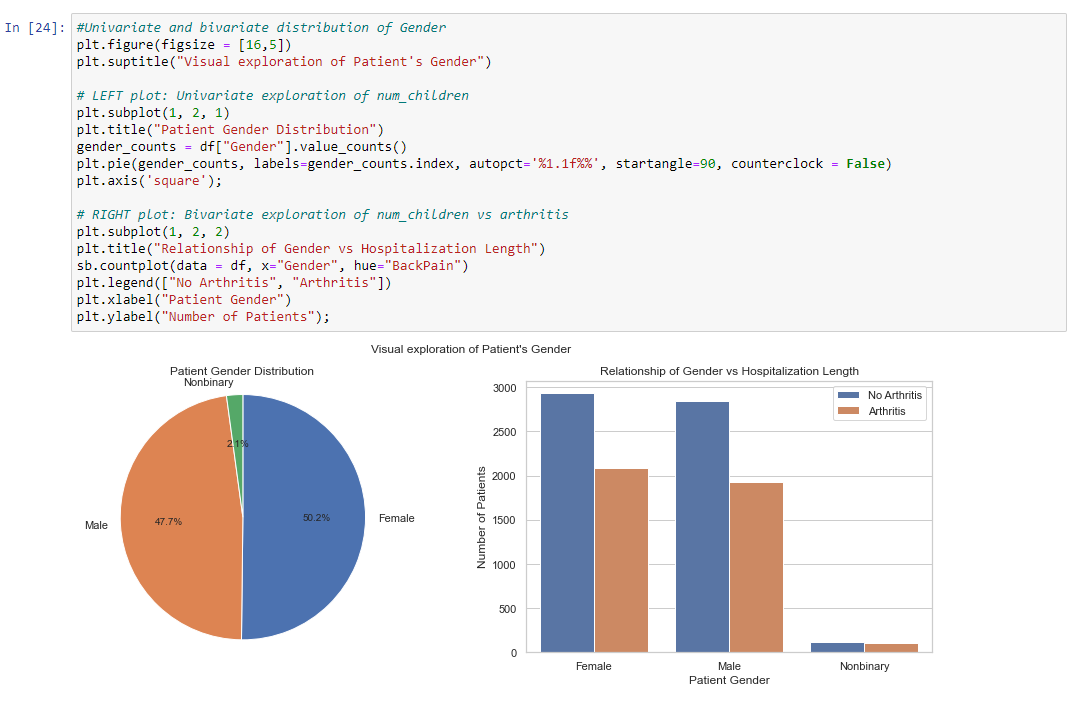
**C4: Univariate and Bivariate Distributions**

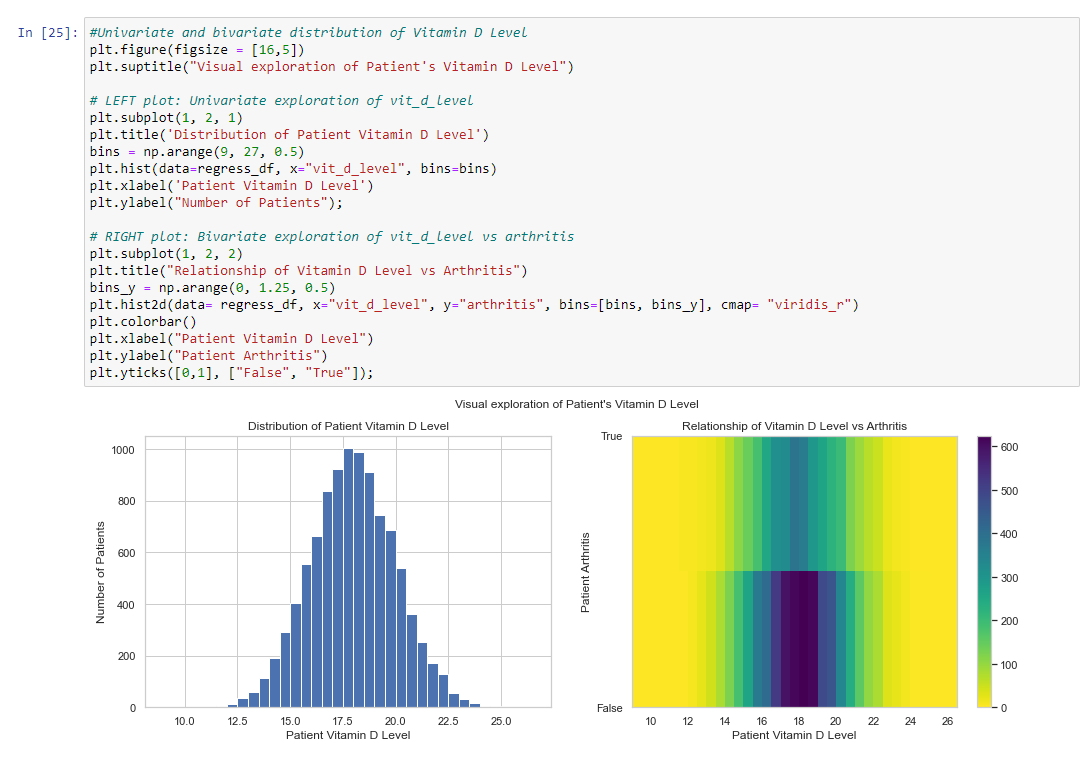
Before examining each of the explanatory variables, presented below is the bivariate distribution dependent variable first.

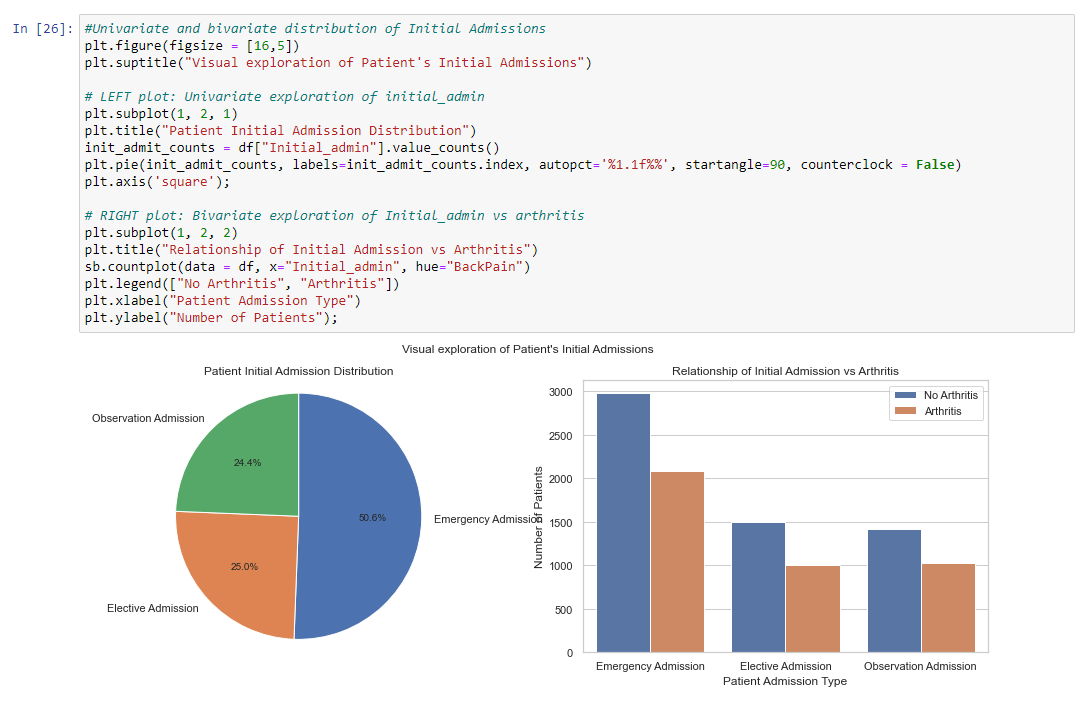






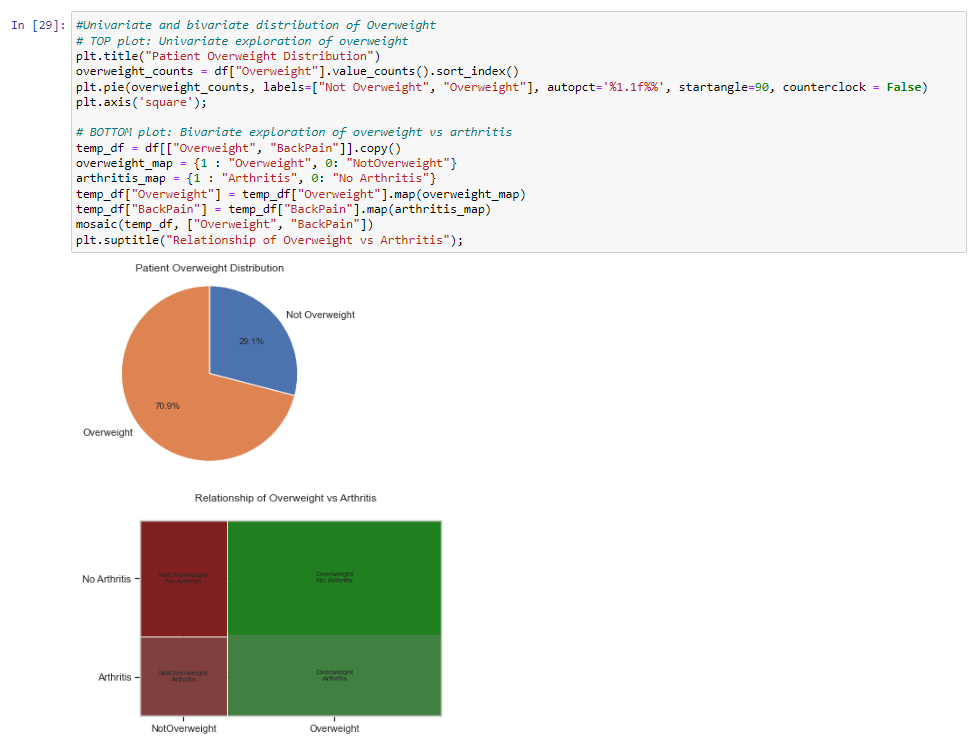


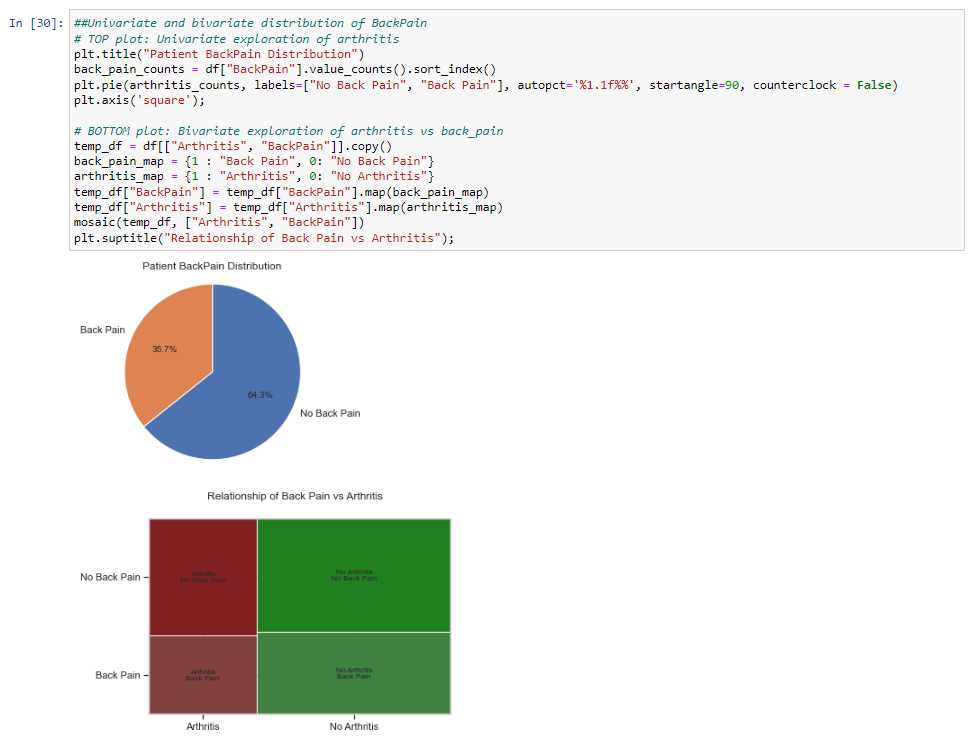


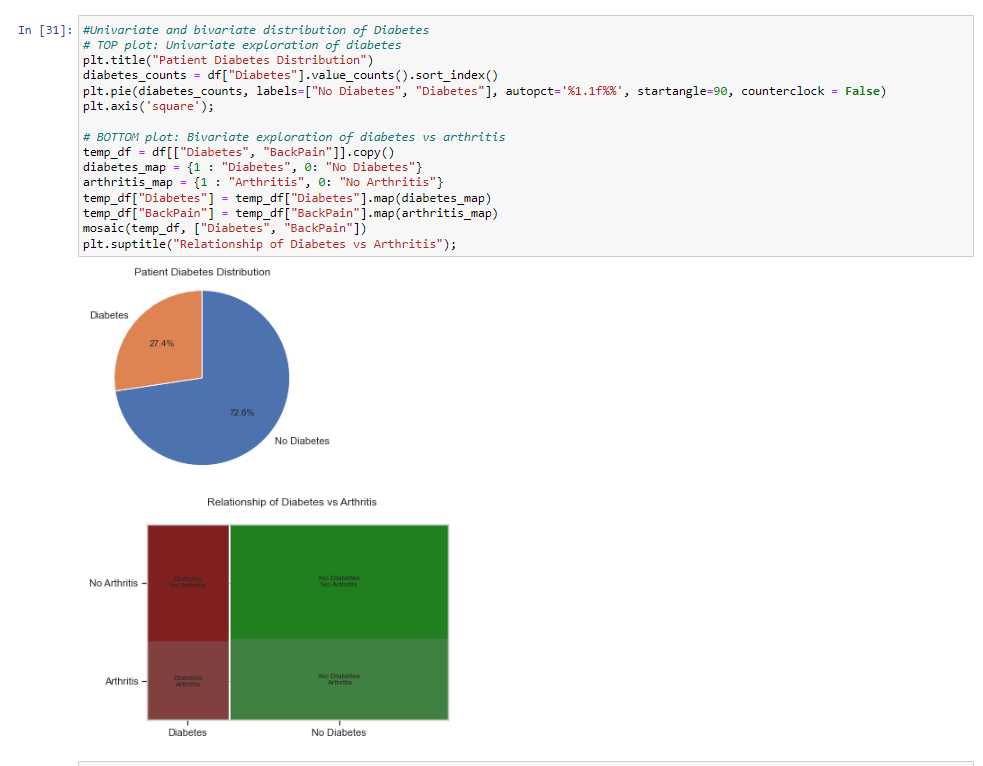


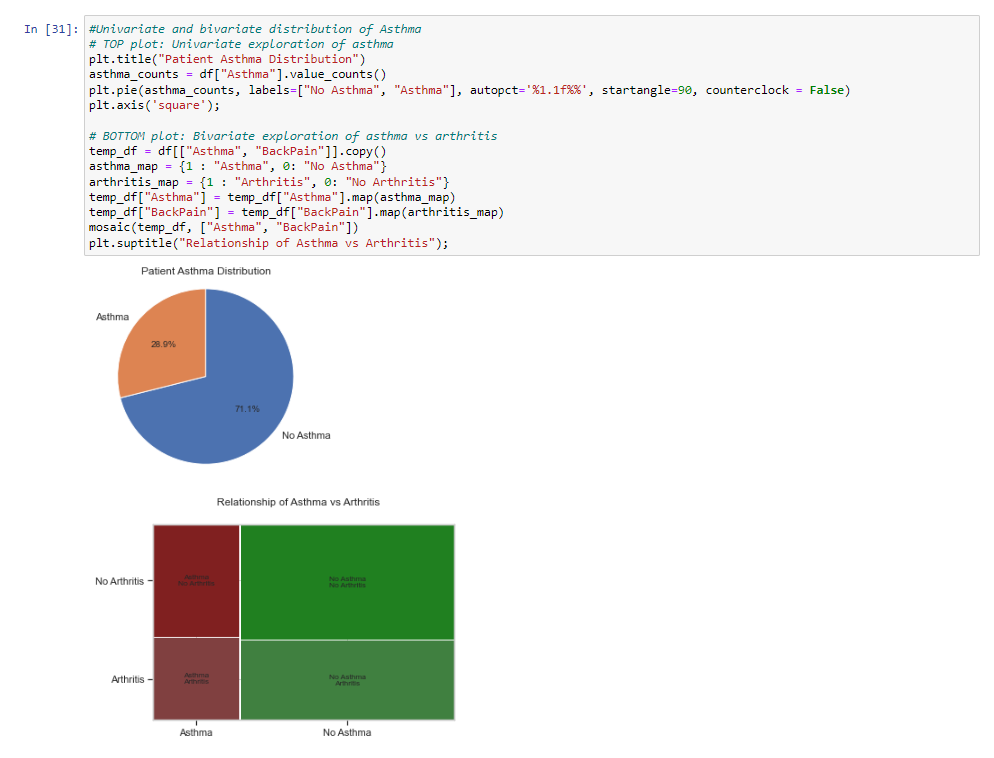


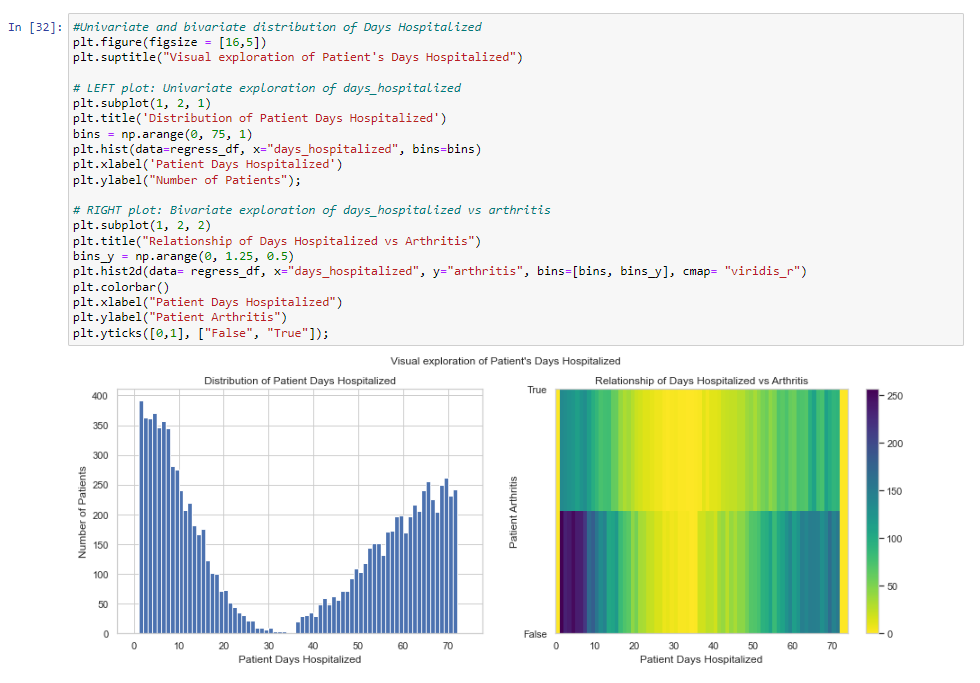












**C5.  Provide a copy of the prepared data set.**



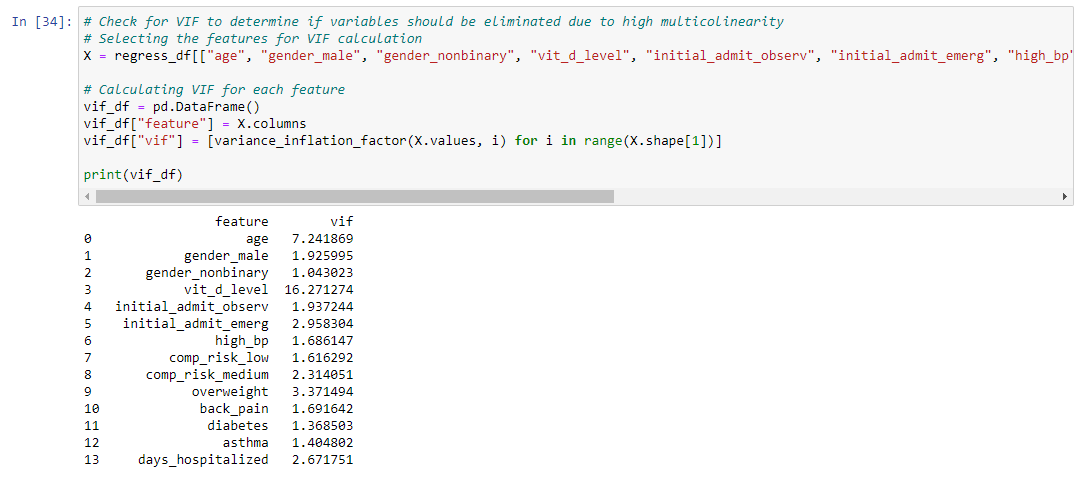
A copy of the cleaned data set, titled “d208task2\_full\_clean.csv” is provided in the task submission.

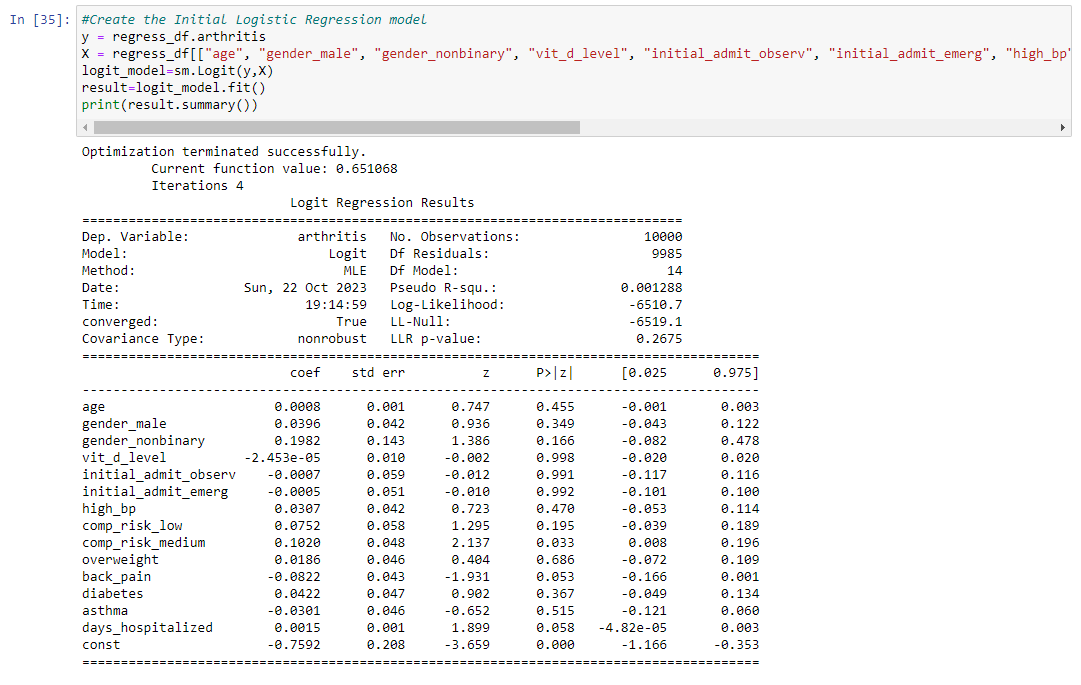
**Part IV: Model Comparison and Analysis**

**D.  Comparison an initial and a reduced logistic regression**

**D1. Initial logistic regression model**

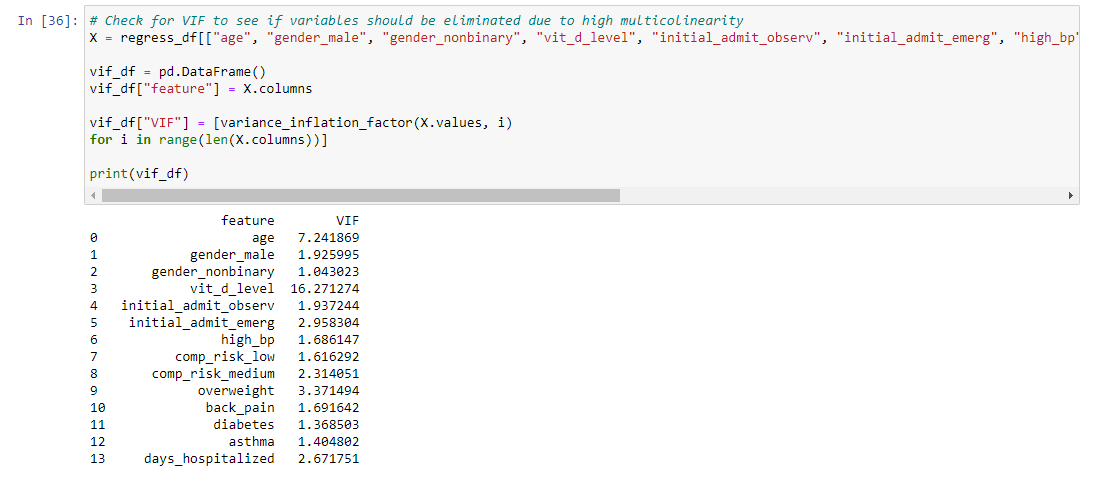
Generating the initial logistic regression model involves incorporating all previously highlighted explanatory variables, with the option to eliminate them later during the model reduction process. Before generating the Initial Logistic Regression Model, I first need to eliminate any variables that would cause high multicollinearity by examining the Variance Inflation Factor (VIF).

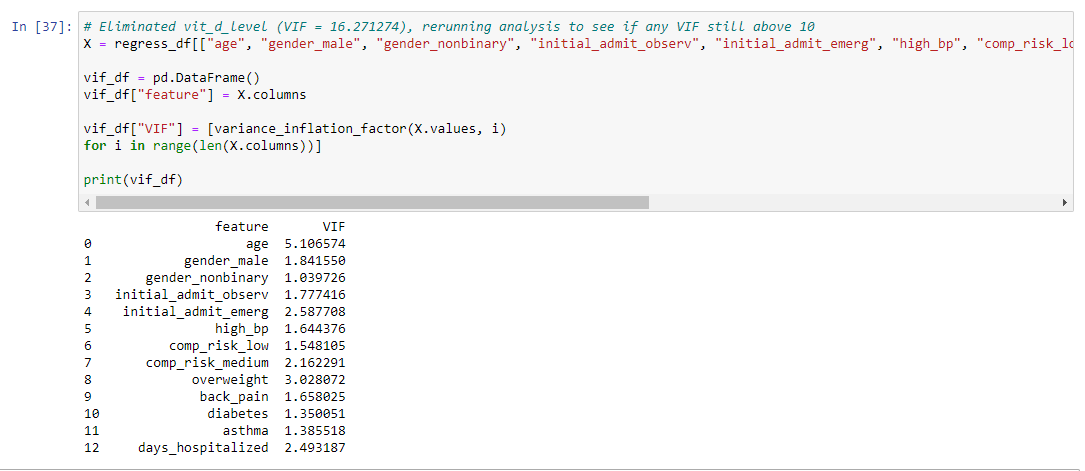
For starters, any VIF value above 10 will be removed. In this case, “vit\_d levels” has a value of 16.27 and will be the first variable removed during the variable reduction process. Additional multicollinearity issues will be addressed during the reduction of the dataset from the initial model. An initial logistic regression model has been constructed, incorporating all predictor values identified in section C2. This initial logistic regression model will undergo reduction to eliminate any observed issues and focus solely on variables that significantly contribute to the explanatory variable.

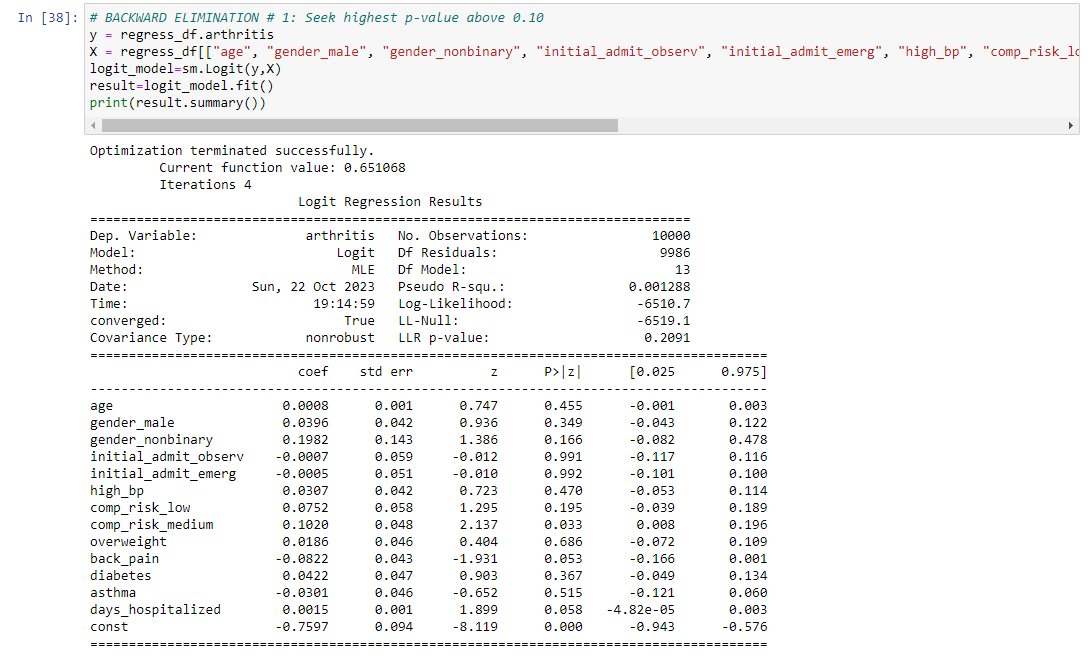


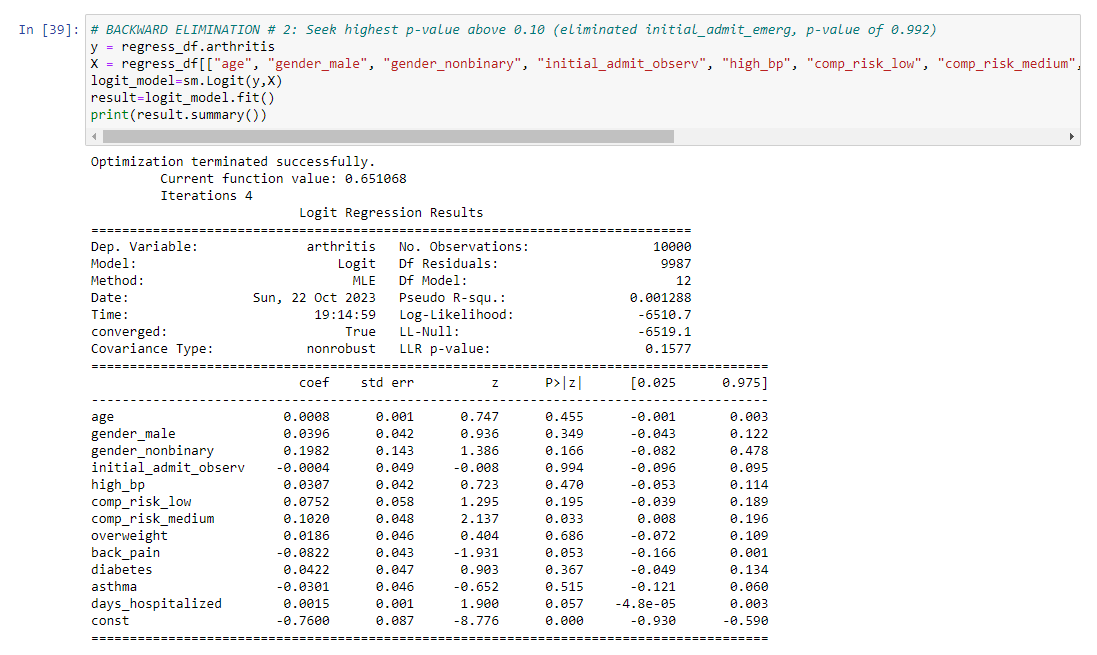
**D2.  Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question**.

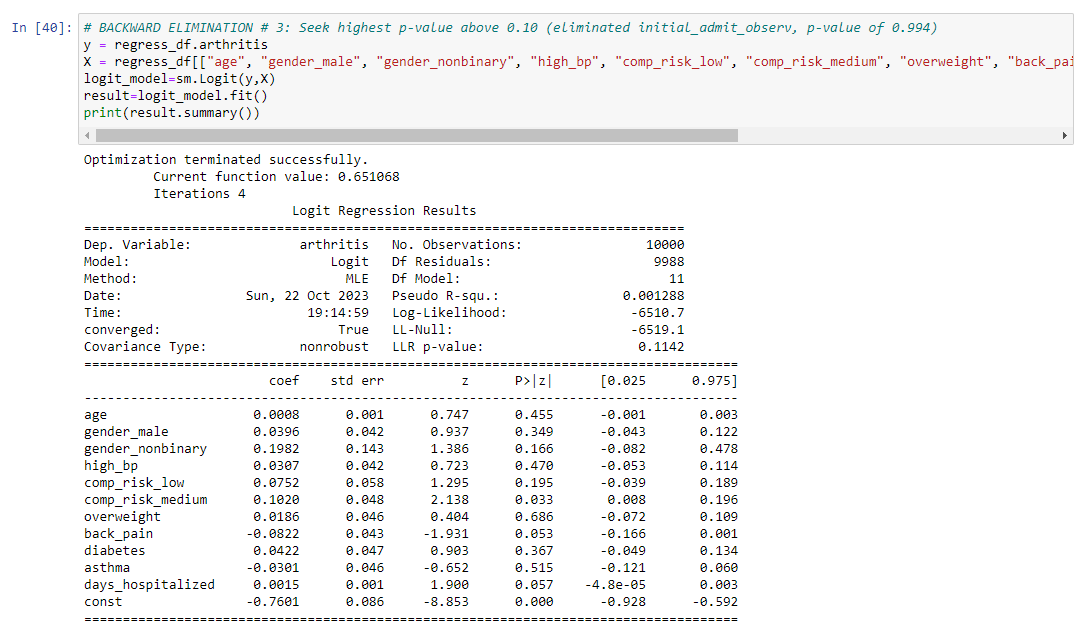
To begin reducing the logistic regression model, I will first begin with eliminate any variables that exhibit substantial multicollinearity. To identify and exclude problematic variables, the process of assessing the Variance Inflation Factor (VIF) will be employed again.

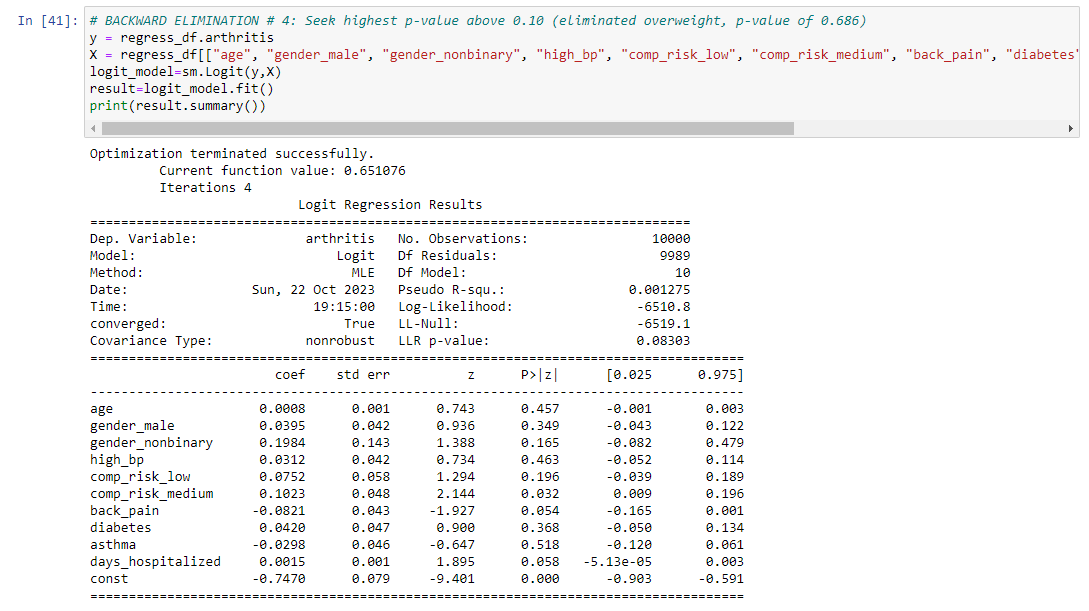


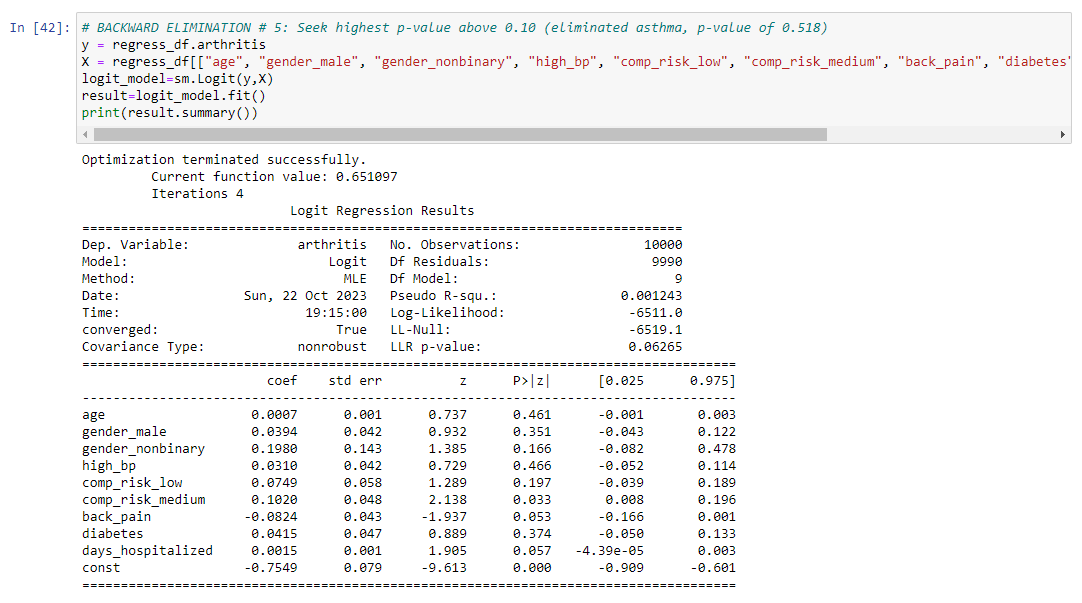


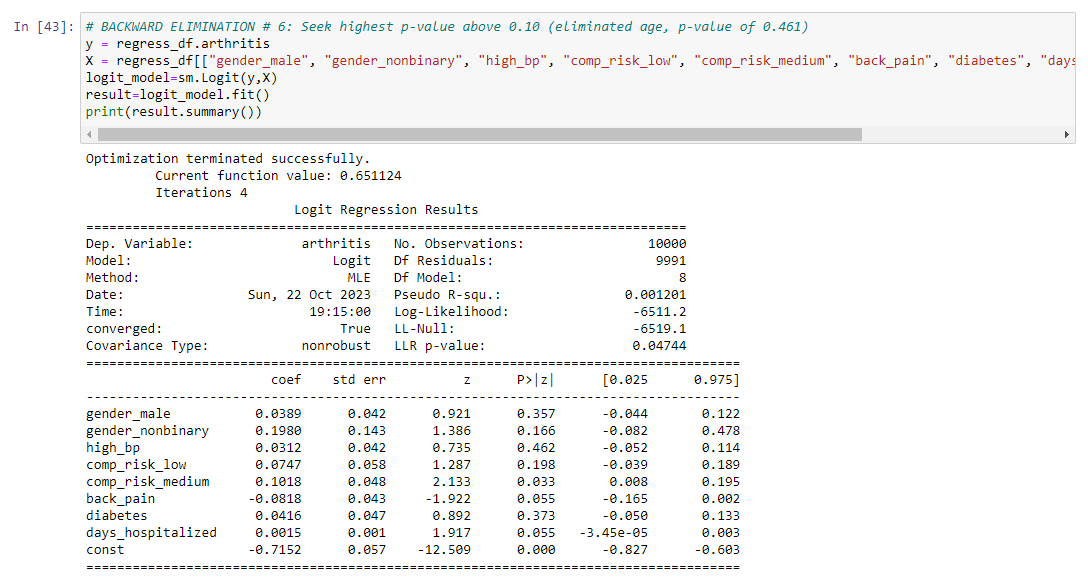
The variable patient vit\_d\_level exhibited a notably high Variance Inflation Factor (VIF) and was removed accordingly. The next step for reducing the model involves removing non-statistically significant variables from the model through Backwards Stepwise Elimination. This method involves generating the logistic regression model and evaluating the subsequent p-values associated with each variable. The focus is on variables that are statistically significant, determined by a threshold (alpha) of 0.10. Any p-value below this threshold is considered significant. The model will be created and the variable with the highest p-value above our significance threshold of 0.10 will be removed. The model will then be regenerated, and this process will be iterated until all p-values for the remaining variables fall below 0.10. I ended up performing Backwards Stepwise Elimination 11 times before arriving at the final reduced logistic regression model. 

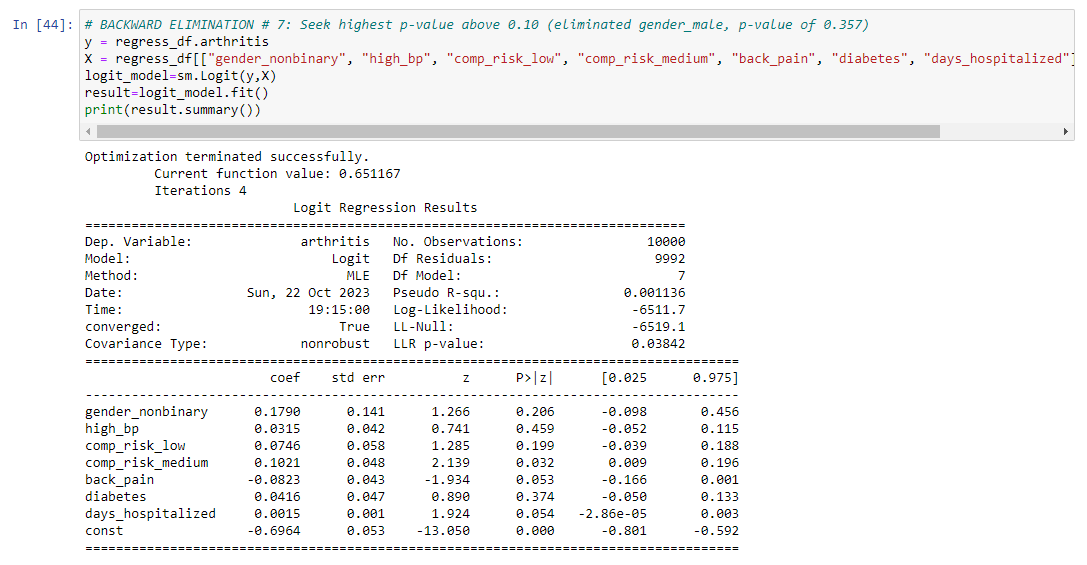


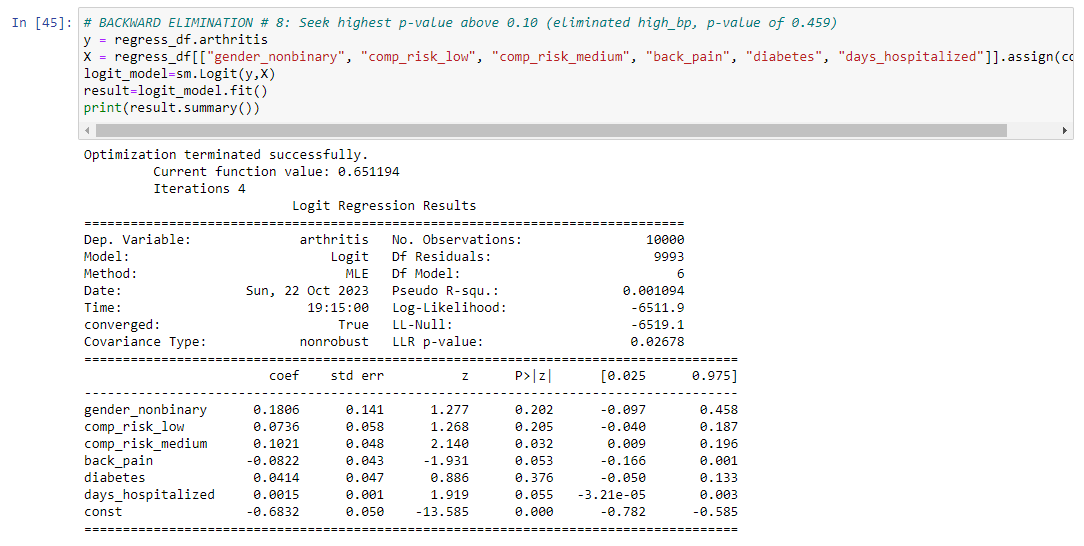


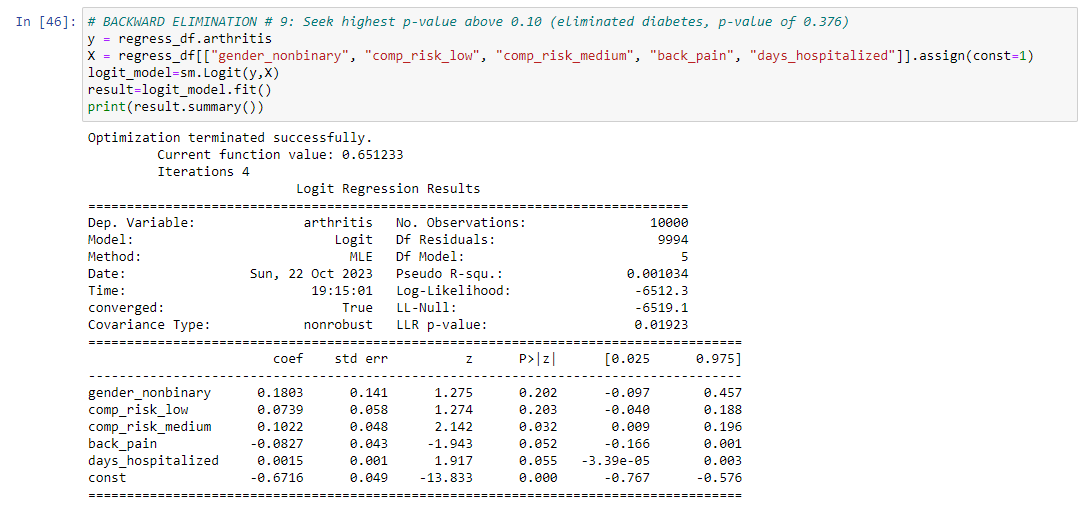


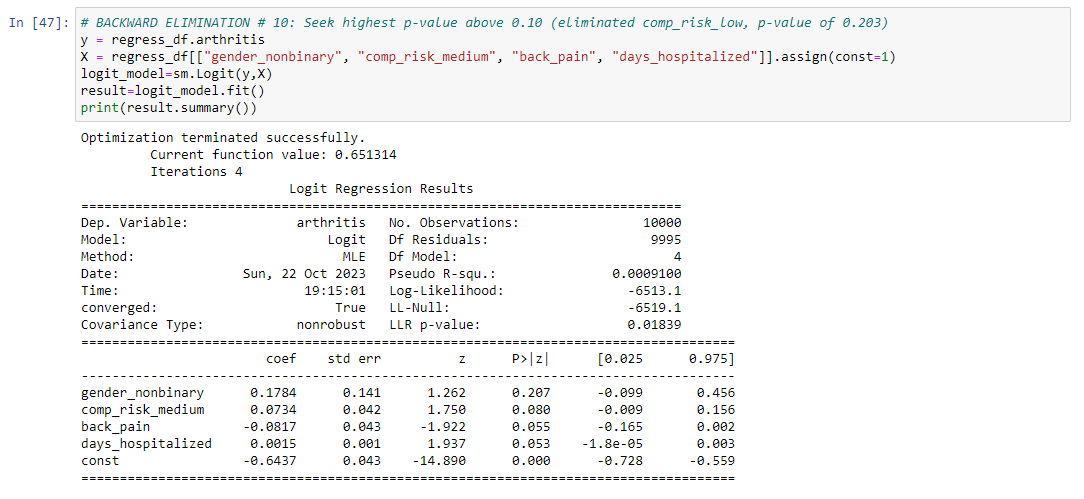


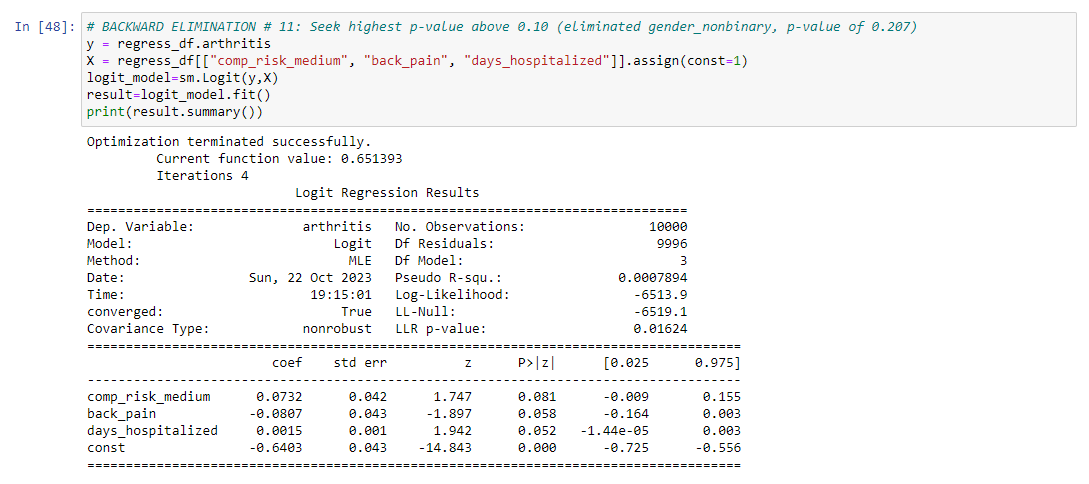










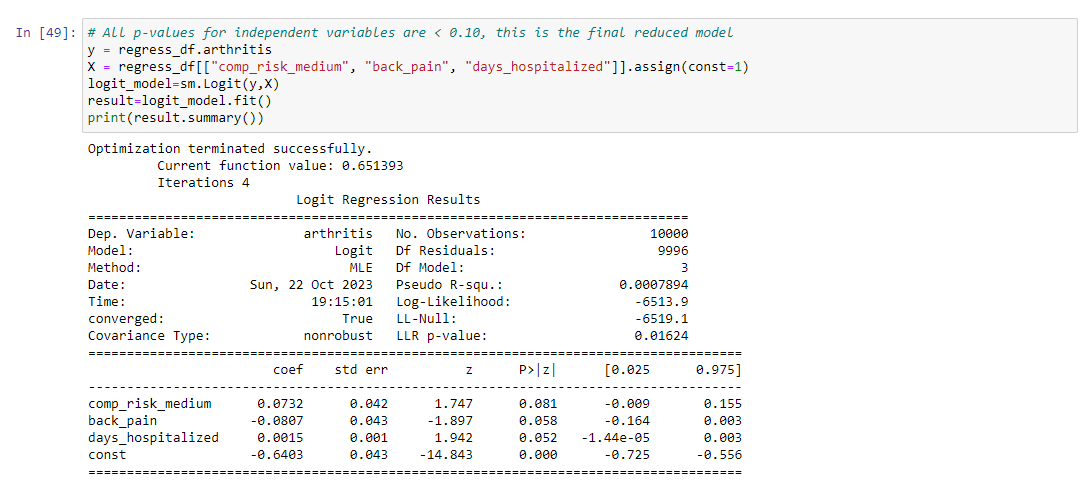


**D3.  Reduced logistic regression model.**

The variables for gender\_nonbinary and vit\_d\_levels were eliminated due to multicollinearity, while the variables for age, gender\_male, initial\_admit\_emerg, initial\_admit\_observ, high\_bp, comp\_risk\_low, overweight, diabetes, and asthma were all eliminated for poor p-values. The variables remaining in the model and having the most impact on the dependent (y) variable of arthritis are:

* complication risk (medium)
* back pain
* days hospitalized

These variables are seen in the final reduced logistic regression model, which has an improved (reduced) LLR p-value compared to the initial model:



**D4. Code used to support the implementation of the logistic regression models.**

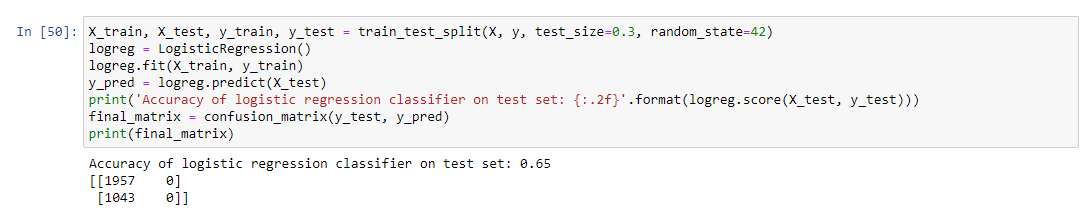
The full code for the project was created in a Jupyter notebook environment, and the notebook is attached in the task submission. Furthermore, a pdf print of the Jupyter notebook used for running the python scripts is also attached. Lastly, a .txt file with the full code used in provided.

**E Analysis of the Dataset using the Reduced Logistic Regression model**

**E1.  Comparison between the initial and reduced logistic regression models**

**Logic of the variable selection technique:** The initial logistic regression model a larger set of variables, not all of which were essential to the model's accuracy. The variables vit\_d\_levels and gender\_nonbinary, were removed due to concerns about multicollinearity. Following their elimination, other variables were systematically removed from the initial model using Backwards Stepwise Elimination. This was done based on the p-value of each remaining variable, and the target threshold was set to a p value below 0.10. A variable's p-value signifies its statistical significance, with lower values indicating greater significance. Non-significant variables were excluded from the model since their presence wasn't necessary. The process of eliminating the variable with the highest (worst) p-value was conducted one at a time, necessitating the rerunning of the model after each elimination. This iterative process continued until all remaining variables had a p-value less than 0.10, indicating their statistical significance. Although a threshold of 0.05 would have been preferable, it would have retained only the age variable. However, the project's requirements mandated the use of both categorical and continuous variables, prompting an adjustment of the threshold to 0.10.

**Model evaluation metric:** The performance of the initial and reduced models can be compared by analyzing their LLR p-values. Similar to p-values in other statistical contexts, a lower LLR p-value signifies a higher likelihood of the observed relationship not being due to chance but rather indicating a real relationship. The initial logistic regression model yielded an LLR p-value of 0.3140, which was too high and indicated that the initial model was not useful in predicting the response variable values. The reduced model in comparison had a significantly lower LLR p-value of 0.01624, after removing all the explanatory variables with high p-values above the chosen threshold. This substantial improvement indicates that the reduced model is more effective in predicting the response variable values compared to the initial model. To determine the reduced model’s accuracy, the following confusion matrix was generated:



This resulting confusion matrix indicates that this logistic regression model has made 1957 correct predictions, and 1043 incorrect predictions. The final resulting model was 65% accurate.

**E2: Model Outputs**

All outputs and calculations, including the confusion matrix are featured in sections D2, D3, and E1.

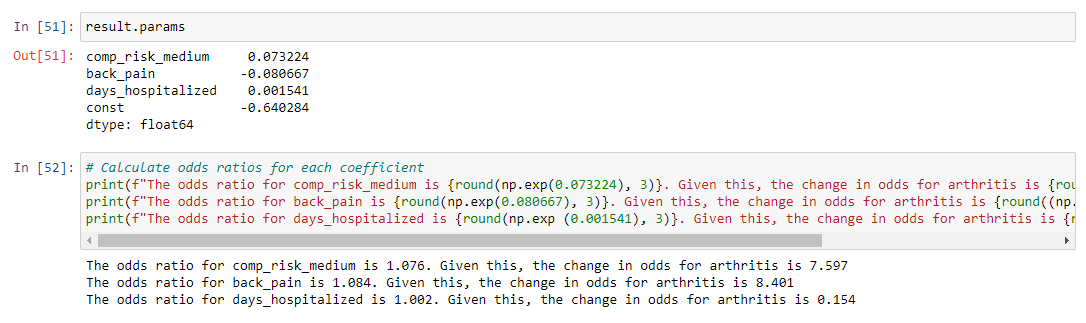
**E3: Model Code**

The code for the generation of the model is showcased in D2 and D3 and is present in the Jupyter notebook as well as the .txt file containing the full code of this assessment.

**Part V: Data Summary and Implications**

**F: Results of the Data Analysis**

**F1. Summary of findings and assumptions**



**Regression equation for the reduced model:**

After conducting the analysis, we created the following logistic regression equation for the reduced model:

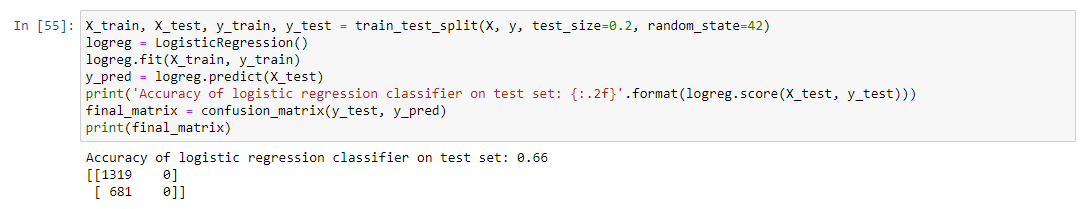
After calculating the odds ratio and using the odds ratio to calculate changes in odds, the following conclusions were drawn about each explanatory variable.

* With all things kept constant, a patient with a medium complication risk has an *increase* in their odds of having arthritis by 7.597%.
* With all things kept constant, a patient with back pain has an *increase* in their odds of having arthritis by 8.4011%.
* With all things kept constant, for one unit increase in days hospitalized, the odds of a patient having arthritis *increase* by 0.154%.

In regards to the statistical significance of this model, the final regression analysis shows an LLR p-value of 0.01624. This value falls below both the alpha threshold of 0.05 and the 0.10 threshold used during the backward stepwise regression for model reduction. This suggests that the model is statistically significant, and the significance is not due to random chance or sampling. The model may lack practical significance due to the low accuracy score of 0.65. One limitation of this analysis was the decision to use an alpha level of 0.10 for eliminating explanatory variables based on their p-values during the regression model reduction process. This decision was made in order to retain both a continuous and categorical variable for the reduced model, and the final p values for the remaining variables were all above 0.05. However, the higher alpha value of 0.10 increases the risk of committing a Type I error, wherein a true null hypothesis is incorrectly rejected in favor of an alternative hypothesis. Another limitation of the dataset was the size. While the dataset has 10000 observations, further testing showed that increasing the data set would be beneficial in improving the model’s accuracy. For every additional 1000 rows the dataset was trained on, the accuracy increased by a factor of 1%.

**F2.  Recommend a course of action based on your results.**

Based on the results of the model, back pain appeared to be the highest indicator of the odds of arthritis. I recommend that patients with back pain and arthritis look into medical services that can be used to treat both conditions. Furthermore, the hospital can investigate the correlation between these conditions and determine whether these are comorbidities. In regards to model accuracy, further testing showed that the model would benefit from a larger data set. The model was trained off of 7000 out of 10000 observations and had an accuracy rate of 65%. The accuracy of the model increased by a factor of 1% for every additional 1000 training rows, as demonstrated below when I increased the testing size to 8000 and got an accuracy rate of 66%.



One recommended course of action would be to increase the size of the data set and then train the model off of a larger number of observations.

**Part VI: Demonstration**

**G. Link to the Panopto Video recording:** <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e668b98c-c5ce-42b4-93d9-b0a40001a3de#>

**H. Sources for third party code:**

<https://intellipaat.com/community/68715/converting-statsmodels-summary-object-to-pandas-dataframe>

<https://stats.stackexchange.com/questions/463324/logistic-regression-failed-in-statsmodel-but-works-in-sklearn-breast-cancer-dat>

<https://thepythonguru.com/python-string-formatting/>

<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

<https://www.kaggle.com/code/alexnystrom/hospital-data-logistic-regression>

<https://www.kaggle.com/code/marshuu/logistic-regression-analysis-breast-cancer>

<https://www.kaggle.com/code/miguelrodriguezolmos/plot-the-logistic-map-with-python-matplotlib/notebook>

<https://www.kaggle.com/code/neisha/heart-disease-prediction-using-logistic-regression>

<https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial/notebook>

<https://www.kaggle.com/code/vipulgandhi/linear-regression/notebook>

<https://www.w3resource.com/pandas/dataframe/dataframe-itertuples.php>

**I. Source References**

Zach. (2020, October 13). The 6 Assumptions of Logistic Regression (With Examples). Statology. Statology.org/assumptions-of-logistic-regression/

Kohli, S. (2019) Understanding a classification report for your machine learning model, Medium. Medium. Available at: <https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>.

Srinidhi, S. (2019) Backward elimination for feature selection in Machine Learning, Medium. Available at: https://towardsdatascience.com/backward-elimination-for-feature-selection-in-machine-learning-c6a3a8f8cef4 (Accessed: 22 October 2023).

Brownlee, J. (2020) Ordinal and one-hot encodings for Categorical Data, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/ (Accessed: 22 October 2023).

Li, S. (2019) Building a logistic regression in Python, step by step, Medium. Available at: <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8> (Accessed: 22 October 2023).

K., D. (2023) A simple interpretation of logistic regression coefficients, Medium. Available at: https://towardsdatascience.com/a-simple-interpretation-of-logistic-regression-coefficients-e3a40a62e8cf (Accessed: 22 October 2023).