Research Question:

The research question that I want to examine is, "Which factors most significantly contribute to patients diagnosed with back pain?" This is a broad topic, but Back painis a problem that affects millions around the world. This is important because this information could aid hospitals in treating patients with unique conditions and could help them get the treatment they need.

The objective of this study is to determine the independent and explanatory variables in the dataset that possess the greatest statistical influence on the dependent variable, or anxiety, utilizing a logistic regression model. Through identifying these variables, the hospital may better adapt its treatment of particular individuals with anxiety to increase their care/treatment efficiency according to each patient's specific risk factors to enhance outcomes for all parties.

4 assumptions

Linearity of logit:

There is a linear relationship between the independent variables and the dependent variable's log chances. This indicates that a linear relationship between the independent variables and the outcome's logit (log-odds) is assumed by the logistic regression model.

It is crucial to verify this assumption by looking at the dependent variable's logit transformation and the linearity of the independent variables.

Error independence:

The residuals, or mistakes, of the model should be independent among observations if observations are to be independent of one another. For the model estimates to be valid, this assumption is essential.

This practically means that autocorrelation should not exist, which is especially important for time series or clustered data.

Incomplete Multicollinearity:

There shouldn't be a perfect correlation between the independent variables. Since it is impossible to distinguish between the individual impacts of the correlated variables, perfect (or nearly perfect) multicollinearity complicates the estimation of the model's coefficients.

Variance inflation factor (VIF) or correlation matrices can be used to test for multicollinearity.

Binary Result:

In logistic regression, the dependent variable needs to be binary, meaning it can have only two possible values, typically represented as 0 and 1. The purpose of logistic regression is to estimate the probability of one of two outcomes by modeling binary outcomes.

Furthermore, although, unlike linear regression, logistic regression does not presuppose homoscedasticity or normality of the residuals, it is still a good idea to look for significant outliers and leverage points, as these can have an excessive impact on the model.

Python offers two advantages when it comes to supporting different stages of the analysis: its vast library and frameworks, as well as its ability to visualize data. We may utilize NumPy and Pandas, plug-ins created specifically for data science, because of the large libraries and frameworks available. Data exploration, pattern recognition, and insight discovery are made easier with support for data visualization. Using Pandas, you can handle a dataset in a manner similar to that of a large table or spreadsheet. NumPy facilitates the execution of particular mathematical operations or the assignment of specific values inside the dataset. MatPlotLib and Seaborn both offer graphing capabilities. Many helpful features for the multiple regression model are provided by SciPy's statistics models, such as the ability to graph residuals and use the variance inflation factor to test for problems such as multicolinearity. We can utilize Sklearn's preprocessing to change our data if needed.

* Explain why multiple logistic regression (task 2) is an appropriate technique to use for analyzing the research question summarized in Part I.

Logistic regression helps clarify the link between a binary dependent variable and several independent variables, which can be either continuous or categorical. This dependent variable often indicates outcomes like True/False or Yes/No but can also include other binary categories like malignant or benign. In this instance, I aim to investigate the relationship between various factors and a patient's diagnosis of anxiety, a binary variable with possible values of Yes and No. Since the outcome of interest is binary, logistic regression is an effective method for analyzing how different explanatory variables impact it.

Data Prep

The dataset still hasn't incorporated the fixes I made in d206, so I'll be bringing in that code to help clean the data. For example, instead of being recorded as qualitative strings, zip codes are treated as whole integers and stored as int64, which removes leading zeros.

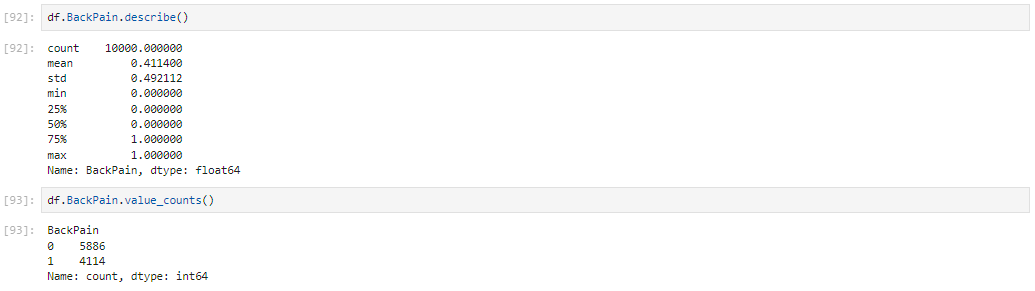
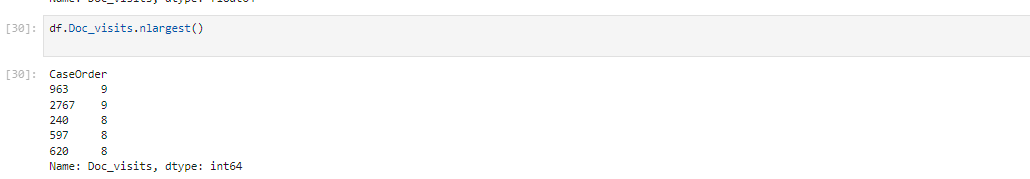
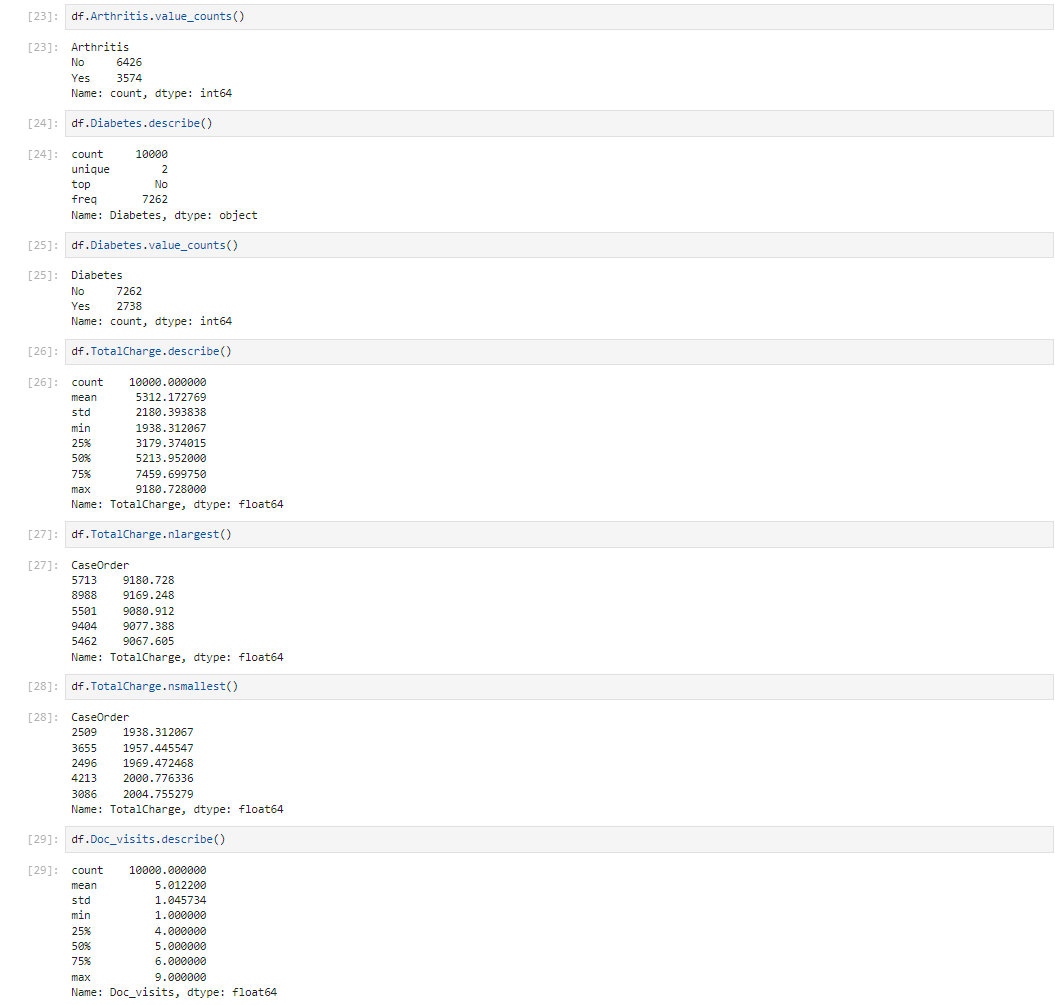
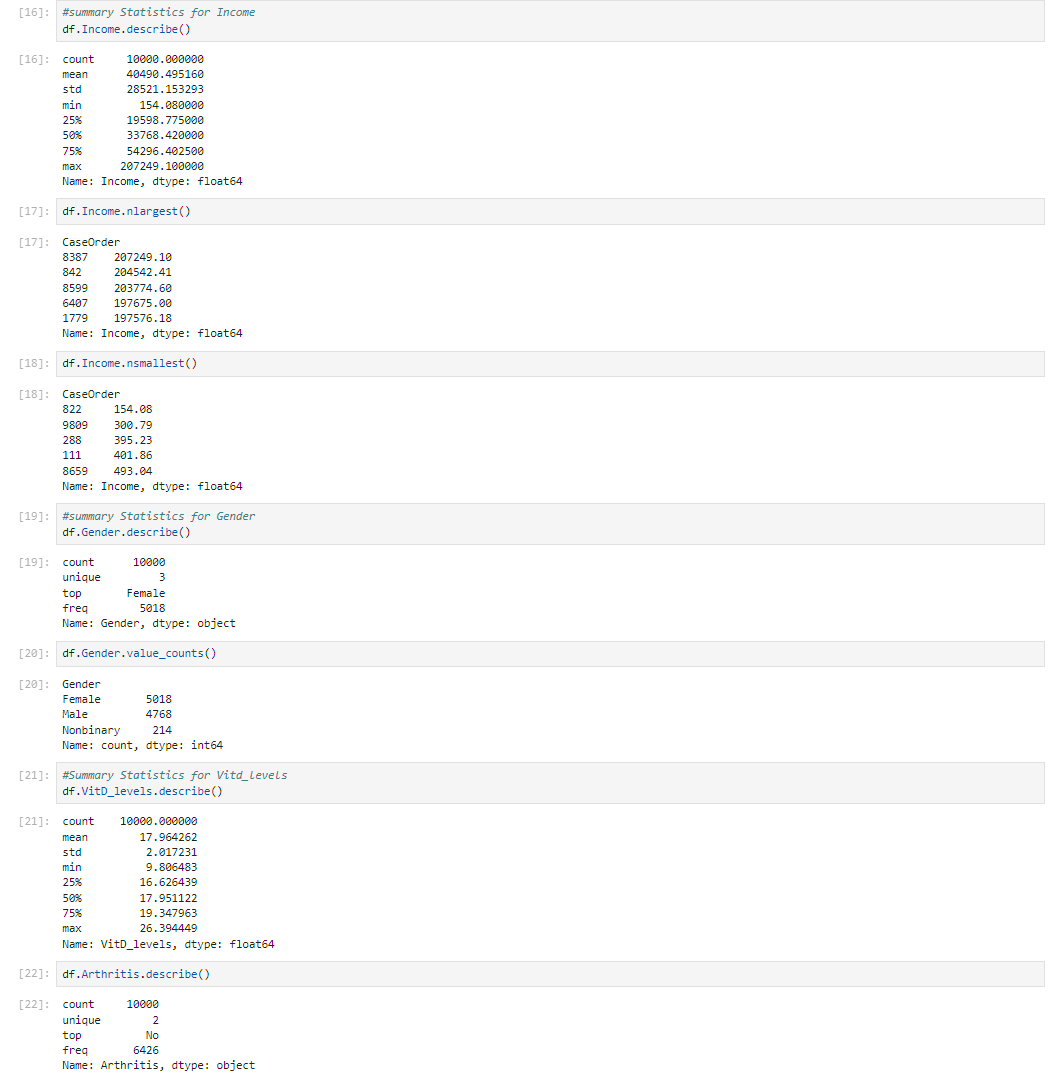
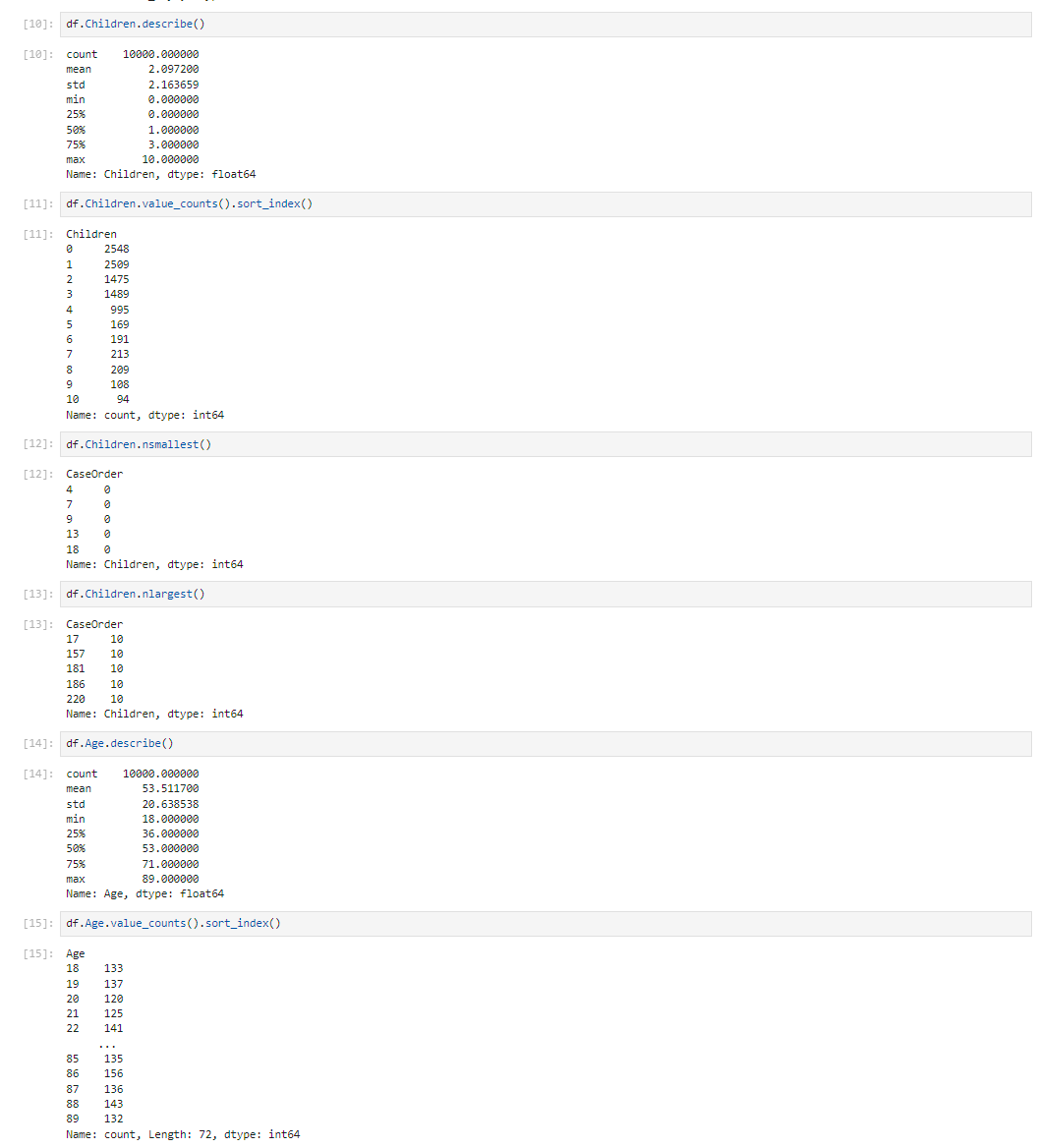
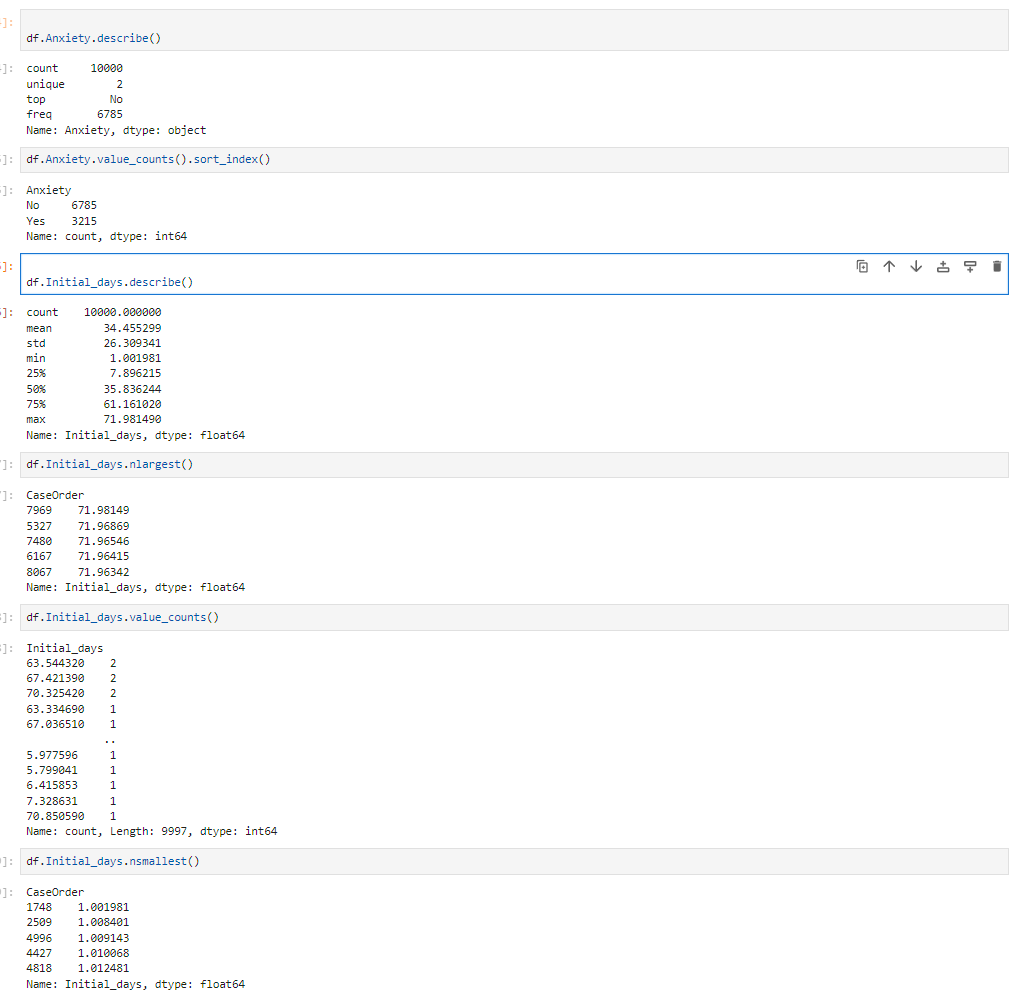
To perform multiple regression analysis, categorical and boolean data need to be in numerical form rather than strings. Converting boolean values is straightforward, with true and false becoming 1 and 0, respectively. However, handling ordinal (where the order of categories matters, like "small," "medium," and "large") and nominal (where the order is irrelevant) categorical variables requires different methods.

This dataset contains more than just boolean categorical data types. Only the columns related to survey scores are ordinal, indicating a range from "most important" to "least important." However, the original numerical form of these scores might not make intuitive sense. To create a more logical pattern—where a lower score means more importance—I plan to remap these variables. These columns will be remapped and converted to an ordered categorical datatype.

For the nominal categorical columns used in this study, I'll create dummy columns using one-hot encoding. This method converts categorical data into a binary numeric format. For instance, the gender column with values of male, female, and nonbinary can be represented numerically by adding two more columns. A 1 in the first column indicates the patient is female, while a 1 in the second column indicates nonbinary. If both columns are 0, the patient is male. This approach leverages Pandas' get\_dummies() function to efficiently handle text data for multiple regression analysis.

These procedures are the main adjustments needed to prepare the dataset for multiple regression analysis. Additional checks will ensure the data is ready, including using info() to confirm no null values, value\_counts() to review all column values, and describe() to summarize numeric columns.

The dataset needs to be modified primarily in these ways in order to be ready for logistic regression analysis. To make sure the data is available, other operations will be performed, like description() to show summary statistics for numeric columns, value\_counts() to check all the values in a column, and info() to make sure there are no null values and to confirm the datatypes for each column.

Summary Statistics



I will be using the same variables from task 1. However, the dependent y variable will be anxiety and Initial days will no longer be one of the independent variables tested. I used my data cleaning/data transformation code/steps from d208 task 1 submission for this research.

Dependent Variable: - BackPain:Of the participants in the dataset, 41% experience chronic back pain discomfort, but the remaining 59% do not.

Anxiety:Of the participants in the dataset, 32% experience anxiety, but the remaining 68% do not.

Independent Variables-

Gender: According to the data, over half of the patients are female, around 48% are male, and slightly more than 2% identify as nonbinary.

VitD\_levels: These numbers show a mean close to 18 and a standard deviation slightly over 2. With an interquartile range of 16.6 to 19.4, the lowest and highest values differ from the mean by fewer than four standard deviations. This implies that the distribution is roughly normal.

Arthritis: In this dataset, 35% of hospitalized patients have arthritis.

Diabetes: In this dataset, diabetes has been identified in 27% of the individuals.

TotalCharge: The average daily rate for patients, excluding "additional" charges, is a little over $5,300. With a standard deviation of roughly $2,200, the minimum fee of approximately $1,900 is nearly 1.5 standard deviations below the mean. The maximum fee, around $9,200, is about 1.8 standard deviations above the mean. This suggests a slight leftward skew, indicating some deviation from normality, although the distribution is generally concentrated around the mean. A quick look at the highest values in the dataset confirms that the maximum fee is within a normal range.

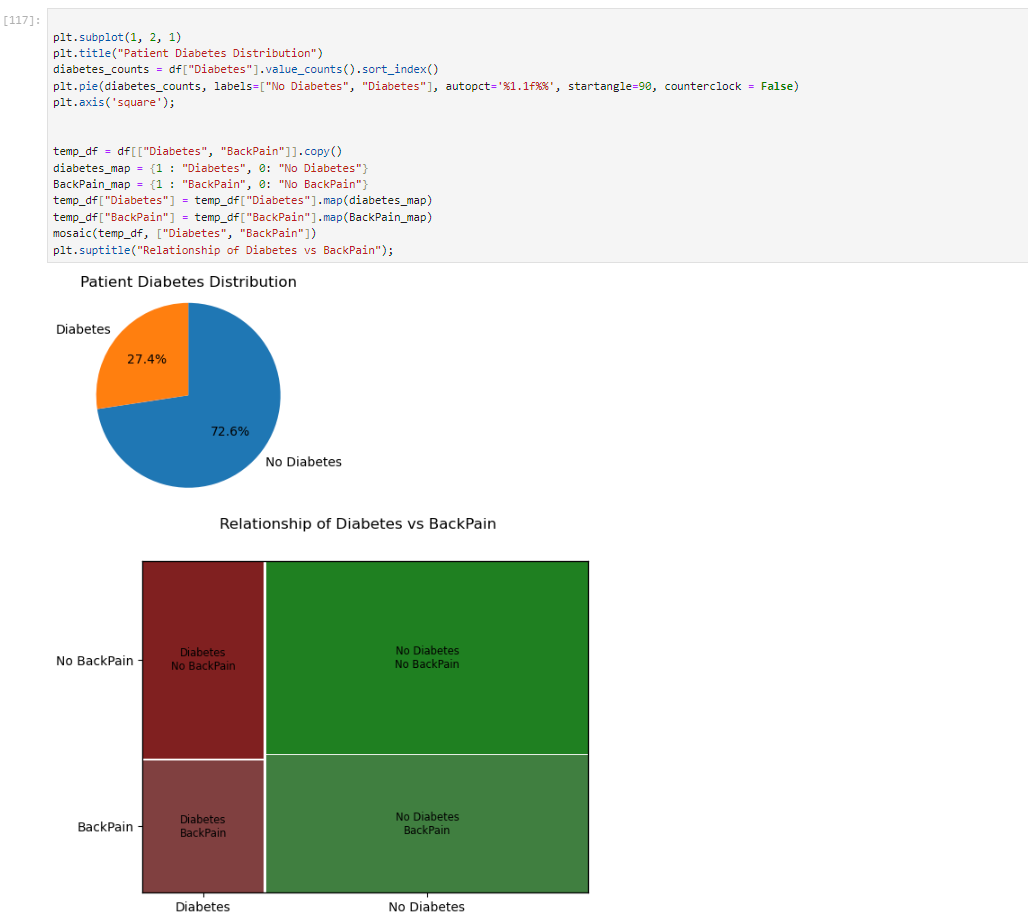
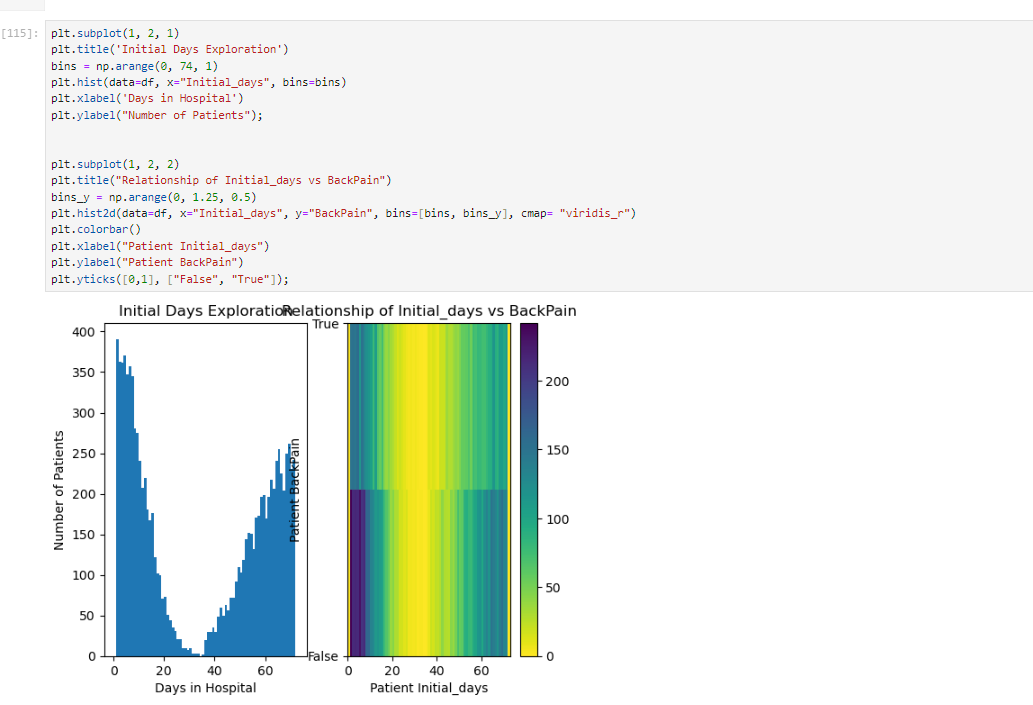
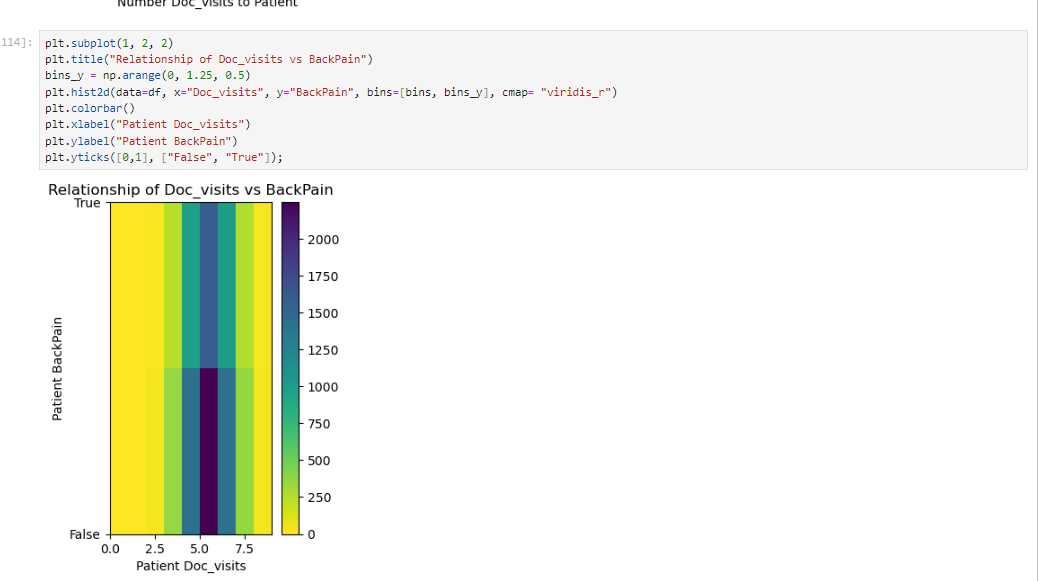
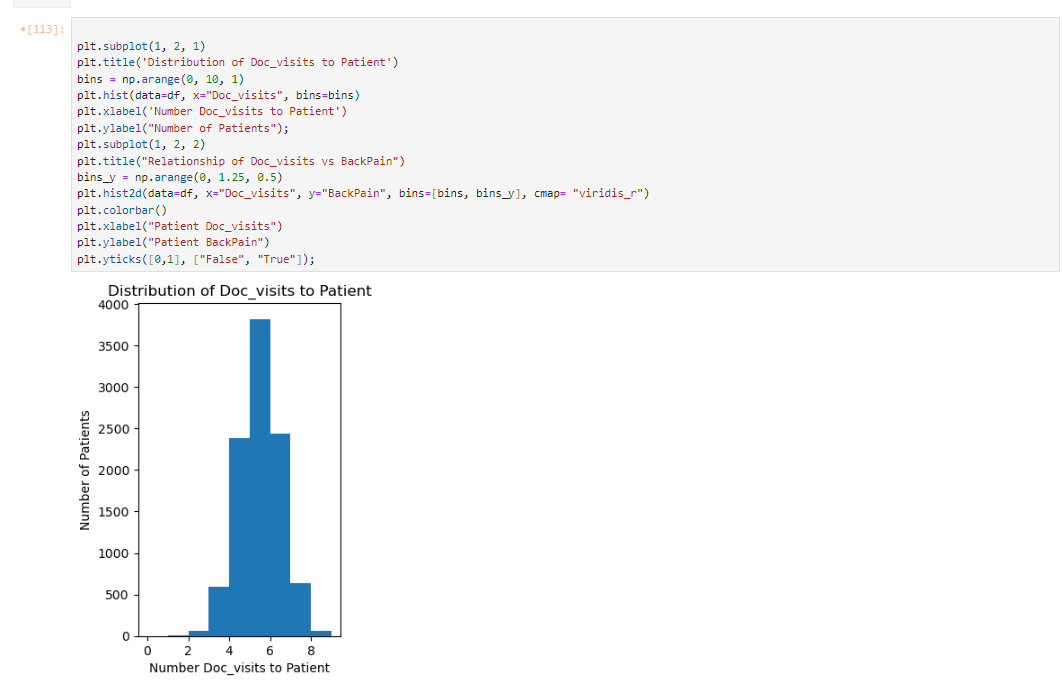
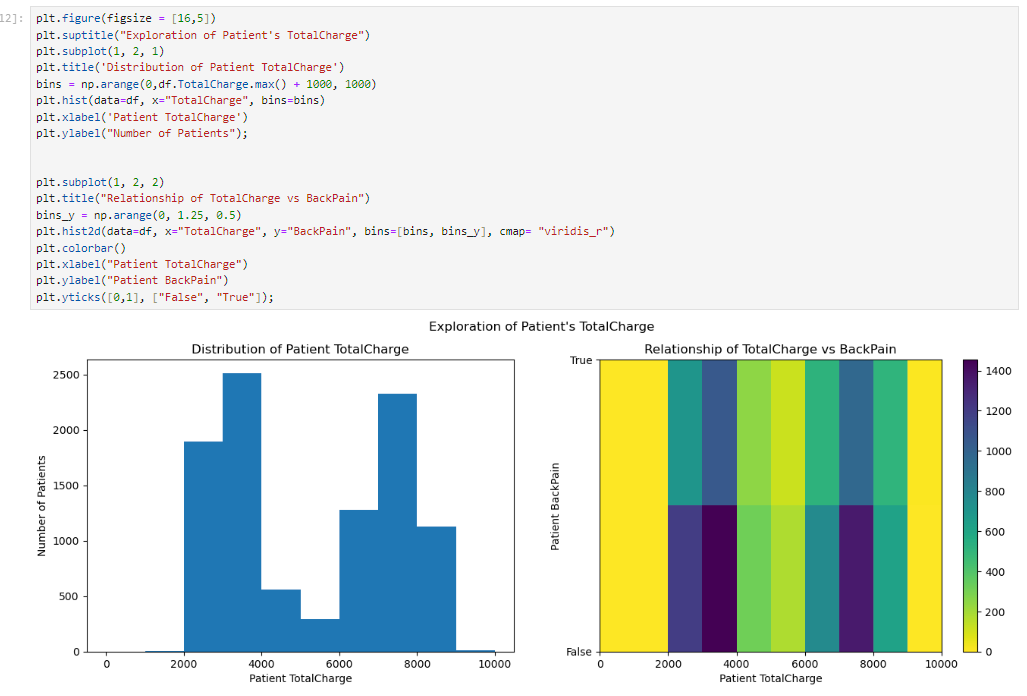
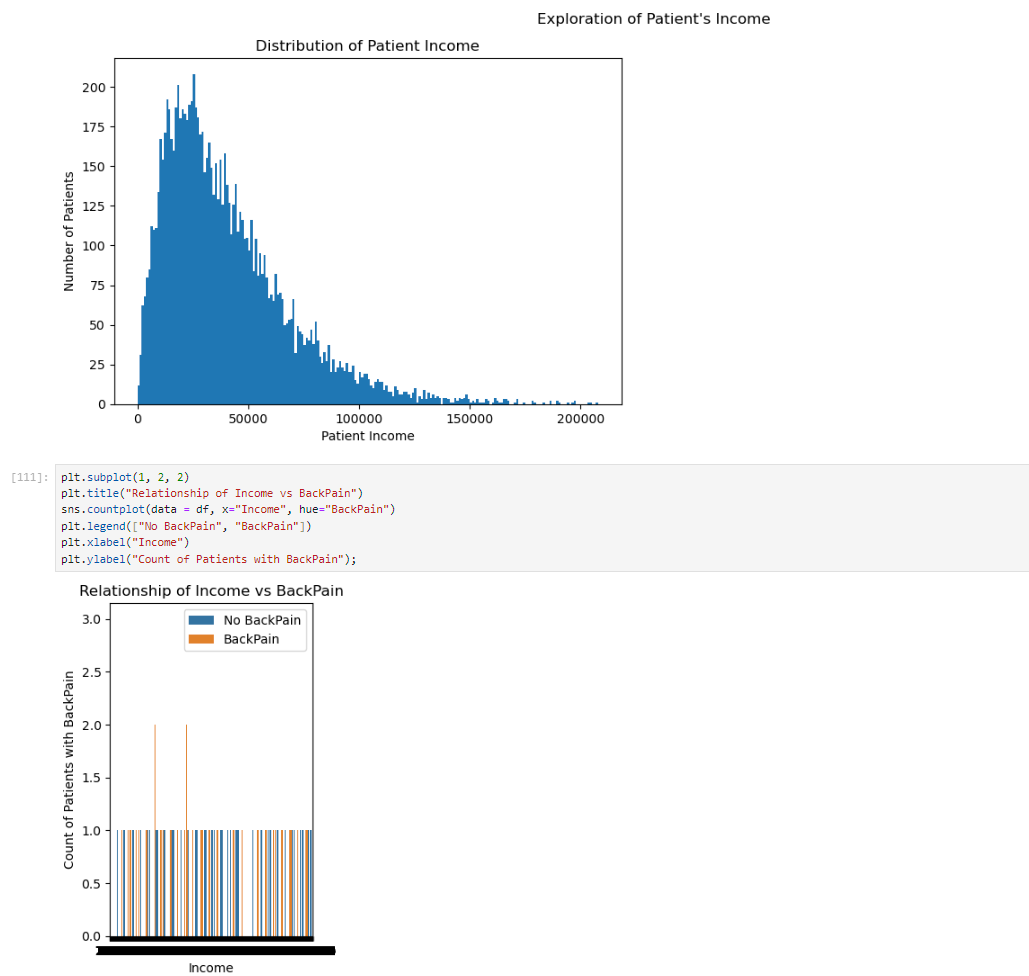
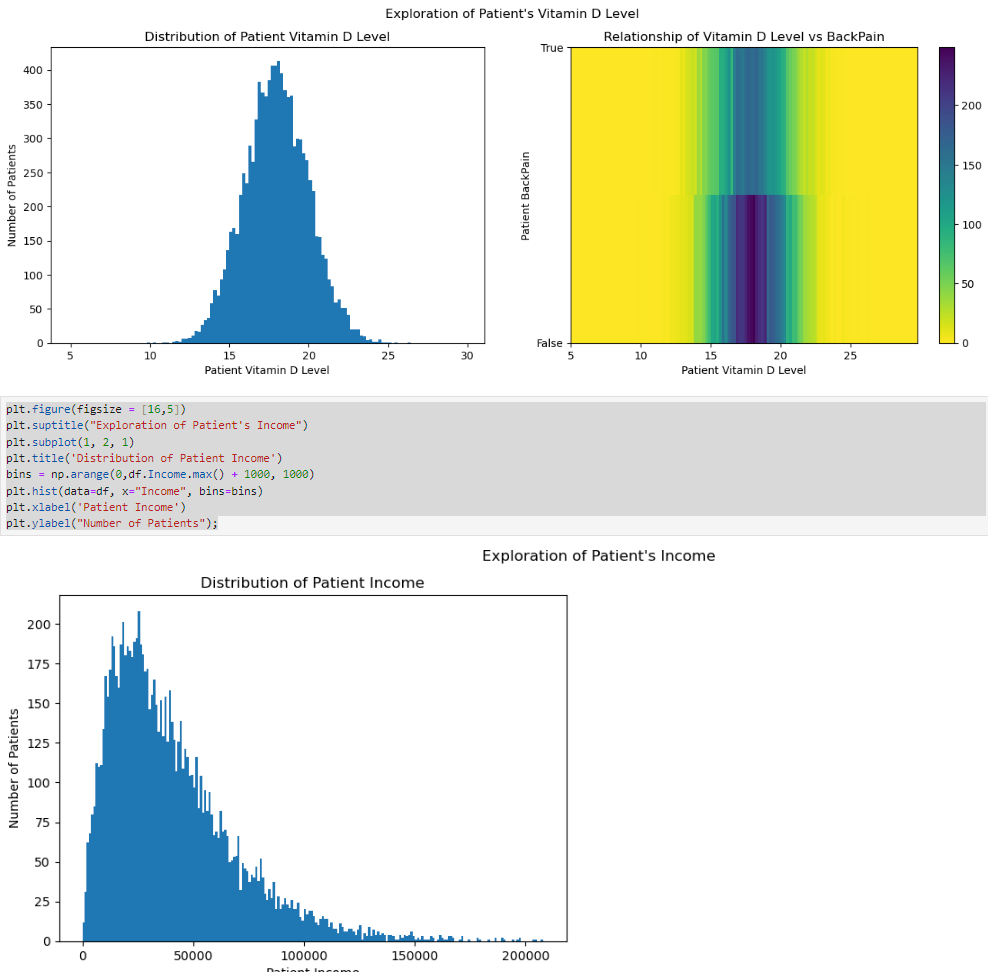
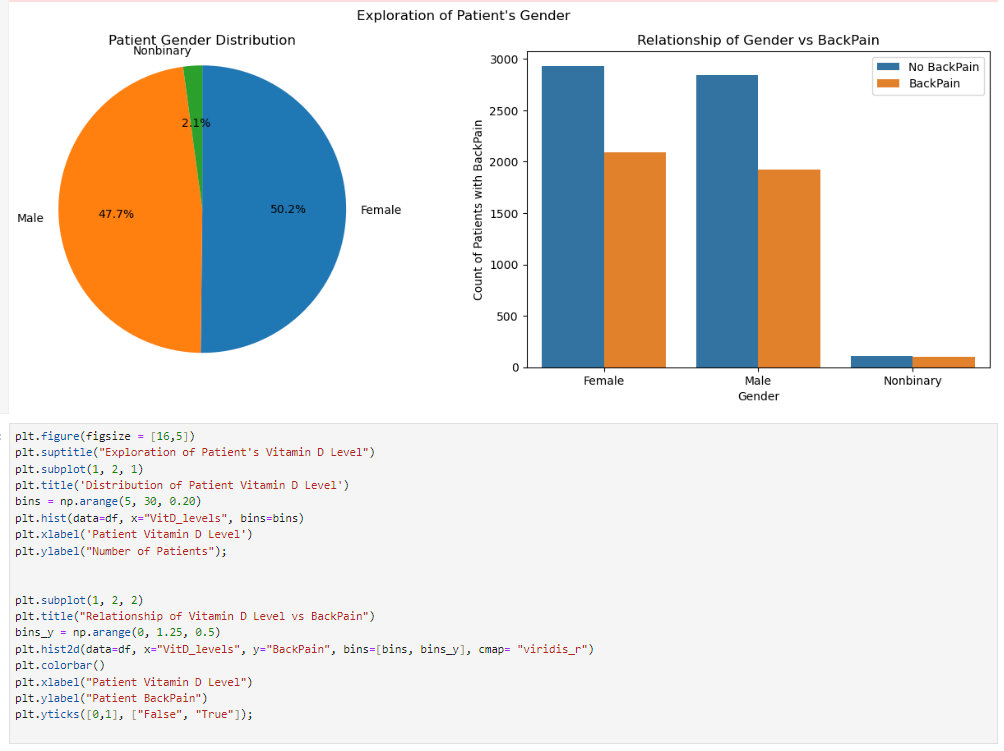
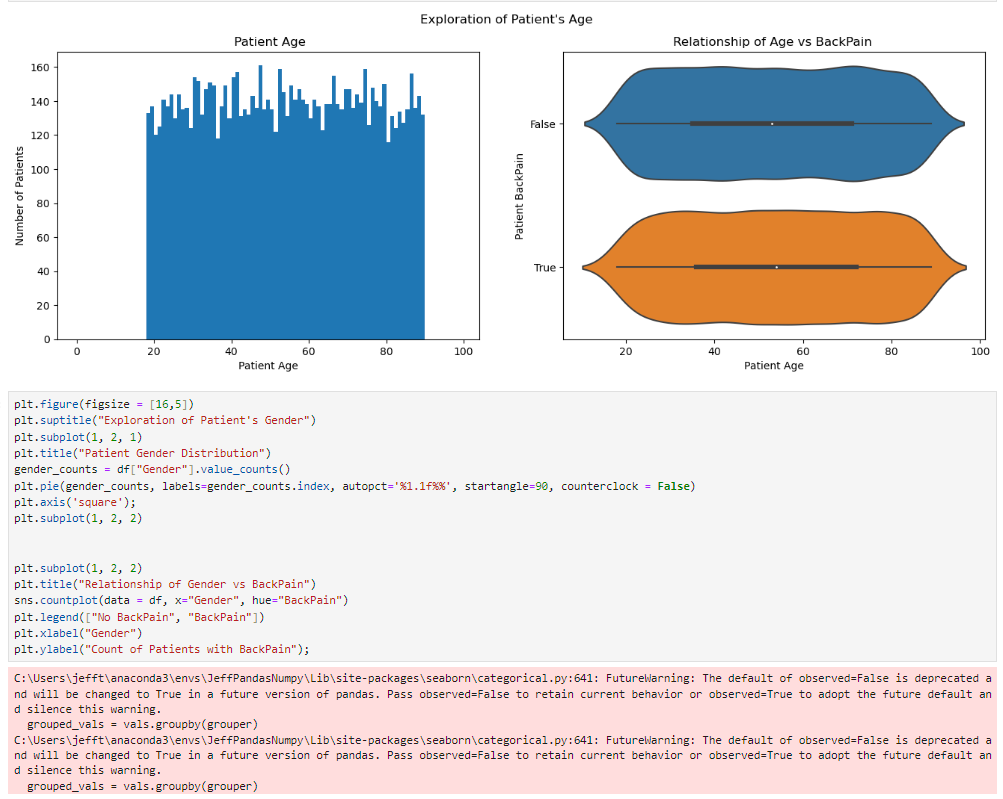
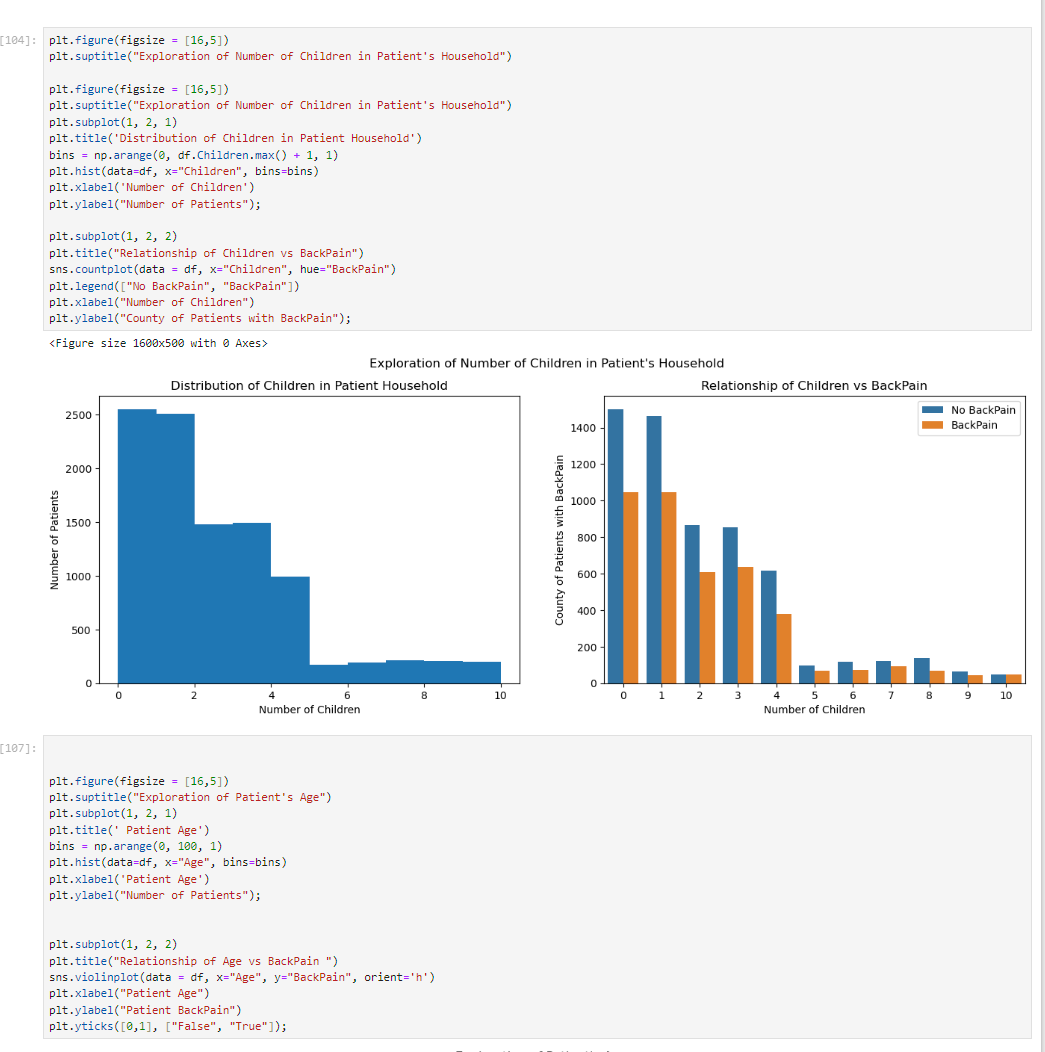
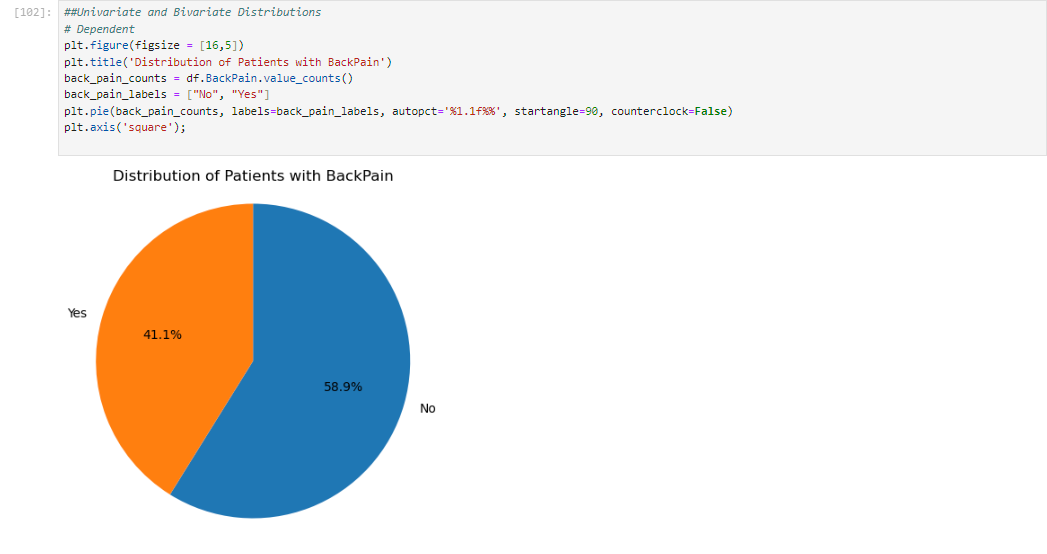
Doc\_visits: The dataset indicates that slightly over 50% of hospitalized patients are admitted for emergencies. This aligns with our expectations, or is possibly even lower than expected, considering that hospital visits often result from emergencies, which we try to avoid. Almost half of the patients are hospitalized for observation or elective surgery.

Age: According to the overall statistics, all our patients are over the age of 18. Since children are admitted to hospitals, our data does not include all hospitalized patients. This omission is important because it raises the possibility that this particular subset of the general population is not represented in our study. If I could, I would address this, but the provided dataset limits my options.

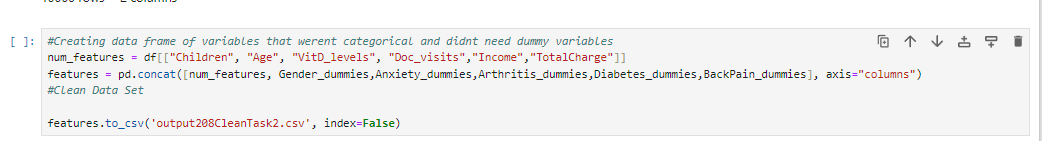
Income: The high end of the income data could be considered an outlier because it is many standard deviations above the mean, whereas the low end is not an outlier since it is within two standard deviations of the mean (within 1.5, even). Although an annual income of $210,000 is significantly above the mean, a sizable portion of the population earns at or above this level. This data is plausible because, with income, there is a floor (0) but no ceiling on the possible values.

Children: The data reveals distinct groups of people with children. Specifically, age ranges from 0-4, 5-6, 7-8, and 9-10 show similar frequencies.

Univariate and Bivariate Exploration



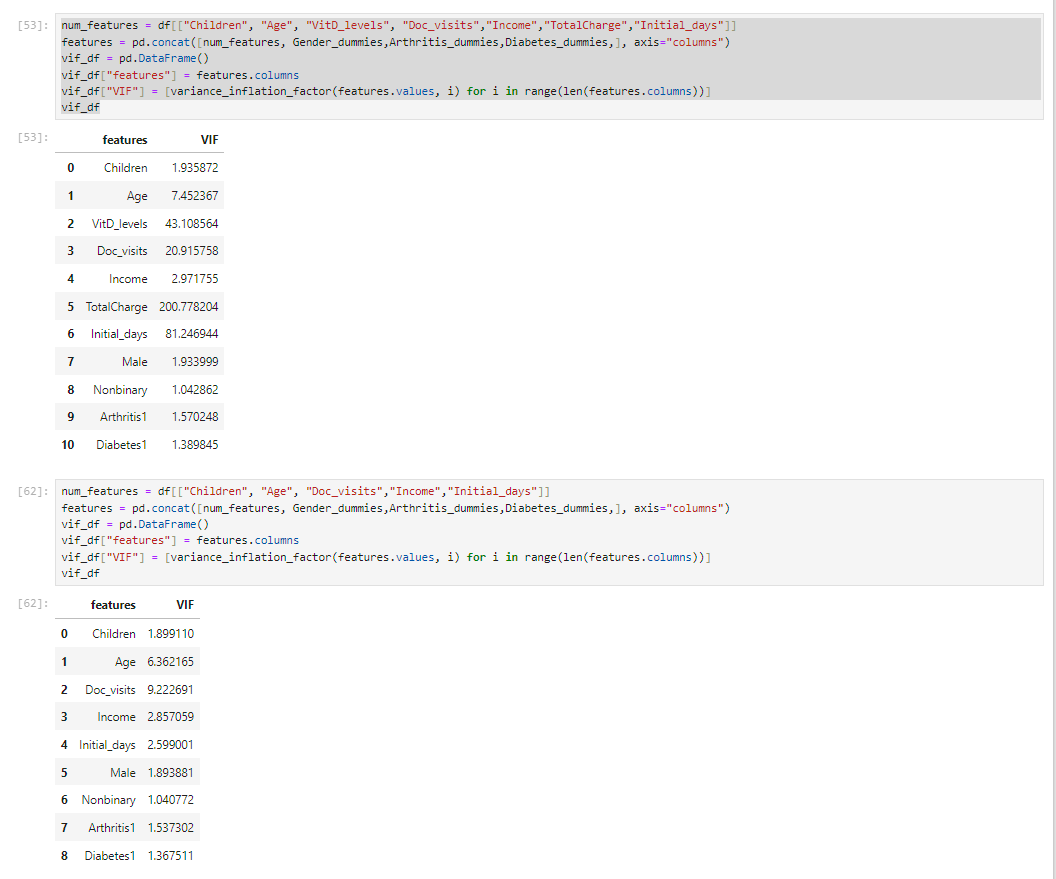
Copy of Prepared Data Set



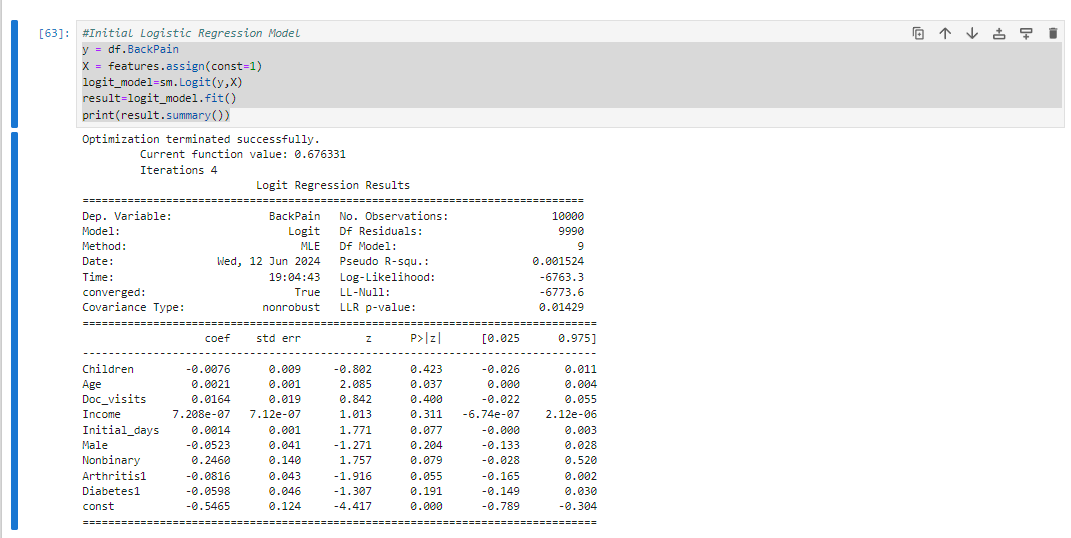
The cleaned copy will be submitted with the rest of the report. The code above shows how it was done.

D1.Initial Logistic Regression Model

I have to check the Variance Inflation Factor (VIF) to make sure there are no variables causing this problem before generating the Initial Logistic Regression Model.

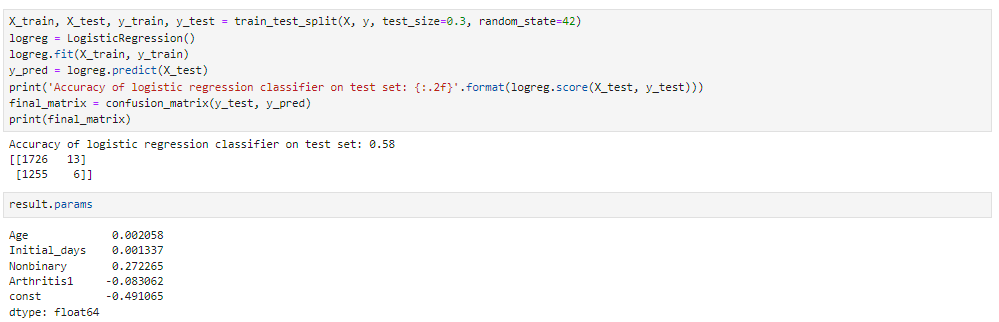
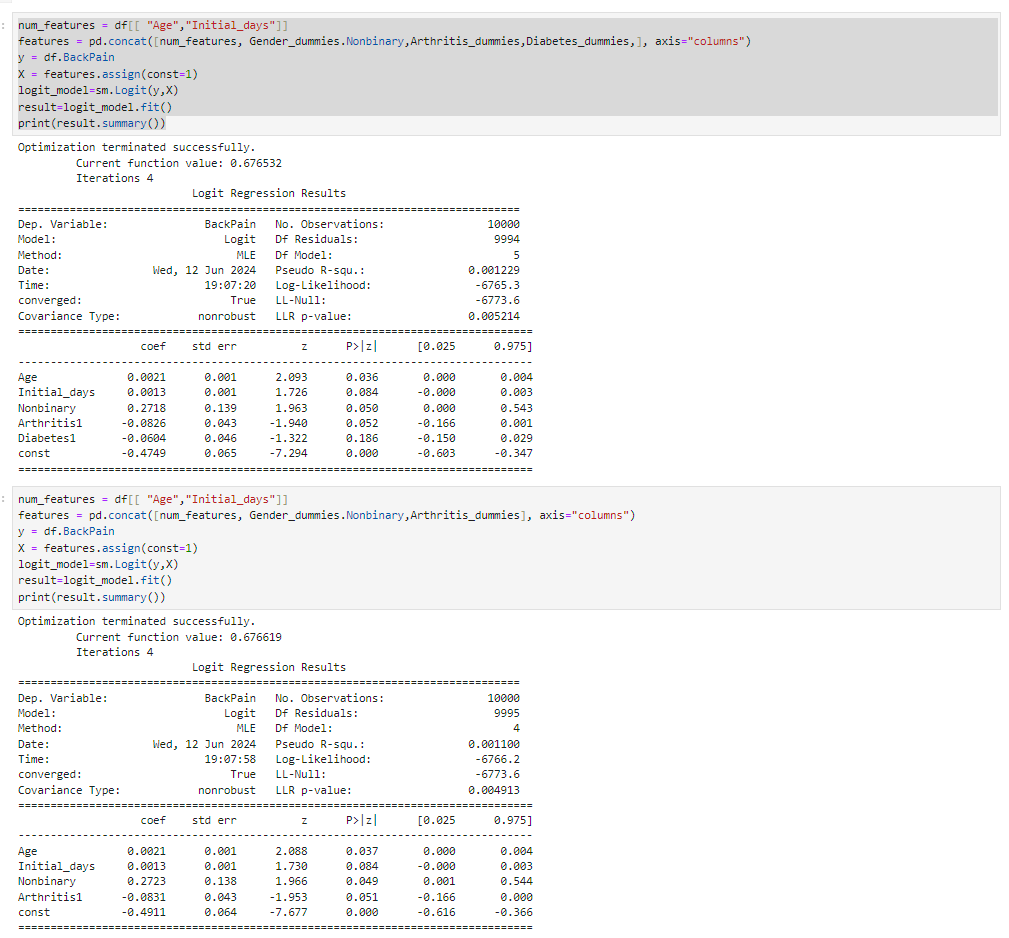
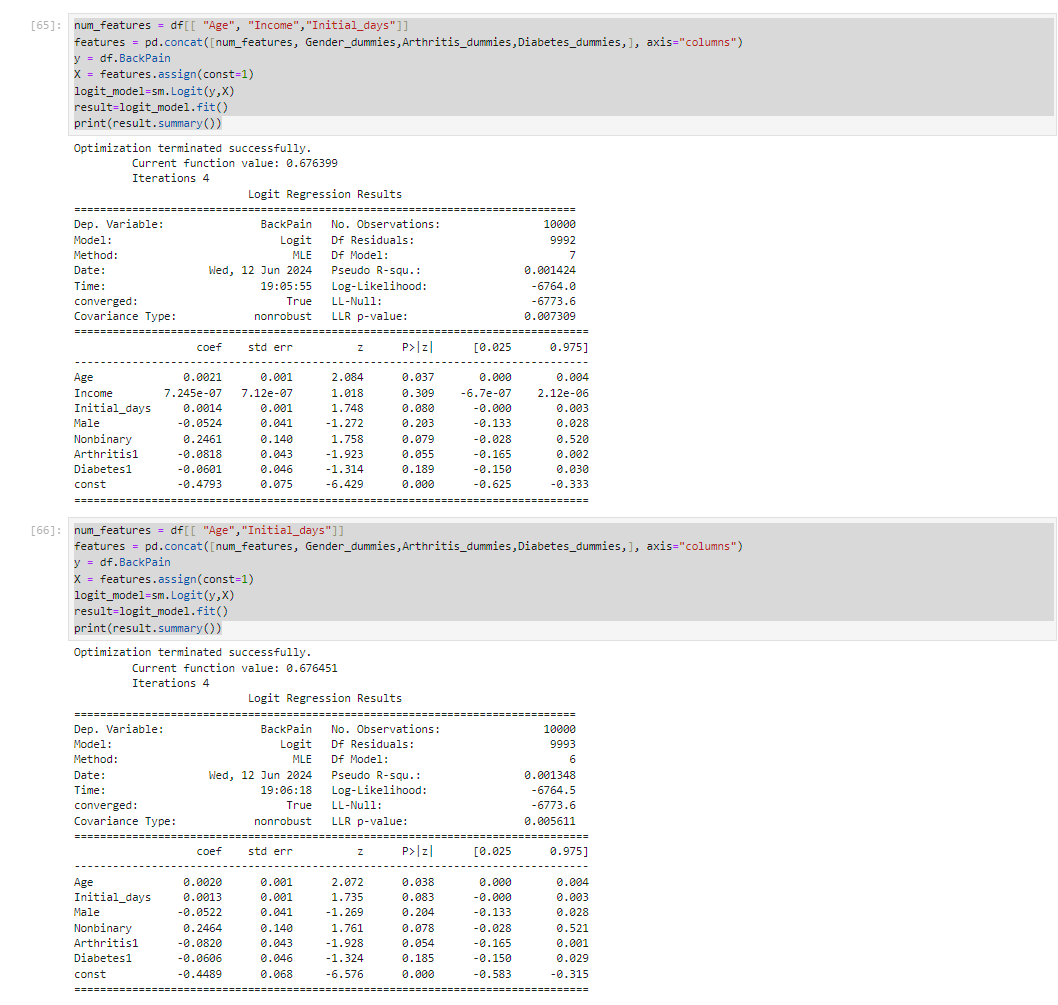
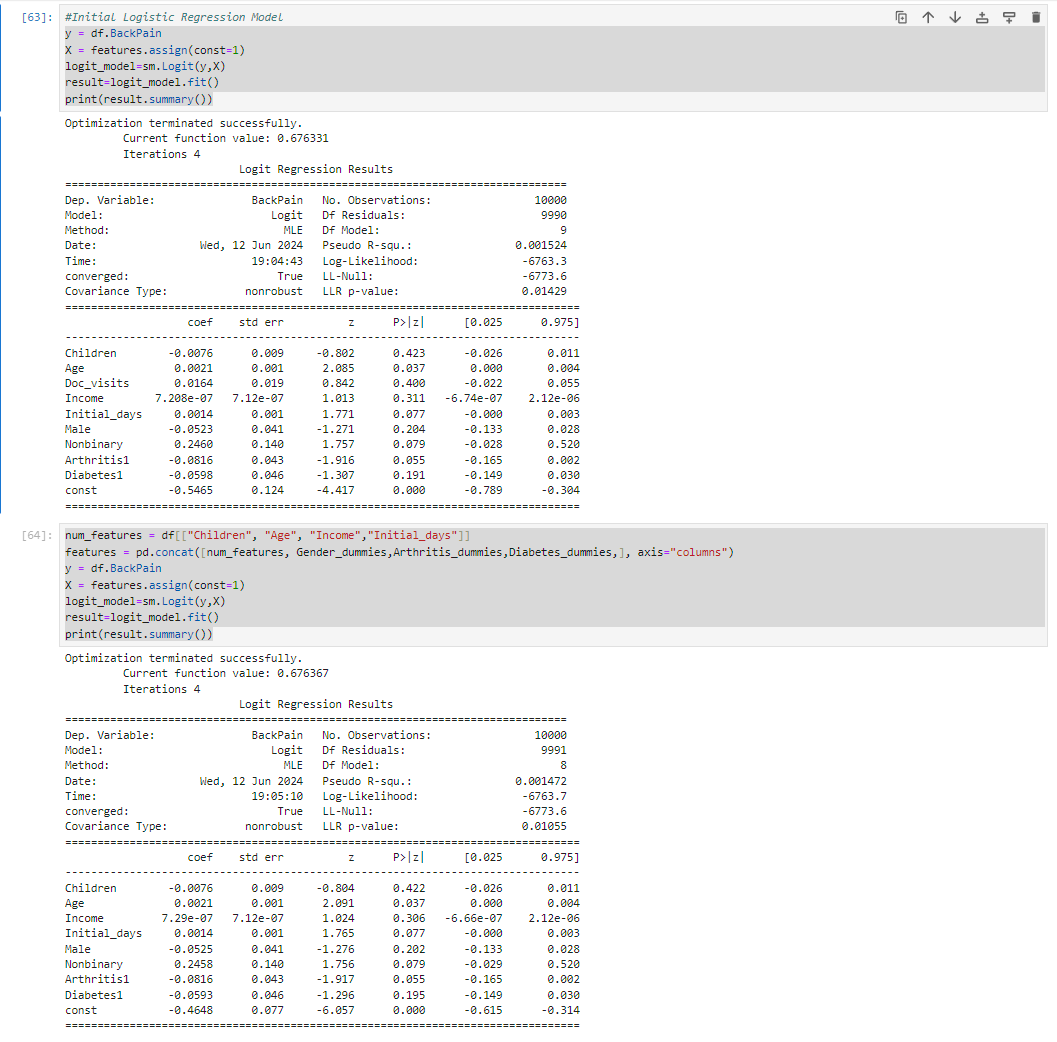


From the above VIF tables, we can observe a particularly high multicollinearity with TotalCharge, so it is removed from the dataset. By removing TotalCharge from the initial model, the model can further focus on variables that have statistical significance to the dependent variable BackPain



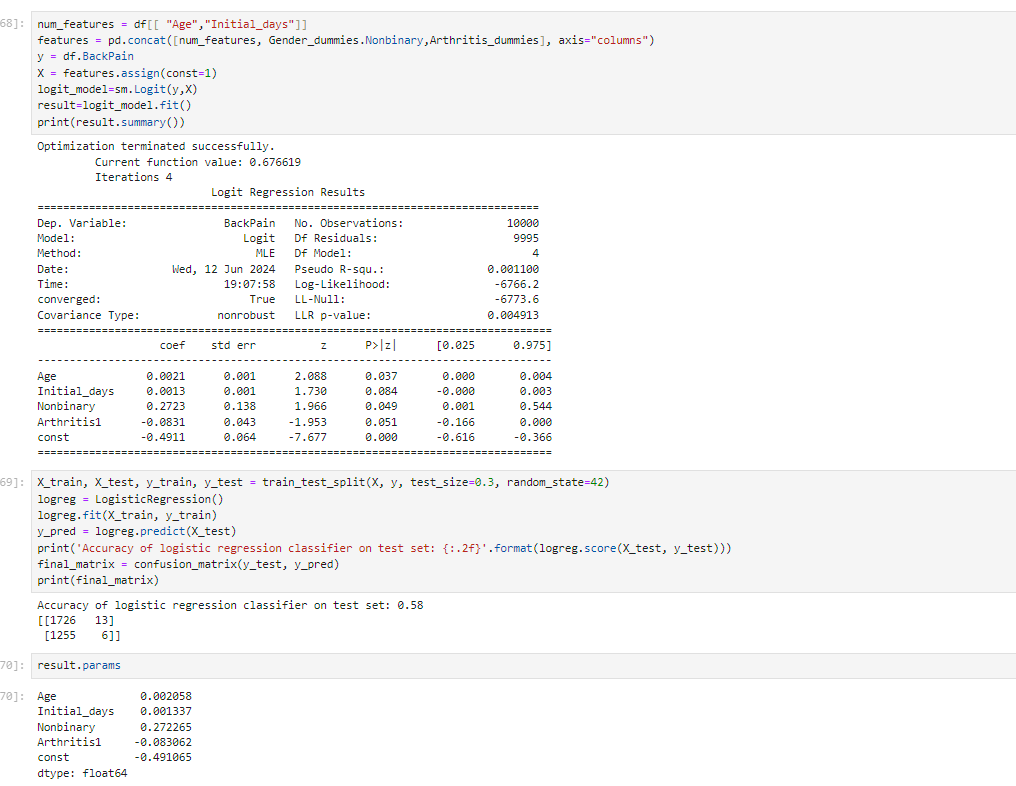
D2:Reduction Justification

Here I eliminated any p values from the model from my logistic models that were higher than.10. If I stuck with p values <.05, I would fail to meet the rubric requirements of keeping a continuous and categorical variable, so I adjusted my requirements accordingly.



D3: Reduced Logistic Regression Model.

The variables remaining in the final reduced model were age, initial days,NonBinary, and arthritis. You can see this above in the model.The original logistic regression model had an LLR p-value of 0.01429, which is below the 0.05 threshold, indicating it’s useful for predicting the response variable. However, several explanatory variables in this model had high p-values, suggesting no correlation with the response variable. After removing these unimportant variables, the simplified model has an LLR p-value of 0.004913, significantly lower than the initial model’s. This improvement indicates the reduced model is more effective in predicting the response variable's values since it no longer includes extraneous or irrelevant variables.



E1:Analysis of Logistic Regression Models

The initial logistic regression model included numerous variables, many of which were not particularly significant. The Variable TotalCharge was removed due to concerns about multicollinearity. Following the elimination of this variable, others were removed from the model using a backward, stepwise elimination process based on their p-values. A variable's p-value indicates its statistical significance, with lower values being more significant. Variables with high p-values, indicating low statistical significance, were eliminated one by one, with the model being re-evaluated after each removal. This process continued until all remaining variables had p-values less than 0.10, indicating statistical significance. Although a threshold of 0.05 is generally preferred, adhering to this would have left only the age variable in the model. The project rubric required the inclusion of both categorical and continuous variables, necessitating the use of a 0.10 threshold. The confusion matrix revealed that the model made 1726 correct predictions and 1255 incorrect ones. This information was comprehended and understood thanks to Susan Li’s Logistic Regression in Python and Statistics.

:Results of Data Analysis

* Age has an odds ratio of 1.0021. In light of this, the probabilities of back pain changing are 0.206.
* For Initial\_days, the odds ratio is 1.0013. In light of this, the probabilities of back pain changing are 0.1338.
* Nonbinary has an odds ratio of 1.3129. In light of this, the probabilities of back pain changing are 31.2935.
* For Arthritis1, the odds ratio is 0.9203. In light of this, the probabilities of back pain changing are -7.9706

Equation

Regression Equation :(-.4911 + 0.0021(Age) +0.0013(Initial\_days)+0.2723(Nonbinary) - 0.0831(Arthritis))

One limitation of this analysis is the age variable, as it excludes patients under 18 and over 89 due to the absence of data for these age groups. Consequently, no conclusions can be drawn for these age ranges. A more comprehensive dataset that includes a wider age range would be more informative.

Another significant limitation is the size of the dataset, particularly the training set used for the logistic regression model. The analysis found that the explanatory variables had a minimal impact on predicting anxiety, resulting in the model rarely predicting anxiety. This limitation could potentially be addressed by expanding the dataset, either by collecting additional patient data or using the SMOTE (Synthetic Minority Oversampling Technique) algorithm. SMOTE, as explained by Susan Li on Towards Data Science, generates synthetic samples from the minority class by selecting one of the k-nearest neighbors and creating a similar but slightly altered observation. Given the hospital system's capacity to provide more patient data, this approach is preferable to generating synthetic data.

F:Reccomend Action

A significant portion of this dataset is too general to be useful for decisions on patient care and treatment. Except in the most exceptional and unique cases, many data points are unrelated to therapy, and most relevant data pertains only to current diagnoses in a binary yes/no format. As stated in D206, the original goal of this dataset was to help identify reasons for patient readmissions. Achieving this objective likely requires a more targeted and detailed dataset, focusing on specific patient healthcare information. Including data on treatment types, initial complaints, and the number of nurse visits would probably be more beneficial than the variables currently provided, despite potential privacy and security concerns related to healthcare information.

Panoptop Recording

Included with Submission

Code References:

Jeffery Bills D208 Task 1 performance assessment

[Susan Li's Logistic Regression in Python Tutorial in Towards Data Science, 2017](https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8#:~:text=Over%2Dsampling%20using%20SMOTE)

[Statology, 2021](https://www.statology.org/assumptions-of-logistic-regression/)