Research Question:

I aim to explore the question: "Which patient variables have the greatest impact on hospital stays?" While this question is open ended, it is crucial because identifying the key factors that lead to longer hospital stays can help hospitals better manage patients and their specific risk factors. By addressing these factors, hospitals can expedite patient recovery and save money and resources, which can be directed to other areas.

## A2: Objectives and Goals of Analysis

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 length of hospital stay, utilizing a multiple regression model. Through identifying these variables, the hospital may better adapt its treatment of particular individuals to each patient's specific risk factors to enhance outcomes for all parties. When more significant contributing factors are found, resources could be directed toward them, expediting patient recovery and reducing hospital stays while reducing expenses on lengthier hospital stays.

Part II: Method Justification

B. Describe multiple linear regression methods by doing the following:

1. Summarize four assumptions of a multiple linear regression model

Linearity

According to the linearity assumption, the dependent variable (outcome) and the independent variables (predictors) have a straight-line connection. This suggests that the relationship between the change in the independent factors and the change in the dependent variable is proportionate.

Biased estimations could result from the linear model's inability to correctly represent the relationship between the variables if this premise is broken. Scatterplots of each predictor against the outcome variable and residual plots can be analyzed for systematic patterns to verify linearity.A linear relationship would indicate that an increase in exam scores is consistently resulting from each more hour of study in a study examining the relationship between study hours (the independent variable) and exam scores (the dependent variable).

1. Independence:

The dataset's observations are said to be independent of one another under the independence assumption. This indicates that there is no correlation between the residuals, or mistakes.Autocorrelation, in which the residuals from two observations are correlated, can occur when this assumption is broken, as in the case of time series data or clustered data. Inflated Type I error rates and underestimated standard errors may result from this.Each patient's recovery time in a hospital dataset should be distinct from the recovery durations of other patients. The independence assumption may be broken if patients are categorized according to their doctor or room.

1. Homoscedasticity:

When residuals are homoscedastic, their variance remains constant at all levels of the independent variables. Stated differently, the distribution or "noise" of the residuals ought to be approximately constant over all anticipated values of the dependent variable.The estimations of the regression coefficients remain unbiased if the homoscedasticity assumption is broken (heteroscedasticity), but the standard errors may be skewed, producing unreliable inferences about the importance of the variables. Residual plots, where the residuals should show a random pattern, can be used to test this.When income is the independent variable and expenditure is the dependent variable in a regression study, homoscedasticity indicates that spending variability should be consistent for both high and low incomes around the anticipated values.

1. Normality of Residuals:

According to the normality assumption, the model's residuals have a normal distribution. This is especially crucial when working with small sample sizes and when putting the regression results through hypothesis testing.The validity of significance tests and confidence intervals may be compromised by breaking this premise. A residuals histogram or a Q-Q plot are frequently used to verify normality, even though the linearity and independence assumptions are more crucial.For accurate inference and trustworthy hypothesis testing, a model that predicts blood pressure based on age and weight should have residuals, or the disparities between observed and predicted blood pressure, that follow a normal distribution.

2. Describe two benefits of using Python or R in support of various phases of the analysis.

Two benefits of using Python in support of various phases of the analysis are its extensive libraries and frameworks and pythons support for data visualization. The extensive libraries and frameworks allow us to use plug ins designed for data science like NumPy and Pandas. The support for data visualization aids in exploring data, noticing patterns, and identifying insights.

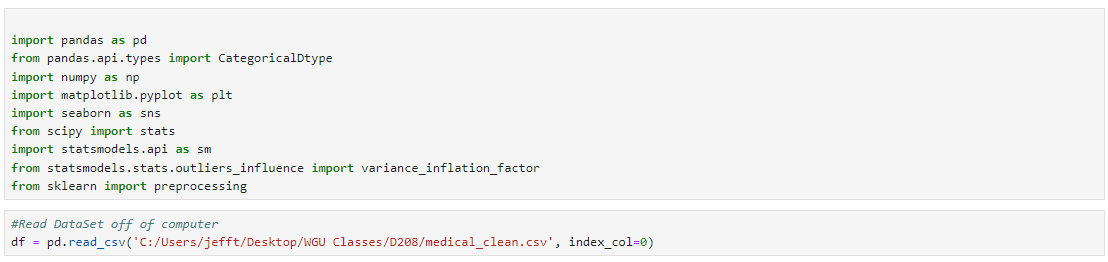
Pandas makes it possible to manage a dataset in a format akin to a big table or spreadsheet.

NumPy enables the assignment of specific values within the dataset or the performance of specific mathematical operations.

Graphing capability is provided by MatPlotLib and Seaborn.

SciPy's statsmodels offer a number of useful features for the multiple regression model, including the ability to graph residuals, as well as for using the variance inflation factor to check for issues like multicolinearity.

When necessary, sklearn's preprocessing can be used to alter our data.



3. Explain why multiple linear regression is an appropriate technique to use for analyzing the research question summarized in part I

I'm using multiple linear regression to see how different factors affect the length of hospital stays. The length of stay is a continuous measure that can vary a lot—there's no fixed upper limit, and it can be tracked with great precision, from days to hours or even minutes. In our data, the length of stay is recorded as a floating-point number, which means it captures both full and partial days accurately. This makes it a good fit for multiple regression, as we can compare it against various other factors to see what influences it.

Part III: Data Preparation

C. Summarize the data preparation process for multiple linear regression analysis by doing the following:

1. Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including your annotated code.

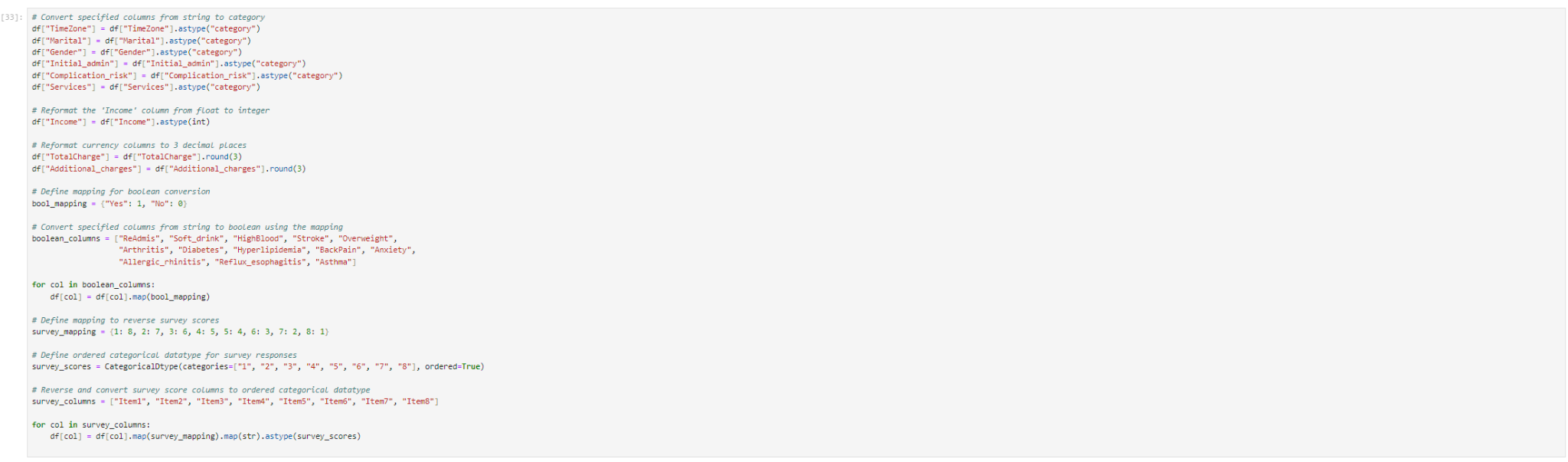
The data set still doesnt have the fixes I made in d206, so I will be importing that code in order to help clean the data. For example, instead of being recorded as the qualitative strings that they should be, zip codes are considered to be whole integers and are stored as an int64, which removes leading zeros.

This dataset's categorical and boolean data require numerical values for multiple regression analysis, not strings. It is simple to convert Boolean values to True or False (1 and 0 correspondingly). Different methods for handling ordinal (i.e., where the order of categories matters, such as "small," "medium," and "large") and nominal (i.e., where the order is irrelevant) categorical variables are required.

There are more than just boolean (already handled) categorical data types in this collection. Of these, only the columns pertaining to survey scores are ordinal, meaning that the values are meant to depict a range from "most important" to "least important." Still, there's a chance that this scale doesn't make intuitive sense in the original numerical form. In order to provide a more logical pattern—where a lower score equates to more importance—I intend to remap these variables. These columns will be remapped, and like in the past, they will be converted to an ordered categorical datatype.

I'll create dummy columns for the nominal categorical columns that are going to be used in this study. This method of converting categorical data into a binary numeric format is called one hot encoding. Consider the values for Male, Female, and Nonbinary in the gender column. The same information can be expressed numerically by adding two more columns. The patient is female if a 1 shows up in the first column and nonbinary if it shows up in the second. The patient is male if there are no 1s in any column (all are 0). This method makes use of Pandas' get\_dummies() function to enable the multiple regression analysis to handle text data efficiently.

The main adjustments required to perform a multiple regression analysis on this dataset are these procedures. The data will undergo additional checks to make sure it is prepared for regression analysis. These checks will include using info() to confirm that there are no null values, value\_counts() to look over all column values, and description() to summarize numeric columns. This was with help from {Tripathi, A. (2019, July 17). *Feature selection techniques in regression model*. Medium. https://towardsdatascience.com/feature-selection-techniques-in-regression-model-26878fe0e24e]



2. Describe the dependent variable and *all* independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.The dependent variable and my y variable for my analysis will be initial\_days. variable. The independent variables or x variables I will use are Children, Anxiety,Age,Income, Gender, Vitd\_levels,arthritis,Diabetes, Total Charge,Doc\_visits, and BackPain.

Initial Days:The dataset indicates a 34-day average hospital stay; however, the standard deviation is a relatively high 26 days. Data points span from one to seventy-two days. As a result, 1.27 standard deviations is the least and 1.5 standard deviations is the biggest departure from the mean. This suggests that the distribution is slightly skewed. Furthermore, a cursory examination of this variable's highest values indicates that the longest stay of roughly 72 days is not an isolated outlier.

Children: There are multiple “groups” of people having kids. 0-4, 5-6,7-8,9-10 have similar frequencies

Anxiety:67.9% oof patients reported not having anxiety, 23.1 reported having anxiety

Age:All of our patients are over the age of 18, according to the overall statistics. 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 take care of this, but the dataset that is offered restricts my possibilities.

Income:The high end could be considered an outlier because it is many standard deviations above the mean, but 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.

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 display a mean that is quite near 18 and a standard deviation that is a little over 2. With an interquartile range of 16.6 to 19.4, the lowest and greatest values differ from the mean by fewer than four standard deviations. This implies that the distribution is roughly normal.

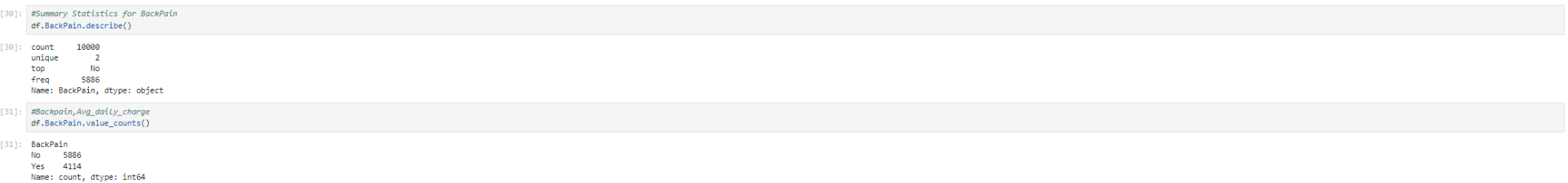
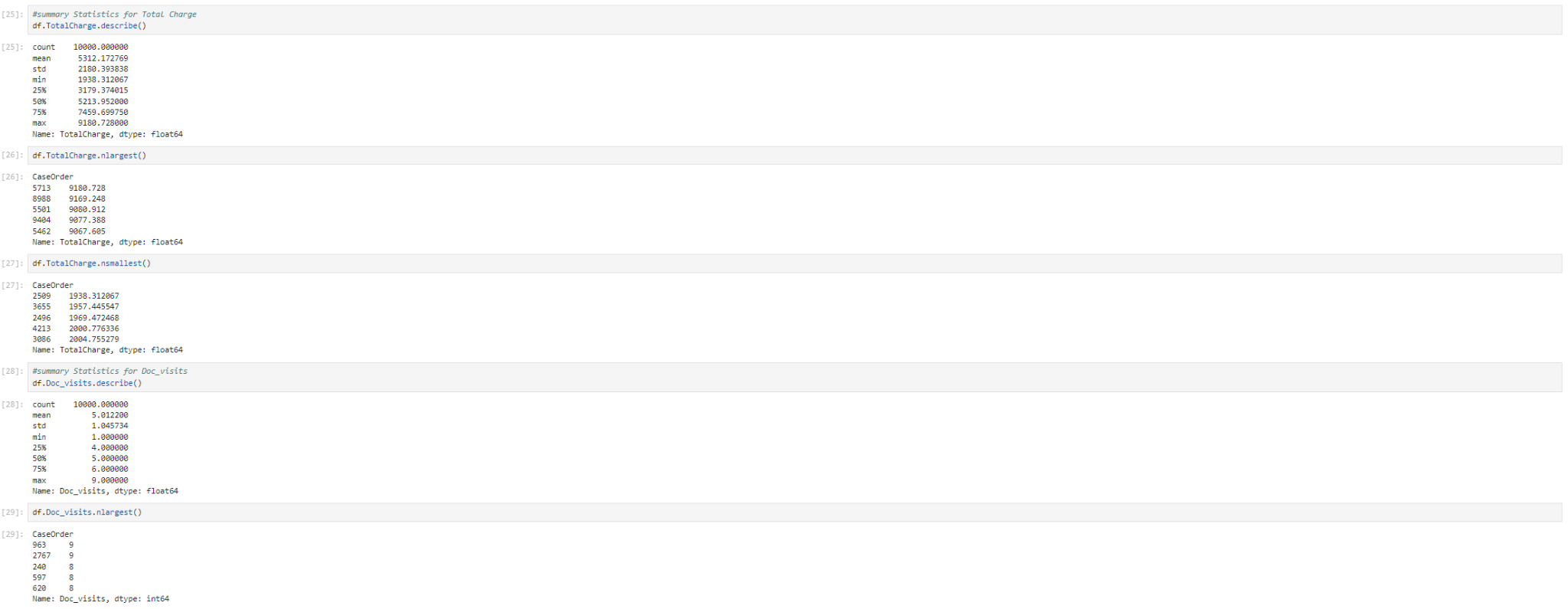
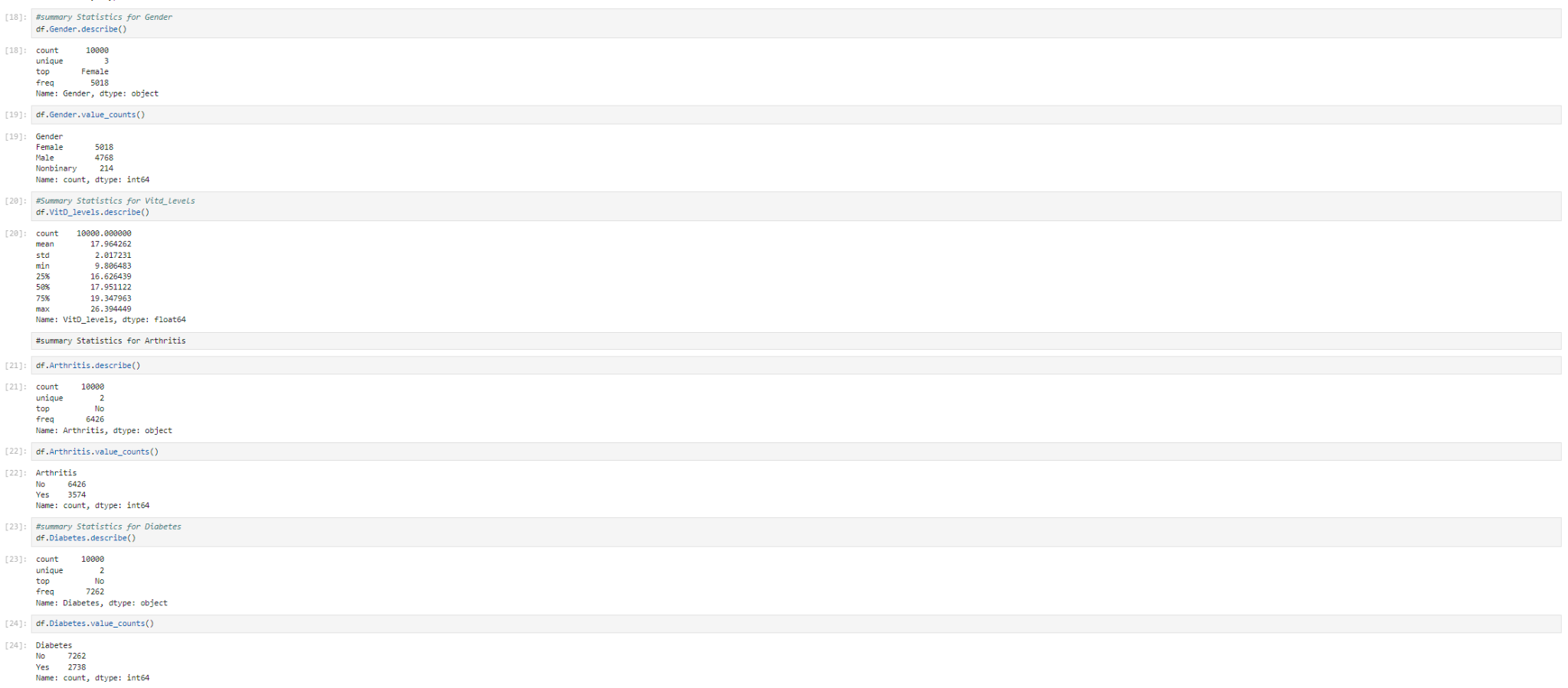
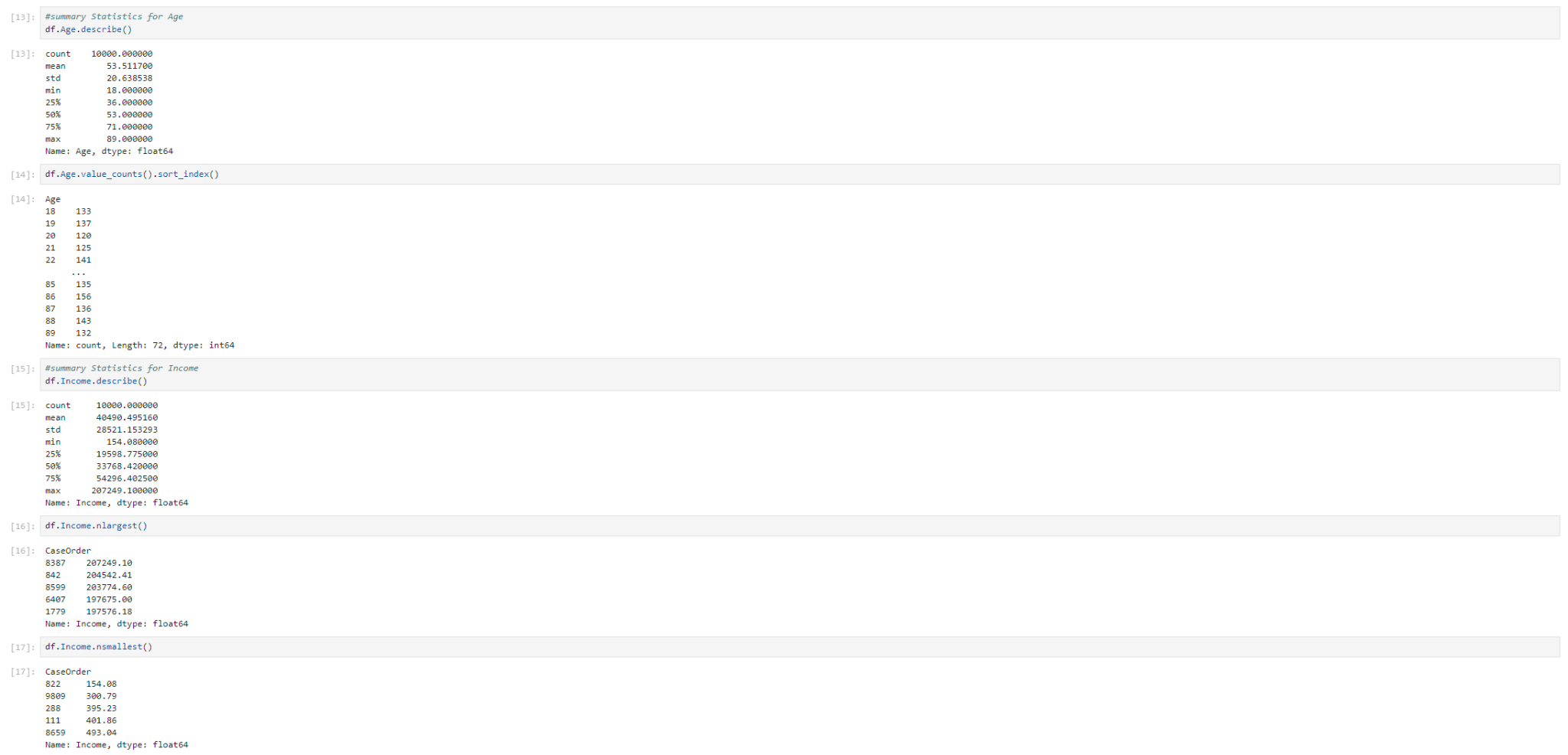
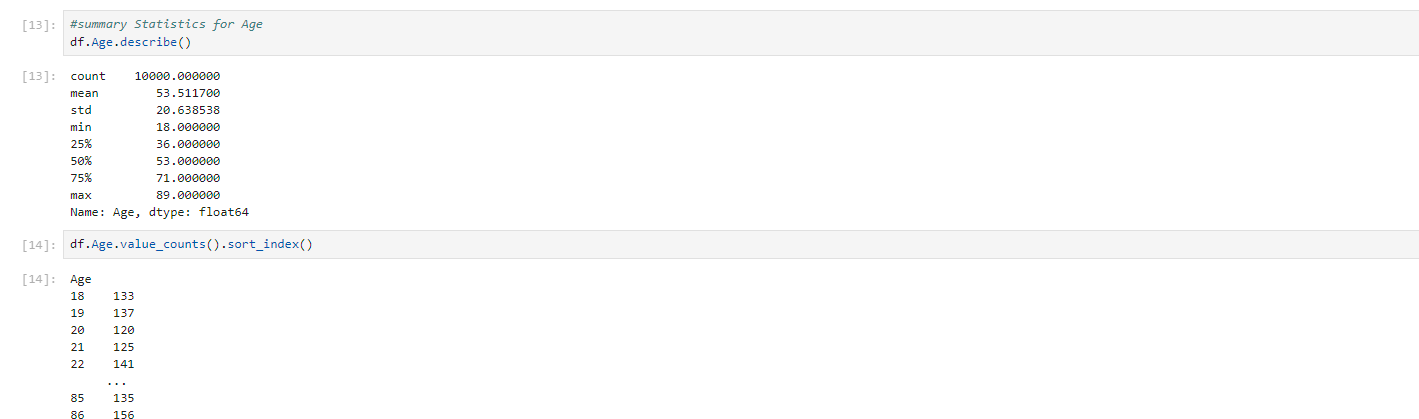
Arthritis: In this dataset, 35% of hospitalized patients have arthritis.

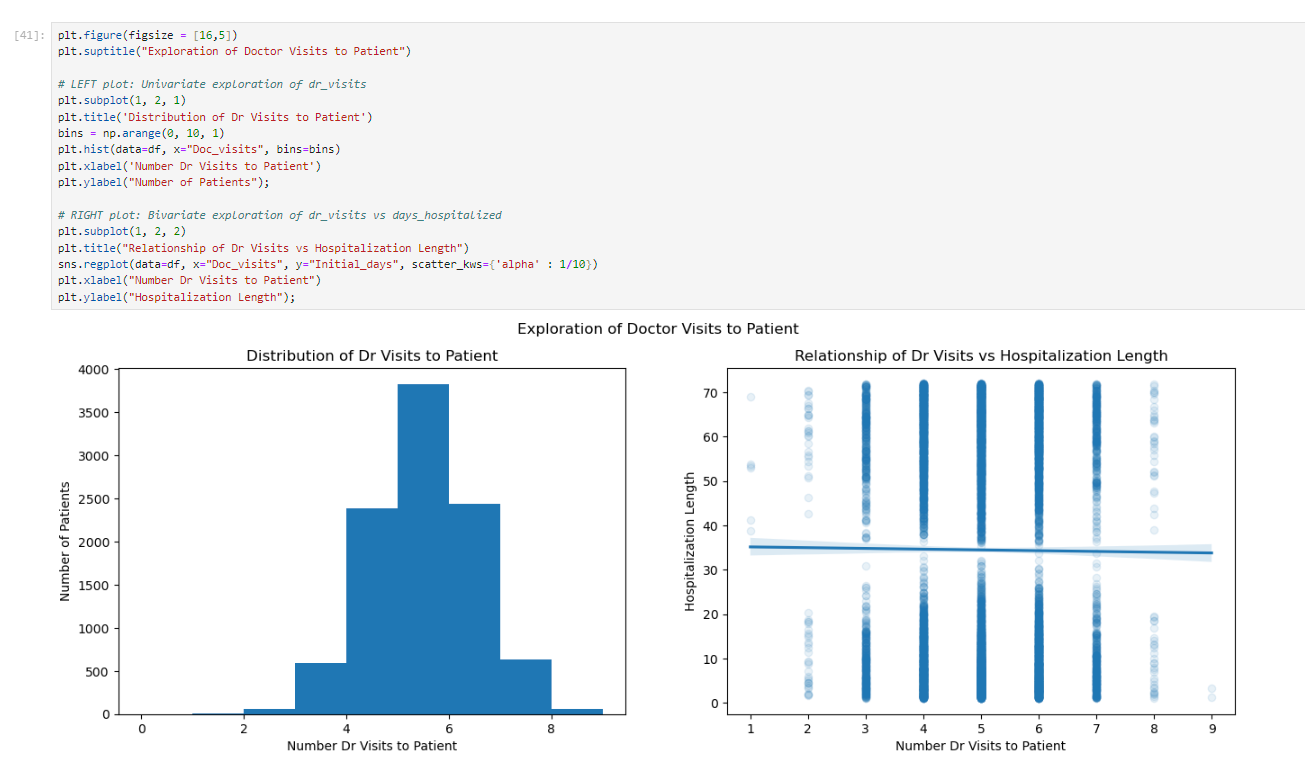
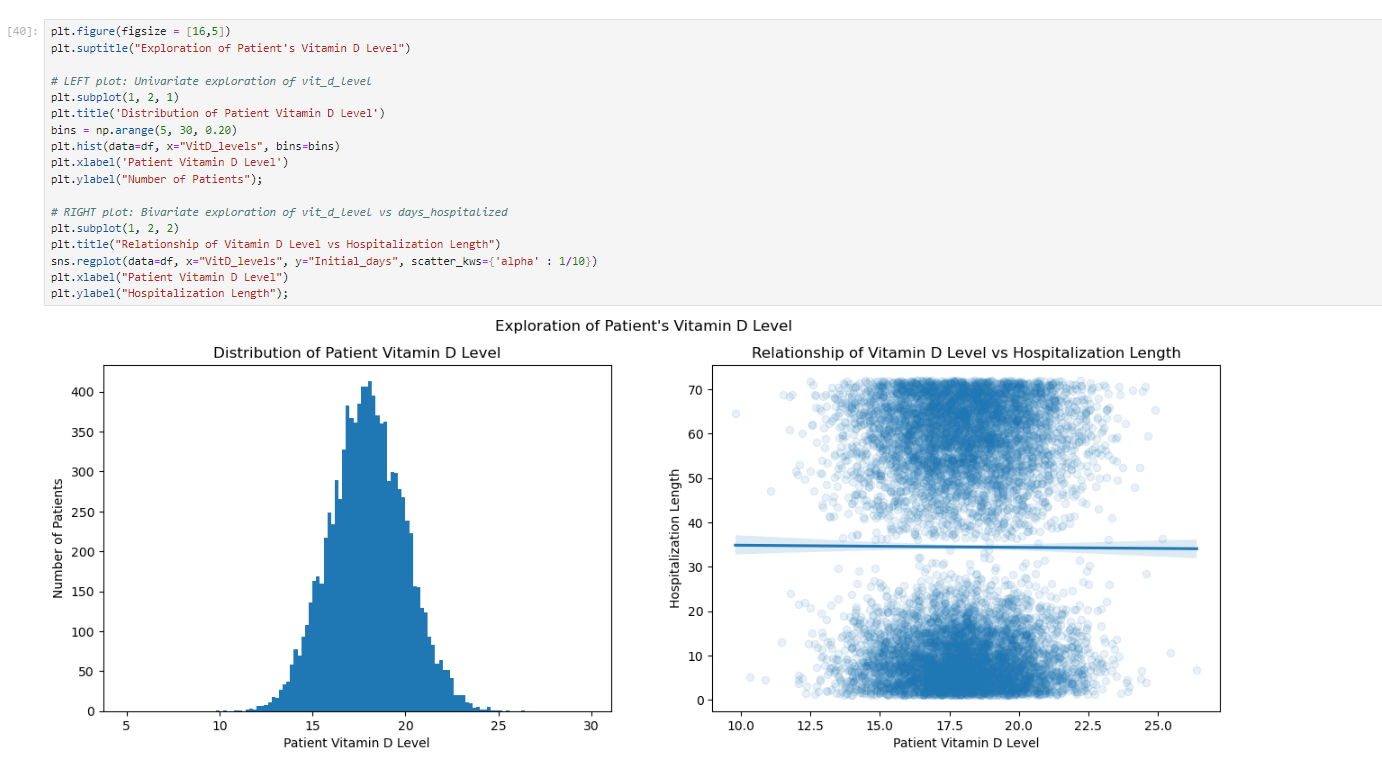
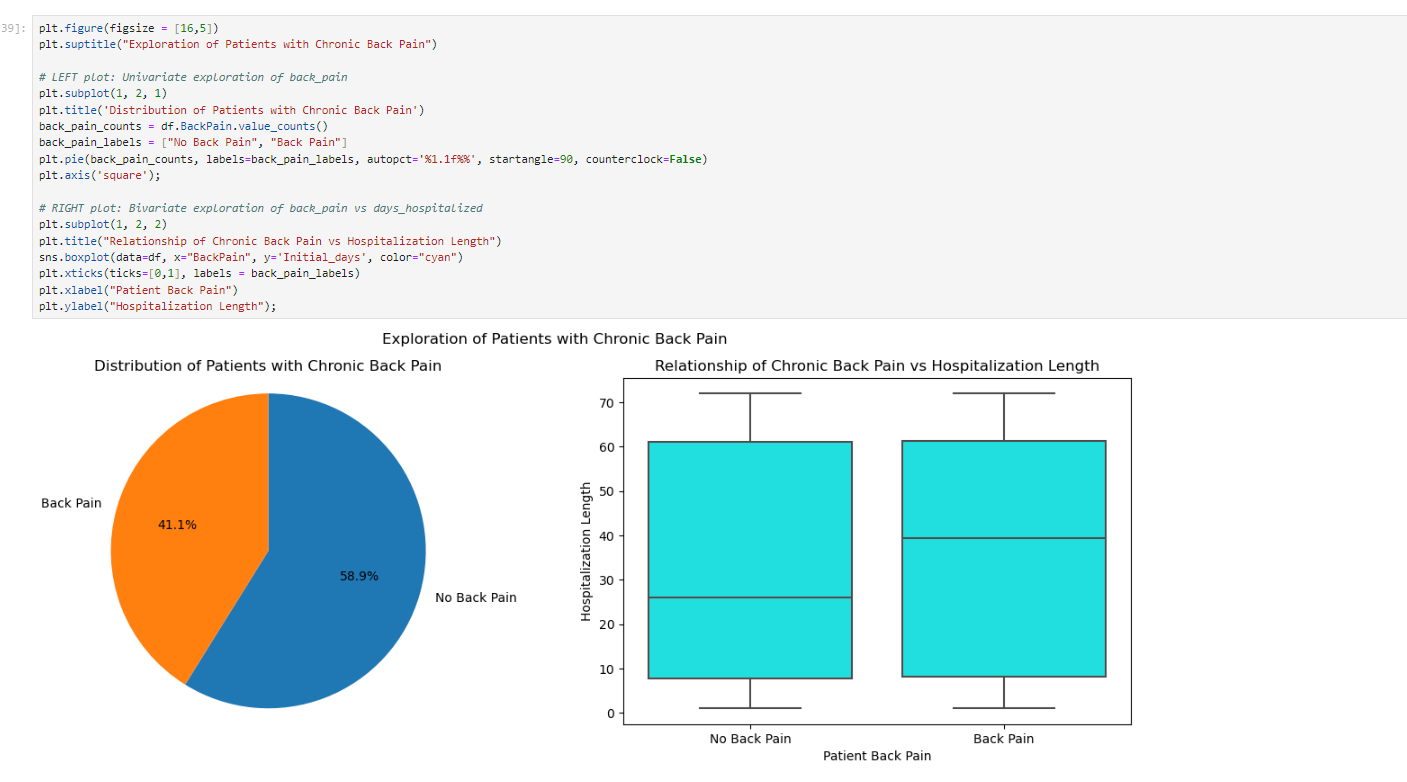
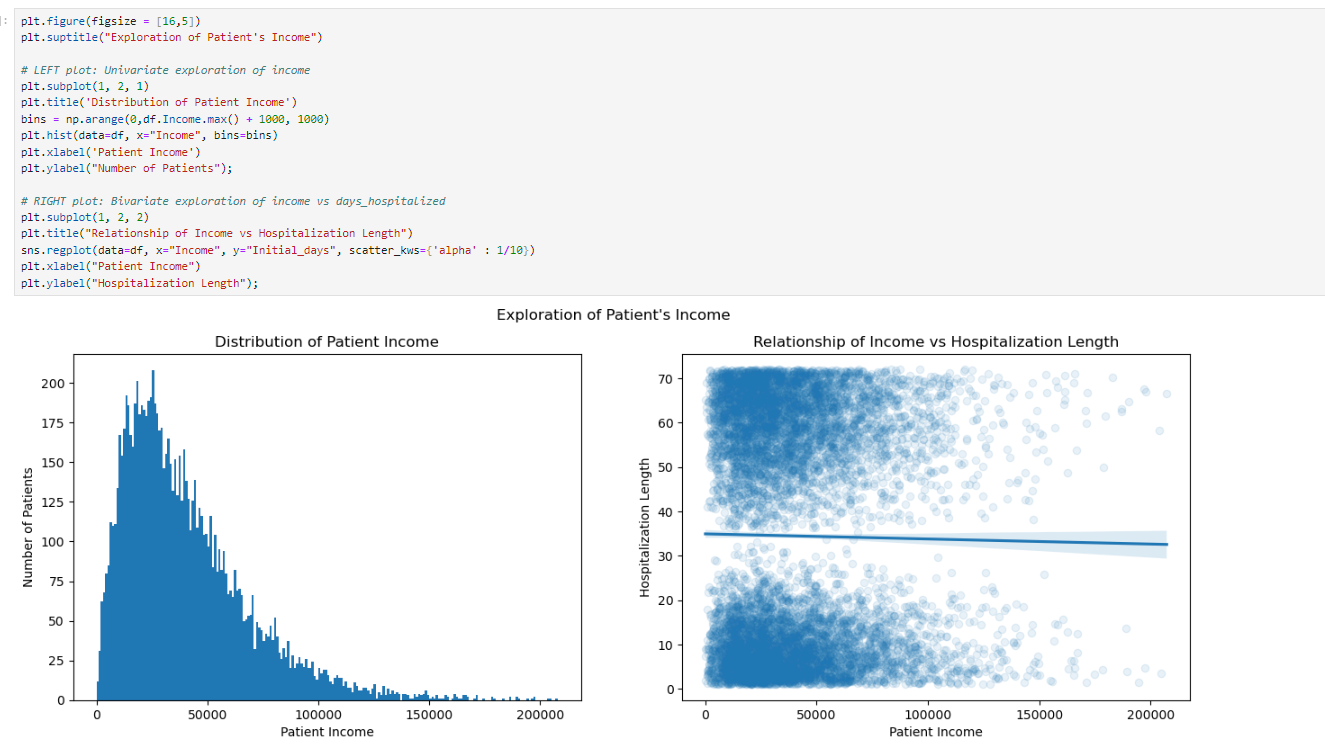
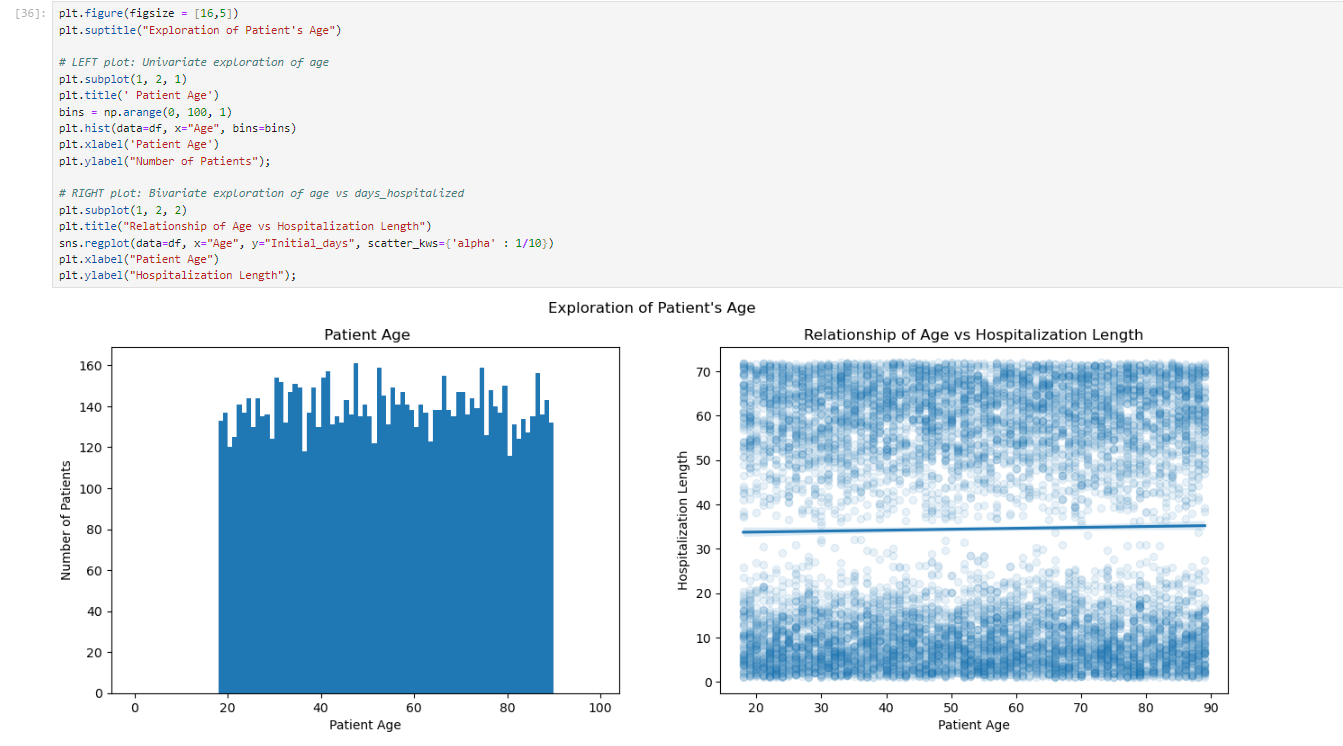
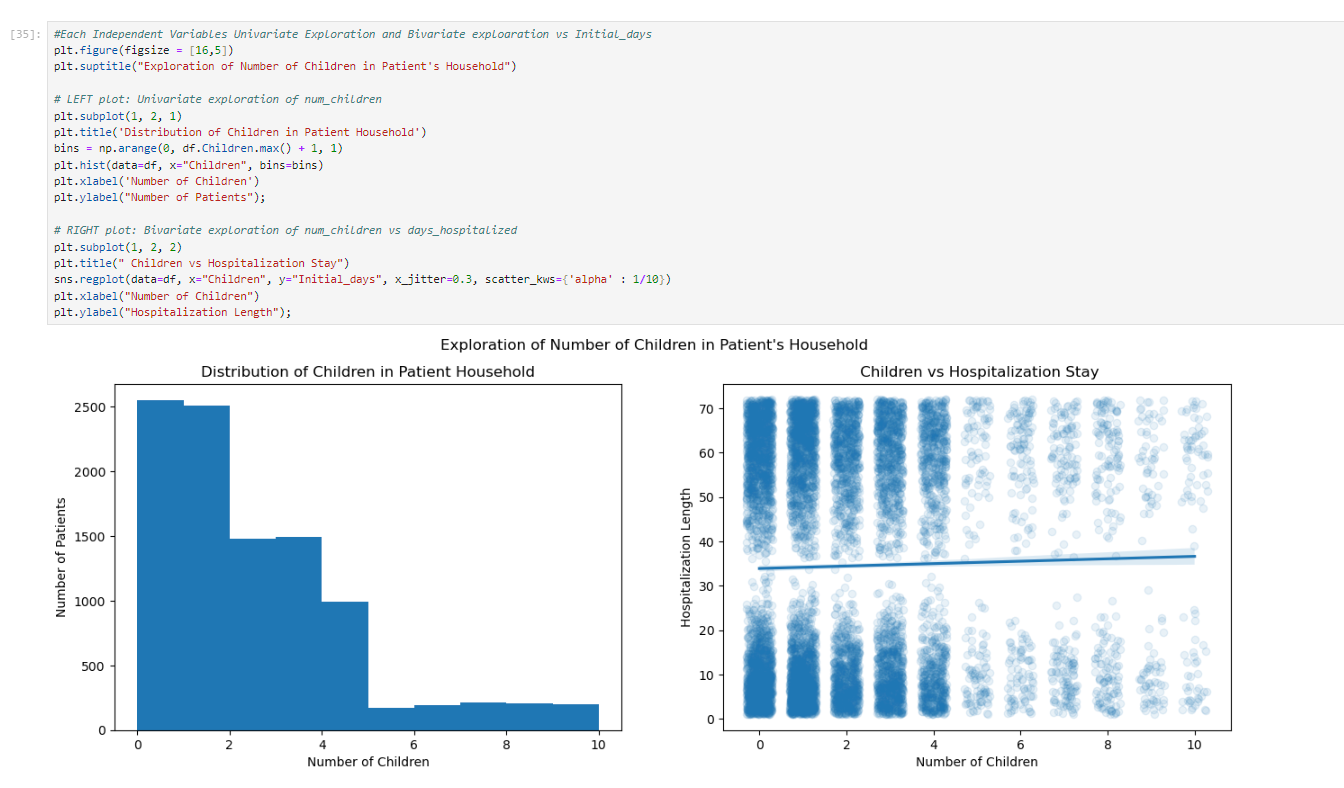
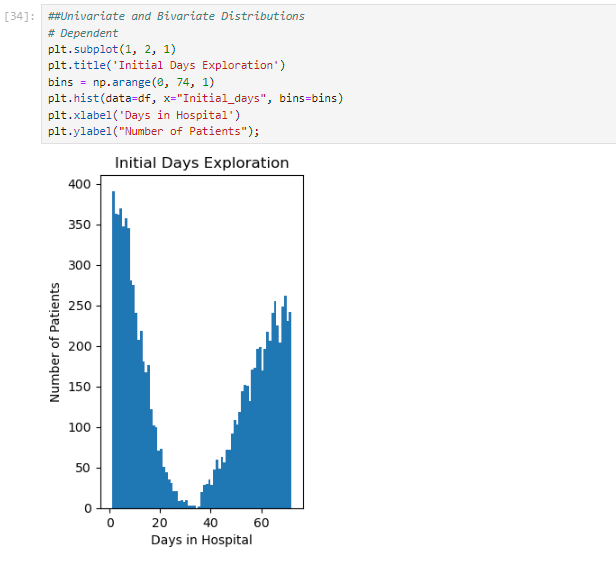
Diabetes: In this dataset, diabetes has been identified in 27 % of the individuals.

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. At around $9,200, the maximum fee is approximately 1.8 standard deviations above the mean. This suggests that there is some deviation from normalcy with a tiny leftward tilt, even if the distribution is often concentrated around the mean. The maximum fee is confirmed to be a normal figure by a quick look at the highest values in the dataset.

Doc\_visits:The dataset indicates that slightly over 50% of hospitalized patients are admitted for emergencies. This is in line with our predictions, or possibly even lower than expected, considering that hospital visits are frequently the result of emergencies, which is something we try to avoid. Hospitalization is required for observation or elective surgery for almost the other half of the patients.

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

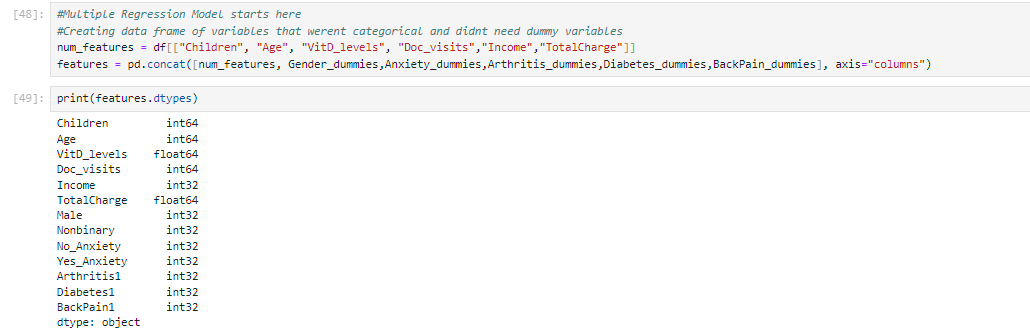
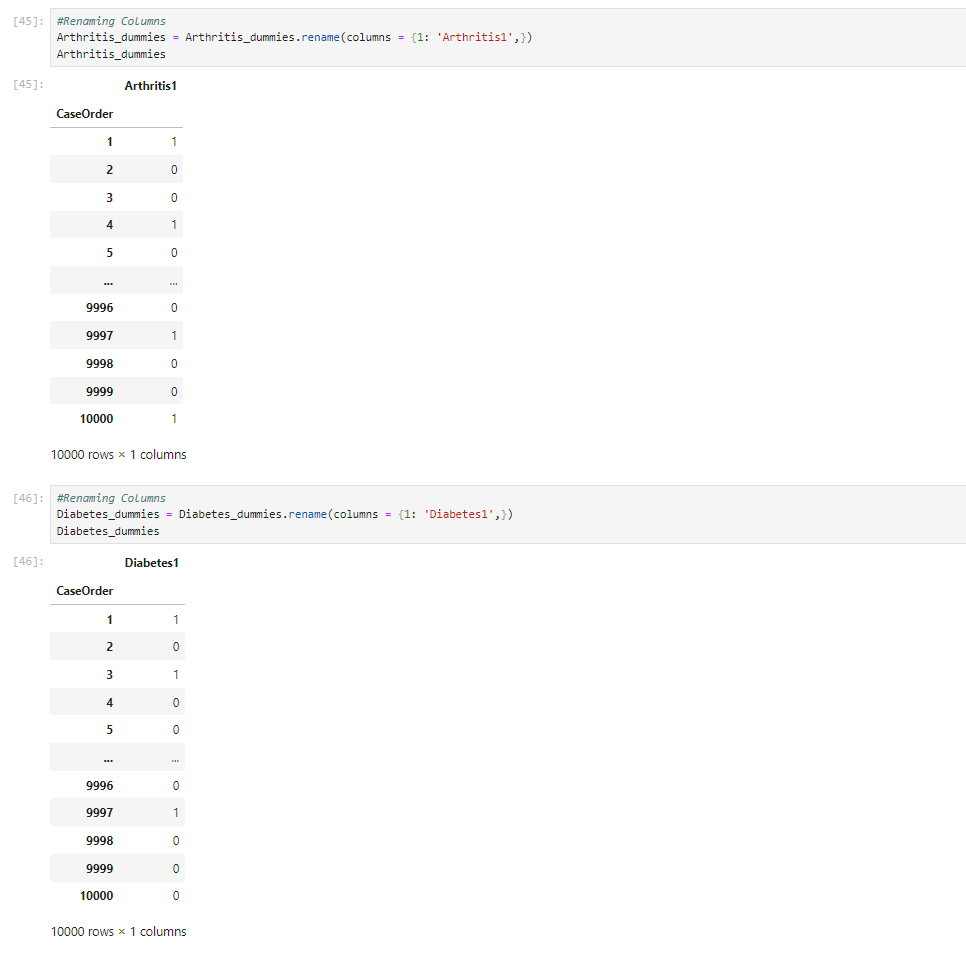
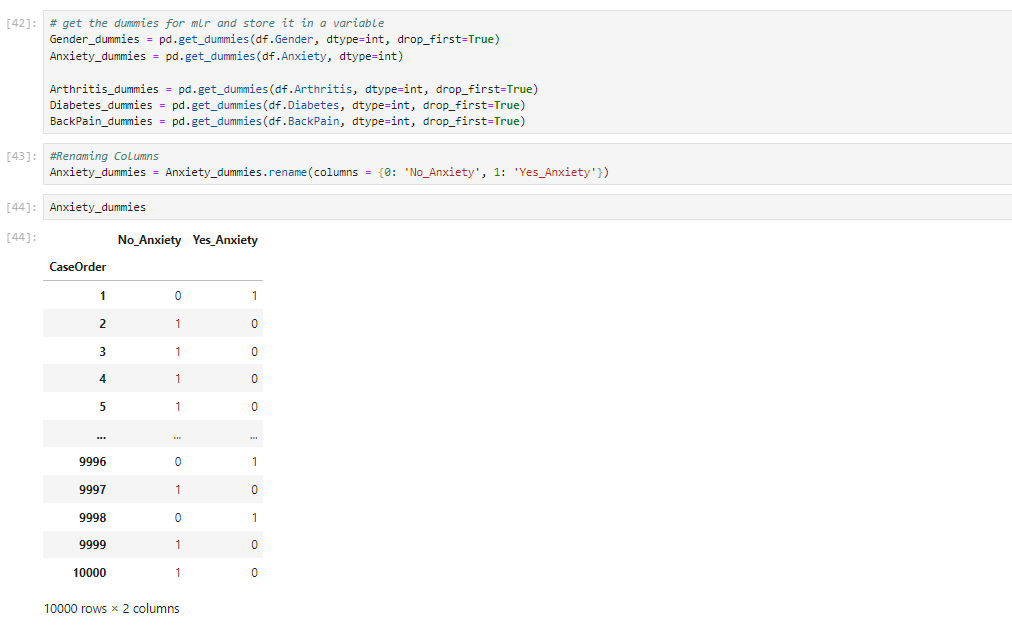


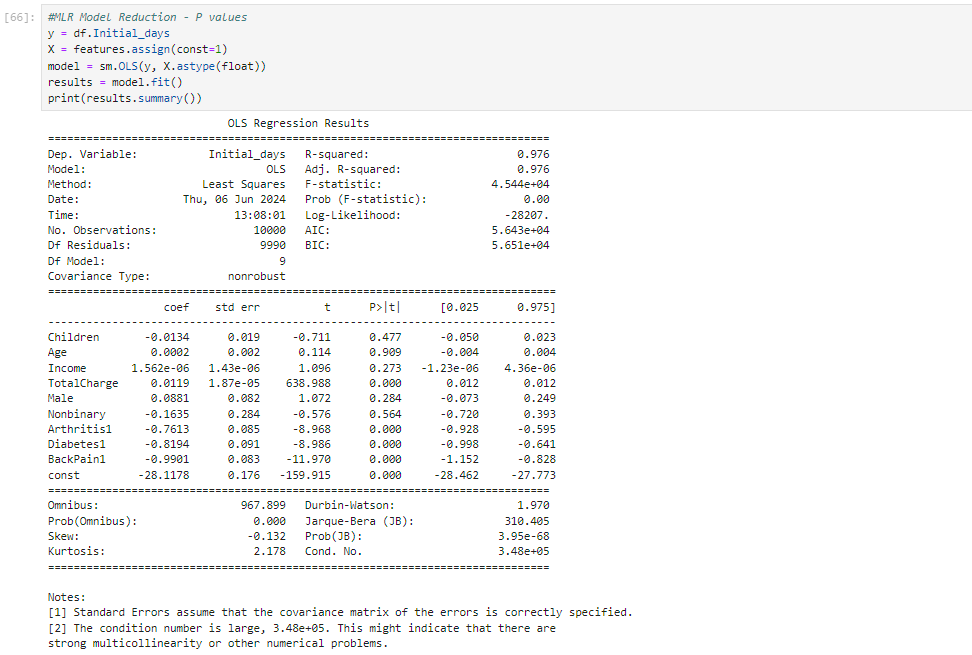
3. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.

5. Provide the prepared data set as a CSV file.



D1 Initial Multiple Regression Model



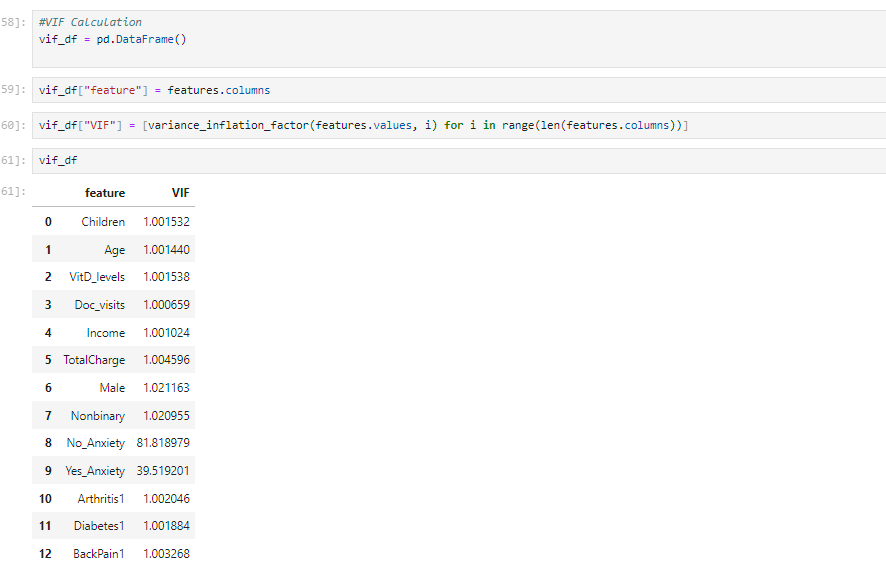


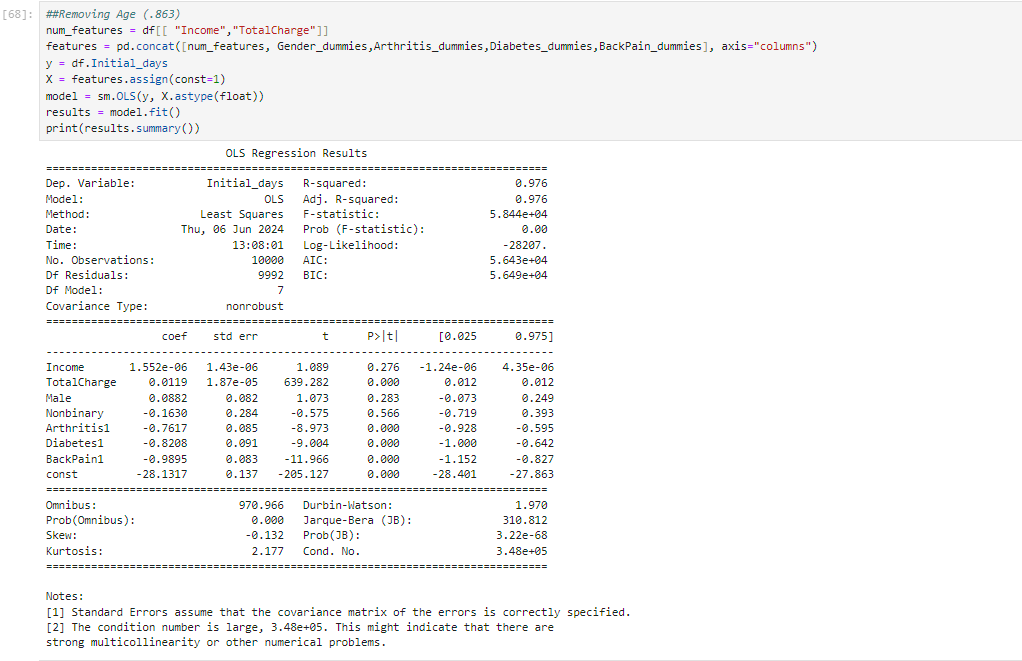
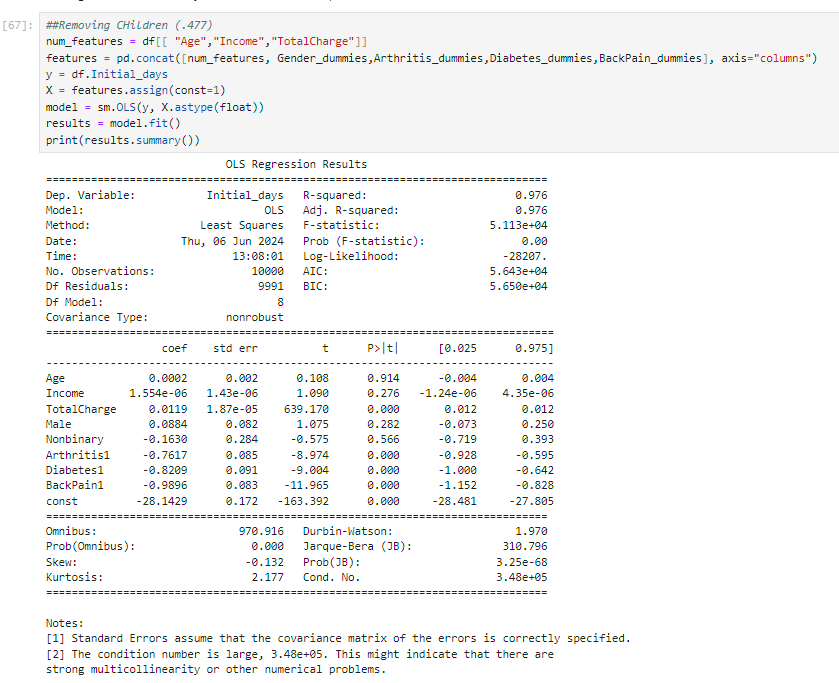
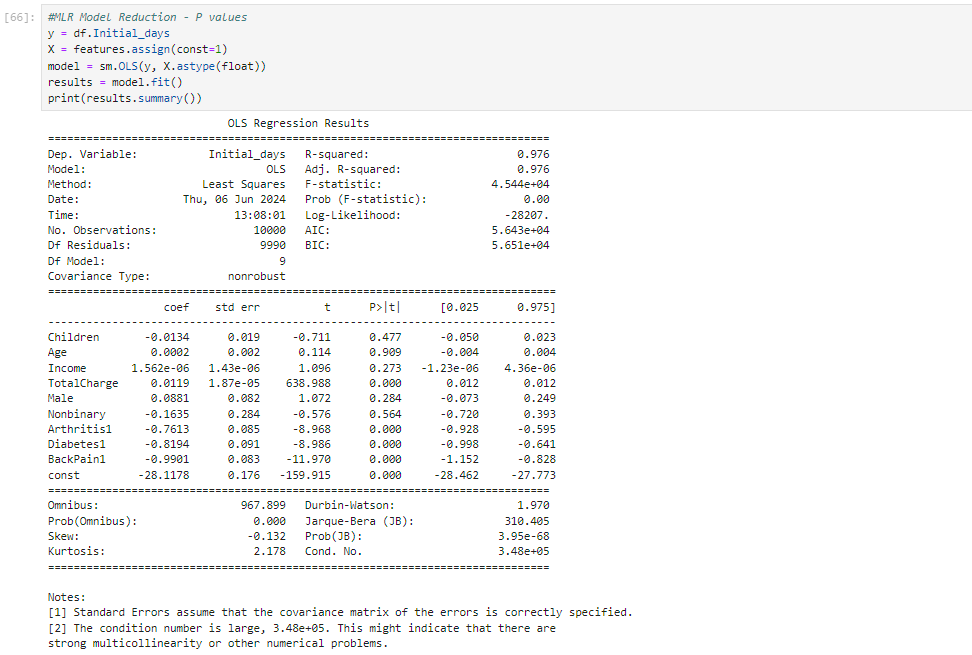
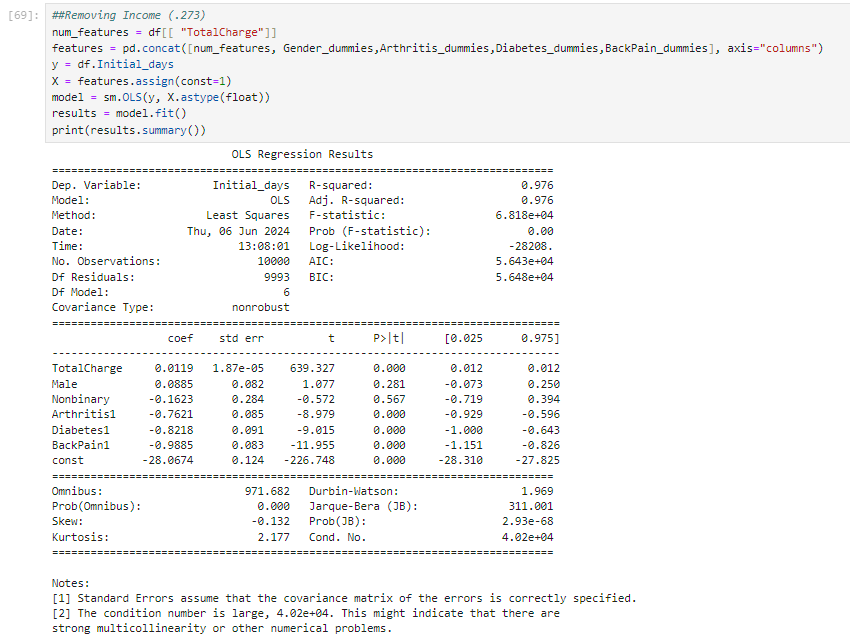
This was performed and comprehended with the help from {YouTube. (2021b, October 11). *Python: Intro to MLR / OLS in statmodels.api*. YouTube. https://www.youtube.com/watch?v=0-fkgpK2knA&list=PLe9UEU4oeAuV7RtCbL76hca5ELO\_IELk4&index=8

YouTube. (2021c, October 11). *Python: MLR, OLS, standardization, normalization*. YouTube. https://www.youtube.com/watch?v=QH\_elD\_JKuc&list=PLe9UEU4oeAuV7RtCbL76hca5ELO\_IELk4&index=10

}

Reduction Justification

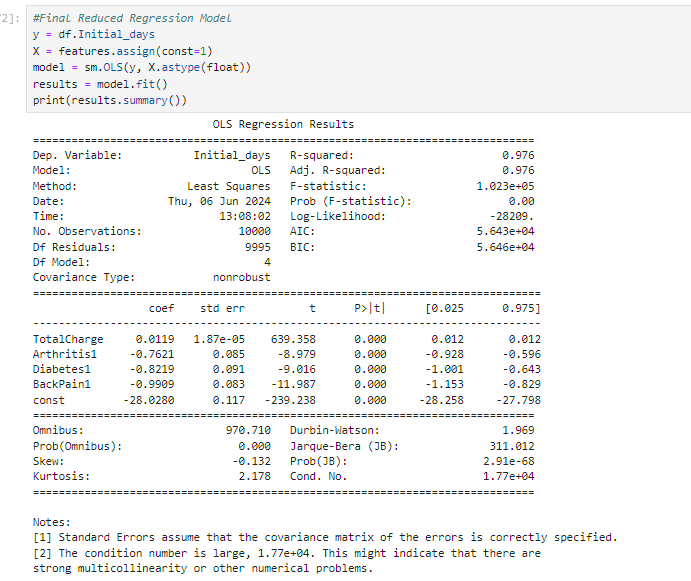
As you can see from the multiple regression findings above, our model may have significant problems with multicollinearity. Using the Variance Inflation Factor (VIF) to check for excessive multicollinearity among our independent variables is a basic step in multiple regression analysis. The elimination of any factor having a VIF higher than 10 is recommended. You must keep doing this until every VIF number is less than 10. I will execute the analysis using the following code (from WGU Courseware Resources) in order to accomplish this. After that, I'll get rid of the factor with the highest VIF, do the analysis again, and keep going until every VIF value is lower than 10.

Next, to reduce variables that are not statistically significant to the model, I will be doing regression and eliminating variables that have a p value higher than .05. This will be repeated until only variables with p values below .05 remain indicating they are statistically significant. I understood these concepts and formulas from Tripathi, A. (2019, July 17). *Feature selection techniques in regression model*. FggsdF1:Regression Equation, Coefficients, etc

Equation:Results of Data Analysis

Y(Initial\_days) =- -28.0280 -0.9909(backpain) - 0.8219(diabetes) - 0.7621(Arthritis) +0.0119(TotalCharge)

Coefficients:.0119 Total Charge, -.7621 Arthritis, -.8219 Diabetes, -.9909 BackPain, -28.0280





With the final model, the statistically significant variables to my research question are TotalCharge,Arthritis,Diabetes and Backpain.

a) Provide a regression equation

Results of Data Analysis

Y(Initial\_days) =- -28.0280 -0.9909(backpain) - 0.8219(diabetes) - 0.7621(Arthritis) +0.0119(TotalCharge)

b) Provide an interpretation - Because "Initial\_days" is a continuous variable, please interpret the coefficients as how many days the dependent variable (Initial\_days) changes if the independent variables change.

Initial day increases by .0119 days if the TotalCharge variable changes.

Initial day decreases by .761 days if the Arthritis1 variable changes.

Initial day decreases by .8219 days if the Diabetes1 variable changes.

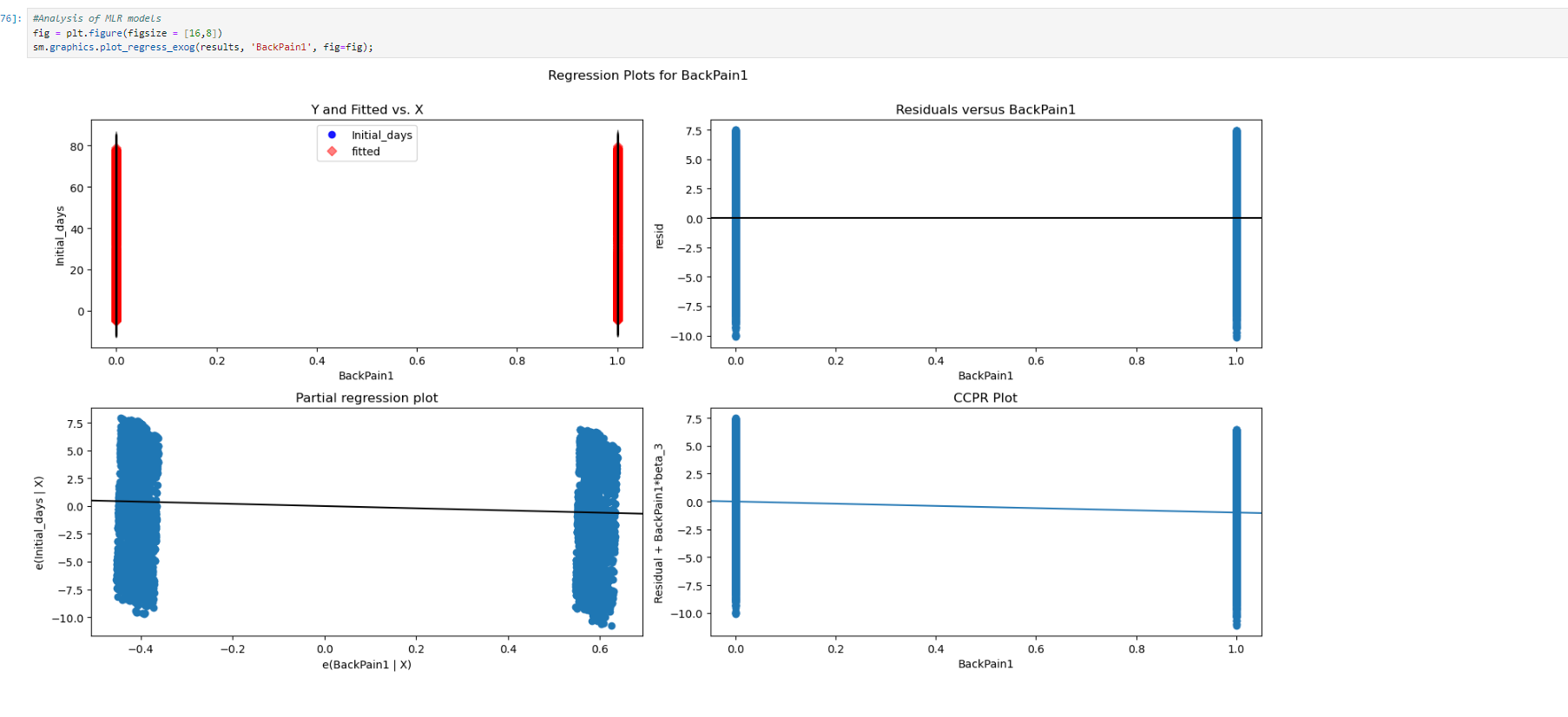
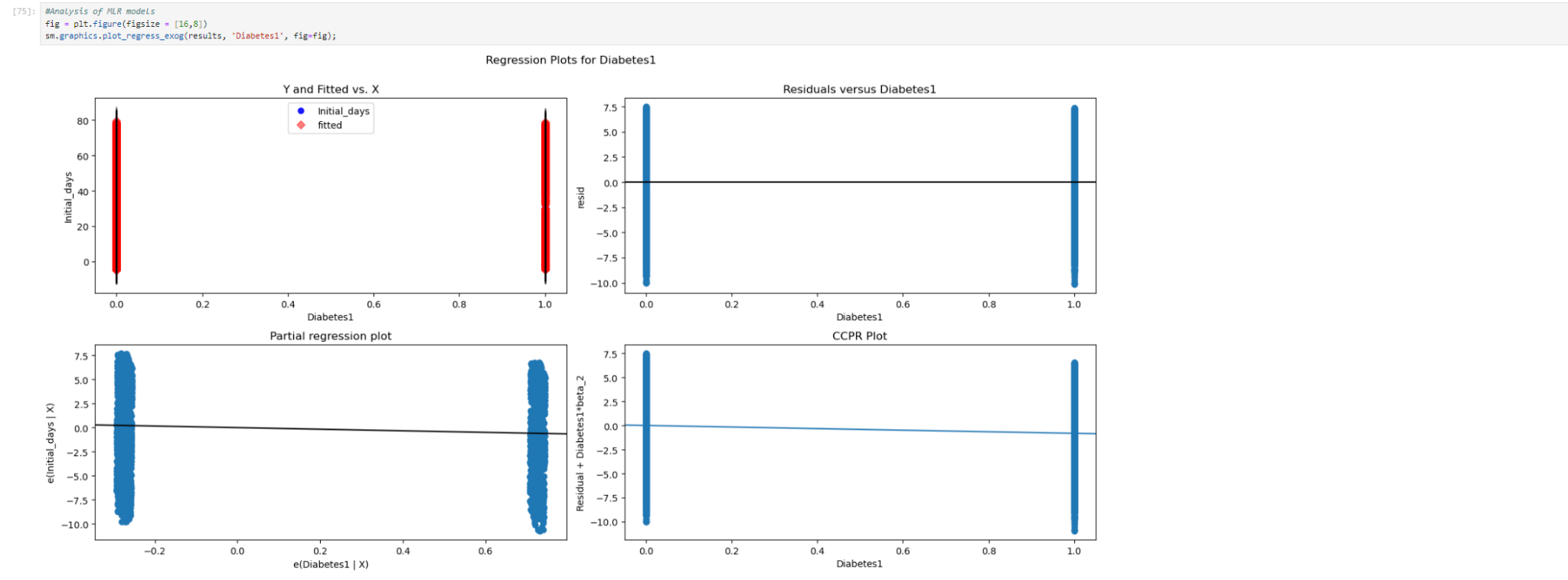
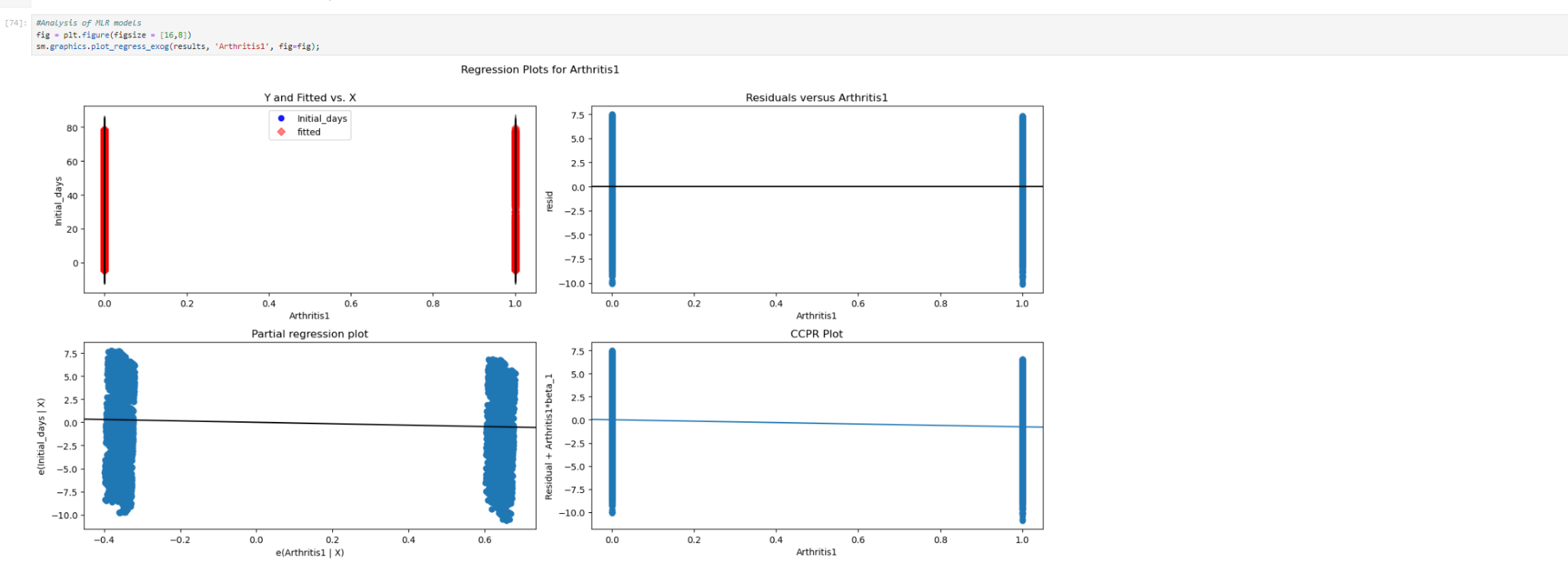
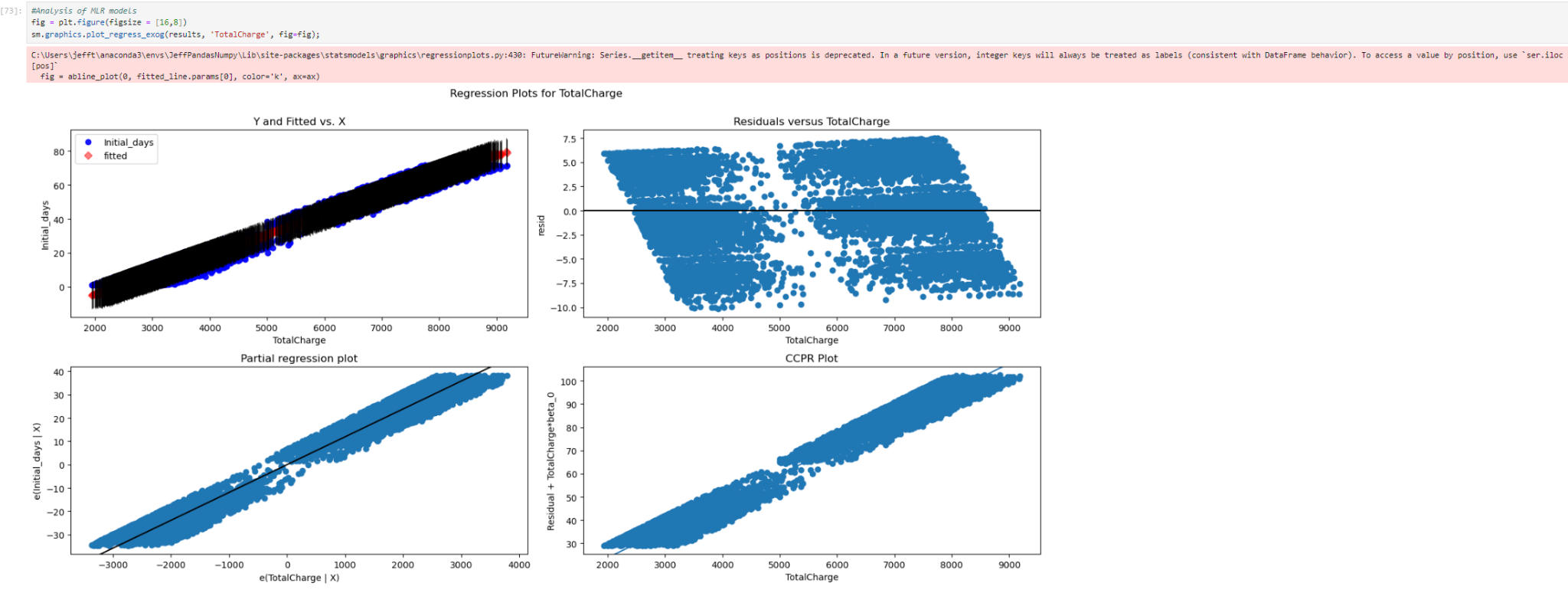
Initial day decreases by .0119 days if the BackPain1 variable changes.

Analysis of Multiple Regression Models/Model Comparison

Many variables were included in the original multiple regression model, some of which were not essential to the model's performance. Two variables were eliminated: dr\_visits and vit\_d\_levels, due to multicollinearity concerns. The remaining variables were then removed from the model using a Backwards Stepwise Elimination procedure in accordance with their p-values. A variable's statistical significance is gauged by the p-value, with lower values denoting more significance. High-p-value variables were eliminated one at a time, recalculating the model after each removal, until all variables that remained had statistically significant p-values of less than 0.05.

The high r-squared values and statistical significance of the initial and reduced regression models were demonstrated by their respective p-values. As a result, the residual standard error—an alternative metric—was employed to compare these models. The discrepancies between the actual and anticipated values of the data points are called residuals, and they are used to evaluate the accuracy of the model. Standard error of these residuals is a measure of the degree of variance between the actual and projected values; a lower value indicates a better model. The residual standard error of the revised model (described in section D3) was 4.0640, while the initial model (discussed in section D1) had a residual standard error of 4.0347. This suggests that, in comparison to the baseline model, the reduced model is less reliable at predicting initial\_days.

The residual plots for each explanatory variable are shown below. The GeeksForGeeks lesson on creating residual plots in Python served as a guide for writing the code that produced these graphs.

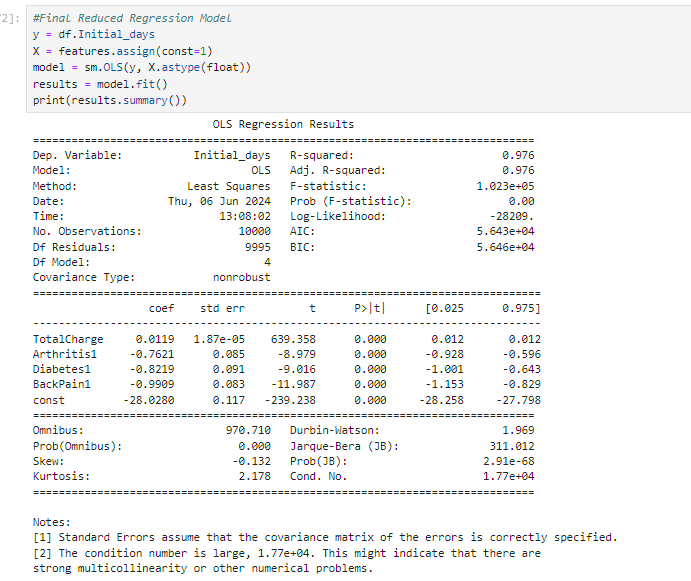


The residual plots indicate that the residuals are not homoscedastic, or at least not optimally.

GeeksforGeeks. (2022, February 21). *How to create a residual plot in Python*. https://www.geeksforgeeks.org/how-to-create-a-residual-plot-in-python/

}. In most of the plots, the residuals are not centered around 0 or the line of best fit; instead, they are mostly below this threshold. This pattern is especially noticeable with the categorical variables.

c) Provide a discussion regarding the statistical significance and practical significance of your reduced model.

· First, you will need to discuss IF your model is/are statistically significant and why or why not.

The model's f-statistic being 0.00 shows that it's statistically significant. This means the result is likely not due to chance, as it's less than 0.05. Although there might be some concerns about the model's accuracy due to deviations from perfect homoscedasticity, I don't see this as a major issue since I don't think the model is practically useful.

The key reason for this is that the significant variables identified by the model are generally beyond the control of the medical system, diminishing its real-world relevance.

The variables in the dataset are about the patients’ and their information. The hospital can't pick who to admit based on certain criteria. So their information is somewhat random, as long as we don't go deep into hospital locations and demographics. On the other hand, for Totalcharge, you can quickly see a relationship as to the longer a patient stays in the hospital, the higher the bill There are no patients under the age of 18 in the sample. The fact that the patterns we discovered might not apply to younger patients could be a problem. If we try to apply best practices or suggestions developed on the basis of this data to kids who were not included in the study, we may run into issues. For older teenagers, this might not be a significant deal, but for younger children or babies, it might be a major issue.

There are no indicators to clarify if the patients died in the hospital after their stays or not. This leaves room for assumption on viewing the results, leading into them being readmitted.

Limitations

Patients with shorter stays are not included in the hospitalization data; only those who stayed for at least 24 hours are included. It's crucial to understand the practical distinction between being admitted formally vs just checking into an emergency department on a temporary basis, since both hospitals and patients want to avoid being hospitalized altogether. The correlations found in the analysis may not apply or may apply differently within the first 24 hours of hospital care as a result of this exclusion, which could also result in different incentives that are not represented in the data.

None of the patients in the sample are under 18 years old. Excluding patients younger than 18 could be challenging, as the associations or patterns identified in this analysis might not be relevant to younger individuals. Developing "best practices" or regulations based on this data could present problems if applied to minors, given that the analysis did not include them. While this might not significantly affect older teenagers close to adulthood, it could be especially problematic for younger children or infants.

Recommended Action

The dataset could be far more detailed in order to create a more realistic analysis/prediction with an emphasis on the patient's healthcare information. There is a lot more that goes into a patient than the ones listed in the dataset. More information can create more accurate results, as long as it is good information. For example, types of treatments, initial cause of admission, quantity of nurse visits, and healthcare services they are able to receive.

Sources

GeeksforGeeks. (2022, February 21). *How to create a residual plot in Python*. https://www.geeksforgeeks.org/how-to-create-a-residual-plot-in-python/

Tripathi, A. (2019, July 17). *Feature selection techniques in regression model*. Medium. https://towardsdatascience.com/feature-selection-techniques-in-regression-model-26878fe0e24e

YouTube. (2021b, October 11). *Python: Intro to MLR / OLS in statmodels.api*. YouTube. https://www.youtube.com/watch?v=0-fkgpK2knA&list=PLe9UEU4oeAuV7RtCbL76hca5ELO\_IELk4&index=8

YouTube. (2021c, October 11). *Python: MLR, OLS, standardization, normalization*. YouTube. https://www.youtube.com/watch?v=QH\_elD\_JKuc&list=PLe9UEU4oeAuV7RtCbL76hca5ELO\_IELk4&index=10