D208 Performance Assessment

Task 1

Adrienne Stafford

02/07/24

**Part I**

A1. Research Question

Can we accurately predict a customer's Total Charge using a list of explanatory variables?

A2. Goal

My main objective is to precisely anticipate a customer's total charge by utilizing a thorough examination of numerous explanatory variables. By analyzing the statistical significance and correlations between the significant factors impacting Total Charge and the target variable, the model will systematically identify and quantify these factors. The ultimate goal is to create a strong and trustworthy predictive framework that provides deep insights into the underlying causes of these charges in addition to achieving high forecasting accuracy for Total Charge.

**Part II**

B1. Assumptions

1. There should be a linear relationship between each independent variable and the dependent variable.
2. For every level of the independent variables, the variance of the residuals, or errors, should be constant.
3. Observations should to stand alone from one another.
4. The residuals should roughly follow a normal distribution, especially when drawing conclusions about the population (such as through hypothesis testing).

B2. Python Benefits

Python's extensive ecosystem and its capacity for automation make it an invaluable tool for data analysis. The language's comprehensive library support is one of its foremost advantages. Libraries such as pandas and NumPy facilitate efficient data manipulation and preprocessing, enabling analysts to handle large datasets with ease. For visualization, Matplotlib, Seaborn, and Plotly provide powerful tools to create detailed and insightful plots, aiding in both exploratory data analysis and the effective communication of results. Another significant benefit of Python is its capability for seamless integration and automation. Python can easily interface with databases, web applications, and other programming languages, ensuring smooth workflow integration. Additionally, its scripting capabilities allow for the automation of repetitive tasks, enhancing efficiency and consistency in data processing and analysis. These attributes underscore Python's status as a critical asset in the arsenal of modern data analysts, empowering them to conduct thorough, efficient, and reproducible analyses (Team, 2024).

B3. Why Linear Regression

Since total charge is a numeric response variable, multiple linear regression is the proper technique for this research question's analysis. I can anticipate the overall charge based on these inputs by modeling the relationship between the response variable and several explanatory variables using linear regression.

**Part II**

C1. Data Goals

Optimal dataset use for multiple linear regression is the aim of data preparation. I'll start by loading the dataset into Python and analyzing it to gain a deeper understanding of its contents. Depending on the data type of each variable, I will first look for any missing or null values. If any are found, I will impute them using the mean, median, or mode. Numerical values will be assigned to columns that have labels for categorical variables. 'State' and the other remaining classified columns will be removed. Lastly, before saving the prepared dataset, I will make univariate and bivariate graphs.

C2. Variable Description

Upon loading the dataset into pandas, I observe that it contains 50 columns, each with 10,000 records. In this analysis, the target variable is the numeric variable 'Total Charge'. I have removed the less meaningful columns that are unlikely to influence our response variable. These discarded columns include less significant data and customer demographic information, such as:

* CaseOrder
* Customer\_id
* Interaction
* UID
* City
* State
* County
* Zip
* Lat
* Lng
* Population
* Area type
* TimeZone
* Job
* Marital
* Gender
* Full\_Meals\_Eaten
* Soft\_Drink
* Services
* Item 1
* Item 2
* Item 3
* Item 4
* Item 5
* Item 6
* Item 7
* Item 8

I looked through the dataset and saw that it was well-cleaned and free of missing data. In addition, 1 and 0, respectively, were substituted for all remaining category values with 'yes' and 'no' fields. After converting the category variables to numerical values, the original columns were eliminated. The remaining columns that came from this method were as follows:

* Children: Numeric, Mean 1.00 Meal range 0 - 7
* Age: Numeric, Mean 53.5 yrs Age range from 18 - 89
* Income: Numeric, Mean 40,490.50 Salary range from 154.08 - 207,249.10
* VitD\_Level: Numeric, Mean 17.96 Levels range 9.81 – 26.39
* Doc\_visits: Numeric, Mean 5 Visit range 1 – 9
* ViD\_supp: Numeric, Mean 0.3989 Supply range 0 - 5
* Initial\_days: Numeric, Mean 34.46 Day range 1.00 – 71.98
* TotalCharge: Numeric, Mean 5,312.17 Total charge range 1,938.31 – 9,180.73
* Additional\_Charges: Numeric, Mean 12,934.53 Charges range 3,125.70 – 30,566.07
* ReAdmis (# of observations: 10,000)

Yes : 37% No: 63%

* Complication\_risk (# of observations: 10,000)

Low : 21% Medium: 45% High: 34%

* Overweight (# of observations: 10,000)

Yes: 29% No: 71%

* Arthritis (# of observations: 10,000)

Yes: 36% No: 64%

* Diabetes (# of observations: 10,000)

Yes: 27% No: 73%

* Hyperlipidemia (# of observations: 10,000)

Yes: 34% No: 66%

* BackPain (# of observations: 10,000)

Yes: 41% No: 59%

* Anxiety (# of observations: 10,000)

Yes: 32% No: 68%

* Allergic\_rhinities (# of observations: 10,000)

Yes: 39% No: 61%

* Reflux\_esophagitis (# of observations: 10,000)

Yes: 41% No: 59%

* Initial\_admin (# of observations: 10,000)

Emergency Admission: 51% Elective Admission: 25% Observation Admission: 24%

* HighBlood (# of observations: 10,000)

Yes: 41% No: 59%

* Stroke (# of observations: 10,000)

Yes: 80% No:20%

* Asthma (# of observations: 10,000)

Yes: 29% No: 71%

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

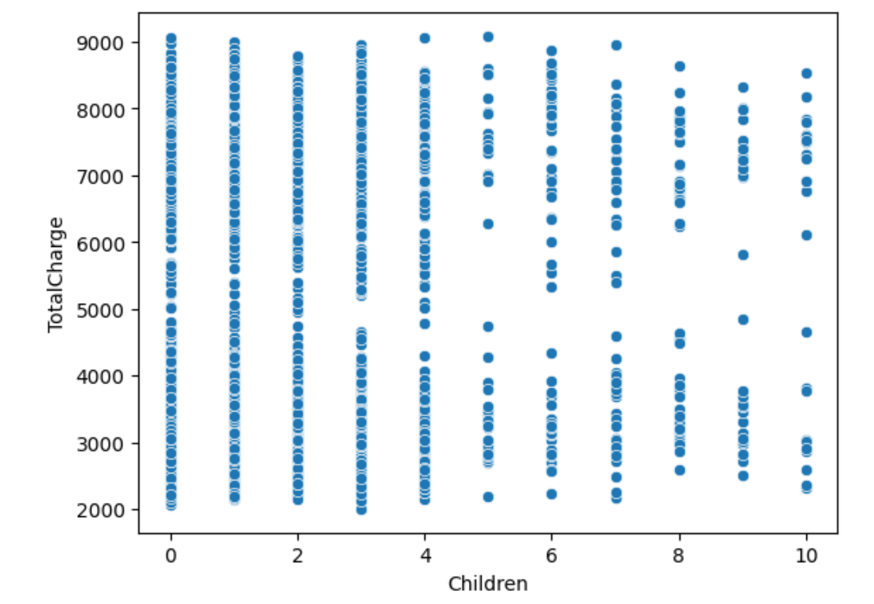
Description automatically generated

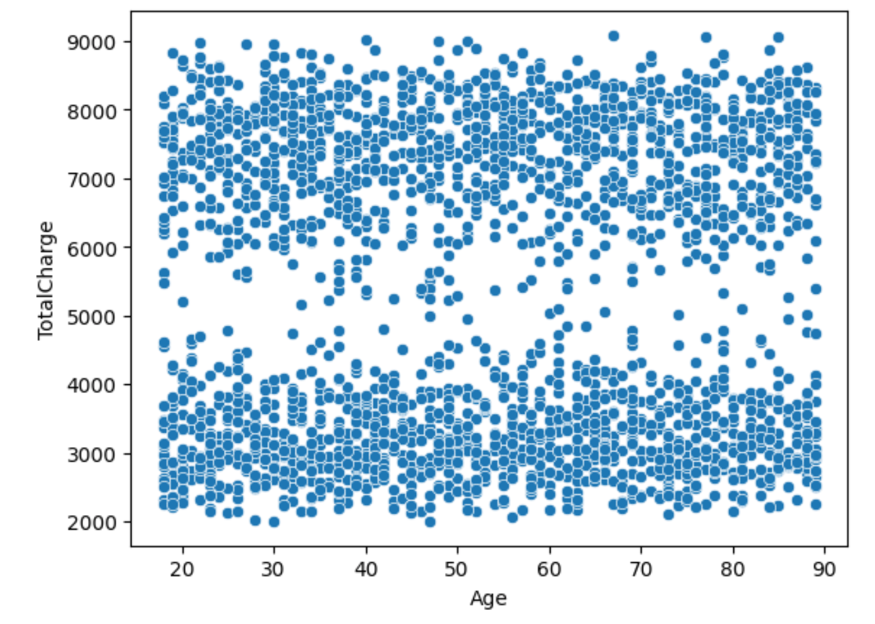
C3. Bivariate and Univariate Graphs

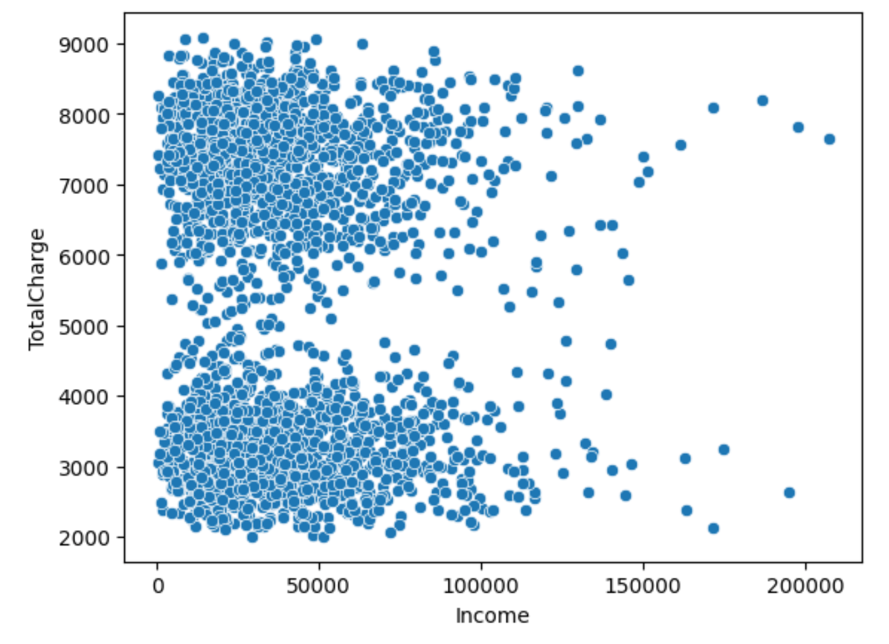
A group of blue bars

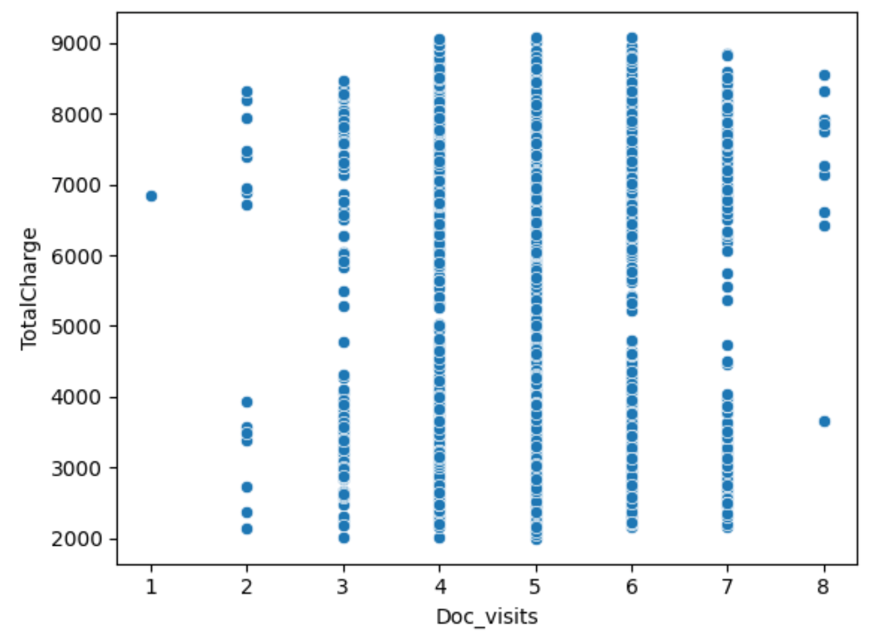
Description automatically generatedA graph of different sizes and colors

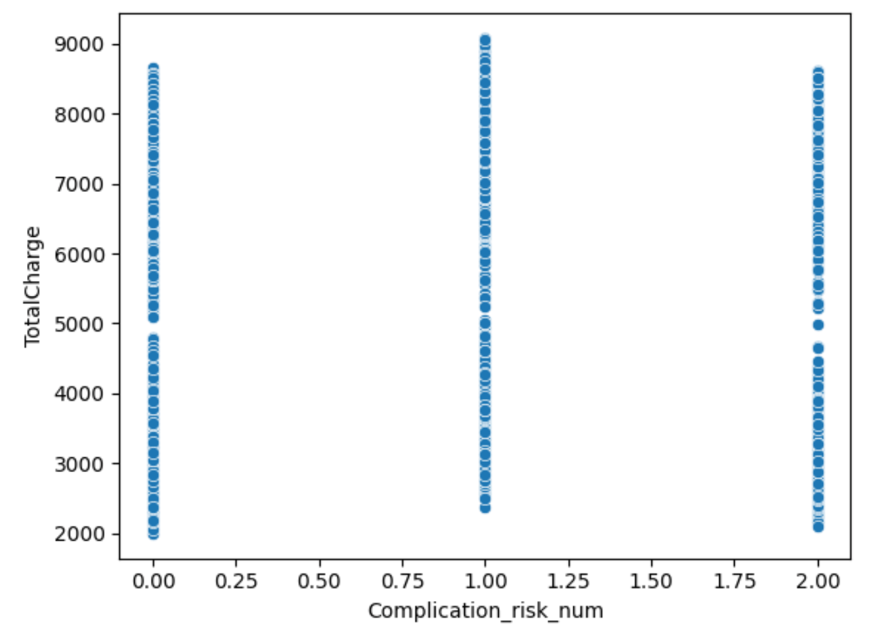
Description automatically generated with medium confidence

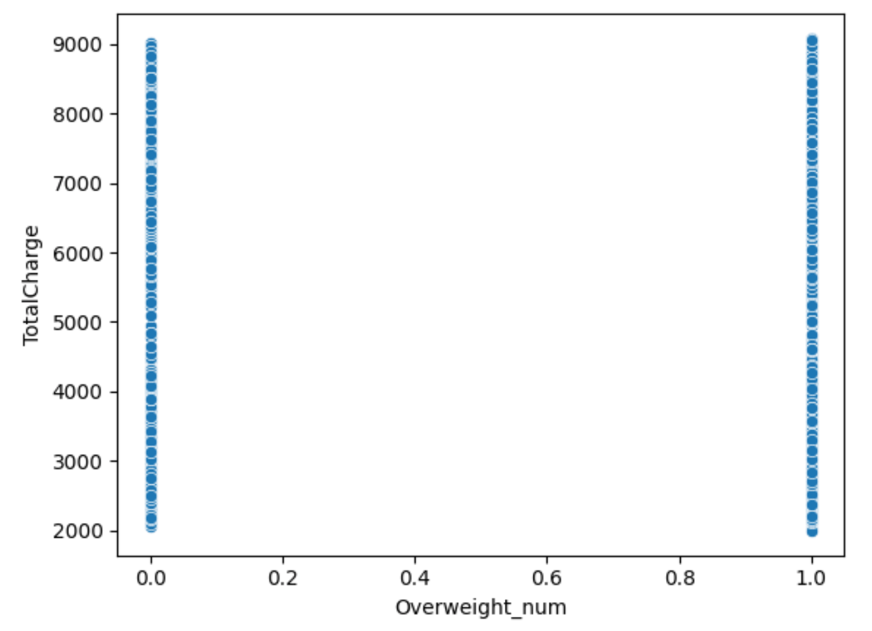


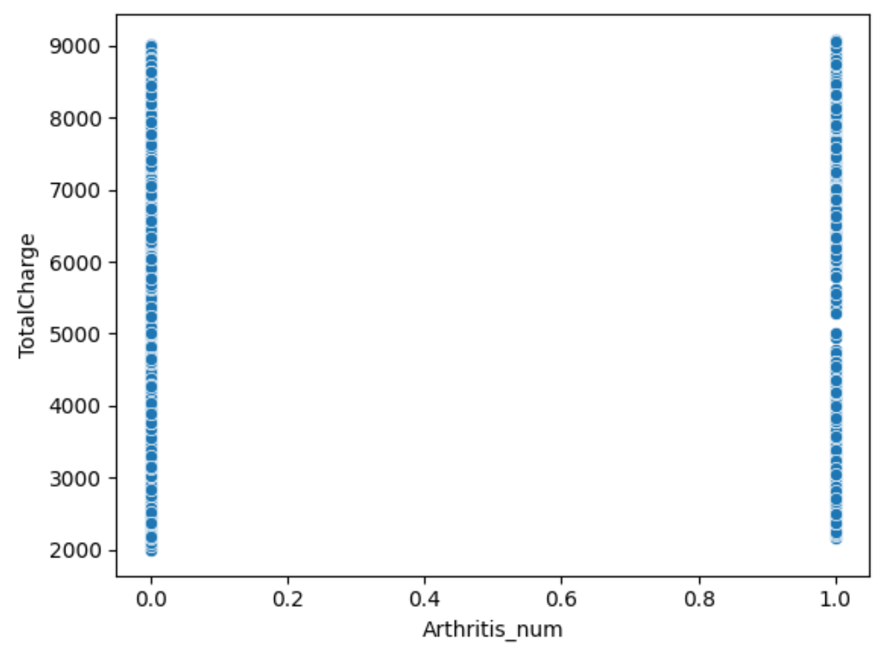






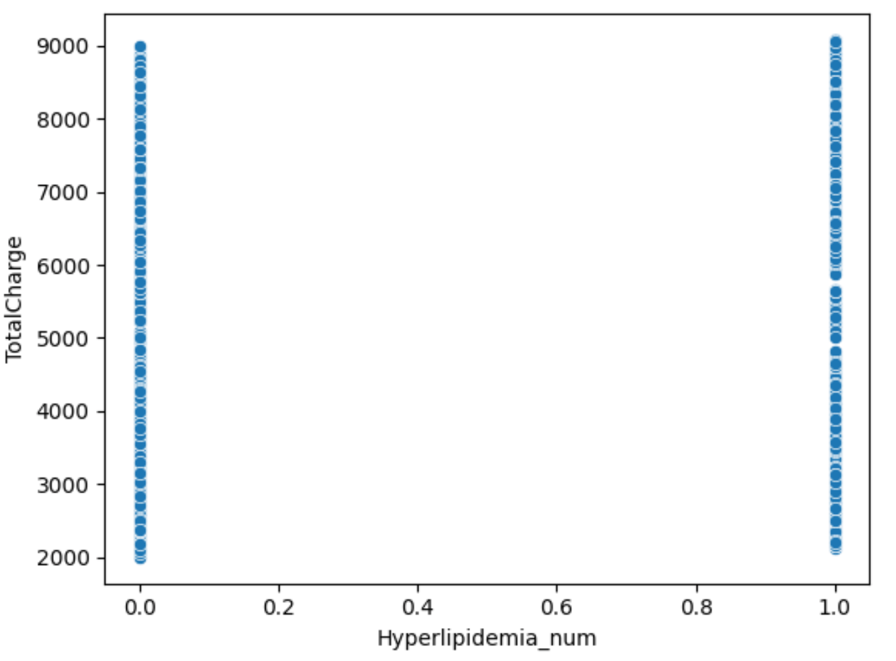


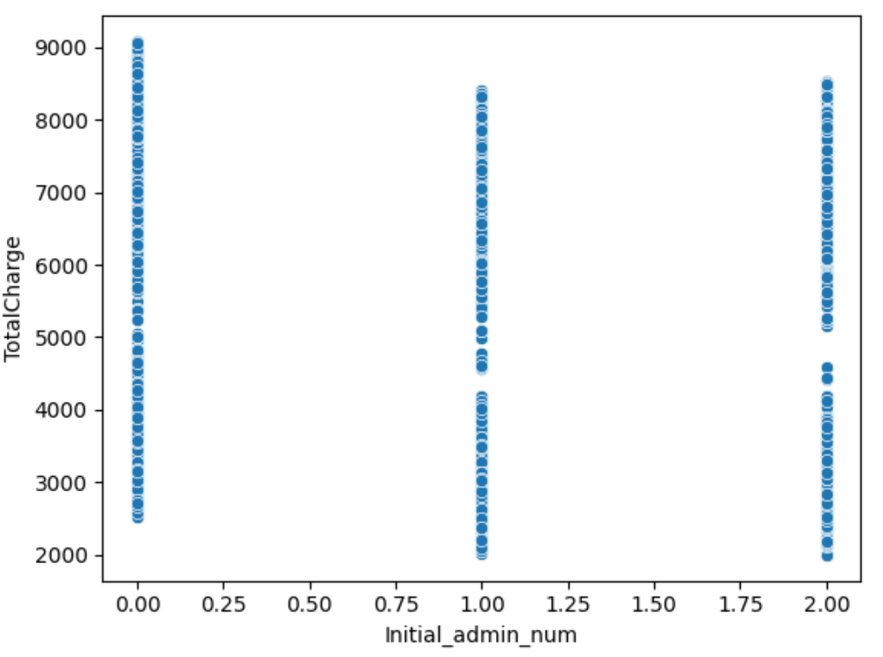


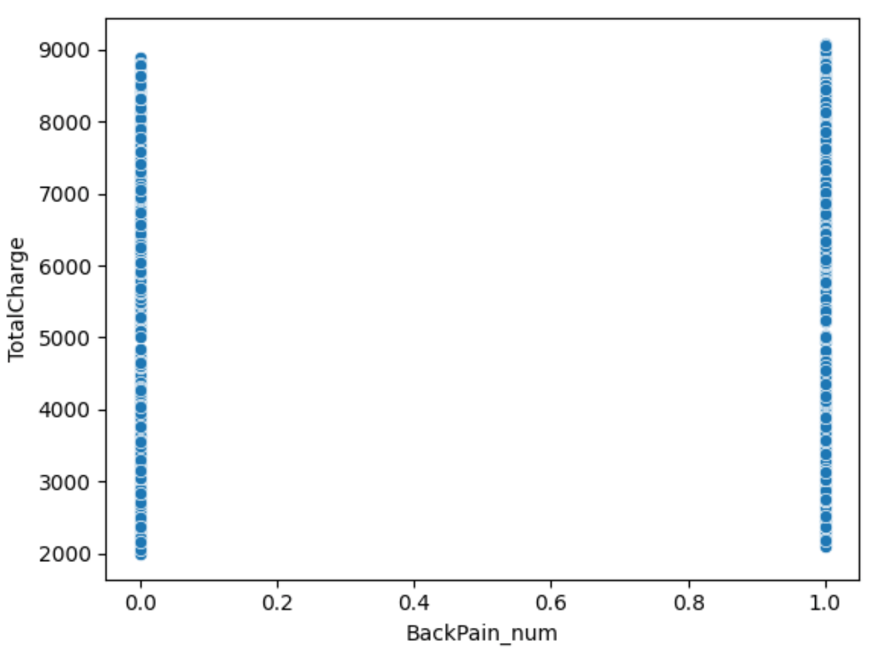


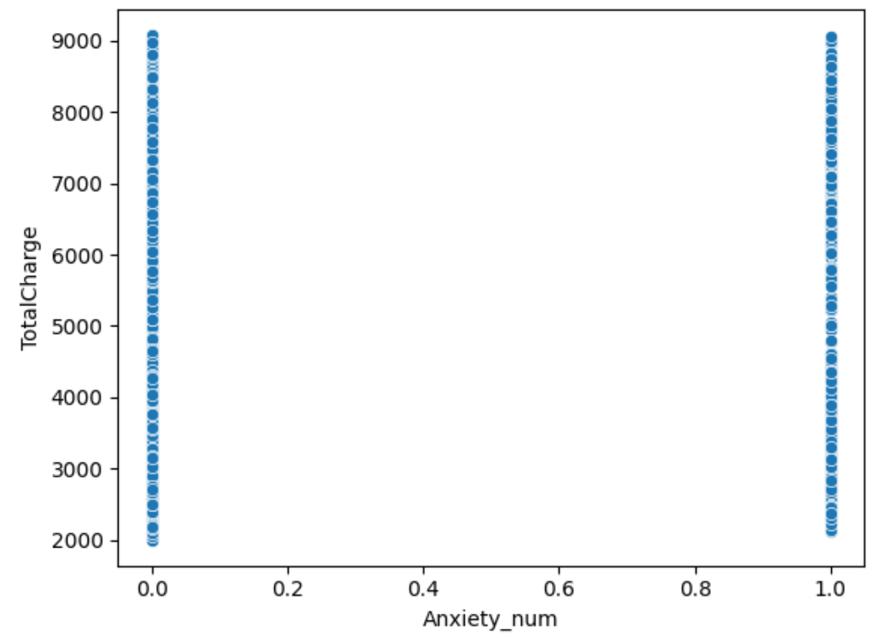
A graph of diabetes

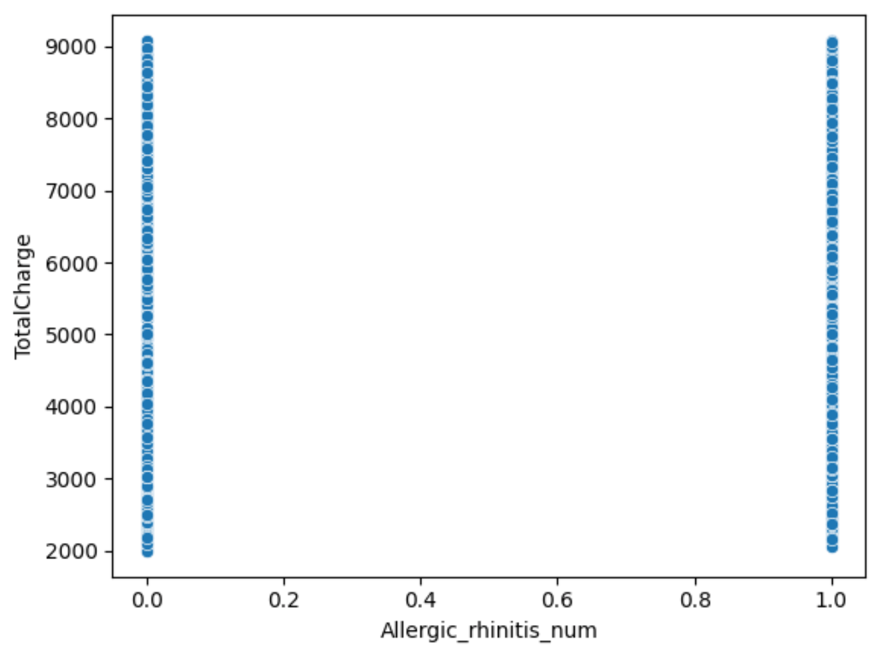
Description automatically generated











A graph of a number of blue dots

Description automatically generated

A graph of a number of blue dots

Description automatically generated

C4. Data Preparation

As detailed in the data preparation goals and necessary manipulation section above, the data will be cleaned using some of the same code from D206, with a few modifications. Categorical columns will be cast as categorical types instead of string objects. Boolean categorical columns will be remapped to integers (1 for True/Yes, 0 for False/No). Currency columns will be rounded to reflect the handling of US currency more accurately, avoiding six decimal places of precision. Survey response columns will be cast as ordered categorical datatypes and their scales will be reversed so that 1 is less than 8, instead of the current process where 1 is greater than 8. Additionally, categorical variables used in multiple regression analysis will be converted to dummy variables using one-hot encoding. A new dataframe will be created for ease of use, containing only the necessary columns for regression and excluding the approximately 20 columns that are not needed.

C5. Prepared Data Set

See attached CSV file

**Part IV: Model Comparison and Analysis**

D1. Initial Multiple Linear Regression Mode

#Create Model

feature = ['Children', 'Age', 'Income', 'Doc\_visits', 'Complication\_risk\_num', 'Overweight\_num',

'Arthritis\_num', 'Diabetes\_num', 'Hyperlipidemia\_num', 'Initial\_admin\_num',

'BackPain\_num', 'Anxiety\_num', 'Allergic\_rhinitis\_num', 'Reflux\_esophagitis\_num',

'Asthma\_num', 'Initial\_days', 'vitD\_supp', 'VitD\_levels', 'Additional\_charges',

'ReAdmis\_num', 'HighBlood\_num', 'Stroke\_num']

X = df[feature]

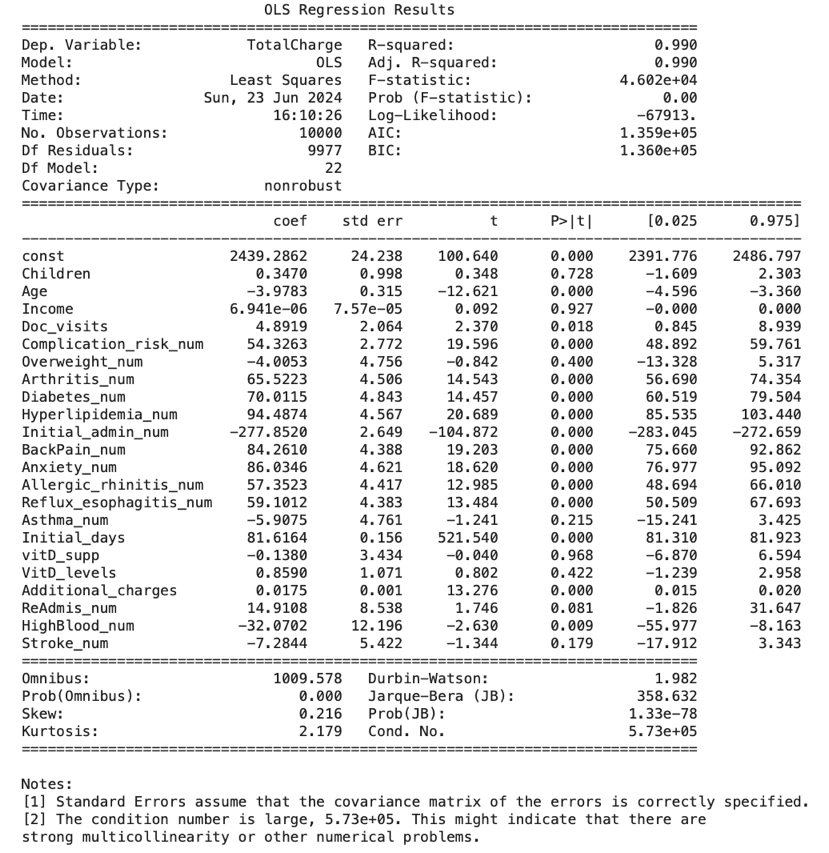
y = df['TotalCharge']

X = sm.add\_constant(X)

#Create model [6]

model = sm.OLS(y, X).fit()

print(model.summary())



D2. The Selection

I carefully decrease the number of predictors by repeatedly eliminating the traits that aren't very important, all the while maintaining the model's ability to forecast the future. This aligns with the research topic as it focuses on the most significant determinants of readmission. Commence with the full model, which includes all of the predictors. At each stage, eliminate the predictor with the highest p-value (greater than a preset significance level, such as 0.05) that is not significantly contributing to the model. Refit the model and evaluate its performance after the predictor has been removed. Continue this process until all remaining predictors have p-values that are statistically significant.

D3. Reducing the Model

The results indicate that the predictors Age, Doc\_visits, Complication\_risk\_num, Arthritis\_num, Diabetes\_num, Hyperlipidemia\_num, Initial\_admin\_num, BackPain\_num, Anxiety\_num, Allergic\_rhinitis\_num, Reflux\_esophagitis\_num, Initial\_days, Additional\_charges, and HighBlood\_num all have significant effects on TotalCharge. The high R-squared value shows the model explains nearly all the variability in TotalCharge, although the diagnostics suggest some issues with normality in the residuals.

#Re-Prepare data

feature = ['Age', 'Doc\_visits', 'Complication\_risk\_num',

'Arthritis\_num', 'Diabetes\_num', 'Hyperlipidemia\_num', 'Initial\_admin\_num',

'BackPain\_num', 'Anxiety\_num', 'Allergic\_rhinitis\_num', 'Reflux\_esophagitis\_num',

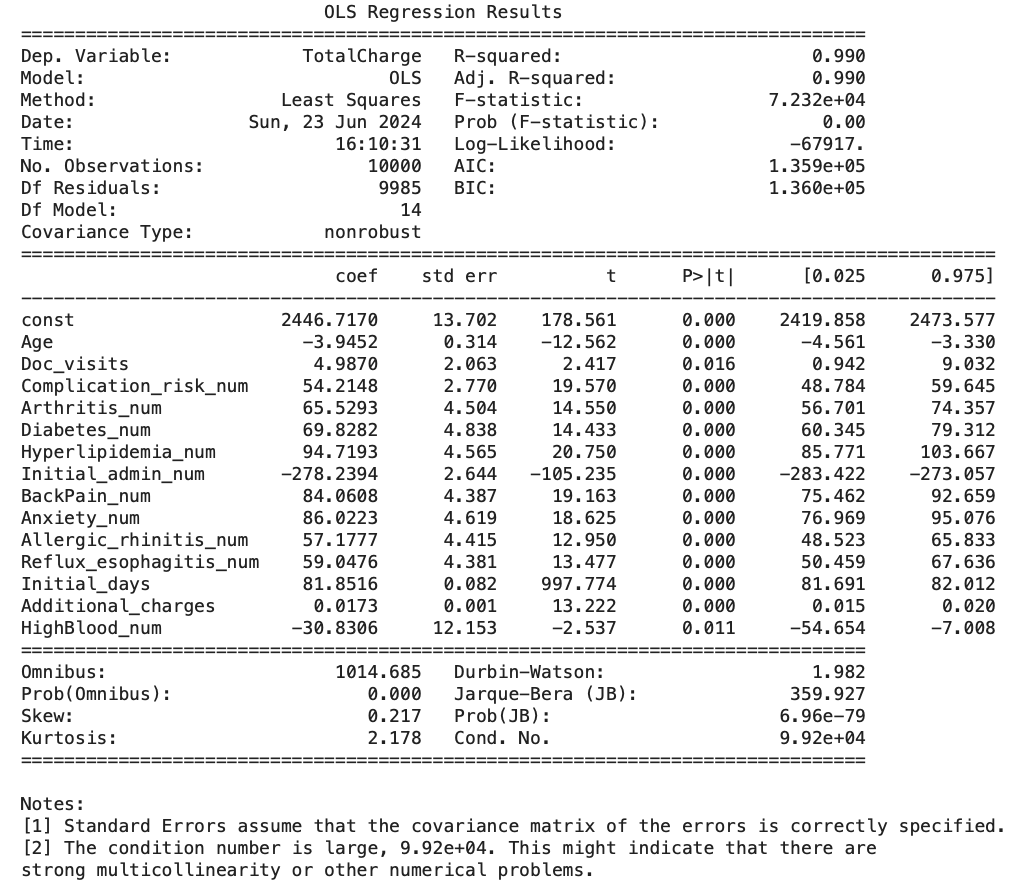
'Initial\_days', 'Additional\_charges',

'HighBlood\_num']

X = df[feature]

y = df['TotalCharge']

X = sm.add\_constant(X)



E1. Reduced Model Explanation

In order to keep good predictive accuracy while improving simplicity and interpretability, a simplified linear regression model was created. Predictors with high p-values were removed from the original model in order to create this simplified model since they were not significantly helpful in explaining TotalCharge. With minimal loss of explanatory power, this method sought to develop a more frugal model. Then, using measures like modified R-squared and AIC, the reduced model was fitted and contrasted with the original model. The simplified model continued to explain 99.0% of the variability in TotalCharge, with an adjusted R-squared of 0.990, the same as the original model, despite the removal of some predictors. Additionally, the AIC of the simplified model was marginally lower, indicating a better trade-off between complexity and model fit.

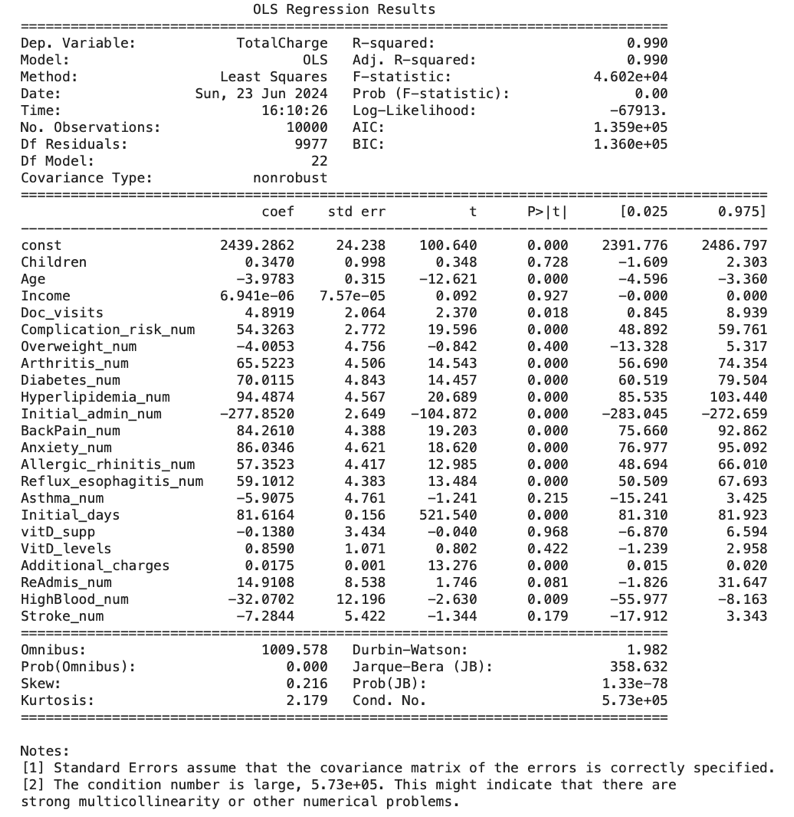
E2. Python Outputs

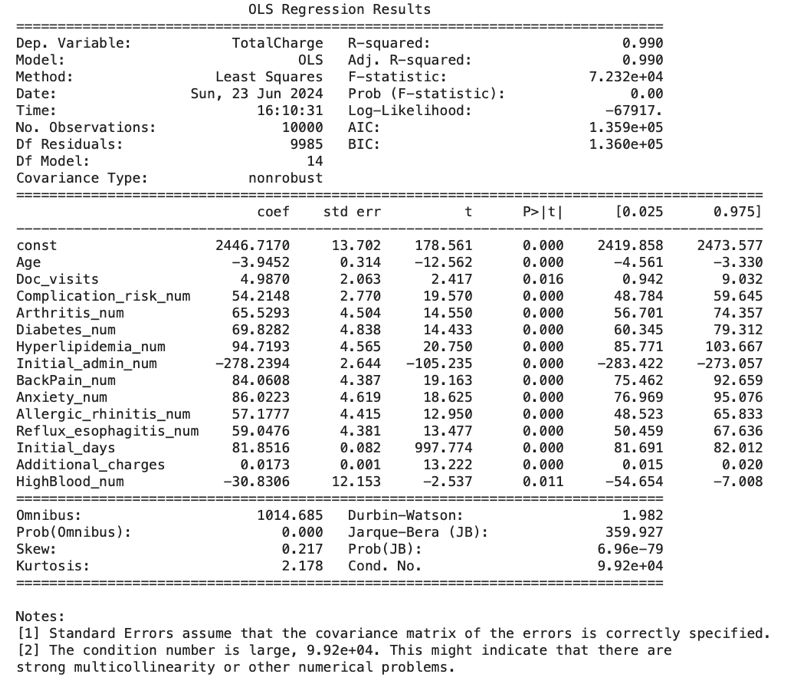
A screenshot of a computer game

Description automatically generated

A blue and white dotted graph

Description automatically generated





E3. Python Code

See attached File

**Part V: Data Summary and Implications**

F1. Results

To explain the dependent variable, TotalCharge, the first multiple linear regression model was created utilizing a wide range of predictors as part of the data analysis procedure. Variables like Age, Doc visits, Complication risk number, Arthritis number, Diabetes number, Hyperlipidemia number, Initial admin number, Back pain number, Anxiety number, Allergic rhinitis number, Reflux esophagitis number, Initial days, Additional charges, and High blood count number were included in this model. Key indicators were used to evaluate the model's performance after Ordinary Least Squares (OLS) regression was used to fit it. With a high R-squared value of 0.990, the model demonstrated that the predictors accounted for 99.0% of the variability in TotalCharge. Even after taking the number of predictors into account, the adjusted R-squared was 0.990, indicating that the model remained robust. The model's validity was further reinforced by the AIC and BIC values, which were 1.359e+05 and 1.360e+05, respectively. Additionally, most predictors had p-values indicating significant contributions to the model.

A simplified linear regression model was then created in order to keep good, predicted accuracy while improving interpretability and simplicity. Predictors with high p-values were removed from the original model in order to create this simplified model since they were not significantly helpful in explaining TotalCharge. With minimal loss of explanatory power, this method sought to develop a more frugal model. Then, using measures like modified R-squared and AIC, the reduced model was fitted and contrasted with the original model. The simplified model continued to explain 99.0% of the variability in TotalCharge, with an adjusted R-squared of 0.990, the same as the original model, despite the removal of some predictors. Additionally, the AIC of the modified model was somewhat lower, indicating a better balance between model fit and complexity.

Overall, the unchanged adjusted R-squared of 0.990 indicates that the reduced model maintains explanatory power while offering a more simplified and easily understood framework. The decreased model is more appropriate for real-world applications while maintaining strong prediction abilities, as seen by the somewhat lower AIC, which also suggests an improved model fit in relation to complexity. This procedure demonstrates how model reduction helps regression analysis strike a compromise between robustness and simplicity.

Intercept (Const): The intercept coefficient is 2446.7170. This represents the expected value of the dependent variable (TotalCharge) when all predictor variables are zero.

Age: The coefficient for Age is -3.9452. This indicates that for each one-unit increase in Age, the TotalCharge decreases by approximately 3.9452 units, holding all other variables constant.

Doc\_visits: The coefficient for Doc\_visits is 4.9870. This suggests that each additional doctor visit is associated with an increase in TotalCharge by approximately 4.9870 units.

Complication\_risk\_num: The coefficient is 54.2148, indicating that higher complication risk is associated with an increase in TotalCharge by approximately 54.2148 units.

Arthritis\_num: The coefficient is 65.5293, indicating that patients with arthritis have a higher TotalCharge by approximately 65.5293 units.

Diabetes\_num: The coefficient is 69.8282, indicating that patients with diabetes have a higher TotalCharge by approximately 69.8282 units.

Hyperlipidemia\_num: The coefficient is 94.7193, indicating that patients with hyperlipidemia have a higher TotalCharge by approximately 94.7193 units.

Initial\_admin\_num: The coefficient is -278.2394. This suggests that initial administration (likely emergency) is associated with a decrease in TotalCharge by approximately 278.2394 units.

BackPain\_num: The coefficient is 84.0608, indicating that patients with back pain have a higher TotalCharge by approximately 84.0608 units.

Anxiety\_num: The coefficient is 86.0223, indicating that patients with anxiety have a higher TotalCharge by approximately 86.0223 units.

Allergic\_rhinitis\_num: The coefficient is 57.1777, indicating that patients with allergic rhinitis have a higher TotalCharge by approximately 57.1777 units.

Reflux\_esophagitis\_num: The coefficient is 59.0476, indicating that patients with reflux esophagitis have a higher TotalCharge by approximately 59.0476 units.

Initial\_days: The coefficient is 81.8516, indicating that each additional initial day is associated with an increase in TotalCharge by approximately 81.8516 units.

Additional\_charges: The coefficient is 0.0173, indicating that each additional charge unit is associated with an increase in TotalCharge by approximately 0.0173 units.

HighBlood\_num: The coefficient is -30.8306. This indicates that patients with high blood pressure have a lower TotalCharge by approximately 30.8306 units.

F2. Recommendation

Regression analysis results for forecasting TotalCharge are used to inform a number of recommendations that enhance model applicability and performance. It is advised to use regularization strategies like Lasso or Ridge regression to address potential overfitting and multicollinearity. This will assist by concentrating on the most important factors and simplifying the model. Additionally, interpretability can be preserved without compromising accuracy by streamlining the model through iterative predictor selection.

Predictive power may be further increased by enlarging the dataset to include a wider variety of patient health and lifestyle characteristics. It is advised to use cross-validation methods such as k-fold cross-validation and out-of-sample testing to guarantee the resilience of the model. These techniques will lessen the chance of overfitting and validate the model's generalizability. Deeper insights and improved model performance in practical applications can also be obtained by concentrating on predictors of practical value and investigating their interactions and transformations.

**Part VI: Demonstration**

G1. Panopto

See Attached Linked

H. Third-party Sorce

Biswal, A. (2023, April 3). *Sklearn Linear Regression*. Simplilearn.com. https://www.simplilearn.com/tutorials/scikit-learn-tutorial/sklearn-linear-regression-with-examples

Python, R. (2023, June 26). *Linear regression in Python*. https://realpython.com/linear-regression-in-python/

*Understanding logistic Regression in Python Tutorial*. (2019, December). DataCamp. Retrieved June 1, 2024, from https://www.datacamp.com/tutorial/understanding-logistic-regression-python

I. Citations

Team, I. C. (2024, May 24). *Python vs. R: What’s the Difference? - IBM Blog*. IBM Blog. https://www.ibm.com/blog/python-vs-r/