**Western Governors University (WGU)**

**D208 Predictive Modeling Performance Assessment**

**Task 1: Linear Regression Modeling**

**Natallia Zimnitskaya | ID: 012247127**

**Master of Science, Data Analytics**

***Part I: Research Question***

***A1. Question***

What factors correlated to Additional\_charges?

***A2. Objectives and Goals for Analysis***

The primary objective of this analysis is to identify and quantify the factors correlated with Additional\_charges. By analyzing variables such as age, gender, number of doctor’s visits, medical conditions, type of services provided, etc, the analysis aims to uncover patterns and insights that can help optimize healthcare resource allocation, improve patient care strategies, and reduce unnecessary costs. Organizations can optimize resource allocation and improve patient care strategies, enhancing operational efficiency. Stakeholders gain valuable insights for informed decision-making regarding policy changes, investments, and financial planning. Patients benefit from personalized care plans that address specific needs, reducing unnecessary costs and improving health outcomes. The goal is to build a multiple regression model that can predict Additional\_charges based on significant variables and to interpret the model’s coefficients to understand the impact of each factor.

***Part II: Method Justification***

***B1. Assumptions***

Regression models describe relationships between variables by fitting a line to the observed data—regression to estimate how a dependent variable changes as the independent variable(s) change.

Multiple linear regression estimates the relationship between two or more independent variables and one dependent variable (*Multiple et al. | A Quick Guide (Examples)*, n.d.).

Here are four assumptions of a multiple linear regression model when creating them:

* Linearity: The relationship between the independent variables (Age, Doc\_visits) and the dependent variable (Additional\_charges) should be linear. For example, as the number of doctor visits increases, the additional charges should increase or decrease in a straight-line manner.
* Independence: The observations should be independent of each other. For instance, the additional charges for one patient should not be influenced by the additional charges for another patient.
* Homoscedasticity: The spread of the errors should be the same no matter the value of the independent variable. For example, whether a patient eats 1 meal or 3 meals a day, the difference between the actual additional charges and the predicted charges should be similar.
* Normality: The residuals (differences between observed and predicted values) should be normally distributed. For example, the differences between the actual additional charges and those predicted by the model should follow a normal distribution.

***B2. Programming Language and Benefits***

After reading "R vs. Python: 12 Key Comparisons," Python was chosen to complete this performance assessment. High-level, general-purpose programming language Python is flexible, object-oriented, open-source, and emphasizes code readability with a clear visual structure and straightforward syntax. Since Python is open source, many people can contribute to its development and improve its libraries and features. Python has many necessary libraries for data science-related tasks, and its integration and control capabilities boost productivity. Python is a standard programming language that is easy for beginners to learn and understand because of its simple syntax; it requires fewer lines of code to be written and is easy to read; for data science projects, Python takes a more streamlined approach; it has a wide range of libraries that allow users to input the action of the library into the code, making it simple to perform matrix computations and optimization (BasuMallick, 2022). However, compared to R, Python has some disadvantages: fewer data science-specific libraries, visualizations that are less visually appealing and informative, more complicated, and require thorough testing because mistakes appear during runtime.

The following Python packages and libraries help with creating a Multiple regression model:

* **Pandas:** used for data analyzing, cleaning, exploring, and manipulating.
* **NumPy:** used to working with arrays.
* **Matplotlib:** used for visualization utility.
* **CategoricalDtype:** helps define categorical data types.
* **Seaborn:** used for advanced visualization.
* **spicy.stats:** used for normalization and statistics.
* **statsmodels.api**: used for conducting statistical tests and exploring statistical data.
* **variance\_inflation\_factor:** used for detection of multicollinearity in the regression model.
* **scipy:** used for optimization, integration, interpolation, eigenvalue problems, and other computations, including ANOVA and point-biserial correlation.
* **statsmodels.api:** used for conducting statistical tests and exploring statistical data.
* **sklearn.preprocessing:** used for scaling and encoding categorical variables.
* **train\_test\_split:** used to split the dataset into training and testing sets for evaluating the performance of the regression model.
* **LinearRegression:** used to fit the regression model to the data.
* **OneHotEncoder:** used to convert categorical variables into a format that can be provided to ML algorithms.
* **mean\_squared\_error:** used to evaluate the performance of the regression model.
* **cross\_val\_score**: used to evaluate the performance of the regression model using cross-validation.

***B3. Justification of using Regression***

Multiple linear regression is a statistical tool used to model the relationship between a continuous responses (dependent) variable and one or more

continuous and/or categorical explanatory (independent) variables (*Redirecting*, n.d., p. 22). It helps understand how each factor individually and collectively influences the additional charges while controlling for the effects of other variables. Using multiple linear regression, we can quantify the impact of each factor, identify significant predictors, and make more accurate predictions about additional charges. This analysis will focus on the following independent variables: Age, Gender, ReAdmis, Doc\_visits, Full\_meals\_eaten, vitD\_supp, Initial\_admin, Stroke, Complication\_risk, Arthritis, Diabetes, Services, and Initial\_days.

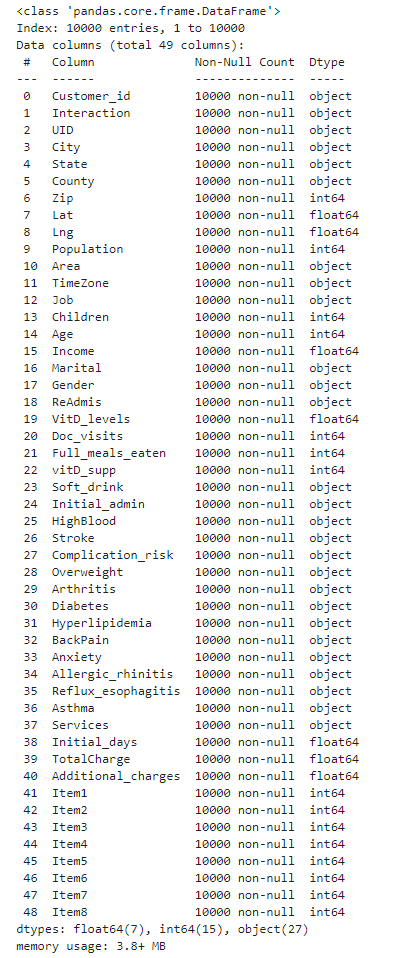
***Part III: Data Preparation and Manipulation***

***C1. Data Cleaning***

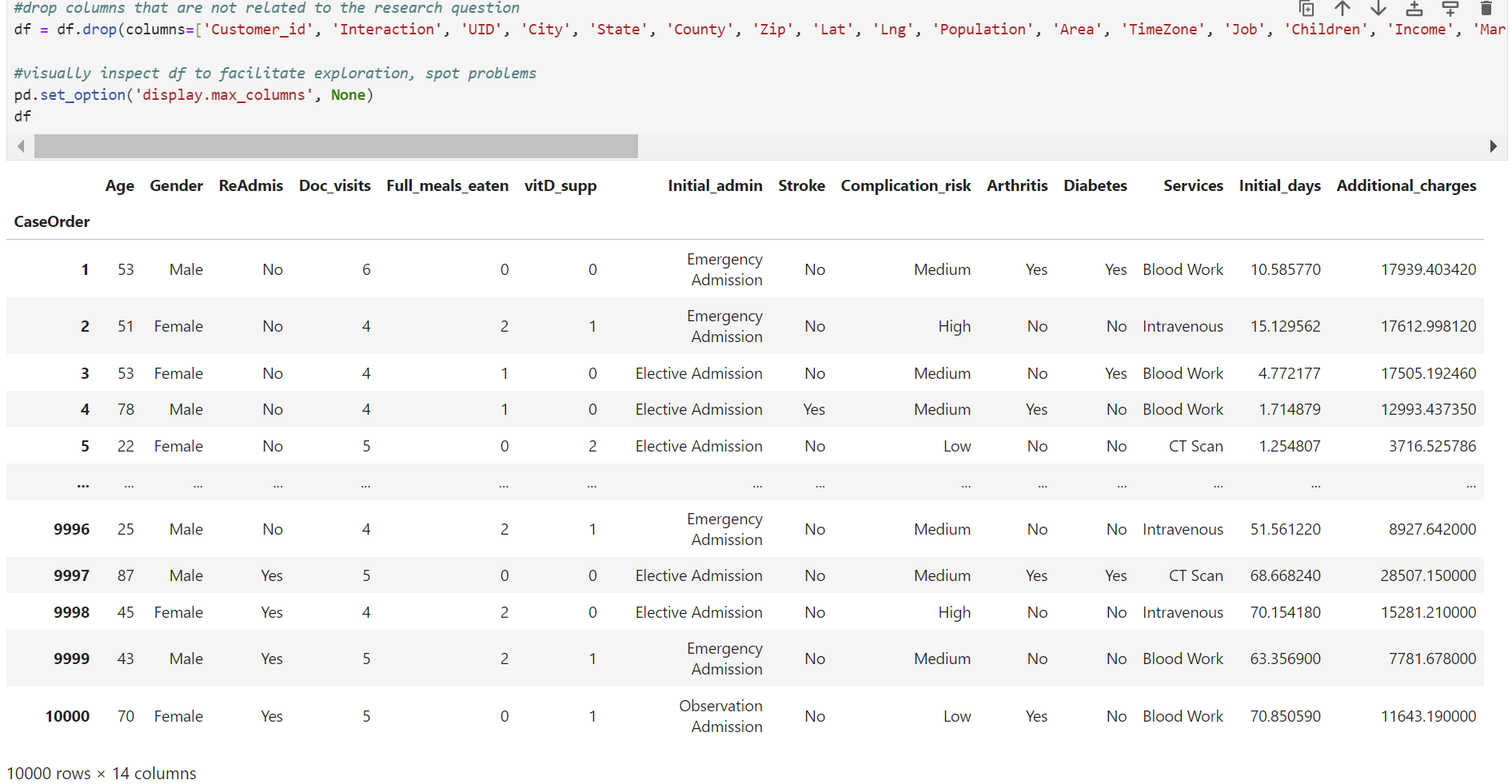
The primary goals for data cleaning are to ensure the dataset is accurate, consistent, and ready for analysis. This involves handling duplicates and missing values, detection and treatment of outliers, correcting errors, and standardizing formats.

The first step is to import required packages and libraries, then using the pd.read\_csv() function, load the medical\_clean.csv file onto the Jupyter Notebook, and lastly, with df.info() function, get the information about the dataset.

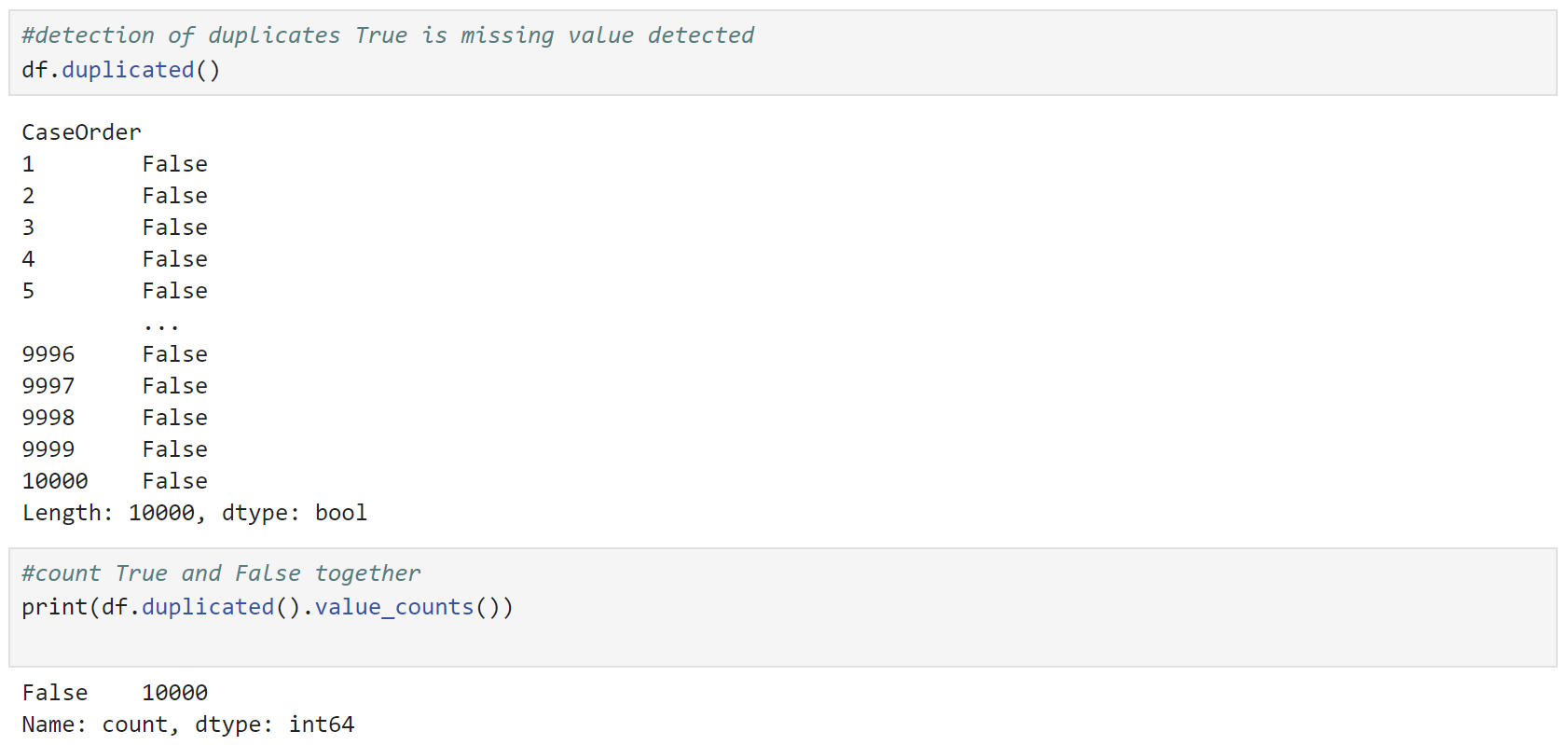




The next step is to drop variables unrelated to the research question using the df.drop() function. With the pd.set\_option() function, visually inspect the dataset to facilitate exploration and spot problems

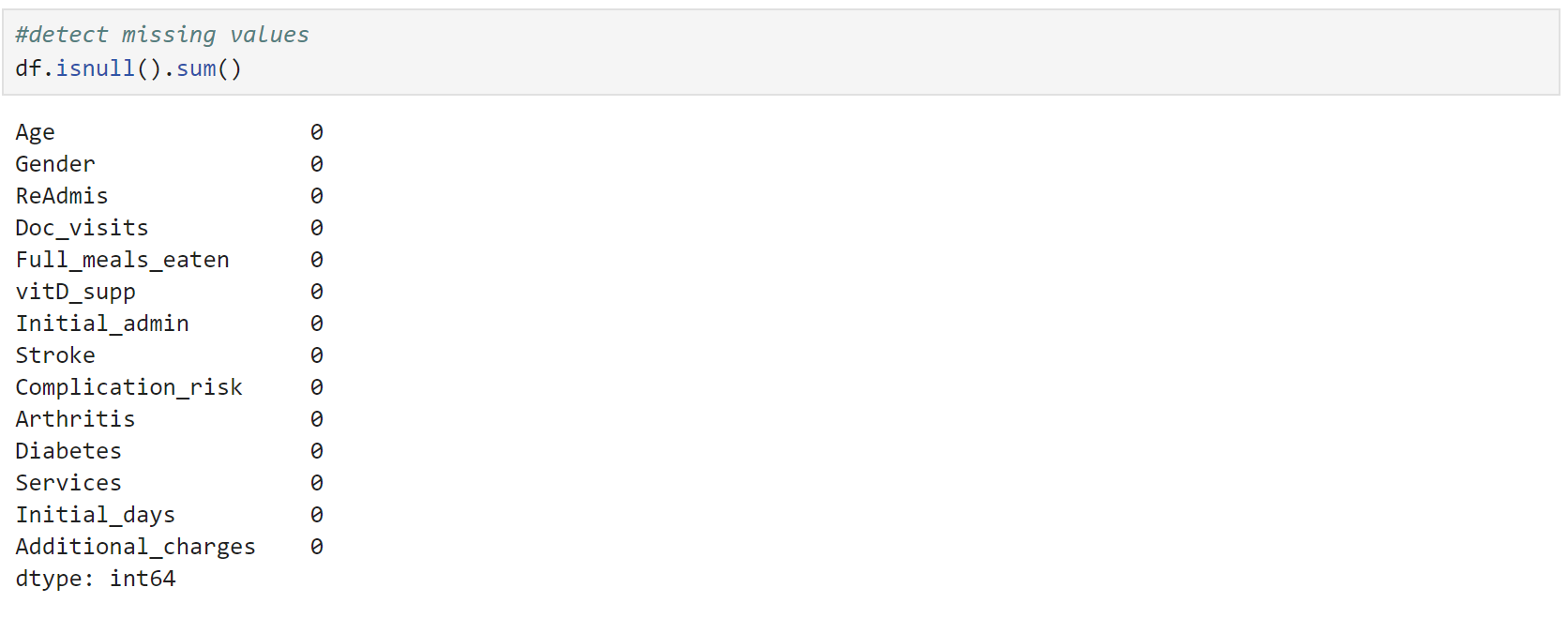


For the detection of duplicates, the duplicated() and duplicated().value\_counts() functions were used.



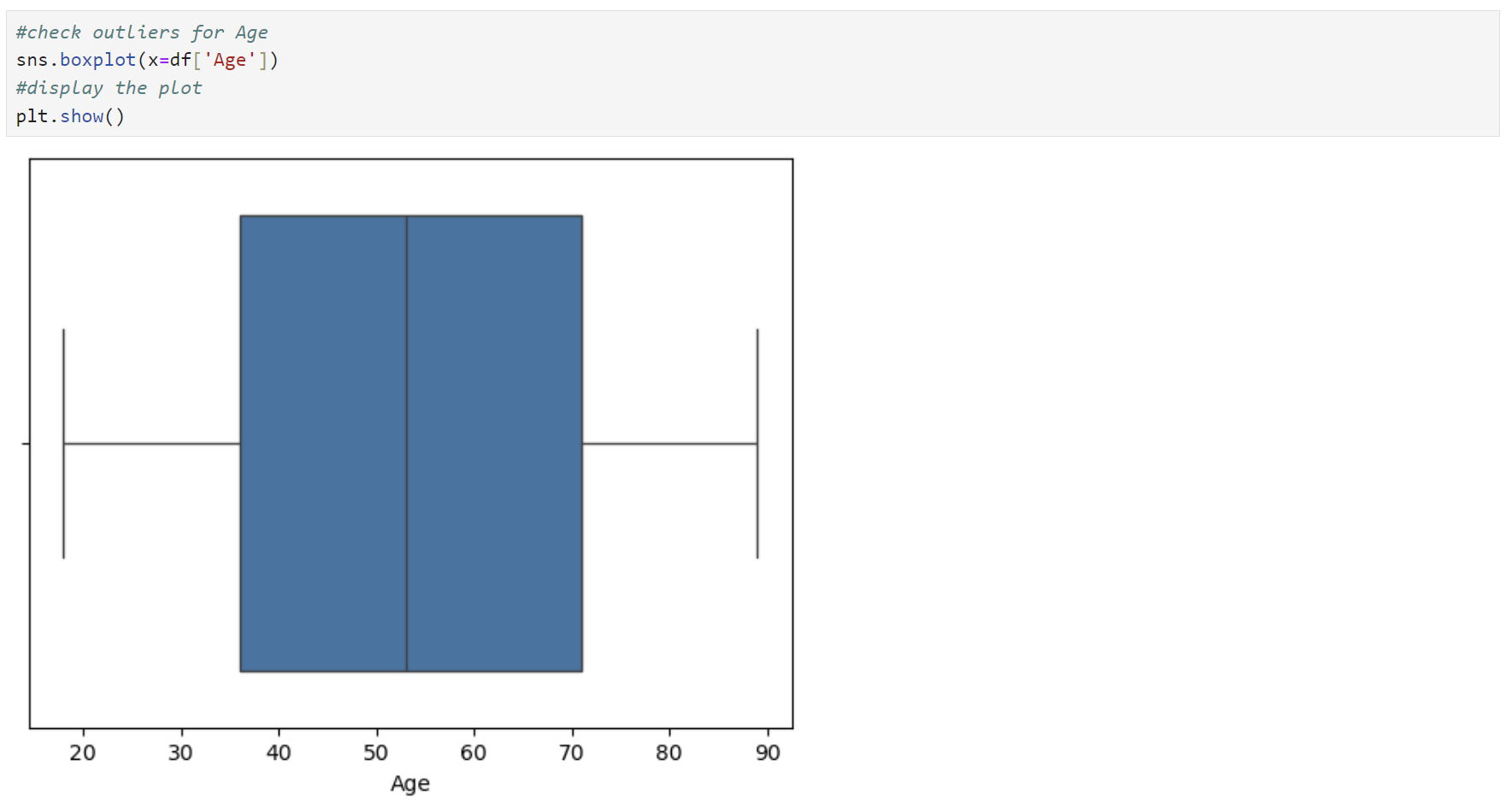
There are no duplicates in the dataset.

Next, the isnull().sum function was used to detect the missing values (qualitative and quantitative variables).

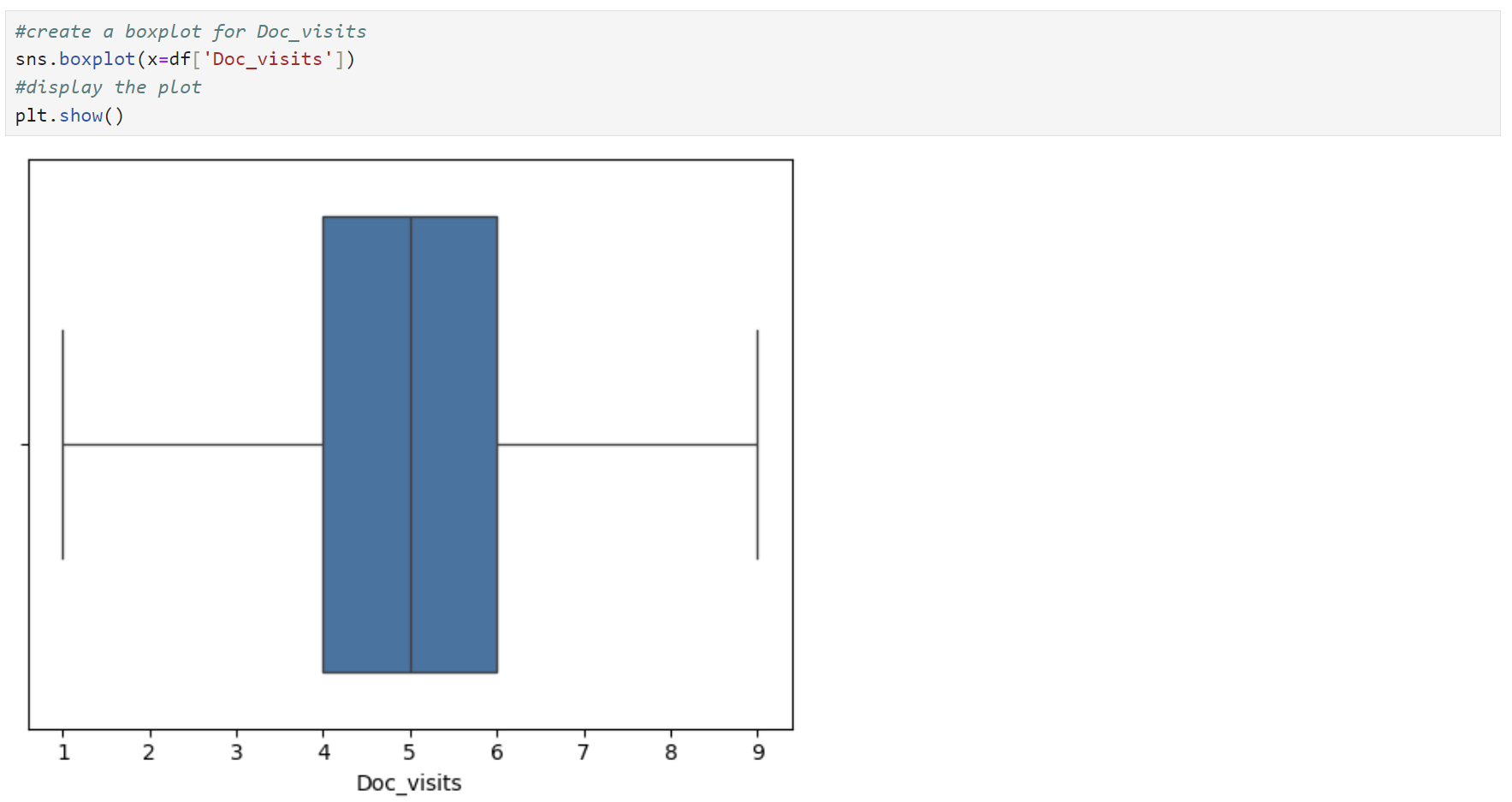


There are no missing values in the dataset.

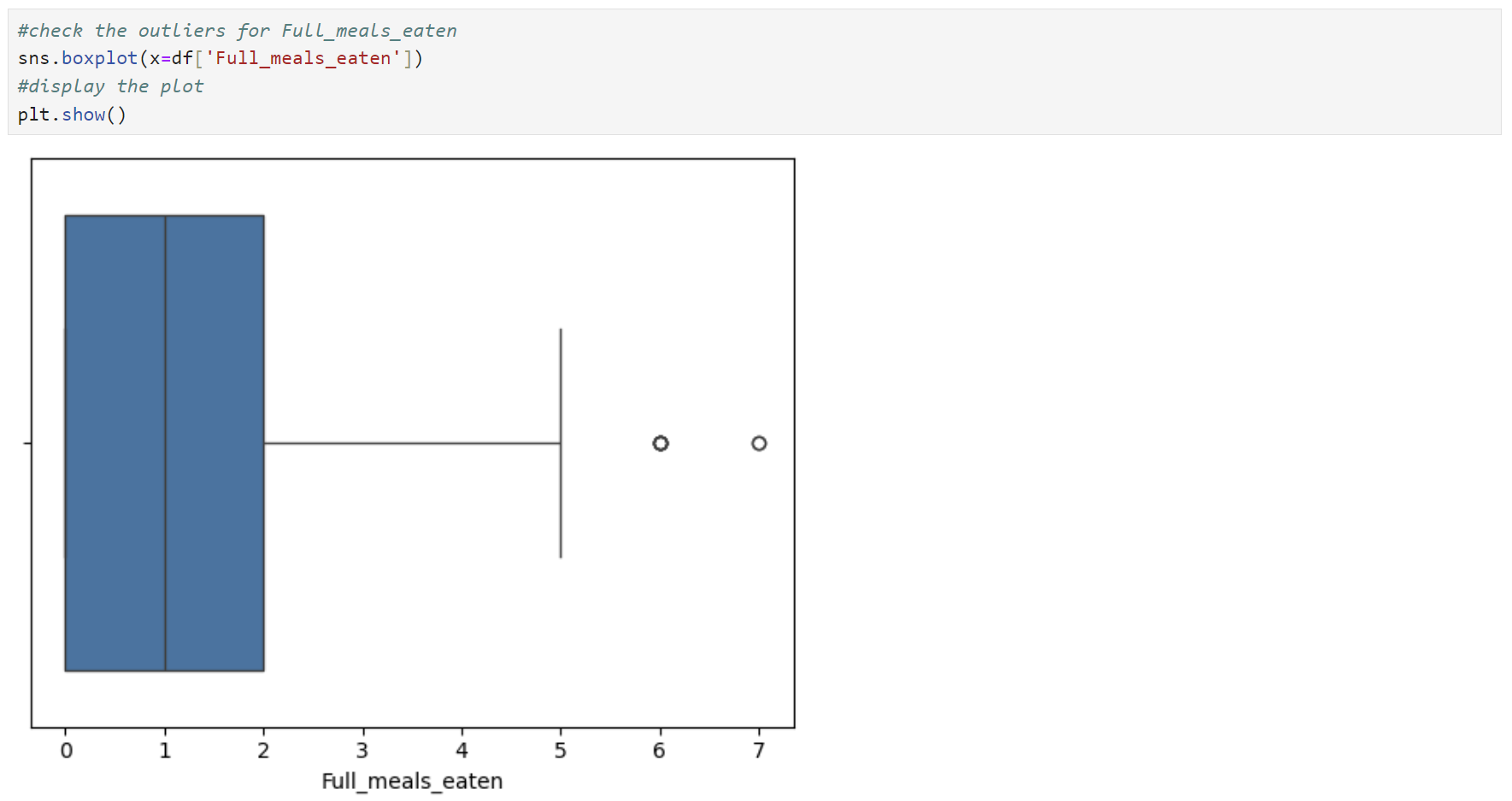
The boxplot and Z\_Score methods detect the outliers for all quantitative variables.



There are no outliers in Age



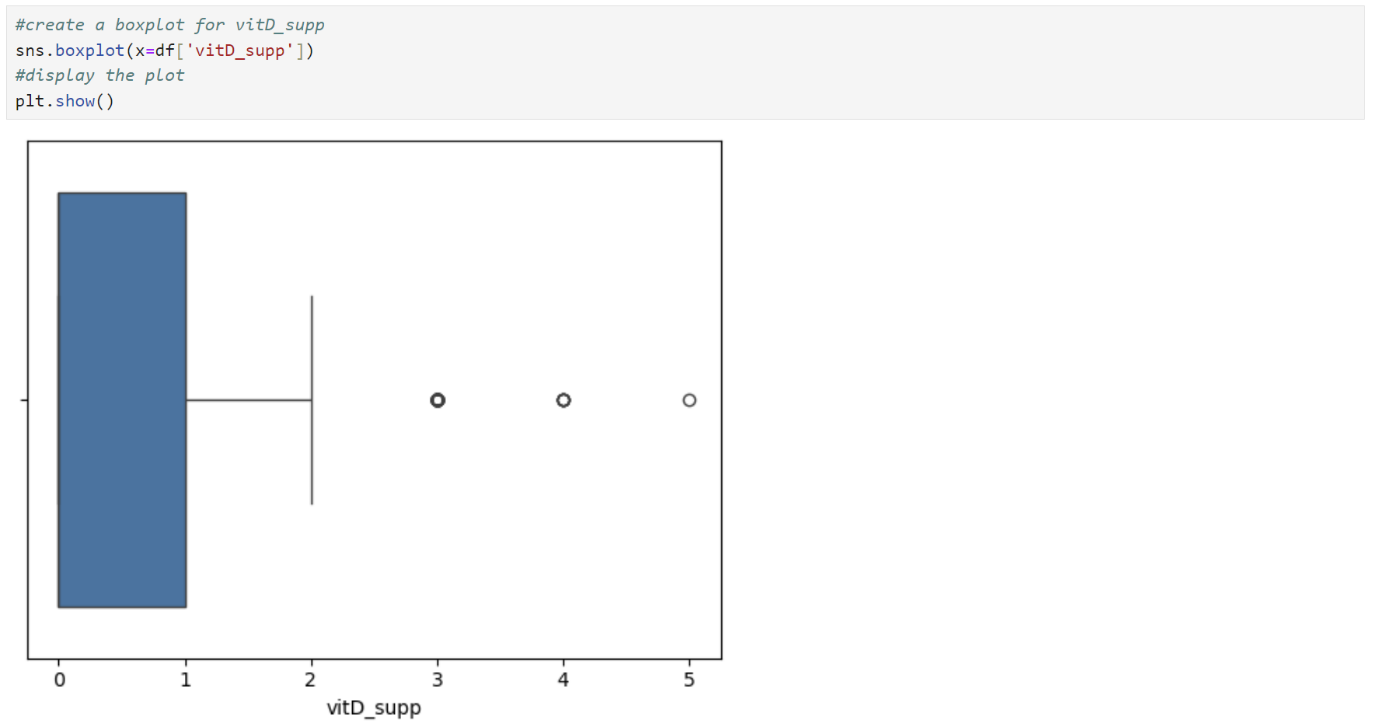
There are no outliers in Doc\_visits.



There are outliers in Full\_meals\_eaten, and the next step is to check the number and the range of outliers.

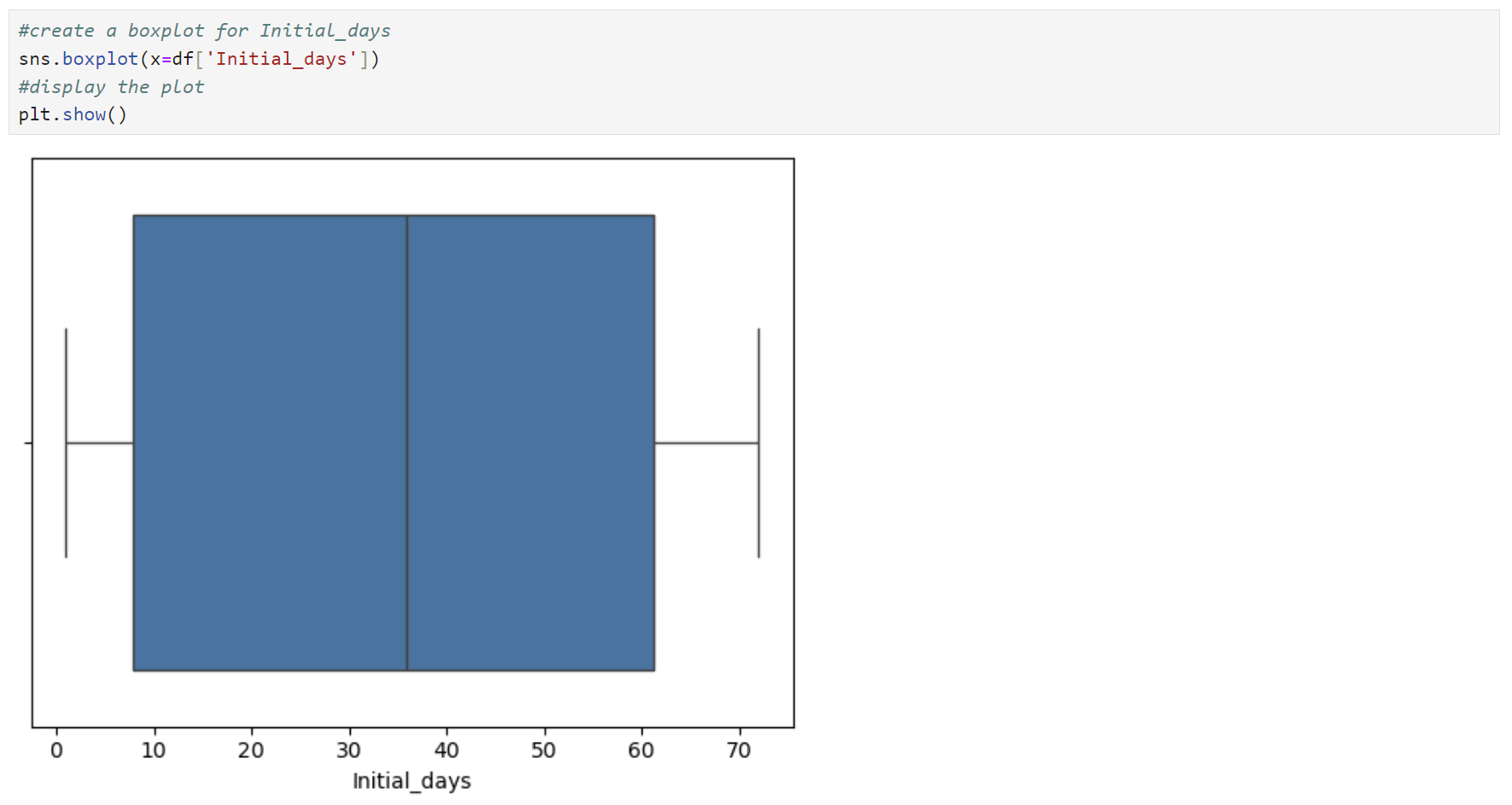


According to the medical\_clean.csv file directory, Full\_meals\_eaten is the number of full meals the patient ate while hospitalized (partial meals count as 0, and some patients had more than three meals per day if requested). The number of outliers (33 out of 10000) is very low, and their range (5, 7) is acceptable. I am going to retain the Full\_meals\_eaten outliers in the dataset.

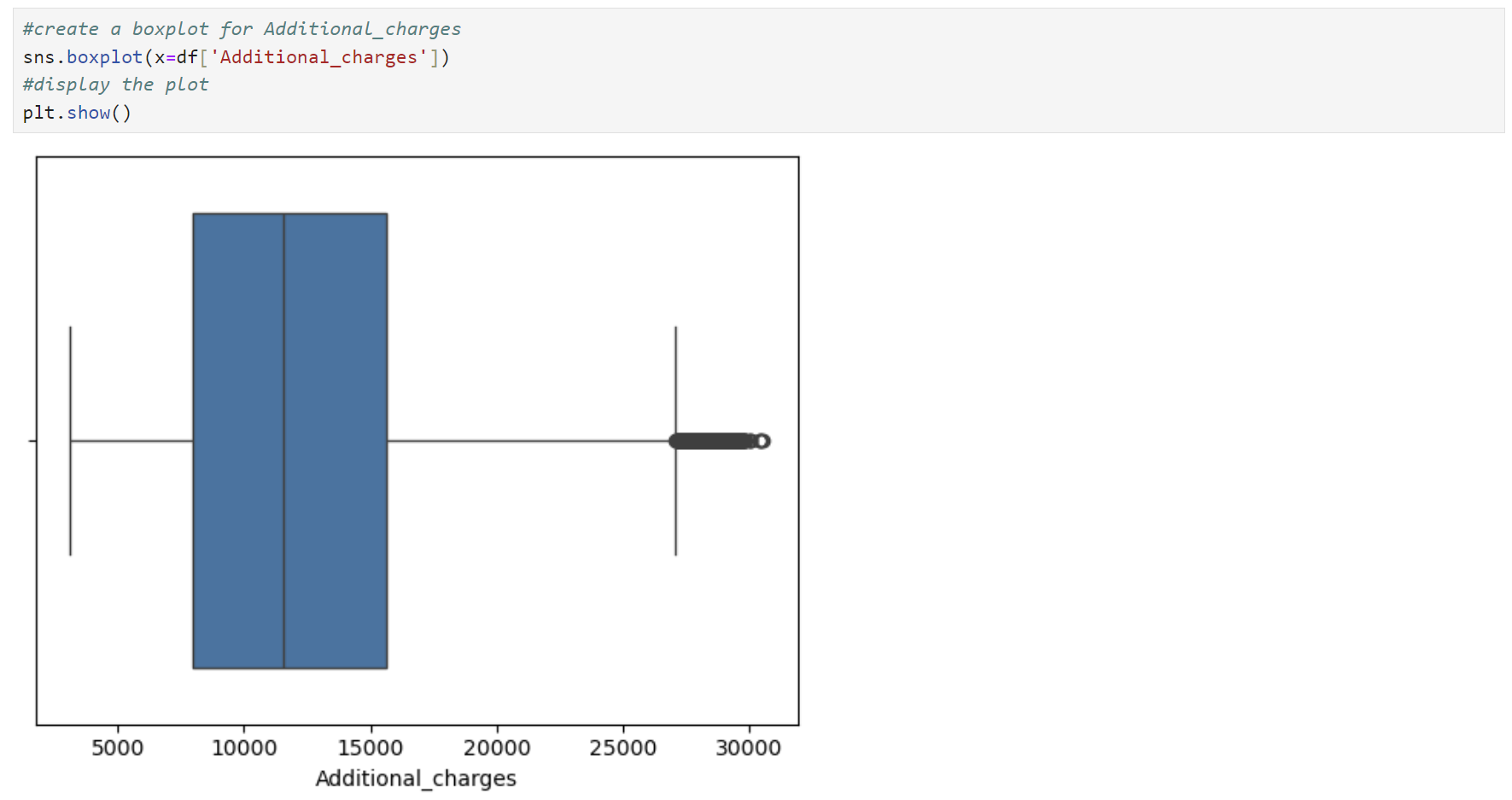


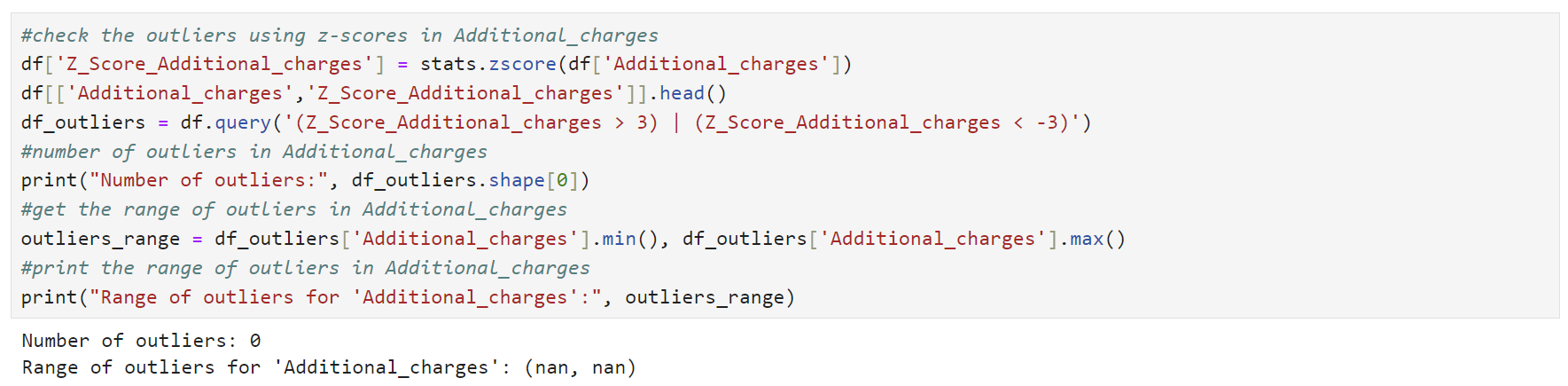


The number of outliers in vitD\_supp is low (70 out of 10000) and the range (3, 5) seems acceptable, so I will retain the outliers in the dataset.



There are no outliers in the Initial\_days.



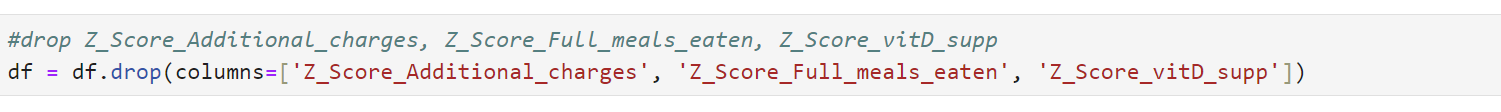


There are no outliers in the Additional\_charges.

In the next step, I will convert ReAdmis, Stroke, Arthritis, Diabetes from string to Boolean, and Gender, Initial\_admis, Complicatiom\_risk, and Services to category.

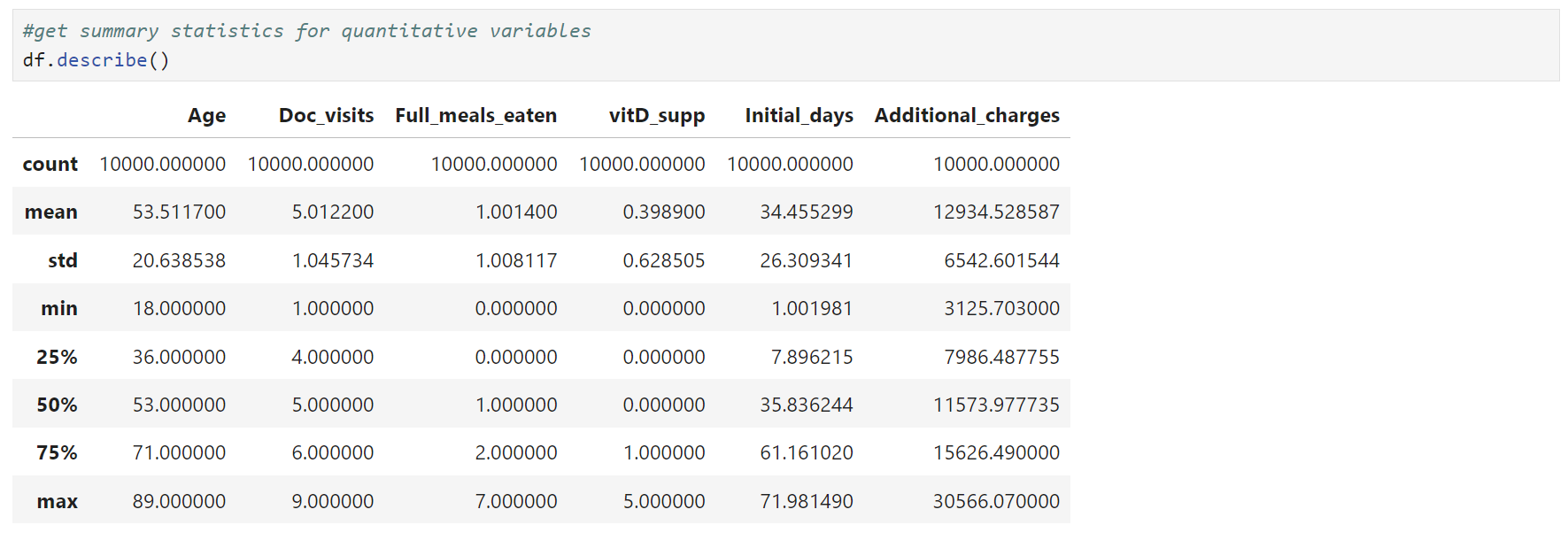


Since I decided to retain all the outliers, I will drop the Z\_Score columns from the dataset.



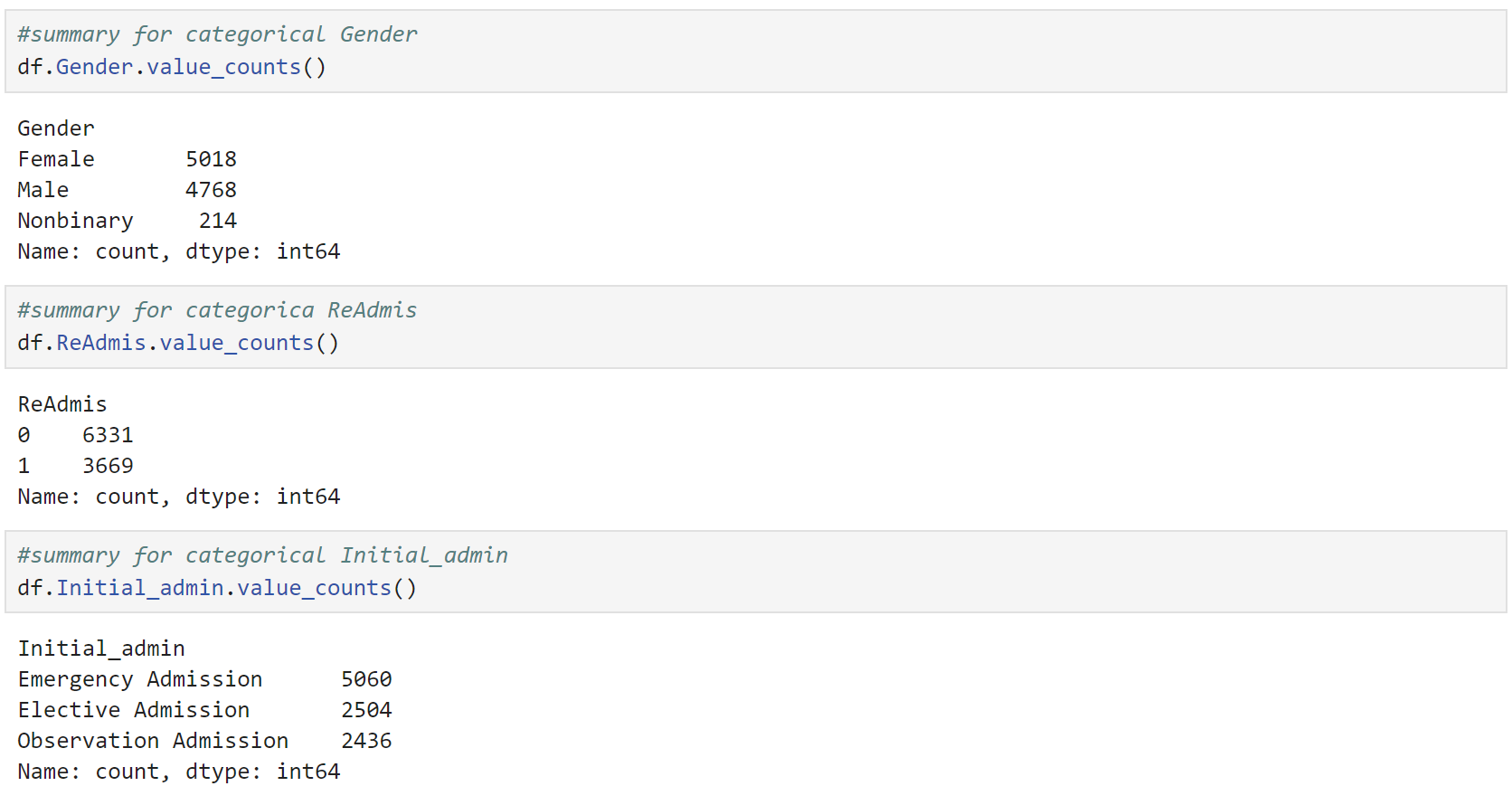
Please see the attached D208Task1\_DataCleaning.ipynb file with the annotated code for cleaning the data.

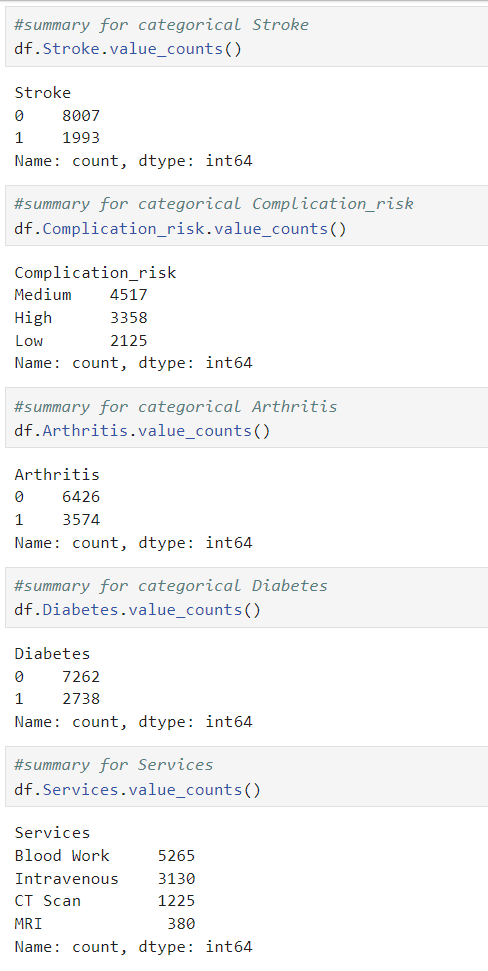
C2. Data Exploration



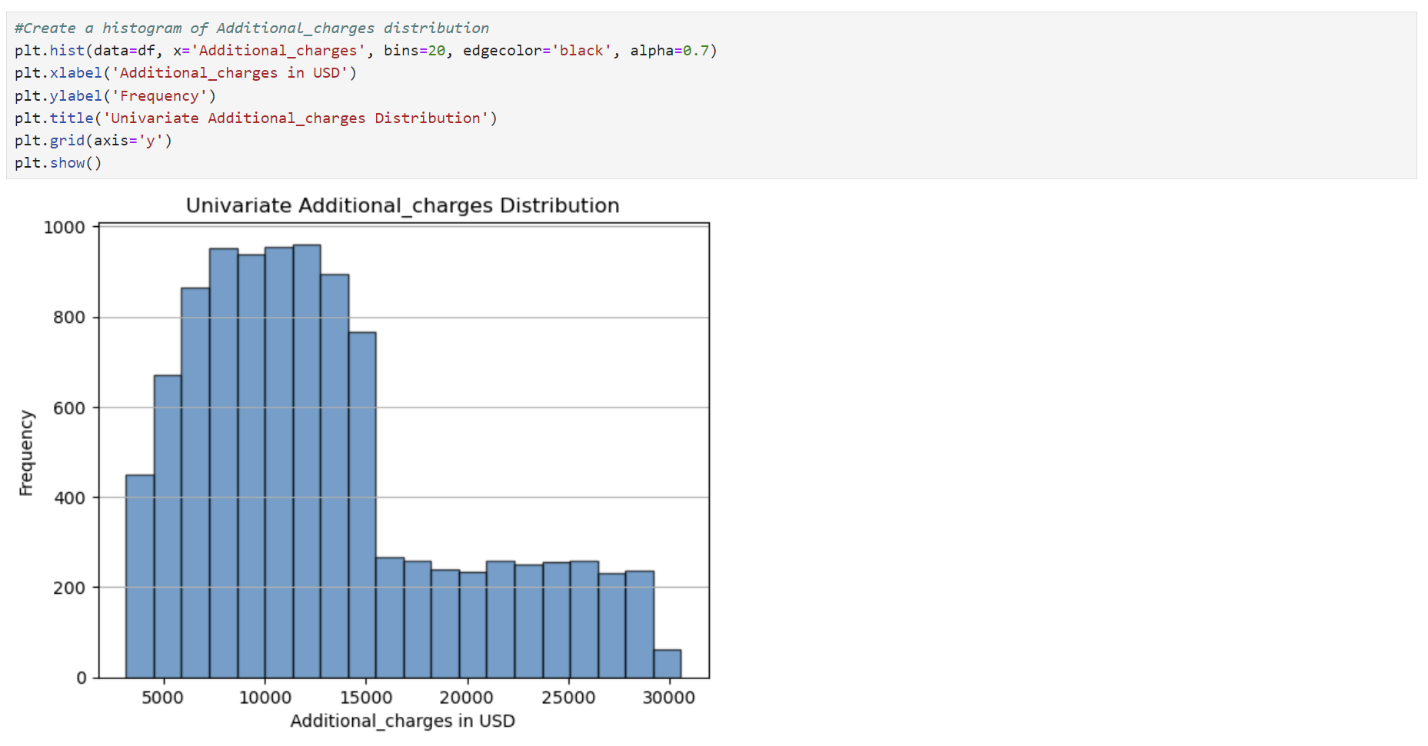
The data shows that the average age of patients is about 53 years, with most being between 36 and 71 years old. The youngest patient is 18, and the oldest is 89. On average, the primary physician visited the patient 5 times during the initial hospitalization, with visits ranging from 1 to 9. Patients ate about 1 full meal a day while hospitalized, though some ate none and others up to 7 meals. Vitamin D supplements were administered less than half the time on average, with most patients not receiving any but some receiving up to 5 doses. The number of days patients stayed in the hospital during the initial visit averaged around 34, ranging from 1 to 72 days. Additional charges for miscellaneous procedures, treatments, medicines, anesthesiology, etc., averaged around $12935 but could vary widely from about $3126 to over $30566. The middle 50% of charges fell between $7986 and $15626.

For categorical variables, the data shows that out of the patients, 5018 are female, 4768 are male, and 214 are nonbinary. Regarding readmissions, 6331 patients were not readmitted, while 3669 were. For initial admissions, 5060 were emergency, 2504 were elective, and 2436 were observation. Most patients did not have a stroke (8007), while 1993 did. The complication risk was medium for 4517 patients, high for 3358, and low for 2125. Arthritis was present in 3574 patients and absent in 6426. Diabetes was present in 2738 patients and absent in 7262. The services provided included blood work for 5265 patients, intravenous treatments for 3130, CT scans for 1225, and MRIs for 380.



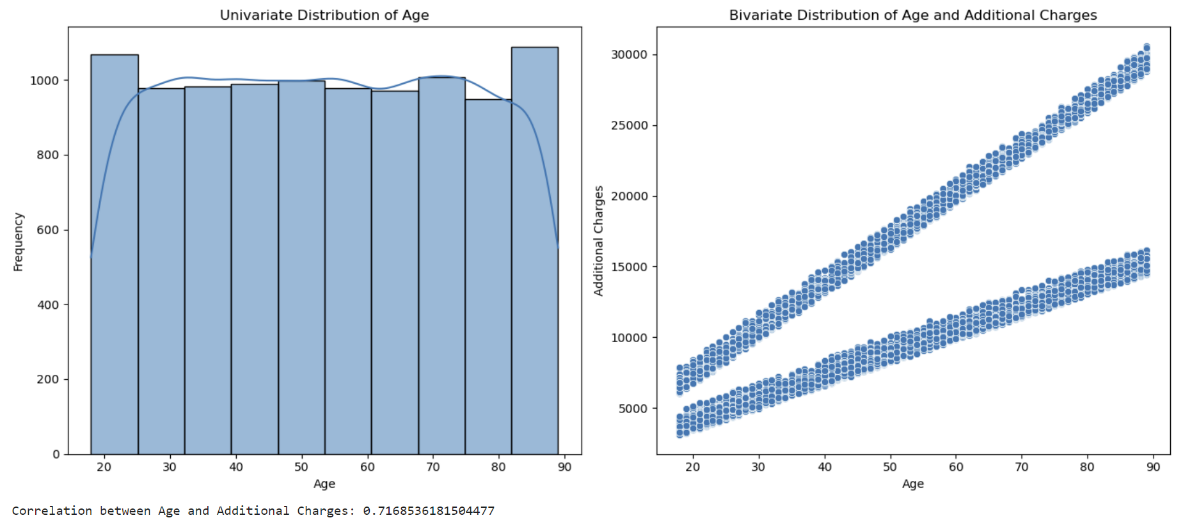


C3. Visualizations



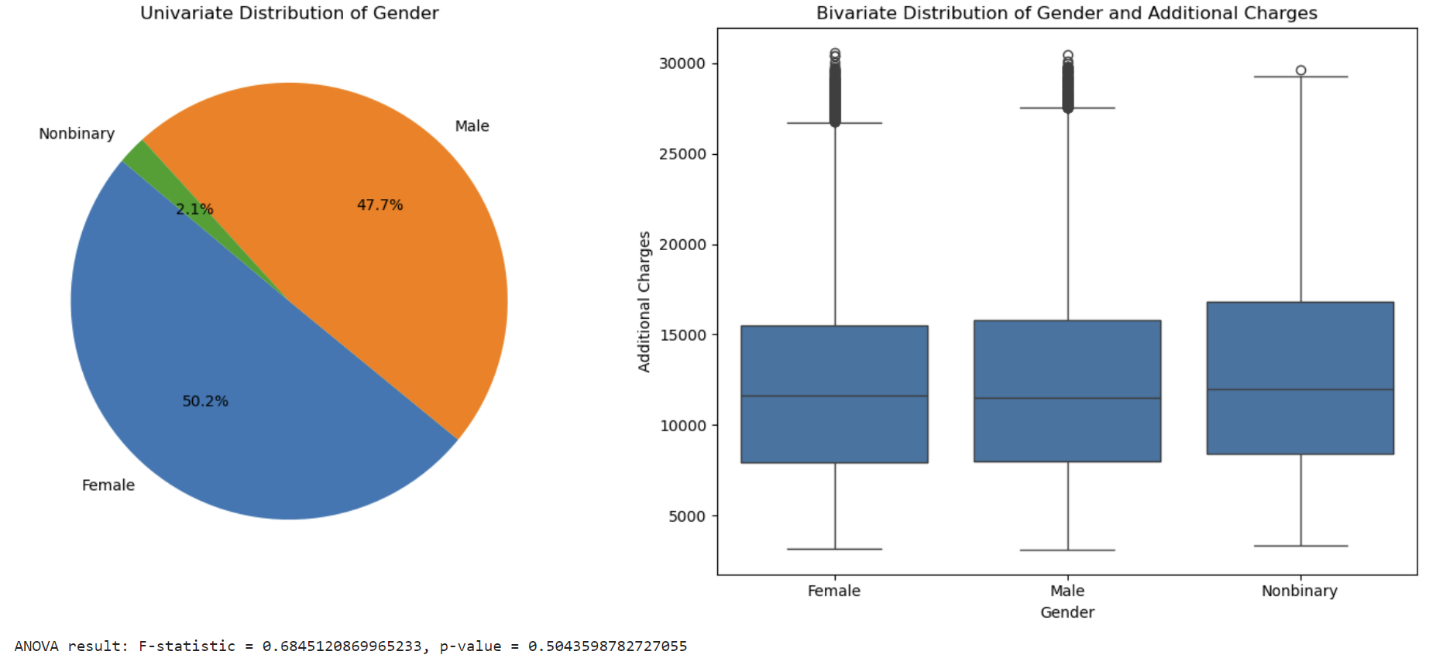
The distribution is right skewed.





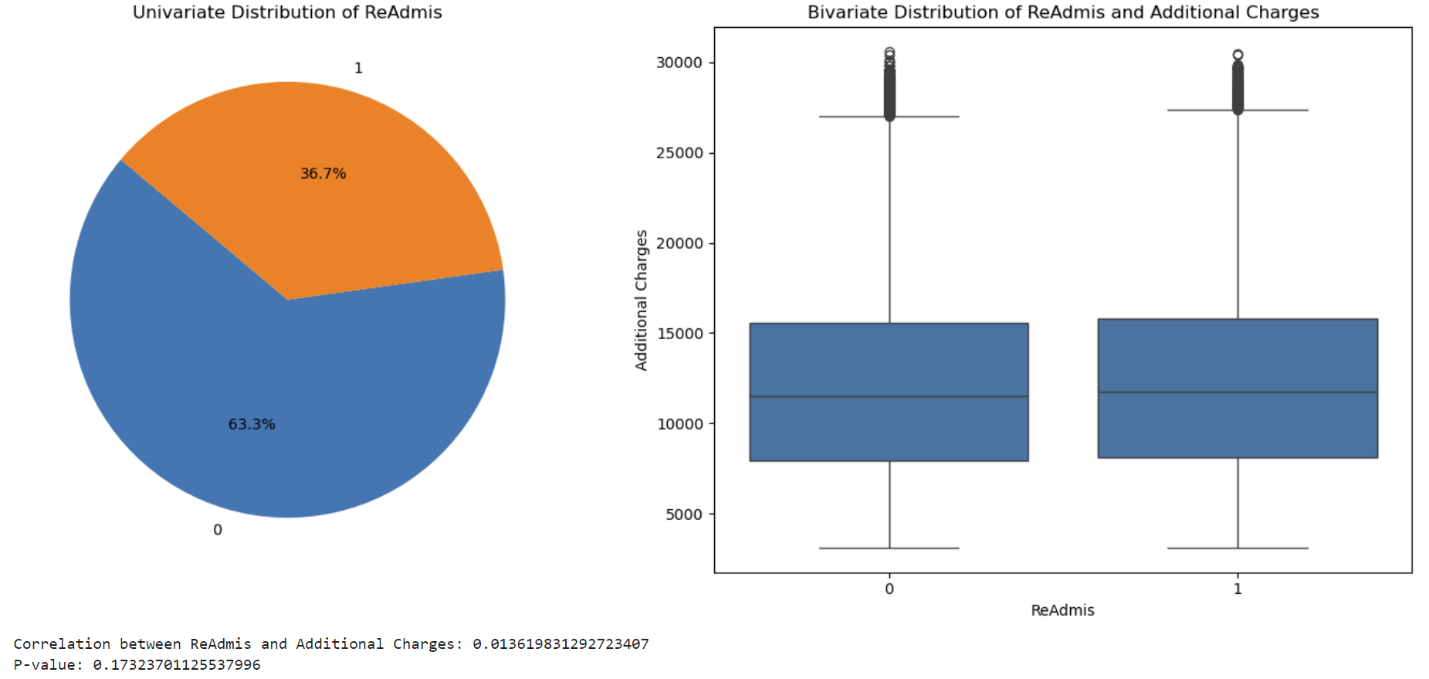
The Age distribution is uniform. The bivariate plot and the correlation of 0.717 demonstrate a relationship between variables Age and Additional\_charges.





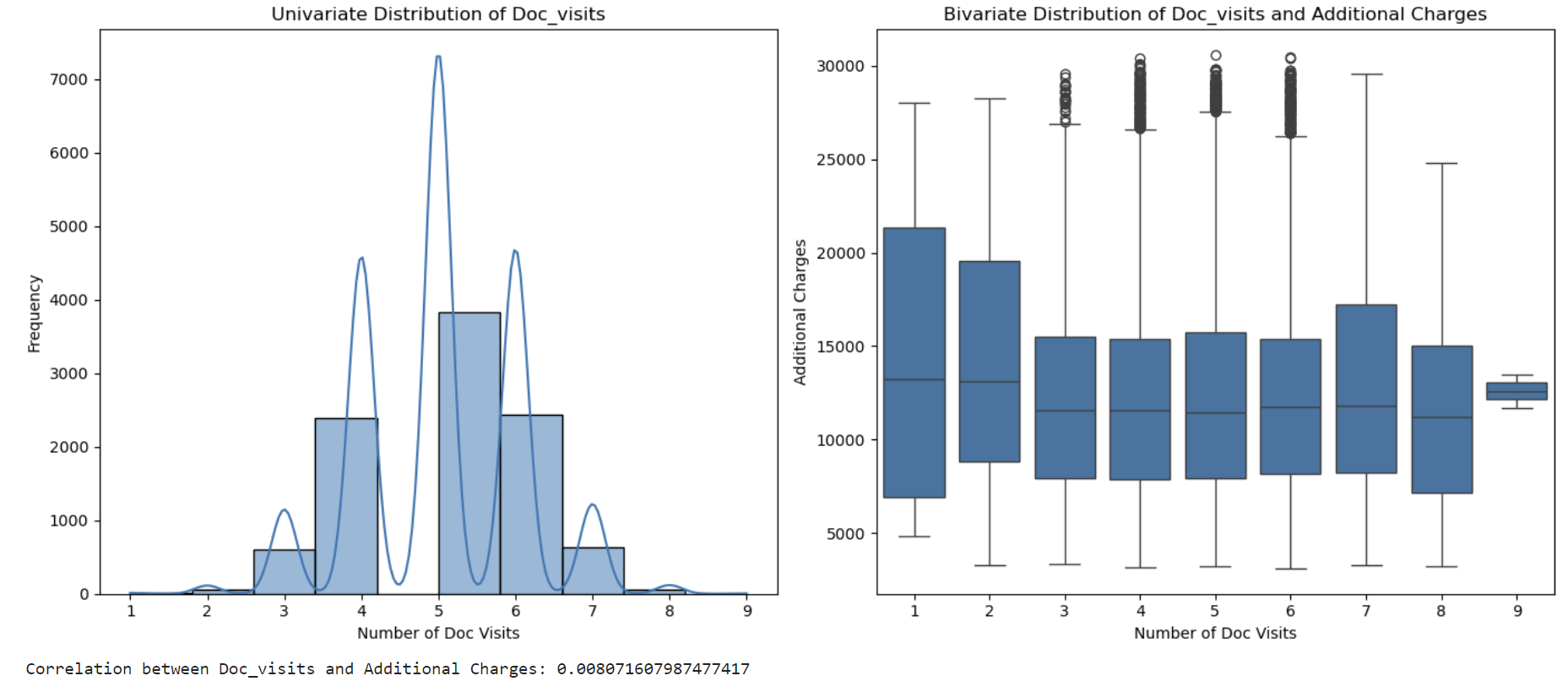
The univariate distribution of Gender shows that 50.2% of patients are Female, 47.7% are Male, and 2.1% are Nonbinary. Bivariate distribution and ANOVA test results demonstrate no relationship between Gender and Additional\_charges.





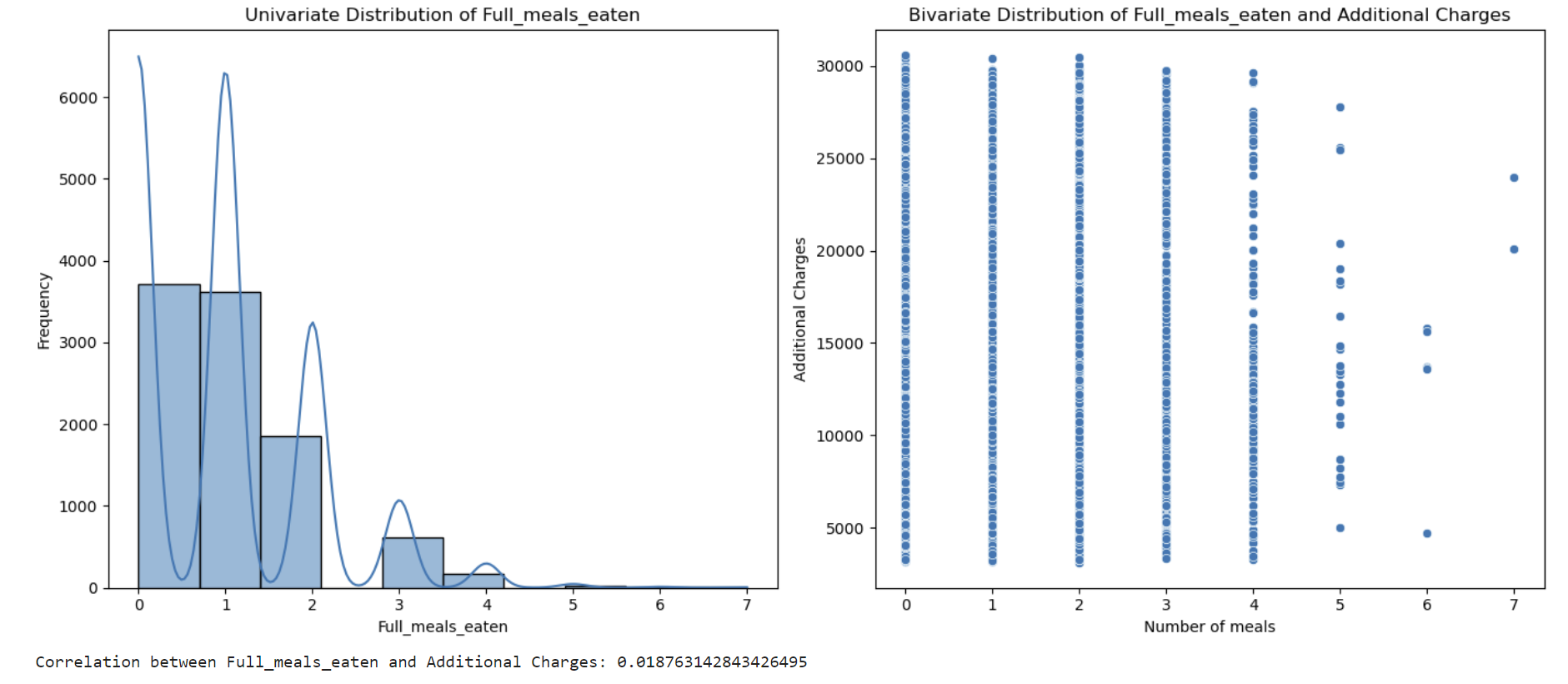
The ReAdmis univariate distribution demonstrates that 63.3% of patients were readmitted, while 36.7% weren’t. The bivariate distribution and results of biserial correlation with a p-value of 0.173 suggest no relationship between ReAdmis and Additional\_charges.





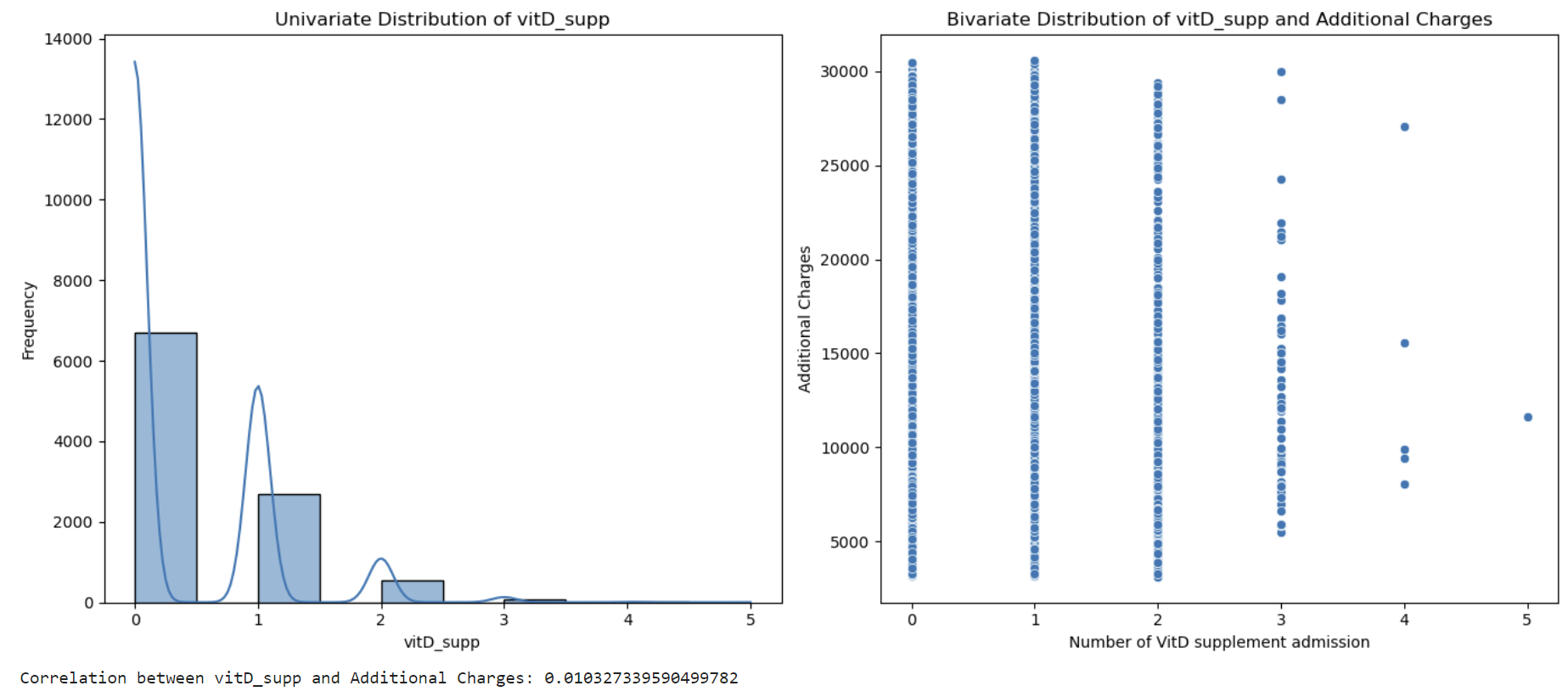
The univariate distribution is normal binomial. The bivariate distribution and correlation of 0.008 suggest that there is no relationship between Doc\_visits and Additional\_charges.





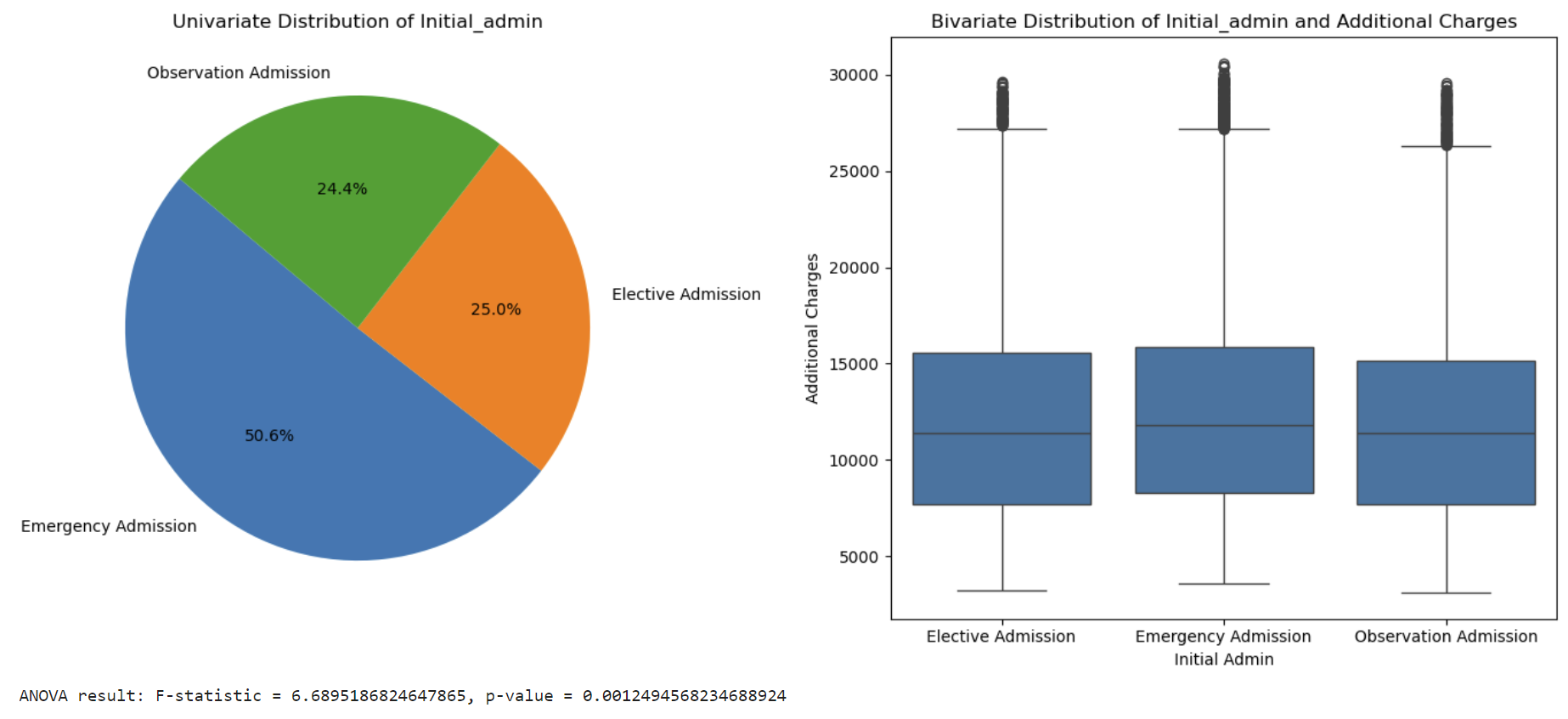
The Full\_meals\_eaten univariate distribution is right skewed. The bivariate distribution and correlation of 0.019 demonstrate no relationship between Full\_meals\_eaten and Additional\_charges.





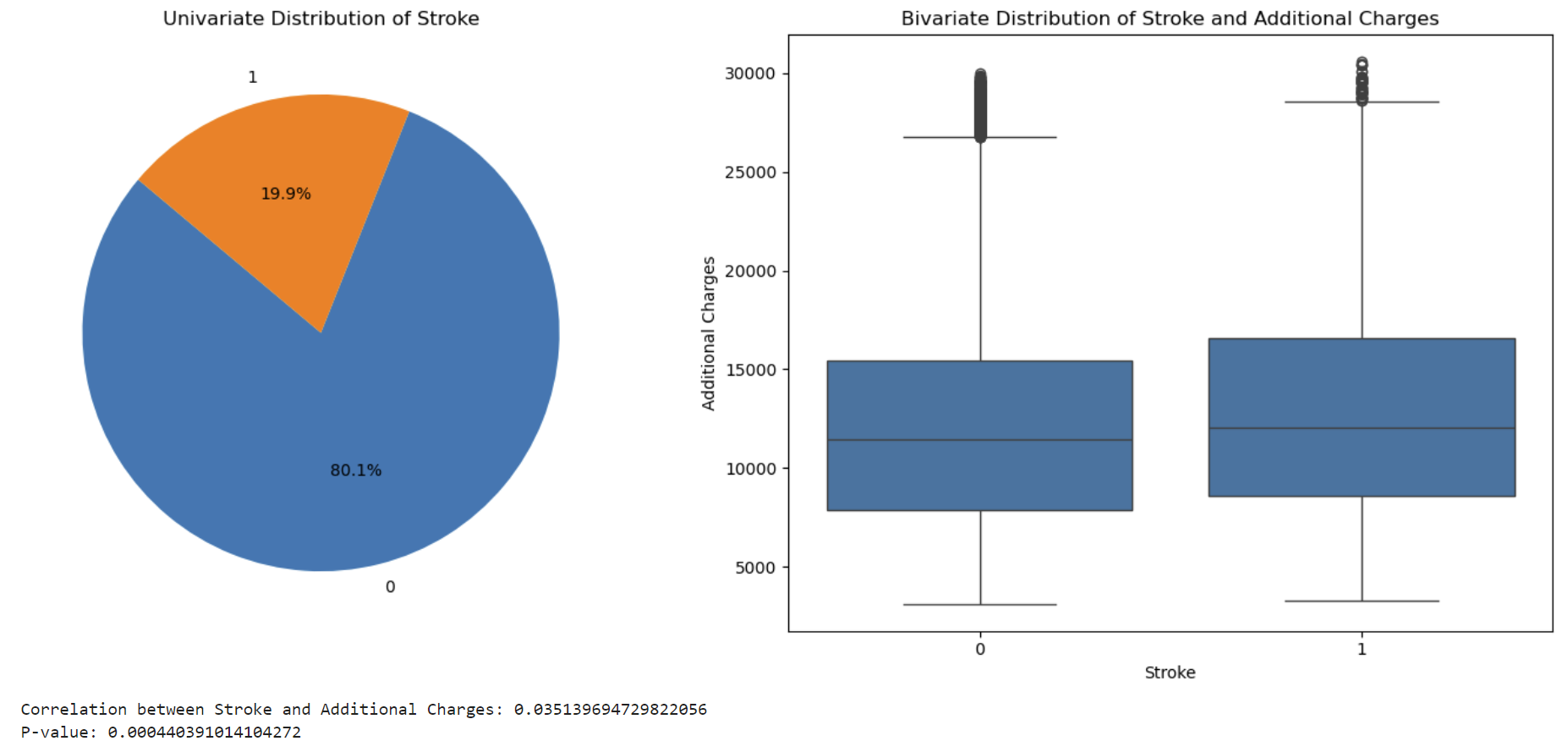
The univariate vitD\_supp distribution is right skewed. The bivariate distribution and correlation of 0.01 suggest that there is no relationship between vitD\_supp and Additional\_charges.





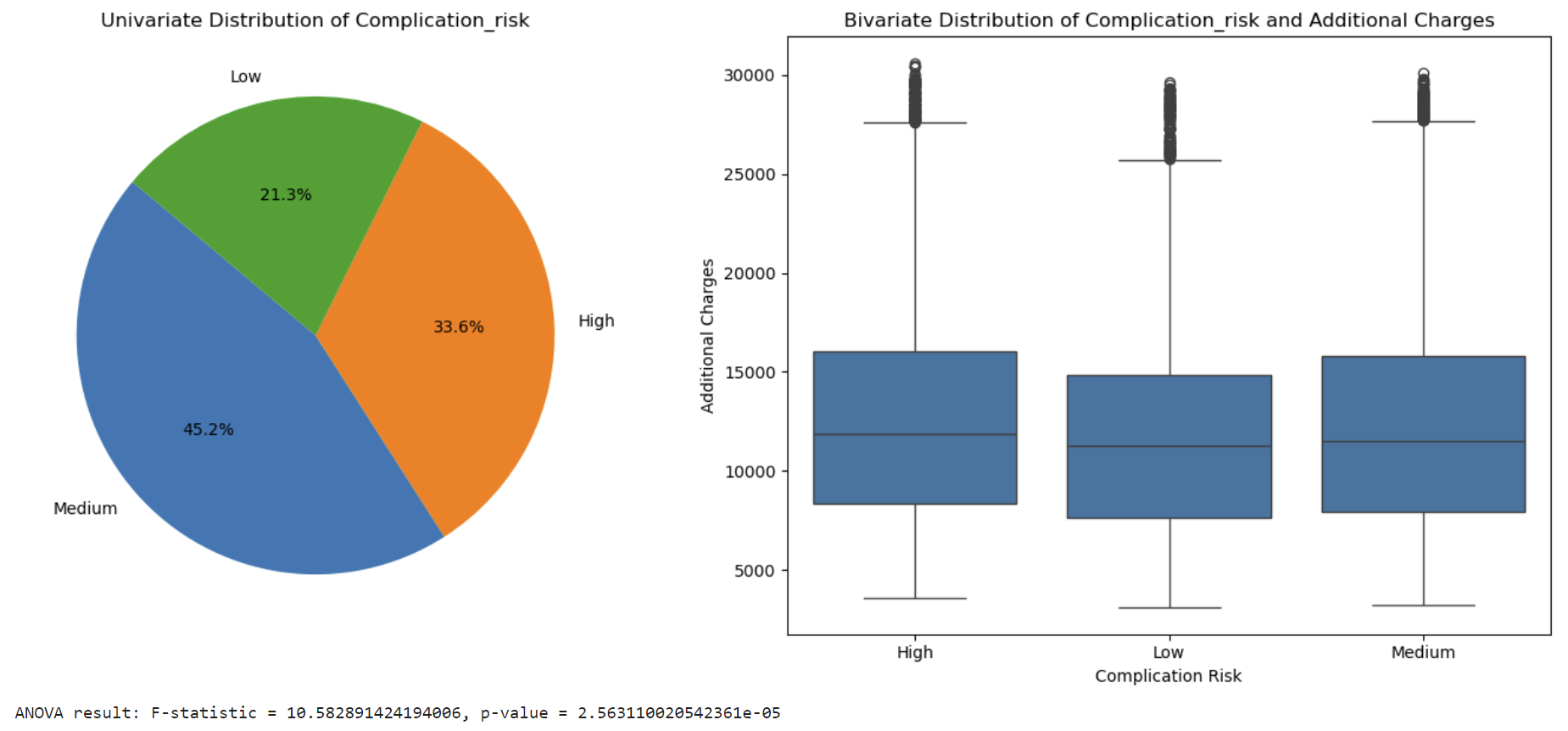
The Initial\_admis univariate distribution demonstrates that Emergency Admission is 50.6%, Elective Admission is 25.0%, and Observation Admission is 24.4%. The bivariate distribution boxplot and p-value 0.0012 suggest that there might be a relationship between Initial\_admin and Additional\_charges.





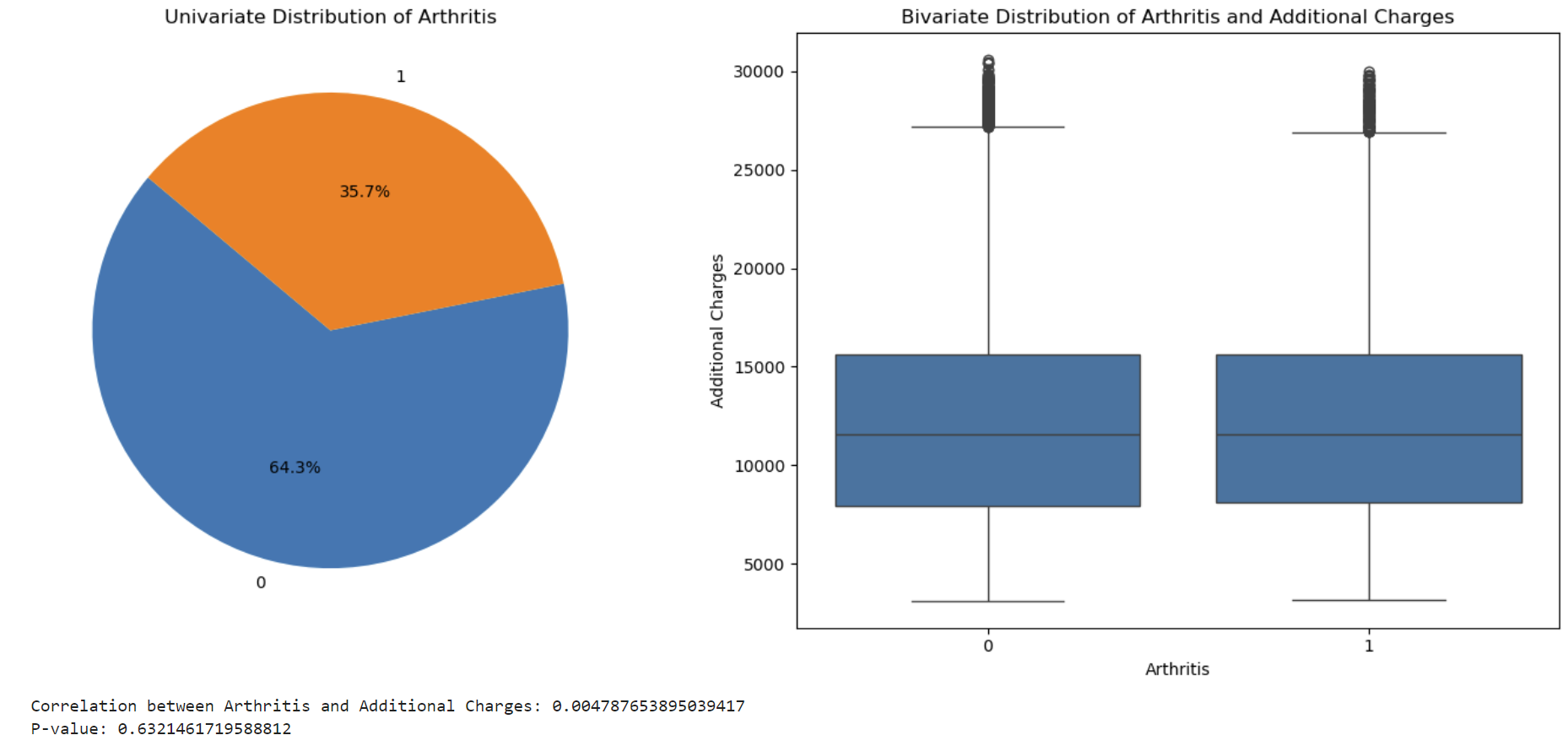
The Stroke univariate distribution shows, that 80.1% of patients didn’t have a stroke, while 19.9% did. The bivariate distribution and p-value 0.0004 demonstrate that there might be a relationship between Stroke and Additional\_charges.





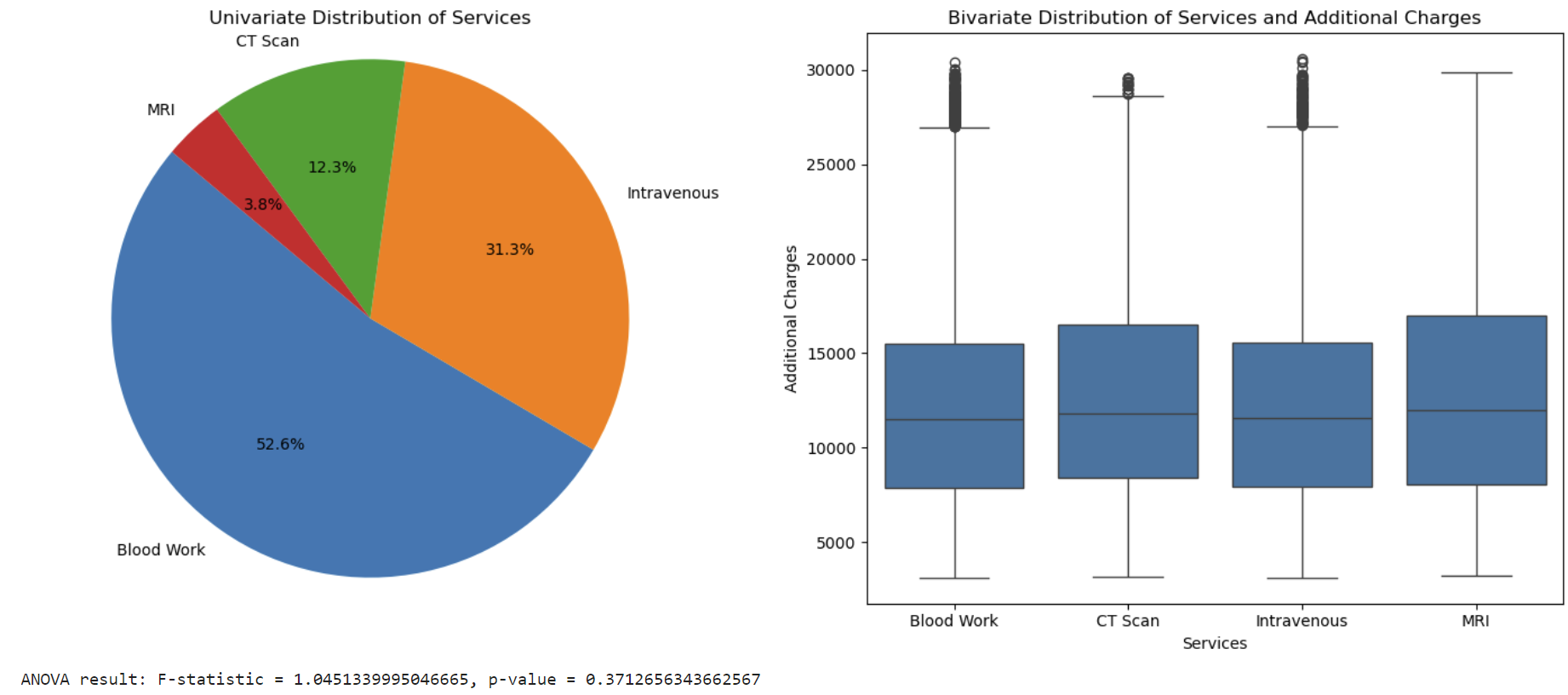
The univariate distribution of Complication\_risk shows that 33.6% is high, 45.2% medium and 21.3% are low risk. Bivariate distribution and low p-value of approximately 0.0000256 demonstrate a relationship between Complication\_rist and Additional\_charges.





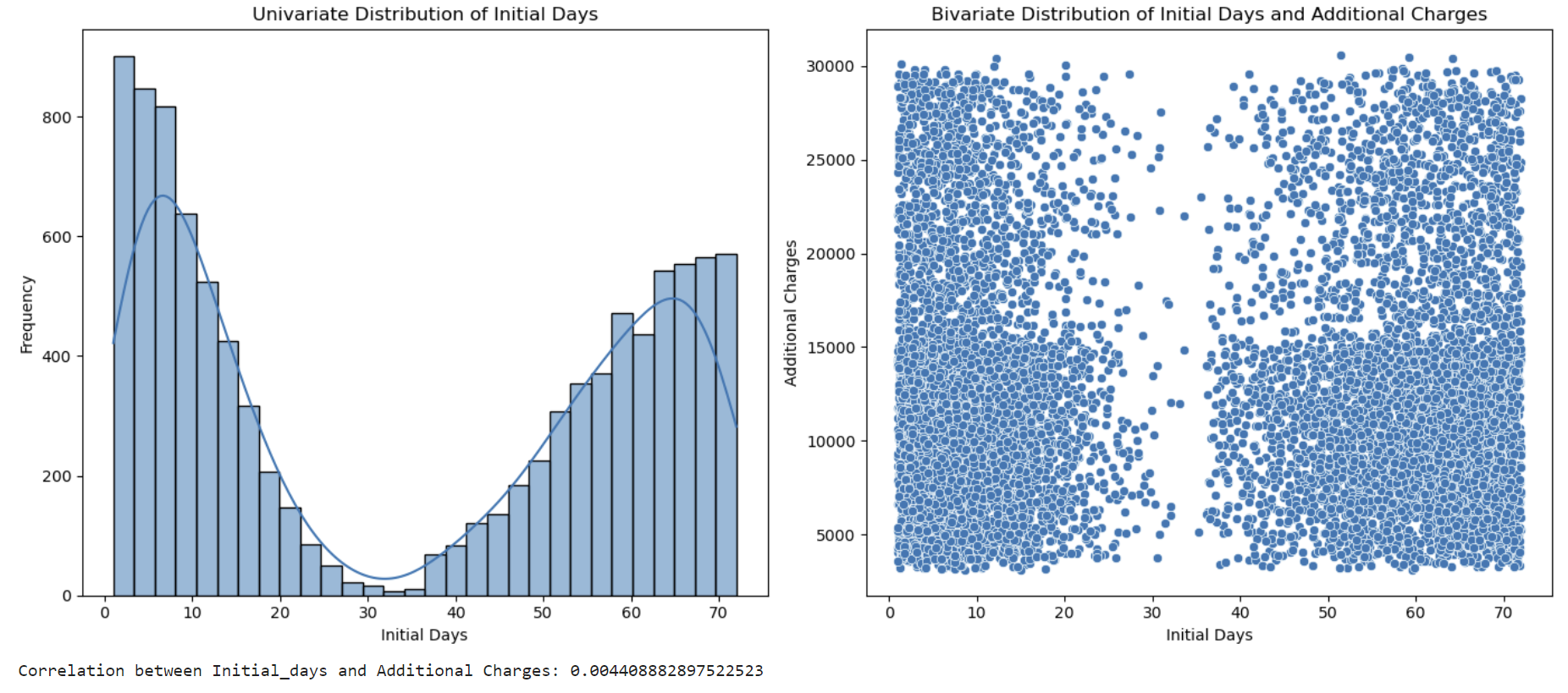
The Arthritis univariate distribution demonstrates that 64.3% don’t have Arthritis, while 35.7% do. The bivariate distribution and results of the biserial correlation test with a p-value of 0.632 suggest no relationship between Arthritis and Adiitional\_charges.





The Services univariate distribution shows that blood Work is the most common service 52.6%, followed by Intravenous 31.3%, while CT Scan 12.3% and MRI 3.8%. The bivariate distribution and results of the ANOVA test with a p-value of 0.371 suggest no relationship between Services and Addition\_charges.





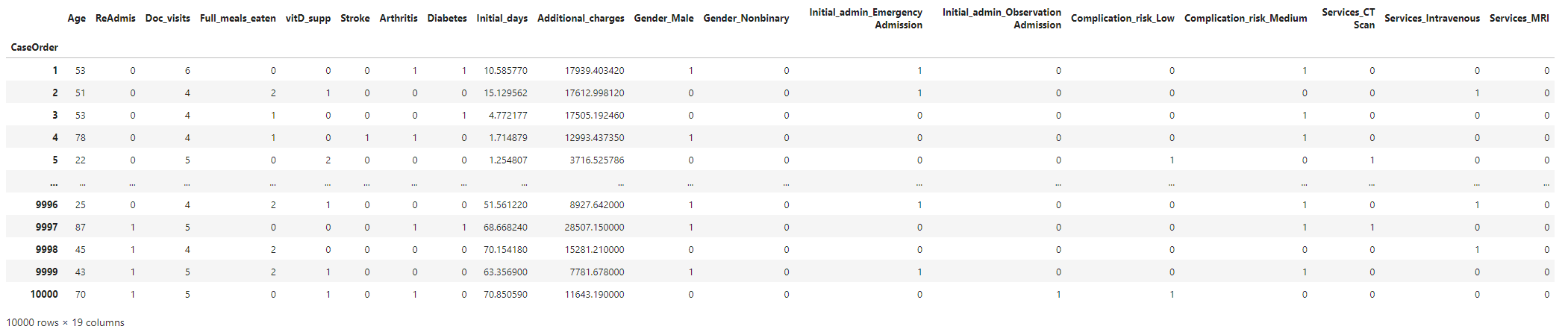
The univariate distribution of Initial\_days is bimodal. The bivariate distribution and correlation 0.0044 suggest no relationship between Initial\_days and Additional\_charges.

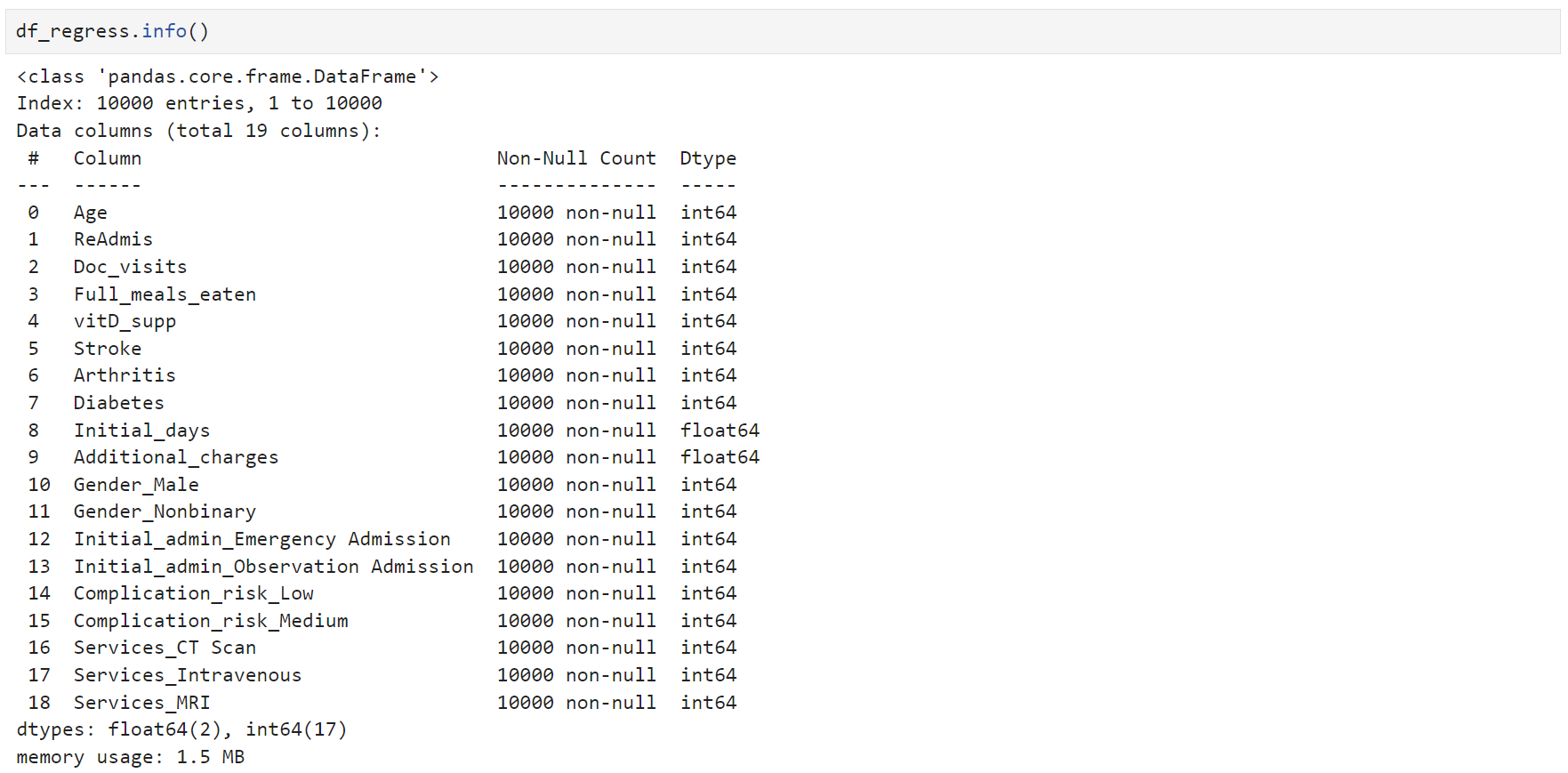
***C4. Data Transformation***

Most statistical methods/machine learning algorithms mining work exclusively with numeric data (*Redirecting*, n.d., p. 28).  One-hot encoding is essential for handling categorical variables when creating a multiple linear regression model. It transforms categorical variables into a set of binary (0 or 1) indicator variables, where each level of the original category becomes a separate feature. This process ensures that the model can learn from the categorical information effectively. For instance, if we have a “Gender” variable with levels “Male” and “Female,” one-hot encoding creates two new features: one indicating male (1 for males, 0 for females) and another indicating female (1 for females, 0 for males).

One-hot encoding was created using the pd.get\_dummies() function (GeeksforGeeks, 2024). The drop\_first=True argument drops the first category for each feature to avoid multicollinearity. Using df\_regress.replace({True: 1, False: 0}), I converted boolean values to numeric (1 and 0) and applied pd.to\_numeric to convert all columns to numeric.







***C5. Prepared Dataset***

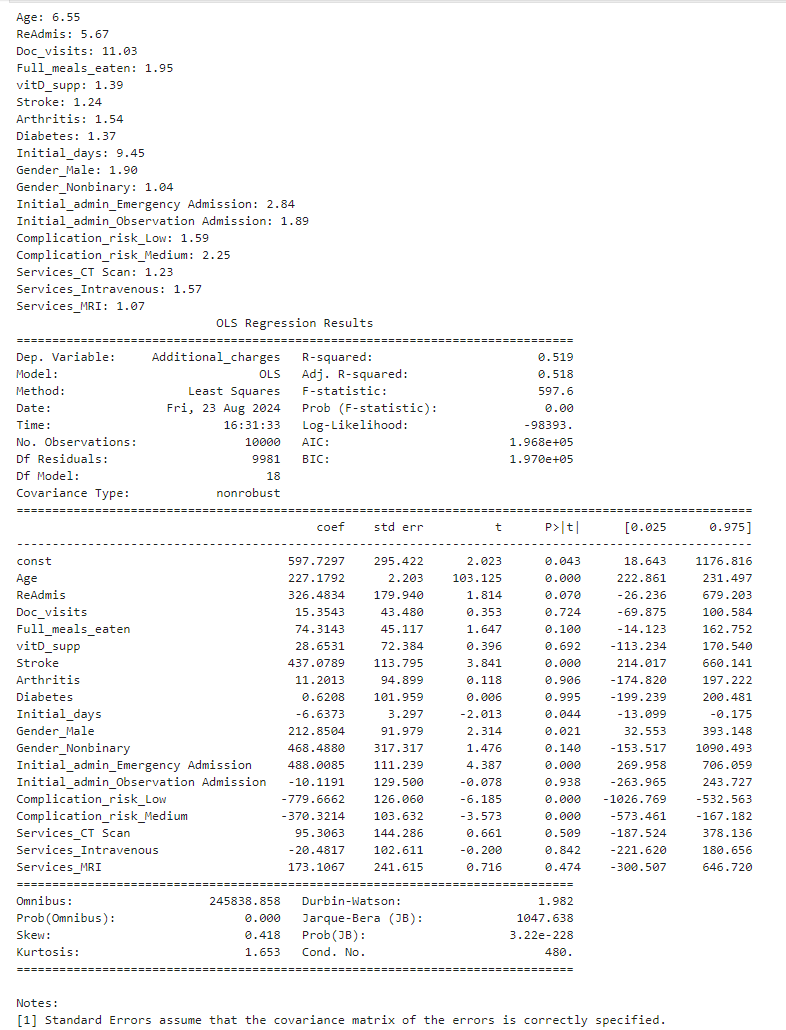
Prepared df\_regress Dataset was saved to a new csv file. Please see the attached “D208\_task1.csv” file.

***Part IV: Model Comparison and Analysis***

***D1. Initial Model***

Multiple linear regression is a statistical method we can use to understand the relationship between multiple predictor variables and a response variable. I set y to the dependent variable Additional\_charges. I created an X dataset with all independent variables (Age, ReAdmis, Doc\_visits, Full\_meals\_eaten, vitD\_support, Stroke, Arthritis, Diabetes, Initial\_days, Gender\_Male, Gender\_Nonbinary, Initial\_admin\_Emergency Admission, Initial\_admin\_Observation Admition, Complication\_risk\_Low, Complicaion\_risk\_Medium, Services\_CT Scan, Services\_Intravenous, Services\_MRI). The variance\_inflation\_factor function is used to compute the VIF for each feature in X. VIF measures the degree of multicollinearity (correlation) between independent variables. X = sm.add\_constant(X) adds a constant term (intercept) to the independent variable matrix X to ensure that the regression model includes an intercept term—the model = sm.OLS(y, X).fit() is used to create an OLS regression model. The fit() method fits the model and the data.



****

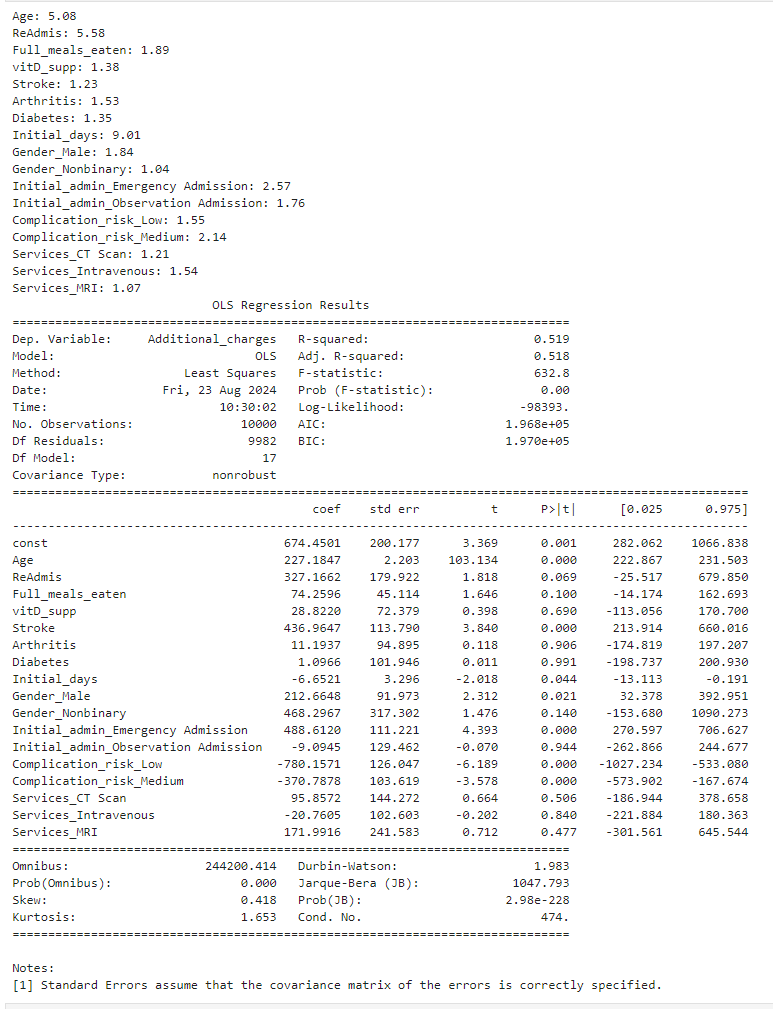
***D2. Model Reduction Method and Justification***

For the model reduction, I will use the "kitchen sink" approach when all variables are thrown into a model, regardless of their relevance to the outcome variable or statistical significance. Next, using the Backward Elimination method, I will begin iteratively removing the least significant variable until no further improvement is achieved. Variance Inflation Factor (VIF) will be calculated for each feature. High VIF indicates multicollinearity. If a variable has a VIF above 10, I will remove it from the model (Michael Parker, 2019). This simplifies the model and reduces redundancy. Next, I look at the p-values for each remaining feature. Low p-values below 0.05 mean the feature is statistically significant. If a feature has a high p-value, I will remove it to keep only the most relevant features.

Additionally, MinMax scaling will be performed to ensure that all features have similar scales between 0 and 1 (Mark Keith, 2021). Without scaling, some features might dominate others just because of their scale (e.g., age in years vs. income in thousands). Scaling helps the model treat all features equally, leading to better predictions.

According to the initial model results, Doc\_visits VIF is 11.03, so I will remove it.

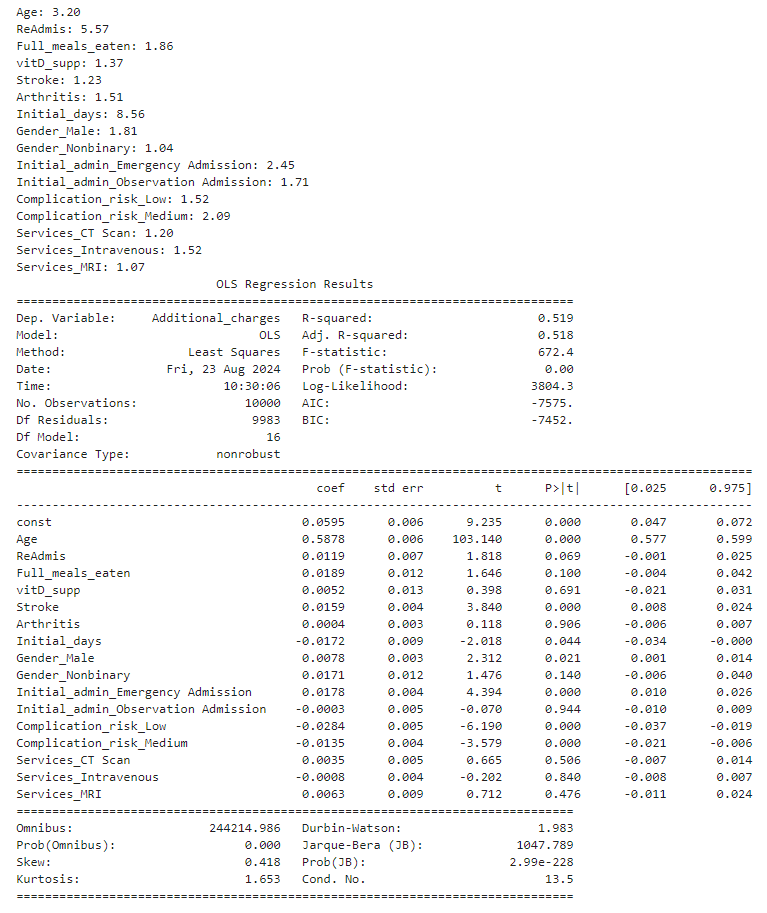




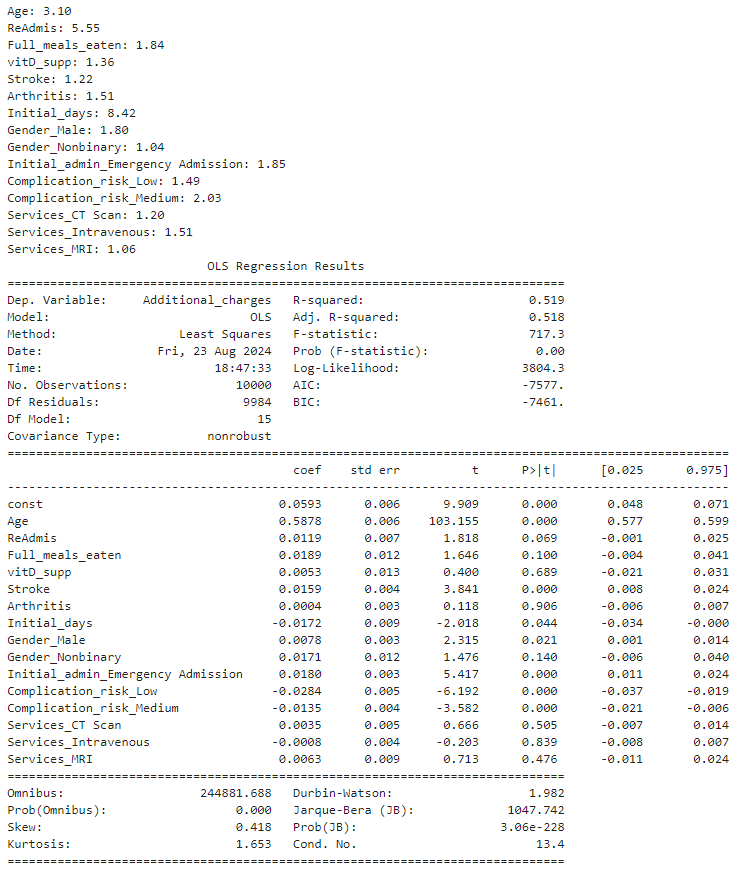
Dataset normalization with MinMaxScaler().



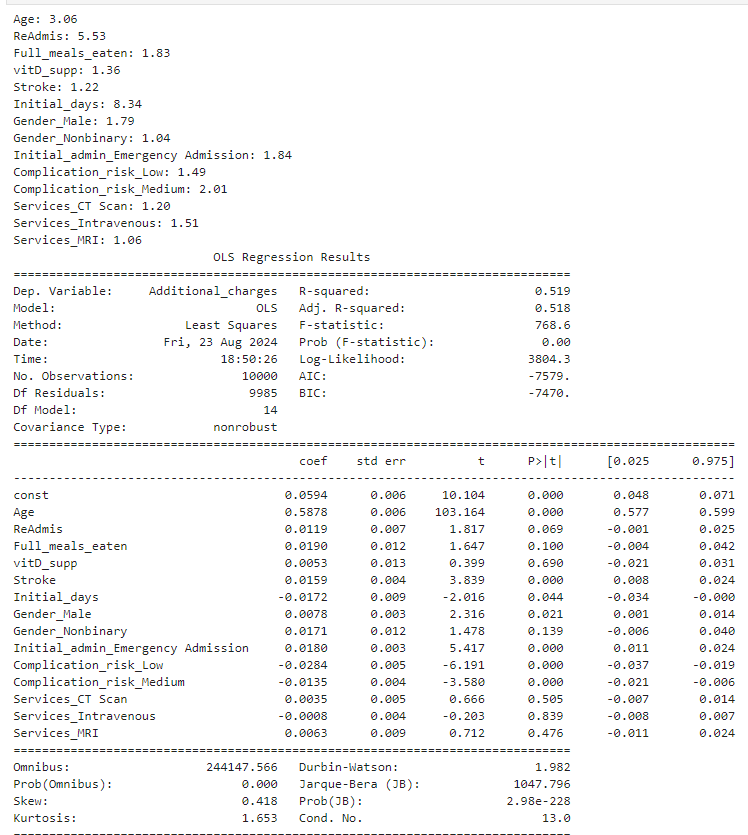
Since in the reduced model, there are no variables with VIF higher than 10, I will start to remove variables based on p-value. I will create a reduced model with removed Diabetes due to p= 0.991



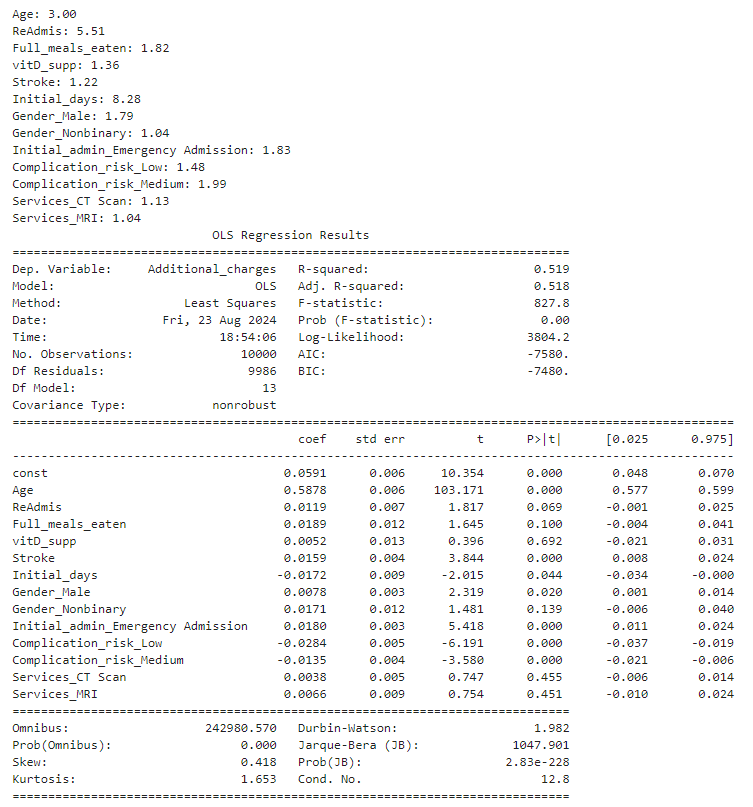
Create a reduced model with removed Initial\_admin\_observation Admission due to p= 0.944



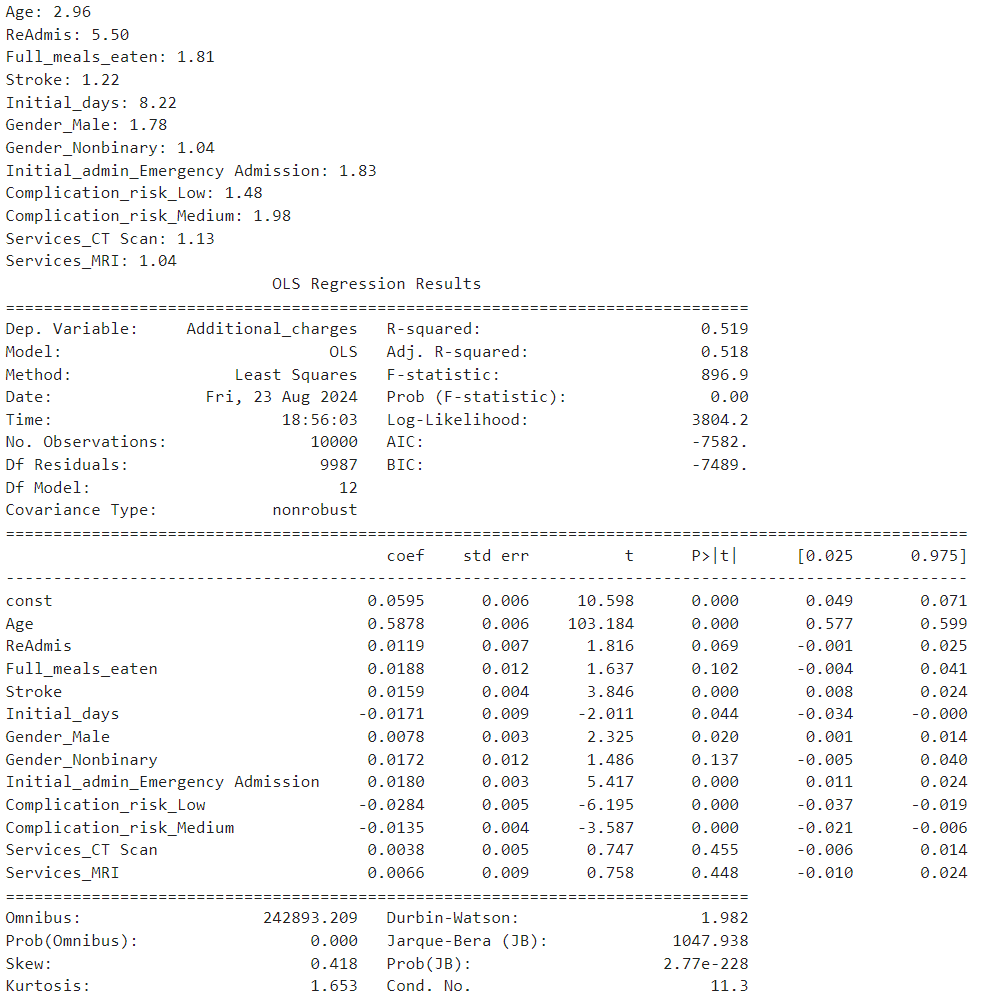
The next step is to remove Arthritis due to a p-value of 0.906



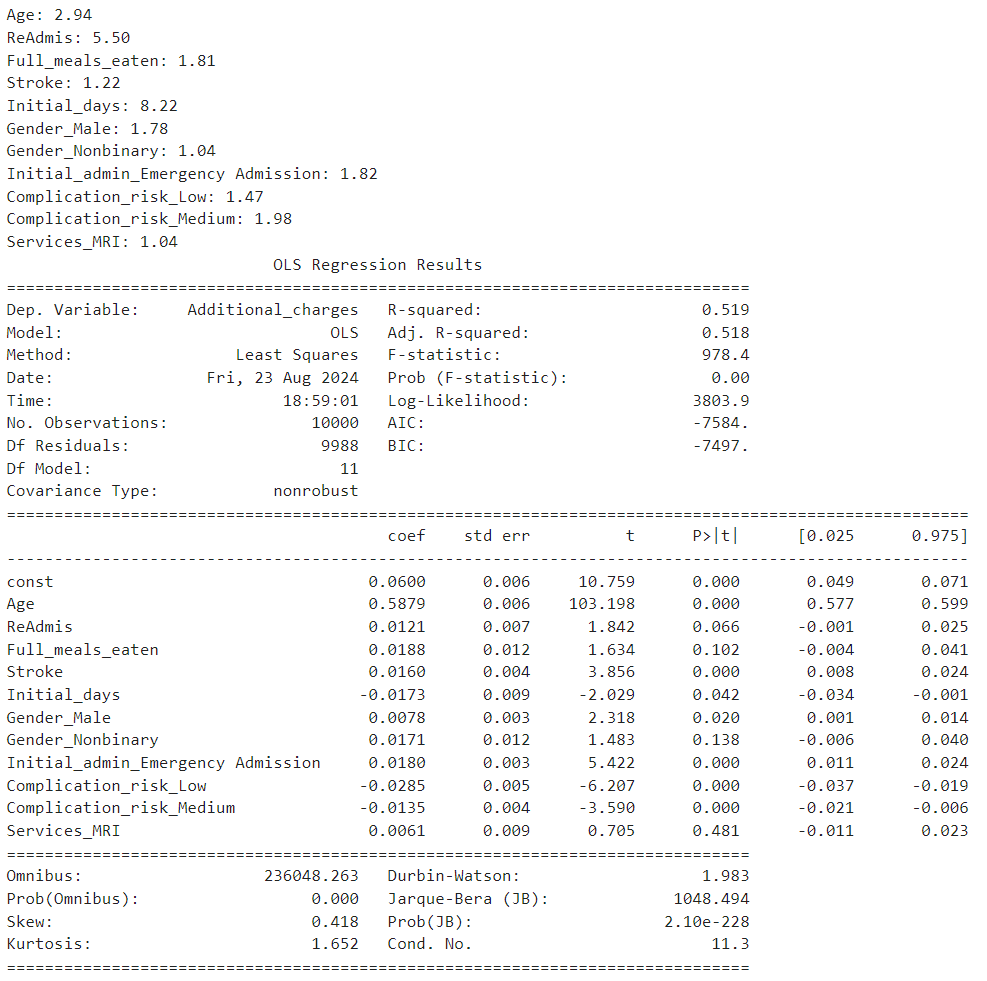
In this reduced model, the Services\_Intravenous has the highest p-value of 0.839. I will remove it from the model.



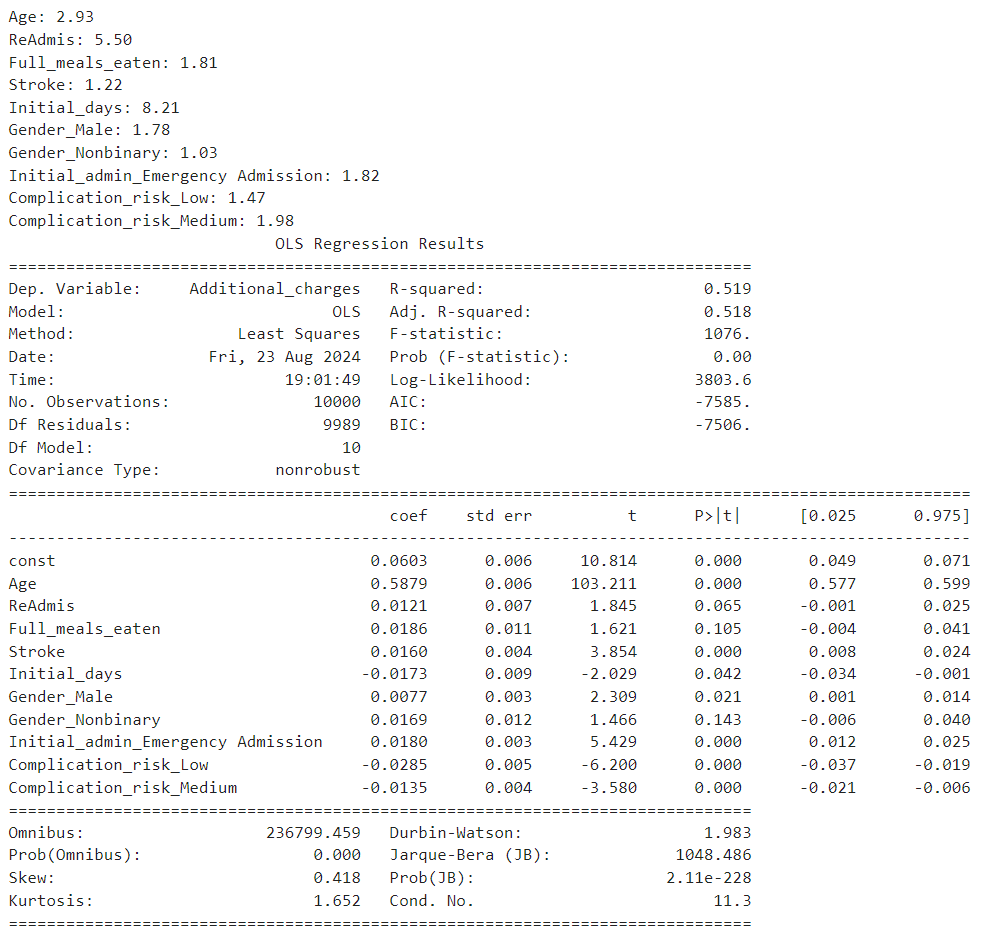
The vitD\_supp will be removed from the model due to the p-value of 0.692.



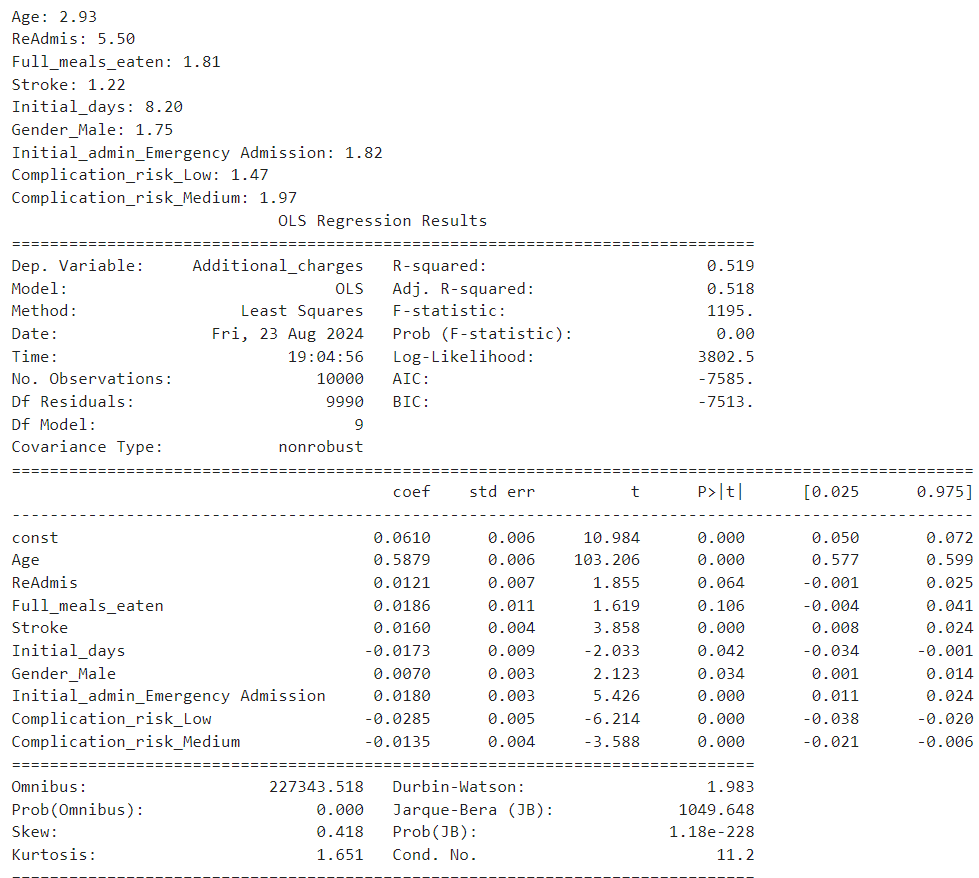
The Services\_CT Scan will be removed due to the p-value of 0.455



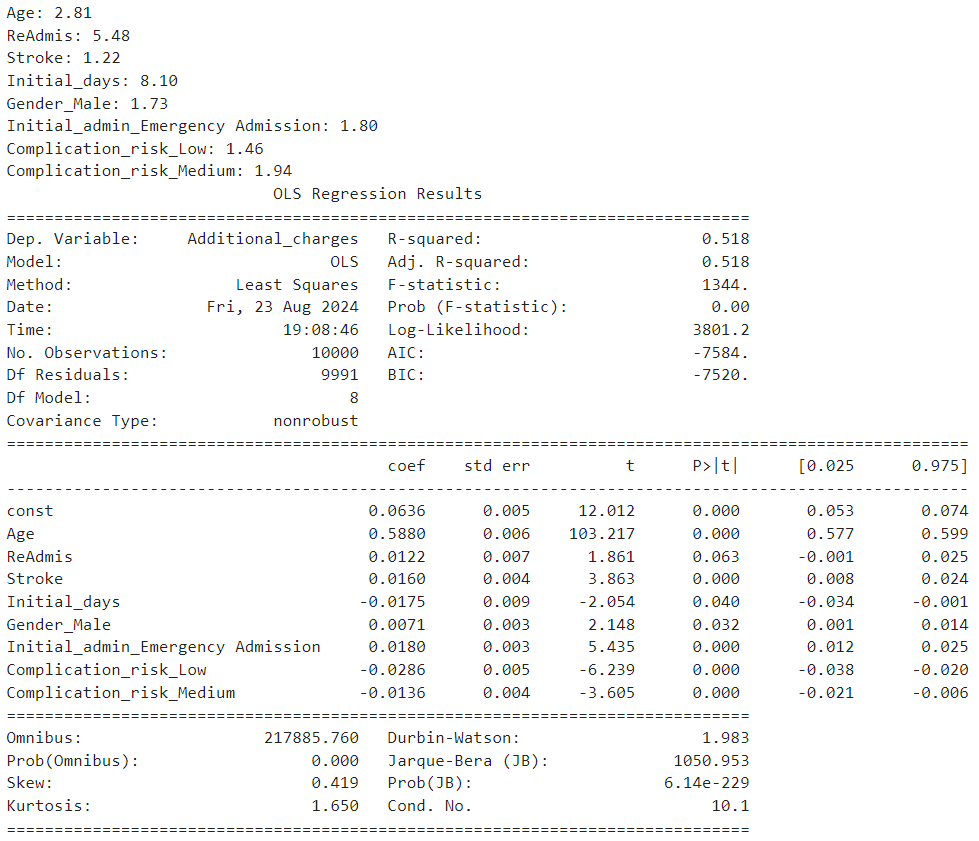
The Services\_MRI p-value is 0.481, so it will be removed.



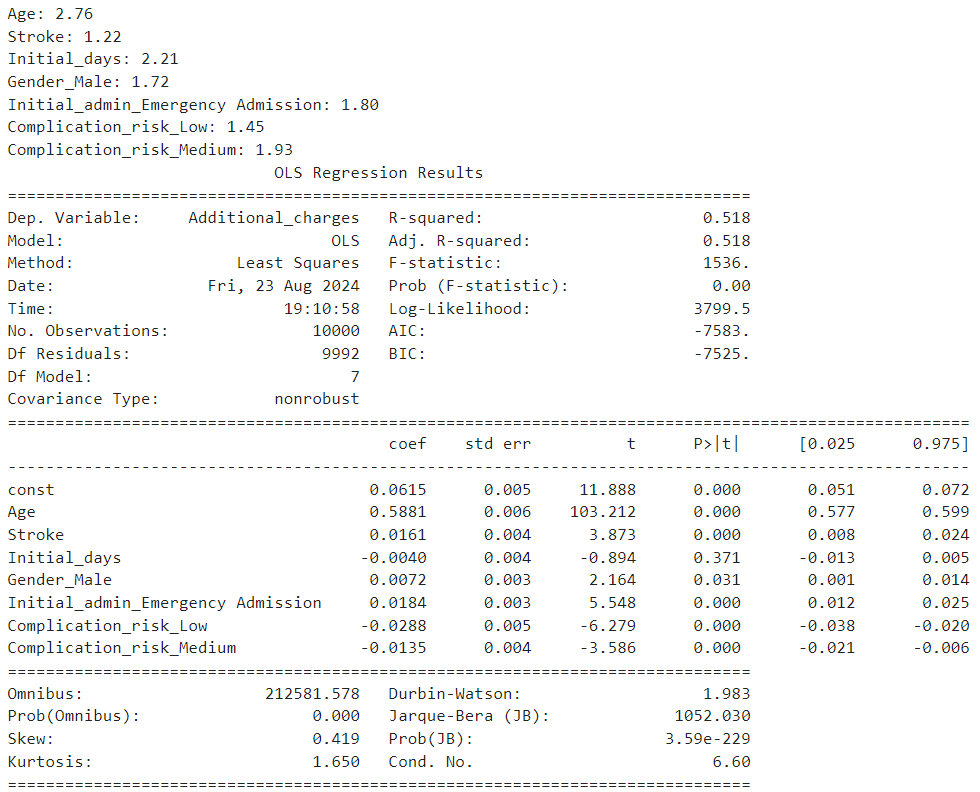
The Gender\_Nonbinary will be removed due to the p-value of 0.143



The Full\_meals\_eaten p-value is 0.106, so it will be removed.



The ReAdmis will be removed due to the p-value of 0.063.

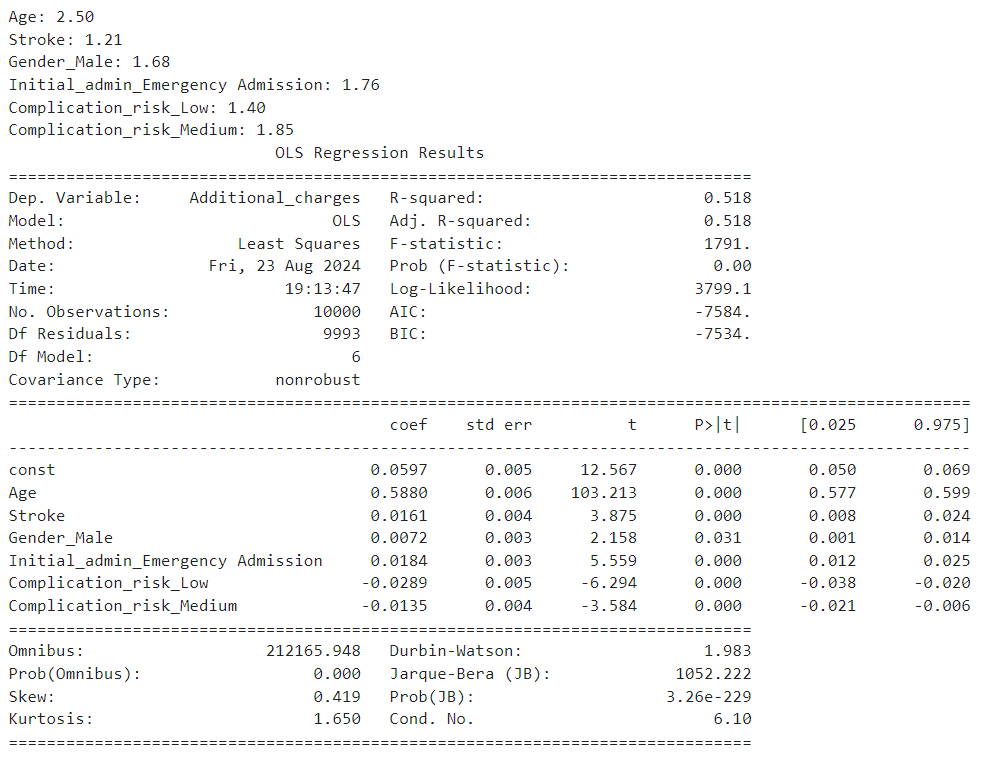


The Initial\_days variable has a p-value of 0.371. It will be removed.

***D3. Reduced Model***

After the last variable, Initial\_days was removed from the model, there were no variables with p-values higher than 0.05.

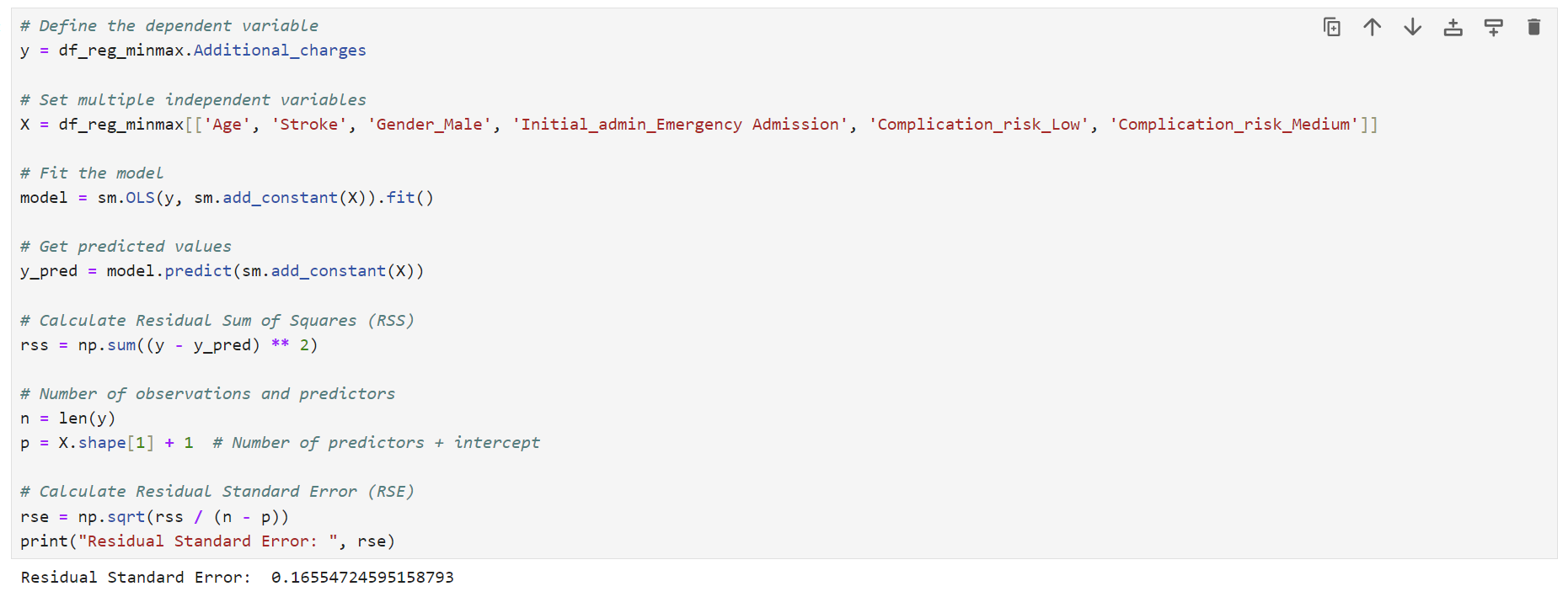
Below is a reduced model.

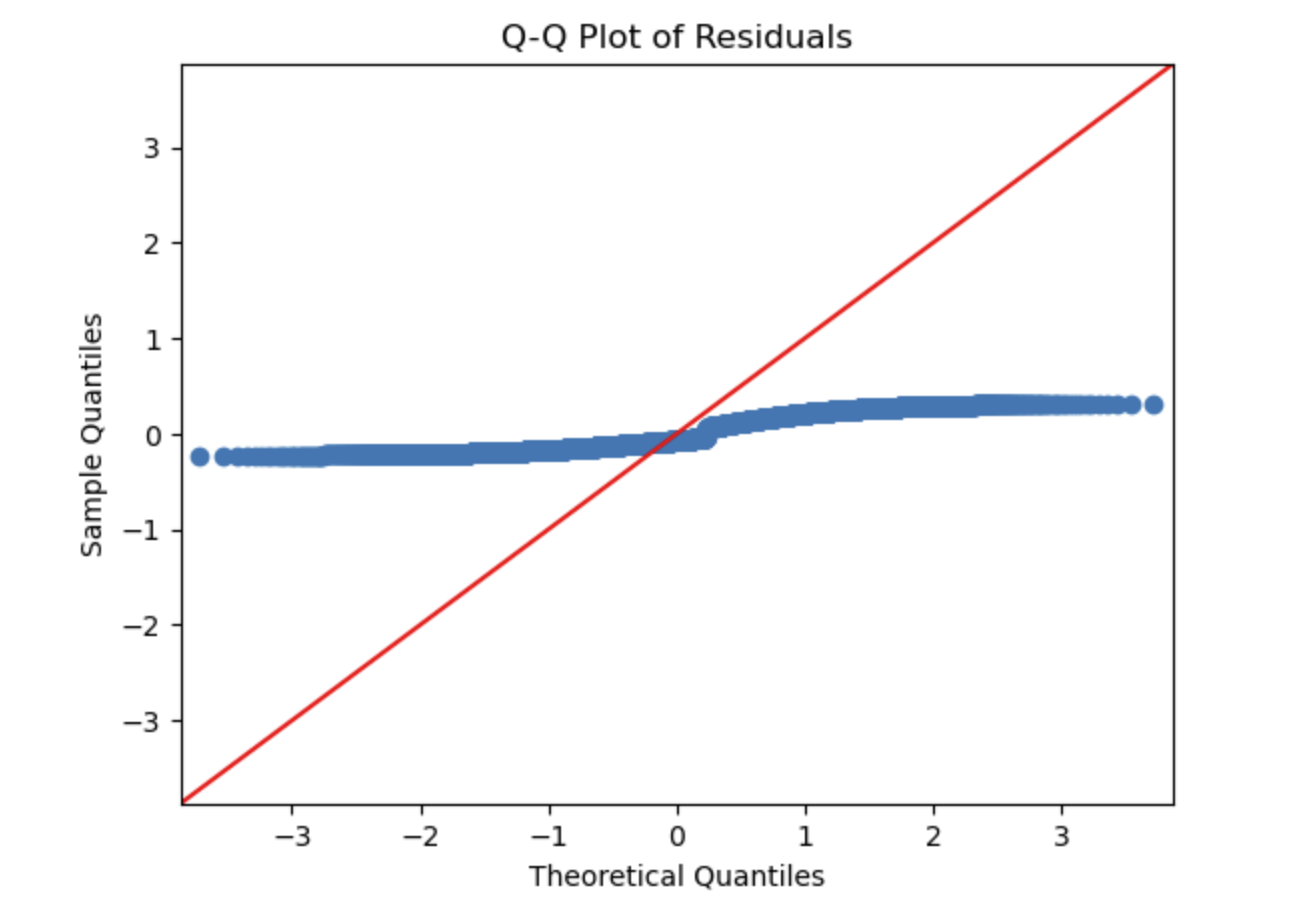


***E1. Model Comparison***

* The initial model has 18 independent variables, while the reduced has 6 variables.
* The initial model had one variable with a VIF higher than 10, which indicates multicollinearity.
* The initial model has variables with p-values higher than 0.05 that suggest that these variables have no significant effect. In the reduced model, there are no variables with a p-value higher than 0.05.
* The initial and reduced models have an R-squared value of 0.518, indicating that approximately 51.8% of the variability in the Additional\_charges can be explained by the independent variables.
* Both models have the same Adj. R-squared (0.518), indicating equivalent explanatory power. However, the reduced model achieves this with fewer predictors, making it more parsimonious (Libretexts, 2022).
* Initial model F-statistic is 597.6 and reduced model F-statistic is 1791.0. In both cases, the F-statistic is significant, indicating that the models are better than an empty model.
* The initial model AIC is 196800, and the reduced model AIC is -7584. Lower AIC values indicate better model fit. The reduced model has substantially lower AIC, implying better performance.
* The initial model is BIC = 197000, and the reduced model BIC is 7534. Similar to AIC, lower BIC values indicate better fit, so the reduced model outperforms the initial model.

***E2. Residual plot and residual standard error for the reduced model.***

******

******

This Q-Q plot indicates that the residuals are not normally distributed. Non-normal residuals can indicate issues like outliers, heteroscedasticity, or model misspecification, but they don’t necessarily invalidate the model’s predictive power.

A Residual Standard Error of 0.166 indicates the typical size of the residuals in this regression model. This value suggests that, on average, the residuals (errors) are relatively small, which is a good sign for the model’s fit.

***E3. Code***

Please see the attached D208Task1.ipynb file.

***Part V: Data Summary and Implications***

***F1. Regression equation, Coefficients, etc***

***a) Regression equation***

*Additional\_charges = 0.0597 + 0.5880 \* Age + 0.0161 \* Stroke + 0.0072 \* Gender\_Male + 0.0184 \* Initial\_admin\_Emergency\_Admission – 0.0289 \* Complicaton\_risk\_Low – 0.0135 \* Complication\_risk\_Medium*

***b) Interpretation of the coefficients of the reduced model***

The coefficients provide insights into how each predictor variable affects the Additional\_charges. Positive coefficients indicate an increase in charges with an increase in the predictor, while negative coefficients indicate a decrease in charges.

* Intercept (0.0597): This is the expected value of Additional\_charges when all the predictor variables are zero. In practical terms, it represents the baseline Additional\_charges when all other factors are absent or at their reference levels.
* Age (0.5880): For each additional unit increase, the Additional\_charges are expected to increase by 0.5880 units, holding all other variables constant. This suggests that older patients tend to incur higher additional charges.
* Stroke (0.0161): For each unit increase in the Stroke variable, the Additional\_charges are expected to increase by 0.0161 units, holding all other variables constant. This indicates that patients with a stroke tend to have slightly higher charges.
* Gender\_Male (0.0072): For each unit increase in the Gender\_Male variable, the Additional\_charges are expected to increase by 0.0072 units, holding all other variables constant. This suggests that male patients tend to have slightly higher additional charges compared to female patients.
* Initial\_admin\_Emergency\_Admission (0.0184): For each unit increase in the Initial\_admin\_Emergency\_Admission variable, the Additional\_charges are expected to increase by 0.0184 units, holding all other variables constant. This indicates that patients initially admitted through emergency admission tend to have higher additional charges.
* Complication\_risk\_Low (-0.0289): For each unit increase in the Complication\_risk\_Low variable, the Additional\_charges are expected to decrease by 0.0289 units, holding all other variables constant. This suggests that patients with a low complication risk tend to have lower additional charges.
* Complication\_risk\_Medium (-0.0135): For each unit increase in the Complication\_risk\_Medium variable, the Additional\_charges are expected to decrease by 0.0135 units, holding all other variables constant. This indicates that patients with a medium complication risk tend to have lower additional charges than those with a high complication risk.

***c) Discussion regarding the statistical significance and practical significance of the model***

Statistical significance refers to the likelihood that the relationships observed in the data are not due to random chance. In this model, several indicators suggest that it is statistically significant:

* The F-statistic is 1791 with a p-value of 0.00. This very low p-value indicates that the overall regression model is statistically significant, meaning that at least one of the predictor variables is significantly related to Additional\_charges.
* All the predictor variables have p-values less than 0.05, indicating they are statistically significant. This means that each predictor variable contributes significantly to the model.
* The R-squared value is 0.518, and the adjusted R-squared is also 0.518. This indicates that the model explains approximately 51.8% of the variability in Additional\_charges. While not exceptionally high, it is a substantial proportion, suggesting that the model fits well.

Practical significance refers to the real-world relevance or importance of the findings. Even if a model is statistically significant, it might not be practically significant if the effect sizes are too small to matter in a real-world context.

The coefficients for Age (0.5880) and Initial\_admin\_Emergency\_Admission (0.0184) are relatively large, suggesting that these variables have a meaningful impact on Additional\_charges. For instance, an increase in age significantly increases the additional charges, which could be important for healthcare cost management. The variables included in the model (e.g., Age, Stroke, Gender\_Male, Initial\_admin\_Emergency\_Admission, Complication\_risk\_Low, Complication\_risk\_Medium) are relevant to healthcare settings. Understanding how these factors influence additional charges can help in resource allocation, policy-making, and improving patient care.

The model can predict additional charges for new patients based on their characteristics. This can help healthcare providers budget and plan. For example, knowing that older patients or those admitted through emergency have higher additional charges can help in better financial planning and resource allocation.

***d) Disadvantages of the methods used to conduct the regression model***

The methods used in this regression analysis have several potential disadvantages and implications, focusing on data preparation, manipulation, and model reduction.

* Retaining outliers in a dataset can significantly impact regression analysis in several ways. Outliers are data points that deviate from other observations and can distort the estimated relationships between variables. When outliers are present, they can lead to biased estimates of the regression coefficients, reducing the model's accuracy. Outliers can also affect the regression model's assumptions, such as the assumption of normally distributed residuals.
* Encoding categorical variables, such as one-hot encoding, increases the dimensionality of the dataset, which can lead to overfitting, especially with a small sample size. It can also introduce multicollinearity if not handled properly, affecting the stability and interpretability of the regression coefficients.
* Using VIF to detect multicollinearity is helpful but does not address the underlying issue. Simply removing variables with high VIF might not always be the best solution, as this method can lead to the exclusion of important variables that are highly correlated with other predictors but still provide valuable information, resulting in a model that lacks interpretability and practical relevance.
* Scaling and normalization help standardize the data but can sometimes obscure the natural relationships between variables. For instance, if the original scale of a variable has practical significance, transforming it might make interpretation less intuitive. Over-reliance on scaling can lead to loss of interpretability and, if not done correctly, can introduce biases, especially if the scaling method is inappropriate for the data distribution.
* Removing variables solely based on p-values can lead to excluding important predictors that might have practical significance, even if they are not statistically significant. This approach can result in a model that is too simplistic and fails to capture the complexity of the data, leading to model instability where small changes in the data result in different sets of selected variables.

***F2. Recommendations***

Based on the regression analysis and findings, here are some recommended actions for an organization, particularly in a healthcare setting:

Allocate more resources to older patients and those admitted through emergency admissions. The regression model indicates that Age and Initial\_admin\_Emergency\_Admission are significant predictors of higher additional charges. The organization can better manage costs and improve patient care by focusing resources on these groups.

Implement preventive measures and early interventions for patients with higher complication risks. Although Complication\_risk\_Low and Complication\_risk\_Medium are associated with lower additional charges, focusing on high-risk patients can help reduce overall costs and improve outcomes. Preventive measures can include regular check-ups, lifestyle interventions, and patient education.

Develop gender-specific health programs to address the unique needs of male patients. The model shows that Gender\_Male is a significant predictor of additional charges, albeit with a small coefficient. Tailored health programs can help address male patients' specific health issues, potentially reducing additional charges.

Enhance stroke management programs to reduce additional charges for stroke patients. The Stroke variable is a significant predictor of additional charges. The organization can potentially reduce these costs by improving stroke management, including rehabilitation and follow-up care.

Continuously monitor the model’s performance and update it with new data. Healthcare data is dynamic, and patient demographics and health trends can change. Regularly updating the model ensures that it remains accurate and relevant, allowing the organization to make informed decisions.

Provide training and education to healthcare staff on the importance of data-driven decision-making. Educating staff on interpreting and using the model’s findings can lead to more informed and effective decision-making, ultimately improving patient care and reducing costs.

Engage patients in their care through education and self-management programs. Educating patients about their health conditions and involving them in their care can lead to better health outcomes and potentially lower additional charges. This is particularly important for managing chronic conditions and preventing complications.

***Part VI: Demonstration***

***G. Panopto video***

Please see the attached link to the Panopto video.

***Sources***

***H. Web sources of third-party code***

Li, S. (2019, February 27). Building a logistic regression in Python, step by step. *Medium*. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

Sharma, N. (2024, March 17). *How to do One Hot Encoding? Transform Your Categorical Data!* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/12/how-to-do-one-hot-encoding/

*Pandas get one hot encodings from a column as booleans*. (n.d.). Stack Overflow. https://stackoverflow.com/questions/56758567/pandas-get-one-hot-encodings-from-a-column-as-booleans

*Quantile-Quantile Plot using SciPy*. (n.d.). Stack Overflow. https://stackoverflow.com/questions/13865596/quantile-quantile-plot-using-scipy

Mark Keith. (2021, October 11). *Python: MLR, OLS, standardization, normalization* [Video]. YouTube. https://www.youtube.com/watch?v=QH\_elD\_JKuc

Mark Keith. (2021a, October 11). *Python: Intro to MLR / OLS in StatModels.API* [Video]. YouTube. https://www.youtube.com/watch?v=0-fkgpK2knA

***I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized***

BasuMallick, C. (2022, December 6). *The 12 vital differences between R and Python*. Spiceworks Inc. <https://www.spiceworks.com/tech/devops/articles/r-vs-python/>

Bobbitt, Z. (2021, November 16). *The five assumptions of multiple linear regression*. Statology. https://www.statology.org/multiple-linear-regression-assumptions/

Bonthu, H. (2024, July 9). *Detecting and Treating Outliers | Treating the odd one out!* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/

*Interpreting regression results*. (2021, June 4). Doc’s Things and Stuff. https://docmckee.com/oer/statistics/section-7/section-7-1/interpreting-regression-results/

Michael Parker. (2019, April 26). *Multiple Linear Regression with High VIF and P-Values* [Video]. YouTube. https://www.youtube.com/watch?v=gxqtU5tgoos

Polat, E. I. (2022, February 26). Hypothesis Testing with Python: Step by step hands-on tutorial with practical examples. *Medium*. https://towardsdatascience.com/hypothesis-testing-with-python-step-by-step-hands-on-tutorial-with-practical-examples-e805975ea96e

*R or Python*. (2023, July 7). Western Governors University. <https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html#transcriptPop>

*Panopto*. (n.d.). Panopto. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=39bbe2db-de7d-4bf5-913b-af5c0003da9d

*Panopto*. (n.d.-b). Panopto. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=15e09c73-c5aa-439d-852f-af47001b8970

*Redirecting*. (n.d.-l). https://westerngovernorsuniversity.sharepoint.com/:w:/r/sites/DataScienceTeam/\_layouts/15/Doc.aspx?sourcedoc=%7B1F520BDD-B52F-4C02-BB62-D369749A4700%7D&file=Dr.%20Middleton%20PA%20Step-by-Step%20Guide%20(NBM3).docx&action=default&mobileredirect=true

AIML.com. (2024). What are the assumptions of linear regression? *AIML.com*. https://aiml.com/what-are-the-assumptions-of-linear-regression/