D208: Predictive Modelling

Task I

Western Governors University

-Oviya Selvaraj

A1. Research question:

How do Vitamin D levels, along with additional factors such as patient demographics, additional charge, and type of disease predict total medical costs for patients?

A2. Goals:

The goals of the data analysis are to identify the most significant factors that impact total medical costs. This involves determining how variables such as vitamin D levels, additional charges and demographic factors (like age, gender and income) relate to and influence the total cost incurred by patients. By analyzing the predictors, we can understand what drives high medical costs. This could provide insights for healthcare providers, policymakers, or insurance companies to focus on cost-control measures or preventive care strategies.

B1. Summary of assumptions:

The four assumptions of a Linear Regression model are:

1. There is a linear relationship between independent and dependent variables.
2. The residuals are independent.
3. In linear regression, homoscedasticity is the assumption that the variance of the error term is constant across all values of the independent variables.
4. The residuals in the model are normally distributed.

B2. Tools Benefit

Language used: Python

1. **Extensive libraries:** Python provides a wide range of libraries for data manipulation, visualization and statistical analysis like Pandas, NumPy, Matplotlib and SciPy.
2. **Easy to learn:** Python has a simple syntax and is easy to understand, making it easier to read and learn.

B3. Appropriate Technique

Multiple Linear Regression is used to understand and predict the influence of multiple independent variables on a single dependent variable. This technique allows us to analyze complex relationships where a single factor cannot fully explain the outcome. According to the research question, the independent features Vitamin D levels, additional charges, and demographic factors (like age, gender and income) are used to predict Total hospital charges. Thus, this method allows us to analyze multiple factors that interact and contribute with the independent factor.

C1. Data Cleaning

I use a boxplot to visualize continuous data such as age, vitamin D levels, income, additional charges, and total charges for the data cleaning process. Upon visualization, it can be observed that Income, Vitamin D levels and Additional charges contain outliers. Using the Interquartile Range Method (IQR), I mitigate the outliers for ‘Income’ and ‘Additional Charges.’ Observing the Vitamin D levels using the ‘describe’ function, it can be seen that the minimum is 9.8, and the maximum value is 26.40, which seems legitimate.

Thus, it will remain. There are no outliers for Age and TotalCharge. Also, there are no missing values or duplicates for any of the data used in my analysis.

C2. Summary Statistics

For the purposes of the analysis, I decided to only use the following data columns 'Income', 'Gender', 'Age', 'VitD_levels', 'HighBlood', 'Stroke','Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain', 'Anxiety', 'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma', 'TotalCharge', 'Additional_charges' and 'Initial_admin'. The dataset has been sufficiently cleaned leaving no null, NAs or missing data points.

Continuous data:

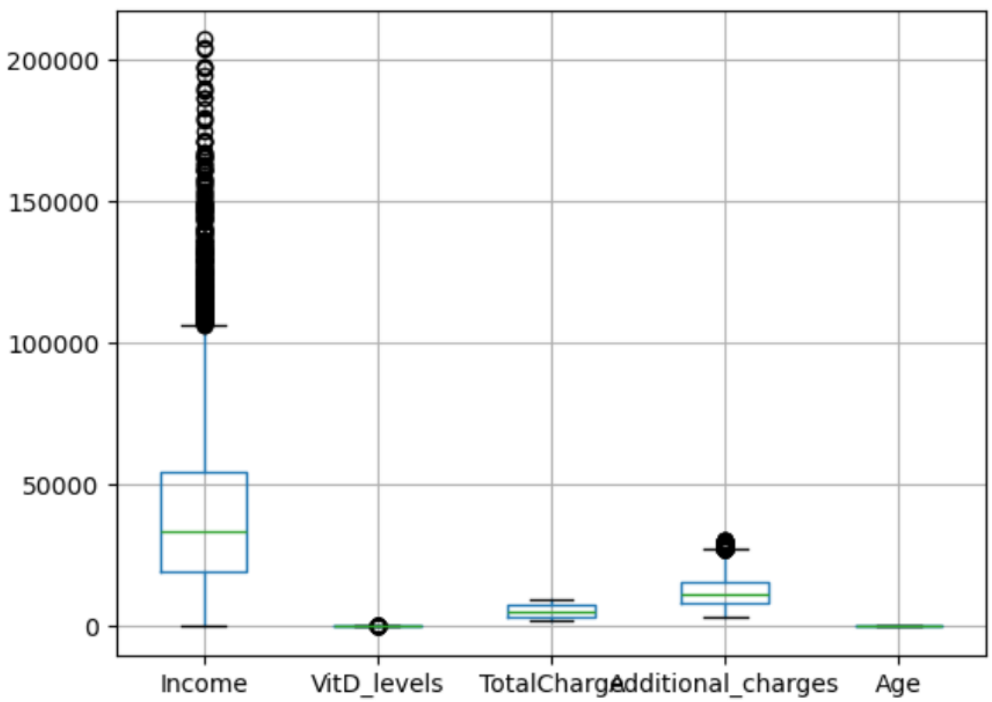
- **Income**: The histogram concludes a Uniform Distribution curve. The mean or average income of patients is $ 39721.119.
- **VitD_levels (Vitamin D Levels)**: The histogram shows a uniform distribution curve. The minimum value is 9.8 and the maximum value of 26.40 seems legitimate.
- **TotalCharge**: The histogram shows a bimodial distribution curve with an average hospital total charge of $ 5312.17.
- **Age**: The histogram shows a uniform distribution with slightly fewer occurrences at the extreme ends (ages below 20 and above 90).
- **Additional_charges**: The histogram shows a concentration of values around 10,000 to 13,000, with a gradual decline in frequency as you move to the right, followed by a small increase near 25,000 (due to the capping effect).

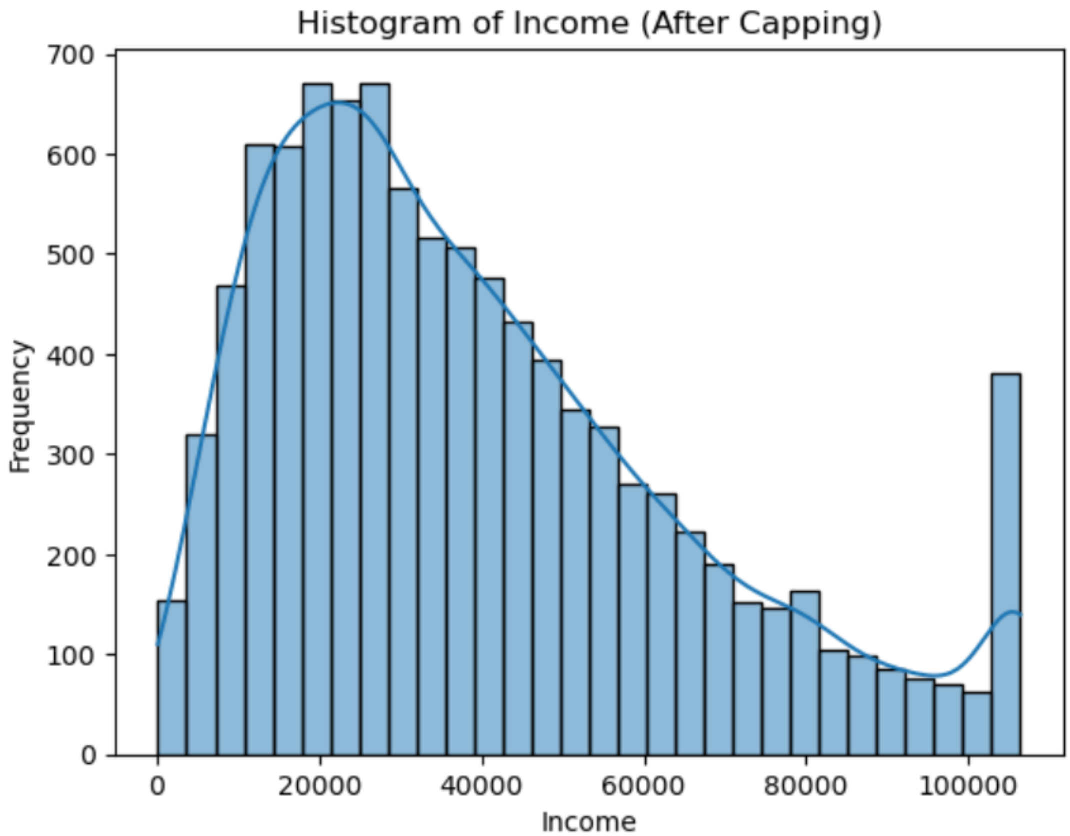
Categorical data: All the categorical data show Bernoulli distribution due to the binary nature except Initial_Admin (Initial Admission) and Gender.

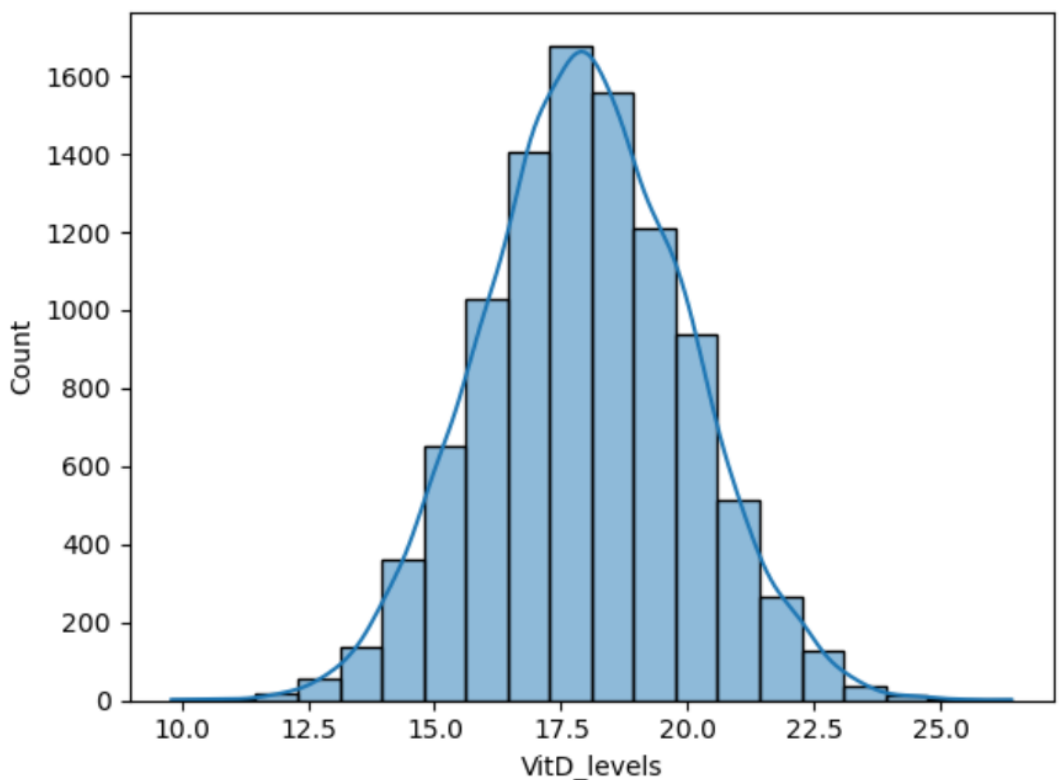
Initial_admin contains three categorical values ‘Emergency admission’, ‘Elective admission’ and ‘Observation’. For Gender there are three categorical values (Male, Female and NonBinary). The frequency table is used to show the count of the categorical values.

C3. Univariate and Bivariate statistics

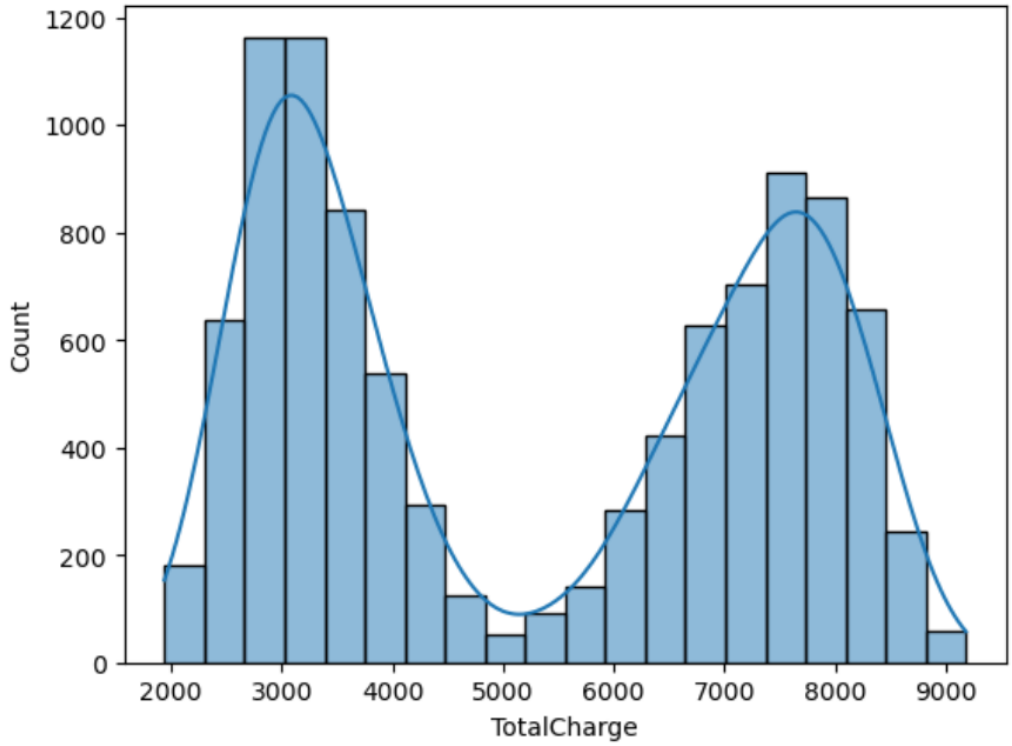
Univariate statistics for continuous and categorical data:

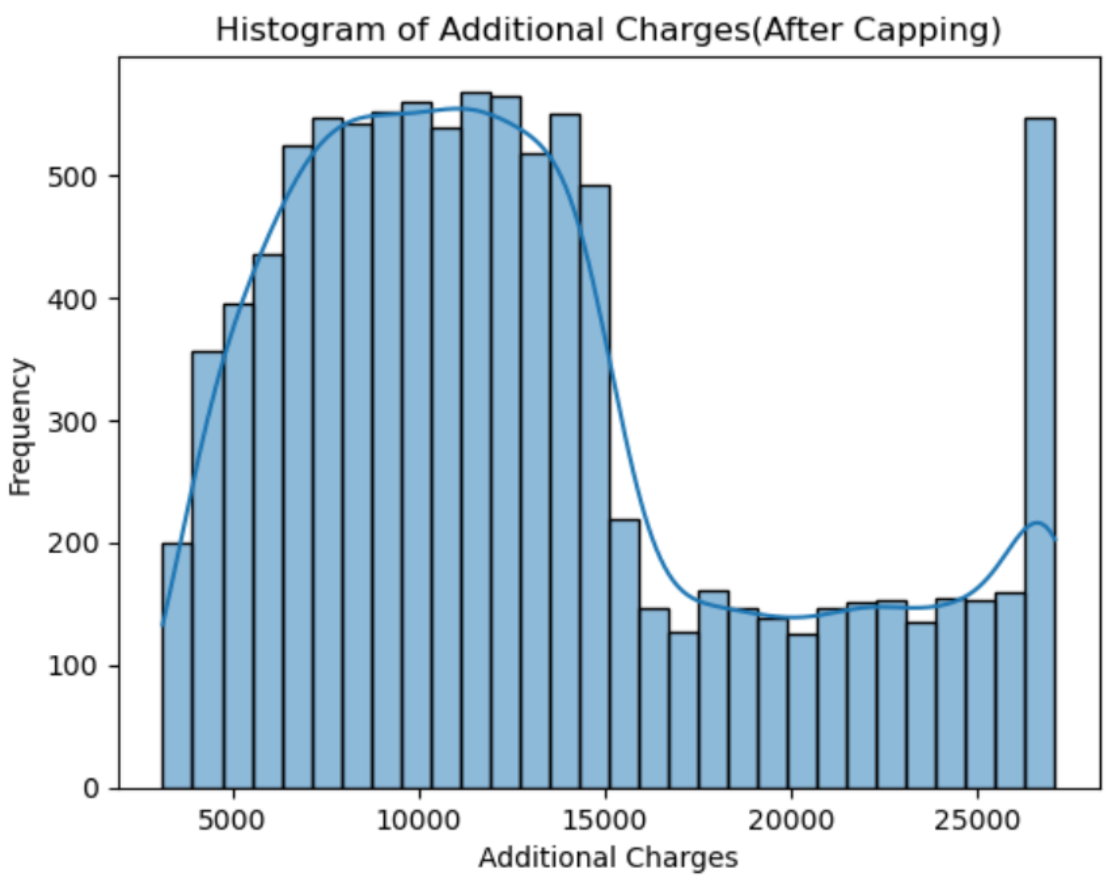


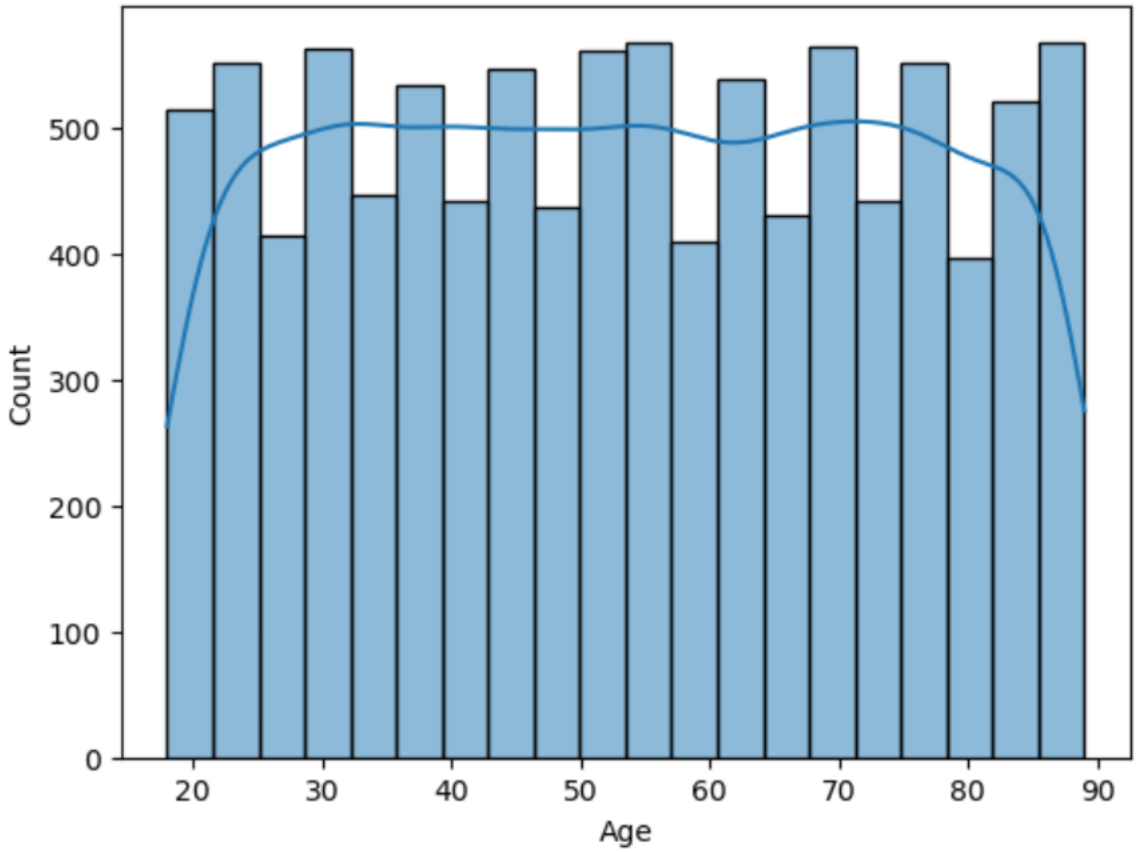

Histogram of Income (After Capping)
Frequency
Income
0
100
200
300
400
500
600
700
0
20000
40000
60000
80000
100000


Count
VitD_levels
0
200
400
600
800
1000
1200
1400
1600
10.0
12.5
15.0
17.5
20.0
22.5
25.0

Total Charges:

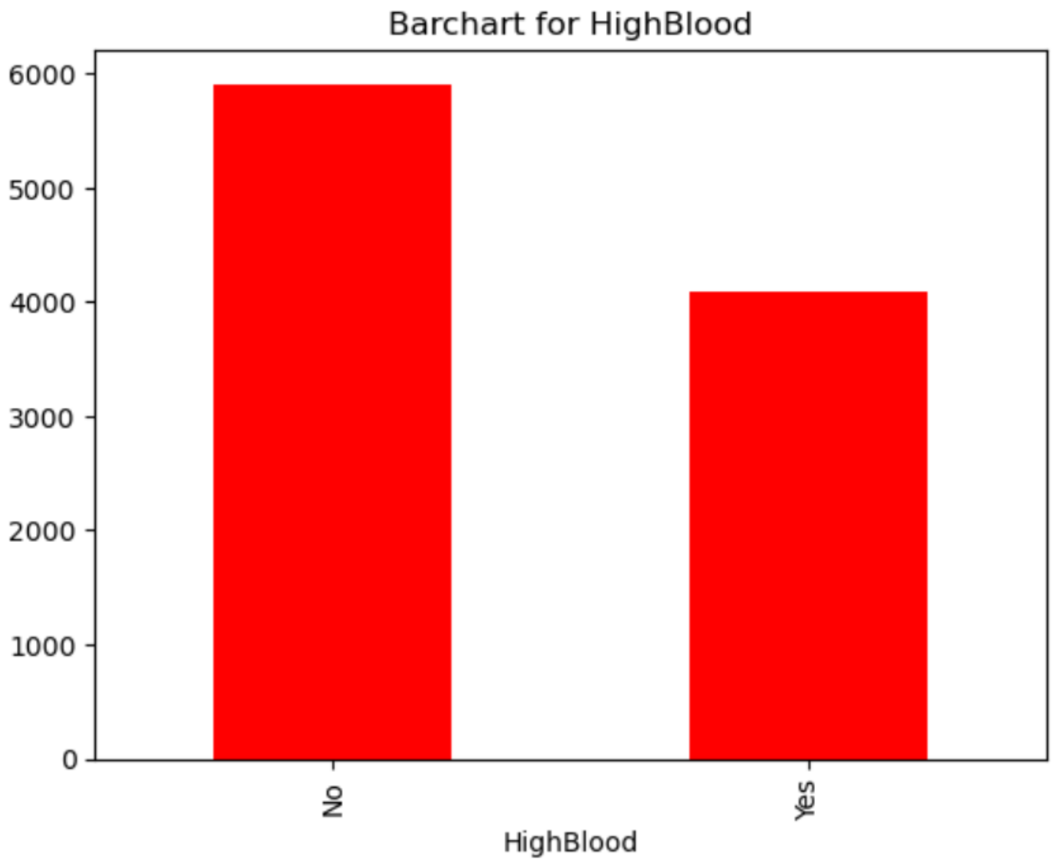




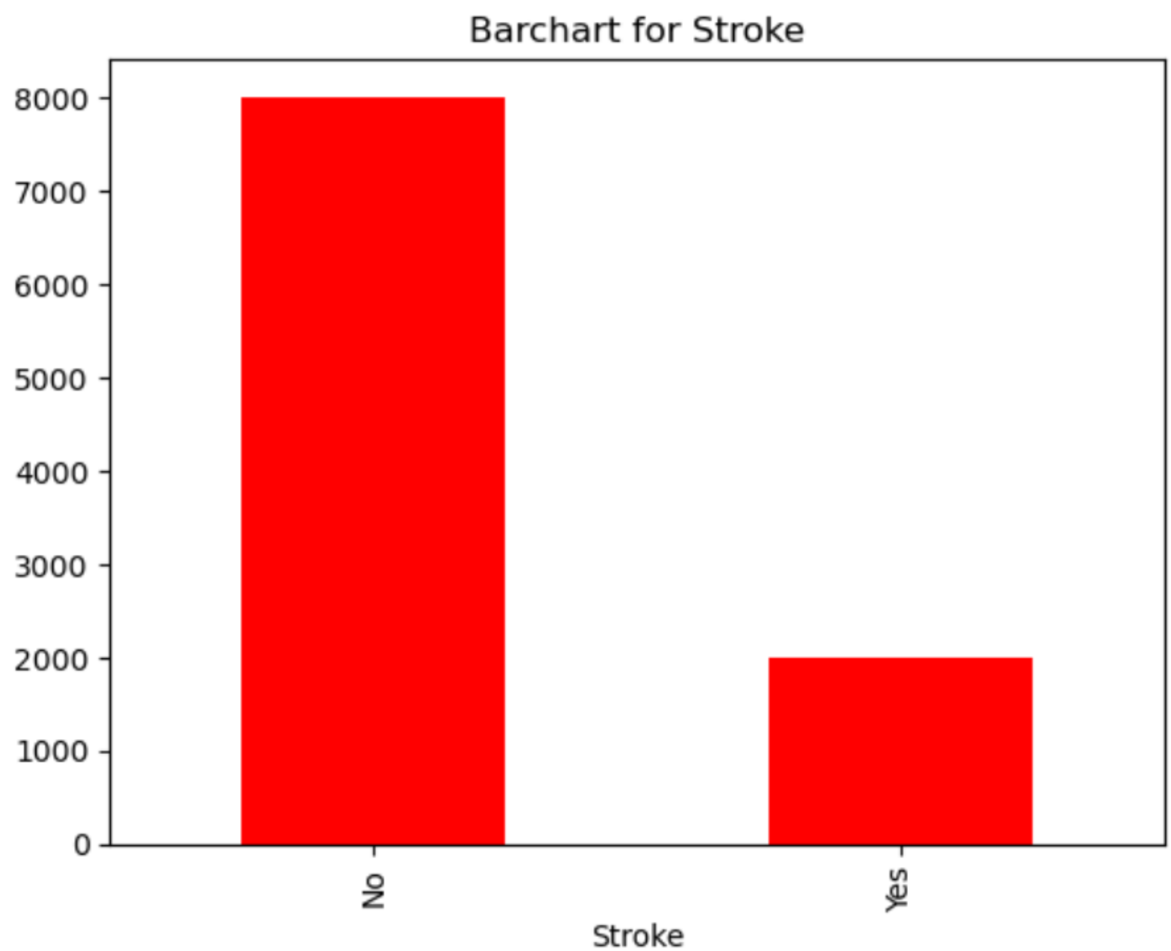


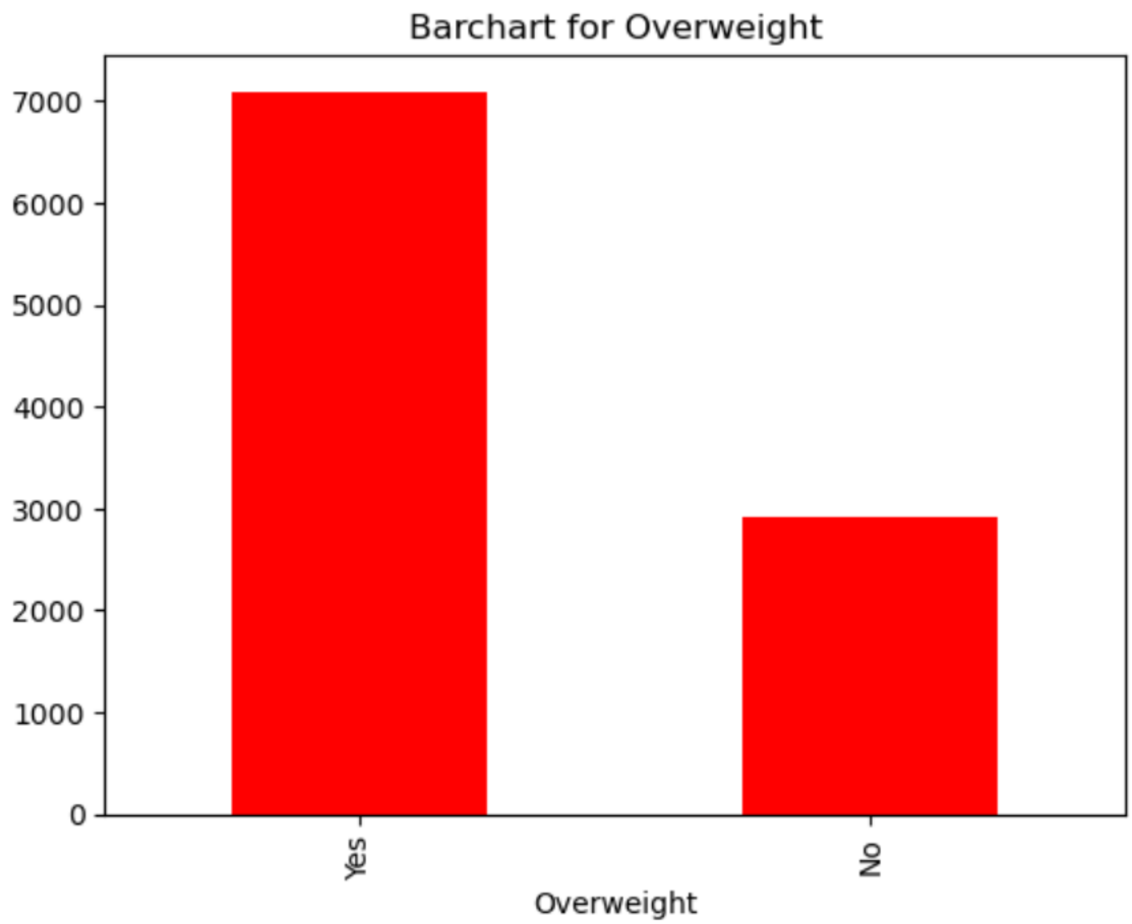
Frequency table for Gender column:

col_0	count
Gender	
Female	5018
Male	4768
Nonbinary	214

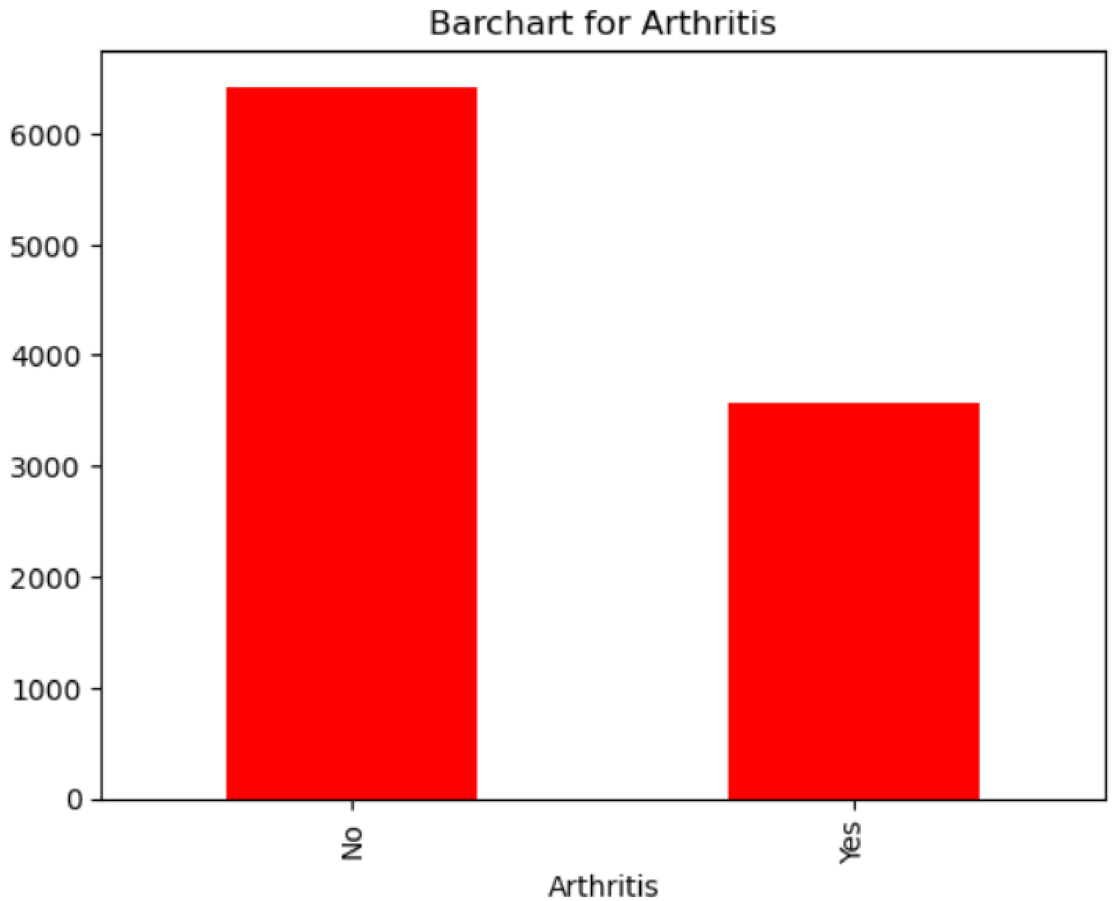


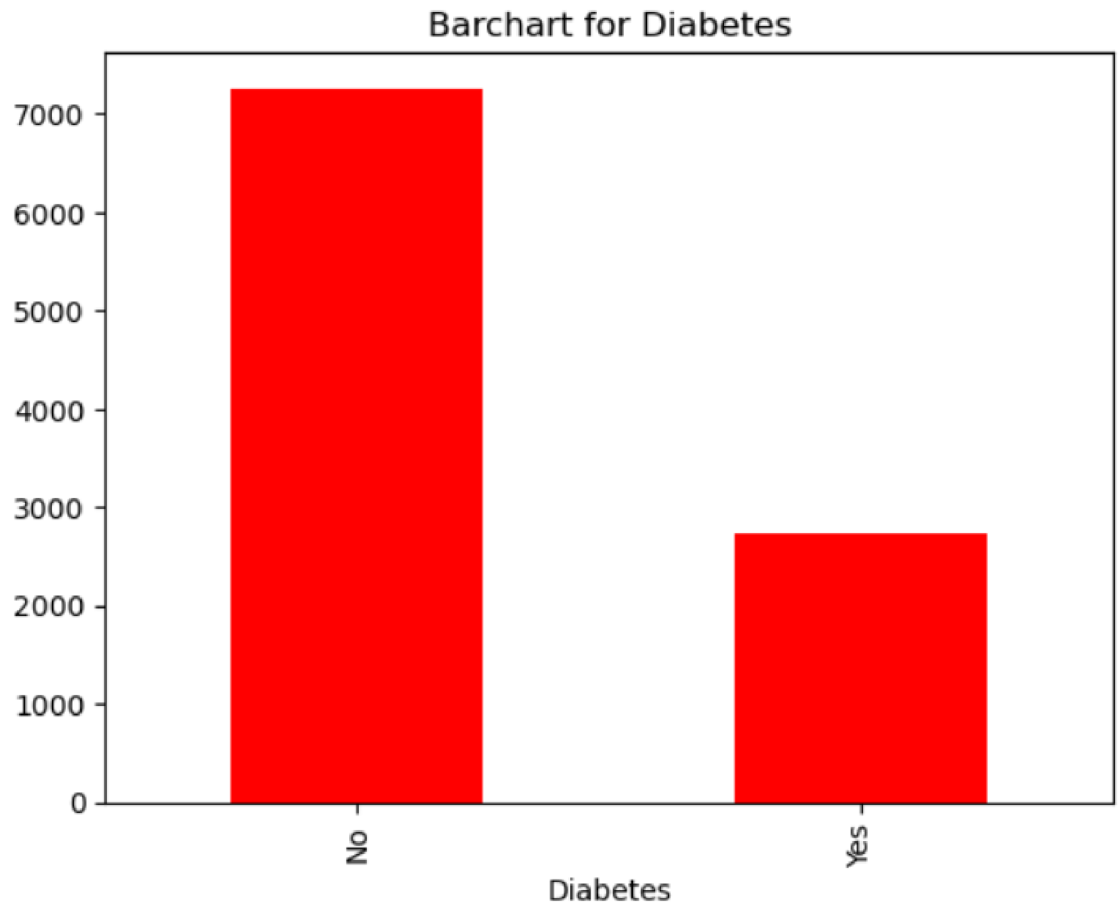
Barchart for HighBlood
6000
5000
4000
3000
2000
1000
0
No
Yes
HighBlood

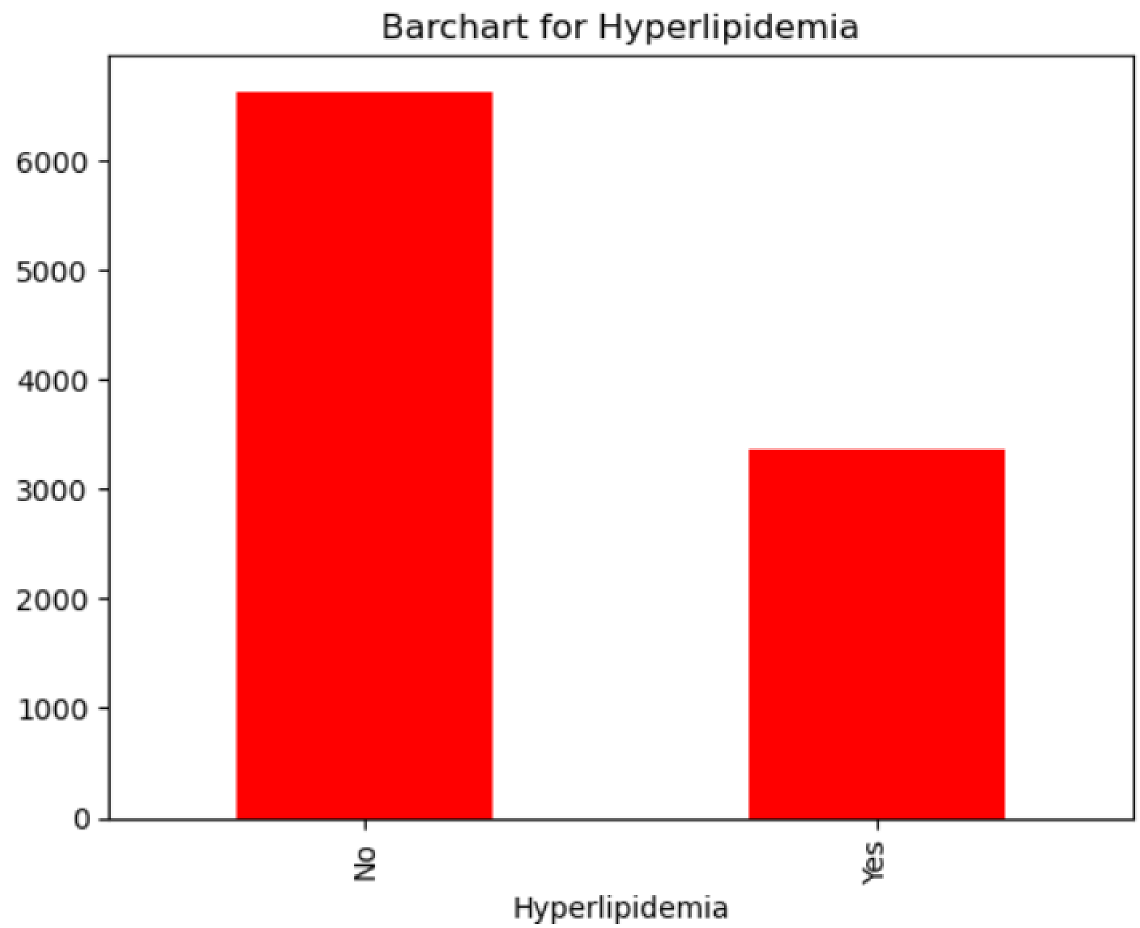

Barchart for Stroke
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Stroke

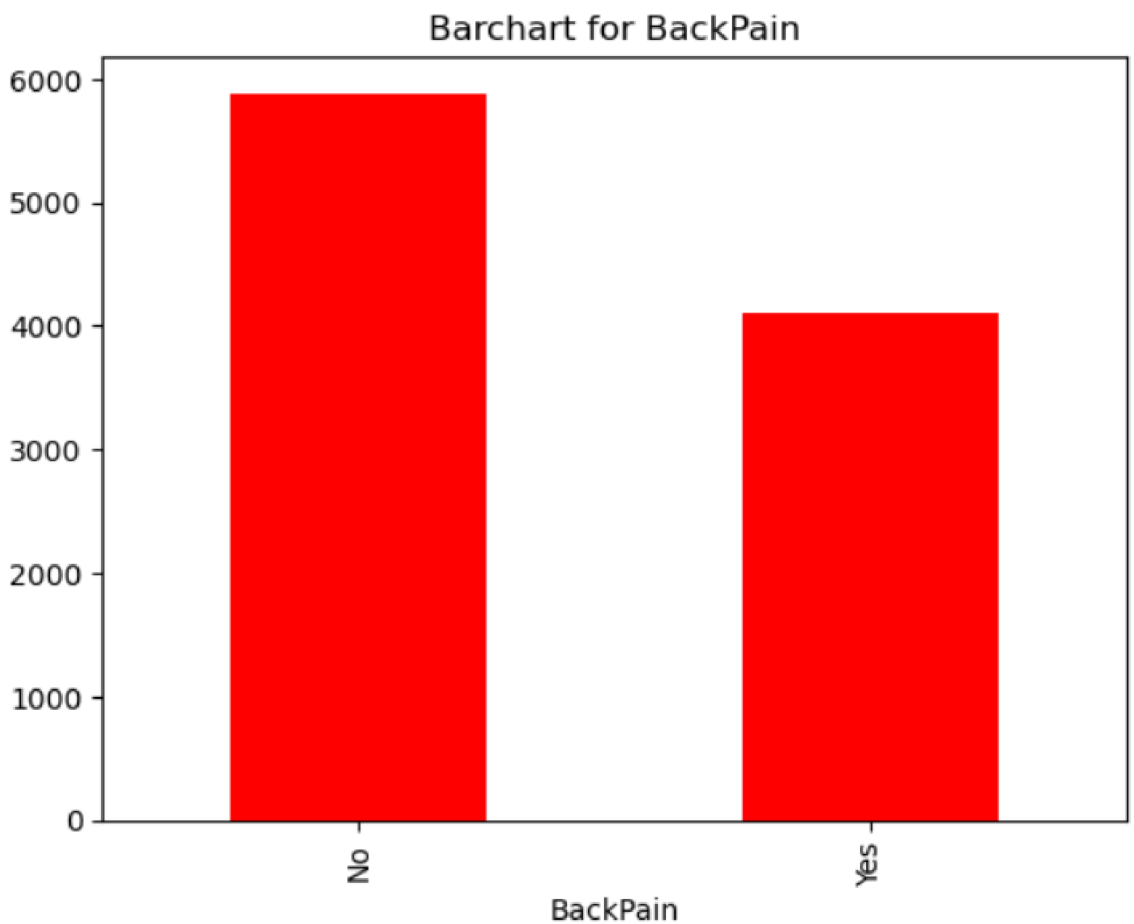


Barchart for Overweight
7000
6000
5000
4000
3000
2000
1000
0
Yes
No
Overweight

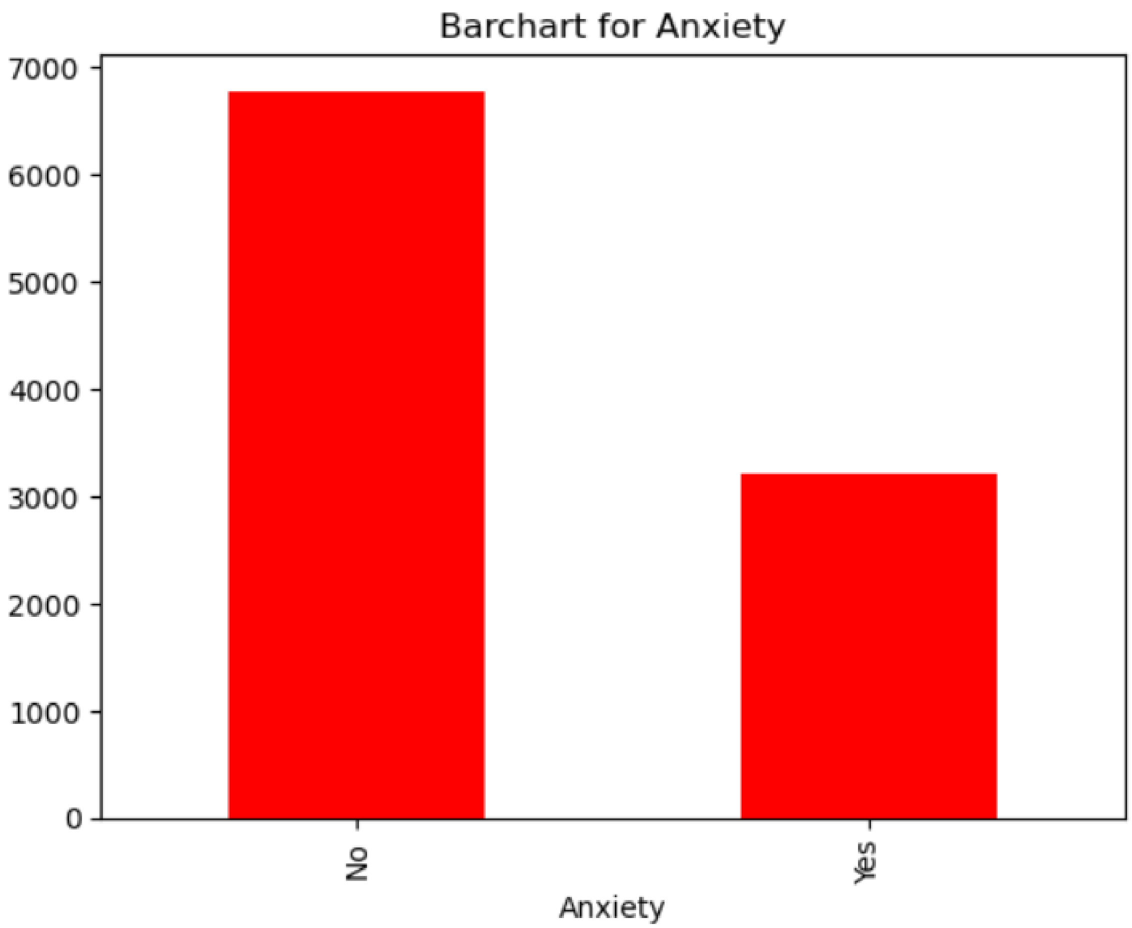

Barchart for Arthritis
6000
5000
4000
3000
2000
1000
0
No
Yes
Arthritis

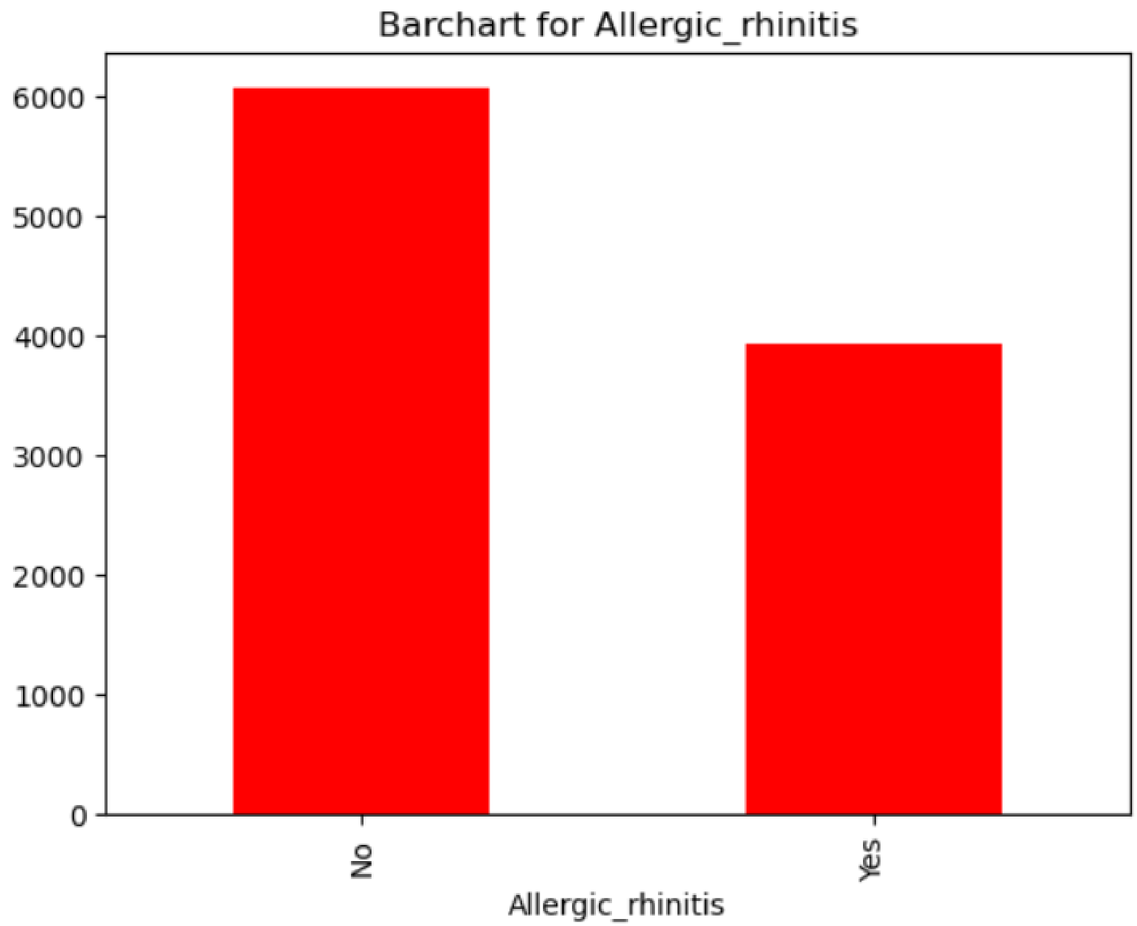

Barchart for Diabetes
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Diabetes

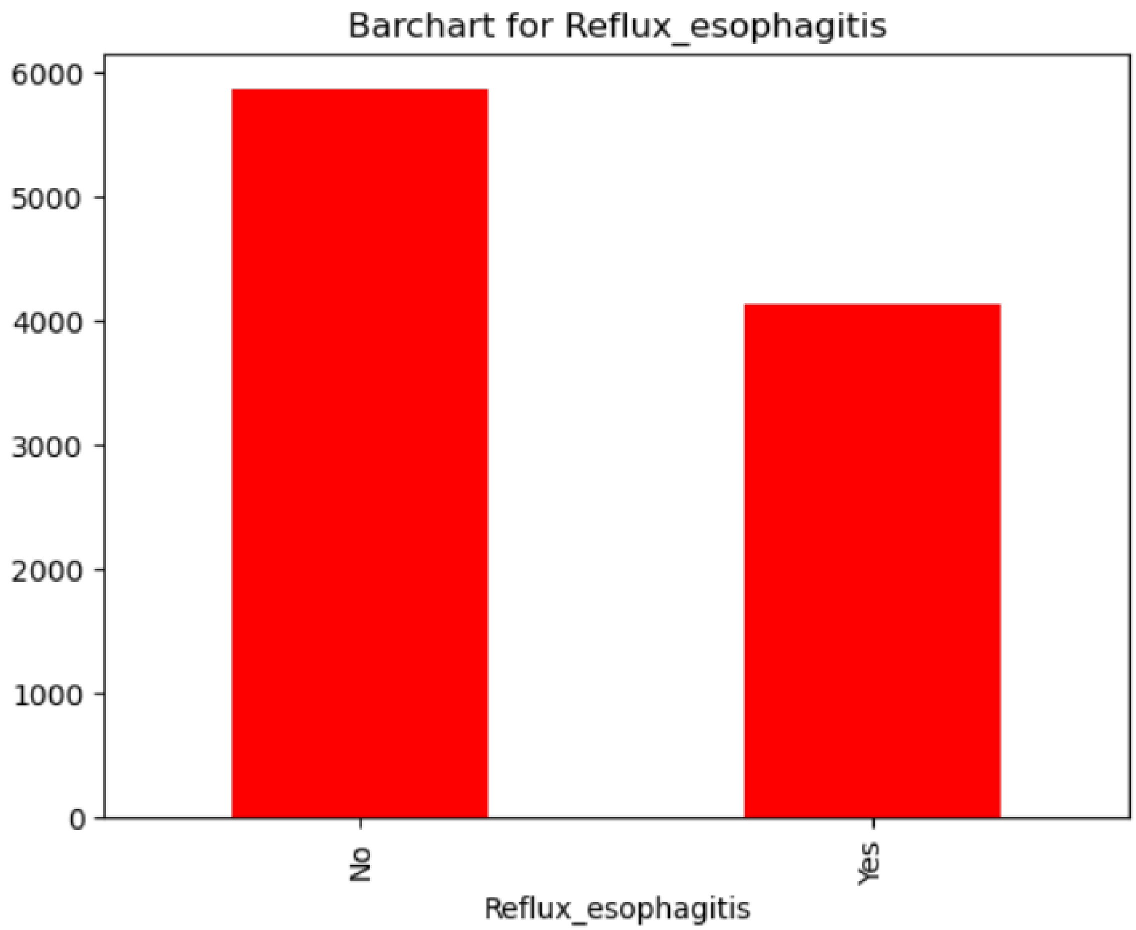

Barchart for Hyperlipidemia
6000
5000
4000
3000
2000
1000
0
No
Yes
Hyperlipidemia

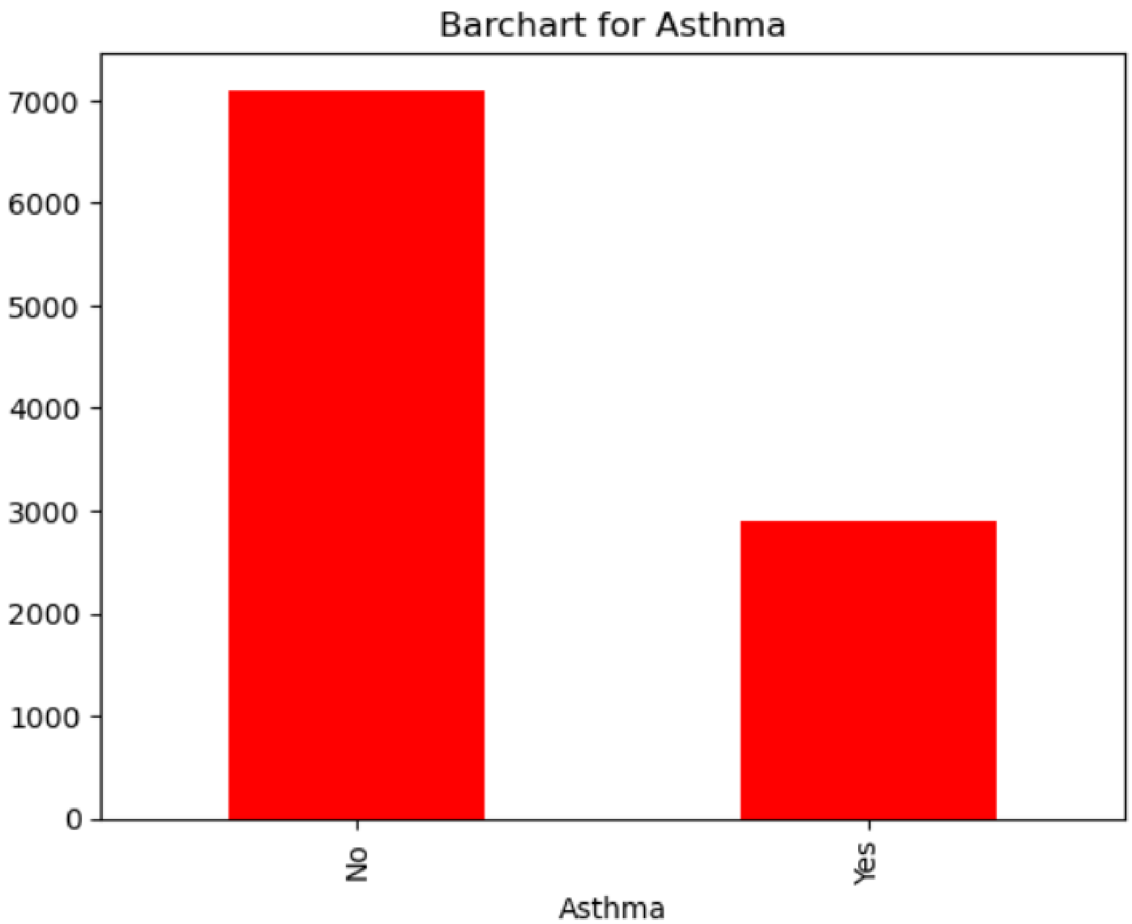


Barchart for BackPain
6000
5000
4000
3000
2000
1000
0
No
Yes
BackPain


Barchart for Anxiety
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Anxiety


Barchart for Allergic_rhinitis
6000
5000
4000
3000
2000
1000
0
No
Yes
Allergic_rhinitis

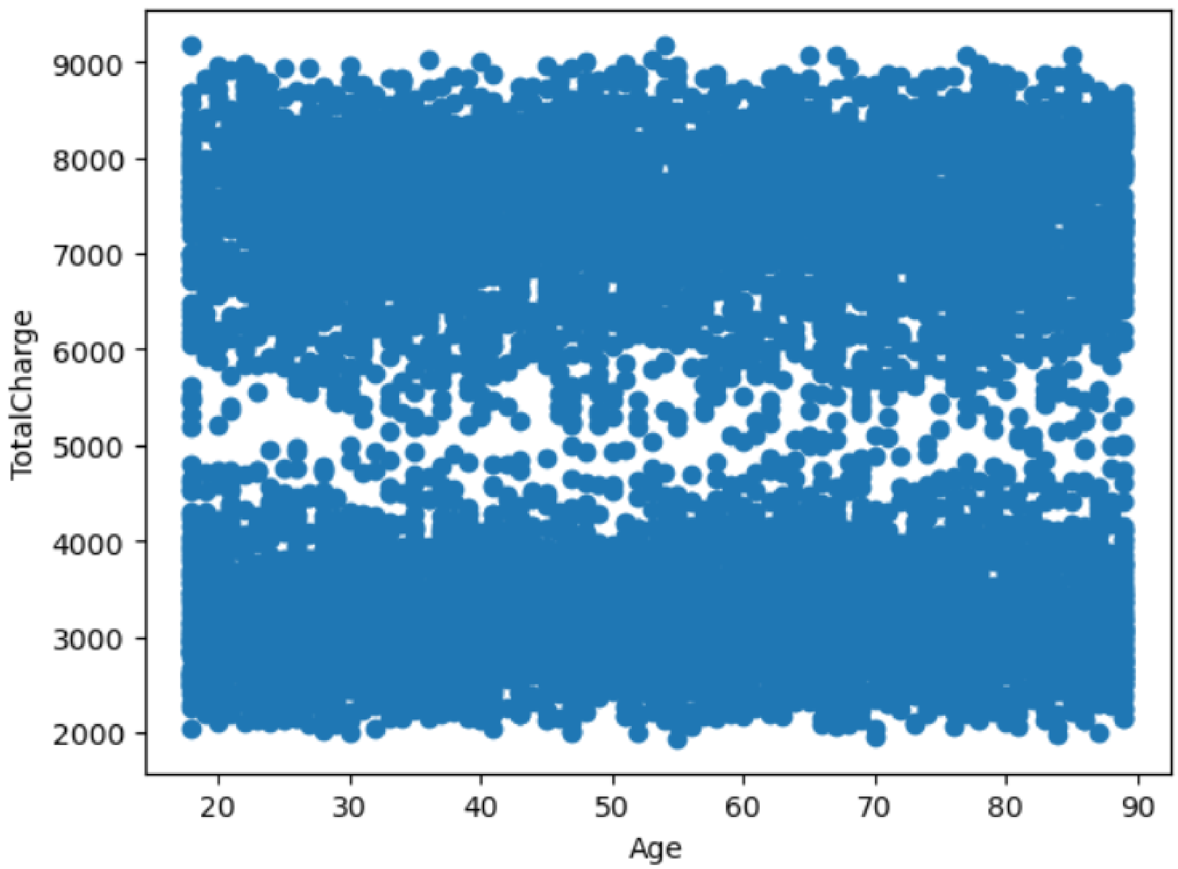

Barchart for Reflux_esophagitis
6000
5000
4000
3000
2000
1000
0
No
Yes
Reflux_esophagitis

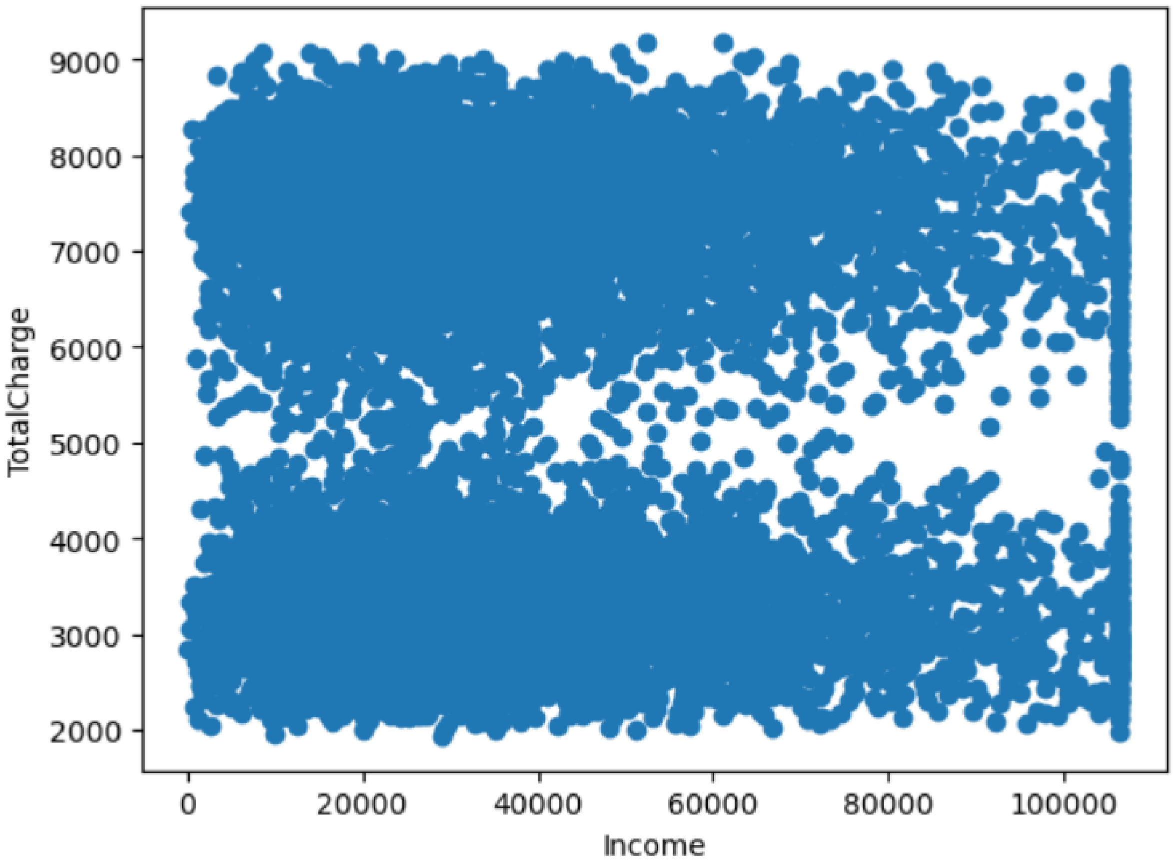


Frequency table for Initial Admission:

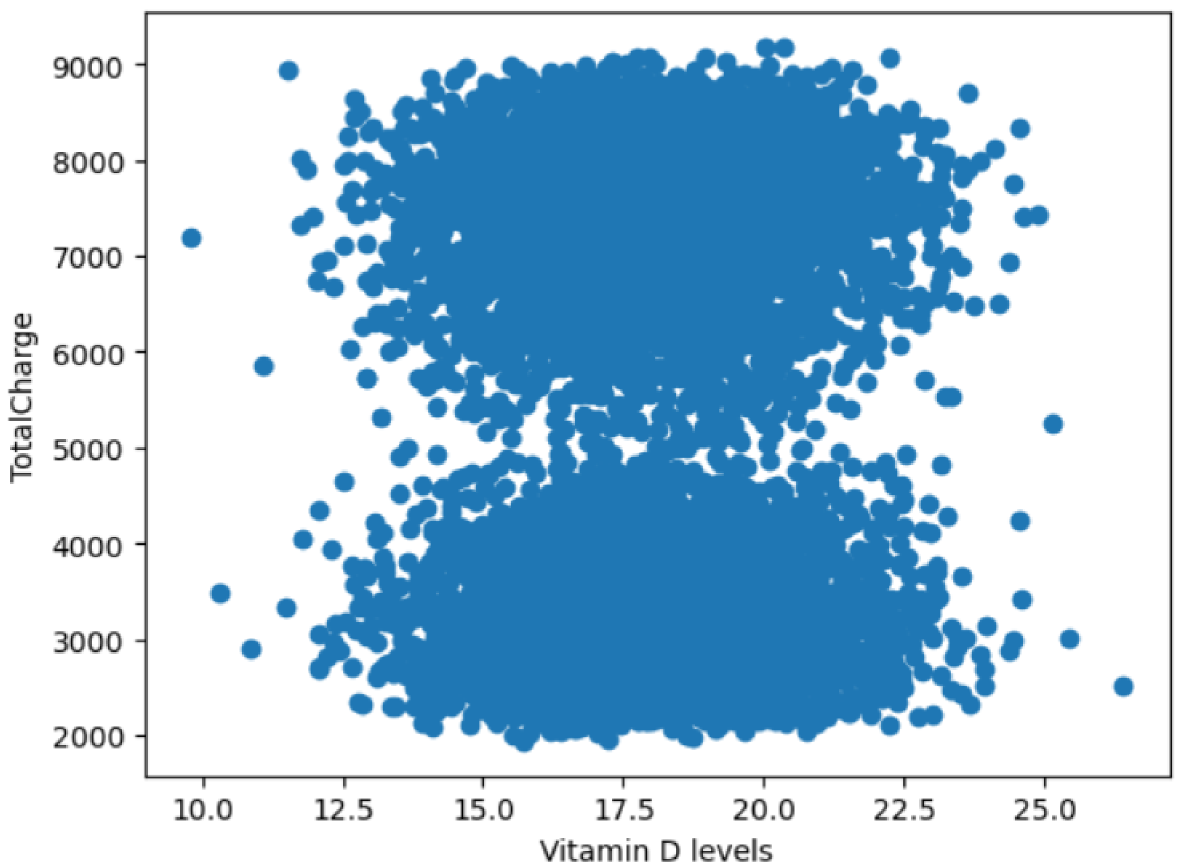
col_0	count
Initial_admin	
Elective Admission	2504
Emergency Admission	5060
Observation Admission	2436

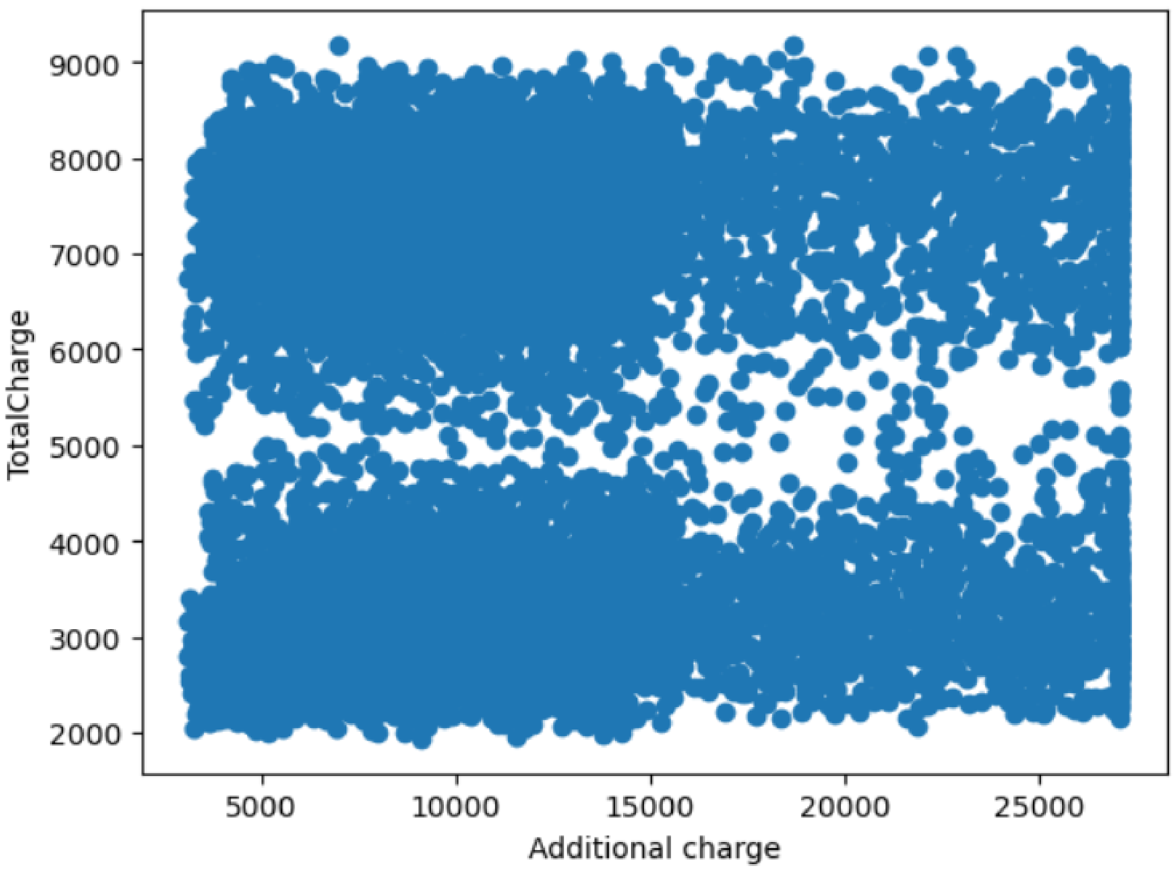
Bivariate Statistics for continuous and categorical data:

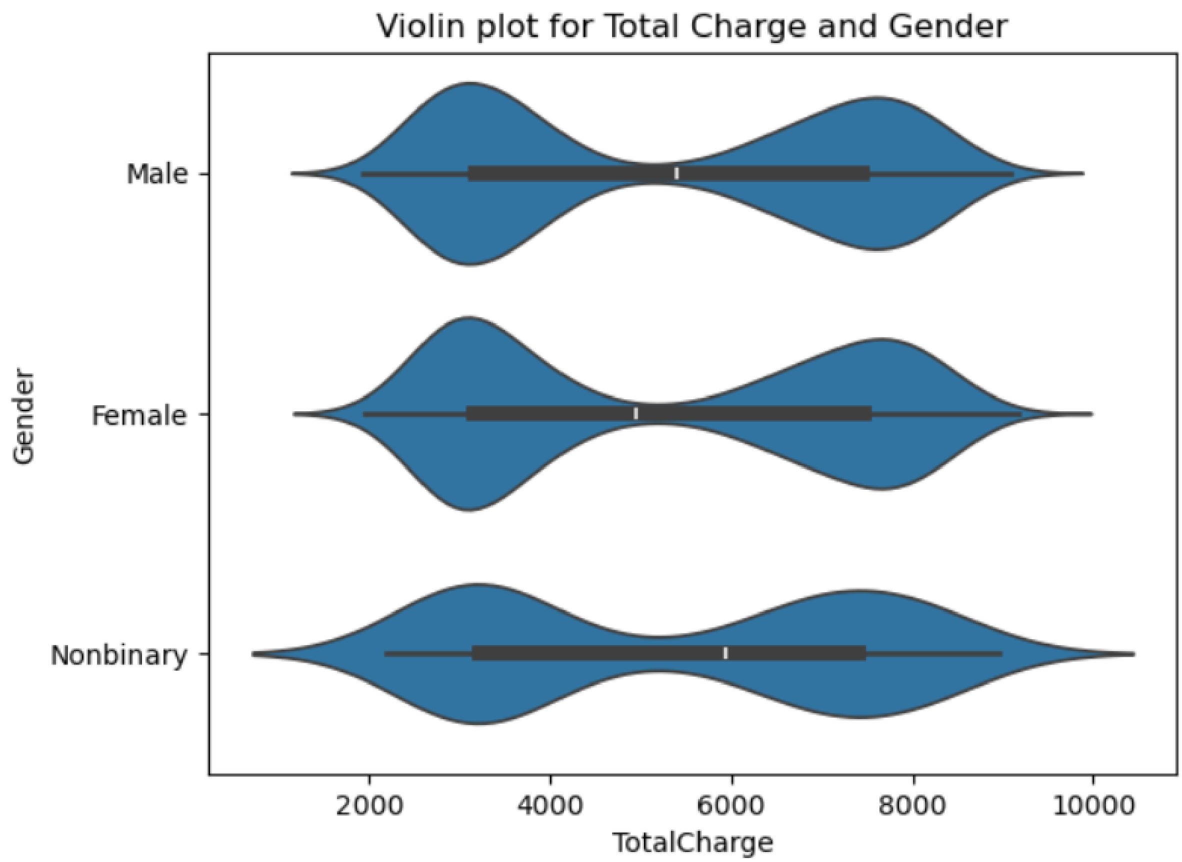


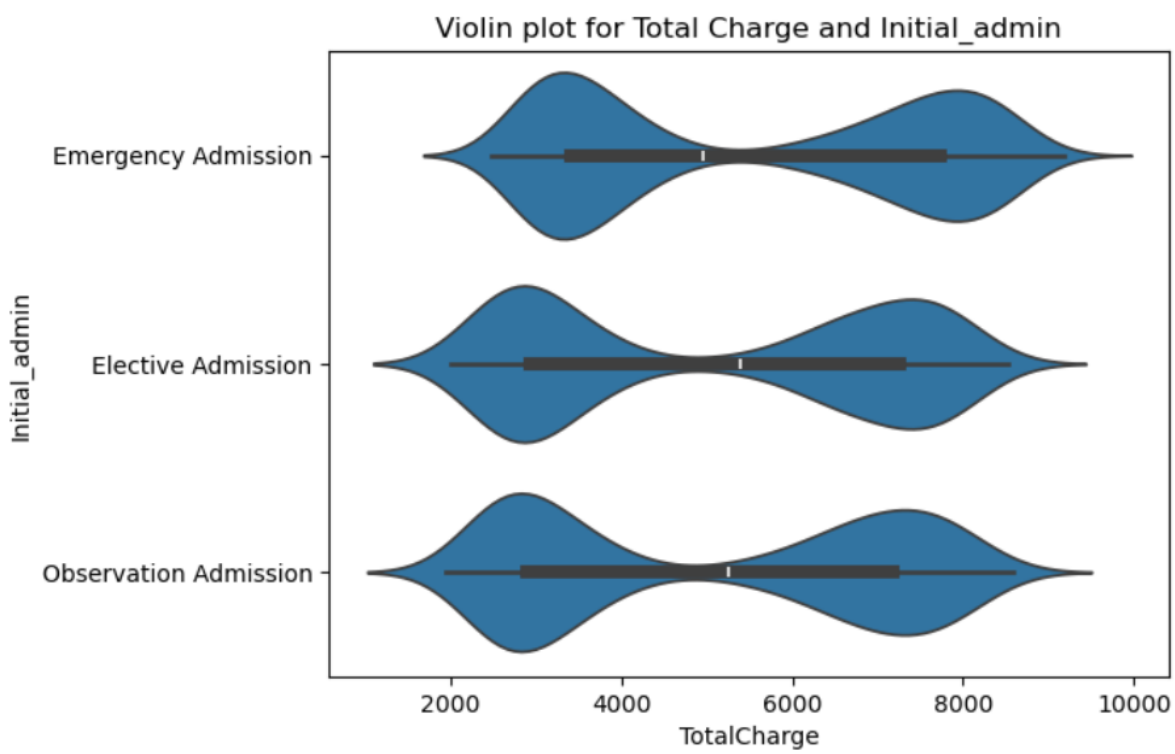


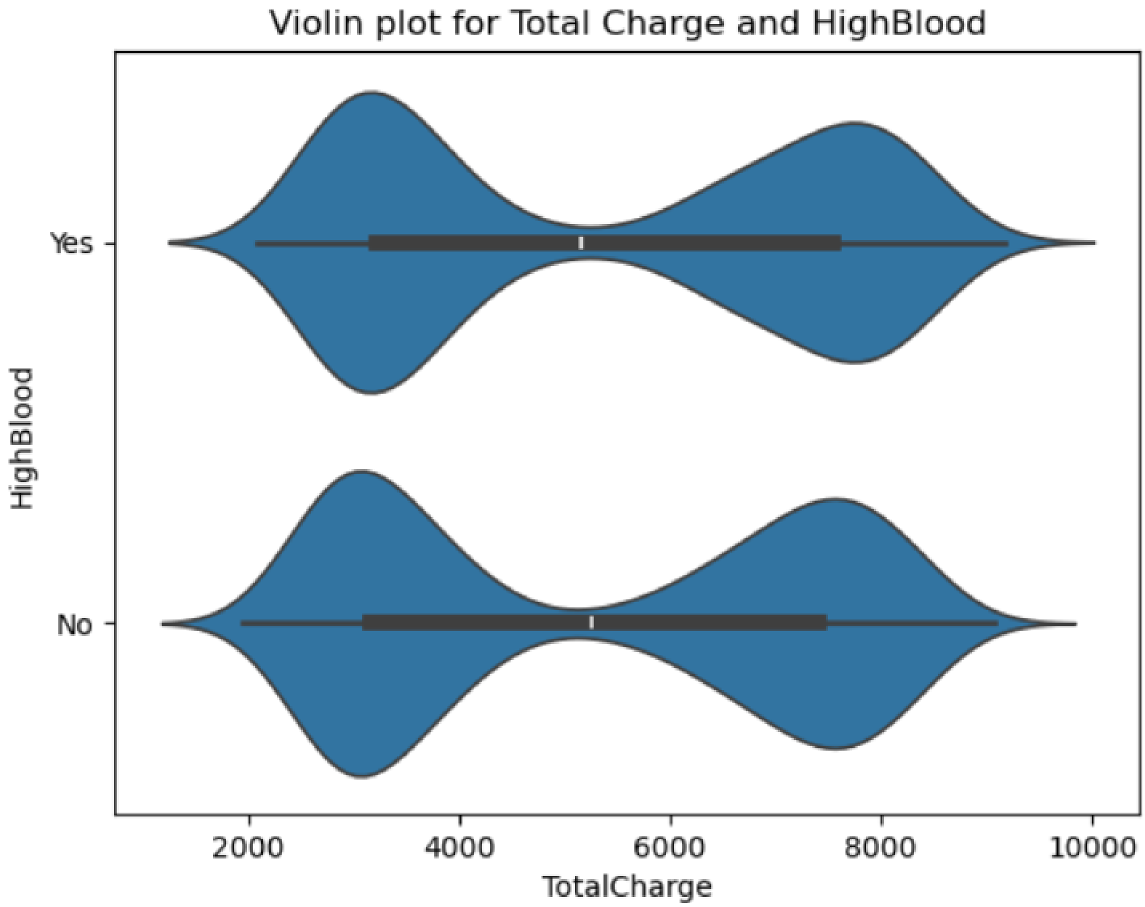
TotalCharge
9000
8000
7000
6000
5000
4000
3000
2000
0
20000
40000
60000
80000
100000
Income

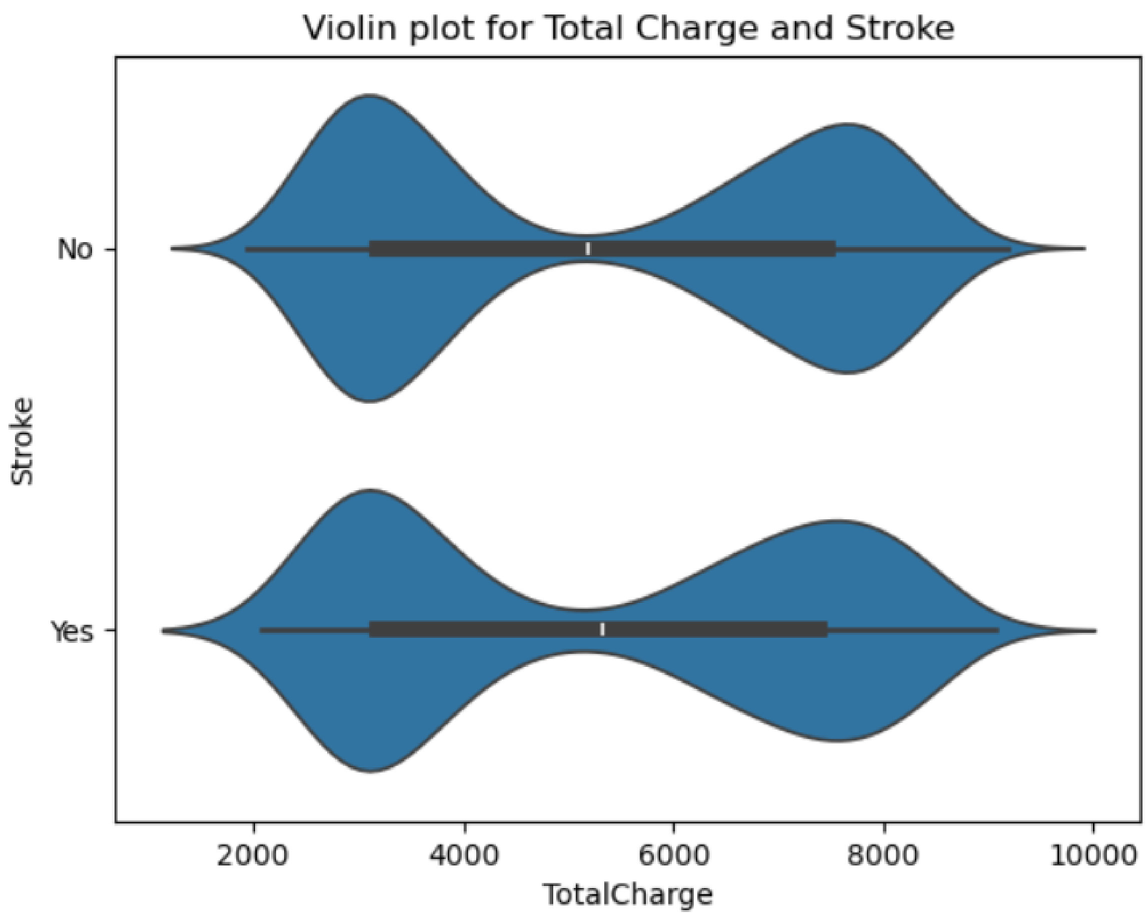

TotalCharge
9000
8000
7000
6000
5000
4000
3000
2000
10.0
12.5
15.0
17.5
20.0
22.5
25.0
Vitamin D levels

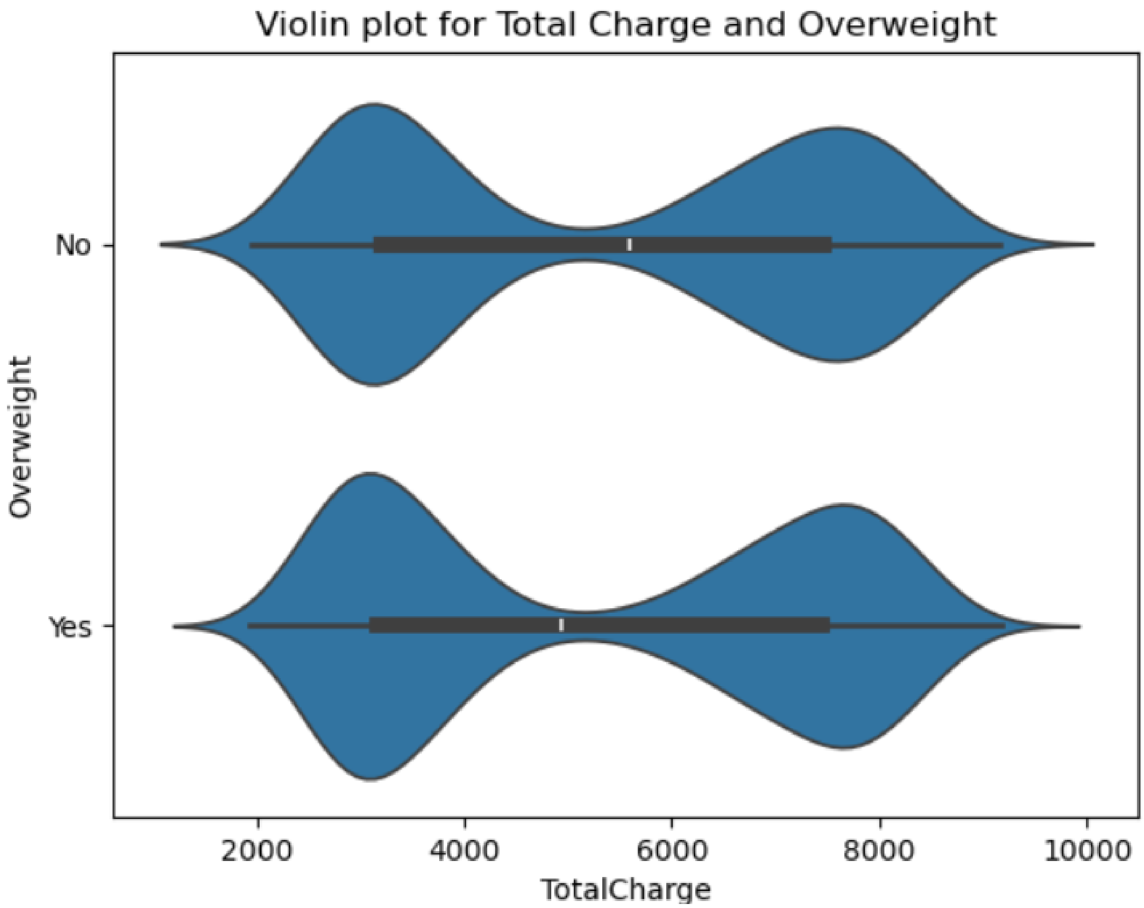

9000
8000
7000
6000
5000
4000
3000
2000
TotalCharge
5000
10000
15000
20000
25000
Additional charge


Violin plot for Total Charge and Gender
Gender
Male
Female
Nonbinary
2000
4000
6000
8000
10000
TotalCharge

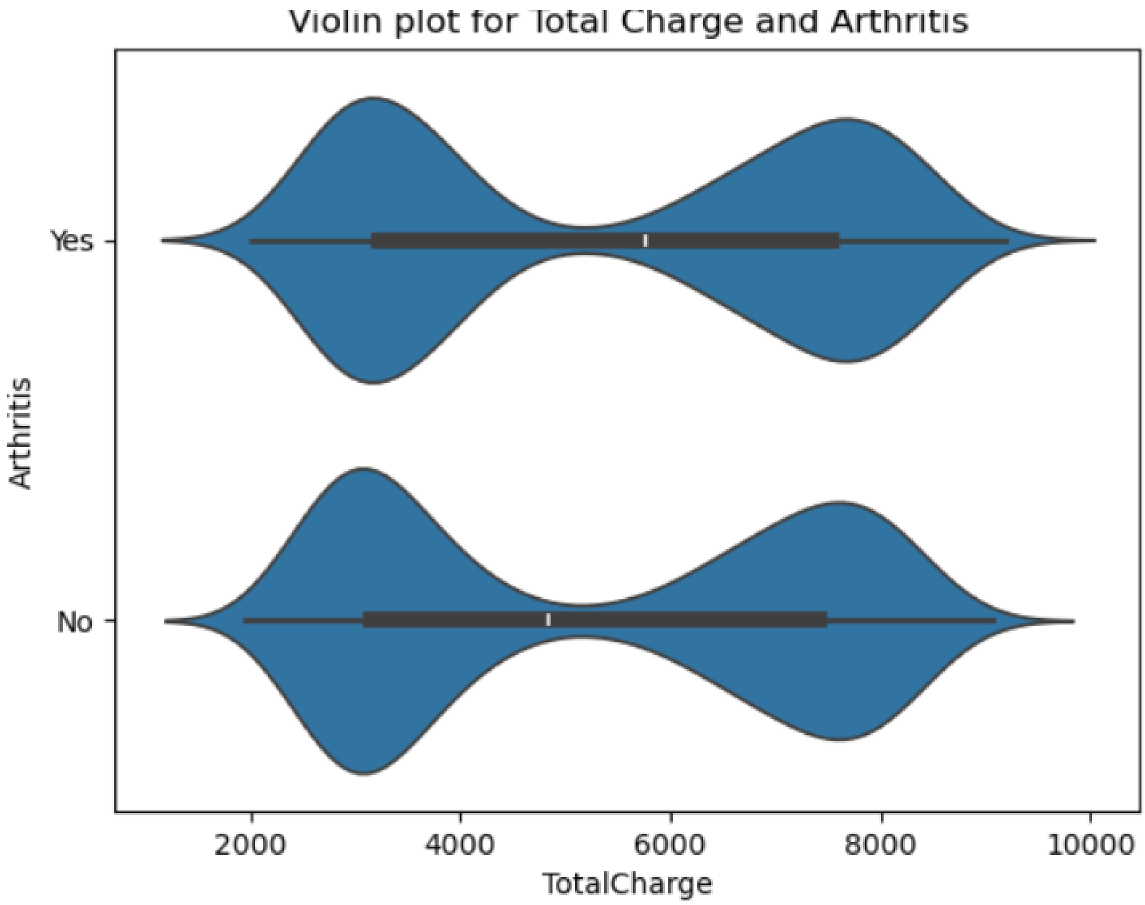

Violin plot for Total Charge and Initial_admin
Initial_admin
Emergency Admission
Elective Admission
Observation Admission
2000
4000
6000
8000
10000
TotalCharge

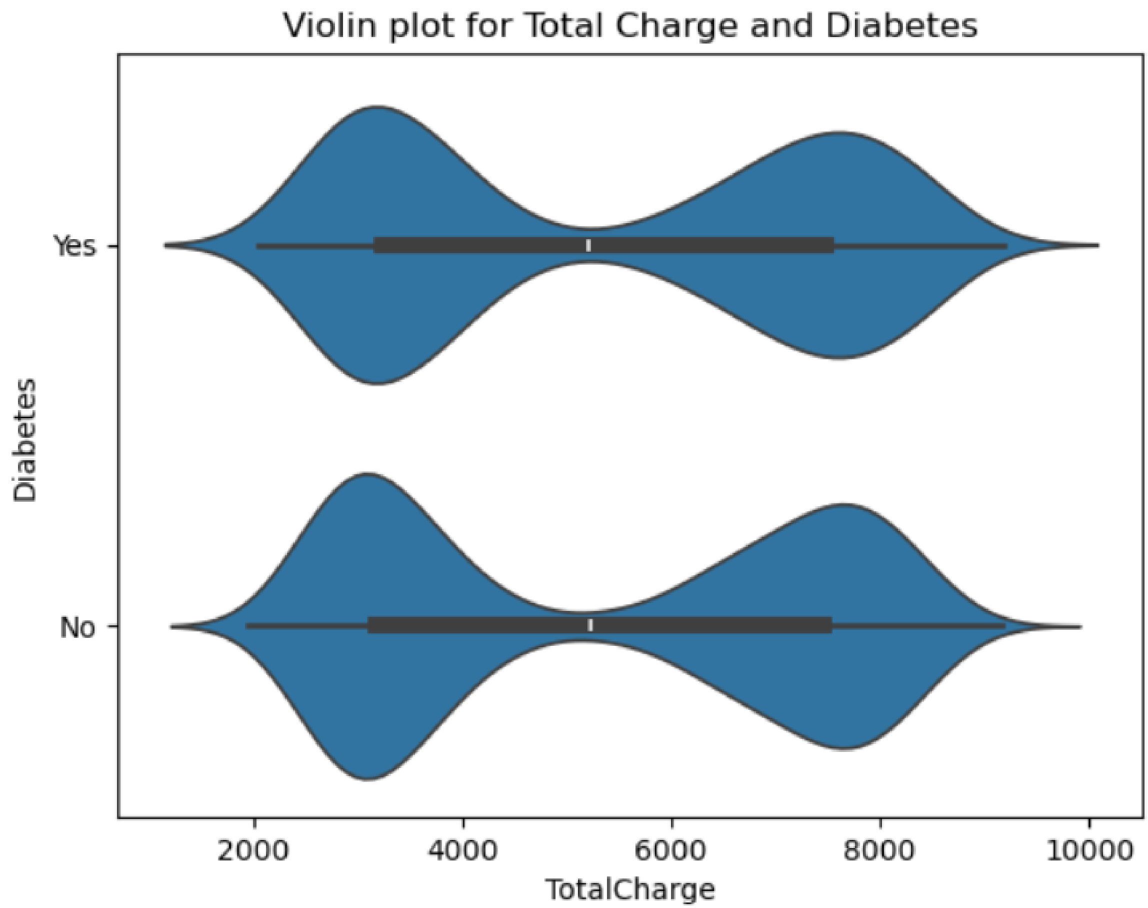

Violin plot for Total Charge and HighBlood
Yes
No
HighBlood
2000
4000
6000
8000
10000
TotalCharge


Violin plot for Total Charge and Stroke
No
Yes
Stroke
2000
4000
6000
8000
10000
TotalCharge

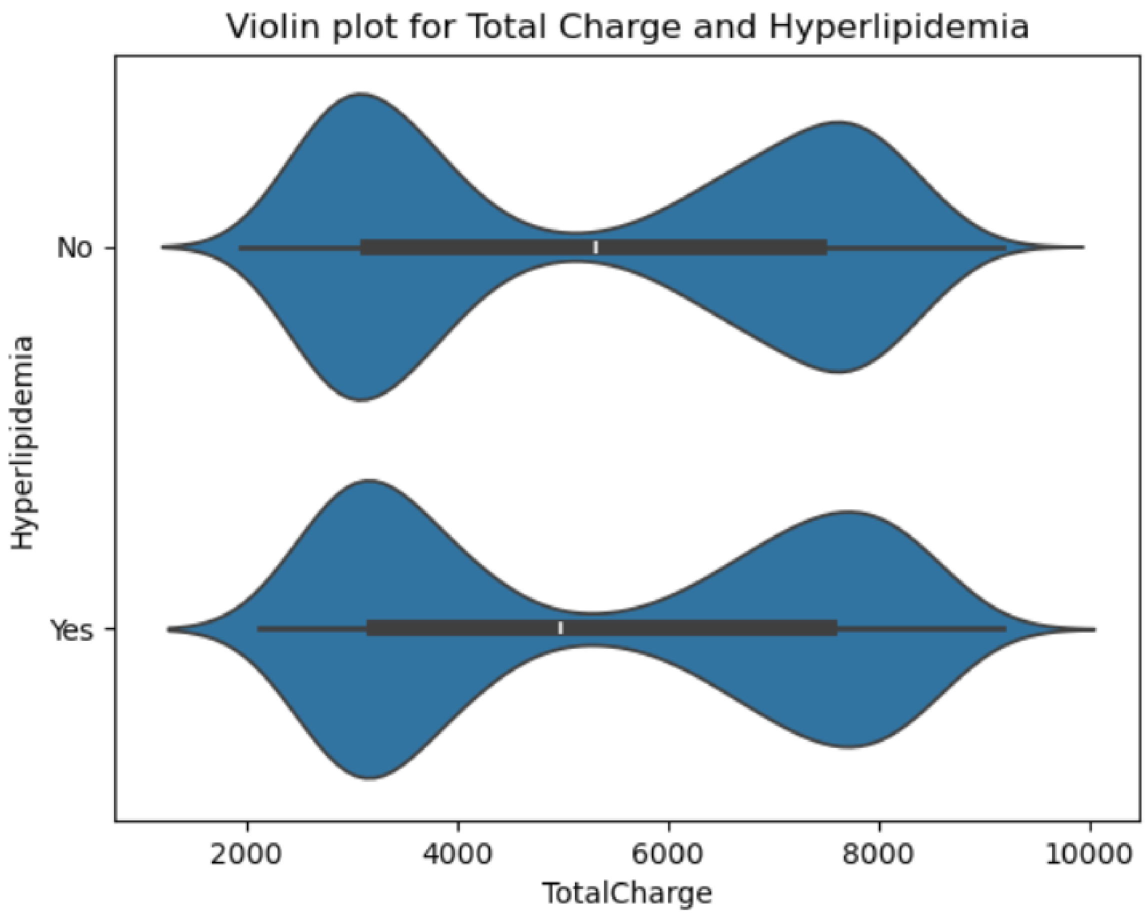


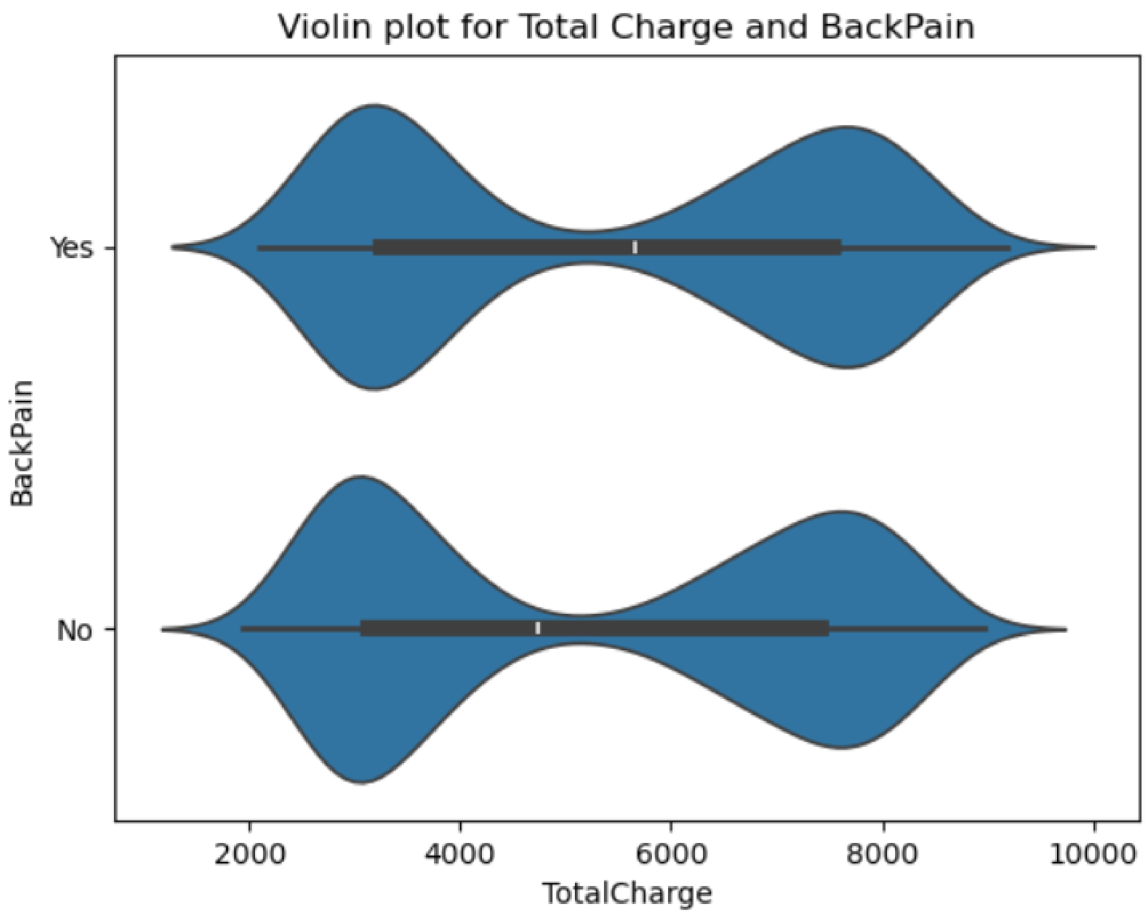
Violin plot for Total Charge and Overweight
No
Yes
Overweight
2000
4000
6000
8000
10000
TotalCharge

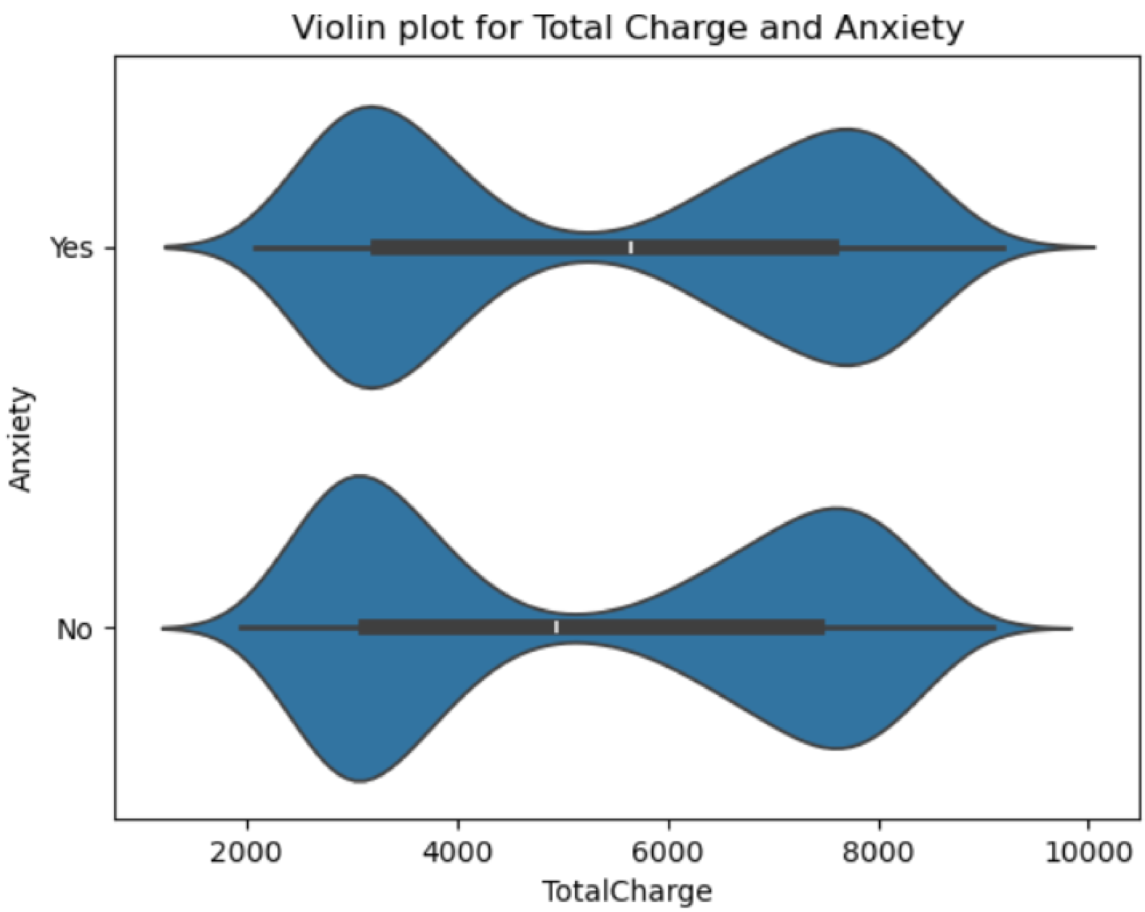

Violin plot for Total Charge and Arthritis
Yes
No
Arthritis
2000
4000
6000
8000
10000
TotalCharge



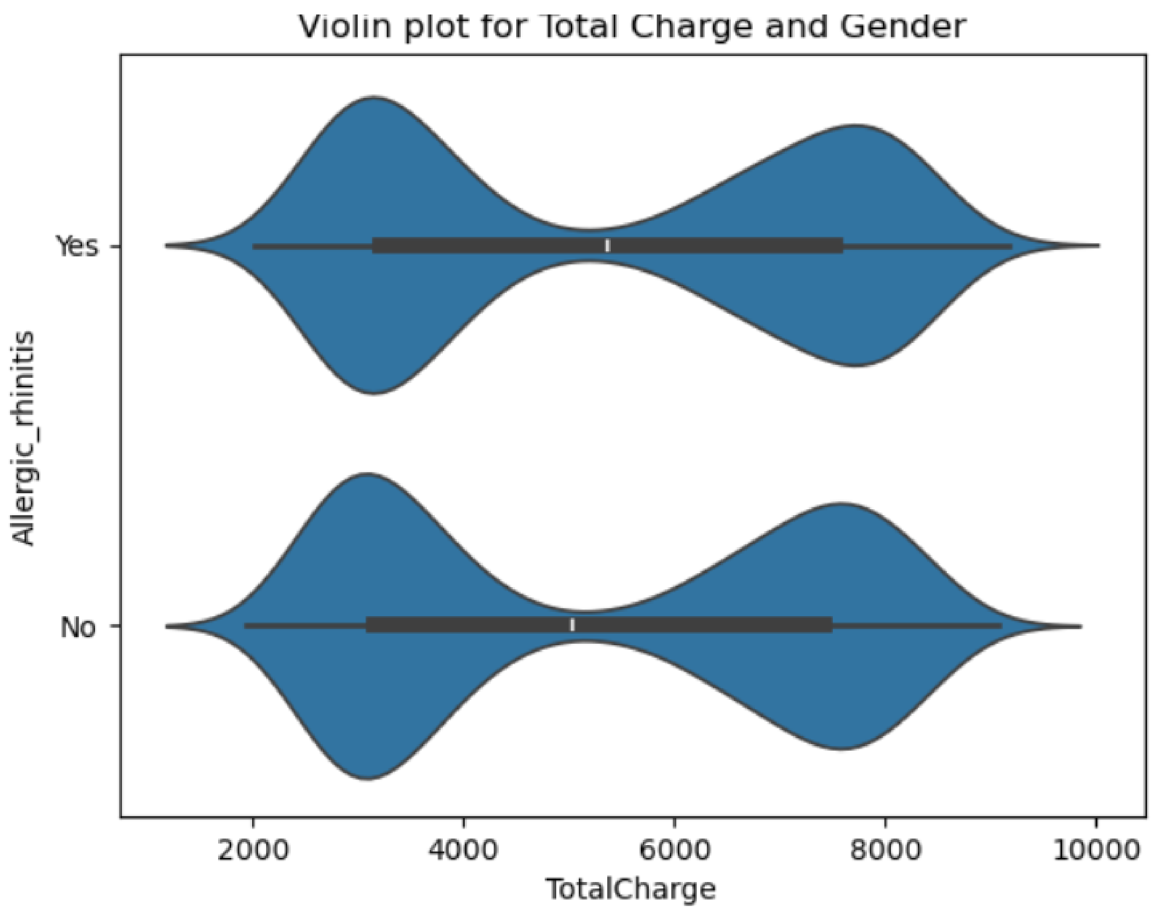
Violin plot for Total Charge and Diabetes
Yes
No
Diabetes
2000
4000
6000
8000
10000
TotalCharge


Violin plot for Total Charge and Hyperlipidemia
No
Yes
Hyperlipidemia
2000
4000
6000
8000
10000
TotalCharge

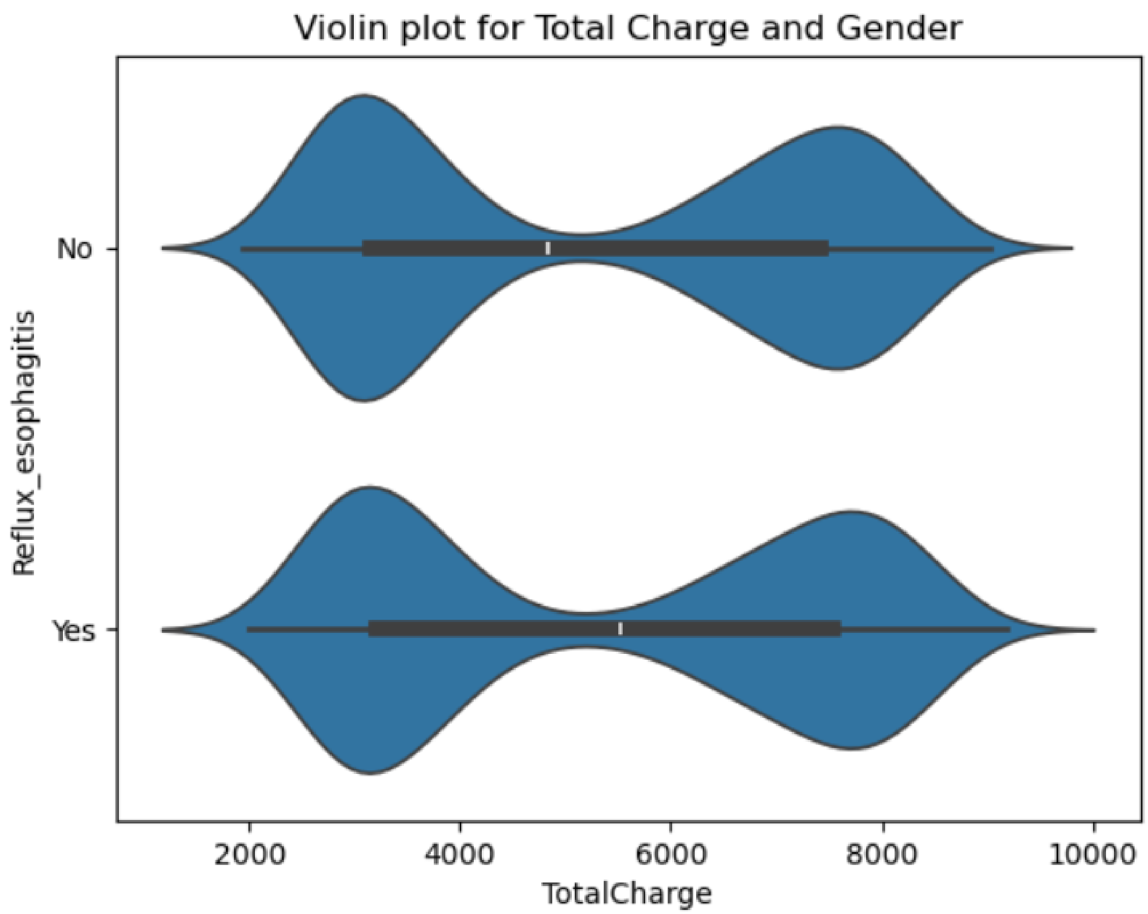

Violin plot for Total Charge and BackPain
Yes
No
BackPain
2000
4000
6000
8000
10000
TotalCharge



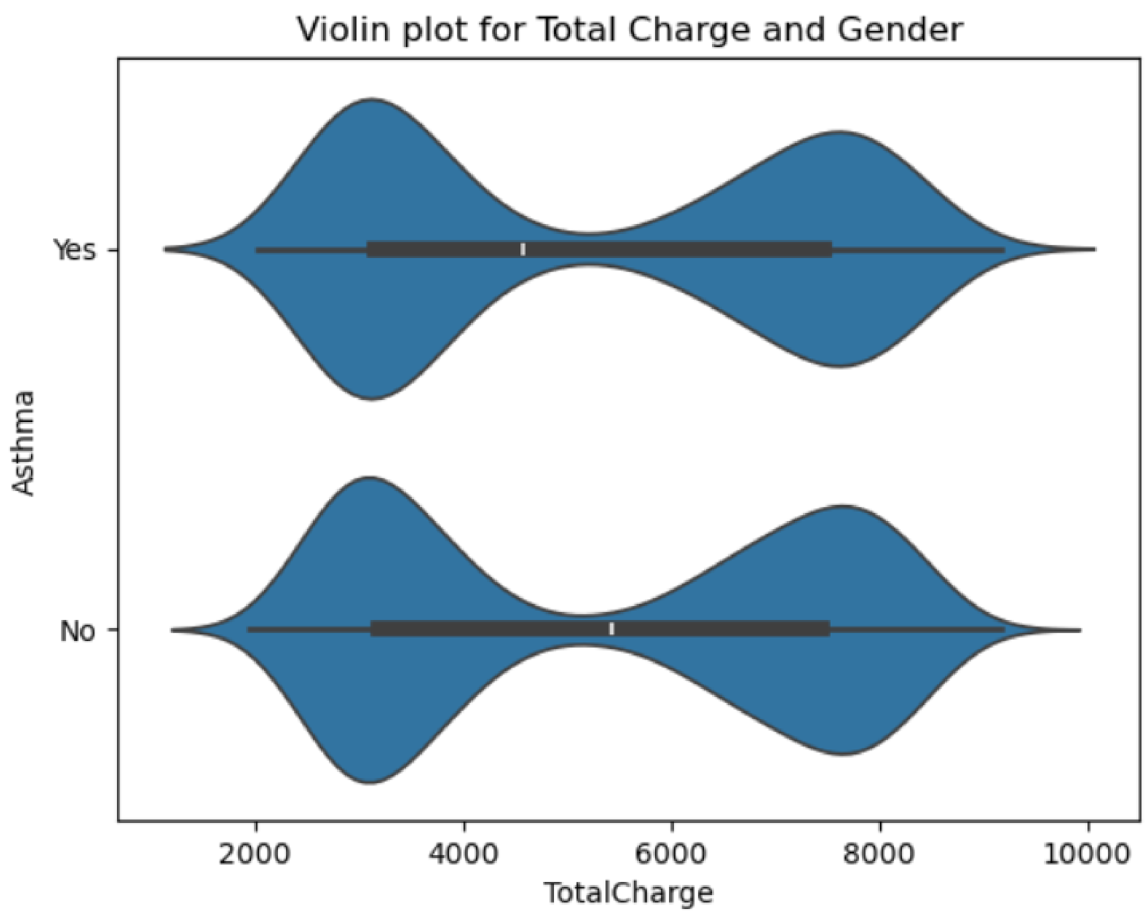
Violin plot for Total Charge and Anxiety
Yes
No
Anxiety
2000
4000
6000
8000
10000
TotalCharge



Violin plot for Total Charge and Gender
Yes
No
Allergic_rhinitis
2000
4000
6000
8000
10000
TotalCharge



Violin plot for Total Charge and Gender
No
Yes
Reflux_esophagitis
2000
4000
6000
8000
10000
TotalCharge



C4. Data Transformation:

The data transformation I performed for my MLR model is one-hot encoding to convert my categorical variables into dummy variables. Categorical variables cannot be used in a regression model as they are not numerical. By converting categorical variables into dummy variables, we create binary (0 or 1) indicators for each category. This allows the model to interpret and process categorical information. The function I used to convert my categorical variables is pd.get_dummies().

C5. Prepared dataset

The final dataset used for my data analysis is ‘medical_clean_1.csv’.

D1. Initial Model

```
                           Results: Ordinary least squares
=====================================================================================
Model:                      OLS                     Adj. R-squared:          0.016
Dependent Variable:         TotalCharge             AIC:                     181987.3834
Date:                       2024-10-11 14:48        BIC:                     182131.5902
No. Observations:           10000                   Log-Likelihood:          -90974.
Df Model:                   19                      F-statistic:             9.305
Df Residuals:               9980                    Prob (F-statistic):      2.48e-27
R-squared:                  0.017                   Scale:                   4.6803e+06
-------------------------------------------------------------------------------------
                                     Coef.   Std.Err.    t     P>|t|    [0.025   0.975]
-------------------------------------------------------------------------------------
const                              4844.2438 220.9818 21.9215 0.0000 4411.0748 5277.4128
Age                                   1.2740   3.2624  0.3905 0.6962   -5.1209    7.6689
Income                               -0.0011   0.0008 -1.2622 0.2069   -0.0027    0.0006
VitD_levels                          -3.7017  10.7408 -0.3446 0.7304  -24.7557   17.3524
Additional_charges                    0.0020   0.0139  0.1421 0.8870   -0.0253    0.0293
Gender_Male                          32.5539  43.8488  0.7424 0.4579  -53.3986  118.5063
Gender_Nonbinary                     87.6044 151.1106  0.5797 0.5621 -208.6028  383.8116
InitialAdmin_Emergency Admission    444.2909  53.2886  8.3374 0.0000  339.8344  548.7474
InitialAdmin_Observation Admission  -36.0924  61.6436 -0.5855 0.5582 -156.9263   84.7415
HB_Yes                               70.4653 126.4607  0.5572 0.5774 -177.4231  318.3538
stroke_Yes                          -10.4179  54.3973 -0.1915 0.8481 -117.0476   96.2118
OverWeight_Yes                      -56.7040  47.6993 -1.1888 0.2346 -150.2043   36.7962
Arthritis_Yes                       146.8464  45.1881  3.2497 0.0012   58.2686  235.4243
Diabetes_Yes                         61.4812  48.5653  1.2660 0.2056  -33.7165  156.6790
Hyperlipidemia_Yes                   73.5321  45.8037  1.6054 0.1084  -16.2524  163.3166
BackPain_Yes                        159.0585  44.0219  3.6132 0.0003   72.7666  245.3504
Anxiety_Yes                         138.8475  46.3505  2.9956 0.0027   47.9911  229.7039
AllergiRhinitis_Yes                  77.9117  44.3038  1.7586 0.0787   -8.9327  164.7561
RefluxEsophagitis_Yes               115.5541  43.9843  2.6272 0.0086   29.3361  201.7721
Asthma_Yes                          -69.6471  47.7458 -1.4587 0.1447 -163.2385   23.9442
-------------------------------------------------------------------------------------
Omnibus:                    42531.113             Durbin-Watson:              0.176
Prob(Omnibus):              0.000                 Jarque-Bera (JB):           1240.477
Skew:                       0.068                 Prob(JB):                   0.000
Kurtosis:                   1.280                 Condition No.:              503922
=====================================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 5.04e+05. This might indicate                that
there are strong multicollinearity or other numerical           problems.
```

Code can be found in D208PA_OviyaSelvaraj.ipynb.

D2. Justification of Model Reduction:

The model reduction technique I used for my performance assessment is the Stepwise Backward elimination method. In this technique, the model starts with all possible predictors and successfully removes non-significant predictors until the stopping criteria are reached. This method suits my model because it contains many potential indicators and reduces the number of variables to those that contribute the most to predicting total medical costs. The variable with the highest p-value is removed from the model, a new model fits, and this process is repeated. This method is also a simple and effective way to select a subset of variables for a linear regression model.

D3. Reduced Linear Regression model

```
#backward Elimination
def backward_elimination(X1, y1, significance_level=0.05):

    model=sm.OLS(y1, X1).fit()
    # Loop through the predictors and remove one at a time based on p-value
    while True:
        max_p_value = model.pvalues.max()  # Get the highest p-value
        if max_p_value > significance_level:
            excluded_variable = model.pvalues.idxmax()  # Identify the variable with the highest p-value
            print(f'Removing {excluded_variable} with p-value {max_p_value}')
            X1 = X1.drop(columns=[excluded_variable])

            # Fit the model again without the excluded variable
            model = sm.OLS(y1, X1).fit()
        else:
            break

    # Return the final model
    return model
```

```python
#Reduced model
final_model = backward_elimination(X, y)
print(final_model.summary2())
```

```
Removing Additional_charges with p-value 0.886969019264995
Removing stroke_Yes with p-value 0.857328303093629
Removing VitD_levels with p-value 0.7276585438463115
Removing Gender_Nonbinary with p-value 0.5638682707041195
Removing InitialAdmin_Observation Admission with p-value 0.5529754850332624
Removing Gender_Male with p-value 0.48705866845603696
Removing OverWeight_Yes with p-value 0.23255844159905126
Removing Income with p-value 0.21524939675738053
Removing Diabetes_Yes with p-value 0.19446016441967912
Removing Asthma_Yes with p-value 0.14408596482766411
Removing Age with p-value 0.10069614240327944
Removing Hyperlipidemia_Yes with p-value 0.0961594313015767
Removing AllergiRhinitis_Yes with p-value 0.07478559993025692
Removing HB_Yes with p-value 0.05017541004227858
                            Results: Ordinary least squares
=======================================================================================
Model:                     OLS                    Adj. R-squared:          0.015
Dependent Variable:        TotalCharge            AIC:                     181979.9984
Date:                      2024-10-11 15:04       BIC:                     182023.2604
No. Observations:          10000                  Log-Likelihood:          -90984.
Df Model:                  5                      F-statistic:             31.22
Df Residuals:              9994                   Prob (F-statistic):      1.20e-31
R-squared:                 0.015                  Scale:                   4.6833e+06
---------------------------------------------------------------------------------------
                                  Coef.    Std.Err.    t      P>|t|    [0.025    0.975]
---------------------------------------------------------------------------------------
const                           4866.7894  45.4732 107.0256 0.0000 4777.6528 4955.9259
InitialAdmin_Emergency Admission 465.4431  43.2868  10.7525 0.0000  380.5922  550.2940
Arthritis_Yes                    149.8178  45.1742   3.3164 0.0009   61.2673  238.3684
BackPain_Yes                     158.1985  43.9937   3.5959 0.0003   71.9620  244.4351
Anxiety_Yes                      138.8286  46.3440   2.9956 0.0027   47.9850  229.6722
RefluxEsophagitis_Yes            112.7159  43.9568   2.5642 0.0104   26.5518  198.8801
---------------------------------------------------------------------------------------
Omnibus:                   42605.480             Durbin-Watson:               0.177
Prob(Omnibus):             0.000                 Jarque-Bera (JB):            1238.269
Skew:                      0.068                 Prob(JB):                    0.000
Kurtosis:                  1.282                 Condition No.:               4
=======================================================================================
```

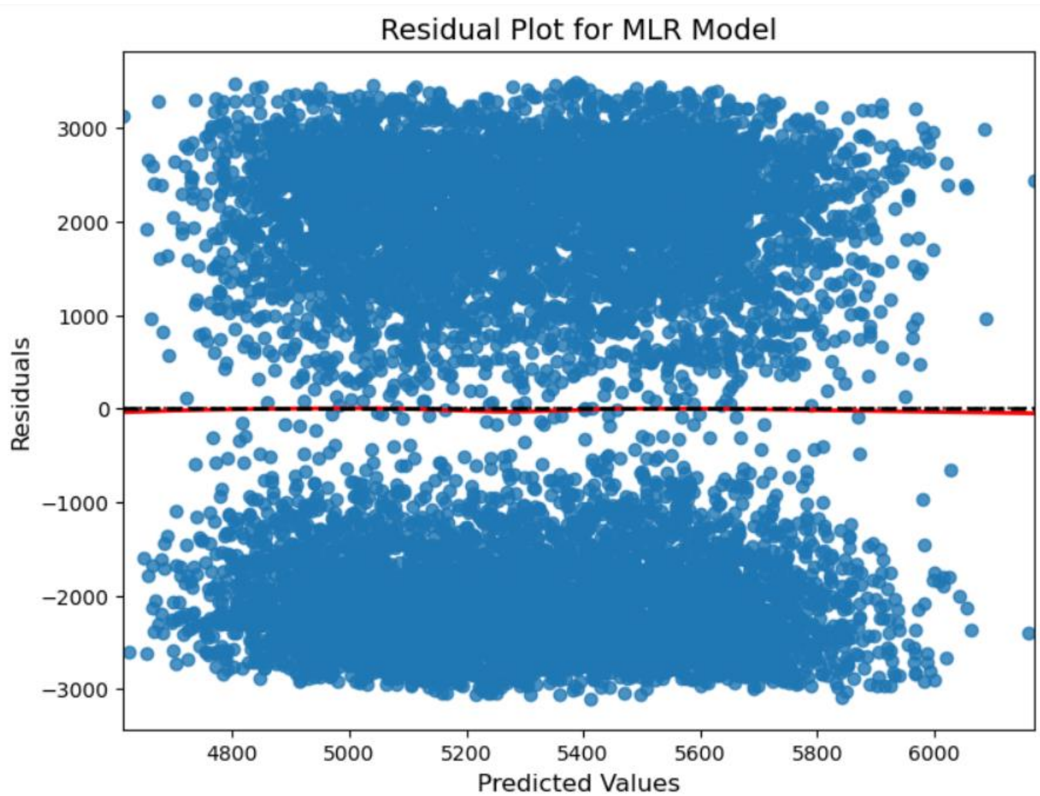
E1. Model Comparison

The R squared for the initial model is 0.016 and the reduced model is 0.015, which is almost identical indicating similar levels of fit to the data even though the models themselves might be different. The reduced model is not a good fit for the data because it fails to capture the underlying relationships between the independent and dependent variable. It can also be that the dependent variable has high variability making it difficult to explain using independent variables. Similarly, the initial models selected variables are not strong predictors or most relevant for the dependent variable.

E2. Output and Calculations

The visualization and calculations of the requirements below can be found in 'D208PA_OviyaSelvaraj.ipynb'

- a residual plot



- the model's residual standard error

```
RSE_summary = np.sqrt(final_model.mse_resid)
print(f'Residual Standard Error (from summary): {RSE_summary:.4f}')
```

```
Residual Standard Error (from summary): 2164.1035
```

E3. Code

Executable code can be found in ‘D208PA_OviyaSelvaraj.ipynb’

F1. Results

Regression equation for the reduced model:

TotalCharge = 4866.7894 +149.8178 * Arthritis_Yes + 158.1985 * BackPain + 138.8286* Anxiety + 112.7159 * Relux_esophagitis_Yes + 465.4431 * Initial_admin_Emergency Admission.

- ★ The intercept 4866.7894 represents the predicted value of TotalCharge when all independent variables are zero.
- ★ The coefficients for each independent variable represent the change in TotalCharge associated with a one-unit increase in that variable, holding all other variables constant.

Interpretation of the coefficients of the reduced model:

- Intercept is 4866.7894 which represents the predicted TotalCharge when all the independent variables are zero. It can be interpreted as the baseline for TotalCharge in the absence of other included factors.
- Arthritis_Yes: We predict that on average patients with Arthritis have an average TotalCharge that is 149.8178 units higher holding other factors constant.
- BackPain_Yes: We predict that on average patients with BackPain have an average TotalCharge that is 158.1985 units higher holding other factors constant.
- Anxiety_Yes: We predict that on average patients with Anxiety have an average TotalCharge that is 138.8286 units higher holding other factors constant.
- Reflux_esophagitis_Yes: We predict that on average patients with Reflux esophagitis have an average TotalCharge of 112.7159 units higher.

- Initial_admin_Emergency Admission: Patients admitted through the emergency room have an average TotalCharge that is 465.4431 units higher compared to those admitted through other means.

Statistical and practical significance of the reduced model:
The reduced model is statistically significant because it shows a 1.5% variance of TotalCharge explained by the model's independent variables. However, this is a low R squared value, meaning the model does not explain much of the variation in the total medical charges. Based on the R squared, the model has little practical significance because the factors only suggest a small portion of the Total Charge variance. It also describes a limited practical predictive power in its current form.

Limitations of the data analysis:
In this model technique, I have considered several categorical variables, which have been converted to dummy variables. MLR is sensitive to how variables are encoded. Too many categories will increase the complexity of the model without improving the predictive ability, leading to overfitting. The model is also sensitive to outliers, and its sensitivity to such data points can lead to skewed results. Thus, it is essential to mitigate these outliers. The low R squared of 1.5% means the model did not capture more of what drives the total charges, which means the MLR can struggle with complex non-linear relationships in the data.

F2. Recommendations:

A low R squared from the analysis above suggests that many factors influencing the outcome are missing. For example, specific medical procedures might not have been captured. Thus, the business team can collaborate with healthcare providers or data teams to gather more granular data that reflects patient behaviors, treatments, or other relevant medical factors.

G. Panopto:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=aa5cc1ef-6b78-442e-8b03-b2080017a07f#

H. References:

In-text citation

Linear regression is a useful statistical method we can use to understand the relationship between two variables, x and y. However, before we conduct linear regression, we must first make sure that four assumptions are met:

Reference entry

Zach Bobbit (January, 2020) The Four Assumptions of Linear Regression https://www.statology.org/linear-regression-assumptions/

In-text citation

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of MLR is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.

Reference entry

Adam Hayes (July, 2024) Multiple Linear Regression (MLR) Definition, Formula, and Example https://www.investopedia.com/terms/m/mlr.asp

JMP (n.d.). Multiple Regression Residuals Analysis and outliers https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-multiple-regression/mlr-residual-analysis-and-outliers.html

Medium (2018, September). Create Multiple Categorical data columns to numerical data columns using Dummy variables.
https://medium.com/@urvashilluniya/convert-multiple-categorical-columns-into-numeric-columns-in-single-line-of-code-577bab825635