D208: Predictive Modelling

Task II

Western Governors University

-Oviya Selvaraj

A1. Research question:

How do lifestyle factors (such as soft drink consumption, and vitamin D levels) and medical conditions (such as overweight, high blood pressure, diabetes, arthritis, and anxiety) predict the likelihood of a patient experiencing a stroke?

A2. Goals:

The goal of the analysis is to quantify the impact of lifestyle factors (such as soft drink consumption, and vitamin D levels) and medical conditions (such as overweight, high blood pressure, diabetes, arthritis, and anxiety) impact stroke on a patient. By developing a logistic regression model, we can better understand the risk factors associated with stroke and potentially guide patient treatment.

B1. Summary of assumptions:

The four assumptions of a Logistic Regression model are:

1. The dependent variable must be binary.
2. In Logistic regression the observations require to be independent of each other.
3. There should be little or no multicollinearity among the independent variables.
4. The independent variables assume linearity and log odds of the dependent variables.

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

Logistic regression is used when the target variable is dichotomous, meaning it has two possible values. According to my research question, I want to predict the factors leading to a patient suffering from a stroke which contains binary variables Yes and No. Thus, logistic regression is an appropriate technique.

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 Total Charge. I decided to round up the decimal values up to 2 decimal places for Income, Vitamin D levels, Total charge and Additional charges. Also, there are no missing values or duplicates for any of the data used in my analysis.

C2. Summary Statistics

Continuous data:

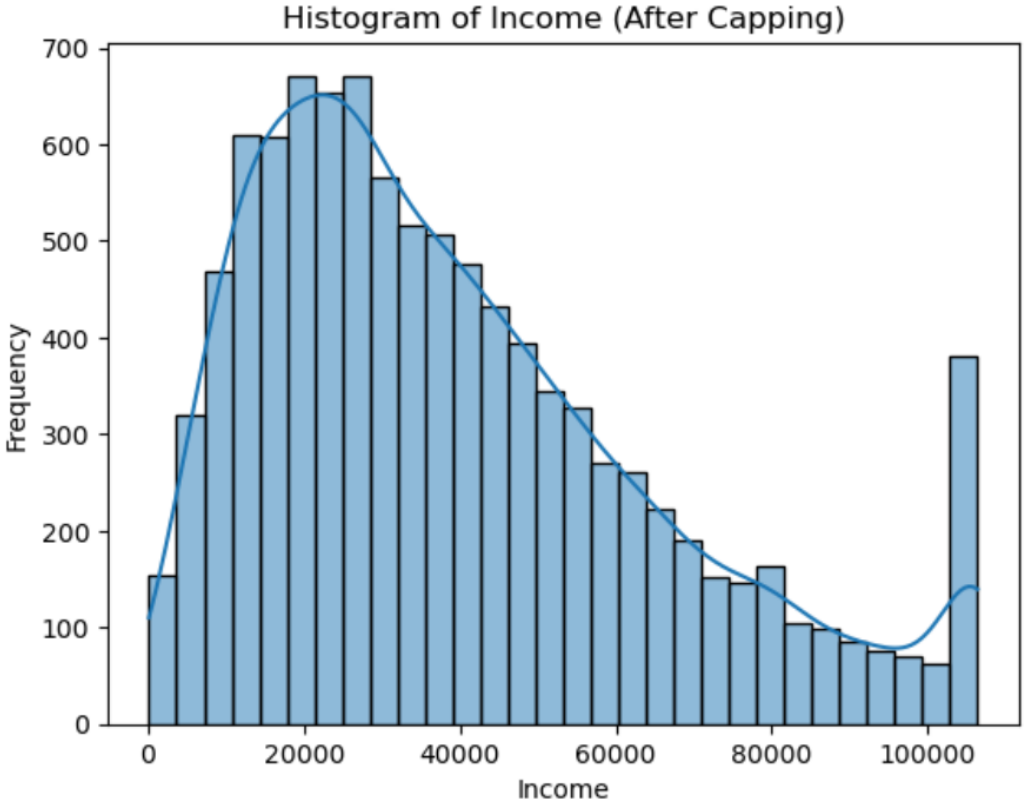
- **Income**: The histogram concludes a Uniform Distribution curve. The mean or average income of patients is $ 39721.119.
- **VitD_levels (Vitamin D Levels)**: The histogram shows a uniform distribution curve. The minimum value is 9.8 and the maximum value of 26.40 seems legitimate.
- **TotalCharge**: 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).

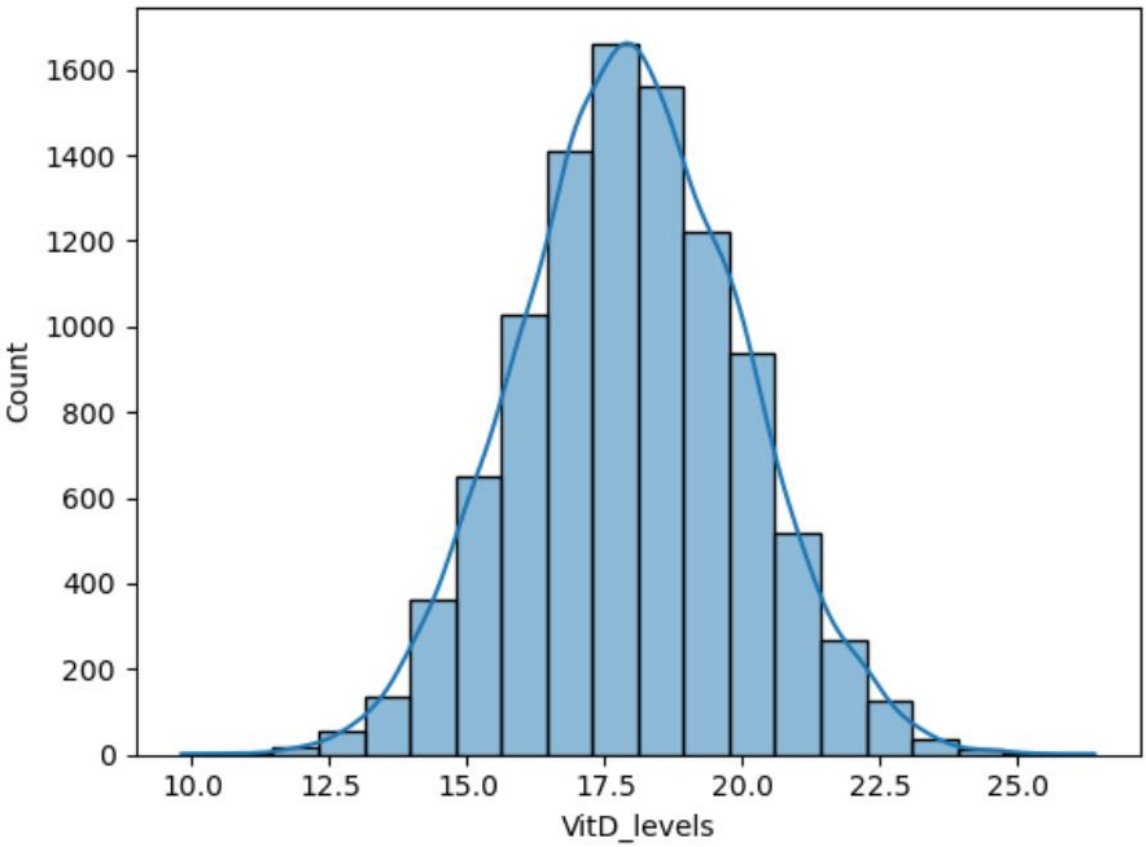
Categorical data:

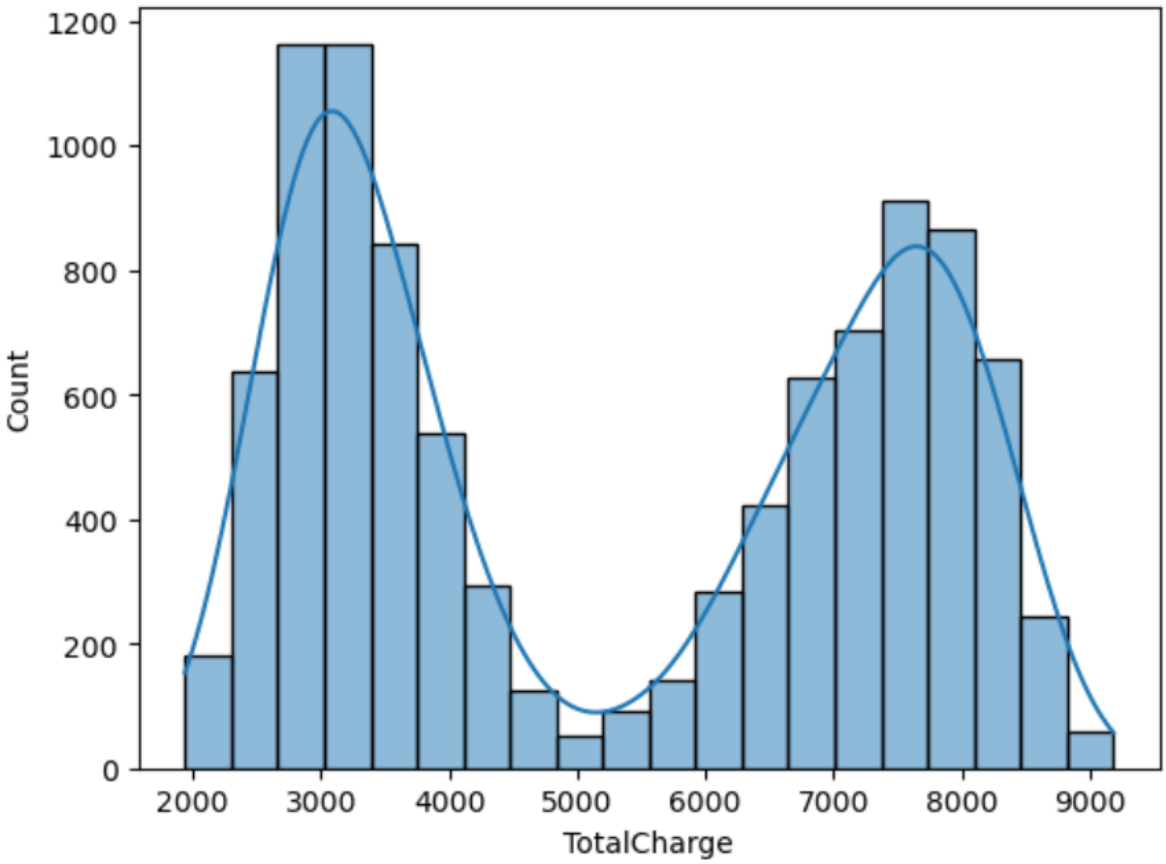
All the categorical data show Bernoulli distribution due to the binary nature. For Gender there are three categorical values (Male, Female and NonBinary). The frequency table is used to show the count of the categorical values.

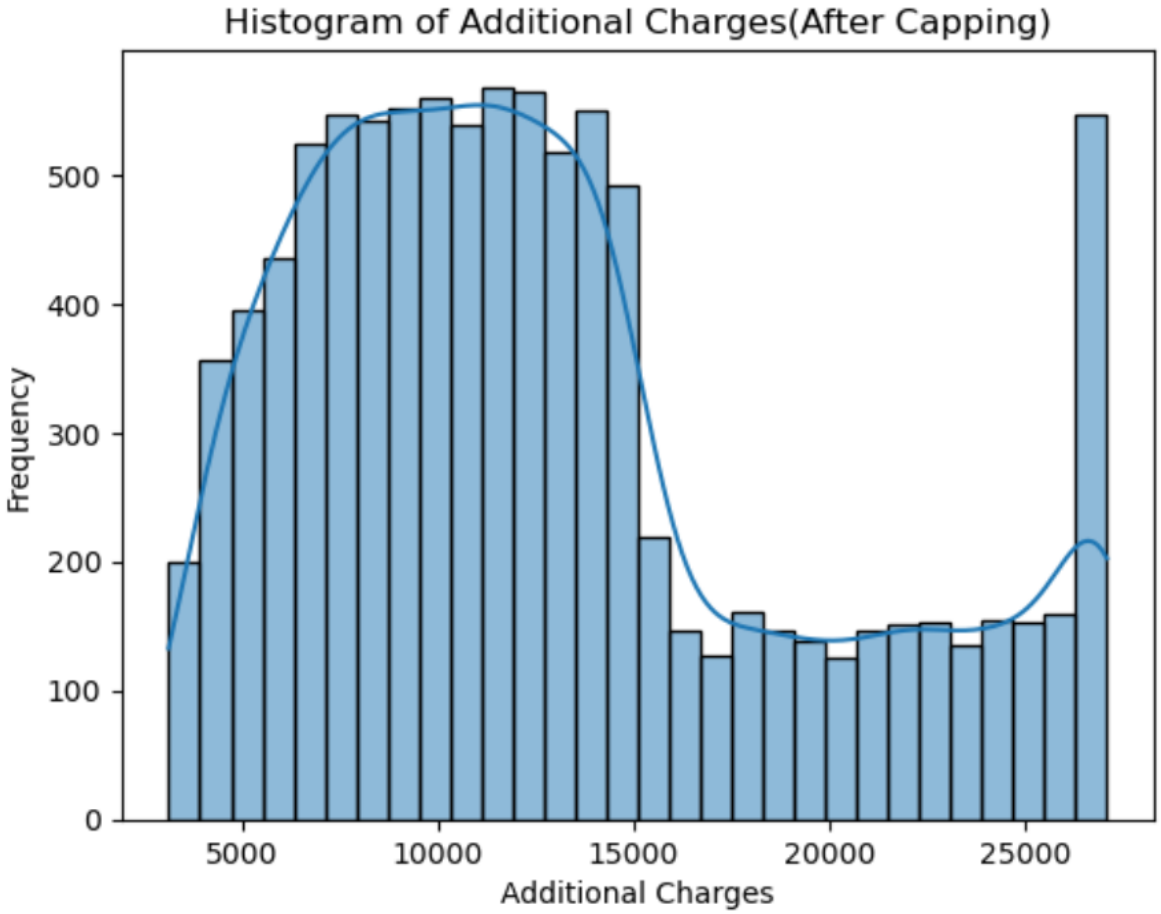
C3. Univariate and Bivariate statistics

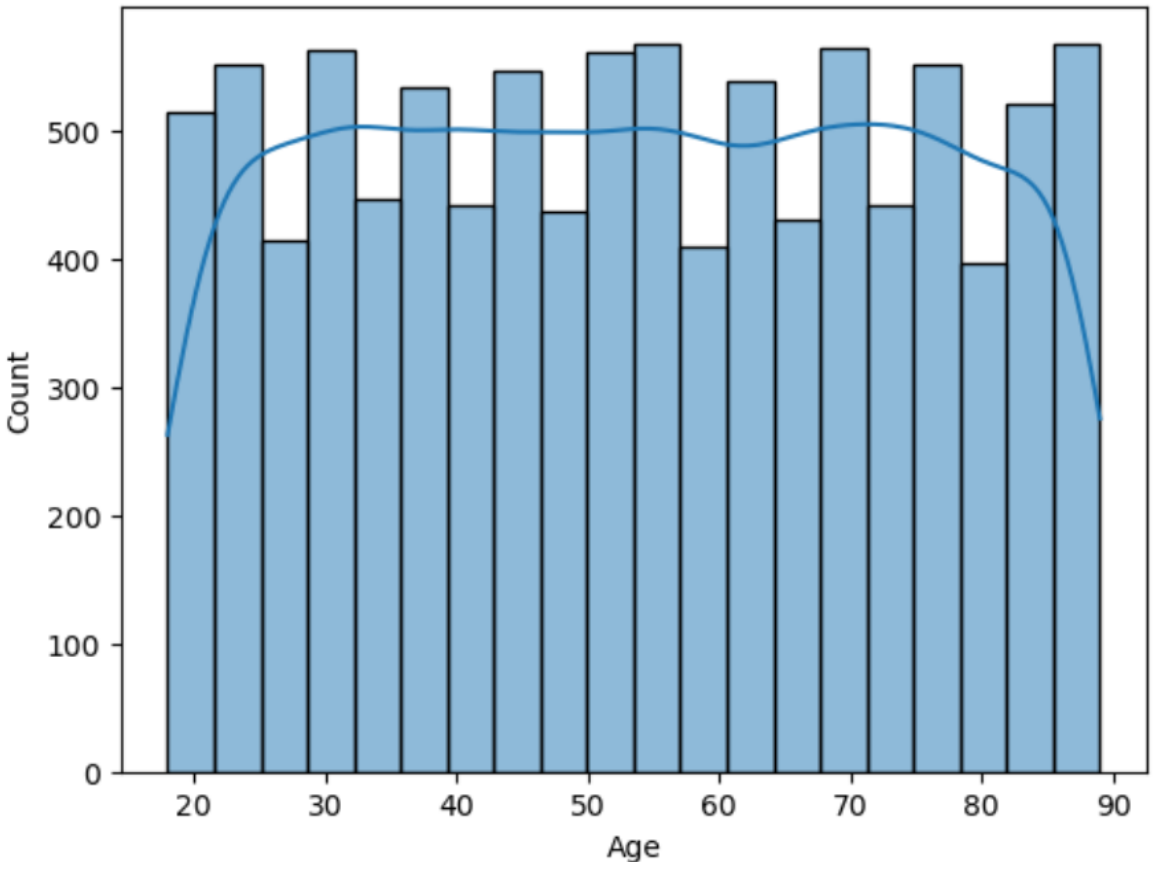
Univariate statistics for continuous and categorical data:




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

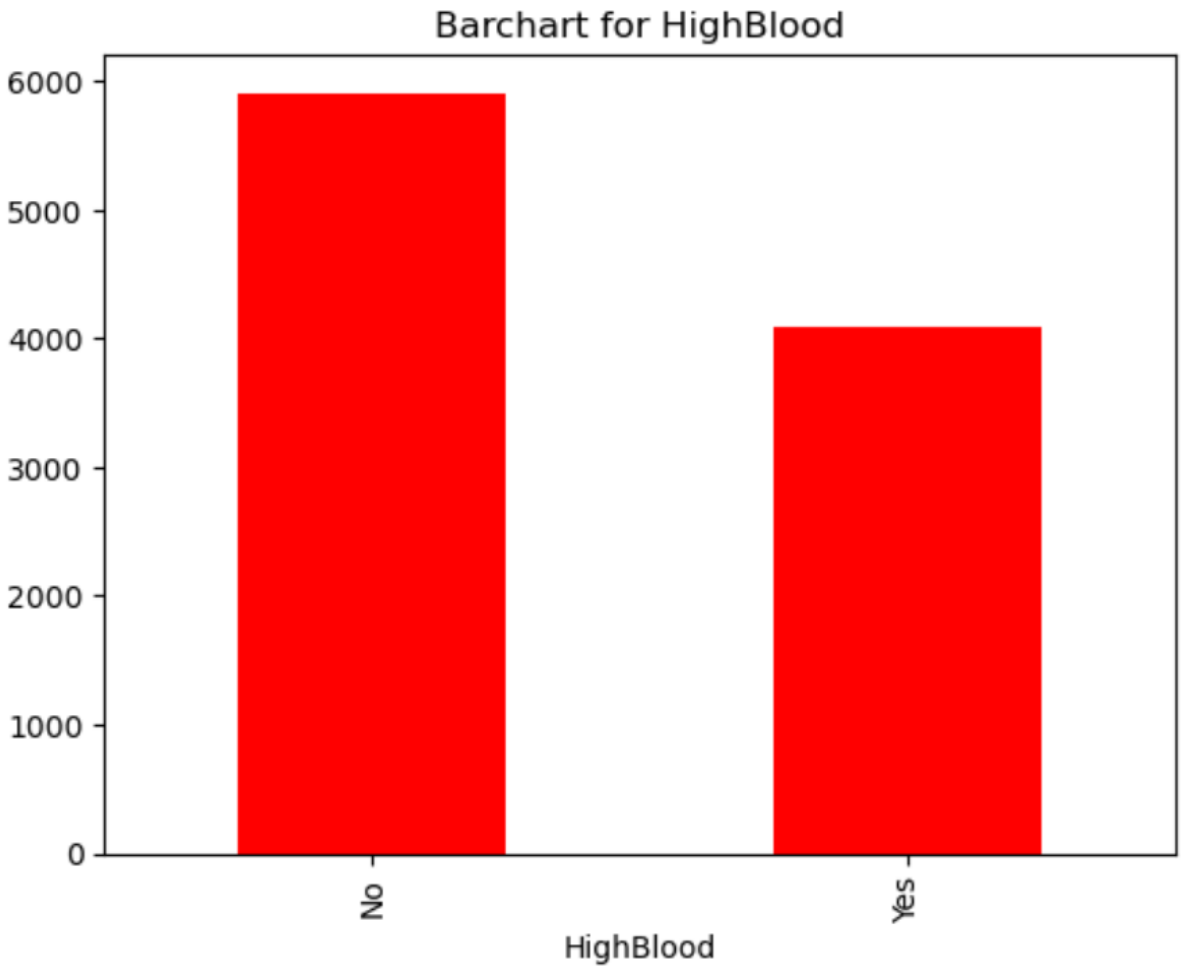

1200
1000
800
600
400
200
0
Count
2000
3000
4000
5000
6000
7000
8000
9000
TotalCharge

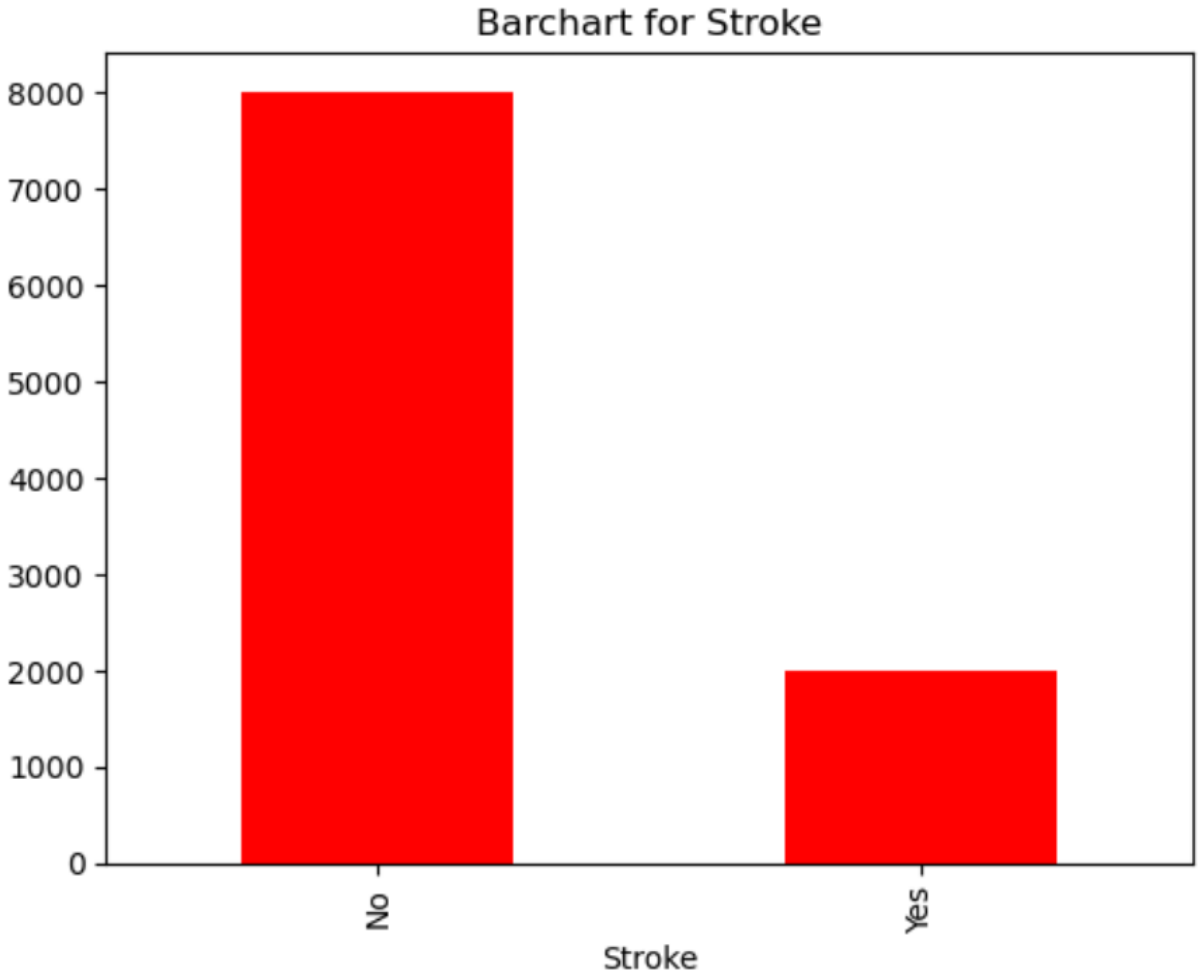

Histogram of Additional Charges(After Capping)
Frequency
0
100
200
300
400
500
5000
10000
15000
20000
25000
Additional Charges

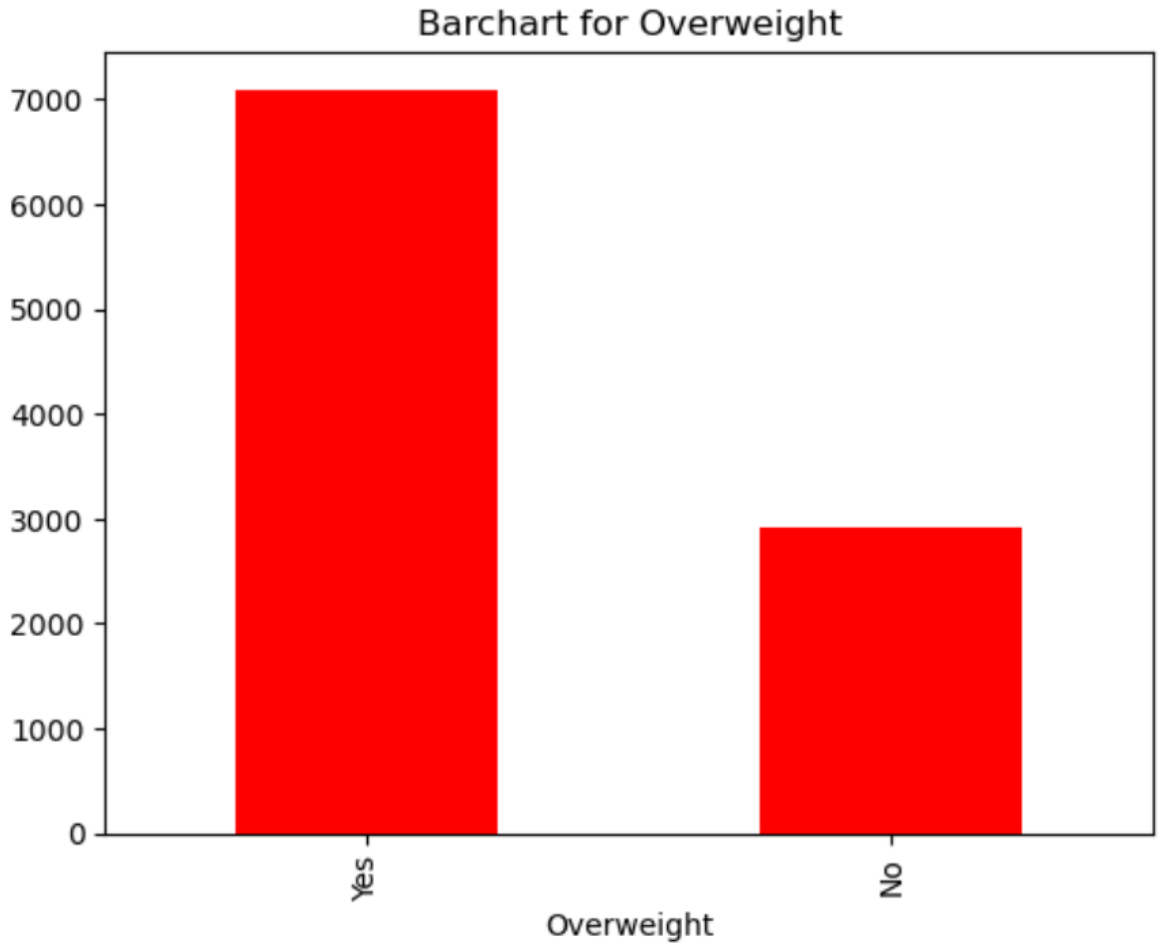


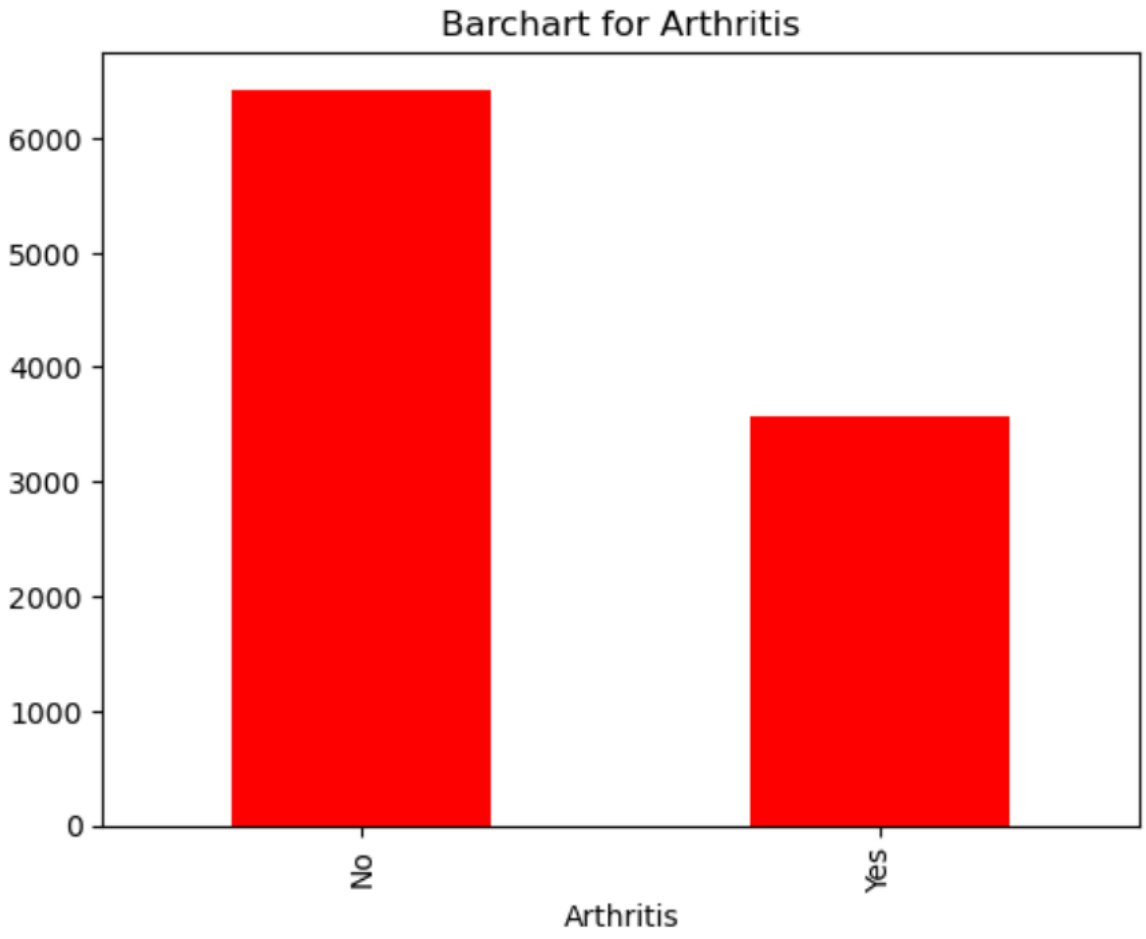
:

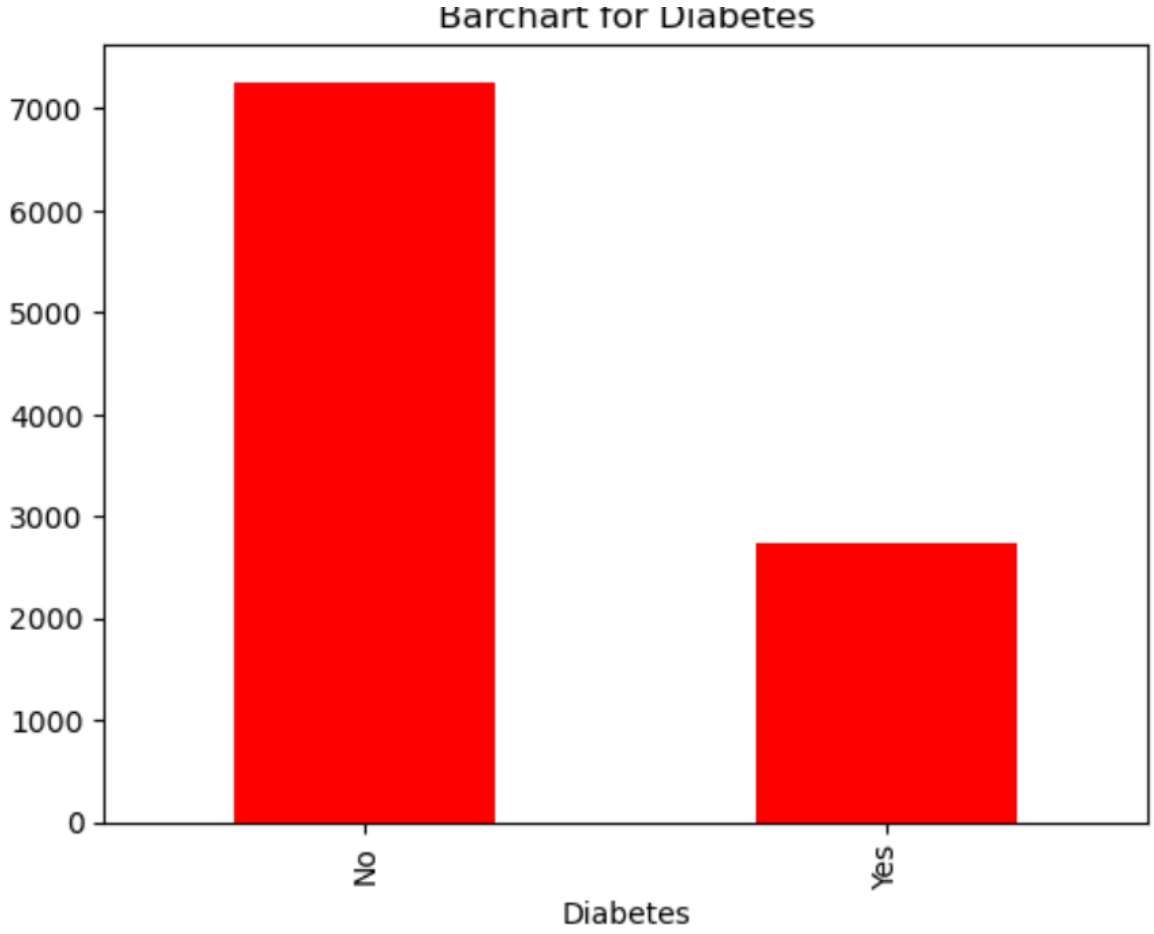
col_0	count
Gender	
Female	5018
Male	4768
Nonbinary	214

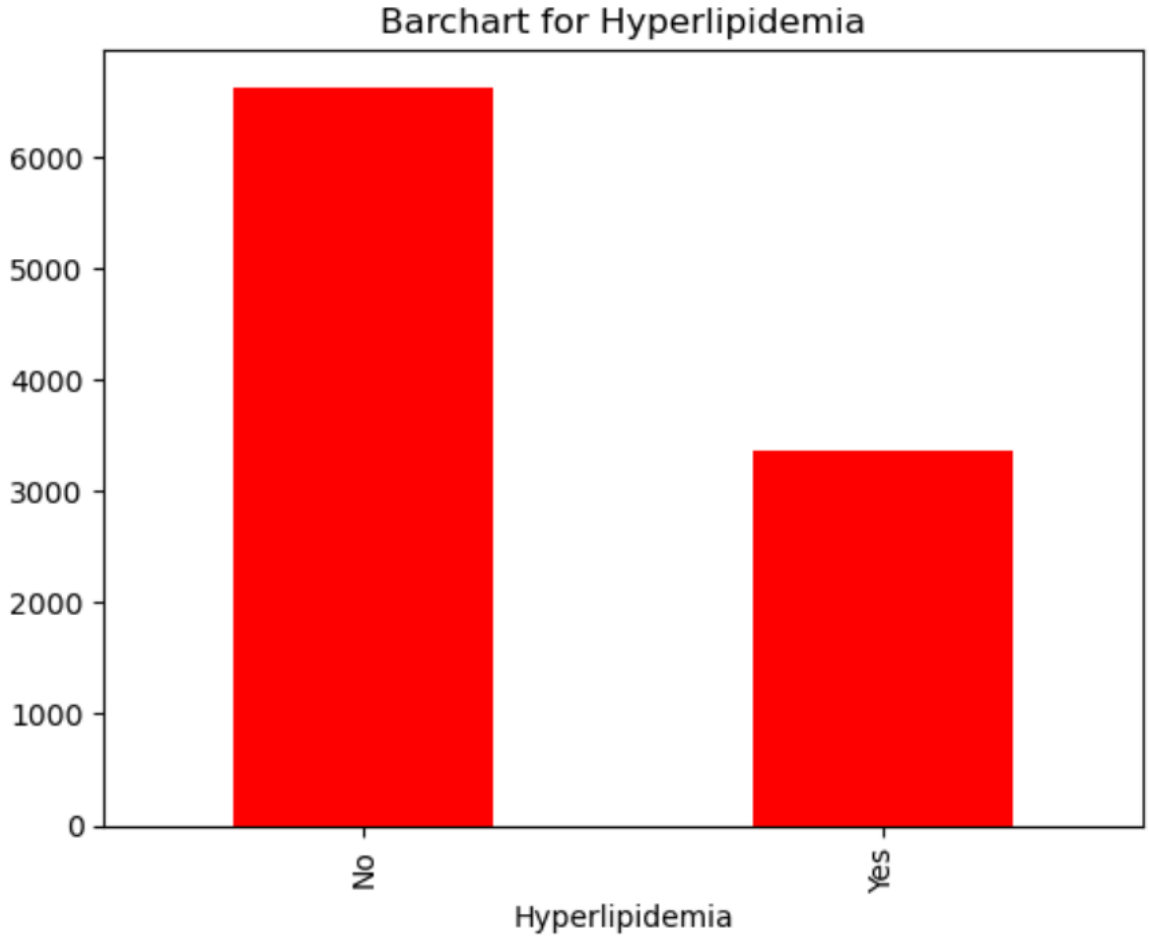

Barchart for HighBlood
6000
5000
4000
3000
2000
1000
0
No
Yes
HighBlood

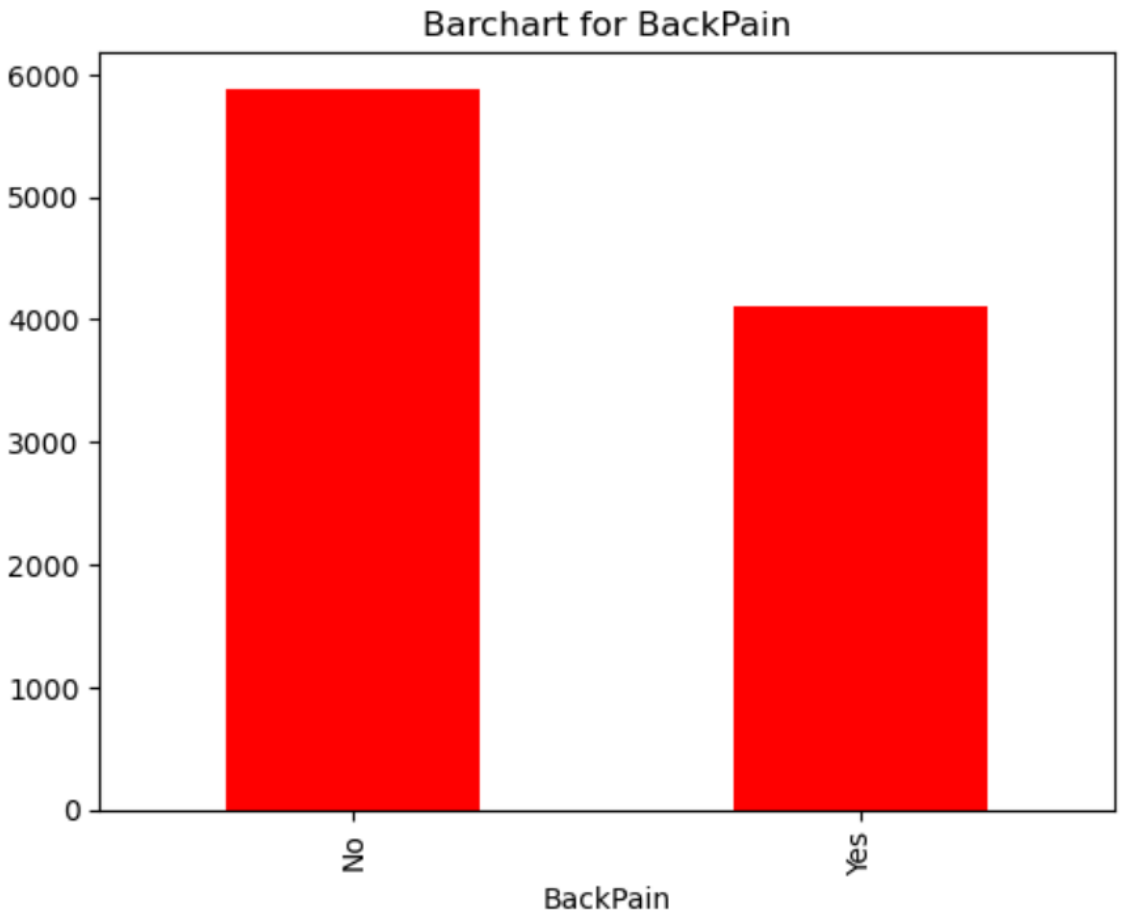

Barchart for Stroke
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Stroke

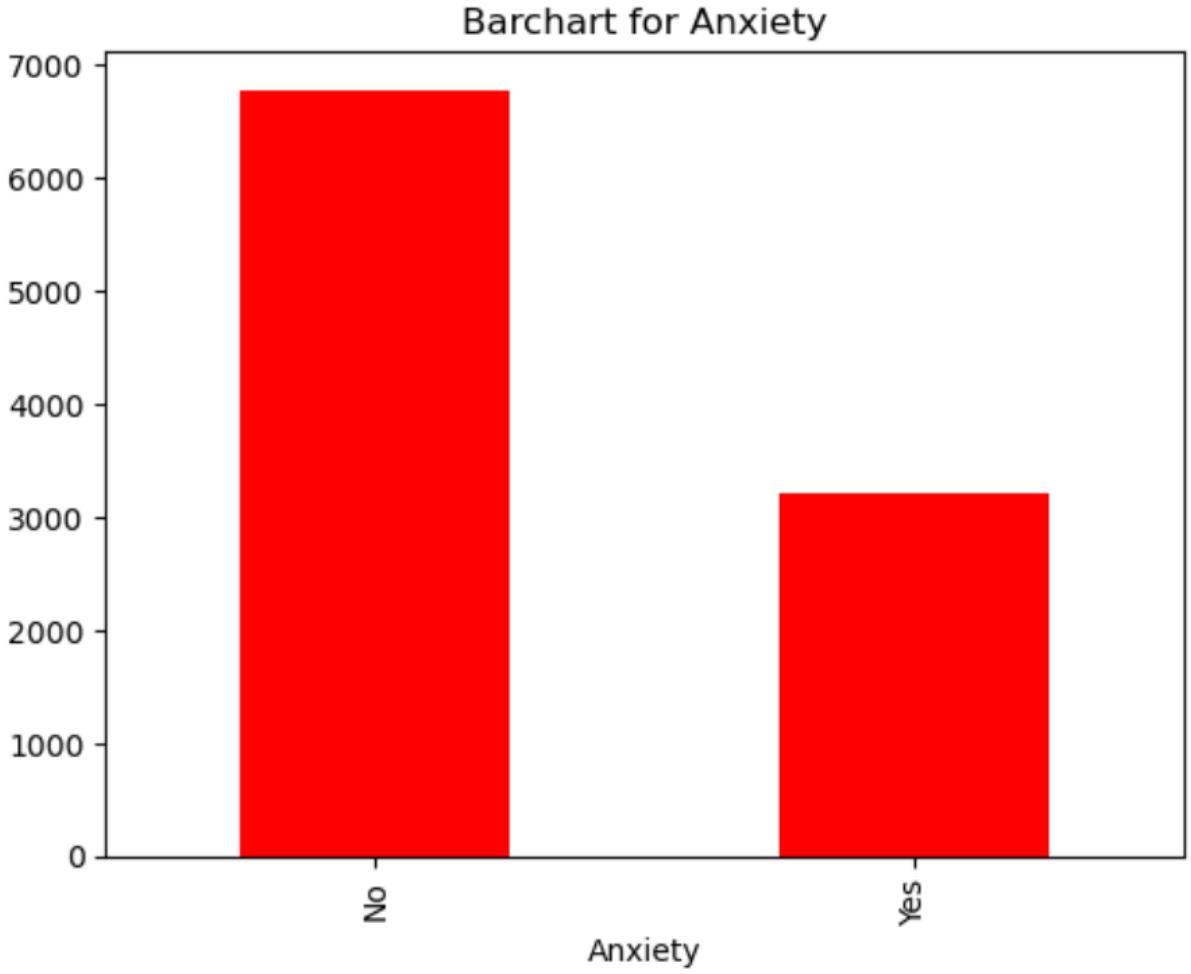

Barchart for Overweight
7000
6000
5000
4000
3000
2000
1000
0
Yes
No
Overweight

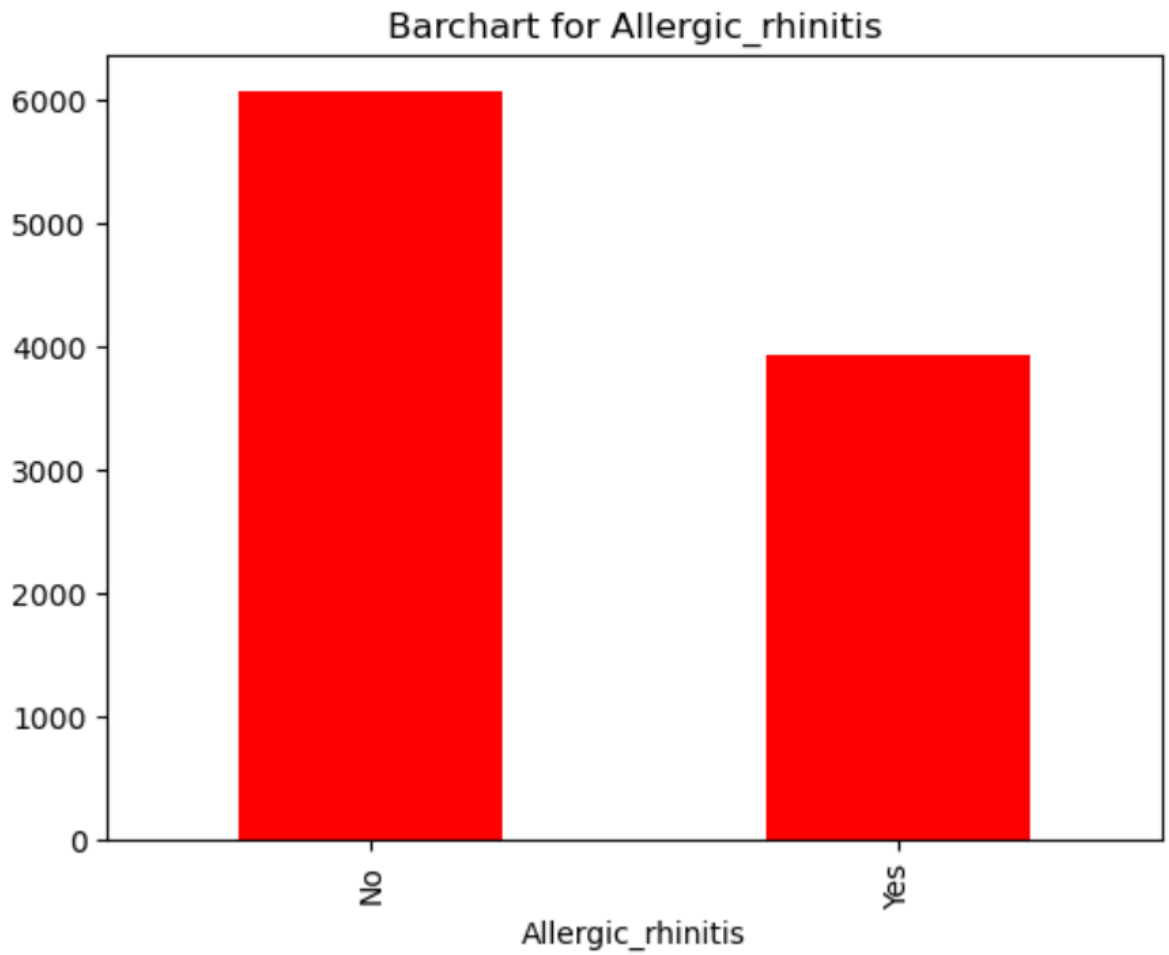

Barchart for Arthritis
6000
5000
4000
3000
2000
1000
0
No
Yes
Arthritis

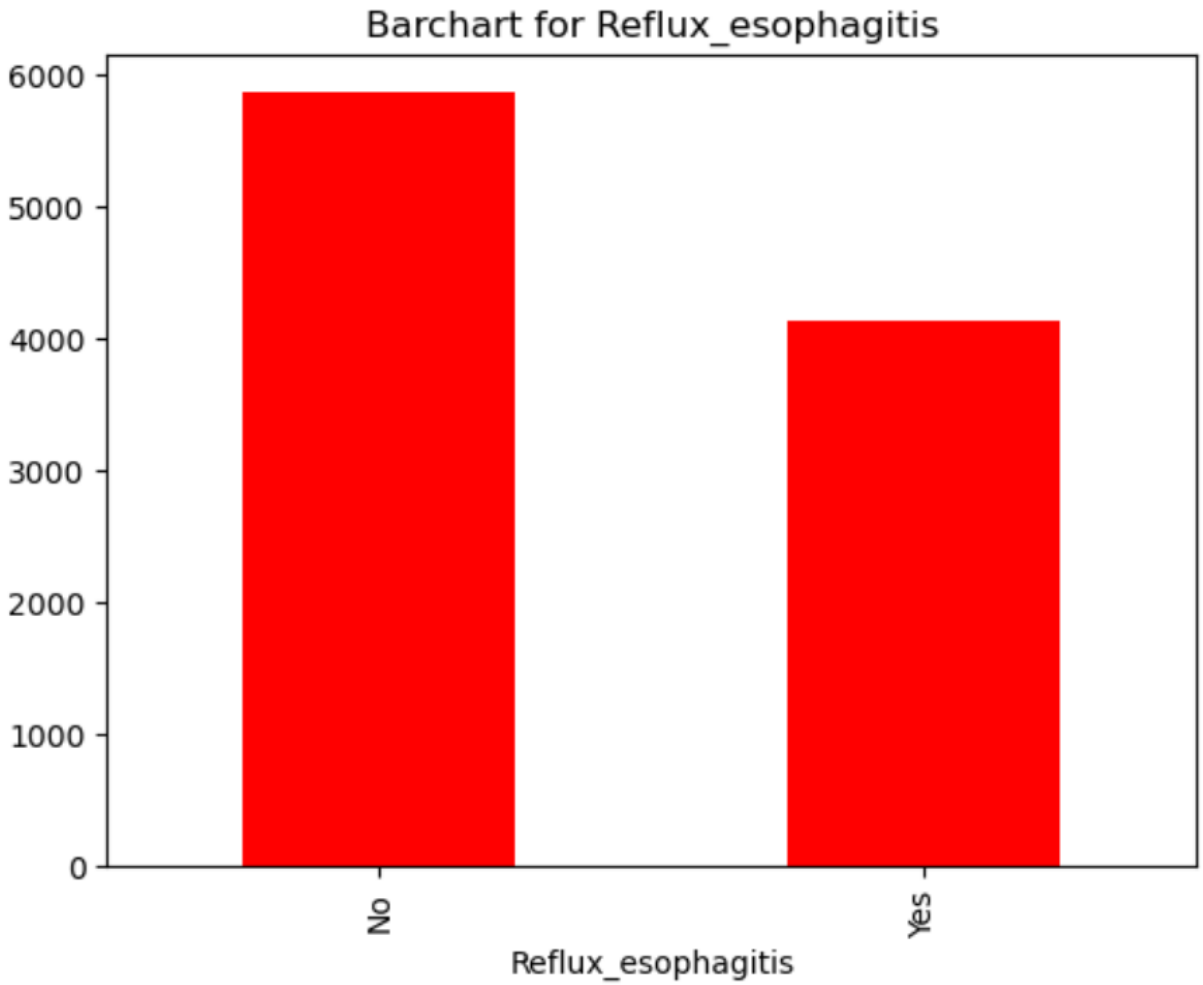

Barchart for Diabetes
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Diabetes

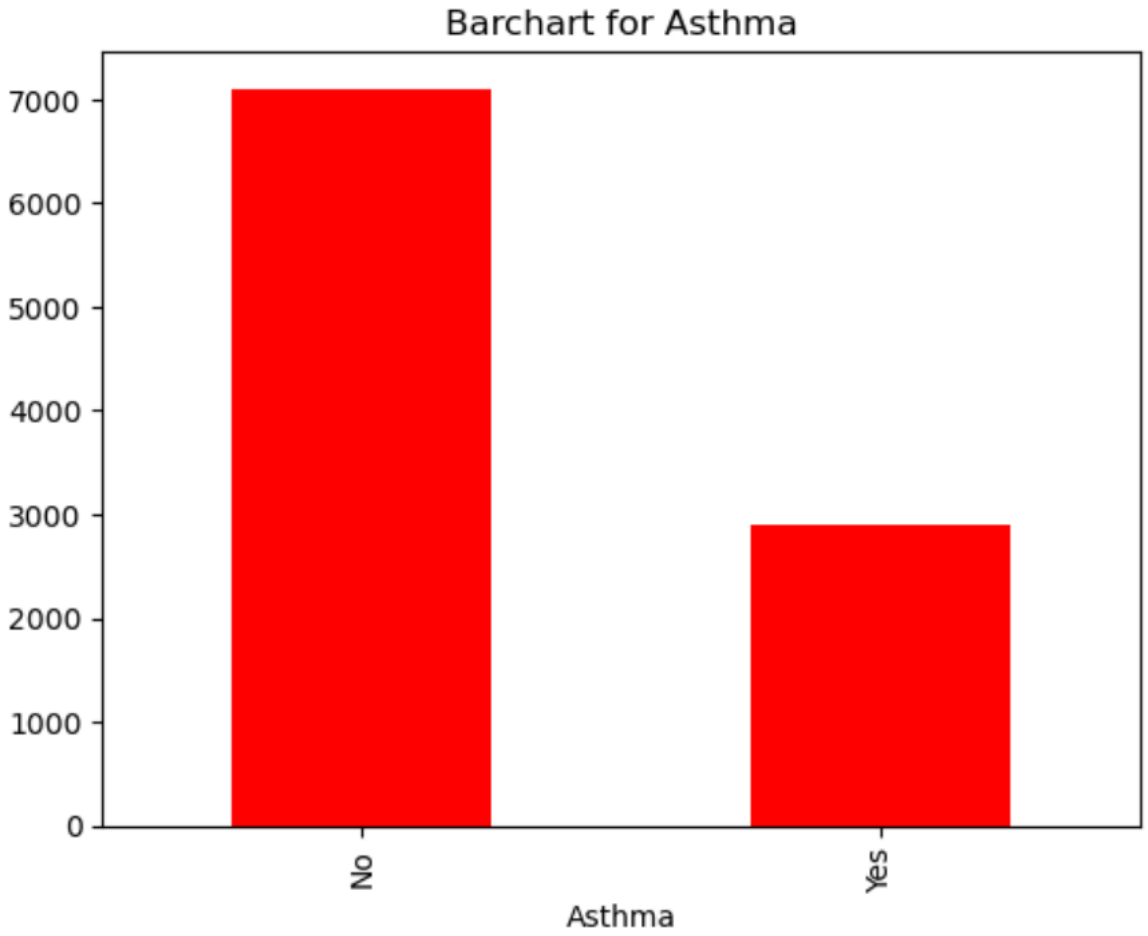

Barchart for Hyperlipidemia
6000
5000
4000
3000
2000
1000
0
No
Yes
Hyperlipidemia


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

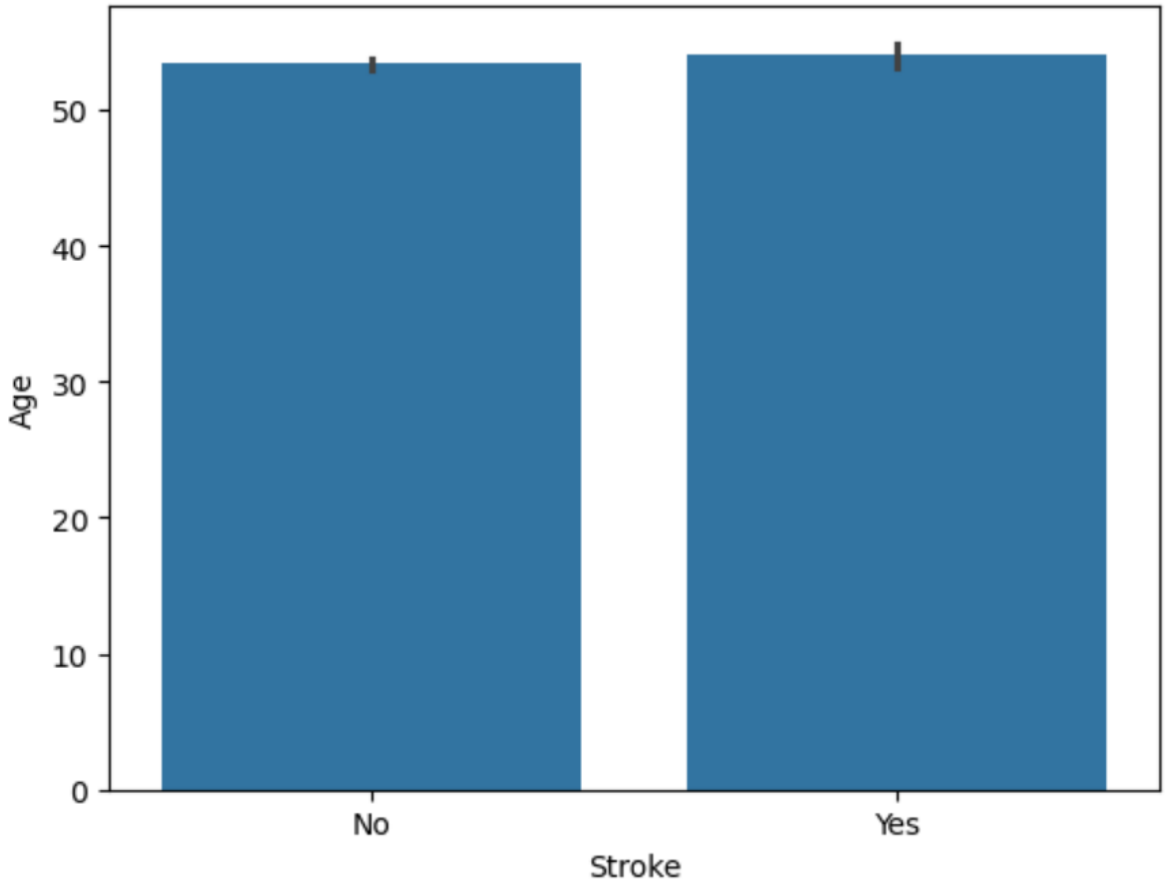

Barchart for Anxiety
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Anxiety

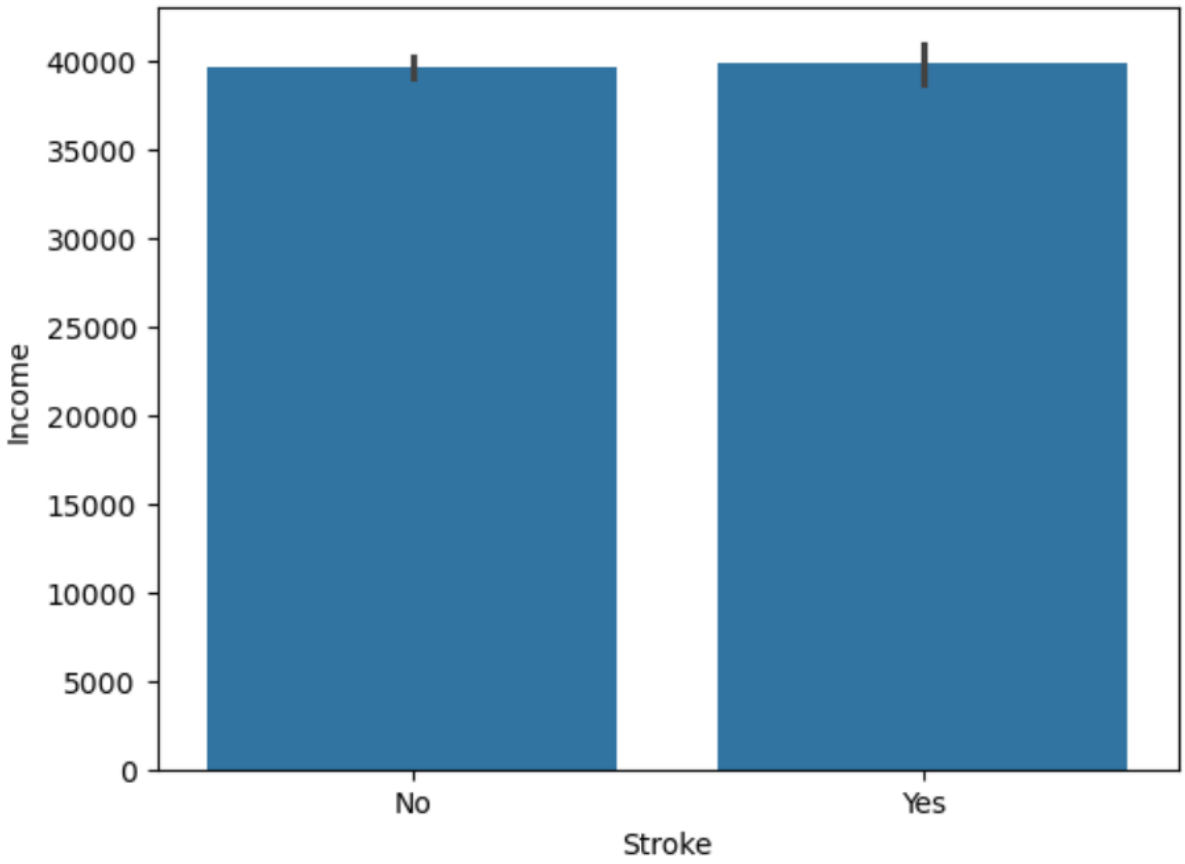

Barchart for Allergic_rhinitis
6000
5000
4000
3000
2000
1000
0
No
Yes
Allergic_rhinitis

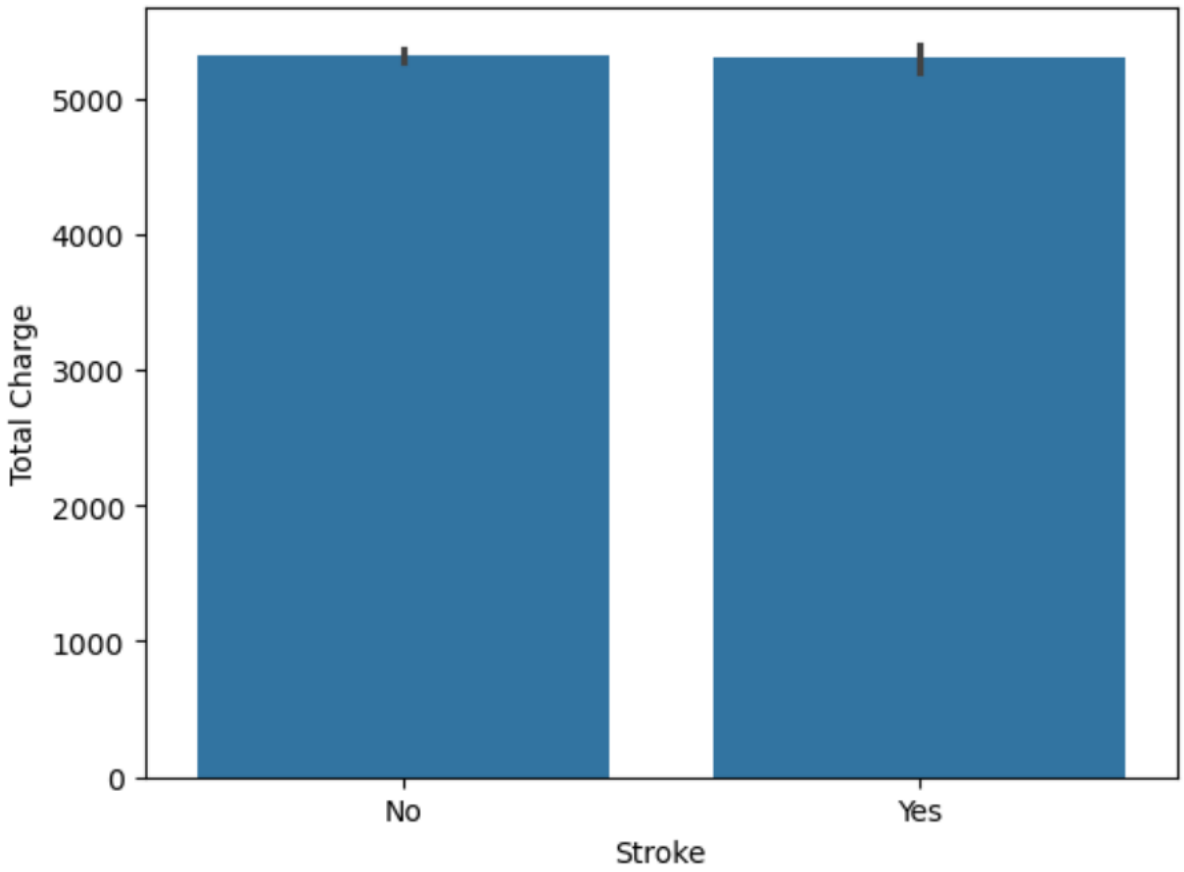

Barchart for Reflux_esophagitis
6000
5000
4000
3000
2000
1000
0
No
Yes
Reflux_esophagitis

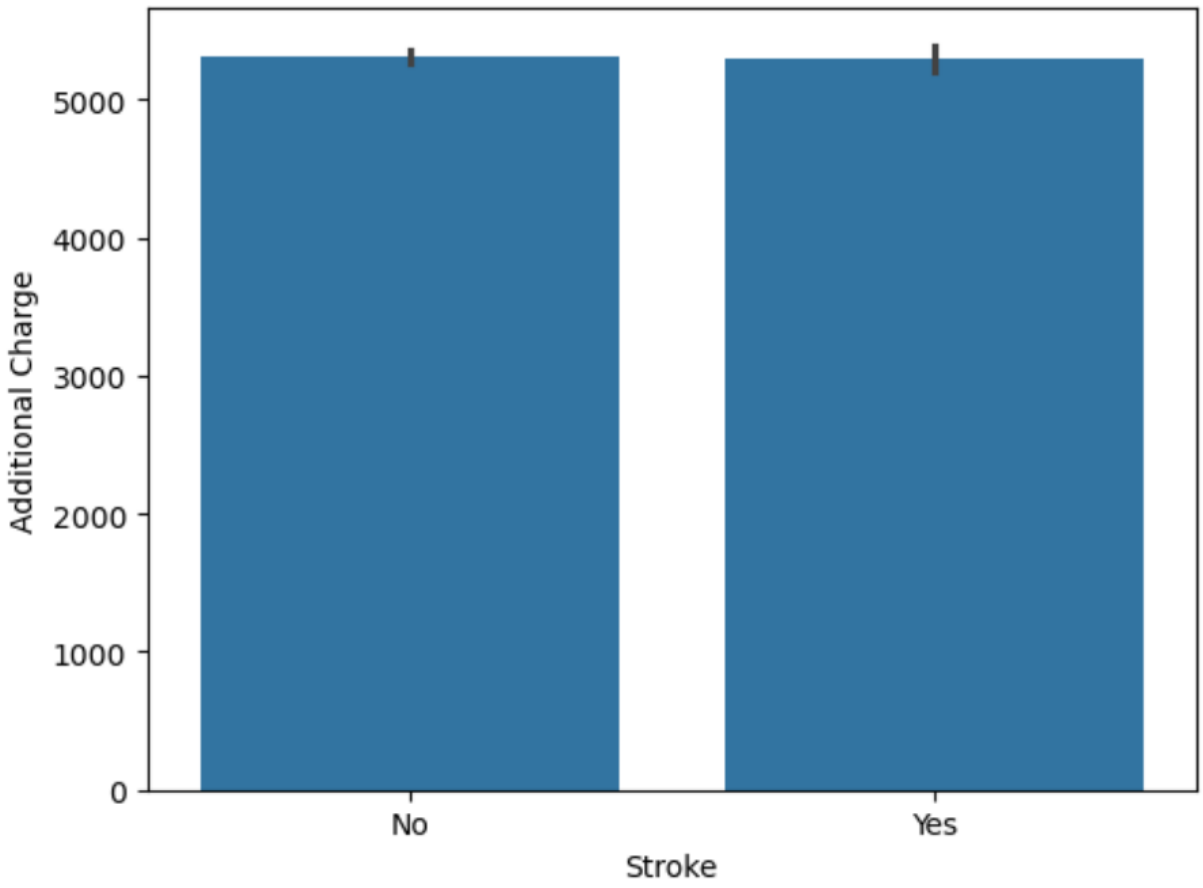

Barchart for Asthma
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Asthma

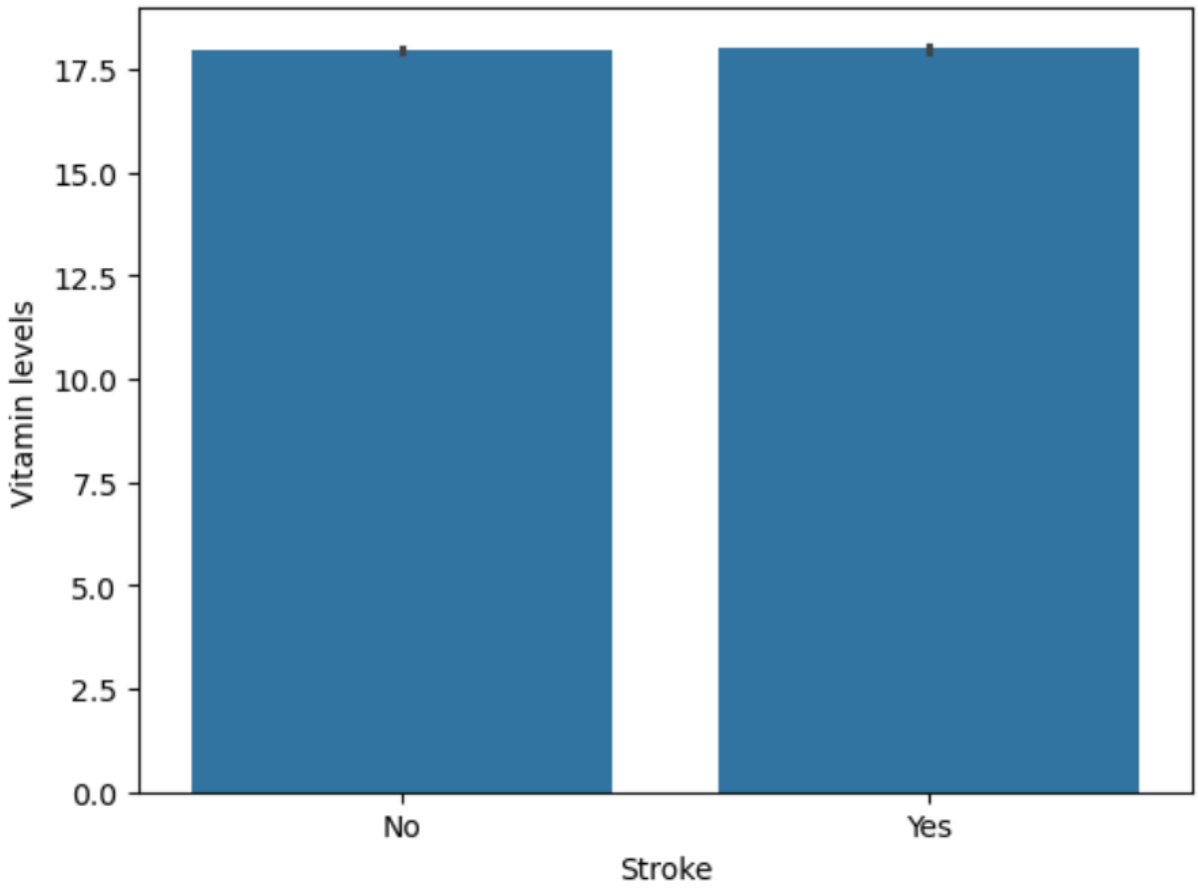
Bivariate Statistics for continuous and categorical data:

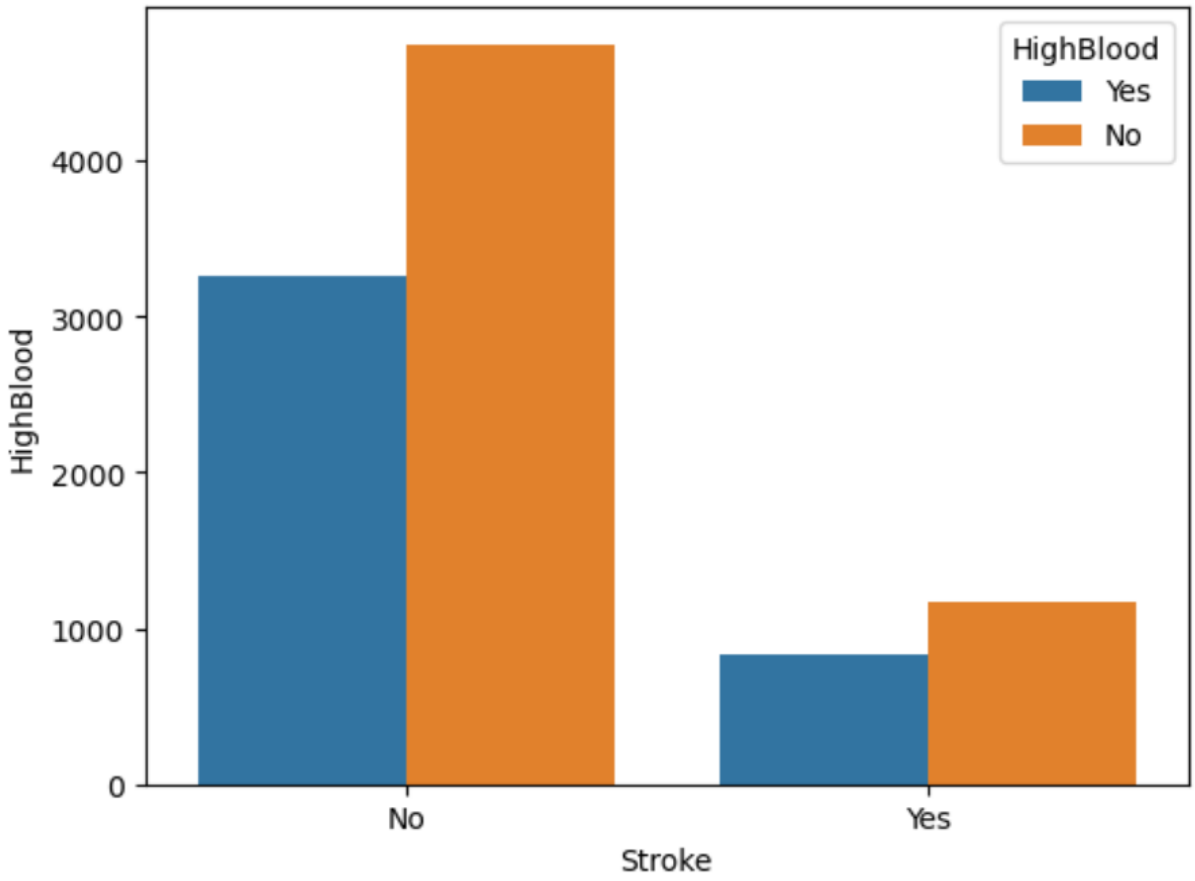


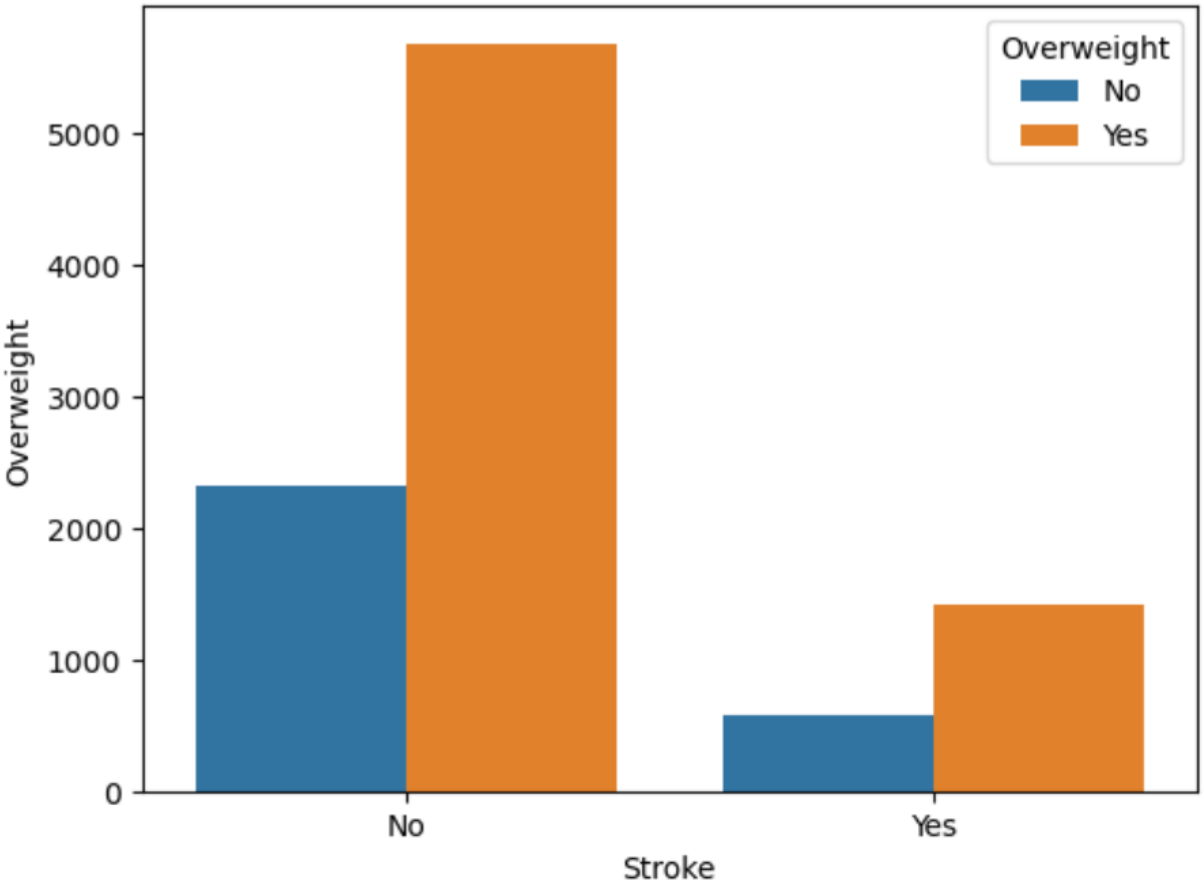

40000
35000
30000
25000
20000
15000
10000
5000
0
Income
No
Yes
Stroke

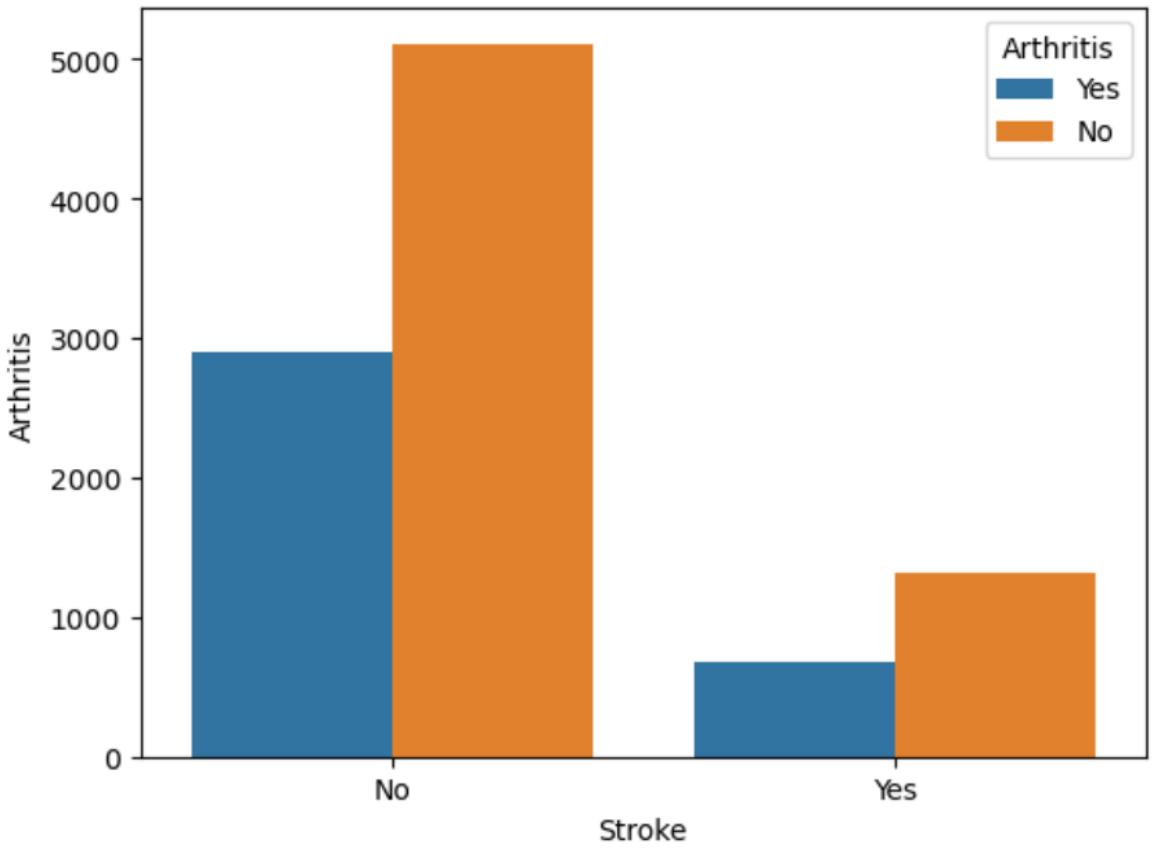

5000
4000
3000
2000
1000
0
Total Charge
No
Yes
Stroke

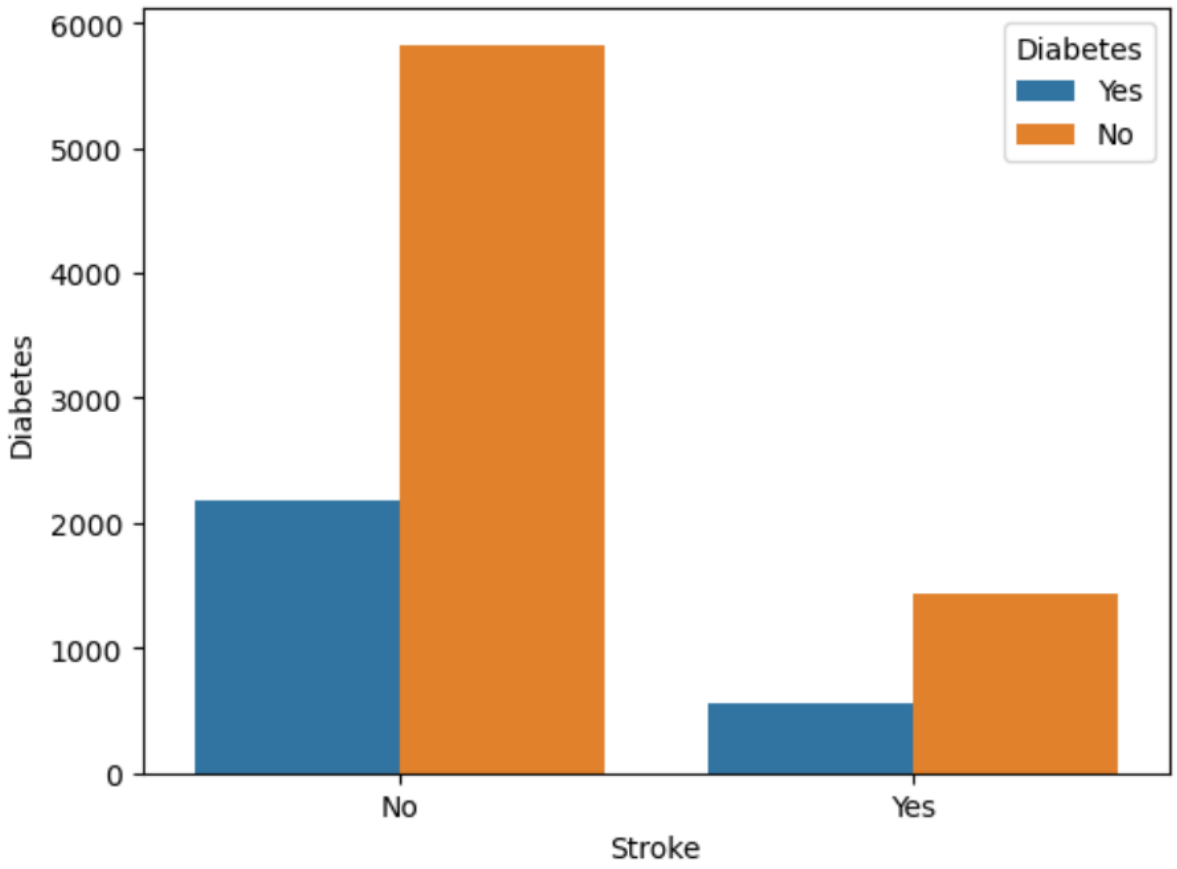

5000
4000
3000
2000
1000
0
Additional Charge
No
Yes
Stroke

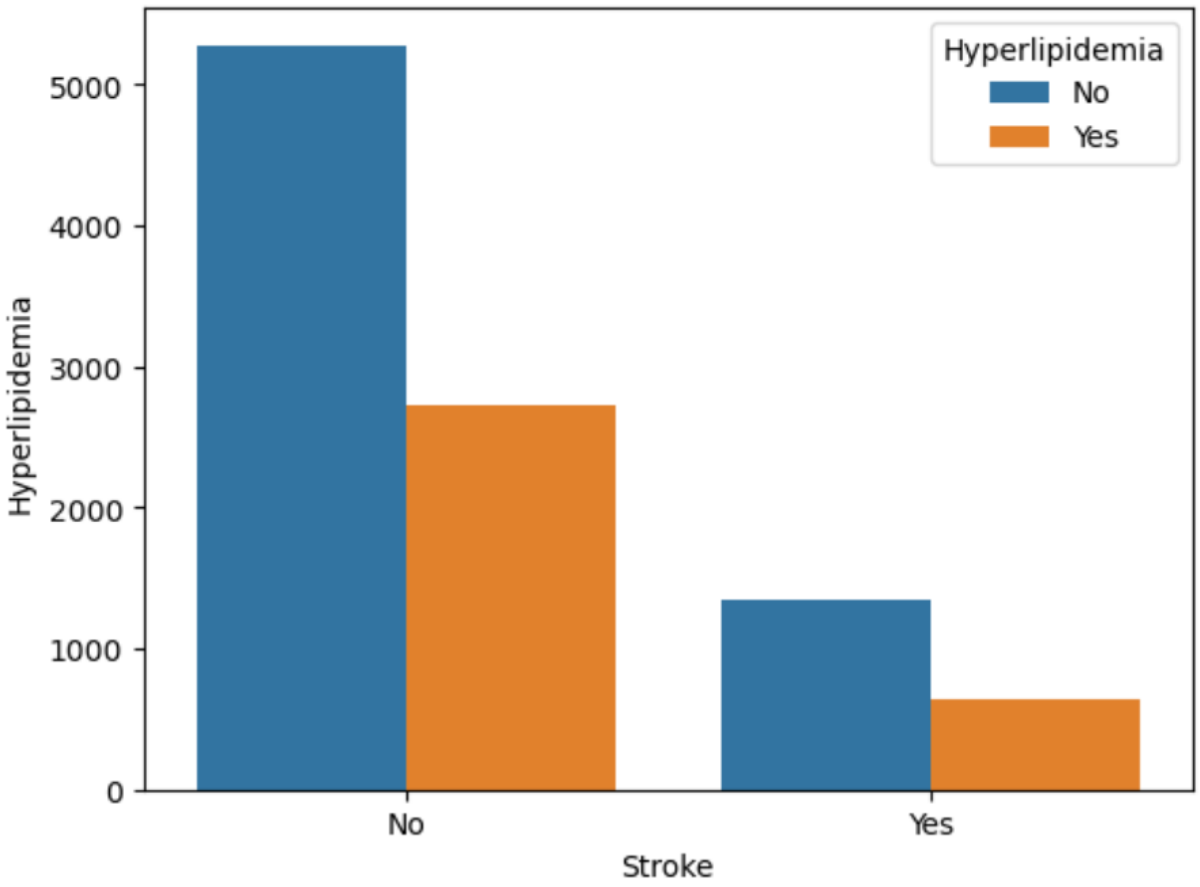

17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Vitamin levels
No
Yes
Stroke

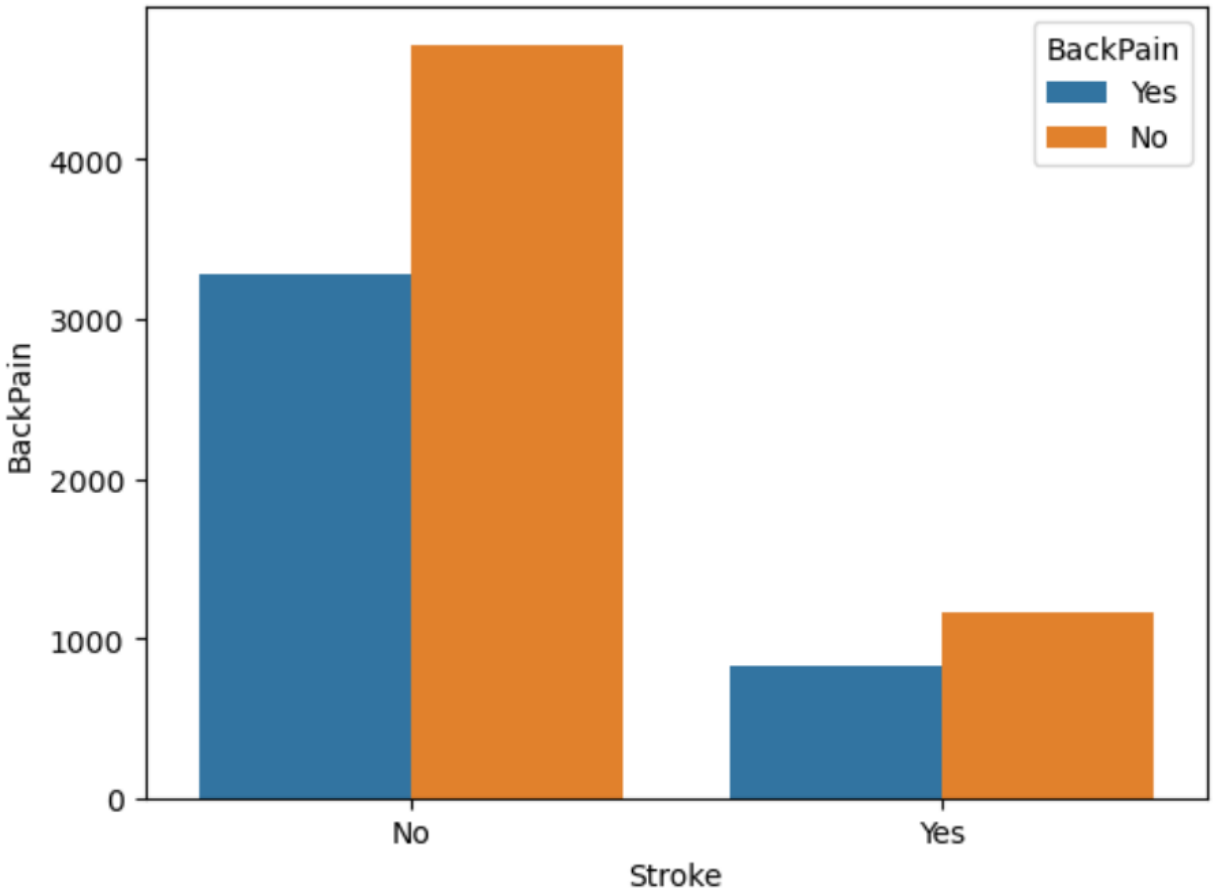

HighBlood
Yes
No
4000
3000
2000
1000
0
HighBlood
No
Yes
Stroke

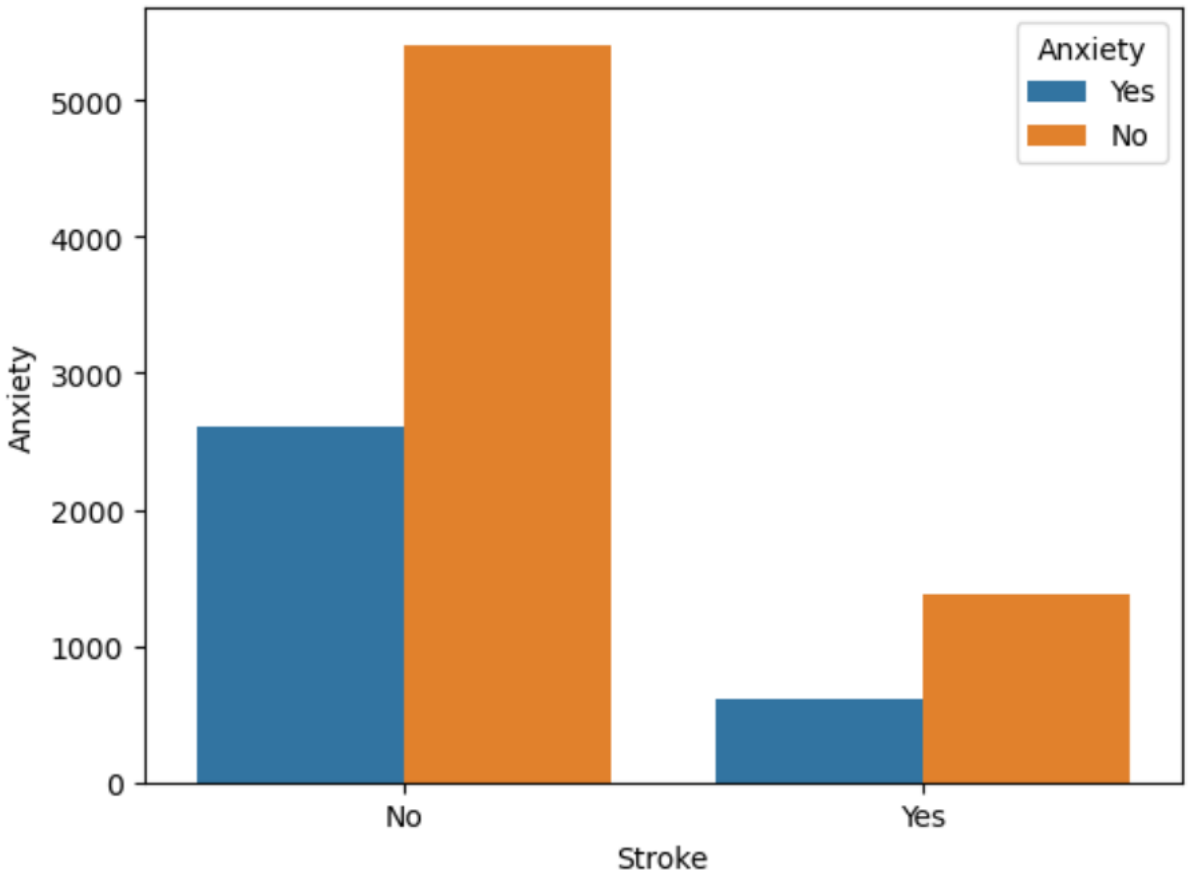

Overweight
No
Yes
5000
4000
3000
2000
1000
0
Overweight
No
Yes
Stroke

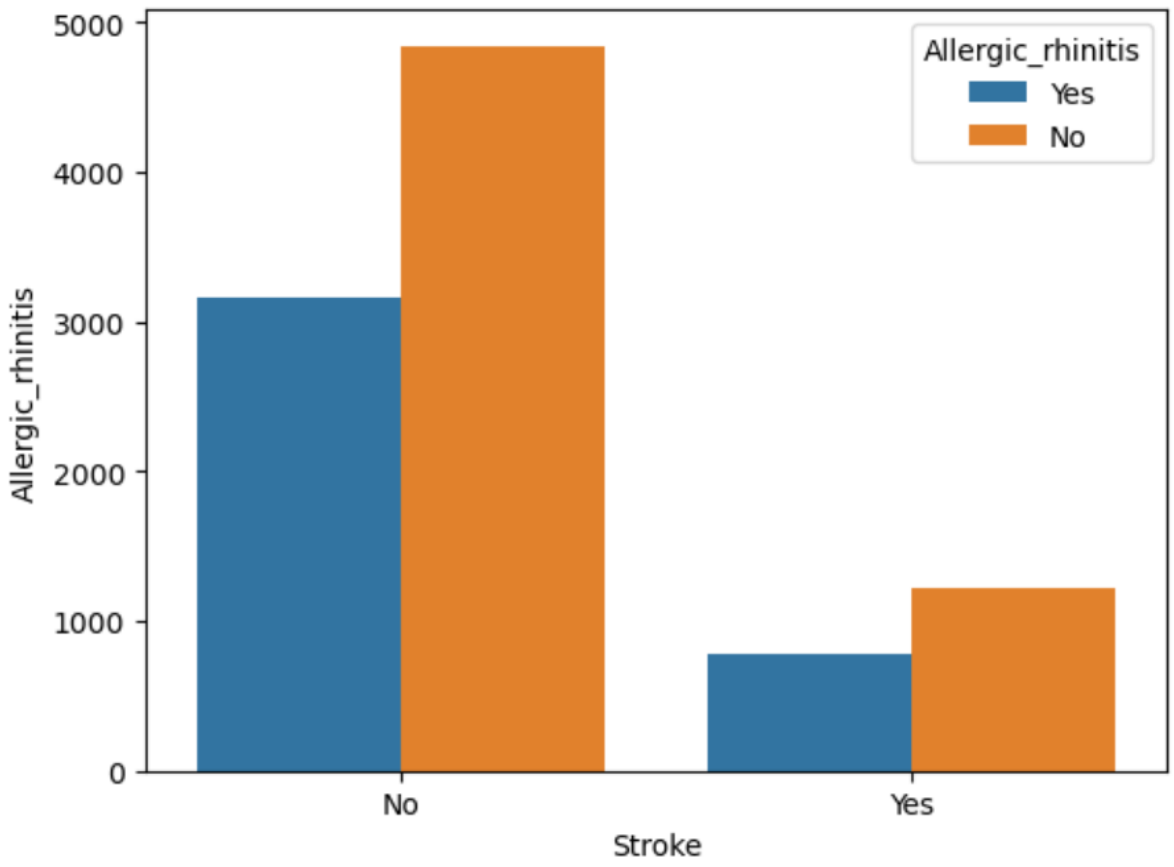

Arthritis
Yes
No
5000
4000
3000
2000
1000
0
Arthritis
No
Yes
Stroke

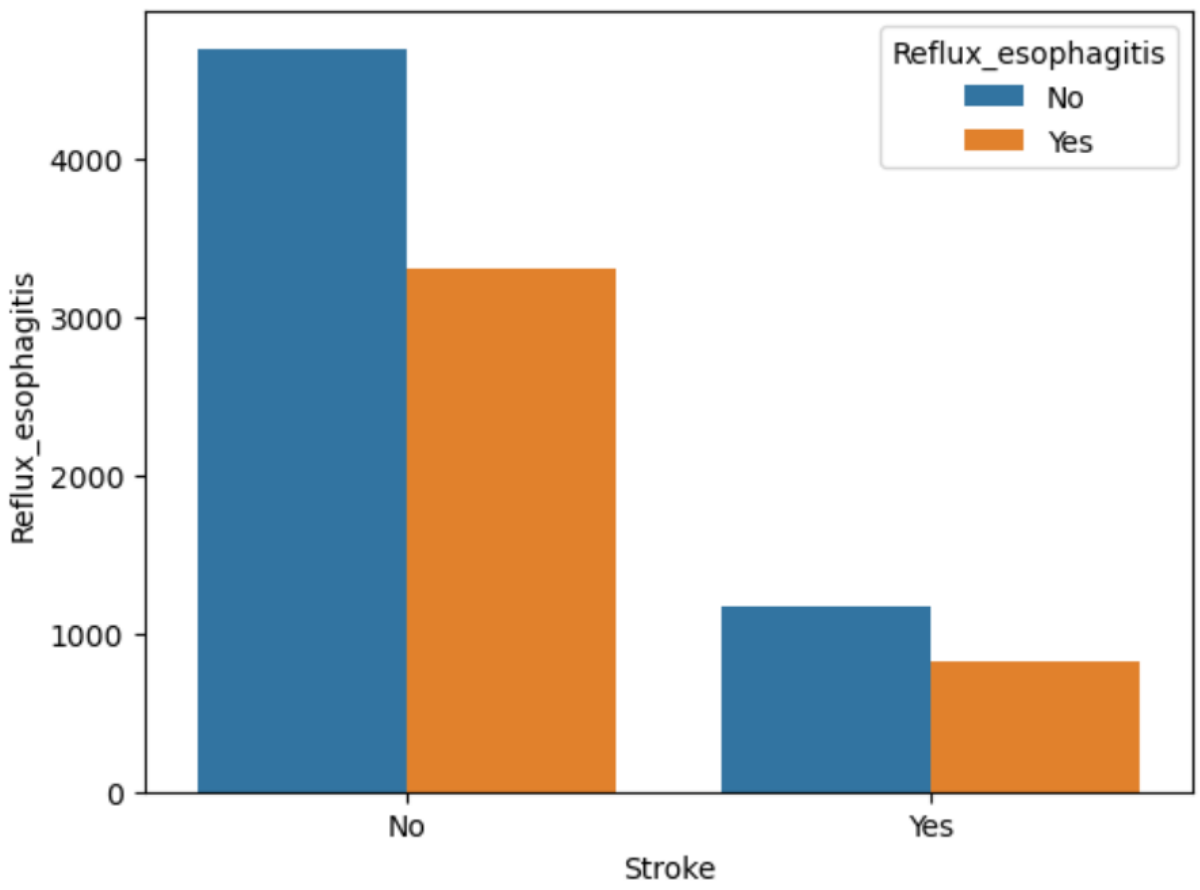

Diabetes
Yes
No
6000
5000
4000
3000
2000
1000
0
Diabetes
No
Yes
Stroke

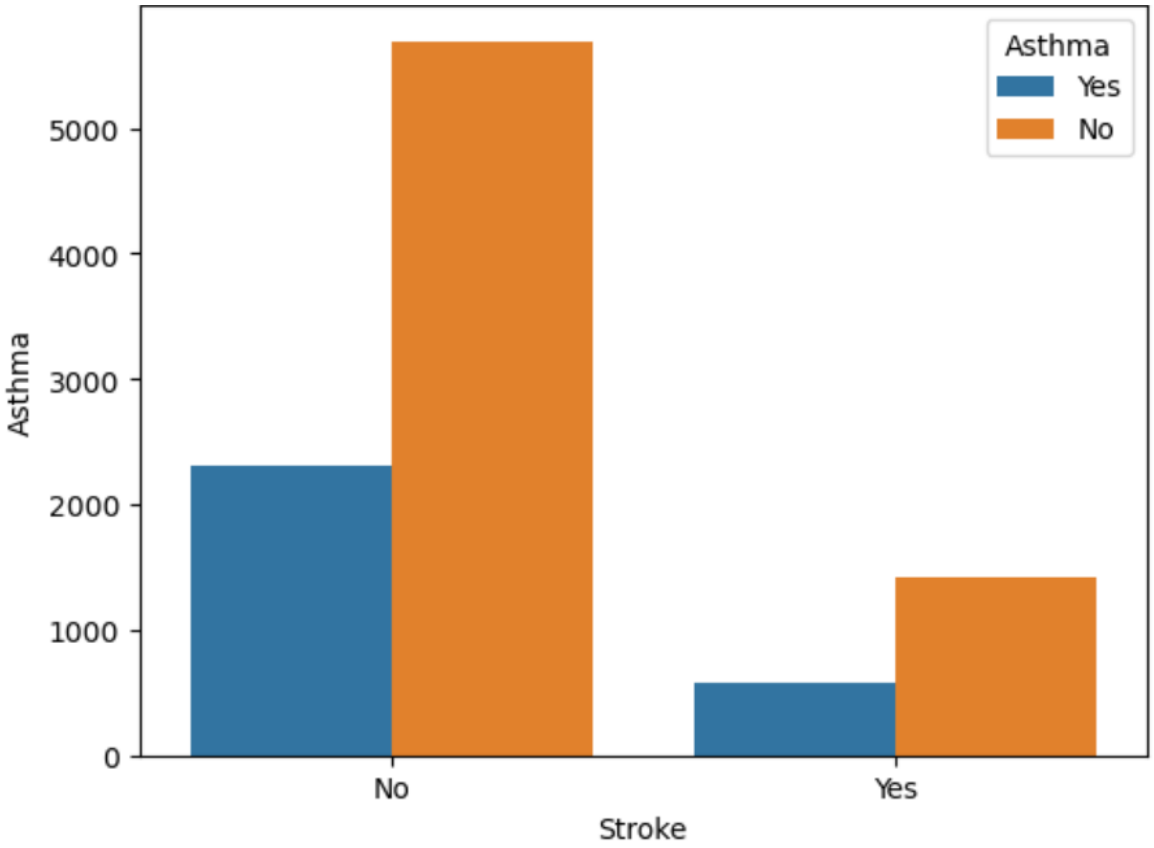

Hyperlipidemia
No
Yes
5000
4000
3000
2000
1000
0
Hyperlipidemia
No
Yes
Stroke


BackPain
Yes
No
4000
3000
2000
1000
0
BackPain
No
Yes
Stroke


Anxiety
Yes
No
5000
4000
3000
2000
1000
0
Anxiety
No
Yes
Stroke


Allergic_rhinitis
Yes
No
5000
4000
3000
2000
1000
0
Allergic_rhinitis
No
Yes
Stroke


Reflux_esophagitis
No
Yes
4000
3000
2000
1000
0
Reflux_esophagitis
No
Yes
Stroke



C4. Data Transformation:

The data transformation I performed for my Logistic Regression model is one-hot encoding to convert my categorical variables into dummy variables. The model requires numerical data and variables with categorical variables more than two categories need to be one – hot encoded. For binary variables you only need to convert them into 0 and 1. Using the get_dummies() function we can convert the categorical variables into numerical form for the model to interpret.

C5. Prepared dataset

The final dataset used for my data analysis is 'medicalT2_clean.csv' .

D1. Initial Model

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9986
Method:	MLE	Df Model:	13
Date:	Mon, 14 Oct 2024	Pseudo R-squ.:	0.0008856
Time:	17:57:13	Log-Likelihood:	-4989.9
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.7845

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3006	0.102	-12.737	0.000	-1.501	-1.100
Income	2.659e-07	9.61e-07	0.277	0.782	-1.62e-06	2.15e-06
TotalCharge	-2.812e-06	1.15e-05	-0.244	0.807	-2.54e-05	1.98e-05
Gender_Male	-0.0189	0.051	-0.373	0.709	-0.118	0.080
Gender_Nonbinary	0.0668	0.172	0.389	0.697	-0.269	0.403
Overweight_Yes	-0.0066	0.055	-0.121	0.904	-0.115	0.101
Arthritis_Yes	-0.0962	0.053	-1.822	0.068	-0.200	0.007
Diabetes_Yes	0.0346	0.056	0.618	0.537	-0.075	0.144
Hyperlipidemia_Yes	-0.0813	0.053	-1.521	0.128	-0.186	0.023
BackPain_Yes	0.0172	0.051	0.338	0.735	-0.083	0.117
Anxiety_Yes	-0.0746	0.054	-1.378	0.168	-0.181	0.031
Allergic_rhinitis_Yes	-0.0243	0.051	-0.473	0.636	-0.125	0.076
Reflux_esophagitis_Yes	0.0005	0.051	0.010	0.992	-0.099	0.100
Asthma_Yes	0.0120	0.055	0.217	0.828	-0.096	0.120

D2. Justification of Model Reduction:

Before creating the model, I want to check for multicollinearity between the variables. The Variance Inflation factor measures how much multi-collinearity is present between the variables in a regression model. Thus, I remove the variables based on the following conditions:

- Below 5: Considered low correlation between variables, indicating minimal multicollinearity.
- Between 5 and 10: May indicate moderate correlation, prompting further analysis.
- Above 10: Suggests high multicollinearity, often requiring adjustments to the model

Based on VIF values Age, VitD_levels, Additional_charges, HighBlood_yes were eliminated.

Next, I use the stepwise backward elimination technique where the variable with the highest p-value is removed from the model, a new model fits, and this process is repeated. The p-value for 'Arthritis_yes' is 0.065; I decided to retain this variable due to it being close to the standard significance level of 0.05. Finally, we are able to get a reduced logistic regression model. This is a simplified way to select a subset of variables.

D3. Reduced Linear Regression model

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9987
Method:	MLE	Df Model:	12
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0008856
Time:	14:10:43	Log-Likelihood:	-4989.9
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.7161

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3004	0.100	-12.975	0.000	-1.497	-1.104
Income	2.661e-07	9.61e-07	0.277	0.782	-1.62e-06	2.15e-06
TotalCharge	-2.809e-06	1.15e-05	-0.244	0.807	-2.54e-05	1.98e-05
Gender_Male	-0.0189	0.051	-0.373	0.709	-0.118	0.080
Gender_Nonbinary	0.0668	0.172	0.389	0.697	-0.269	0.403
Overweight_Yes	-0.0067	0.055	-0.121	0.904	-0.115	0.101
Arthritis_Yes	-0.0962	0.053	-1.822	0.068	-0.200	0.007
Diabetes_Yes	0.0346	0.056	0.618	0.537	-0.075	0.144
Hyperlipidemia_Yes	-0.0813	0.053	-1.521	0.128	-0.186	0.023
BackPain_Yes	0.0172	0.051	0.339	0.735	-0.083	0.117
Anxiety_Yes	-0.0746	0.054	-1.378	0.168	-0.181	0.031
Allergic_rhinitis_Yes	-0.0243	0.051	-0.473	0.636	-0.125	0.076
Asthma_Yes	0.0120	0.055	0.217	0.828	-0.096	0.120

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9988
Method:	MLE	Df Model:	11
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0008841
Time:	14:10:43	Log-Likelihood:	-4989.9
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.6375

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3053	0.092	-14.241	0.000	-1.485	-1.126
Income	2.685e-07	9.61e-07	0.279	0.780	-1.61e-06	2.15e-06
TotalCharge	-2.791e-06	1.15e-05	-0.242	0.809	-2.54e-05	1.98e-05
Gender_Male	-0.0189	0.051	-0.373	0.709	-0.118	0.080
Gender_Nonbinary	0.0667	0.172	0.389	0.697	-0.270	0.403
Arthritis_Yes	-0.0962	0.053	-1.823	0.068	-0.200	0.007
Diabetes_Yes	0.0346	0.056	0.619	0.536	-0.075	0.144
Hyperlipidemia_Yes	-0.0813	0.053	-1.521	0.128	-0.186	0.023
BackPain_Yes	0.0172	0.051	0.337	0.736	-0.083	0.117
Anxiety_Yes	-0.0745	0.054	-1.377	0.169	-0.181	0.032
Allergic_rhinitis_Yes	-0.0243	0.051	-0.473	0.636	-0.125	0.076
Asthma_Yes	0.0119	0.055	0.216	0.829	-0.096	0.120

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9989
Method:	MLE	Df Model:	10
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0008795
Time:	14:10:44	Log-Likelihood:	-4989.9
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.5527

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3019	0.090	-14.421	0.000	-1.479	-1.125
Income	2.706e-07	9.61e-07	0.282	0.778	-1.61e-06	2.15e-06
TotalCharge	-2.831e-06	1.15e-05	-0.246	0.806	-2.54e-05	1.97e-05
Gender_Male	-0.0188	0.051	-0.372	0.710	-0.118	0.081
Gender_Nonbinary	0.0668	0.172	0.390	0.697	-0.269	0.403
Arthritis_Yes	-0.0963	0.053	-1.824	0.068	-0.200	0.007
Diabetes_Yes	0.0348	0.056	0.623	0.534	-0.075	0.144
Hyperlipidemia_Yes	-0.0814	0.053	-1.523	0.128	-0.186	0.023
BackPain_Yes	0.0173	0.051	0.341	0.733	-0.082	0.117
Anxiety_Yes	-0.0744	0.054	-1.375	0.169	-0.180	0.032
Allergic_rhinitis_Yes	-0.0243	0.051	-0.472	0.637	-0.125	0.076

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9990
Method:	MLE	Df Model:	9
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0008734
Time:	14:10:44	Log-Likelihood:	-4989.9
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.4631

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3163	0.069	-19.195	0.000	-1.451	-1.182
Income	2.742e-07	9.6e-07	0.285	0.775	-1.61e-06	2.16e-06
Gender_Male	-0.0189	0.051	-0.373	0.709	-0.118	0.080
Gender_Nonbinary	0.0666	0.172	0.388	0.698	-0.270	0.403
Arthritis_Yes	-0.0967	0.053	-1.833	0.067	-0.200	0.007
Diabetes_Yes	0.0347	0.056	0.620	0.535	-0.075	0.144
Hyperlipidemia_Yes	-0.0816	0.053	-1.527	0.127	-0.186	0.023
BackPain_Yes	0.0169	0.051	0.332	0.740	-0.083	0.117
Anxiety_Yes	-0.0748	0.054	-1.383	0.167	-0.181	0.031
Allergic_rhinitis_Yes	-0.0245	0.051	-0.477	0.633	-0.125	0.076

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9991
Method:	MLE	Df Model:	8
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0008653
Time:	14:10:44	Log-Likelihood:	-4990.0
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.3733

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3054	0.057	-22.920	0.000	-1.417	-1.194
Gender_Male	-0.0189	0.051	-0.372	0.710	-0.118	0.080
Gender_Nonbinary	0.0666	0.172	0.388	0.698	-0.270	0.403
Arthritis_Yes	-0.0967	0.053	-1.834	0.067	-0.200	0.007
Diabetes_Yes	0.0345	0.056	0.617	0.537	-0.075	0.144
Hyperlipidemia_Yes	-0.0815	0.053	-1.526	0.127	-0.186	0.023
BackPain_Yes	0.0170	0.051	0.334	0.738	-0.083	0.117
Anxiety_Yes	-0.0748	0.054	-1.383	0.167	-0.181	0.031
Allergic_rhinitis_Yes	-0.0246	0.051	-0.478	0.632	-0.125	0.076

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9992
Method:	MLE	Df Model:	7
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0008541
Time:	14:10:44	Log-Likelihood:	-4990.0
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.2881

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2982	0.053	-24.636	0.000	-1.401	-1.195
Gender_Male	-0.0191	0.051	-0.377	0.706	-0.118	0.080
Gender_Nonbinary	0.0676	0.172	0.394	0.693	-0.269	0.404
Arthritis_Yes	-0.0971	0.053	-1.841	0.066	-0.200	0.006
Diabetes_Yes	0.0343	0.056	0.613	0.540	-0.075	0.144
Hyperlipidemia_Yes	-0.0815	0.053	-1.526	0.127	-0.186	0.023
Anxiety_Yes	-0.0746	0.054	-1.380	0.168	-0.181	0.031
Allergic_rhinitis_Yes	-0.0245	0.051	-0.477	0.634	-0.125	0.076

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9993
Method:	MLE	Df Model:	6
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0008399
Time:	14:10:44	Log-Likelihood:	-4990.1
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.2110

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3074	0.047	-27.941	0.000	-1.399	-1.216
Gender_Nonbinary	0.0769	0.170	0.453	0.650	-0.256	0.410
Arthritis_Yes	-0.0973	0.053	-1.845	0.065	-0.201	0.006
Diabetes_Yes	0.0343	0.056	0.614	0.539	-0.075	0.144
Hyperlipidemia_Yes	-0.0818	0.053	-1.532	0.125	-0.187	0.023
Anxiety_Yes	-0.0745	0.054	-1.377	0.168	-0.180	0.032
Allergic_rhinitis_Yes	-0.0245	0.051	-0.476	0.634	-0.125	0.076

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9994
Method:	MLE	Df Model:	5
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0008196
Time:	14:10:44	Log-Likelihood:	-4990.2
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.1462

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3059	0.047	-27.982	0.000	-1.397	-1.214
Arthritis_Yes	-0.0970	0.053	-1.839	0.066	-0.200	0.006
Diabetes_Yes	0.0343	0.056	0.613	0.540	-0.075	0.144
Hyperlipidemia_Yes	-0.0816	0.053	-1.527	0.127	-0.186	0.023
Anxiety_Yes	-0.0743	0.054	-1.375	0.169	-0.180	0.032
Allergic_rhinitis_Yes	-0.0246	0.051	-0.479	0.632	-0.125	0.076

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9995
Method:	MLE	Df Model:	4
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0007966
Time:	14:10:44	Log-Likelihood:	-4990.3
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.09316

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3154	0.042	-31.168	0.000	-1.398	-1.233
Arthritis_Yes	-0.0972	0.053	-1.843	0.065	-0.201	0.006
Diabetes_Yes	0.0341	0.056	0.610	0.542	-0.075	0.144
Hyperlipidemia_Yes	-0.0813	0.053	-1.523	0.128	-0.186	0.023
Anxiety_Yes	-0.0745	0.054	-1.377	0.169	-0.180	0.032

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9996
Method:	MLE	Df Model:	3
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0006057
Time:	14:10:45	Log-Likelihood:	-4991.3
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.1092

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3391	0.039	-34.639	0.000	-1.415	-1.263
Arthritis_Yes	-0.0980	0.053	-1.859	0.063	-0.201	0.005
Diabetes_Yes	0.0343	0.056	0.613	0.540	-0.075	0.144
Hyperlipidemia_Yes	-0.0804	0.053	-1.505	0.132	-0.185	0.024

Logit Regression Results

Dep. Variable:	Stroke_Yes	No. Observations:	10000
Model:	Logit	Df Residuals:	9997
Method:	MLE	Df Model:	2
Date:	Thu, 17 Oct 2024	Pseudo R-squ.:	0.0005682
Time:	14:10:45	Log-Likelihood:	-4991.5
converged:	True	LL-Null:	-4994.3
Covariance Type:	nonrobust	LLR p-value:	0.05856

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3299	0.036	-37.395	0.000	-1.400	-1.260
Arthritis_Yes	-0.0977	0.053	-1.854	0.064	-0.201	0.006
Hyperlipidemia_Yes	-0.0800	0.053	-1.498	0.134	-0.185	0.025

```
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:             Stroke_Yes   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9998
Method:                           MLE   Df Model:                            1
Date:                Thu, 17 Oct 2024   Pseudo R-squ.:               0.0003423
Time:                        14:11:33   Log-Likelihood:                -4992.6
converged:                       True   LL-Null:                       -4994.3
Covariance Type:            nonrobust   LLR p-value:                   0.06444
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -1.3566      0.031    -43.885      0.000      -1.417      -1.296
Arthritis_Yes  -0.0972      0.053     -1.843      0.065      -0.200       0.006
==============================================================================
```

E1. Mode Comparison:

The initial LLR p-value of 0.7845 suggests that the full model does not fit the variables significantly better than the base model. The predictors in the initial model have contributed little. The final/reduced model has a lower p-value of 0.06444, which suggests the variables fit the model better. Also, the value is closer to 0.05, which means the model is starting to explain the variation of the outcome.

E2. Output and Calculations

```
Confusion Matrix:
[[8007    0]
 [1993    0]]
```

Confusion Matrix:

- 8007 True Negative: The model correctly predicted No Stroke (0).
- 0 False Positive: The model incorrectly predicted Stroke (1).
- 1993 False Negative: The model incorrectly predicted No Stroke (0).

- 0 True Positive: The model correctly predicted Stroke (1).

This model is poorly performing model and is biased towards predicting the majority class 'No Stroke' (0).

Accuracy calculation:
Accuracy = (TP + TN) / (TP + TN + FP + FN)

TN= 8007
FP=0
FN=1993
TP=0

Accuracy = 8007/10000 = 0.8007

E3. Code

Executable code can be found in 'D208PATask2_OviyaSelvaraj.ipynb'

F1. Results

Regression equation for the reduced model:
log(p/1-p) = -1.3566 - 0.0972 * (Arthritis_Yes)

Interpretation of the coefficients of the reduced model:
Keeping all things constant, a patient with arthritis has a decrease in their odds of having stroke by 10.21%.

Statistical and practical significance of the reduced model:
The final/reduced model is a better fit than the initial model. Even though the reduced model showed a p-value slightly higher than the threshold, it is on the verge of being statistically significant. The reduced model explains the

factors that lead to meaningful outcomes that can lead to actionable recommendations that suggest practical significance.

Limitations of the data analysis:

Logistic regression is designed to model binary classification and outcomes, so it is not suitable for non-binary variables without modification. The model is sensitive to outliers, and its sensitivity to such data points can lead to skewed results. Thus, it is essential to mitigate these outliers.

F2. Recommendations:

A slightly higher p – value above the threshold for the final model shows that many factors influencing the outcome are missing. For example, specific diseases may not be captured influencing patient condition. 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=0ad7efaa-31f8-4086-80e8-b2100114d4fe&edit=true#

H. References:

In-text citation

Binary Logistic Regression Major Assumptions

Reference entry

Statistics Solution (n.d.). What is Logistic Regression https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/

In-text citation

Binary Logistic Regression Major Assumptions

Reference entry

Statistics Solution (n.d.). What is Logistic Regression https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/

GeeksforGeeks (August, 2024) Advantages and Disadvantages of logistic regression

https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

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