Predicting Patient Readmission (Logistic Regression)

Ashley Traore

Research question:

What factors influence patient readmission status most significantly?

Goals of the Analysis:

My analysis aims to gain greater insight into determining what patient factors correlate to whether a patient gets readmitted to the hospital. The company can use this to predict which patients are at higher risk for readmission and make more informed business decisions.

Assumptions:

Four key assumptions that a logistic regression model operates under are: Firstly, it assumes a linear relationship between each explanatory variable and the logit of the response variable. Secondly, it assumes that the independent variables are not highly correlated, a condition known as multicollinearity, which can create challenges in determining the specific variable contributing to the variance in the dependent variable. Thirdly, it assumes the dependent variable is binary (Bobbitt, 2020). Fourthly, the model assumes that there are no extreme outliers influencing the results.

Programming Language and Benefits:

Python was chosen as the programming language for this project for two key reasons. Firstly, Python has simple and consistent syntax, which facilitates easier coding and debugging. Secondly, Python has a large and active community, providing an abundance of resources and tutorials for support and knowledge sharing.

Several modeling and data analysis libraries were utilized for this project:

- **Pandas** - The main library for data handling, it facilitates easy manipulation and cleaning of datasets with its powerful DataFrame structure. It was used to handle and prepare the data for analysis.
- **Scikit-learn** - A machine learning library. It provides efficient tools for data analysis and modeling. Specifically, the LabelEncoder, LogisticRegression, accuracy_score, and train_test_split classes were used. LabelEncoder was used to encode the categorical variables as numeric, LogisticRegression for

implementing the logistic regression model, accuracy_score to get the accuracy score of the model, and train_test_split for splitting the data into training and testing sets.

- **Matplotlib/Seaborn** - Both libraries were used for visualizing the complex dataset, aiding in understanding the data through various charts and plots.
- **Numpy** - Essential for scientific computing, it efficiently supports large multi-dimensional arrays and matrices, along with a suite of mathematical functions to operate on them. It was used for numerical operations.
- **Statsmodels** - Provides classes and functions for the estimation of statistical models, and for conducting statistical tests. Specifically, the formula.api and api modules were used to create the logit models. Statsmodels.stats.outliers_influence was utilized to check for multicollinearity.

Justification of using Logistic Regression:

Logistic Regression is the appropriate technique to answer the research question because it quantifies the relationship between the dependent variable and multiple independent variables. This technique helps understand how changes in various factors influence whether or not a patient gets readmitted. It allows for the identification of significant factors, which can guide improvements in patient care. Since the dependent variable, ReAdmis, is binary, Logistic Regression is the most suitable for answering this research question.

Data Cleaning:

To clean the medical dataset, I identified and addressed nulls, outliers, and duplicates. There were no duplicates or nulls identified in the dataset, so no cleaning was needed for those aspects. However, outliers were present, which can significantly impact logistic regression models in various ways, some of which include distorted results and inflated errors. To mitigate these issues, I applied z-score capping with a threshold set at 3 standard deviations from the mean. This method is straightforward and easy to implement, efficiently capping extreme values without removing them from the dataset, thus retaining as much data as possible.

Data Exploration (EDA):

	Children	Initial_days
count	10000.000000	10000.000000
mean	2.097200	34.455299
std	2.163659	26.309341
min	0.000000	1.001981
25%	0.000000	7.896215
50%	1.000000	35.836244
75%	3.000000	61.161020
max	10.000000	71.981490

Shown in the screenshot above are the summary statistics of the continuous variables used in the analysis. Listed below are each of the calculated values and a corresponding description.

- **Count**: Total number of non-null values
- **Mean**: Average value
- **STD:** Standard deviations from the mean. This indicates how spread out the data is compared to the mean.
- **Min**: Minimum value
- **25%**: First quartile
- **50%**: Second quartile
- **75%**: Third quartile
- **Max**: Maximum value.

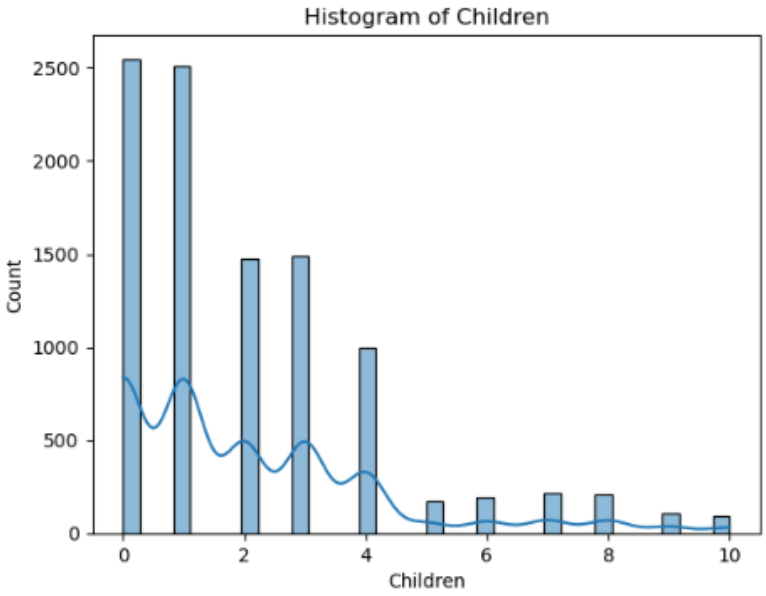
Summary statistics apply exclusively to quantitative variables, rendering the mean, median, or mode unsuitable for qualitative variables. Instead, for each categorical variable, I provide below the percentage distributions of every distinct value within each column.

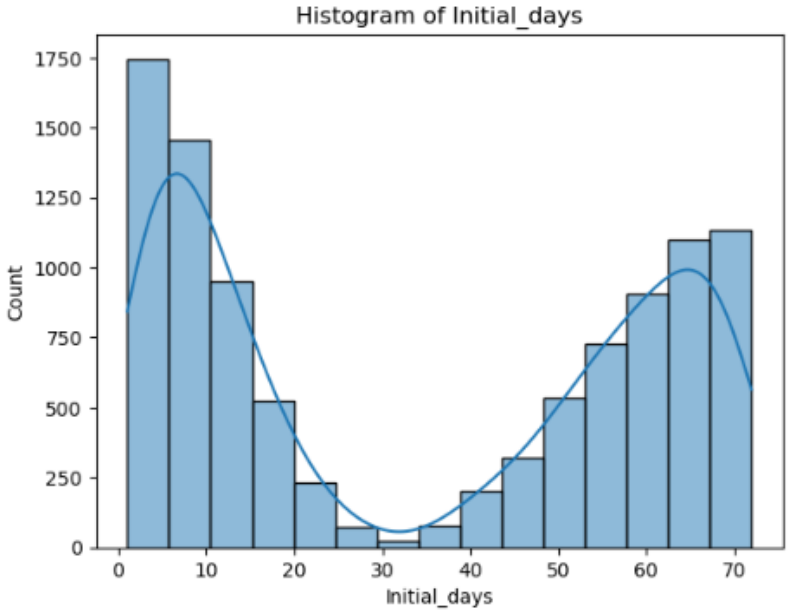
- ReAdmis
    - No 63.31%
    - Yes 36.69%
- Soft_drink
    - No 74.25%
    - Yes 25.75%
- HighBlood
    - No 59.1%
    - Yes 40.9%
- Stroke
    - No 80.07%
    - Yes 19.93%
- Complication_risk
    - Medium 45.17%

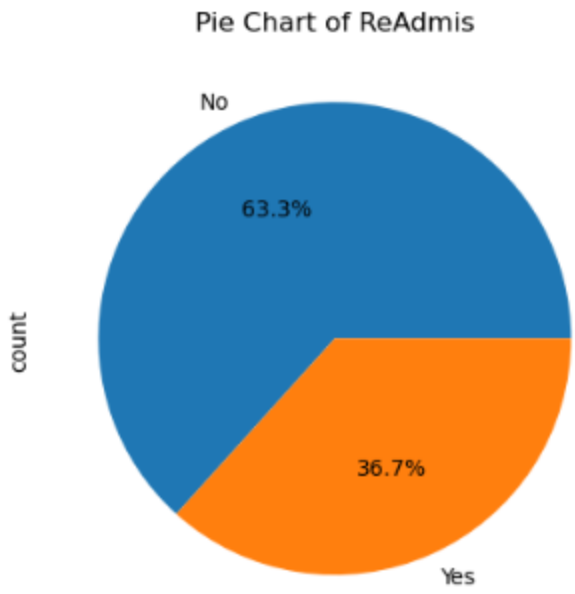
    - High 33.58%
    - Low 21.25%
- Overweight
    - Yes 70.94%
    - No 29.06%
- Arthritis
    - No 64.26%
    - Yes 35.74%
- Diabetes
    - No 72.62%
    - Yes 27.38%
- Hyperlipidemia
    - No 66.28%
    - Yes 33.72%
- BackPain
    - No 58.86%
    - Yes 41.14%
- Anxiety
    - No 67.85%
    - Yes 32.15%
- Asthma
    - No 71.07%
    - Yes 28.93%
- Services
    - Blood Work 52.65%
    - Intravenous 31.3%
    - CT Scan 12.25%
    - MRI 3.8%

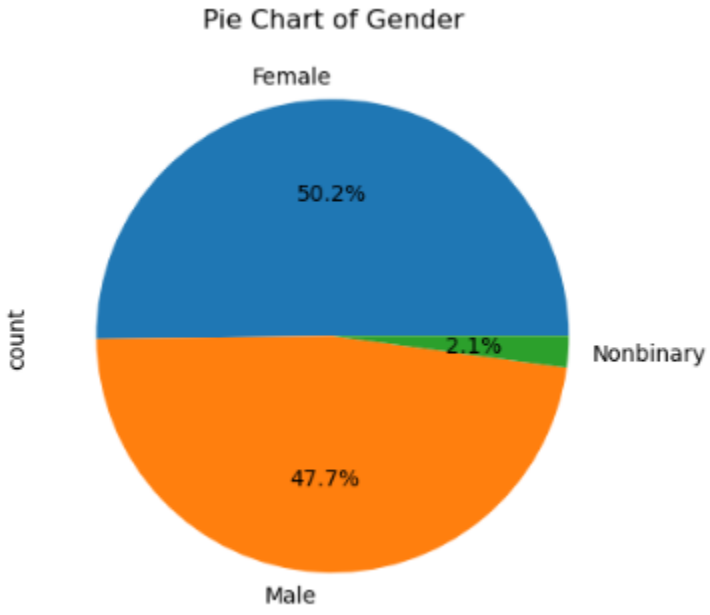
Visualizations:

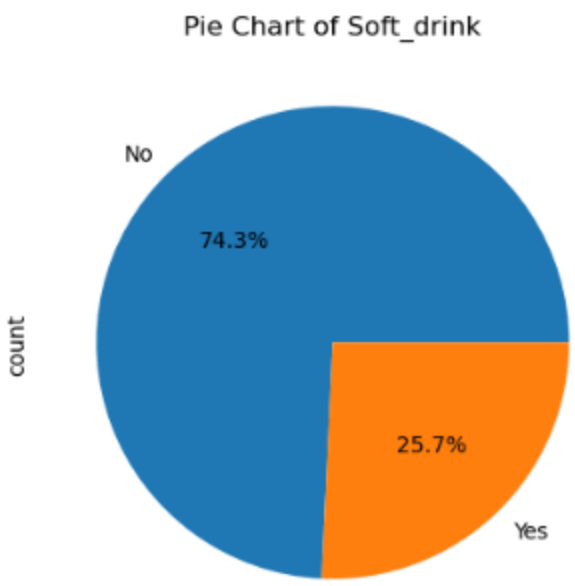
Univariate visuals:

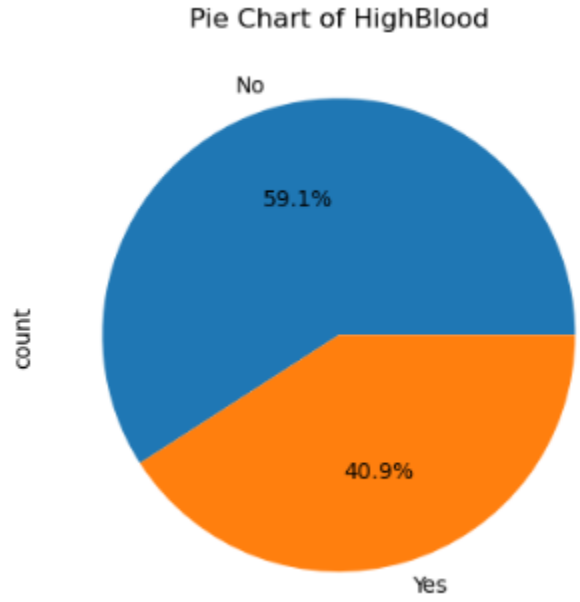

Histogram of Children
Count
Children
0
500
1000
1500
2000
2500
0
2
4
6
8
10

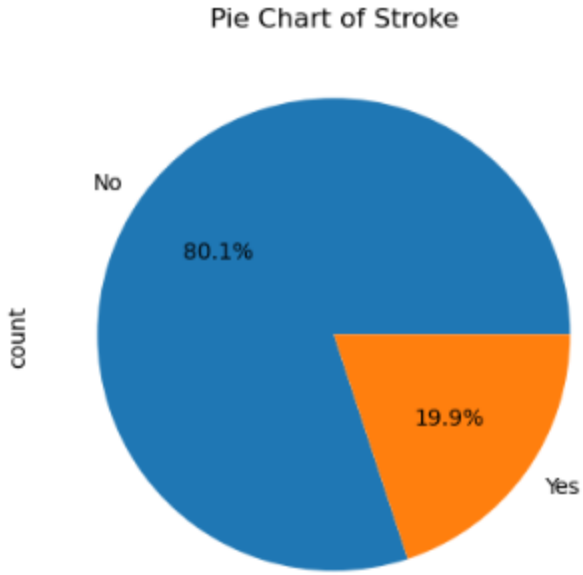

Histogram of Initial_days
Count
Initial_days
0
250
500
750
1000
1250
1500
1750
0
10
20
30
40
50
60
70

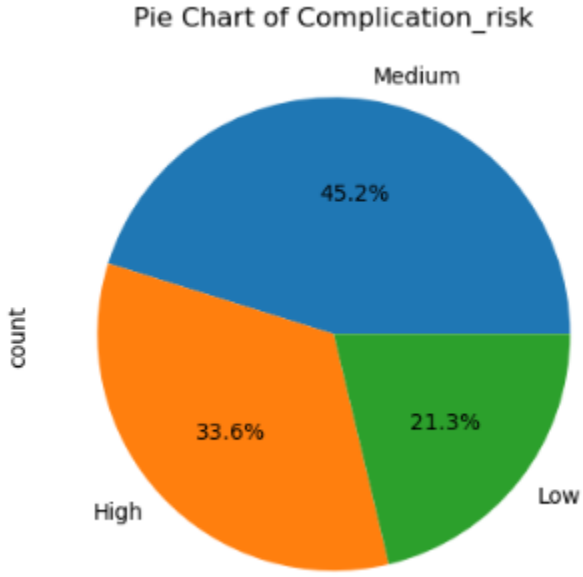

Pie Chart of ReAdmis
No
63.3%
count
36.7%
Yes

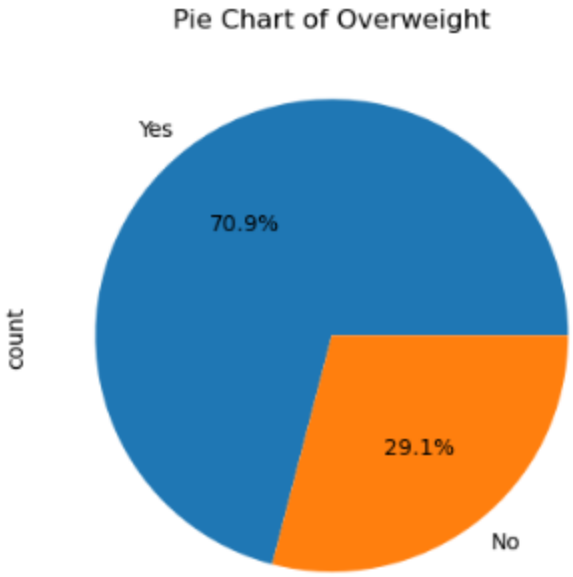

Pie Chart of Gender
Female
50.2%
count
2.1%
Nonbinary
47.7%
Male

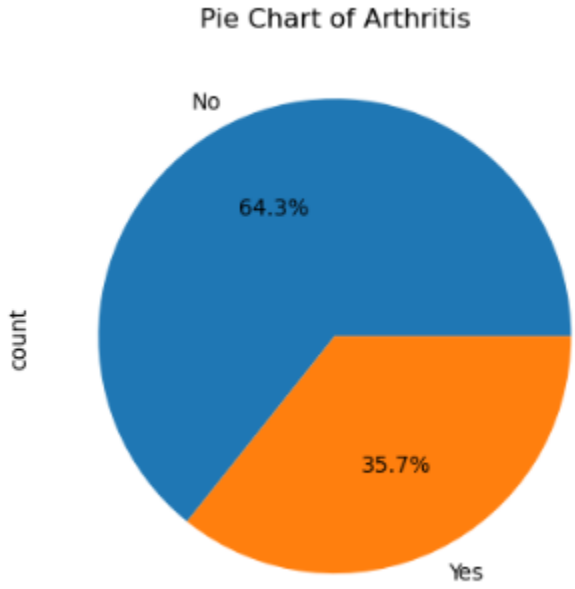

Pie Chart of Soft_drink
No
74.3%
count
25.7%
Yes

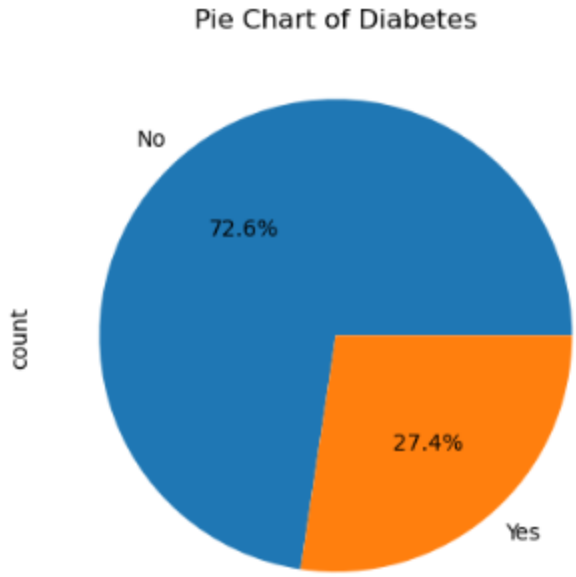

Pie Chart of HighBlood
No
59.1%
count
40.9%
Yes

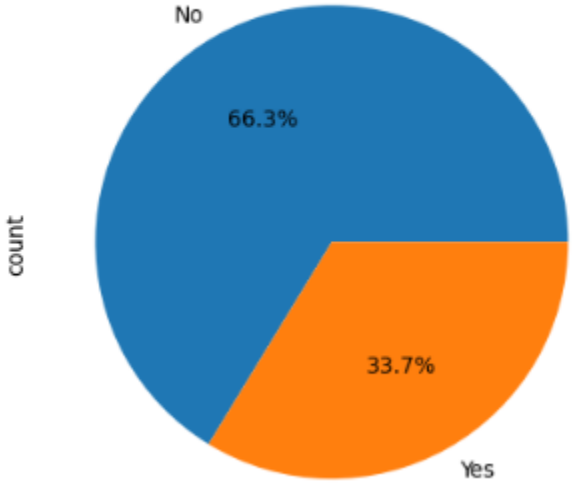

Pie Chart of Stroke
No
80.1%
count
19.9%
Yes

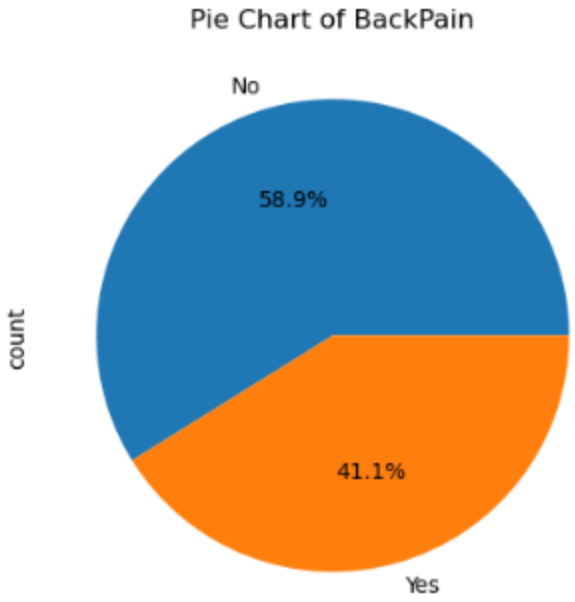

Pie Chart of Complication_risk
Medium
45.2%
count
33.6%
21.3%
High
Low

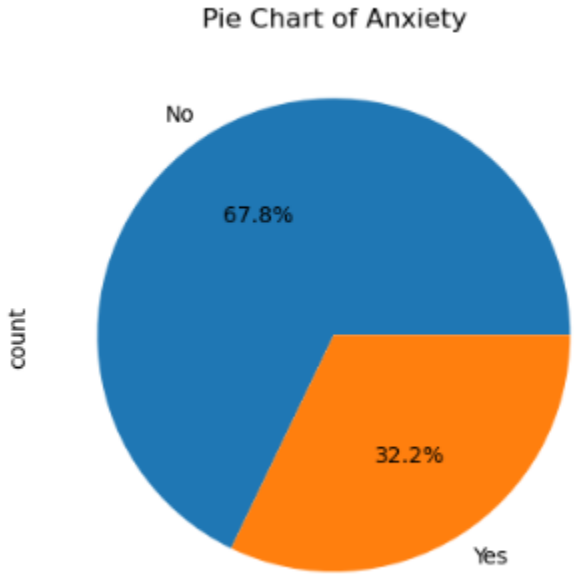

Pie Chart of Overweight
Yes
70.9%
count
29.1%
No

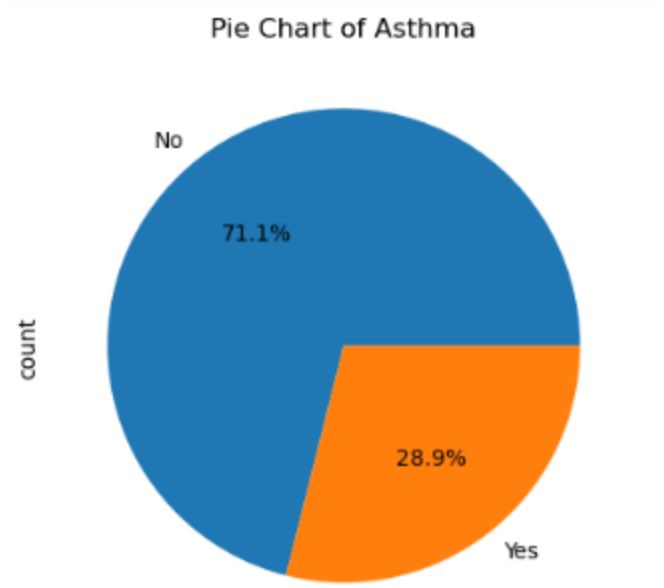

Pie Chart of Arthritis
No
64.3%
count
35.7%
Yes

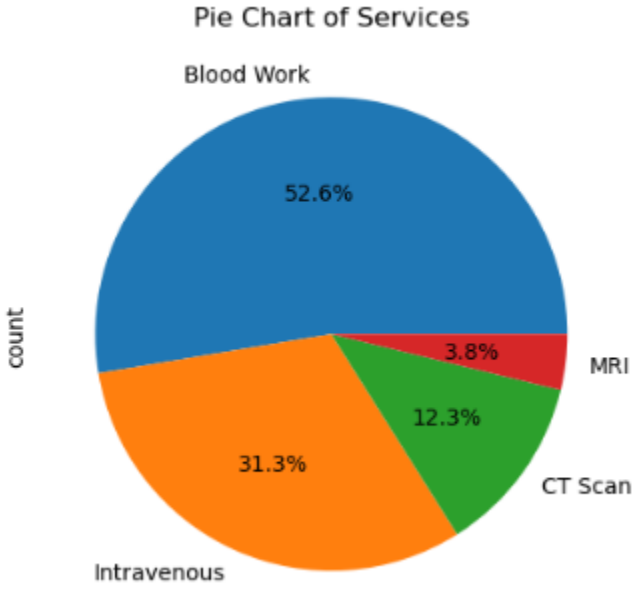

Pie Chart of Diabetes
No
72.6%
count
27.4%
Yes


Pie Chart of Hyperlipidemia
No
66.3%
count
33.7%
Yes

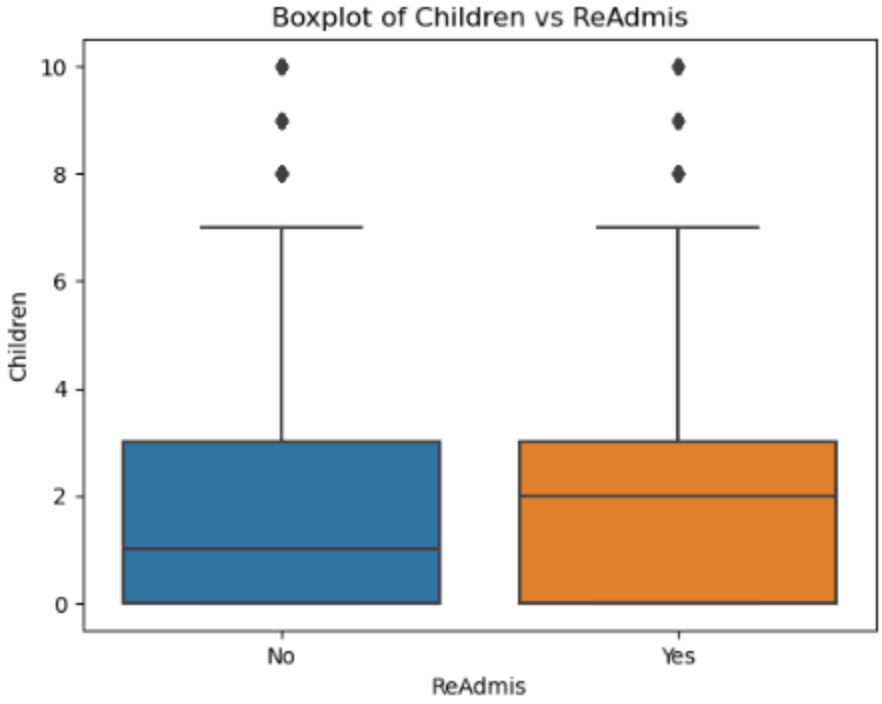

Pie Chart of BackPain
No
58.9%
count
41.1%
Yes

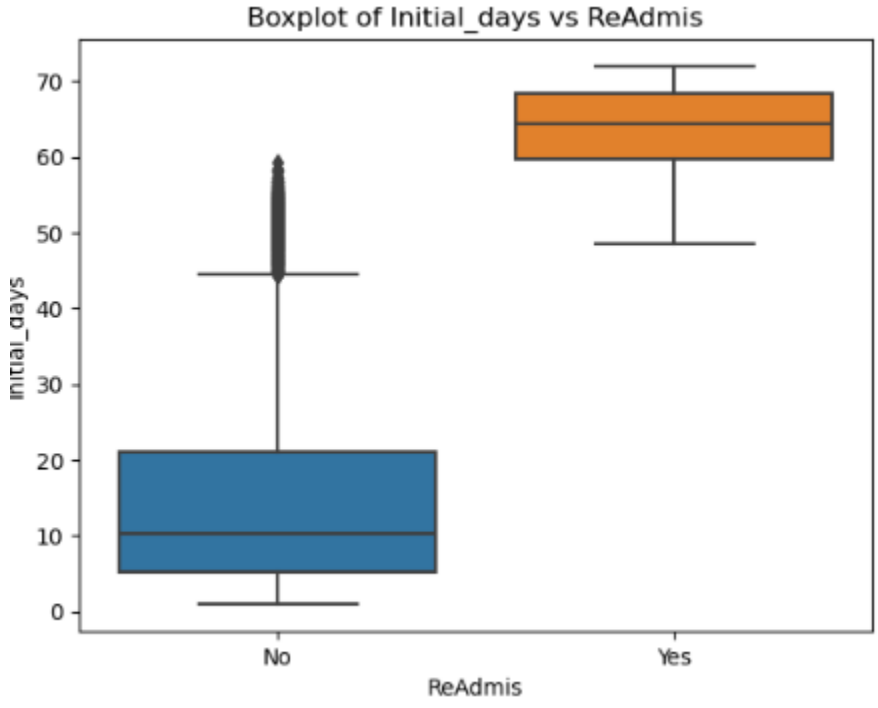

Pie Chart of Anxiety
No
67.8%
count
32.2%
Yes

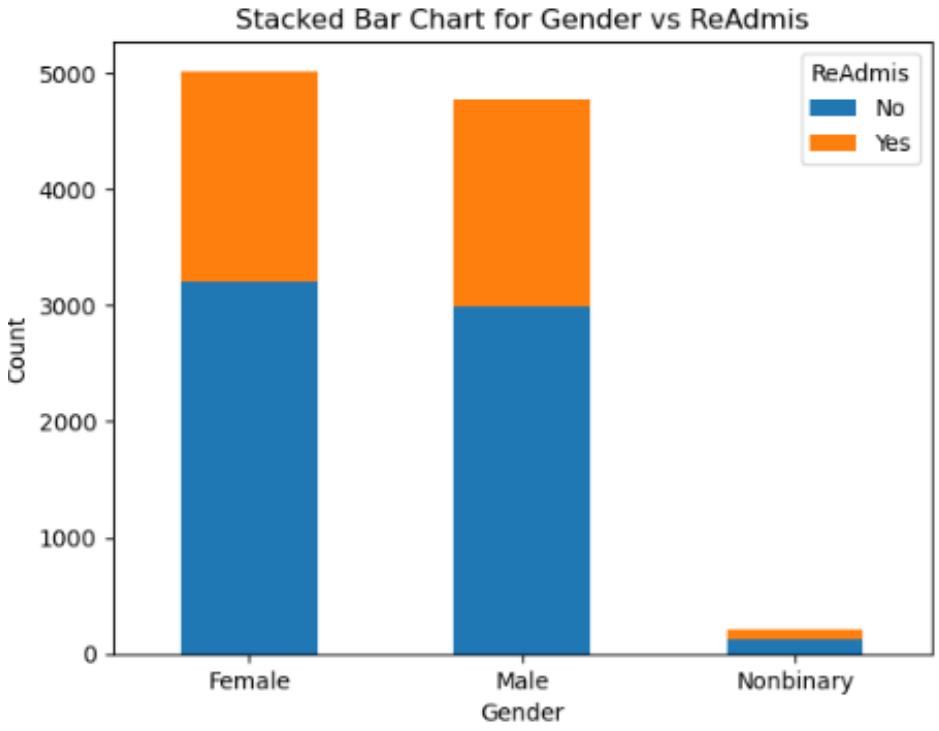


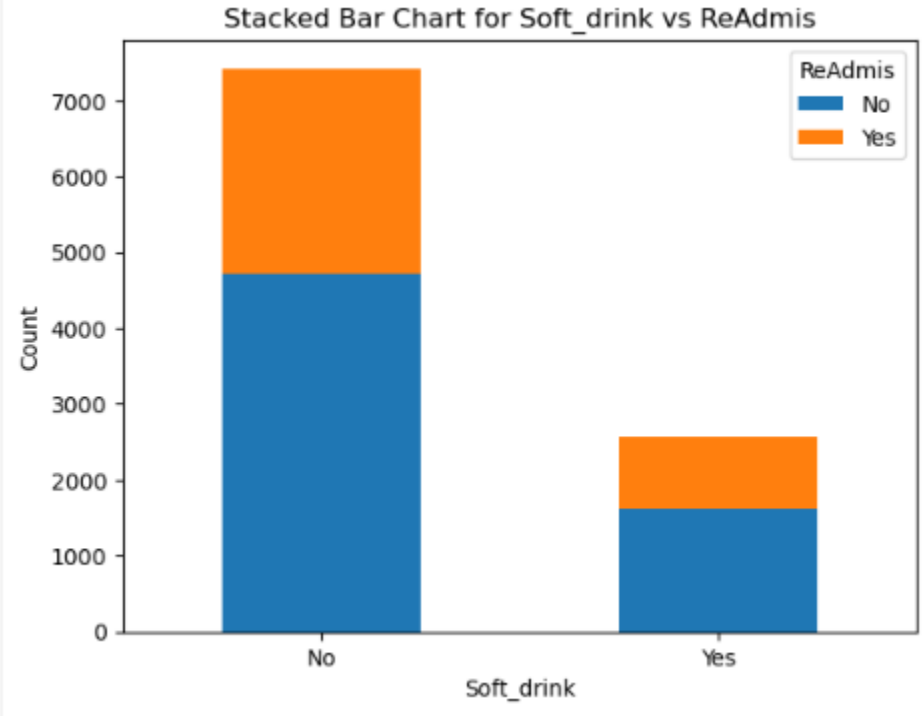


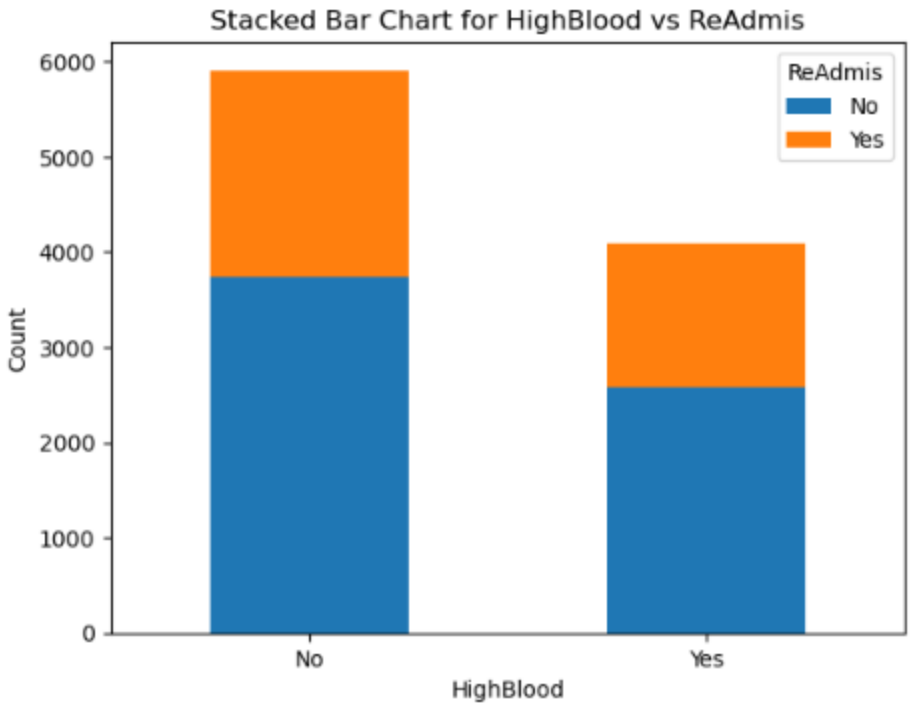
Bivariate Visuals:

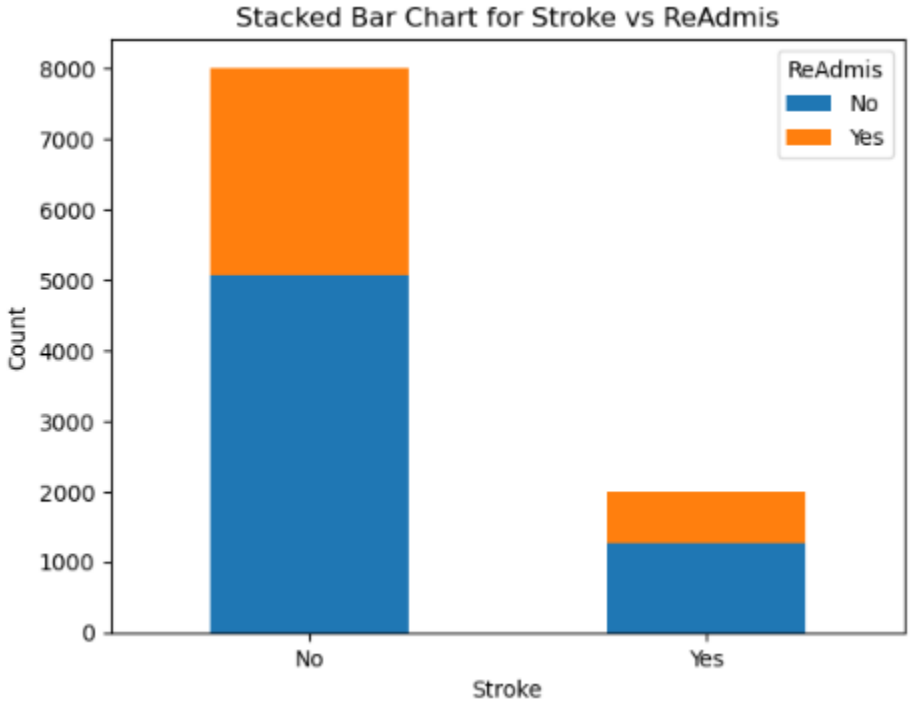

Boxplot of Children vs ReAdmis
Children
ReAdmis
No
Yes
0
2
4
6
8
10


Boxplot of Initial_days vs ReAdmis
Initial_days
ReAdmis
No
Yes
0
10
20
30
40
50
60
70

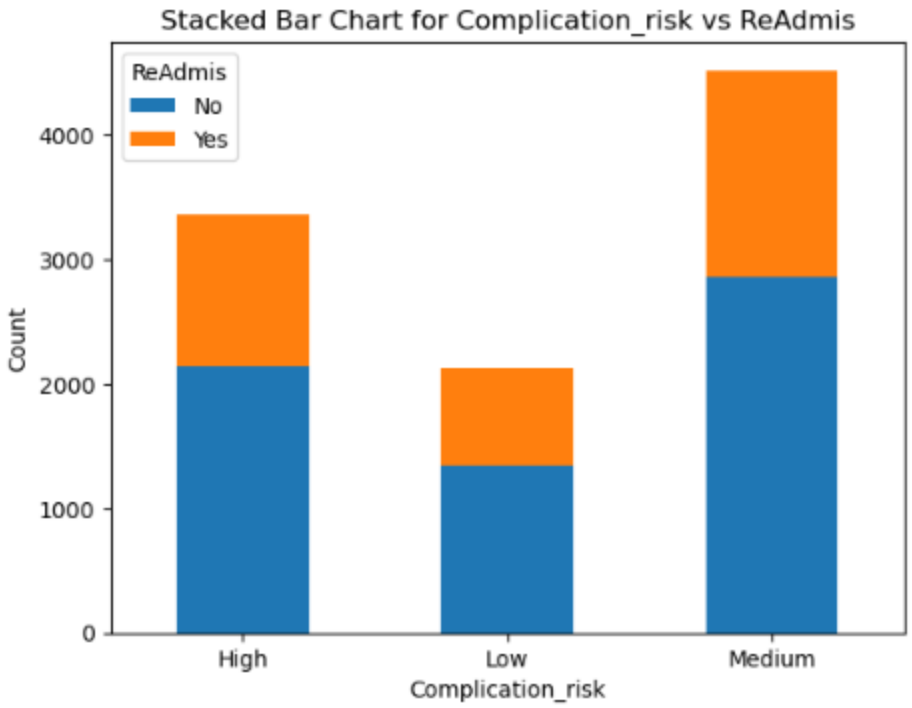

Stacked Bar Chart for Gender vs ReAdmis
ReAdmis
No
Yes
5000
4000
3000
2000
1000
0
Count
Female
Male
Nonbinary
Gender

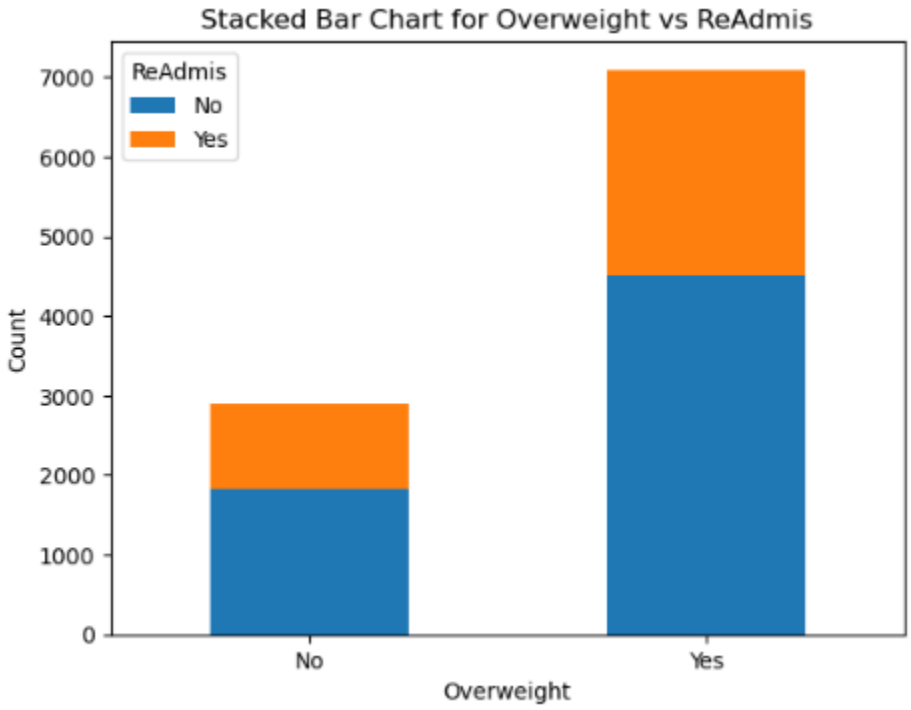

Stacked Bar Chart for Soft_drink vs ReAdmis
ReAdmis
No
Yes
7000
6000
5000
4000
3000
2000
1000
0
Count
No
Yes
Soft_drink

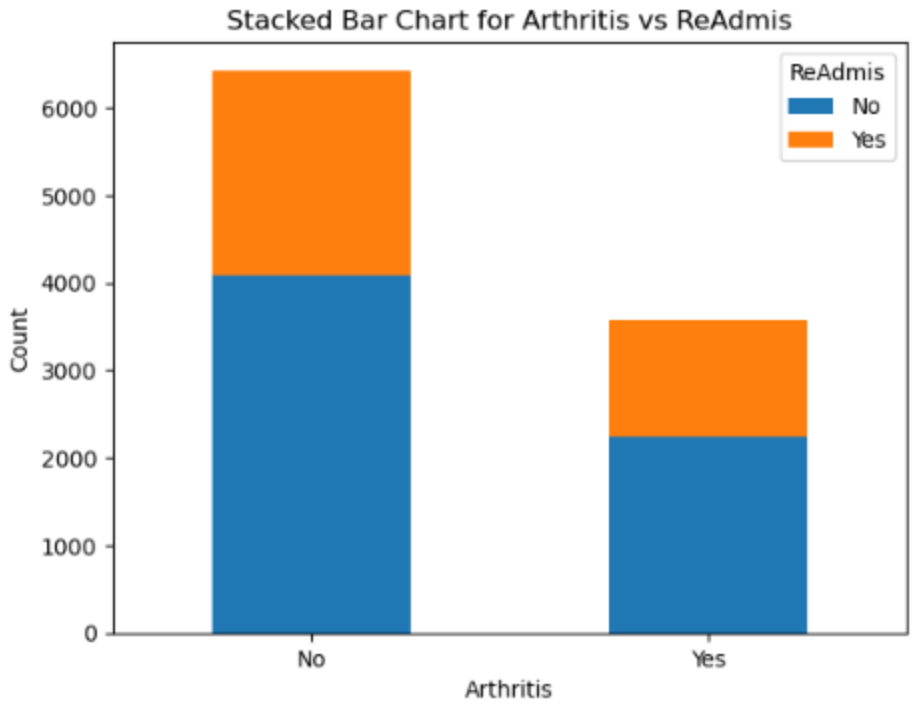

Stacked Bar Chart for HighBlood vs ReAdmis
ReAdmis
No
Yes
6000
5000
4000
3000
2000
1000
0
Count
No
Yes
HighBlood

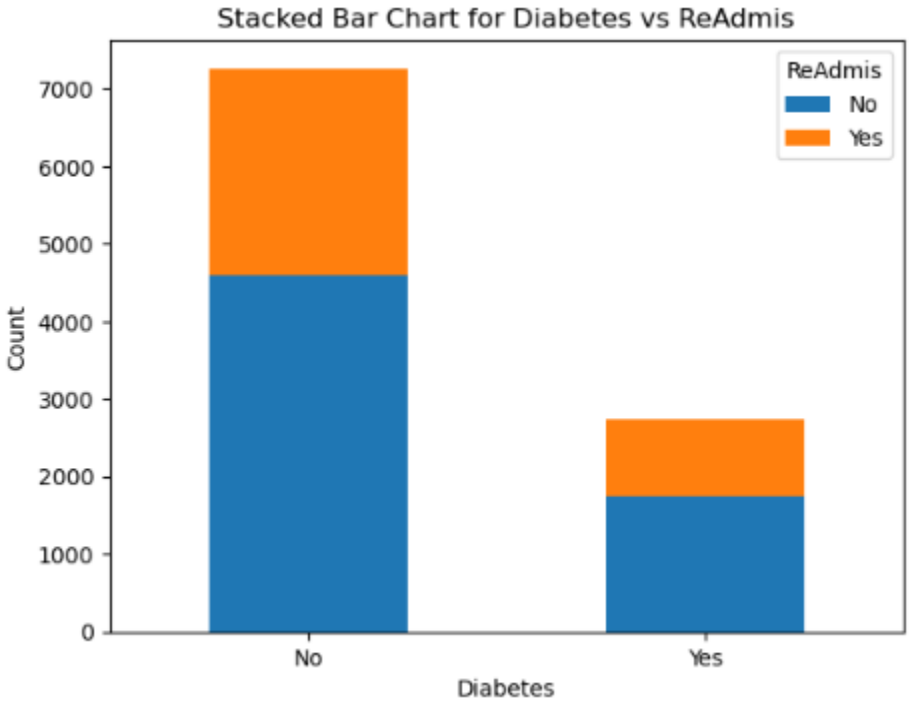

Stacked Bar Chart for Stroke vs ReAdmis
ReAdmis
No
Yes
8000
7000
6000
5000
4000
3000
2000
1000
0
Count
No
Yes
Stroke

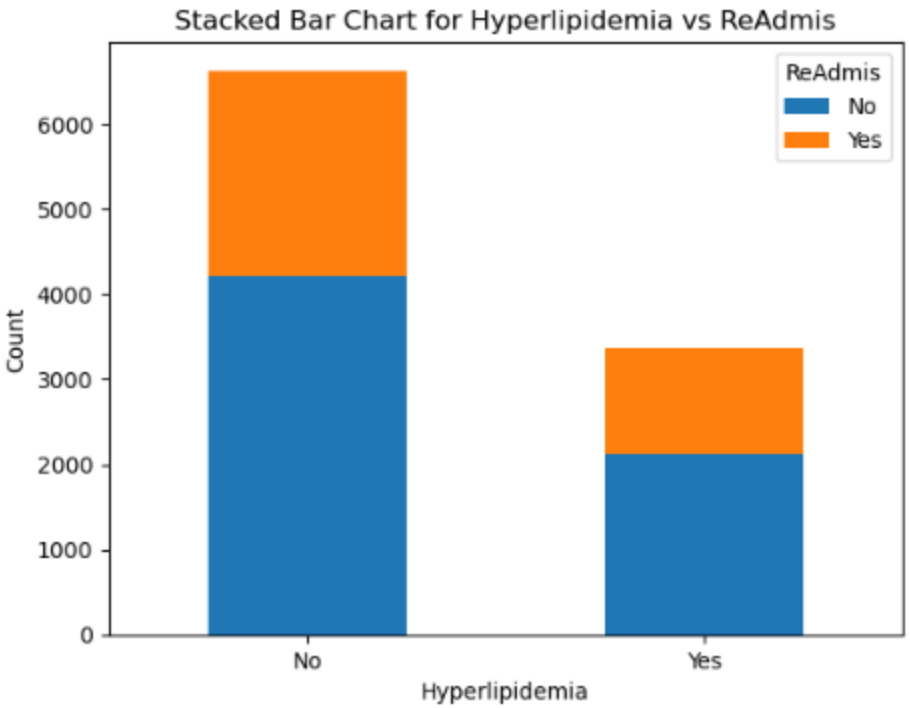
Stroke

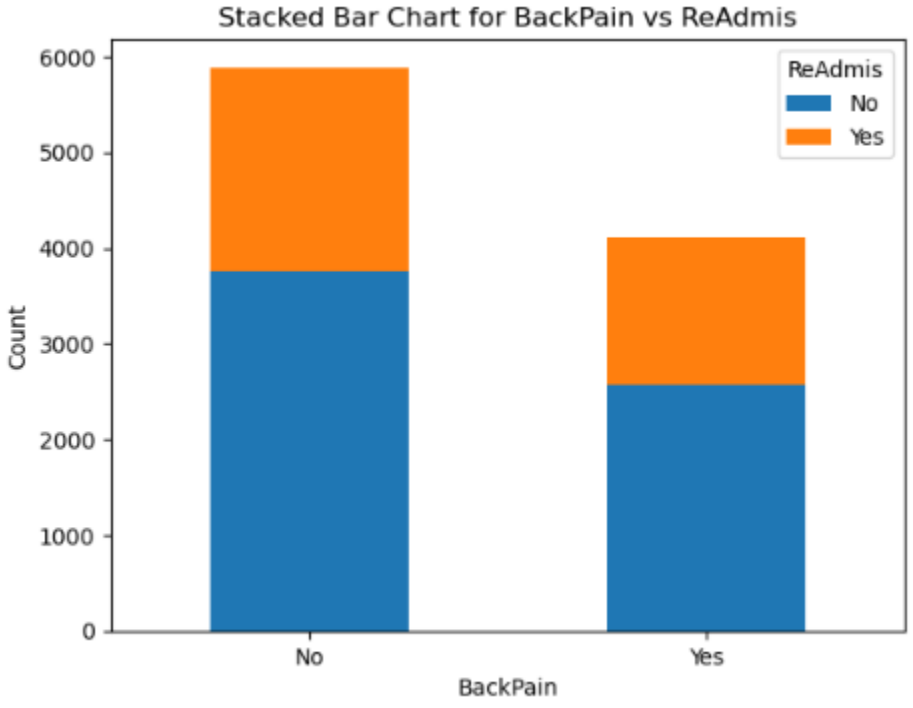

Stacked Bar Chart for Complication_risk vs ReAdmis
ReAdmis
No
Yes
4000
3000
2000
1000
0
Count
High
Low
Medium
Complication_risk

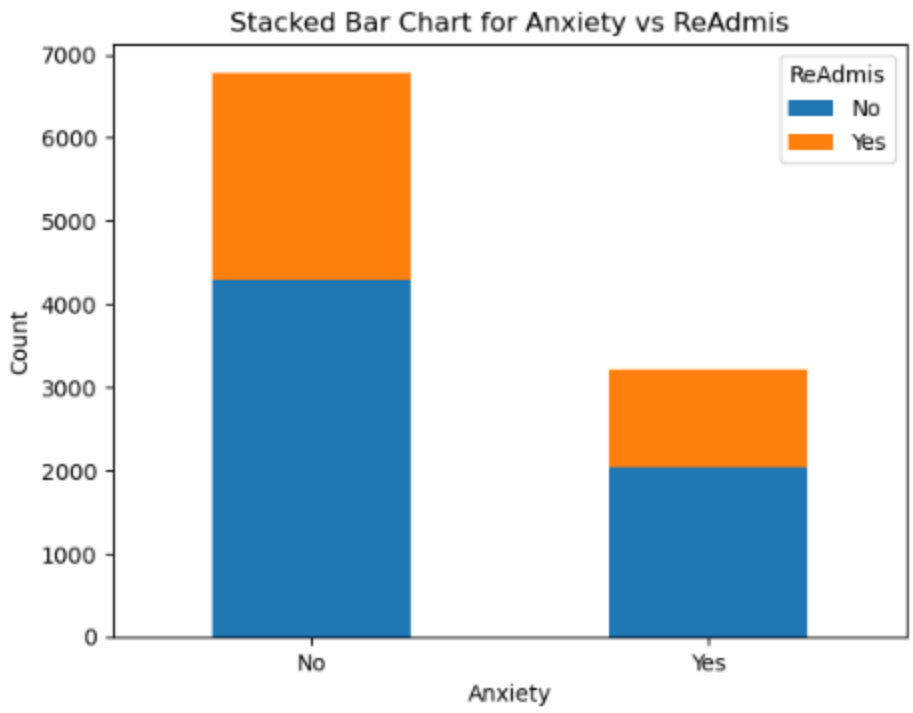

Stacked Bar Chart for Overweight vs ReAdmis
ReAdmis
No
Yes
7000
6000
5000
4000
3000
2000
1000
0
Count
No
Yes
Overweight

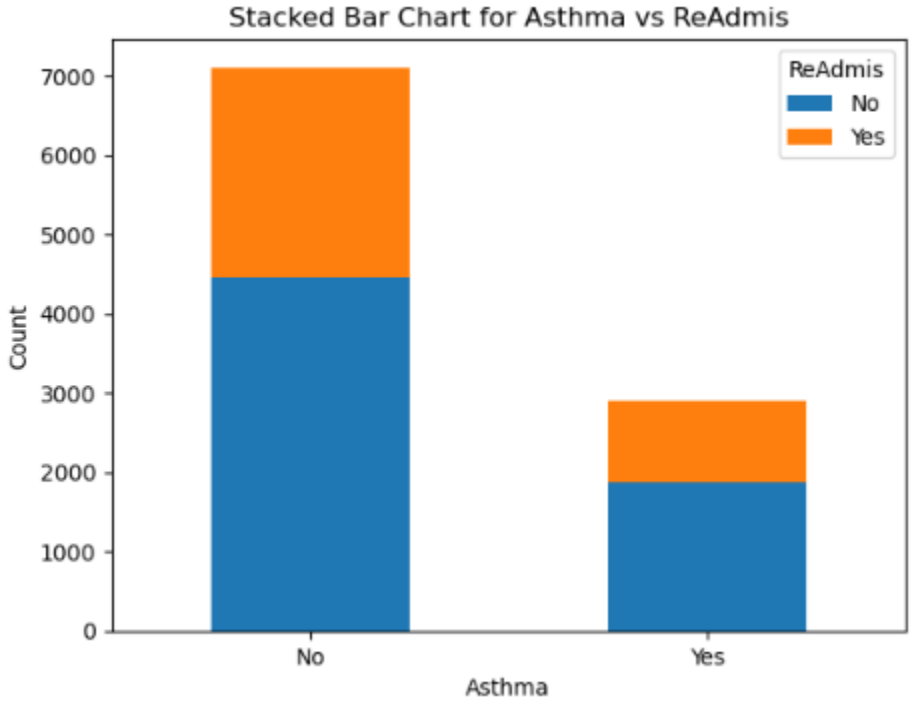

Stacked Bar Chart for Arthritis vs ReAdmis
ReAdmis
No
Yes
6000
5000
4000
3000
2000
1000
0
Count
No
Yes
Arthritis

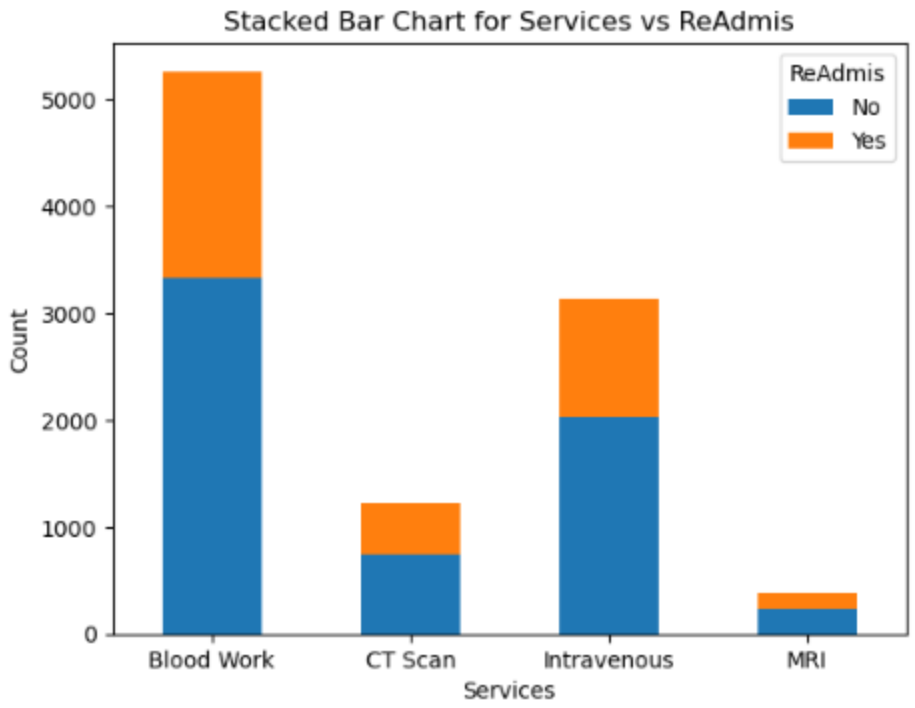

Stacked Bar Chart for Diabetes vs ReAdmis
ReAdmis
No
Yes
7000
6000
5000
4000
3000
2000
1000
0
Count
No
Yes
Diabetes


Stacked Bar Chart for Hyperlipidemia vs ReAdmis
ReAdmis
No
Yes
6000
5000
4000
3000
2000
1000
0
Count
No
Yes
Hyperlipidemia


Stacked Bar Chart for BackPain vs ReAdmis
ReAdmis
No
Yes
6000
5000
4000
3000
2000
1000
0
Count
No
Yes
BackPain


Stacked Bar Chart for Anxiety vs ReAdmis
ReAdmis
No
Yes
Count
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Anxiety


Stacked Bar Chart for Asthma vs ReAdmis
ReAdmis
No
Yes
Count
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Asthma



Data Transformation (Data Wrangling):

Logistic regression models require all variables to be numeric. Therefore, I converted categorical variables into numerical values. For yes/no variables, I employed a loop and the replace() function to transform them into 1/0. For other categorical variables that are non-binary, I utilized the LabelEncoder() function to assign numeric values to each category.

Initial Model:

Before performing multiple linear regression, it is important to check for multicollinearity. Multicollinearity occurs when independent variables, also known as predictor variables, in a regression model, are highly correlated, which can distort the results of the model and make it difficult to determine the exact variable responsible for the change in the dependent variable. To detect multicollinearity, I utilized the Variance Inflation Factor (VIF) method. VIF measures the extent of variance inflation due to multicollinearity among the predictor variables in the model (The Investopedia Team). A VIF value greater than 5 indicates multicollinearity; however, since all VIF scores for the independent variables used in the model are below 5, there is no multicollinearity issue.

	Variable	VIF
0	Children	1.860855
1	Gender	1.798041
2	Soft_drink	1.320949
3	HighBlood	1.618029
4	Stroke	1.223859
5	Complication_risk	2.327080
6	Overweight	2.834397
7	Arthritis	1.504382
8	Diabetes	1.345838
9	Hyperlipidemia	1.452097
10	BackPain	1.618195
11	Anxiety	1.427927
12	Asthma	1.372286
13	Services	1.660538
14	Initial_days	2.383951

Having verified the absence of multicollinearity, the next step is to develop the initial model. This model will encompass all the variables previously mentioned. Below are the Logit Regression Results.

Logit Regression Results

Dep. Variable:	ReAdmis	No. Observations:	10000
Model:	Logit	Df Residuals:	9984
Method:	MLE	Df Model:	15
Date:	Thu, 06 Feb 2025	Pseudo R-squ.:	0.9356
Time:	12:57:46	Log-Likelihood:	-422.96
converged:	True	LL-Null:	-6572.9
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-63.5858	3.164	-20.099	0.000	-69.786	-57.385
Children	0.0836	0.041	2.018	0.044	0.002	0.165
Gender	0.0809	0.165	0.492	0.623	-0.242	0.403
Soft_drink	0.0701	0.207	0.338	0.735	-0.336	0.477
HighBlood	0.7521	0.187	4.020	0.000	0.385	1.119
Stroke	1.4176	0.233	6.097	0.000	0.962	1.873
Complication_risk	-0.0970	0.101	-0.961	0.336	-0.295	0.101
Overweight	-0.1231	0.196	-0.627	0.531	-0.508	0.262
Arthritis	-0.8880	0.191	-4.641	0.000	-1.263	-0.513
Diabetes	0.4397	0.199	2.205	0.027	0.049	0.830
Hyperlipidemia	0.3957	0.189	2.091	0.037	0.025	0.767
BackPain	0.3184	0.180	1.765	0.078	-0.035	0.672
Anxiety	-0.7735	0.193	-4.018	0.000	-1.151	-0.396
Asthma	-0.9691	0.199	-4.878	0.000	-1.358	-0.580
Services	0.1668	0.088	1.889	0.059	-0.006	0.340
Initial_days	1.1650	0.058	20.140	0.000	1.052	1.278

Model Reduction Method:

The feature selection procedure used to reduce the initial model is the Backward Stepwise Elimination procedure with an alpha-to-remove value set at 0.02. Backward Stepwise Elimination is a variable elimination approach that begins with a full model and gradually removes variables based on the p-value. For each step, the alpha-to-remove value is used to determine which variables to eliminate. This method helps ensure that only the most significant predictors remain in the final model, therefore enhancing its accuracy and interpretability. Additionally, Backward Stepwise Elimination simplifies the model, improves predictive performance by focusing on the most impactful variables, and provides an automated, unbiased approach to variable selection.

Reduced Model:

Logit Regression Results

Dep. Variable:	ReAdmis	No. Observations:	10000
Model:	Logit	Df Residuals:	9993
Method:	MLE	Df Model:	6
Date:	Thu, 06 Feb 2025	Pseudo R-squ.:	0.9339
Time:	12:58:04	Log-Likelihood:	-434.19
converged:	True	LL-Null:	-6572.9
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-61.4344	3.018	-20.354	0.000	-67.350	-55.519
HighBlood	0.7271	0.182	3.986	0.000	0.370	1.085
Stroke	1.2998	0.226	5.757	0.000	0.857	1.742
Arthritis	-0.8475	0.188	-4.518	0.000	-1.215	-0.480
Anxiety	-0.7333	0.188	-3.895	0.000	-1.102	-0.364
Asthma	-1.0116	0.196	-5.169	0.000	-1.395	-0.628
Initial_days	1.1363	0.056	20.389	0.000	1.027	1.246

Model Comparison:

The model evaluation metric employed to compare the initial model to the reduced model is the Bayesian Information Criterion (BIC) metric. BIC is a comprehensive metric that accounts for both goodness-of-fit and the complexity of the model (Kumarl, 2023). Lower BIC values represent a better fit. The initial model has a BIC of 996.55, whereas the reduced model has a BIC of 935.62. The lower BIC of the reduced model indicates a better fit with fewer predictor variables. This suggests that

the reduced model is the preferred choice, as it mitigates overfitting and decreases the inclusion of unnecessary variables in the analysis.

```
Initial Model BIC: 993.2944283817587
Reduced Model BIC: 932.8552190779789
```

Confusion Matrix and Accuracy Calculation:

```
Confusion Matrix:
[[6229.  102.]
 [  91. 3578.]]

Accuracy Score: 0.98
```

Regression Equation and Interpretation of Coefficients:

$$LN\left(\frac{P}{(1-P)}\right) = -61.4344 + 0.7271 + 1.2998 - 0.8475 - 0.7333 - 1.0116 + 1.1363$$

Coefficient Explanations:

- **HighBlood**: the log odds for readmission increase by 0.7271 for patients with high blood pressure, keeping all other variables constant.
- **Stroke**: the log odds for readmission increase by 1.2998 for patients that have had a stroke, keeping all other variables constant.
- **Arthritis**: the log odds for readmission decrease by 0.8475 for patients that have arthritis, keeping all other variables constant.
- **Anxiety**: the log odds for readmission decrease by 0.7333 for patients that have anxiety, keeping all other variables constant.
- **Asthma**: the log odds for readmission decrease by 1.0116 for patients that have asthma, keeping all other variables constant.
- **Initial_days**: the log odds for readmission increase by 1.1363 for every additional day the patient initially spent in the hospital, keeping all other variables constant.

Statistical Significance:

Yes, the model is statistically significant. The reduced model has an LLR p-value of 0.00 which is less than the 0.05 threshold, it can be confidently stated that the predictor variables significantly influence the outcome. This result indicates that the findings are not due to chance or random variation.

Practical Significance:

Yes, the model is practically significant. Several coefficients in the model are substantial enough to be of value. For example, the Initial_days variable has a coefficient of 1.16. This indicates that for each additional day the patient initially spends in the hospital, their log odds for readmission increase by 1.16. Similarly, for patients who have had a stroke, their log odds of readmission increase by 1.41. This information is valuable to the company as it can identify patients who are at higher risk of being readmitted and develop targeted care plans to reduce readmission rates.

Limitations of Methods Used:

While logistic regression is straightforward and easy to understand, it assumes a linear relationship between each explanatory variable and the logit of the response variable. The model may not perform well if the relationship is not linear (GeeksforGeeks, 2025). It is also prone to overfitting if not carefully managed, meaning it memorizes the data set instead of learning patterns. Backward stepwise elimination reduces the number of independent variables by iteratively removing the least significant ones, but it can lead to the exclusion of variables that may be important in combinations, which can result in an overly simplistic model. Z-score capping of outliers is a technique to mitigate the impact of extreme values. However, this can lead to the loss of valuable information by limiting the range of the data. Capping outliers can skew the results of an analysis, which can give an inaccurate representation of the data distribution.

Recommendations:

Based on the results of the logistic regression model, it is recommended that the company develop targeted patient care plans for individuals who are more likely to be readmitted to reduce readmission rates. Additionally, the company should

consider implementing follow-up care for patients who have had a stroke and those who spend longer in the hospital as these patients are more likely to be readmitted.

Code Sources:

Logisticregression. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Sources:

Johnathan, S. (n.d.). *Evaluating a logistic regression and its features: Data Science for Journalism*. investigate.ai: Data Science for Journalists. https://investigate.ai/regression/evaluating-logistic-regressions/

GeeksforGeeks. (2025a, January 2). *Advantages and disadvantages of logistic regression*. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

KumarI, A. (2023, November 10). *AIC & BIC for selecting regression models: Formula, examples*. Analytics Yogi. https://vitalflux.com/aic-vs-bic-for-regression-models-formula-examples/