PART I: RESEARCH QUESTION

A1. RESEARCH QUESTION

What factors contribute to the likelihood of hospital readmission?

A2. GOALS

The goal of this analysis is risk assessment. The organization should understand the factors that contribute to hospital readmission and focus on the high-risk groups to implement targeted interventions.

PART II: METHOD JUSTIFICATION

B1. SUMMARY OF ASSUMPTIONS

There are assumptions to keep in mind for logistic regression.

Firstly, logistic regression assumes that the relationship between the independent variables and the log-odds of the dependent variable is linear. The log-odds should change linearly with each predictor variable. Moreover, there should be little or no multicollinearity among the independent variables. Multicollinearity makes it challenging to isolate the individual effect of each variable. It is also essential that the observations are independent of each other. Dependency affects the probability of an event happening from one observation to another. Lastly, there is an assumption that the data set is a large enough sample size for reliable estimates.

B2. TOOL BENEFITS

For this analysis, I used R. This language has several data-cleaning features and capabilities.

Firstly, R has a vast ecosystem of packages designed explicitly for data cleaning. In this assessment, **naniar**, **dplyr,** and **plyr** packages were used. Naniar was used for finding missing data. Dplyr was used for data manipulation and analysis. Plyr was used to convert and revalue variables. All these packages were essential to accomplish a clean dataset.

Lastly, R has easy-to-use data visualization capabilities. Visualization can aid in identifying outliers, missing values, and patterns between variables.

B3. APPROPRIATE BENEFIT

Logistic regression is an appropriate technique to find my research question for multiple reasons.

Firstly, hospital readmission is a binary outcome. Logistic regression is suitable for modeling binary outcomes and estimating the probability of an event. Moreover, logistic regression provides interpretable results. Coefficients can be exponentiated to obtain the odds ratio needed for my research. Furthermore, logistic regression models the probability of an event occurring. This provides the context for understanding the likelihood of an outcome happening. Lastly, logistic regression is a well-established statistical method. Its results are widely received in various fields.

PART III: DATA PREPARATION

C1. DATA CLEANING

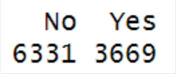
Data cleaning is vital to data preparation. This ensures the quality, accuracy, and reliability of the data used for analysis. Firstly, find missing values. Ignoring missing data can lead to inaccurate conclusions and affect the study's trustworthiness. Then, find duplicates. Duplicates can introduce redundancy to the analysis. It can lead to overestimation of specific trends, affecting insights' reliability. Lastly, count the outliers. Outliers can affect statistical measures. This leads to distorted results, affects data distribution, and leads to misleading conclusions.

To find the missing values, use the **naniar** package. **Naniar** provides valuable functions to visualize and handle missing data. Meanwhile, R makes it easier to find duplicates using **duplicated().** This function identifies duplicate rows. Lastly, outliers can be identified using the interquartile range. Values beyond the lower and upper bound are considered outliers.

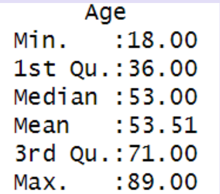
The code for detection is attached.

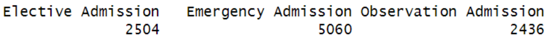
C2. SUMMARY STATISTICS

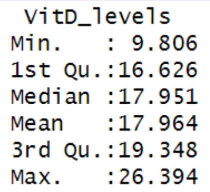
Variable	Data type	Variable type	Summary for Variables

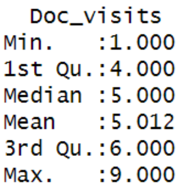
Dependent Variable:
ReAdmis
	Qualitative	Categorical	
			`  No  Yes`
			`6331 3669`

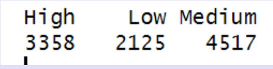
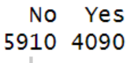
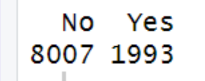
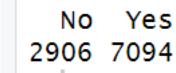
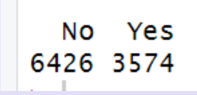
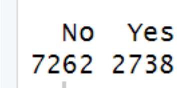
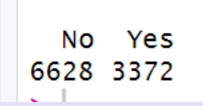
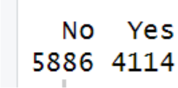
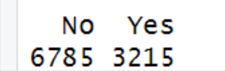
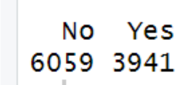
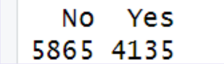
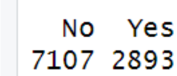
Independent Variables			

Age	Quantitative	Discrete	
			`      Age`
			`Min.   :18.00`
			`1st Qu.:36.00`
			`Median :53.00`
			`Mean   :53.51`
			`3rd Qu.:71.00`
			`Max.   :89.00`

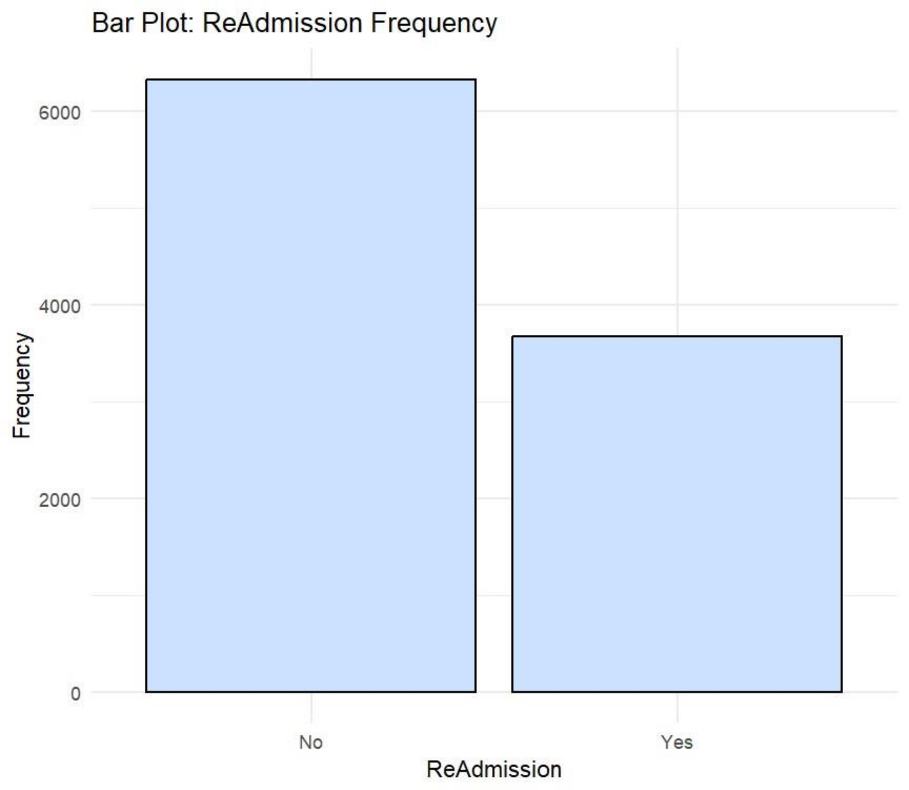
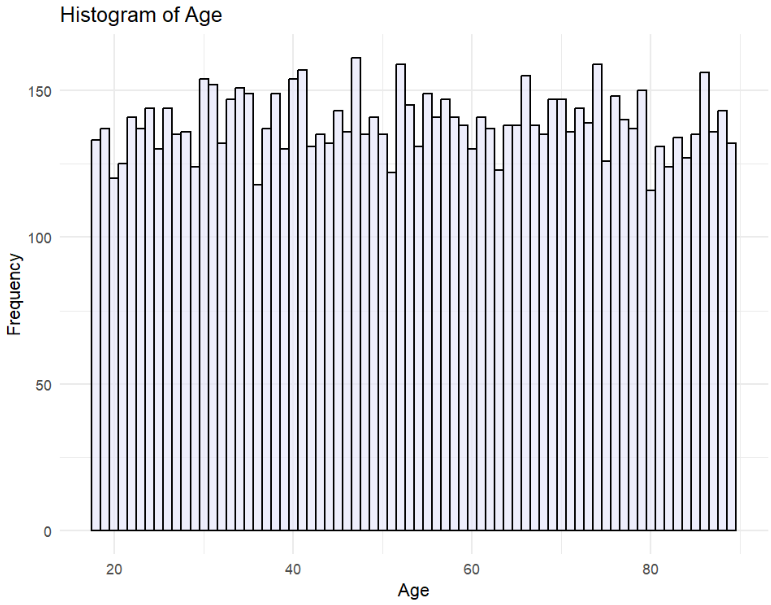
Initial_admin	Qualitative	Categorical	
			`Elective Admission   Emergency Admission Observation Admission`
			`              2504                  5060                  2436`

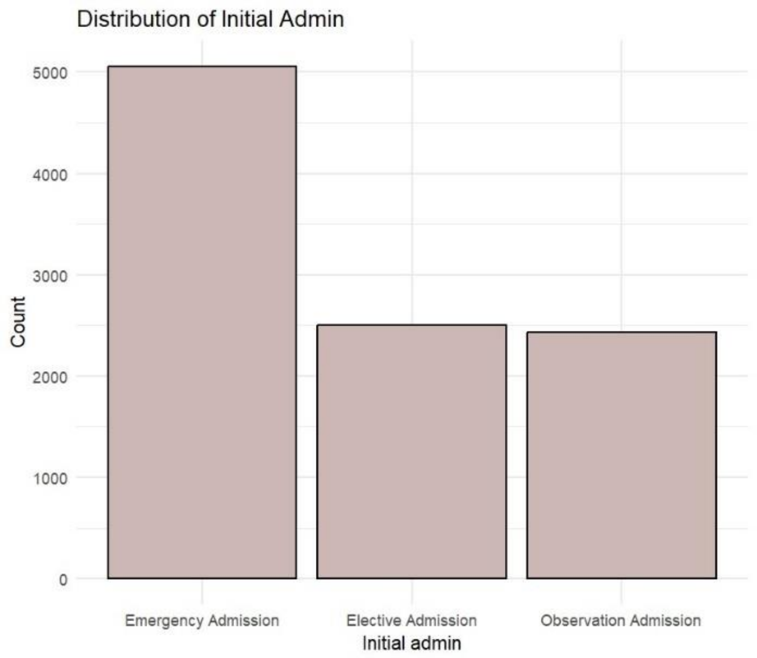
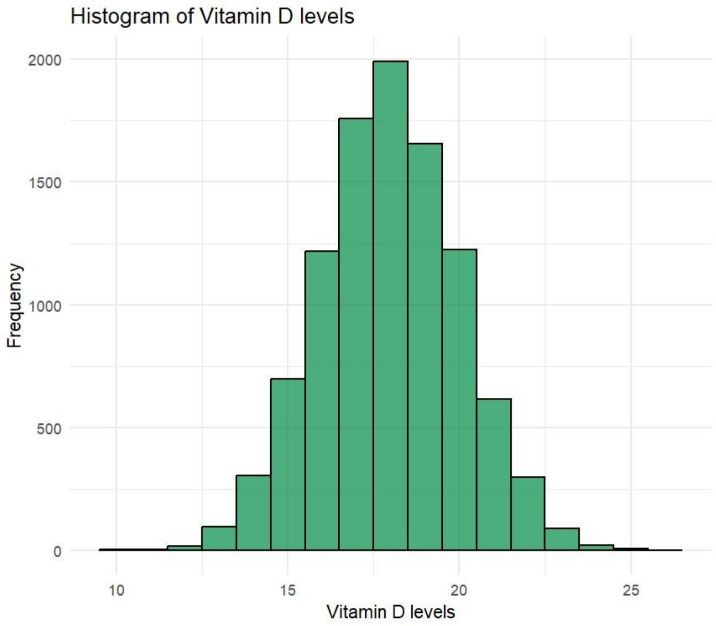
VitD_levels	Quantitative	Continuous	
			` VitD_levels`
			`Min.   : 9.806`
			`1st Qu.:16.626`
			`Median :17.951`
			`Mean   :17.964`
			`3rd Qu.:19.348`
			`Max.   :26.394`

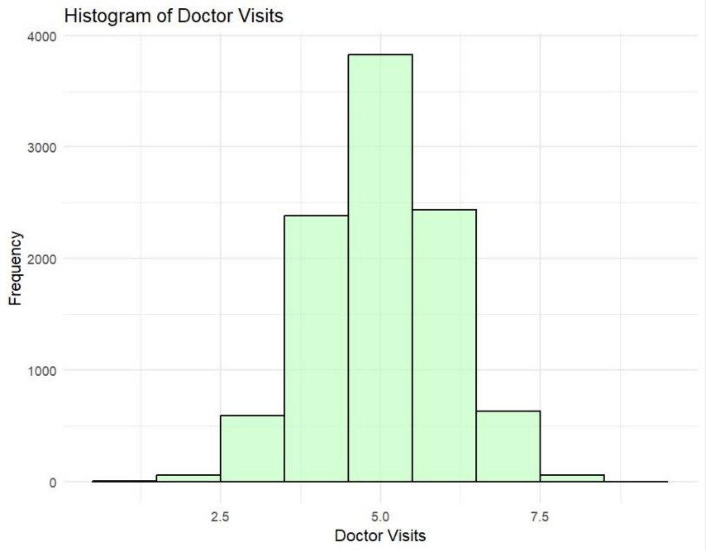
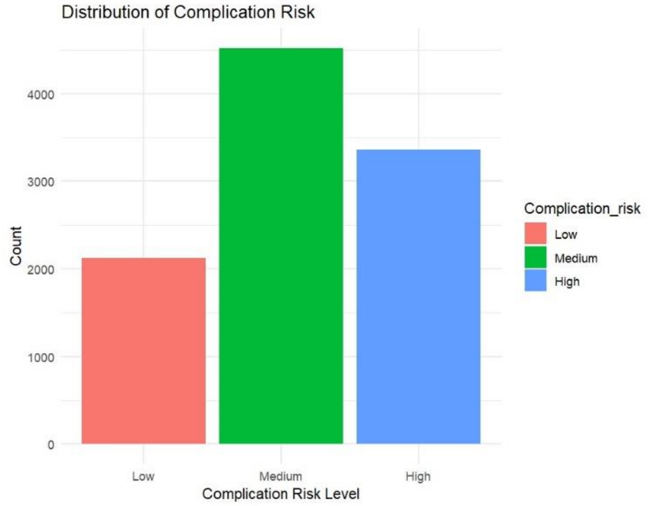
Doc_visits	Quantitative	Discrete	
			`   Doc_visits`
			`Min.   :1.000`
			`1st Qu.:4.000`
			`Median :5.000`
			`Mean   :5.012`
			`3rd Qu.:6.000`
			`Max.   :9.000`

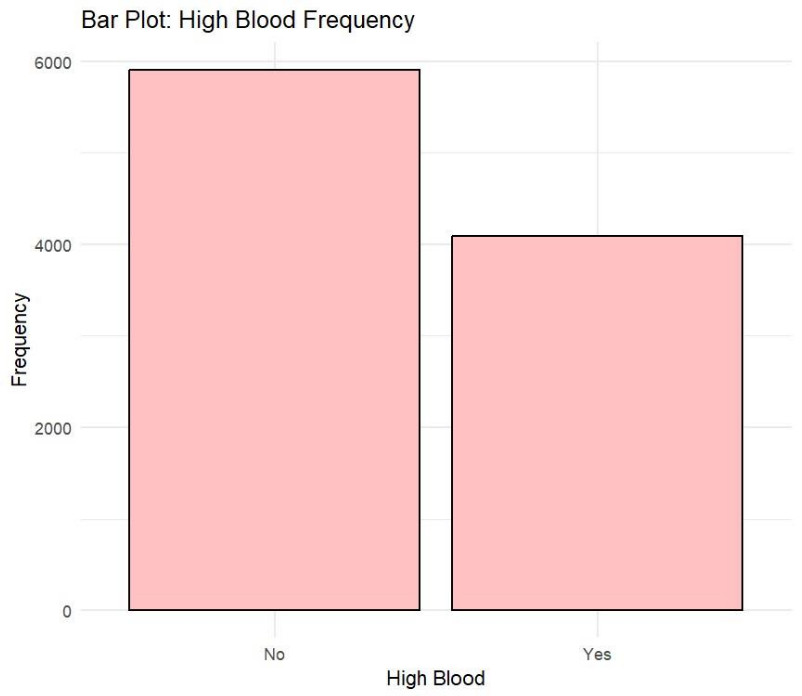
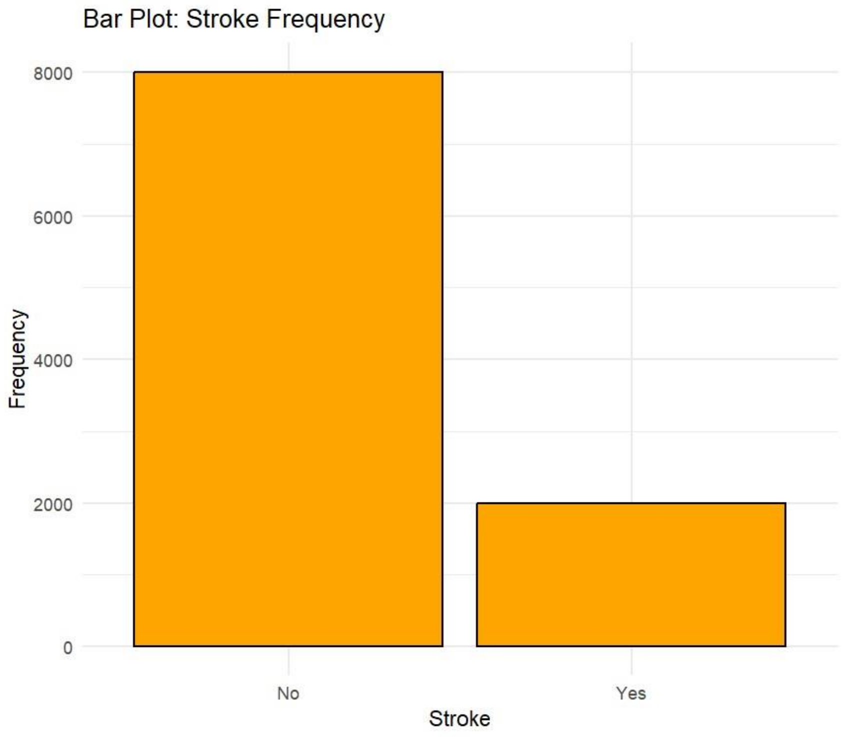
Complication_risk	Qualitative	Categorical	 High Low Medium 3358 2125 4517
HighBlood	Qualitative	Categorical	 No Yes 5910 4090
Stroke	Qualitative	Categorical	 No Yes 8007 1993
Overweight	Qualitative	Categorical	 No Yes 2906 7094
Arthritis	Qualitative	Categorical	 No Yes 6426 3574
Diabetes	Qualitative	Categorical	 No Yes 7262 2738
Hyperlipidemia	Qualitative	Categorical	 No Yes 6628 3372
BackPain	Qualitative	Categorical	 No Yes 5886 4114
Anxiety	Qualitative	Categorical	 No Yes 6785 3215
Allergic_rhinitis	Qualitative	Categorical	 No Yes 6059 3941
Reflux_esophagitis	Qualitative	Categorical	 No Yes 5865 4135
Asthma	Qualitative	Categorical	 No Yes 7107 2893

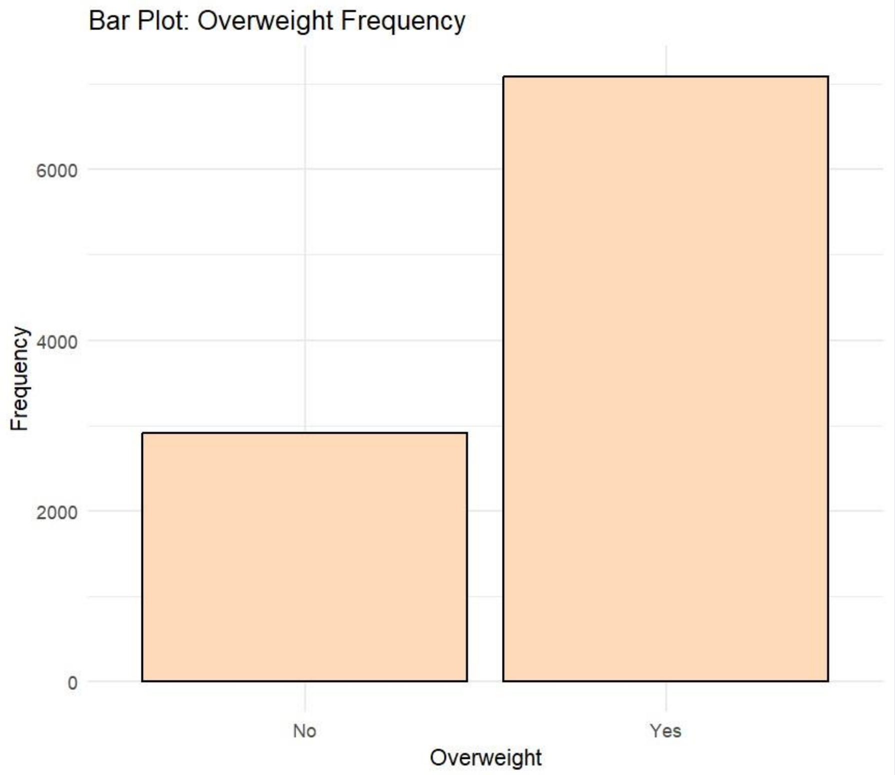
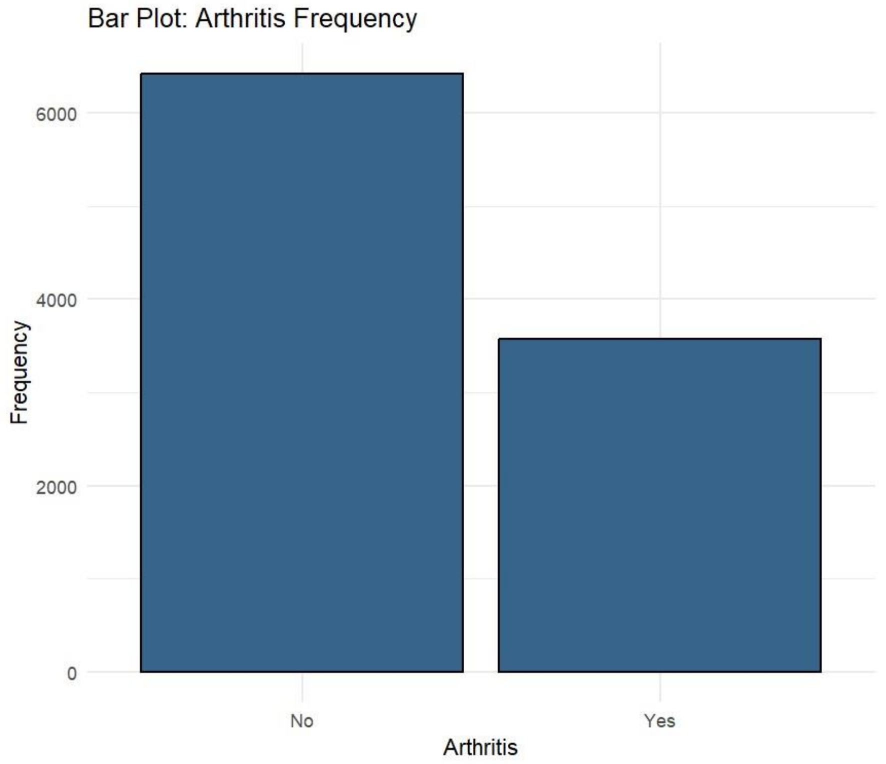
C3. VISUALIZATIONS

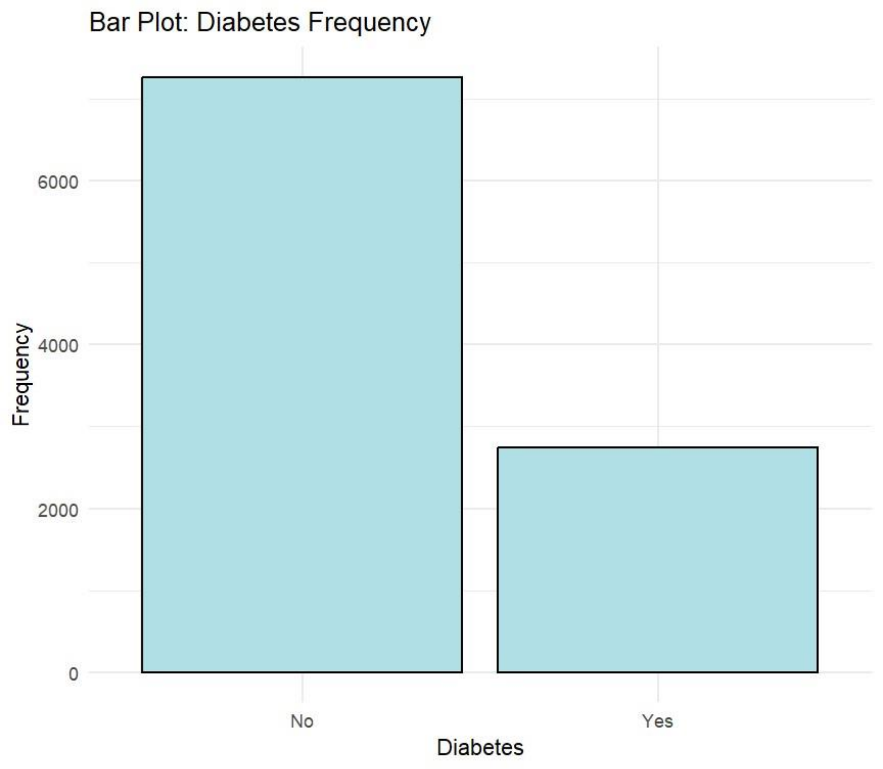
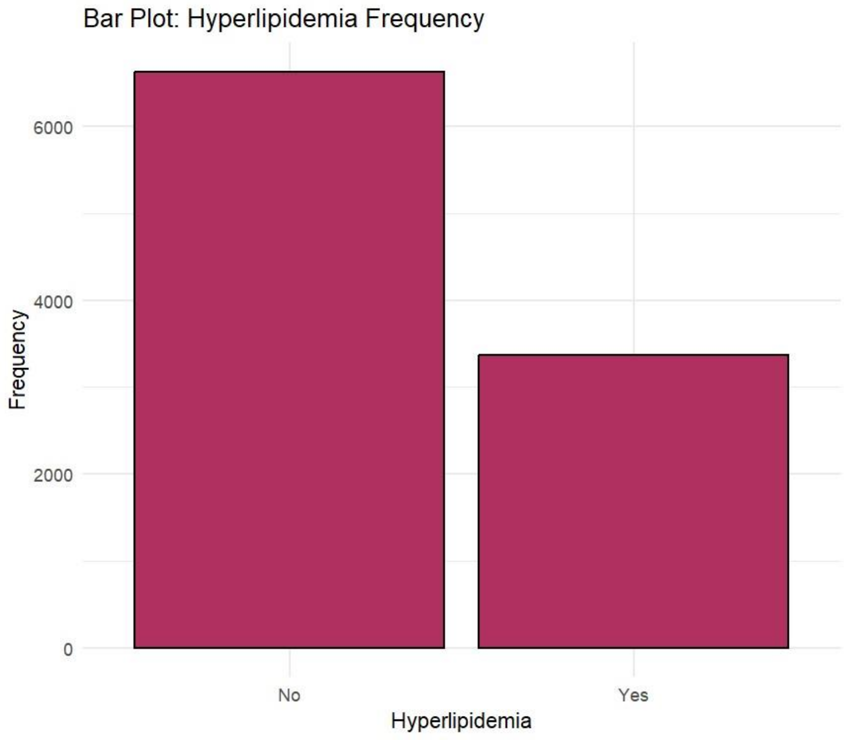
Variables	**Univariate Visualizations**
Dependent Variable: **ReAdmis**	
Age	

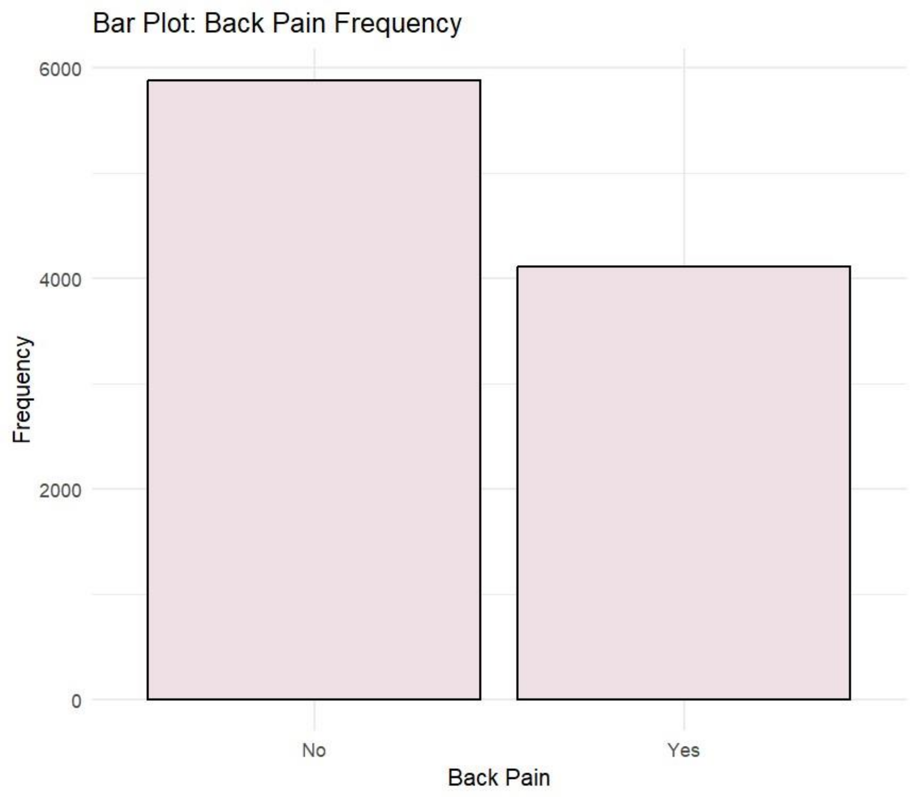
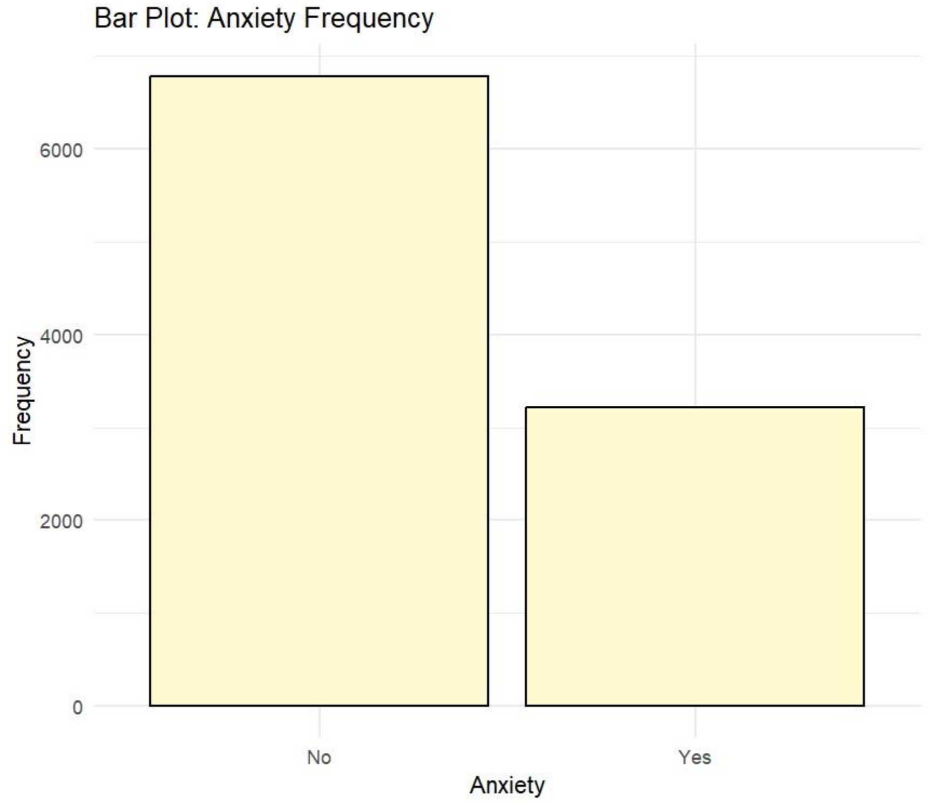
Initial_admin	
VitD_levels	

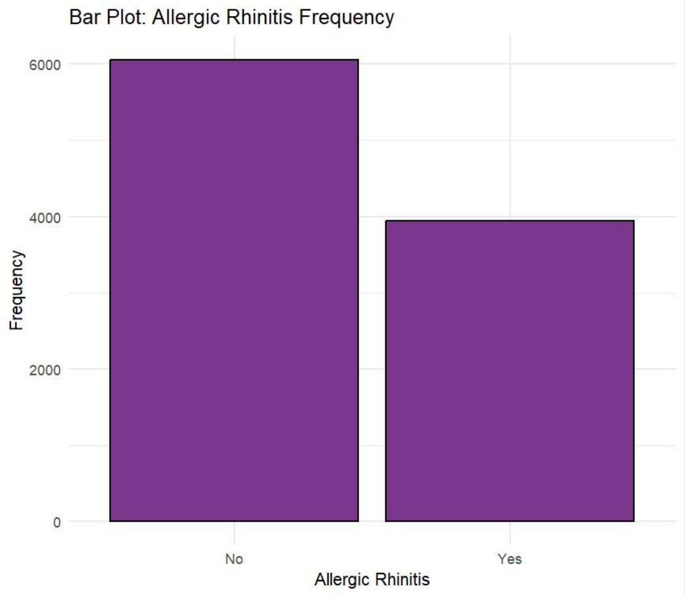
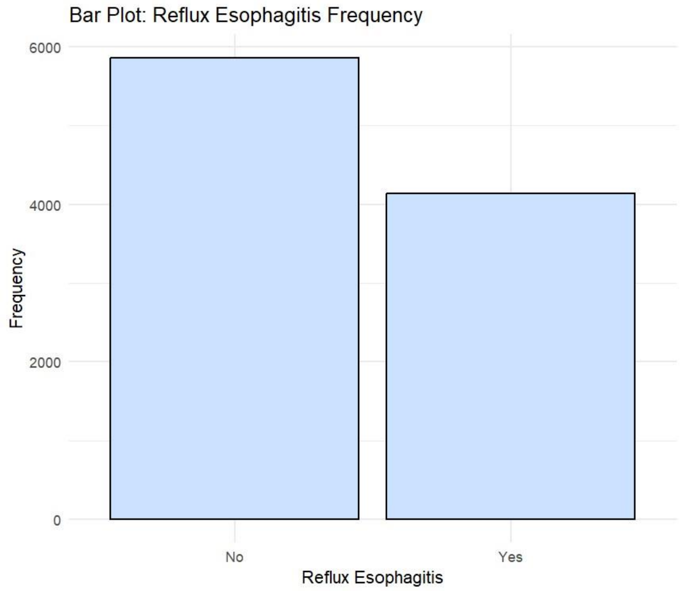
Doc Visits	
Complication Risk	

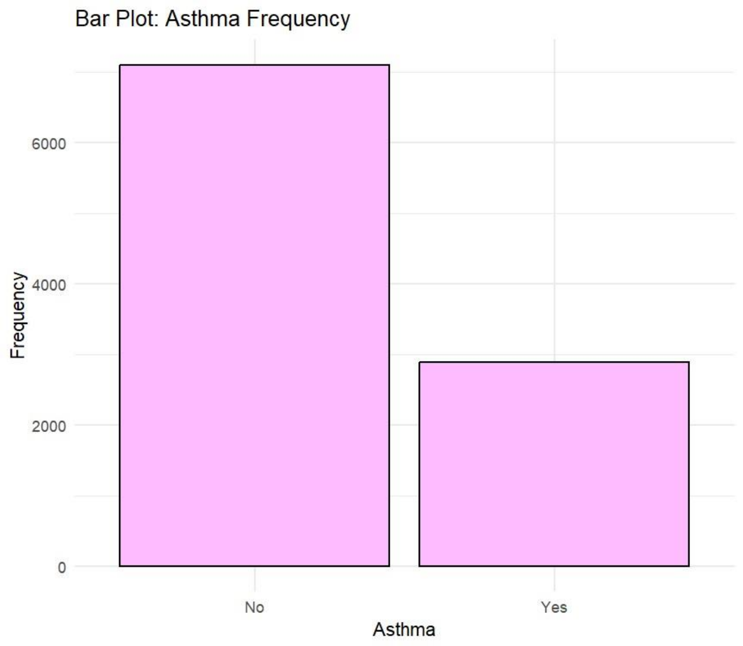
High Blood	
Stroke	

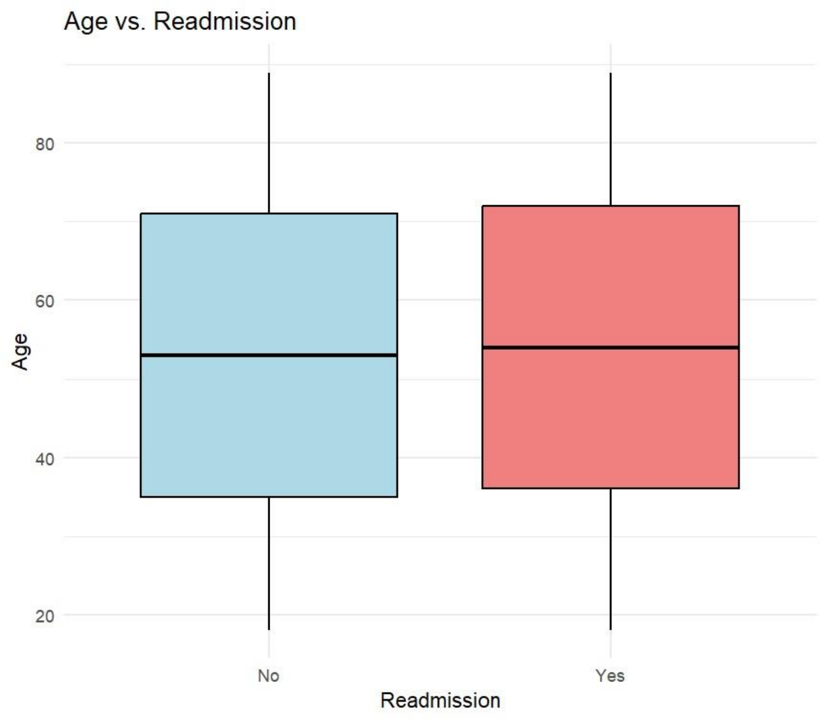
Overweight	
Arthritis	

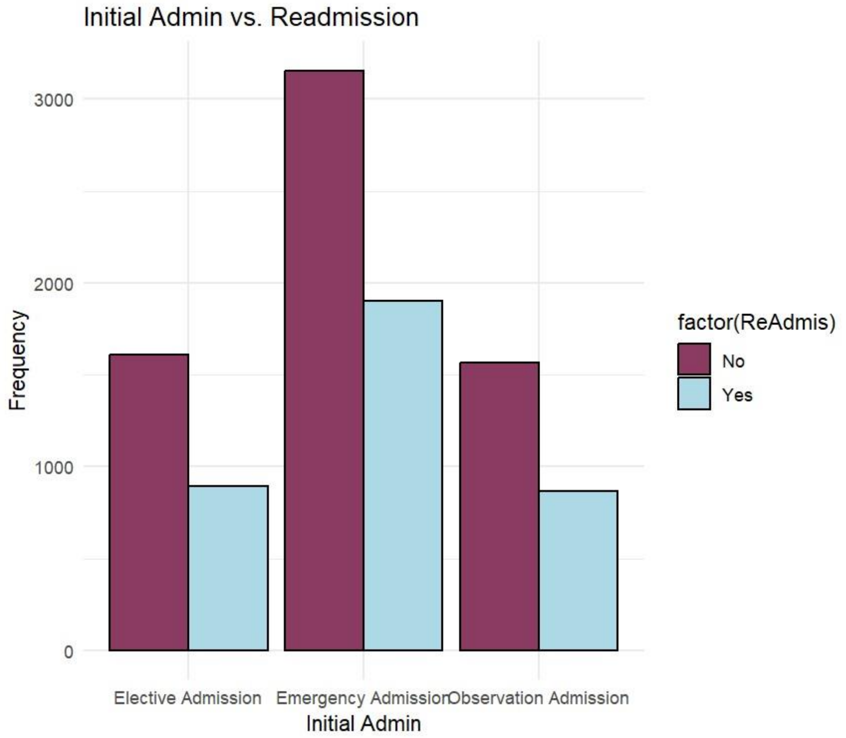
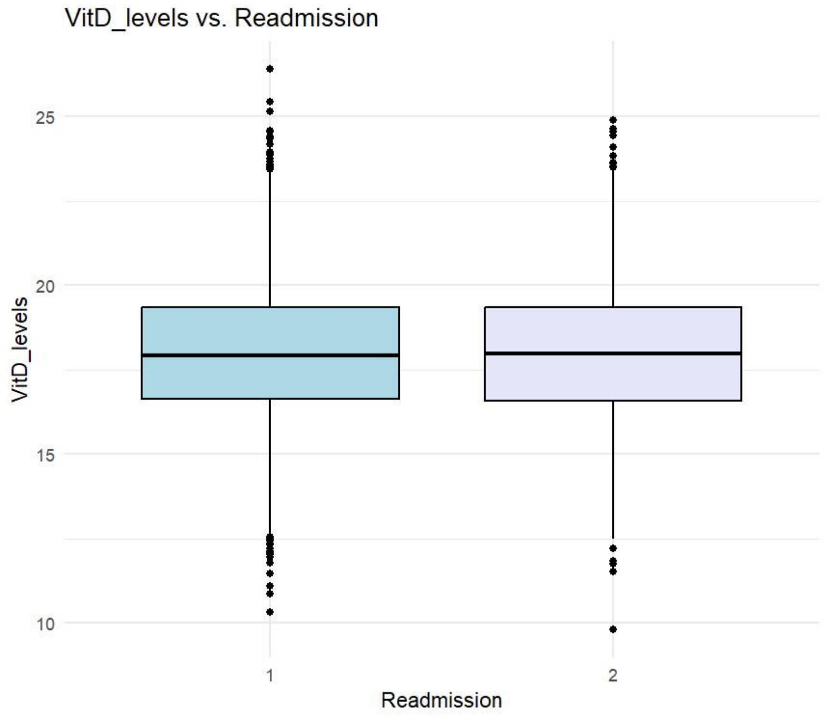
Diabetes	
Hyperlipidemia	

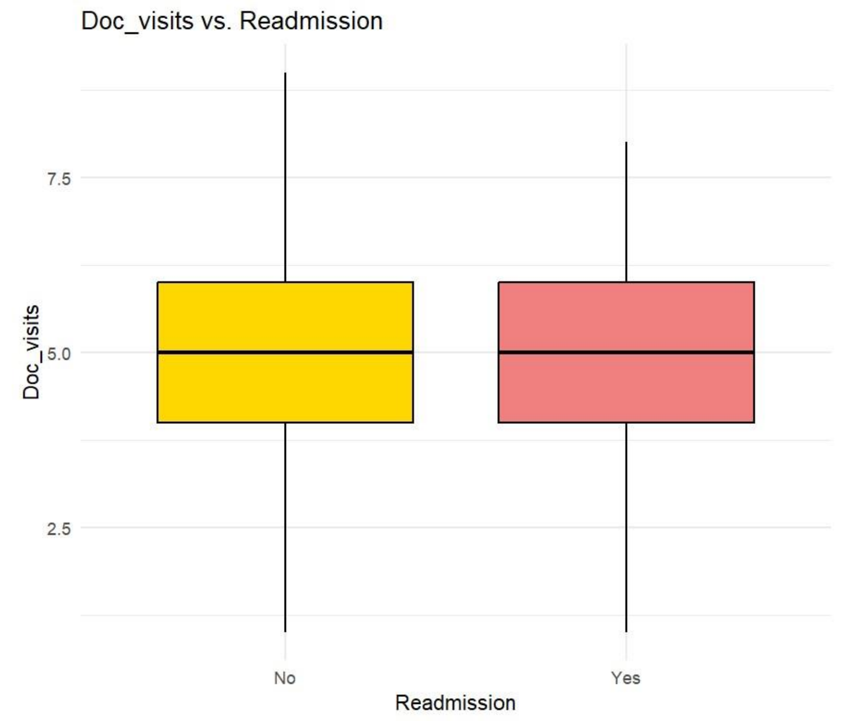
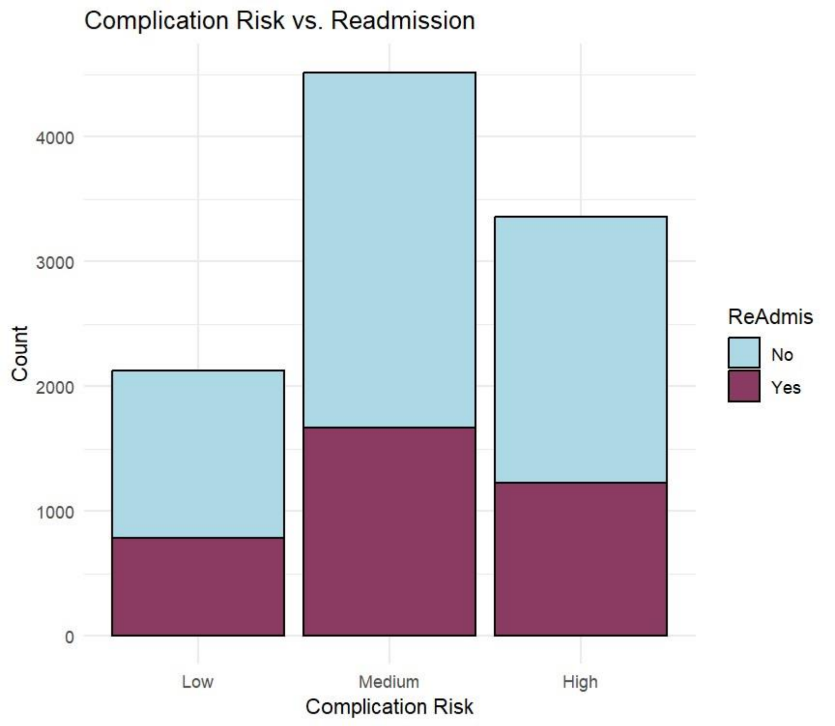
Backpain	
Anxiety	

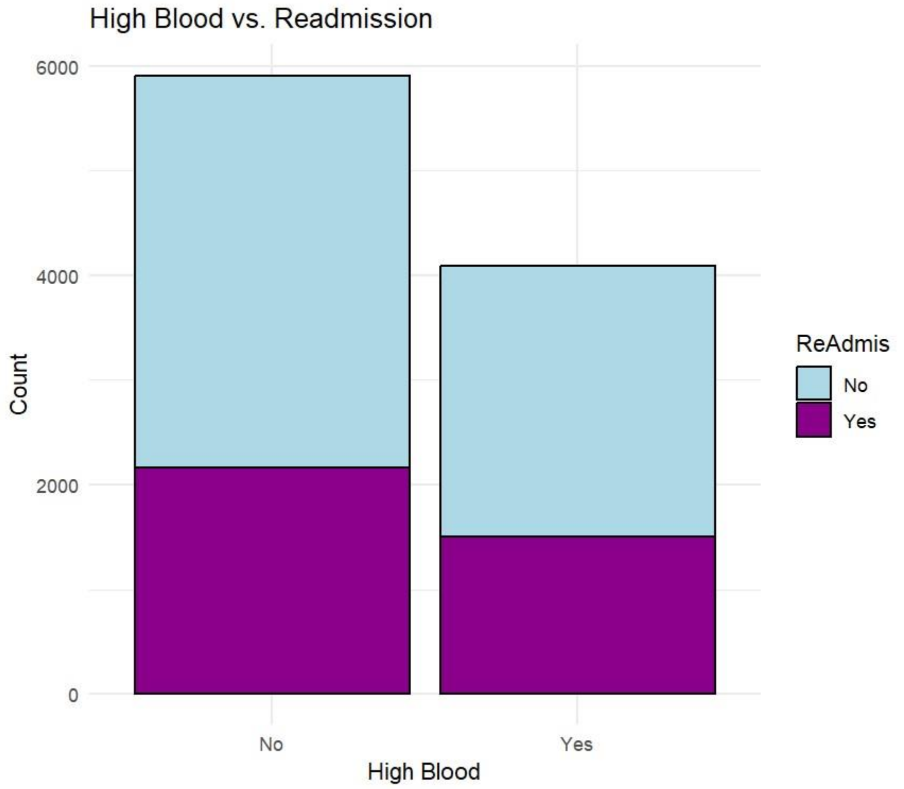
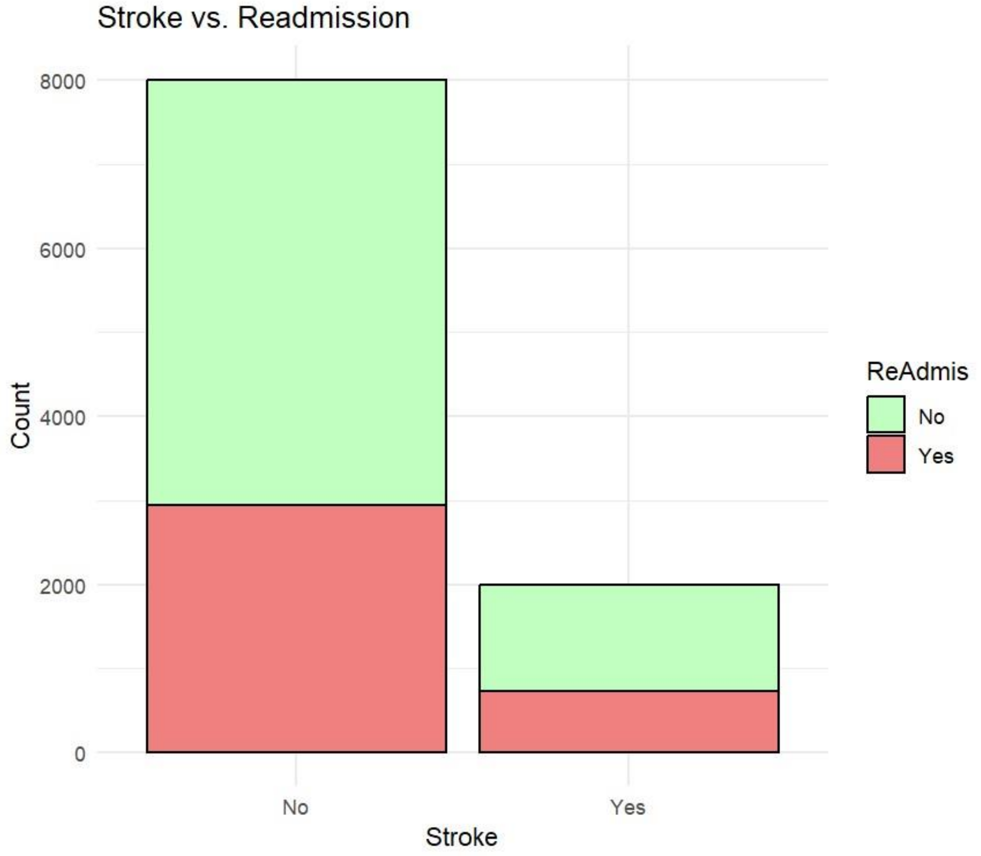
Allergic rhinitis	
Reflux esophagitis	

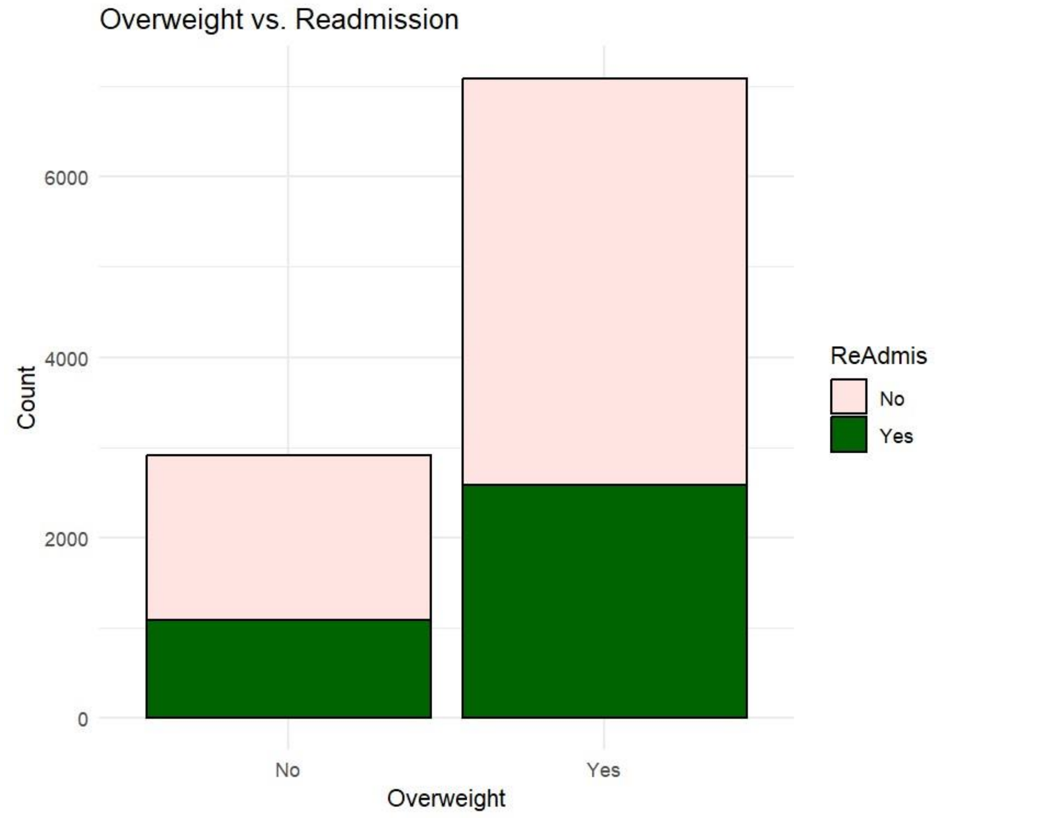
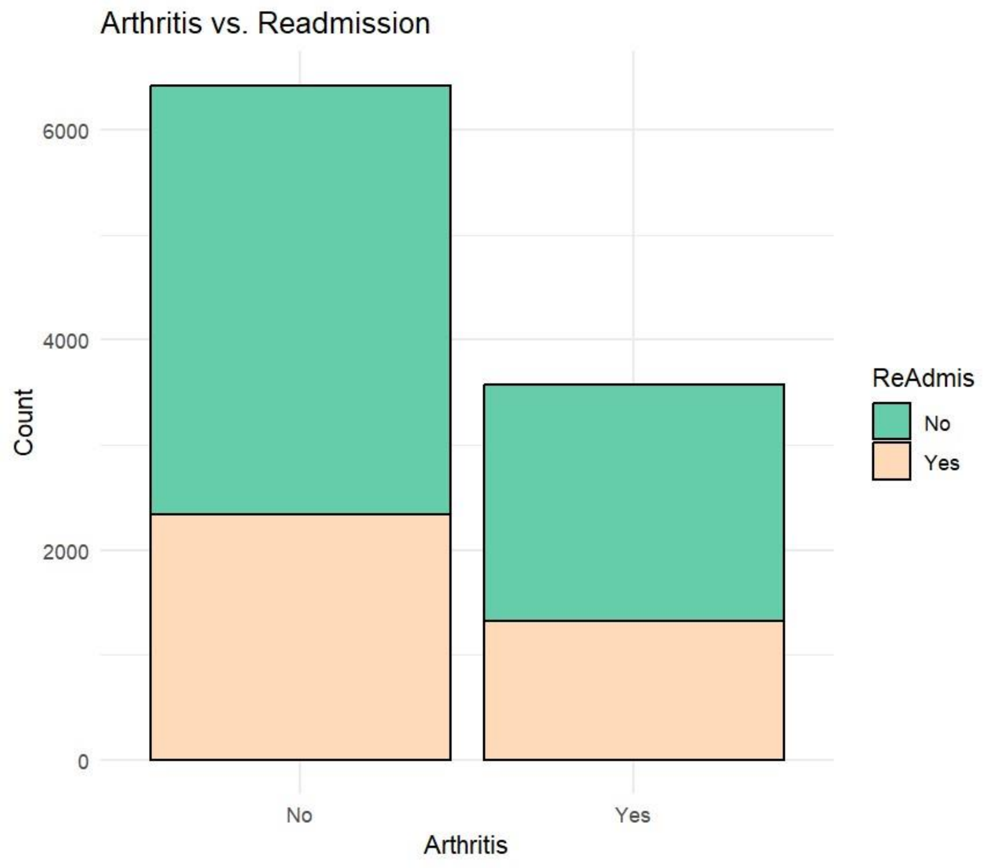
Asthma	

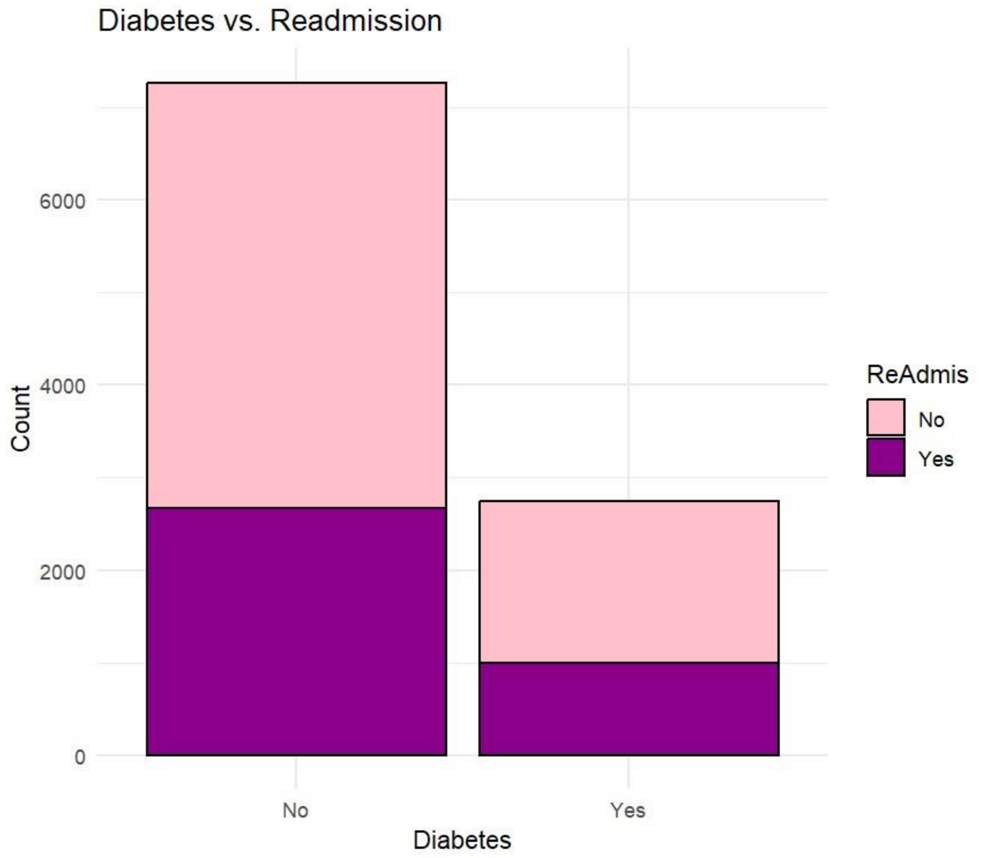
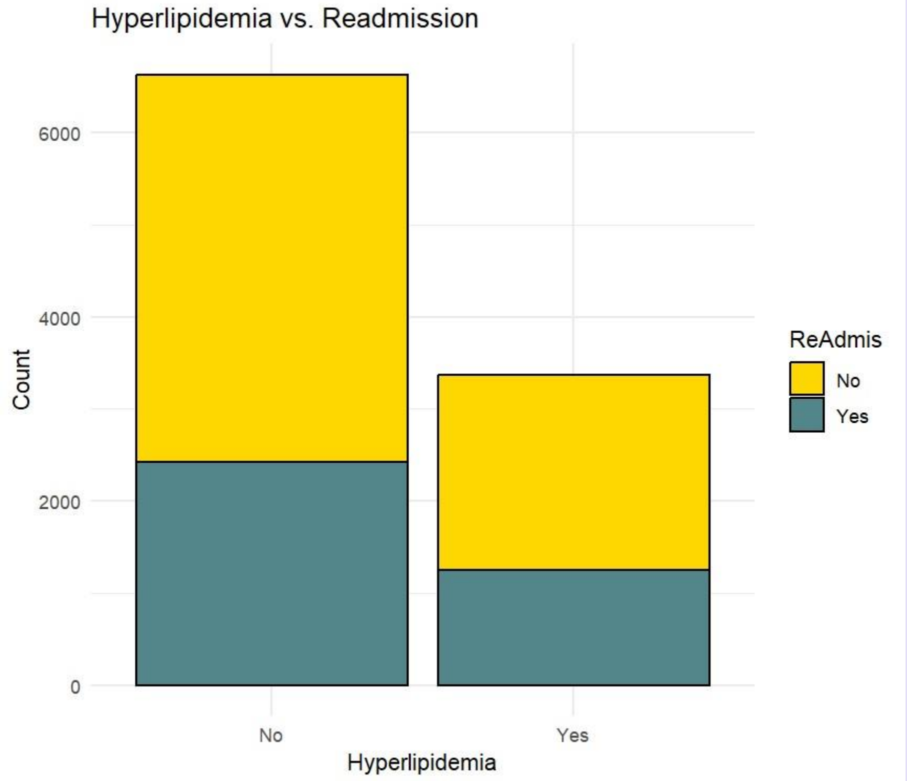
Variables	**Bivariate Visualizations**
Age	

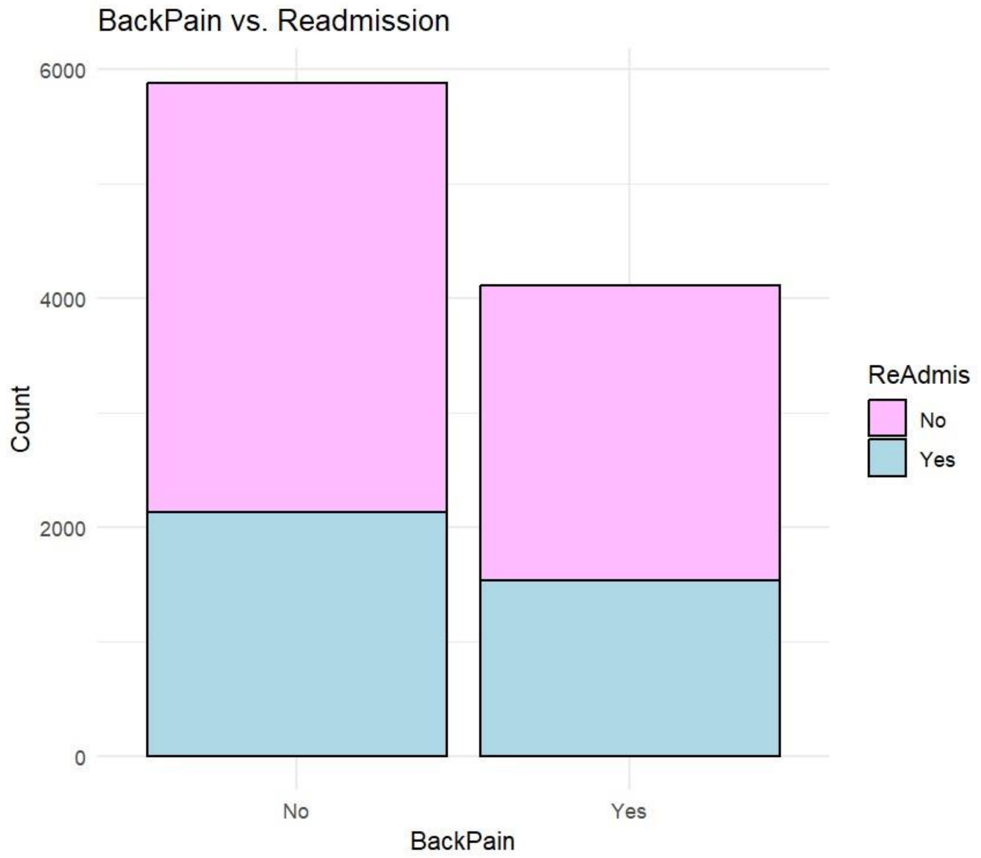
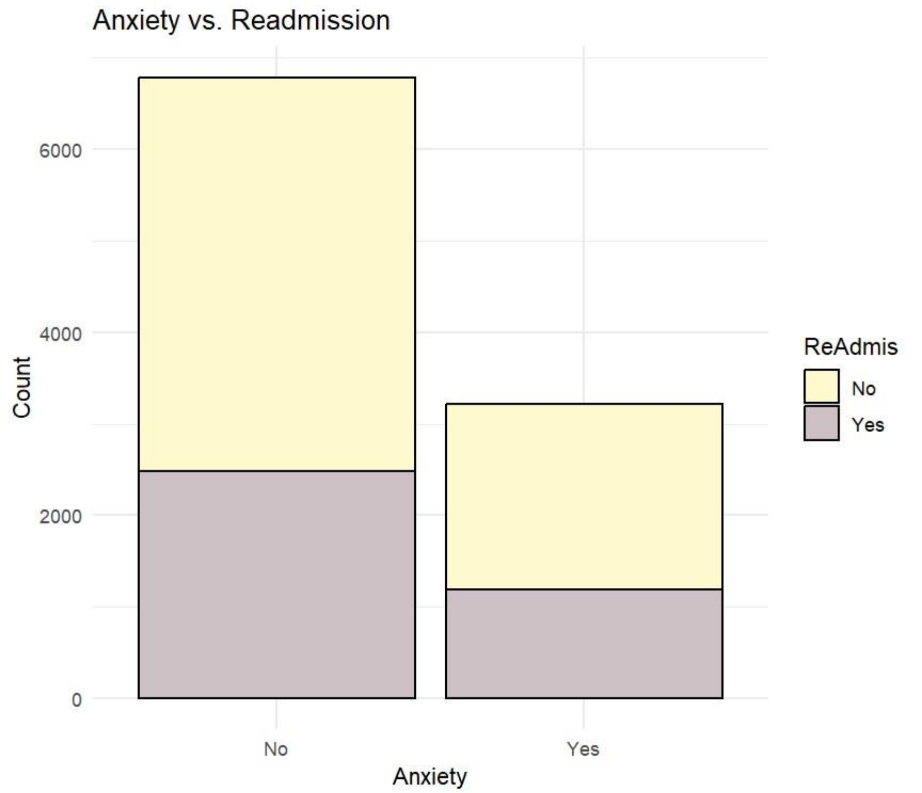
Initial_admin	
VitD_levels	

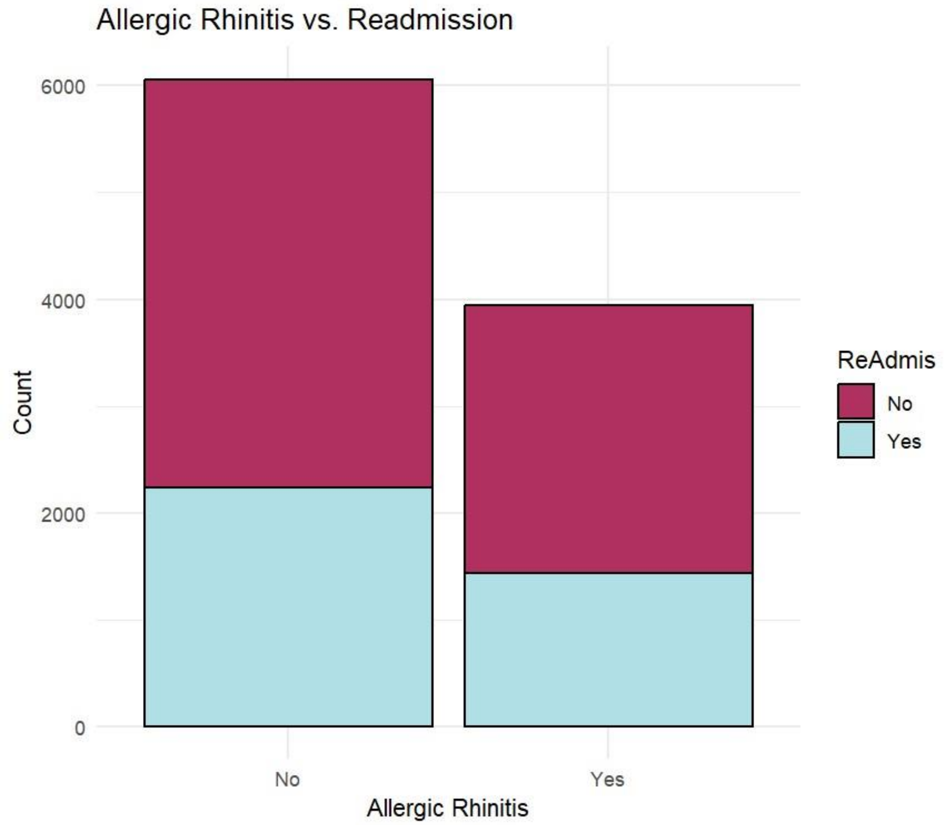
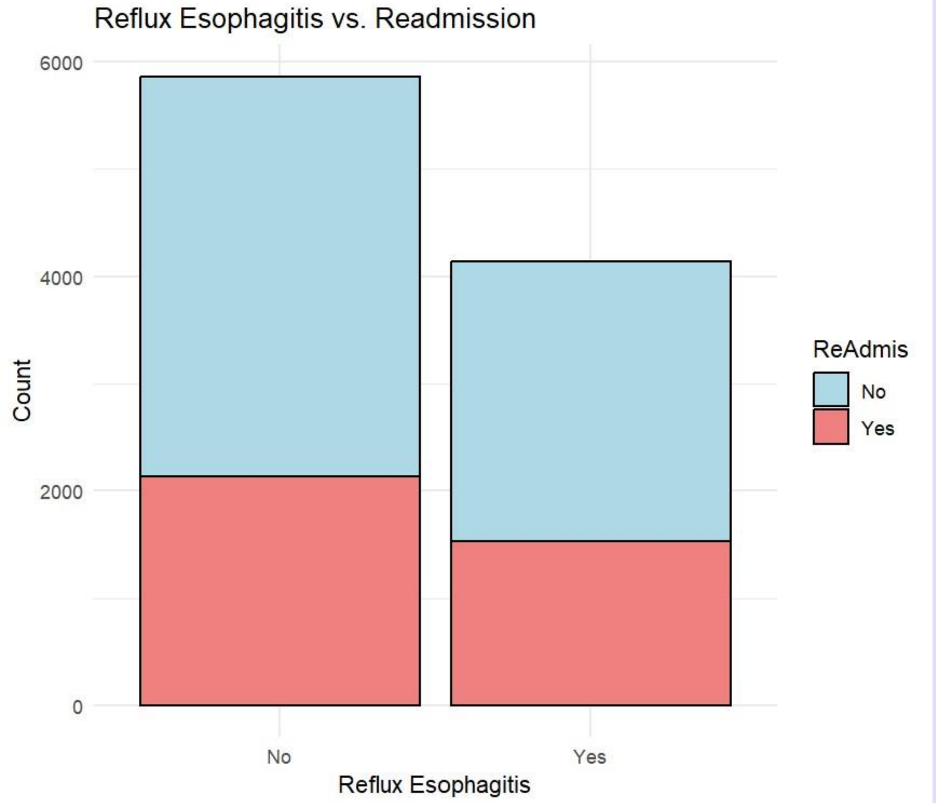
Doc Visits	
Complication Risk	

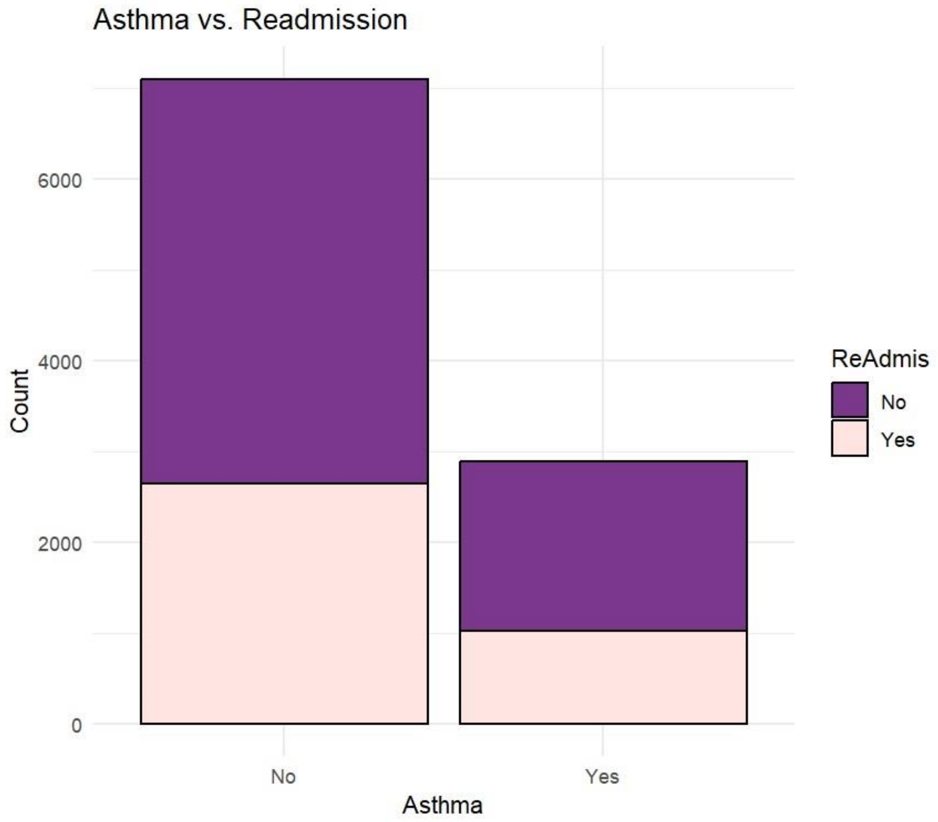
High Blood	
Stroke	

Overweight	
Arthritis	

Diabetes	
Hyperlipidemia	

Backpain	
Anxiety	

Allergic rhinitis	
Reflux esophagitis	

Asthma	

C4. DATA TRANSFORMATION

The data wrangling activities performed on the dataset are converting categorical variables to numeric and factor-level adjustments.

Several categorical variables, such as readmission, high blood, stroke, overweight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, reflux esophagitis, and asthma, were converted into numeric factors. This conversion was achieved using the plyr package with the revalue function. The conversion turned Yes to 1, No to 0, and NA to NA. The variable complication risk was adjusted to Low, Medium, and High levels. The variable initial admin was assigned with Elective Admission, Emergency Admission, and Observation Admission as its factors. This was done with factor().

Performing the conversion and adjustment is essential to answer my research question. Firstly, this ensures consistency in data types and makes it easier to work with the dataset. Numeric representation of categorical variables helps maintain uniformity across the data set. Furthermore, adjusting factor levels is essential for ensuring that categorical variables are treated correctly in statistical models. Overall, data wrangling activities contribute to data quality. Uniform and well-structured data are necessary for obtaining reliable and meaningful insights from analysis.

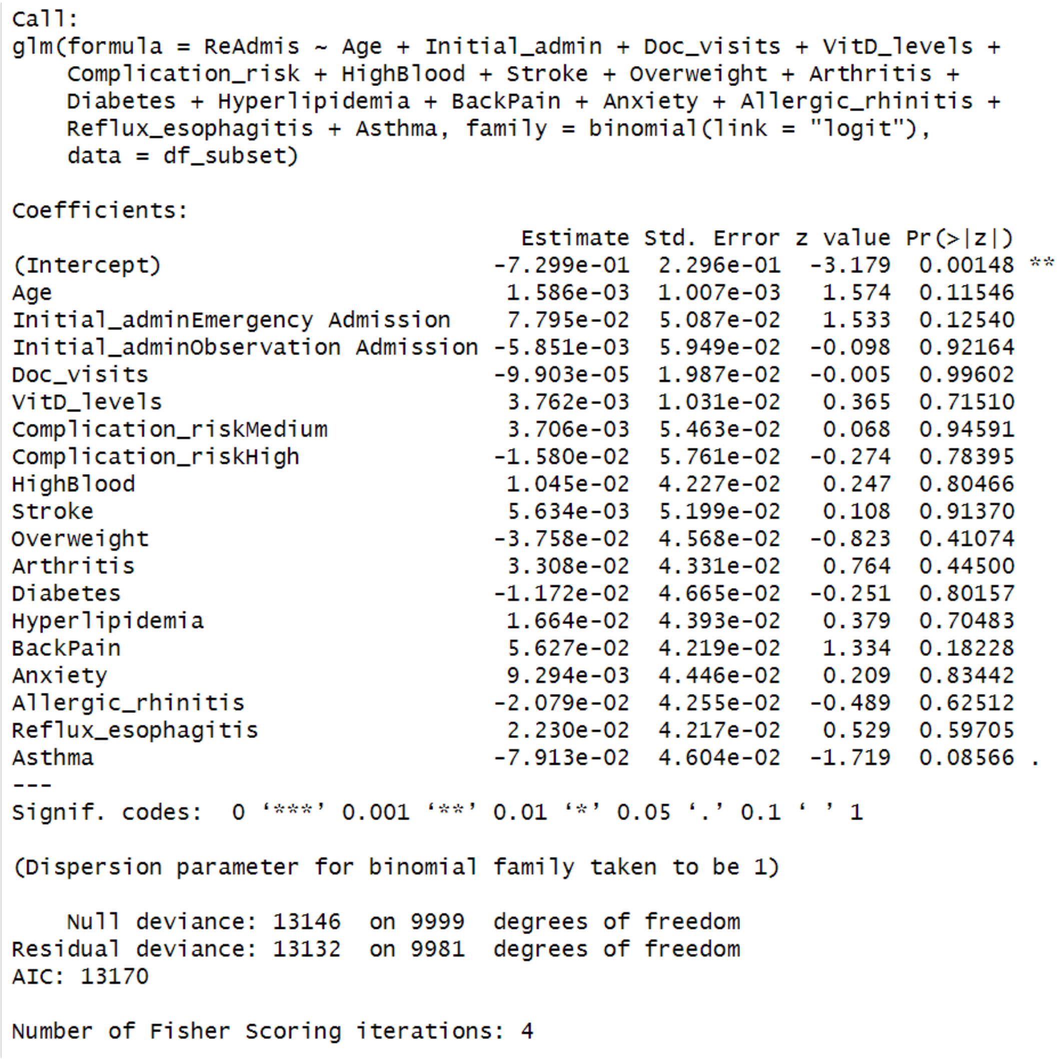
Code for wrangling is attached.

C5. PREPARED DATA SET

.csv file attached

PART IV: MODEL COMPARISON & ANALYSIS

D1. INITIAL MODEL



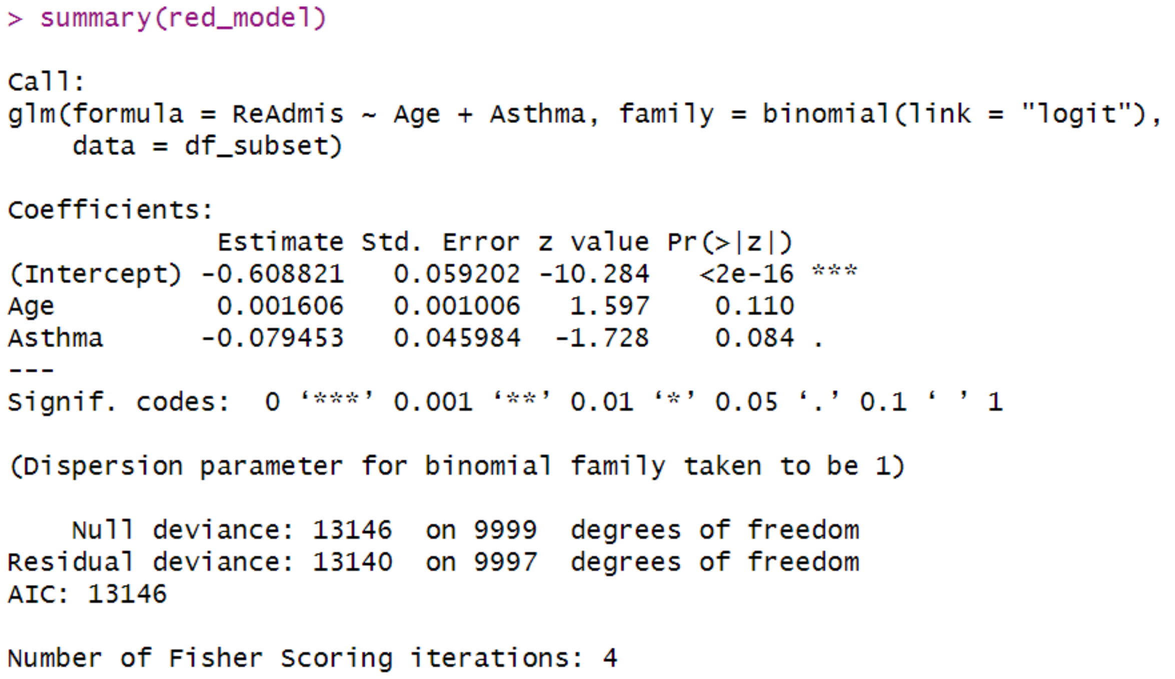
```
Call:
glm(formula = ReAdmis ~ Age + Initial_admin + Doc_visits + VitD_levels +
    Complication_risk + HighBlood + Stroke + Overweight + Arthritis +
    Diabetes + Hyperlipidemia + BackPain + Anxiety + Allergic_rhinitis +
    Reflux_esophagitis + Asthma, family = binomial(link = "logit"),
    data = df_subset)

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -7.299e-01  2.296e-01  -3.179  0.00148 **
Age                                 1.586e-03  1.007e-03   1.574  0.11546
Initial_adminEmergency Admission    7.795e-02  5.087e-02   1.533  0.12540
Initial_adminObservation Admission -5.851e-03  5.949e-02  -0.098  0.92164
Doc_visits                         -9.903e-05  1.987e-02  -0.005  0.99602
VitD_levels                         3.762e-03  1.031e-02   0.365  0.71510
Complication_riskMedium             3.706e-03  5.463e-02   0.068  0.94591
Complication_riskHigh              -1.580e-02  5.761e-02  -0.274  0.78395
HighBlood                           1.045e-02  4.227e-02   0.247  0.80466
Stroke                              5.634e-03  5.199e-02   0.108  0.91370
Overweight                         -3.758e-02  4.568e-02  -0.823  0.41074
Arthritis                           3.308e-02  4.331e-02   0.764  0.44500
Diabetes                           -1.172e-02  4.665e-02  -0.251  0.80157
Hyperlipidemia                      1.664e-02  4.393e-02   0.379  0.70483
BackPain                            5.627e-02  4.219e-02   1.334  0.18228
Anxiety                             9.294e-03  4.446e-02   0.209  0.83442
Allergic_rhinitis                  -2.079e-02  4.255e-02  -0.489  0.62512
Reflux_esophagitis                  2.230e-02  4.217e-02   0.529  0.59705
Asthma                             -7.913e-02  4.604e-02  -1.719  0.08566 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13146  on 9999  degrees of freedom
Residual deviance: 13132  on 9981  degrees of freedom
AIC: 13170

Number of Fisher Scoring iterations: 4
```

D2. JUSTIFICATION OF MODEL REDUCTION

For this model, I used the backward stepwise elimination. Firstly, construct the initial regression model with all the predictor variables in your dataset. Then, the **step()** function does all the work. It utilizes multiple iterations that evaluate the contribution of each predictor variable. Every predictor variable with the highest p-value and thus contributes the least to the model is removed. The function stops when the iteration with the lowest AIC is obtained.

The backward stepwise elimination method is vital to my research question since it enables me to select the statistically significant variables. These variables are considered the most important contributors to the increase in additional charges.

D3. REDUCED LINEAR REGRESSION MODEL



```
> summary(red_model)

Call:
glm(formula = ReAdmis ~ Age + Asthma, family = binomial(link = "logit"),
    data = df_subset)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.608821   0.059202 -10.284   <2e-16 ***
Age          0.001606   0.001006   1.597    0.110    
Asthma      -0.079453   0.045984  -1.728    0.084 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13146  on 9999  degrees of freedom
Residual deviance: 13140  on 9997  degrees of freedom
AIC: 13146

Number of Fisher Scoring iterations: 4
```

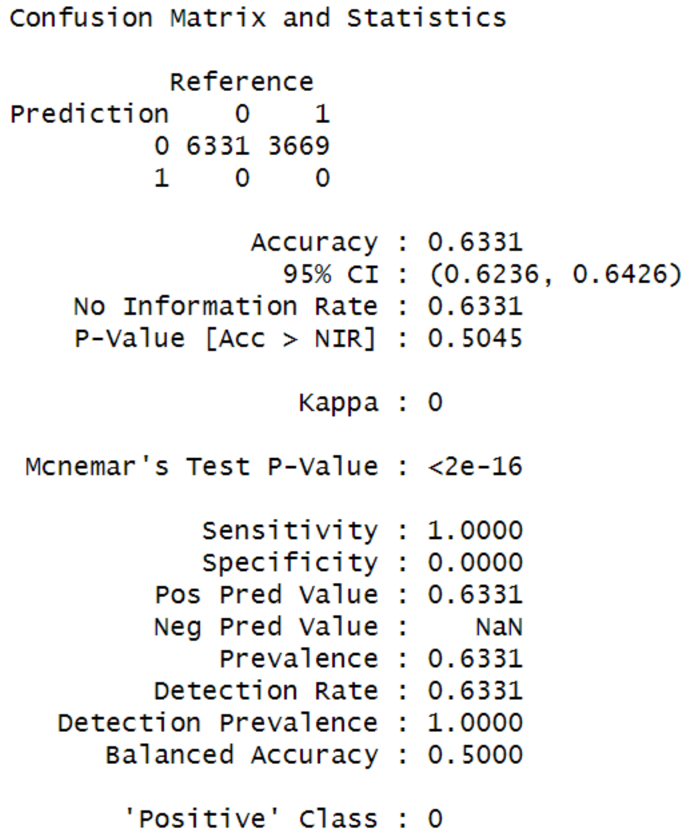
E1. MODEL COMPARISON

	Initial model	**Reduced model**
No. of predictors	18	2
Null deviance	13126 on 9999 degrees of freedom	13146 on 9999 degrees of freedom
Residual deviance	13122 on 9981 degrees of freedom	13140 on 9999 degrees of freedom
AIC	13170	13146

The predictors are simplified with fewer predictors in the reduced model. In contrast, the initial model includes a comprehensive set of predictors. Null deviance represents the difference between the null model and the current model. In both models, the null deviance is in the comparable range. This indicates that both models fit the data similarly. The residual deviance represents the difference between the model with predictors and the null model. The reduced model has a slightly higher residual deviance. This suggests that the initial model first makes the data marginally better regarding deviance. The most significant indicator is the AIC (Akaike Information Criterion). This criterion is a measure that balances the goodness of fit and model complexity. A lower AIC indicates a better-fitting model. In this analysis, the reduced model has the lower AIC of the two models. Therefore, the reduced model is the better-fitting model of the two models.

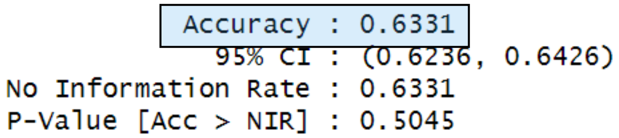
E2. OUTPUT & CALCULATIONS

CONFUSION MATRIX OF THE REDUCED MODEL



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6331 3669
         1    0    0

               Accuracy : 0.6331
                 95% CI : (0.6236, 0.6426)
    No Information Rate : 0.6331
    P-Value [Acc > NIR] : 0.5045

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.6331
         Neg Pred Value :    NaN
             Prevalence : 0.6331
         Detection Rate : 0.6331
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

ACCURACY CALCULATION FOR THE REDUCED MODEL



```
               Accuracy : 0.6331
                 95% CI : (0.6236, 0.6426)
    No Information Rate : 0.6331
    P-Value [Acc > NIR] : 0.5045
```

E3. CODE

.R code attached

PART V: DATA SUMMARY & IMPLICATIONS

F1. RESULTS

REGRESSION EQUATION

$$\text{log-odds}(\hat{p}) = \beta_0 + \beta_1 * \text{ Age } - \beta_2 * \text{Asthma}$$

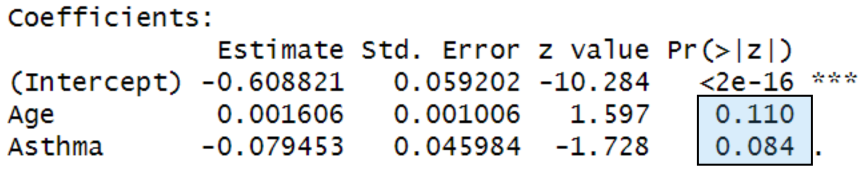
INTERPRETATION

$$\text{log-odds}(\hat{p}) = \text{-}0.608821 + 0.001606 * \text{ Age } - 0.079453 * \text{Asthma}$$

The intercept is the log-odds of readmission when all predictor variables are zero. For every one-unit increase in age, the log-odds of readmission increase by 0.001606, assuming all other variables are held constant. For patients with asthma, the log-odds of readmission decrease by 0.079452 compared to those without, assuming all other variables are held constant.

Furthermore, to interpret these coefficients, use the logistic function to convert log-odds to probabilities. Then, find the percentage change in odds for a one-unit change in each predictor variable. Subtract 1 to the odds ration then multiply by 100 to solve for the percentage change. The intercept has an odds ratio of 0.544, which means a reduction of approximately 45.6% in the odds of readmission compared to when all predictor values are at their reference level. The odds ratio of 1.002 indicates a marginal increase in the odds ratio of approximately 0.16% for every one-unit increase in age. There is a marginal increase in the odds of readmission age increases. Lastly, the odds ratio of asthma is 0.924. this implies approximately 7.74 lower odds of readmission for individuals with asthma compared to those without. The -7.64 % is a negative percentage change that reflects the decrease in readmission odds for individuals with asthma.

STATISTICAL SIGNIFICANCE

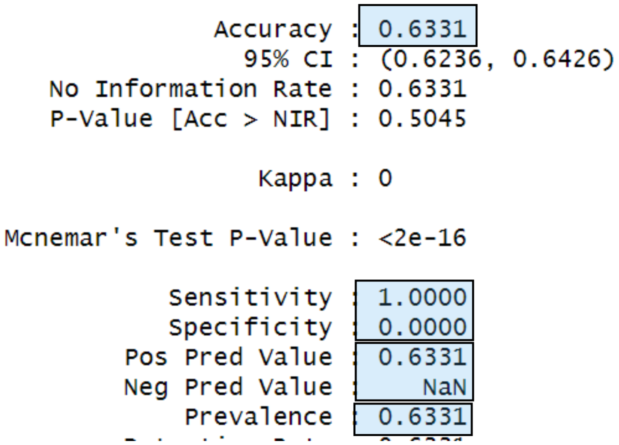


```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.608821   0.059202 -10.284   <2e-16 ***
Age          0.001606   0.001006   1.597    0.110    
Asthma      -0.079453   0.045984  -1.728    0.084 .  
```

The significance level of the p-value is often set at 0.05. Researchers choose this threshold to determine what to consider as statistically significant. It is essential to look at the associated p-value to determine the significance. Both age and asthma are greater than 0.05, indicating that in this model, there is insufficient evidence to declare that these variables significantly influence the likelihood of readmission. When interpreting the p-values, it is crucial to remember that a nonsignificant result does not prove the absence of an effect.

PRACTICAL SIGNIFICANCE

Practical significance is about the importance of the observed effects. In this logistic regression, practical relevance can be determined by considering the magnitude and impacts of the coefficients of the predictors. The odds ratio for age is 1.002, while asthma is 0.924. Considering these results, both age and asthma show relatively small effects on the odds of readmission.



```
               Accuracy : 0.6331
                 95% CI : (0.6236, 0.6426)
    No Information Rate : 0.6331
    P-Value [Acc > NIR] : 0.5045

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.6331
         Neg Pred Value :    NaN
             Prevalence : 0.6331
         Detection Rate : 0.6331
```

Practical significance can be assessed by examining the performance metrics derived from the confusion matrix. Firstly, the accuracy indicates the overall correctness of the model's prediction. The accuracy is 63.31%. This means that about 63.31% of the predictions made by the model are correct. Next, analyze the sensitivity (true positive rate) and specificity (true negative rate). The sensitivity is 100%, which means the model correctly identifies all instances of class 1 hospital readmissions. Meanwhile, the specificity is 0%, which means the model incorrectly classifies all instances of class 0 (no hospital readmission). The 0% is a significant concern since this signifies scenarios of false positives.

Moreover, the positive predictor value (PPV) and the negative predictor value (NPV) highlight the dataset's imbalance. The PPV is 63.31%, while the NPV is NaN. The NaN implies that no instances are predicted as negative (no hospital readmission), making it impossible to calculate NPV. This means the model has an inability to make negative predictions.

In a medical context, this model is not practically significant. The lack of specificity and inability to make negative predictions suggest that it may not be practically useful.

LIMITATIONS OF THE ANALYSIS

It is vital to be aware of the limitations of the analysis. The transformation of categorical variables in regression analysis is one of the potential limitations. Firstly, the dataset may lose some information inherent in the categories. In this analysis, this transformation is evident for variables like high blood and stroke and potential complication risk. Moreover, recoding categorical variables into numerical values might lead to a loss of meaningful information. Assigning 1 to Yes and 0 to No may not capture the nuances of each category. This can lead to a risk of misinterpretation. Users of the model may interpret the binary if not adequately communicated.

It is crucial to note the imbalanced dataset. There are twice the number of No or 0 compared to Yes or 1. Imbalanced datasets can lead to biased models. Based on the confusion matrix, the algorithm performed well in predicting the majority but poorly in predicting the minority class. Moreover, the confusion matrix shows perfect sensitivity but zero specificity. This imbalance is due to issues with the modeling approach or the dataset itself. A model that labels all instances as positive achieves perfect sensitivity but is not practically useful.

Understanding the limitations of the analysis is crucial for informed interpretation.

F2. RECOMMENDATIONS

Based on the regression analysis results, where I identified the factors contributing to readmission, I have several recommendations for the organization.

Firstly, the hospital should launch patient education initiatives targeting asthma management and general health awareness. The organization can provide patients with the knowledge and tools to participate actively in their healthcare. This leads to better health outcomes and reduced readmission rates. Moreover, the organization should foster collaborations with community resources to provide holistic support for patients post-discharge. This might include partnerships with community health organizations or local support groups to address health before and after readmission.

Furthermore, the organization can strengthen collaboration among healthcare providers. This is to ensure seamless communication between different specialties involved in patient care. This could involve shared electronic health records and interhospital care teams to provide comprehensive teams. Regular training programs can also help healthcare staff to improve patient care. This ensures that the latest evidence-based practices are incorporated.

Lastly, establish a system for continuous model refinement based on real-time data. Regularly update the predictive model to enhance accuracy and relevance. Evolving patterns in patient demographics, health conditions, or treatment should be considered. Most importantly, establish a feedback mechanism among healthcare providers, patients, and data analysts. The organization should encourage open communication to gather insights and gauge the effectiveness of implemented interventions.

It is essential to take a proactive and targeted approach to hospital readmission. The model can enhance patient outcomes and optimize healthcare resources when it is continuously maintained. By implementing these recommendations, the organization can improve patients' hospital readmission.