Logistic Regression

Matthew E. Heino

D208 Predictive Modeling

Logistic Regression

Introduction

The purpose of the assessment is to be able to produce a logistic regression from a set of chosen features. This paper will work through a series of concepts that will be needed to create the logistic regression. The steps that will covered will be data preparation, data transformation, the creation of an initial logistic regression, and the subsequent reduction of the features within the model.

Background.

The background for this data is the following. The data is for a medical organization. The file is stored in a CSV format and it consists of 50 columns and 10,000 rows of observations. The columns on the dataset are both categorical and continuous. The columns store information about the patient. This information is contained if this patient has a history of illnesses like the following:

- Does the patient have high blood pressure?
- Does patient the have diabetes?

These are just a few of the questions that are included in the dataset. For a more exhaustive look, it will be beneficial to look at the complete CSV and the accompanying data dictionary.

All the code that was used to create this assessment can be found in the accompanying Jupyter Notebook. In this notebook, you will find some additional notes that were included to help explain the process that was used to create the logistic regression model.

Part 1. The Research Question

In this section, there will be a discussion about the research question and what are the goals of the research question. This section will go over what the research question means and how it will be of benefit to the organization and the patient. The goals will be discussed as they pertain to the research question and the organization.

A1. Question Summary. The question that proves beneficial to the medical organization is the following: "Which factors lead to the patient the patient being readmitted?" The definition of readmitted is any patient who has been readmitted within the last thirty days. It will be beneficial to see what factors the patient had in the past or is currently suffering from the see if the patient will be readmitted again.

A2. Goal of the Data Analysis. The goal of the data analysis is to see if will be possible to predict the probability of whether a patient will be readmitted based on a set of features. These features will be a set of categorical values that represent the patient's previous history. Overall goals are the following:

1. A model that will be able to predict whether a patient with certain features will face the possibility of being readmitted.
2. After the model and analysis are completed is there anything the medical institution can do to reduce the likelihood of re-admittance?
3. If a course of action can be offered based on the analysis, is there a possibility that the organization can use this analysis to help categorize the patient for further assistance to reduce the chance of the patient's re-admittance?

After the research has been answered we would like to see an avenue that will allow for definitive action to occur. If certain features seem to be pertain. It may be beneficial to see what resources are available to make sure the patient will not be readmitted in the future.

Part II. Justification of the Method

This section will look at the assumptions that must be true to create a logistic regression model. There will be a discussion about why Python has been employed as the language of choice for this form of analysis. There will be an explanation as to why logistic regression will be the appropriate model to answer the question that was proposed in Section A.

B1. The Assumptions of a Logistic Regression. To have a successful logistic regression model there needs to be an adherence to a few assumptions that make logistic regression possible. The assumptions are the following:

1. The target variable **y** must be categorical.
    - There are a few forms of logistic regression. There is multinomial logistic, binary logistic, and ordinal logistic regression. For this paper, the binomial form is the regression form that will be used to create the model. The model's **y** variable must be categorical (Straw, n.d.).
2. The logistic regression has linearity with the logit(p).
    - This means that the continuous **x** variable **x** and the logit(p). The form of the **x** variables where there were no transformations, raised to a power, or other methods used to transform the **x** (Straw, n.d.). The logit function is used to map probability values of **y** rather than the actual value of **y**. The **y** variable is only binary and not a

continuous value. This is in contrast to the linear regression model where we are looking for a predicted numeric value and not a binary outcome.

3. The logistic regression exhibits low or no multicollinearity.
   - This is the relationship between the predictor variables and the other predictor variables in the dataset. If there were high degrees of multicollinearity among the **x** variables that would affect the accuracy of the calculation of the coefficients (Straw, n.d.).
   - If there is a high degree of multicollinearity among the predictor variables it will make it difficult for the model to find the relationships between the predictor variable (**x**) and the target variable (**y**). The relationship between the **x** and **y** can be observed faithfully if there are no other variables impacting the relationship. The relationship must be free of influence from the other variables in the model. This will not happen if there is a high degree of multicollinearity among the variables.
4. The observations of the dataset must be independent.
   - This assumption means that the observations are not correlated with the other observations within the dataset. If the observations are not independent we will have the problem of multicollinearity as discussed in a previous assumption. If observations are not independent there will be "repeating patterns in the errors " (Straw, n.d.).

B2. Benefits of Using Python. While there are many choices for creating a logistic regression model, the language that has been chosen is Python. Python offers a range of packages that can be used to meet the requirements of data preparation and the creation of the logistic regression model. The benefits of the language are that it is open source and many independent packages are available. It is fully supported and updated regularly. In the table, you will see some of the packages that will be utilized to prepare the data and create the model.

Python Package	Description
matplotlib	Plotting functions for the graphs that aid in visualizing the data in the dataframe.
missingno	Used to visualize if there are any missing values in the dataset.
pandas	Provides methods that can be used to create a dataframe. It also has a function that is used to manipulate the content within it.
sklearn	For the creation of the confusion matrix.
seaborn	Used to plot and create the graphs that are found in the assessment. Used to create histograms and other visuals.
statsmodels	Used to create the model for the logistic regression regression.
sklearn	Used to create testing sets for testing the model that was created.

Table 1: Python Packages

B3. Why is Logistic Regression the Appropriate Technique? Why should we use logistic regression? To answer this question we need to look at the components that make up the question's outcome. We are looking at a target variable that only has two possible values. The target variable only contains Yes or No. We want to see the probability of a patient being readmitted within one month. We want to look at a set of features that can be used to determine the probability of the event, in this case, the likelihood of readmission to the hospital.

The logistic regression model is the best technique because it allows the question to be answered. The three things that the logistic regression model is used for in data analysis is that it can be used to forecast the effects of changes within a scope of features. It can be used to look for trends and future values. We are looking to predict if a given set of predictor features which ones seems to be most likely to influence the outcome. This outcome is either the patient will have a probability of being readmitted within a month or the patient will not be readmitted within the month. When the model is completed we can look at the "strength" of the predictors by looking at the coefficients and see how they positively or negatively affect the model (Thanda, 2023).

These are some of the reasons that utilizing the regression model is appropriate for answering the research question. A linear regression would not be appropriate here as it requires the target to be a continuous one. The chosen target variable ReAdmis is categorical which is not continuous.

As stated earlier the two models do not produce the same types of results. If we wanted to use a linear regression we need to change what we are looking to answer. We would need to ask a question where the response is not a dichotomous one. We would need to ask a question that yields a numeric one (Kanade, 2022).

Part III. Data Preparation

C1. Description of Data Cleaning Goals. The goal of data cleaning is to make sure the data is in a state that will allow for successful analysis. While the given file is stated as being "clean" it is always advisable to look at the file and see if there are areas that are not usable in their current state. The steps that will be followed will be the following:

1. To look for any data that has duplicates within the dataset. If there are duplicates will be removed from the set.   The inclusion of these duplicates may influence the possible imputation of missing values. The inclusion of these duplicate rows may affect the statistics, e.g., mean, median, and mode.
2. The next step is to look at missing values. There will be a visualization using the missingno package to see if there are any missing values. This along with code to see if there are a small number of missing numbers that might not be apparent in the missingno graph.
3. In this step, the goal is to look for outliers before performing any imputation of the missing values. If they are found they will be removed from the dataset. The method that will be used is the interquartile range method. The code for this will be found in Step 3 of the Jupyter Notebook.
4. In this step, the data that is not in a form that is suitable for the creation of the logistic model will be transformed. The candidate method for this is the **get_dummies()** method that is found in the pandas package. Please note that is transformation will be done after the creation of statistics that are to be discussed in subsequent sections of this

paper. It was to make it easier to understand the summary statistics that are required in Section C2.

5. In this step, there will be a look at whether the predictor variables exhibit multicollinearity. If there are predictor variables that exhibit this quality then they will be removed from consideration and not included in the reduced set of variables.

All the code that is required to accomplish these steps can be found in section C of the Jupyter. You will find that there are some additional notes as well as the proper citations for the code.

C2. Description of Variables. The features that are required are an assortment of categorical and continuous variables. The table below will show what are the candidate variables for inclusion in the logistic regression.

	Variable	Data Type	Purpose
1	Age	Continuous	Predictor
2	Gender	Categorical	Predictor
3	VitD_levels	Continuous	Predictor
4	vitD_supp	Categorical	Predictor
5	HighBlood	Categorical	Predictor
6	Initial_admin	Categorical	Predictor
7	Complication_risk	Categorical	Predictor
8	ReAdmis	Categorical	Target
9	Diabetes	Categorical	Predictor

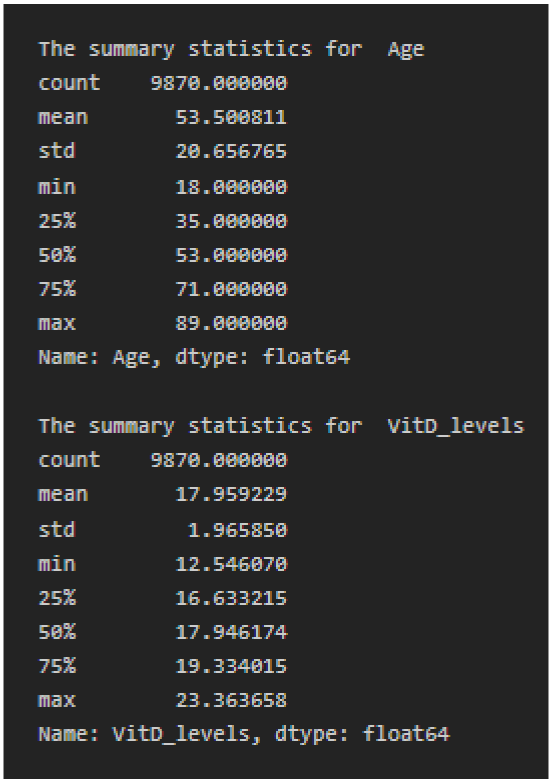
10	BackPain	Categorical	Predictor
11	Stroke	Categorical	Predictor
12	Initial_days	Continuous	Predictor
13	Asthma	Categorical	Predictor

Table 2: List of Required Variables

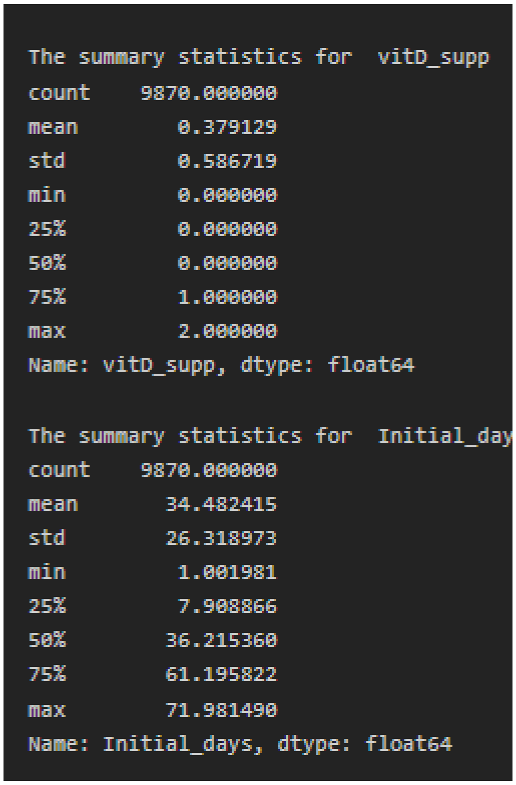
Summary Statistics of the Variables.

This section will show summary statistics of the candidate variables that will be used in the creation of the logistic regression model. Please note that these statistics have been calculated before any sort of transformation has been performed on the variables. The only data preparation that has been performed is that any outliers have been removed from the dataset. This was accomplished in section C1 of the Jupyter Notebook.

The first group of summary statistics is for the continuous variables that are to be included in the model.



```
The summary statistics for  Age
count    9870.000000
mean       53.500811
std        20.656765
min        18.000000
25%        35.000000
50%        53.000000
75%        71.000000
max        89.000000
Name: Age, dtype: float64

The summary statistics for  VitD_levels
count    9870.000000
mean       17.959229
std         1.965850
min        12.546070
25%        16.633215
50%        17.946174
75%        19.334015
max        23.363658
Name: VitD_levels, dtype: float64
```



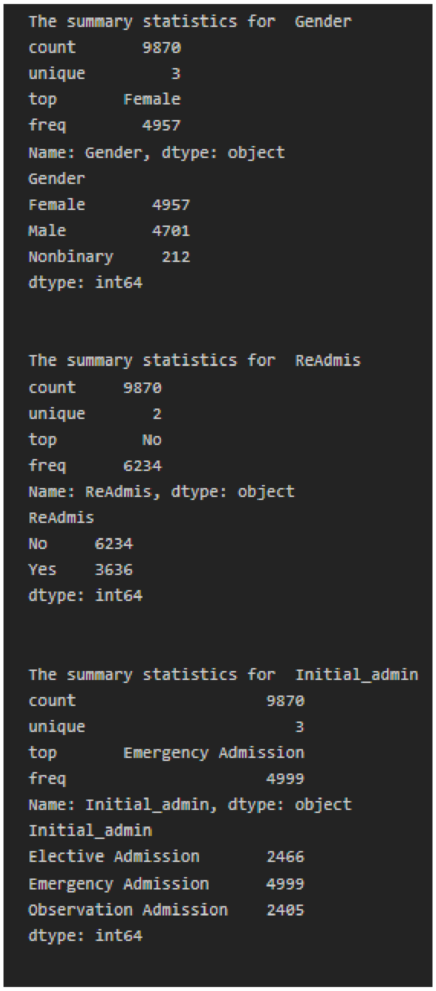
```
The summary statistics for  vitD_supp
count    9870.000000
mean        0.379129
std         0.586719
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         2.000000
Name: vitD_supp, dtype: float64

The summary statistics for  Initial_day
count    9870.000000
mean       34.482415
std        26.318973
min         1.001981
25%         7.908866
50%        36.215360
75%        61.195822
max        71.981490
Name: Initial_days, dtype: float64
```

Some observations about the data using summary statistics, looking at the Age column the youngest seems to be 18. This does not seem like it would be representative of hospital admissions. Ages less than this age may not be adequately modeled since the test data will not include age groups below this age. Age groups greater than 89 are not represented. Based on this there may be problems modeling data that falls outside these ranges.

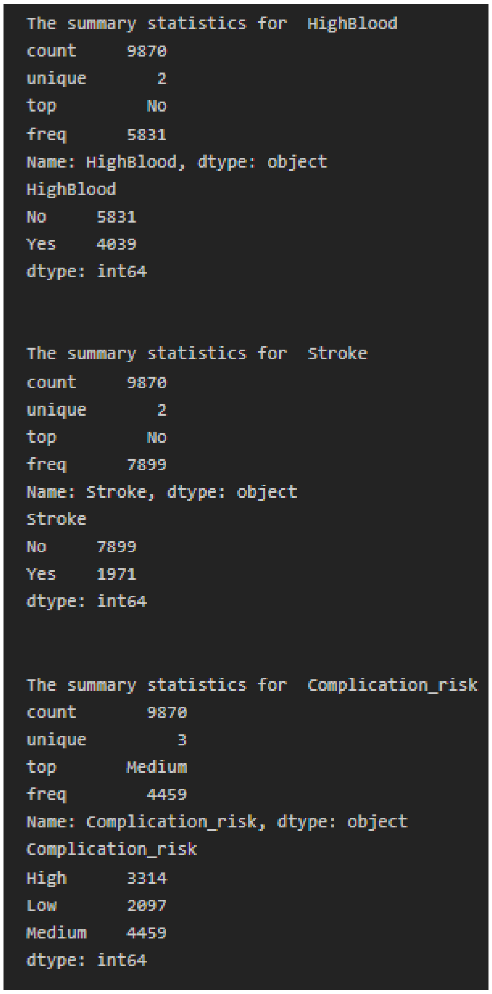
VitD features seem to have no anomalies in terms of mean and standard deviation. All the values seem to be in line with what is expected and no further inference can be offered.

Examining Initial_days there is a wide standard deviation the standard deviation is 26 days. It does seem like a lot but it will encompass most values within two standard deviations. Of note the time in the hospital where the mean does seem quite high.   Looking at available statistics this is quite a bit higher than the average of 5.7 days (Statista, 2023). This may be an avenue for future research.

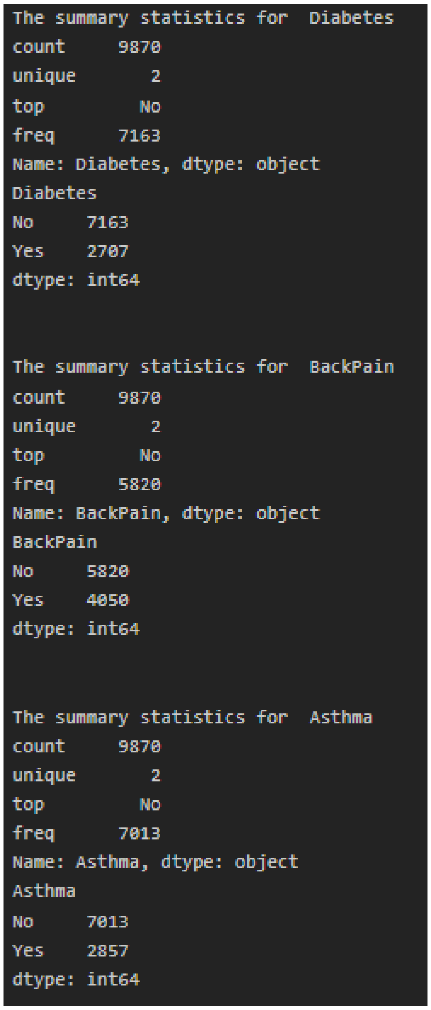
The next group of statistics is for the categorical variables.



```
The summary statistics for  Gender
count       9870
unique         3
top       Female
freq        4957
Name: Gender, dtype: object
Gender
Female       4957
Male         4701
Nonbinary     212
dtype: int64


The summary statistics for  ReAdmis
count     9870
unique       2
top         No
freq      6234
Name: ReAdmis, dtype: object
ReAdmis
No     6234
Yes    3636
dtype: int64


The summary statistics for  Initial_admin
count                        9870
unique                          3
top       Emergency Admission
freq                         4999
Name: Initial_admin, dtype: object
Initial_admin
Elective Admission       2466
Emergency Admission      4999
Observation Admission    2405
dtype: int64
```



```
The summary statistics for  HighBlood
count     9870
unique       2
top         No
freq      5831
Name: HighBlood, dtype: object
HighBlood
No     5831
Yes    4039
dtype: int64


The summary statistics for  Stroke
count     9870
unique       2
top         No
freq      7899
Name: Stroke, dtype: object
Stroke
No     7899
Yes    1971
dtype: int64


The summary statistics for  Complication_risk
count       9870
unique         3
top       Medium
freq        4459
Name: Complication_risk, dtype: object
Complication_risk
High      3314
Low       2097
Medium    4459
dtype: int64
```



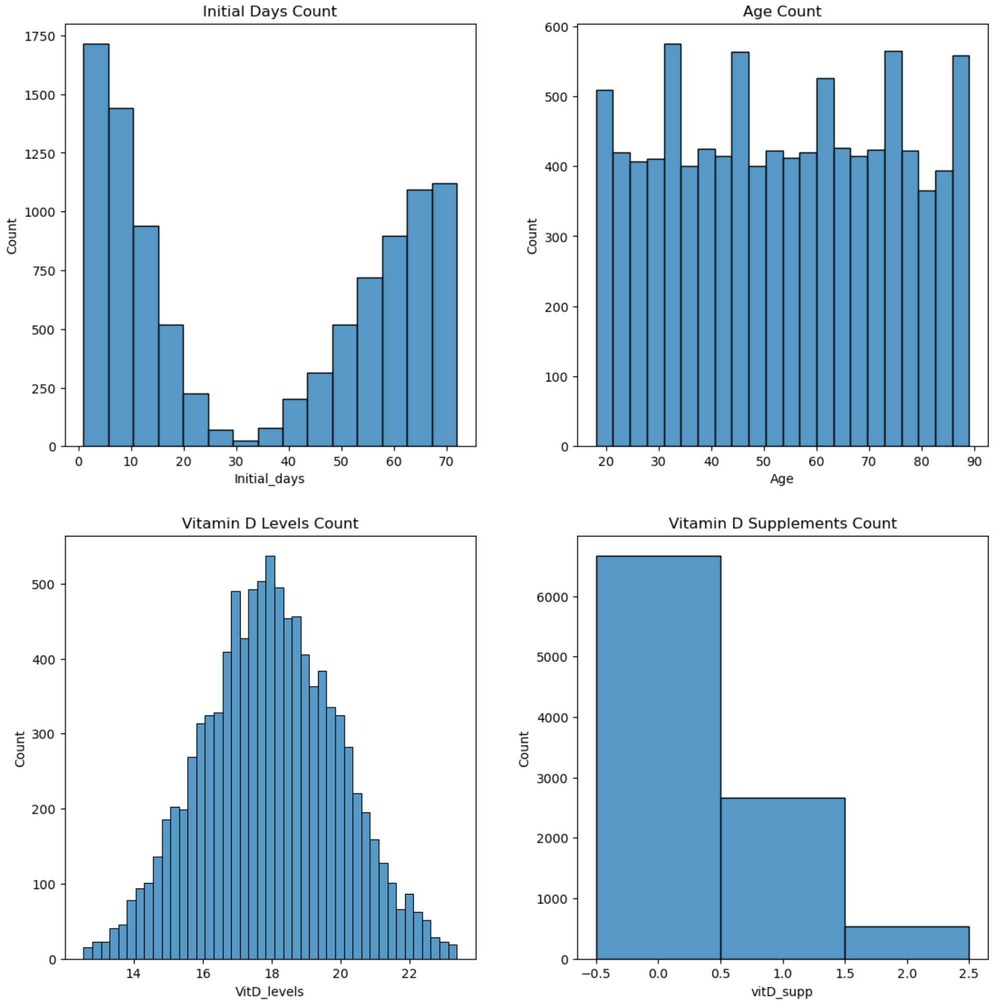
```
The summary statistics for  Diabetes
count     9870
unique       2
top         No
freq      7163
Name: Diabetes, dtype: object
Diabetes
No     7163
Yes    2707
dtype: int64


The summary statistics for  BackPain
count     9870
unique       2
top         No
freq      5820
Name: BackPain, dtype: object
BackPain
No     5820
Yes    4050
dtype: int64


The summary statistics for  Asthma
count     9870
unique       2
top         No
freq      7013
Name: Asthma, dtype: object
Asthma
No     7013
Yes    2857
dtype: int64
```

Please note that all the code can be found in the C2. Data Exploration section of the Jupyter Notebook. You will also find some additional notes in the code section. These summary statistics were created using the method **describe()**.

C3. Univariate and Bivariate Visualizations.   In this section, there will be visualizations of the continuous and categorical variables that will be used in the logistic regression. Where needed there will be some comments on what is observed in the graphs. These observations may be useful in the analysis section that is in the following section of the document.

Note: Please note that some of the variables will not be used in creating the final model. The VIF stage has not been completed. This stage will be completed at the end of the section. This is to better keep in line with the assessment sections. Implementing the VIF procedure will change the values that are exhibited to the reader. Since these values will need to be transformed into numerics. This transformation may make it hard to understand what the graphs are depicting when it comes to depicting the categorical variables.

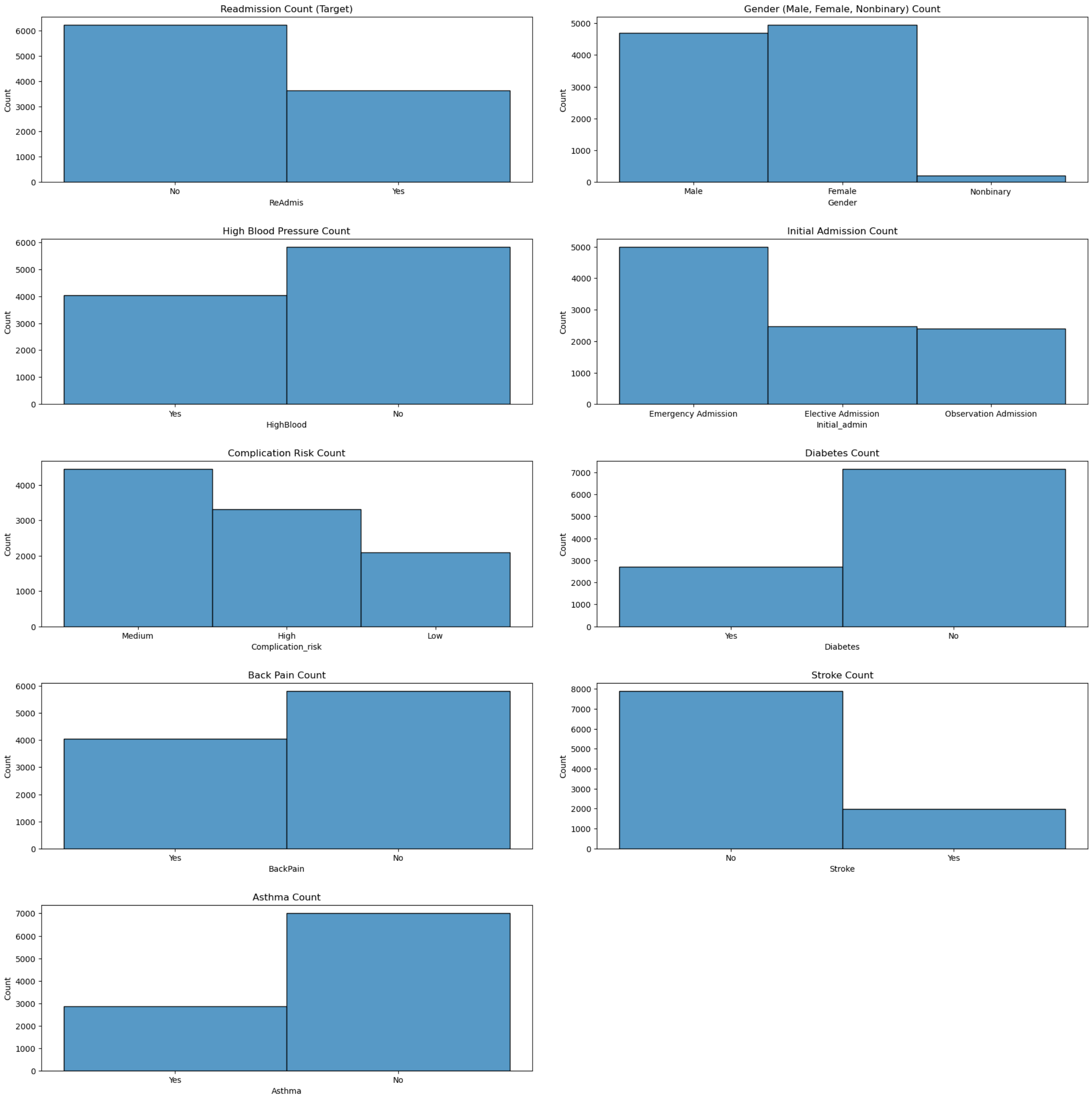
Continuous Univariate Visualization of Variables. This section contains the graph of the distribution of the continuous variables that are used in the model.



A few observations can be made by looking at the continuous variables. Age seems to exhibit a very close uniform distribution. And as noted earlier there are no observations for patients below 18 included or after 89. This may cause issues if this feature is included in the model.

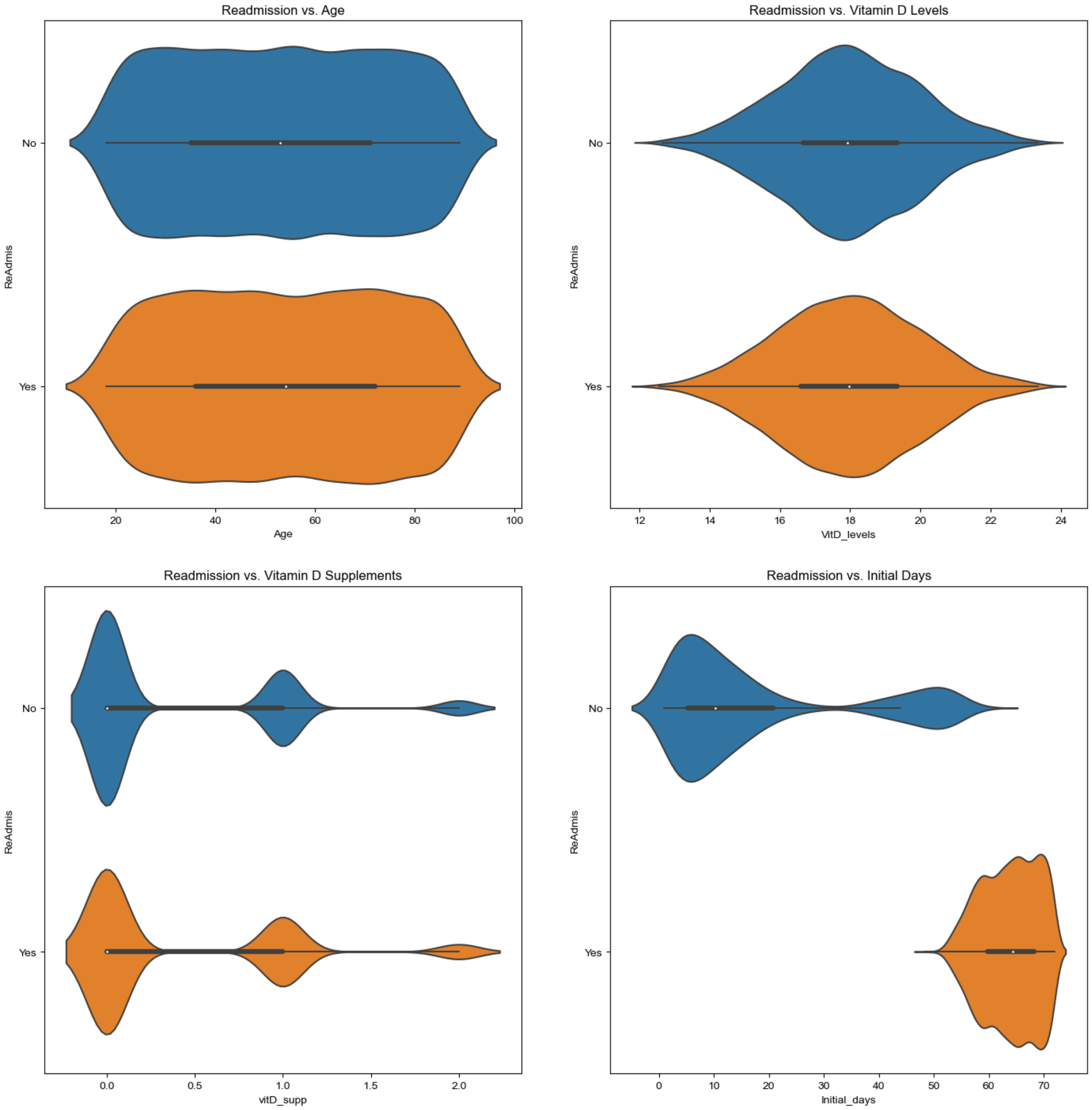
There is one other observation looking at the graph for Initial_days. There seems to be an almost lack of observations for days spent in the hospital in the 30 to 40-day range. This might cause problems and it could indicate some flaws in data collection. It would seem likely that there should be more observations in this range.

Categorical Univariate Visualization of Variables. This section contains a graph of the distribution of the categorical variables that are used in the model.



Bivariate Continuous Visualizations.

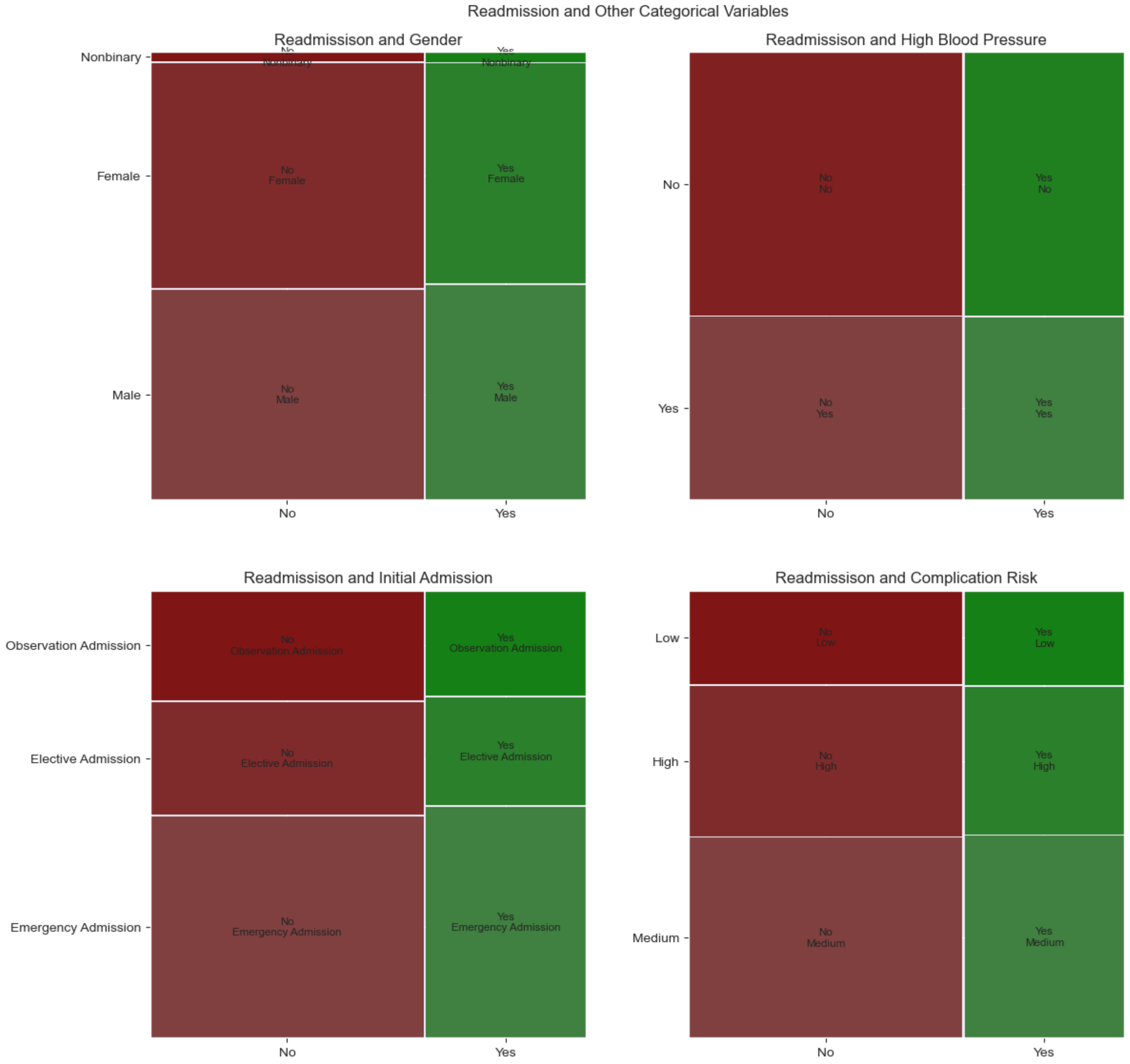
This section will contain visualizations that show the Readmis variable plotted against the continuous variables that are candidates

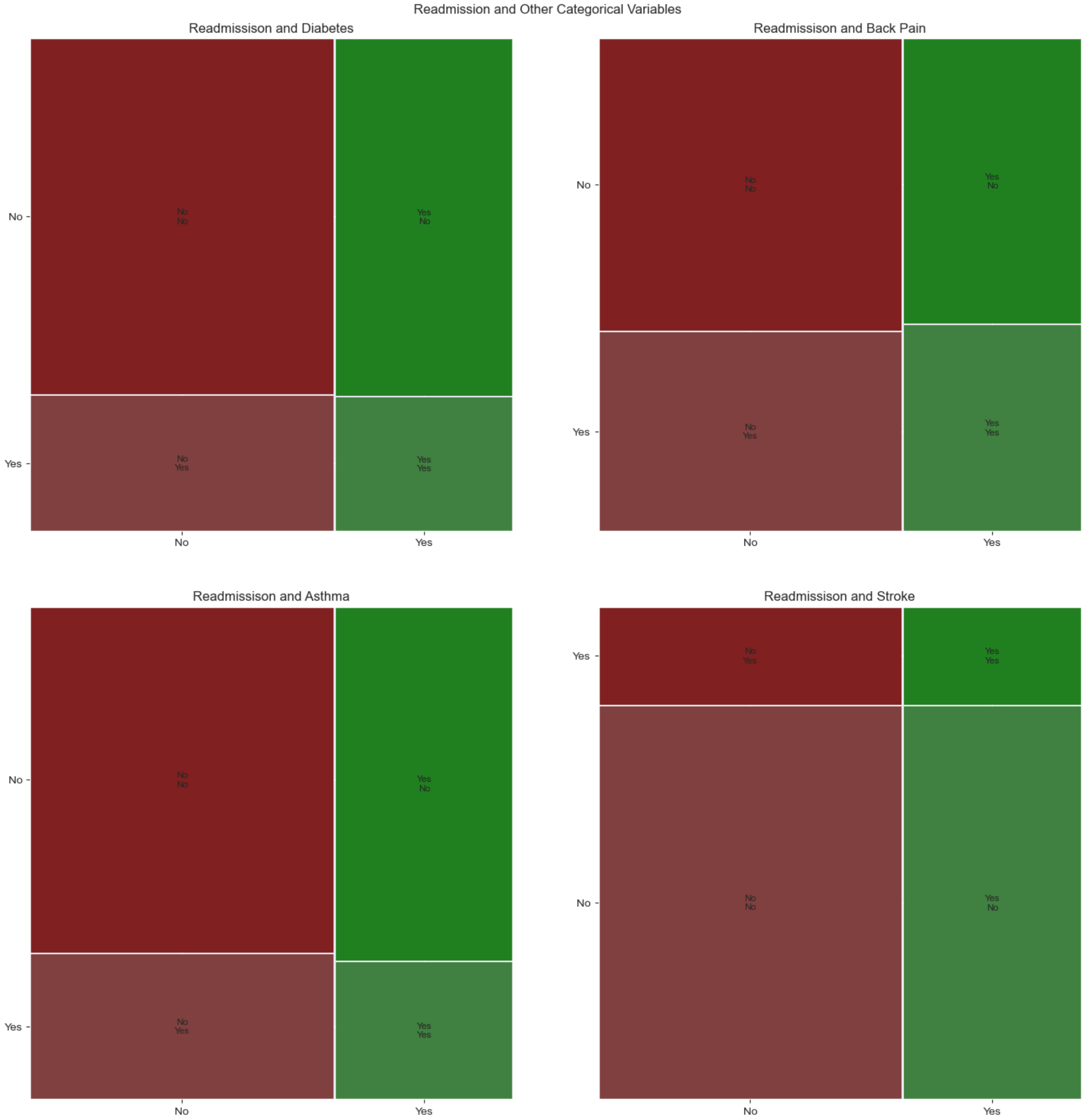


Violin plots were chosen as a way to view a few things that were not available in a standard boxplot. It was a way to view the distribution along with data that is found in the boxplot. You can see the interquartile range along with the median (the white dot) (Carron, 2021).

Bivariate Categorical Visualizations.

In this section, the mosaic graph was used as a way to show a chart to help visualize the categorical variables (*Mosaic Plot*, n.d.).





The code that was used to create the mosaic plot can be found in section C3 of the accompanying Notebook.

C4. Transformation Goals. The goal of data transformation is the render the data in a state that can be used for the proposed model. The model that was proposed for the question was the logistic regression. This model needs to have the variables that are not numeric in nature transformed. Some of the explanatory variables that will be used are categorical and use strings to represent the data in that column or feature. These need to be changed for these categorical features to be used in the model. The method that will be used is to use get_dummies() method. This method will encode the categorical using a binary scheme.

For example, if the variable is composed of three different sub-categories there will be a scheme where it will encode in the following manner. Using complication_risk as an example. Complication_risk has the following values Low, Medium, and High. The encoding table will be something like the following:

	Complication_risk_Low	Complication_risk_medium	Complication_risk_high
Low	1	0	0
Medium	0	1	0
High	0	0	1

Table 3: Example get_dummies encoding.

This will yield a boolean to determine what the complication risk is. This is a binary value and it can be used for the regression model. When coded it is customary to drop the first column as a way to reduce the dimensionality of the dataframe. Rather than having K columns the reduction will result in K-1 columns. This is helpful in a large dataset or when there are a lot of sub-categories.

As stated in a previous section we will look at the VIF or variance inflation factor for the candidates for the initial regression model. Any feature candidate will need to have a score of less than 10 to be included in the prepared dataset.

C5. Prepared Dataset.

The prepared dataset can be found in the following file:

Heino_reduced_medical_task2.csv.

This will include all features that have passed with a score of less than 10 for the VIF value. This is to facilitate the creation of the model without the possible inclusion of variables that exhibit multicollinearity. Further feature reduction will be carried out in a subsequent section of this document.

Part IV. Comparison of the Model and Analysis.

In this section, there will be an initial model created using the variables that were identified in previous sections of the document. There was an initial viewing of the VIFs in the previous section. There will be no further discussion of this as it has been viewed again in Section D but will be omitted since none of the included variables have a score that is above 10. To view the code for this please view Section D.

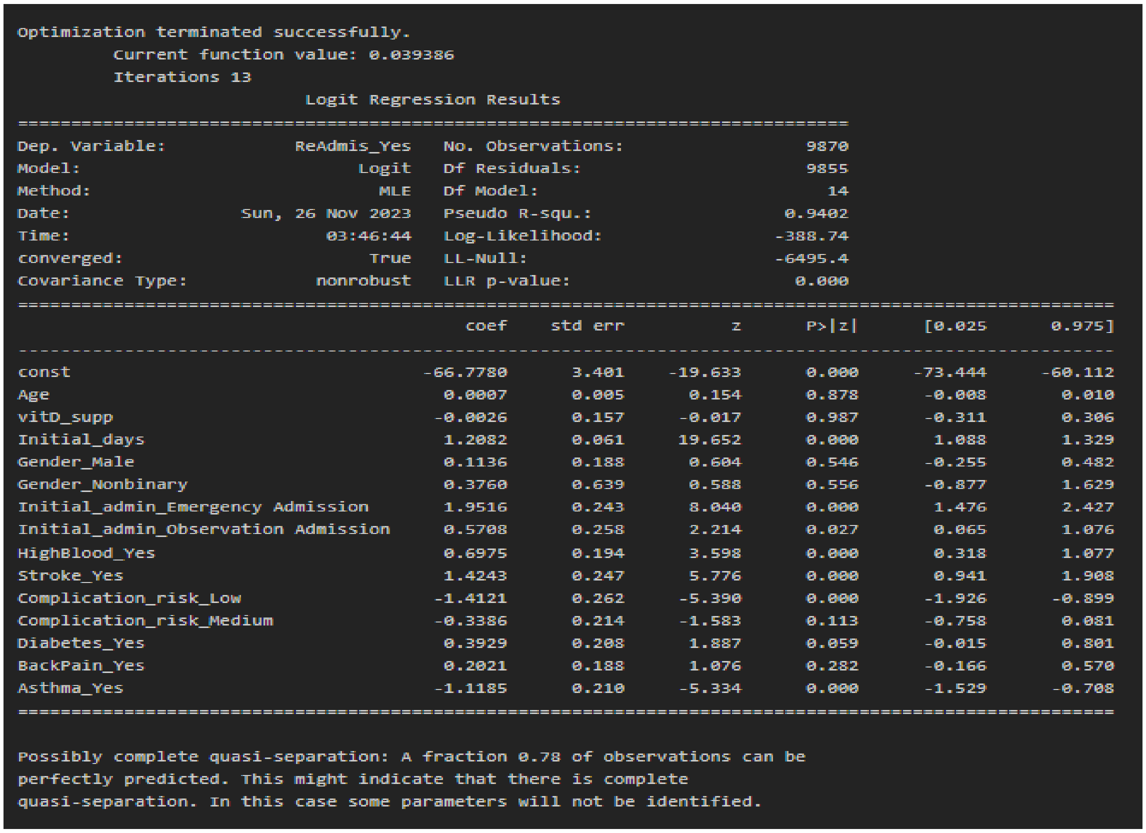
This section there will be a discussion of the features that will compose the final model. This will be the reduced model. There will be an equation created along with an explanation of the coefficients of the mean in rearguards to the question that was asked in Section A. The final section, E2, will show the accuracy of the logistic regression model.

D1. Initial Logistic Regression Model. After data cleaning and preparation it is now possible to create a model with the variables that were identified in a previous section of this document. There was one feature that was dropped after looking at the VIF in section C of this paper. The VitD_levels have been removed from consideration for the model. The code that was used to verify this is found in the Jupyter Notebook in section C.

The variables that will compose the initial model are the following:

- Age
- vitD_supp
- Initial_days
- Gender_Male
- Gender_Nonbinary
- Initial_admin_Emergency Admission'
- Initial_admin_Observation Admission
- HighBlood_Yes
- Stroke_Yes
- Complication_risk_Low
- Complication_risk_Medium
- Diabetes_Yes
- BackPain_Yes
- Asthma_Yes

Take note that there are a few additional columns that do not appear in the original CSV file. These columns were created during the use of the **get_dummies()** method. This process was briefly illustrated in the previous section. Creating the initial model yielded the following model summary.



```
Optimization terminated successfully.
         Current function value: 0.039386
         Iterations 13
                           Logit Regression Results
==============================================================================
Dep. Variable:            ReAdmis_Yes   No. Observations:                 9870
Model:                          Logit   Df Residuals:                     9855
Method:                           MLE   Df Model:                           14
Date:                Sun, 26 Nov 2023   Pseudo R-squ.:                  0.9402
Time:                        03:46:44   Log-Likelihood:                -388.74
converged:                       True   LL-Null:                       -6495.4
Covariance Type:            nonrobust   LLR p-value:                     0.000
=======================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
const                                 -66.7780      3.401    -19.633      0.000     -73.444     -60.112
Age                                     0.0007      0.005      0.154      0.878      -0.008       0.010
vitD_supp                              -0.0026      0.157     -0.017      0.987      -0.311       0.306
Initial_days                            1.2082      0.061     19.652      0.000       1.088       1.329
Gender_Male                             0.1136      0.188      0.604      0.546      -0.255       0.482
Gender_Nonbinary                        0.3760      0.639      0.588      0.556      -0.877       1.629
Initial_admin_Emergency Admission       1.9516      0.243      8.040      0.000       1.476       2.427
Initial_admin_Observation Admission     0.5708      0.258      2.214      0.027       0.065       1.076
HighBlood_Yes                           0.6975      0.194      3.598      0.000       0.318       1.077
Stroke_Yes                              1.4243      0.247      5.776      0.000       0.941       1.908
Complication_risk_Low                  -1.4121      0.262     -5.390      0.000      -1.926      -0.899
Complication_risk_Medium               -0.3386      0.214     -1.583      0.113      -0.758       0.081
Diabetes_Yes                            0.3929      0.208      1.887      0.059      -0.015       0.801
BackPain_Yes                            0.2021      0.188      1.076      0.282      -0.166       0.570
Asthma_Yes                             -1.1185      0.210     -5.334      0.000      -1.529      -0.708
=======================================================================================================

Possibly complete quasi-separation: A fraction 0.78 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

D2. Justification of Feature Selection. The model as it is has a few features that can be removed based on the initial summary of the model. Since VIF was performed in a previous step we can use the p-values to see which of the remaining variables can be removed from the model. Using the p-value with a value of 0.05 we can see that there are a few variables that will be removed from the model. The method that will be performed to remove the variables is the Backwards Step-wise Elimination. This will be carried out until all the variables have a p-value that is less than 0.05 as stated earlier. The remaining features will be used to create the reduced logistic regression in the following section.

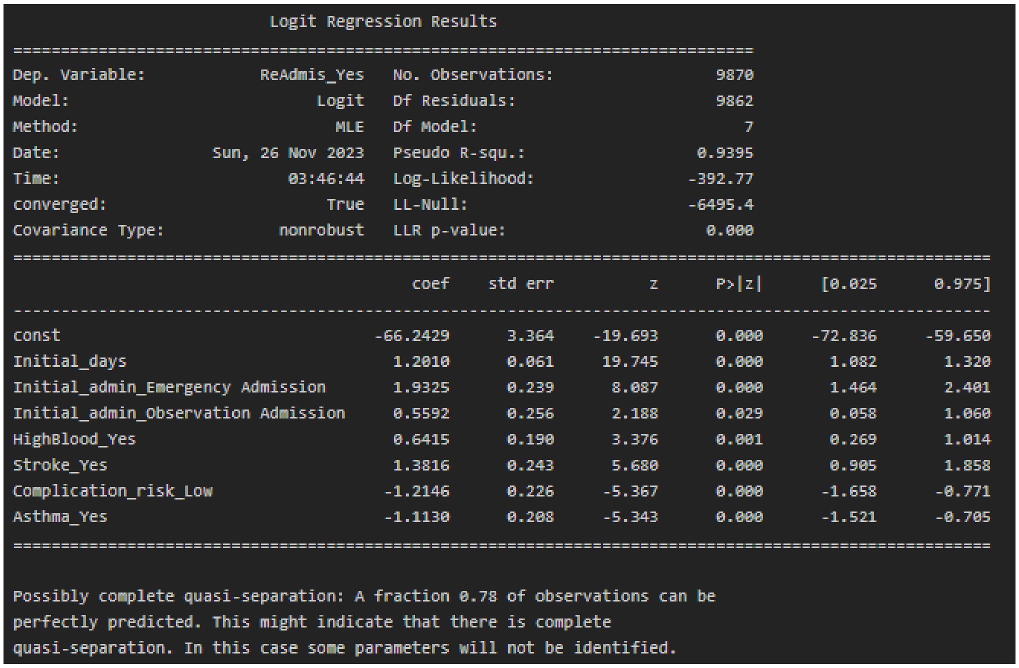
D3. The Reduced Logistic Regression Model. After performing the Backwards Step-wise elimination the following variables were removed from the model.

- vitD_supp
- Gender_Male
- Gender_Nonbinary
- BackPain_Yes
- Complication_risk_Medium
- Diabetes_Yes
- Age

The model will be composed of the following:

- Initial_days
- Initial_admin_Emergency Admission
- Initial_admin_Observation Admission
- HighBlood_Yes
- Stroke_Yes
- Complication_risk_Low
- Asthma_Yes

Please note that there are no longer any variables included in the predictor variables with a p-value greater than 0.05. The p-values were well above the accepted threshold of 0.05 and could not be included in the model based on this metric. Below is the screenshot of the reduced model. You will notice that there has been a substantial reduction in the features. If you look at the P>|z| column all the features in the column have a p-value less than 0.05.

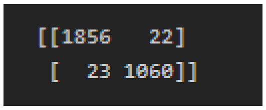


```
                           Logit Regression Results
==============================================================================
Dep. Variable:            ReAdmis_Yes   No. Observations:                 9870
Model:                          Logit   Df Residuals:                     9862
Method:                           MLE   Df Model:                            7
Date:                Sun, 26 Nov 2023   Pseudo R-squ.:                  0.9395
Time:                        03:46:44   Log-Likelihood:                -392.77
converged:                       True   LL-Null:                       -6495.4
Covariance Type:            nonrobust   LLR p-value:                     0.000
=======================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
const                                 -66.2429      3.364    -19.693      0.000     -72.836     -59.650
Initial_days                            1.2010      0.061     19.745      0.000       1.082       1.320
Initial_admin_Emergency Admission       1.9325      0.239      8.087      0.000       1.464       2.401
Initial_admin_Observation Admission     0.5592      0.256      2.188      0.029       0.058       1.060
HighBlood_Yes                           0.6415      0.190      3.376      0.001       0.269       1.014
Stroke_Yes                              1.3816      0.243      5.680      0.000       0.905       1.858
Complication_risk_Low                  -1.2146      0.226     -5.367      0.000      -1.658      -0.771
Asthma_Yes                             -1.1130      0.208     -5.343      0.000      -1.521      -0.705
=======================================================================================================

Possibly complete quasi-separation: A fraction 0.78 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

The code for this can be found in section D3. Reduced Model of the Jupyter Notebook.

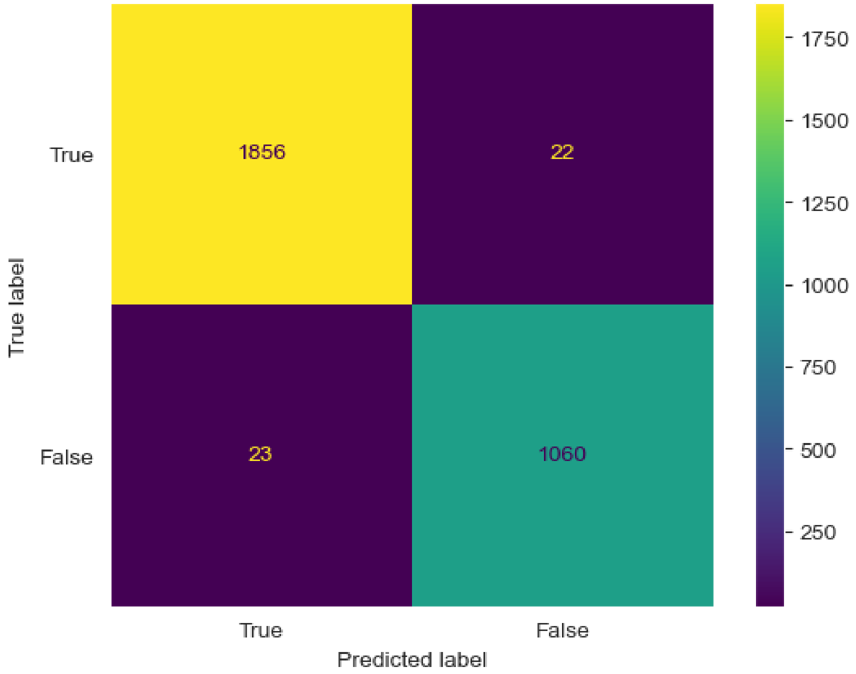
E1. Explanation of the Data Analysis Process. Comparing the models of the initial and the reduced regression are almost on par with one another. Using the LLR p-value for the models they are the same, but the key difference is the reduced model uses far fewer features to get the same result. It might be beneficial to reduce the model using a p-value that is below 0.05. This is not currently required as the values that are left are below the usual bonds of 0.05. The upside of using the reduced model is that we have pared down the number of required features that are required to make the model. This process was to simply remove the values of the features that were above the p-value of 0.05. This was done until the reduced equation was composed of components that were no longer above the accepted p-value of 0.05. This resulted in the features that can be reliably used in the model and they are significant to the model.

E2. Output of All Calculations. Using the reduced model that was created in the previous section of E. The calculation of the confusion matrix is done using the **confusion_matrix** method that is found in the **sklearn** package. The result is the following:



```
[[1856   22]
 [  23 1060]]
```

The confusion matrix graph looks like the following:



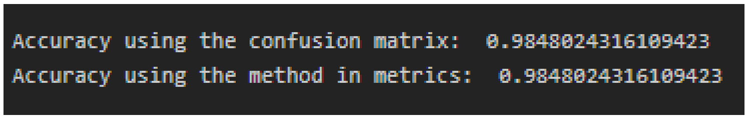
The code for these calculations can be found in the Notebook in Section E2. Logistic Regression. The model accuracy using the values from the reduced and is calculated using the following formula:

accuracy = (TN + TP) / (TN + FN+ FP + TP), where,

TN	True negative
TP	True positive

FN	False negatives
FP	False positives

The accuracy for this model is 98.4 percent. Below is a screenshot of the calculation both using the confusion matrix and the method **accuracy_score()** which is found in sklearn.



```
Accuracy using the confusion matrix:  0.9848024316109423
Accuracy using the method in metrics:  0.9848024316109423
```

E3. Code for the Assessment.

The code for the assessment can be found in the following file. It is a Jupyter Notebook.

Heino D208 Predictive Modeling Task 2.ipynb

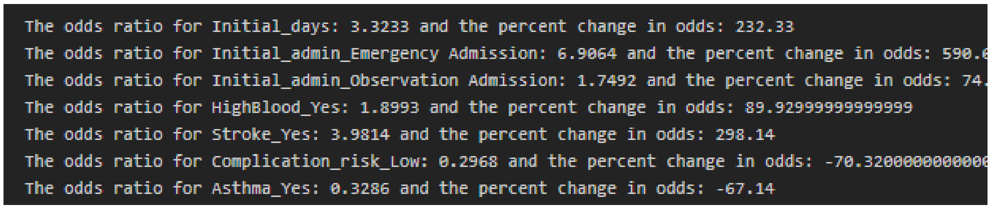
Part V. Summary and Implications

F1. Findings and Assumption.

The equation for the logistic equation is the following:

ln(p_hat / (1 - p_hat)) = -66.2429 + 1.2010(*Initial_days*) + 1.9325(*Initial_admin_Emergency Admission*) + 0.5592(*Initial_admin_Observation Admission*) + 0.6415(*HighBlood_Yes*) + 1.3816(*Stroke_Yes*) - 1.2146(*Complication_risk_Low*) - 1.1130(*Asthma_Yes*)

To interpret what the individual coefficients mean in the logistic regression model, it is beneficial to look at the odds ratio and look at the change in odds percentage (Zach, 2021). The calculations yielded the following results.



```
The odds ratio for Initial_days: 3.3233 and the percent change in odds: 232.33
The odds ratio for Initial_admin_Emergency Admission: 6.9064 and the percent change in odds: 590.6
The odds ratio for Initial_admin_Observation Admission: 1.7492 and the percent change in odds: 74.
The odds ratio for HighBlood_Yes: 1.8993 and the percent change in odds: 89.92999999999999
The odds ratio for Stroke_Yes: 3.9814 and the percent change in odds: 298.14
The odds ratio for Complication_risk_Low: 0.2968 and the percent change in odds: -70.3200000000000
The odds ratio for Asthma_Yes: 0.3286 and the percent change in odds: -67.14
```

An interpretation of the features is:

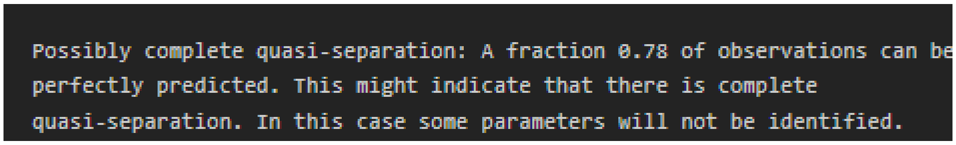
- Keeping all things constant, for a one-unit increase in days spent the odds of a patient being readmitted increase by 232.33%.
- Keeping all things constant, a patient being admitted as an emergency admission has an increase in the patient's odds of being readmitted by 590.64%.
- Keeping all things constant, for a patient being admitted as an observation admission has an *increase* in the patient's odds of being readmitted by 74.92%.
- Keeping all things constant, a patient with high blood pressure has an *increase* in the patient's odds of being readmitted by 89.93%.
- Keeping all things constant, a patient who has had a stroke has an *increase* in the patient's odds of being readmitted by 298.14%.
- Keeping all things constant, for a patient being a low complication risk there is a *decrease* in the patient's odds of being readmitted by 70.3%.
- Keeping all things constant, a patient who has asthma has a de*crease* in the patient's odds of being readmitted by 67.14%.

The code for these calculations can be found in the Jupyter Notebook.

Some limitations have been encountered with this model. Looking at the dataset there are gaps in data that may make it difficult to predict if the patient's features fall within the data gap.

For example, there aren't any patients whose age is below 18. It would be beneficial to see data from the segment of patients. Also looking at the initial days there is a distinct grouping. Either below approximately 30 days and approximately 40 days are the two groups of stays in the data set. Curiously, there is such a distinct grouping of the data. It would be logical that there would be more data in the range. This data range should not have been affected by the removal of the outliers, as it is well within the bounds of the IQR.

It is also of note that the creation of the model led to the following warning.



```
Possibly complete quasi-separation: A fraction 0.78 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Based on research the easiest best course of action is to do nothing. The rationale here is that the other predictor variables are still applicable and valid (*FAQ What Is Complete or Quasi-complete Separation in Logistic/Probit Regression and How Do We Deal With Them?*, n.d.). Since this is just a warning the other methods to reduce the model's features meet the objectives of the analysis. The warning can be noted and dealt with as described above.

F2. Course of Action. Analyzing the outcomes of the analysis has yielded the following insights. Patients who were admitted as an emergency seem to show a high likelihood of being readmitted to the hospital. This is not an overly surprising finding. How does this translate into a course of action for the organization? There will need to be further analysis as to what were the other factors that could have contributed to the readmission. It might be beneficial to look at other features that are more tangible in nature looking at the continuous variables that gauge the health of the patient. Looking at full meals, vitD_levels (even if originally removed), and other variables that can quantify the care that was received while in the hospital.

Most of the features that have been included are not within the control of the hospital. For example, type of admission or high blood pressure.

While it is impossible to control the type of admission, it is possible to control the high blood pressure. The hospital can institute a program that may inform patients about the consequences of high blood pressure. The hospital can do more proactive treatment of this condition while the patient is in the hospital by information prescribing the right course of medications and making sure the follow up with the patient in a reasonable time frame.

This course of action can also be applied to patients who have a history of stroke. Patients who have had a stroke are very likely to return to the hospital. This is not surprising. Patients who had a stroke have more medical conditions besides the ones that are covered in the dataset. There is a need to look for more patient data to help these patients. This would also help with patients with high blood pressure.

The last course of action that should be undertaken is that there should be an attempt to get a more broad dataset. Observations for the gaps in data discussed earlier should be remedied. It will make the models more accurate when trying to predict those groups that exhibit a bimodal distribution. It is highly unlikely that this occurred and there have been omissions in the data for these groups.

Part VI. Panopto Demonstration.

G1. Panopto. In this section, you will find the link to the Panopto video. The link for the Panopto video is the following:

References

H. Web Sources.

This section will contain links to sites that were used. Note that this was in addition to the resources that were provided by WGU and the DataCamp videos. Links for all the code used in the assessment can be found below.

Creating multiple subplots using plt.subplots — Matplotlib 3.8.2 documentation. (n.d.). https://matplotlib.org/stable/gallery/subplots_axes_and_figures/subplots_demo.html

GeeksforGeeks. (2023, January 10). Detecting Multicollinearity with VIF   Python. https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

pandas.DataFrame.select_dtypes — pandas 2.1.3 documentation. (n.d.). Retrieved November 21, 2023, from https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.select_dtypes.html

Python Machine Learning - Confusion Matrix. (n.d.). Retrieved November 23, 2023, from https://www.w3schools.com/python/python_ml_confusion_matrix.asp

Regression Plots - statsmodels 0.15.0 (+73). (n.d). https://www.statsmodels.org/dev/examples/notebooks/generated/regression_plots.html

seaborn.histplot — seaborn 0.13.0 documentation. (n.d.). Retrieved November 21, 2023, from https://seaborn.pydata.org/generated/seaborn.histplot.html

Seaborn.violinplot() method. (n.d.). Retrieved November 21, 2023, from https://www.tutorialspoint.com/seaborn/seaborn_violinplot_method.html statsmodels.graphics.mosaicplot.mosaic - statsmodels 0.15.0 (+73). (n.d.). Retrieved November 22, 2023, from https://www.statsmodels.org/dev/generated/statsmodels.graphics.mosaicplot.mosaic.html

I. In-text Citations

In this section will find the citations that were used in the text of the document. These sources were not used in the creation of the code. These sources were only used within the text of this document.

Carron, J. (2021b, December 13). *Violin Plots 101: Visualizing Distribution and Probability Density | Mode*. Retrieved November 26, 2023, from https://mode.com/blog/violin-plot-examples/#:~:text=A%20violin%20plot%20is%20a,the%20density%20of%20each%20variable.

FAQ What is complete or quasi-complete separation in logistic/probit regression and how do we deal with them? (n.d.). UCLA. Retrieved November 27, 2023, from https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/#:~:text=If%20it%20

Kanade, V. (2022, June 10). *Linear vs. Logistic Regression*. Spiceworks. Retrieved November 26, 2023, from https://www.spiceworks.com/tech/artificial-intelligence/articles/linear-regression-vs-logistic-regression/

is%20quasi%2Dcomplete,predicts%20the%20outcome%20variable%20effectively.

Mosaic plot. (n.d.). Introduction to Statistics | JMP. Retrieved November 26, 2023, from https://www.jmp.com/en_us/statistics-knowledge-portal/exploratory-data-analysis/mosaic-plot.html

Statista. (2023, November 2). *Average length of stay in U.S. community hospitals from 1993 to 2020*. Retrieved November 26, 2023, from https://www.statista.com/statistics/183916/average-length-of-stay-in-us-community-hospitals-since-1993/

Straw, E. (n.d.). *Dr. Straw's Rx - 5 Linear Regression*. Retrieved December 2, 2023, from https://professorstraw.com/Rx/chapter5.html

Thanda, A. (2023, May 11). What is Logistic Regression? A Beginner's Guide [2023]. *CareerFoundry*. Retrieved November 26, 2023, from https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

Zach. (2021, May 19). *How to interpret an odds ratio less than 1*. Statology. https://www.statology.org/interpret-odds-ratio-less-than-1/