JANUARY 10, 2024

D208 PERFORMANCE ASSESSMENT

TASK TWO: LOGISTIC REGRESSION

ROBERT PATTON
WESTERN GOVERNORS' UNIVERSITY
Rpatt33@wgu.edu

Table of Contents

Part I: Purpose of Analysis

A. Research Question and Goals of Analysis

1. Do patients' health conditions correlate to the likelihood of hospital readmissions?
2. The goal of the hospital should be to reduce the number of hospital readmissions, thus improving patients' quality of care. In doing so, this implies that patients are receiving thorough examinations for all medical conditions stated on initial admission. For those medical conditions that are treatable at home, providing accurate diagnosis and treatment plans can decrease readmission rates from complications due to varying medical conditions. Extensive readmission rates tend to impact hospital revenue, deter current and future patients from seeking services, and can cause fines based on criteria via the Center for Medicare and Medicaid Services (Upadhyay et al. 2019). Therefore, it is imperative to identify which medical conditions are associated with higher rates of readmission and take more aggressive treatment measures to prevent such from happening.

Part II: Method Justification

B. Describe Logistic Regression and Python Benefits

1. The four assumptions of logistic regression are as follows:
   - The dependent variable must be a binary categorical variable. That is, the dependent variable should have a yes/ no, male/female, or any combination where there are only two values that cannot be ranked.
   - Logistic regression assumes a Bernoulli distribution of the dependent variable, meaning it does not require a linear relationship between the dependent and independent variables.
   - Predictor values are restricted to nominal values.
   - The independent variables do not experience high multicollinearity.
2. Python is a great program coding language that is user-friendly. It is a common tool used by a majority of data analysts who deal with large datasets, allowing them to work in an interactive development environment that makes analytic workflow processes easier to manage. I chose to use

Python because it will be a program that I will use when I complete the MSDA program and begin my new career. The Python language is a great tool because it contains packages that allow a user to access specifically designed libraries for statistical analysis, visualization of data, and other analytic tools that can provide stakeholders with insight into a research question.

3. Logistic regression is an appropriate technique to use when a research question involves a binary dependent variable, such as readmission, where the values are yes/no. This method allows an analyst to figure out whether a specific set of independent variables is related to the outcome of the selected dependent variable.

Part III: Data Preparation

C. Data Preparation Process

1. We need to clean and transform the data in a way that ensures all our variables are accurate and meaningful. By doing so, we can then perform various analytic tests and create a predictive model that will produce accurate results about the research question.

Data preparation techniques are as follows:

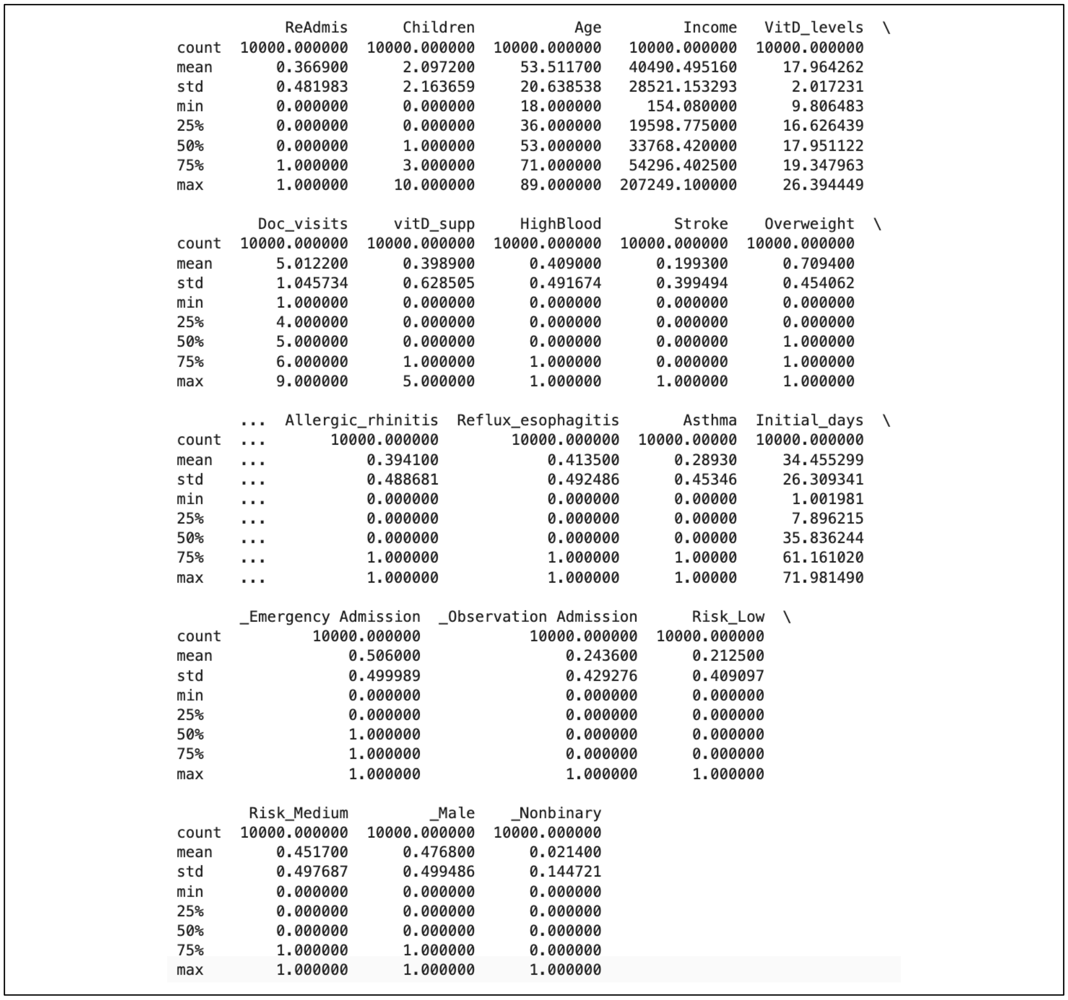
- Import all necessary and required Python libraries for statistical information and data visualization.
- Read in our CSV file.
- Examine our data types, structure, and variables.
- Detect and eliminate any duplicates.
- Detect and address any null/missing values by using the replace method with mean, median, or mode.
- Rename any variables to improve the clarity of the data frame.
- Examine the variables as univariate and bivariate visualizations.
- Detect any outliers in the data variables.
- Drop any variables I feel are irrelevant to my research question.
- Transform categorical data into numerical data types with dummy codes.

- *See code attached with submission*

2. **Summary Statistics**

- The dataset originally consisted of 10000 rows and 50 columns. After cleaning the data and selecting the variables for linear regression, the dataset consisted of 10000 rows and 25 columns.
- For this performance assessment, my dependent variable is ReAdmis. ReAdmis is a binary object data type, presenting as a value organized into a category, such as yes or no. It is being used to determine whether or not a patient's readmission is correlated to their overall health.
- My independent variables, such as children, age, Income, VitD_levels, Doc_visits, vitD_supp, high blood, stroke, overweight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic_rhinitis, reflux_esophagitis, asthma, initial_days, _emergency admission, _observation admission, risk_low, risk_medium, _male, and _nonbinary, are all observations taken from the data that I feel may contribute to a patients readmission probability due to their overall health.
- Children, age, Doc_visits, and vitD_supp are all integer data types, meaning they are values that present as whole numbers.
- Income, VitD_levels, and Initial_days are all float/continuous data types. Meaning they are continuous numbers that include a decimal.
- Gender, Initial_admin, HighBlood, Stroke, Overweight, Complication_risk, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic_rhinitis, Reflux_esophagitis, and Asthma are all object data types. This mean that they present as values that are organized into categories, such as yes/no, male/female/nonbinary, or as having a low, medium, or high complication risk.
- After preparing all the data as described in the steps above, I then obtained my summary of statistics for non-categorical/quantitative

variable types. Reviewing dataset summary statistics is valuable to an analyst because it can provide information on a few things. It can help determine where the midpoint, or central tendency, in a dataset, using the mean, median, and mode. We can also determine the dispersion in a dataset using standard deviation, this can tell an analyst what values may be at one, two, or three standard deviations away from our mean. Summary statistics also tell an analyst the shape of a dataset’s distribution, whether it is normal, skewed, bimodal, or uniform. Furthermore, by looking at the summary statistics for this performance assessment, I was able to see what the minimum and maximum values were for each variable and the 25%, 50%, and 75% interquartile ranges. Interquartile ranges help an analyst determine the spread of the middle portion of a dataset, helping to determine outliers and treat them.

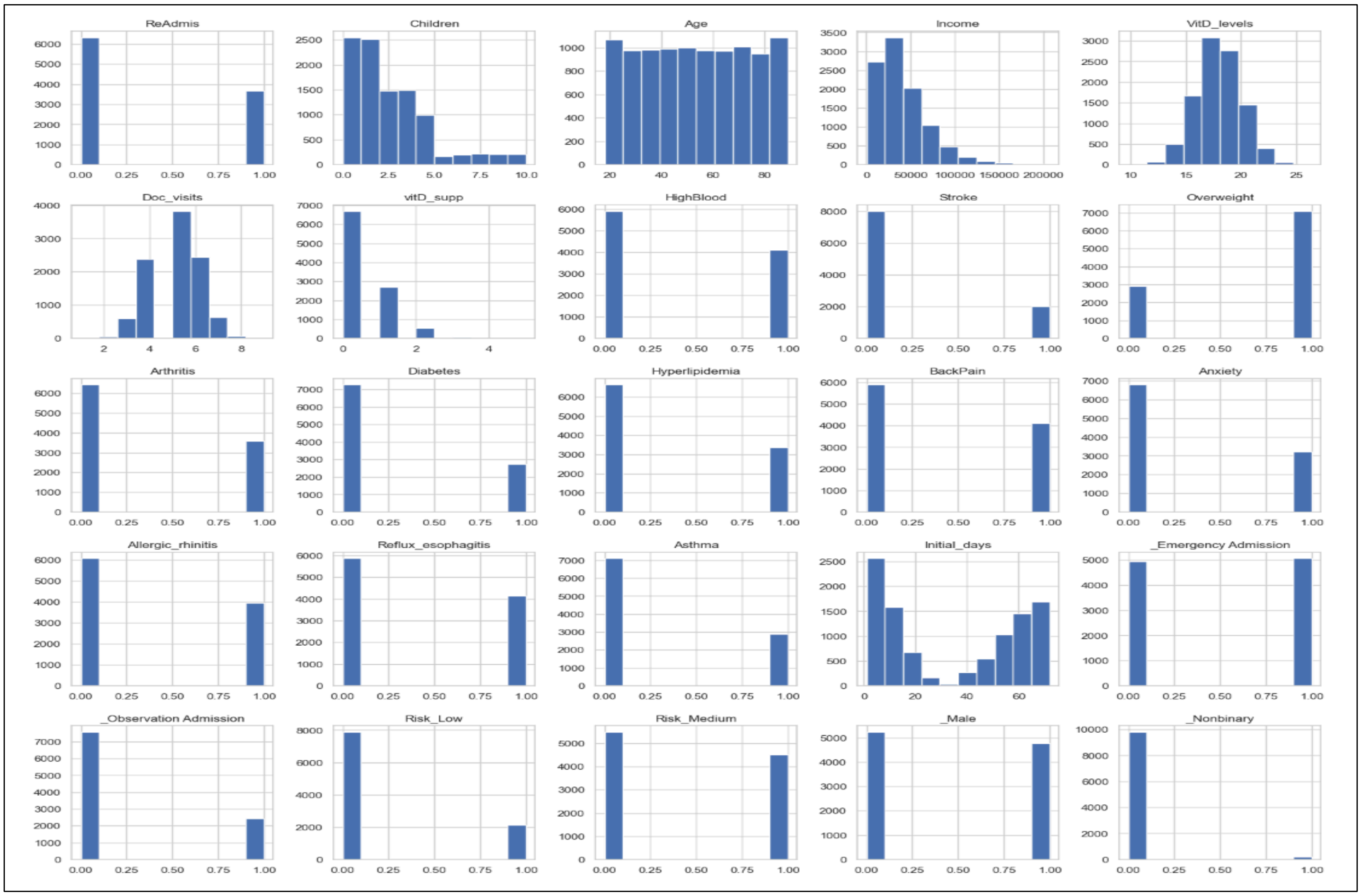


```
          ReAdmis      Children           Age         Income   VitD_levels  \
count  10000.000000  10000.000000  10000.000000   10000.000000  10000.000000
mean       0.366900      2.097200     53.511700   40490.495160     17.964262
std        0.481983      2.163659     20.638538   28521.153293      2.017231
min        0.000000      0.000000     18.000000     154.080000      9.806483
25%        0.000000      0.000000     36.000000   19598.775000     16.626439
50%        0.000000      1.000000     53.000000   33768.420000     17.951122
75%        1.000000      3.000000     71.000000   54296.402500     19.347963
max        1.000000     10.000000     89.000000  207249.100000     26.394449

        Doc_visits     vitD_supp     HighBlood        Stroke    Overweight  \
count  10000.000000  10000.000000  10000.000000  10000.000000  10000.000000
mean       5.012200      0.398900      0.409000      0.199300      0.709400
std        1.045734      0.628505      0.491674      0.399494      0.454062
min        1.000000      0.000000      0.000000      0.000000      0.000000
25%        4.000000      0.000000      0.000000      0.000000      0.000000
50%        5.000000      0.000000      0.000000      0.000000      1.000000
75%        6.000000      1.000000      1.000000      0.000000      1.000000
max        9.000000      5.000000      1.000000      1.000000      1.000000

       ...  Allergic_rhinitis  Reflux_esophagitis       Asthma  Initial_days  \
count  ...       10000.000000        10000.000000  10000.00000  10000.000000
mean   ...           0.394100            0.413500      0.28930     34.455299
std    ...           0.488681            0.492486      0.45346     26.309341
min    ...           0.000000            0.000000      0.00000      1.001981
25%    ...           0.000000            0.000000      0.00000      7.896215
50%    ...           0.000000            0.000000      0.00000     35.836244
75%    ...           1.000000            1.000000      1.00000     61.161020
max    ...           1.000000            1.000000      1.00000     71.981490

       _Emergency Admission  _Observation Admission      Risk_Low  \
count          10000.000000            10000.000000  10000.000000
mean               0.506000                0.243600      0.212500
std                0.499989                0.429276      0.409097
min                0.000000                0.000000      0.000000
25%                0.000000                0.000000      0.000000
50%                1.000000                0.000000      0.000000
75%                1.000000                0.000000      0.000000
max                1.000000                1.000000      1.000000

        Risk_Medium         _Male    _Nonbinary
count  10000.000000  10000.000000  10000.000000
mean       0.451700      0.476800      0.021400
std        0.497687      0.499486      0.144721
min        0.000000      0.000000      0.000000
25%        0.000000      0.000000      0.000000
50%        0.000000      0.000000      0.000000
75%        1.000000      1.000000      0.000000
max        1.000000      1.000000      1.000000
```

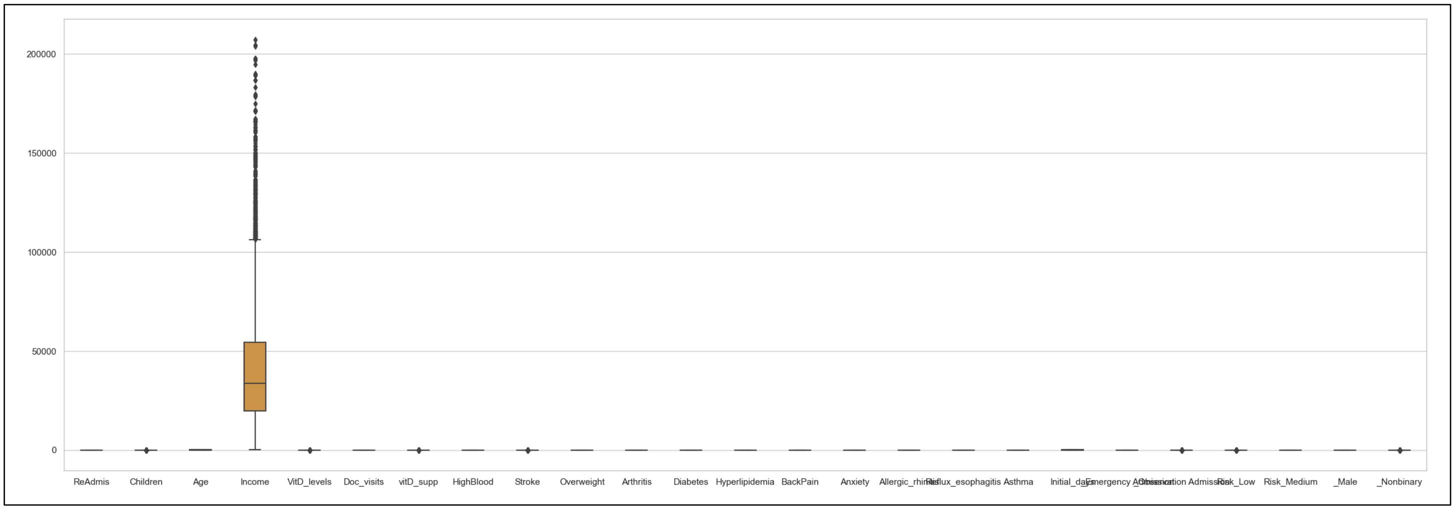
- Using the .value_counts() function, I can obtain the exact number of values for each response in the categorical/qualitative variables.
    1. Initial_admin has three values, emergency admission with 5060 responses, observation admission with 2436 responses, and elective admission with 2504 responses.
    2. HighBlood had 5910 no response and 4090 yes responses.
    3. Stroke had 8007 no responses and 1993 yes responses.
    4. Complication_risk has three values, low with 2125 responses, medium with 4517 responses, and high with 3358 responses.
    5. Overweight had 2906 no responses and 7094 yes responses.
    6. Arthritis had 6426 no responses and 3574 yes responses.
    7. Diabetes had 7262 no responses and 2738 yes responses.
    8. Hyperlipidemia had 6628 no responses and 3372 yes responses.
    9. BackPain had 5886 no responses and 4114 yes responses.
    10. Anxiety had 6785 no responses and 3215 yes responses.
    11. Allergic_rhinitis had 6059 no responses and 3941 yes responses.
    12. Reflux_esophagitis had 5865 no responses and 4135 yes responses.
    13. Asthma had 7107 no responses and 2893 yes responses.
    14. Gender had three values as well, male with 4768 responses, female, with 5018 responses and non-binary with 214 responses.
    15. ReAdmis has 6331 no responses, and 3669 yes responses.

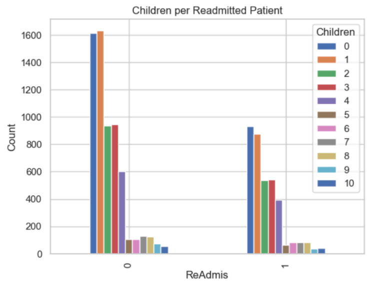
3. **Generate Univariate and Bivariate Visualizations**

- Both univariate and bivariate visualizations can help an analyst determine which variables will be used for multiple linear regression. They enable an analyst to see the types of distributions for each data type and provide insight into how the dependent variable is compared to the potential predictor variables.

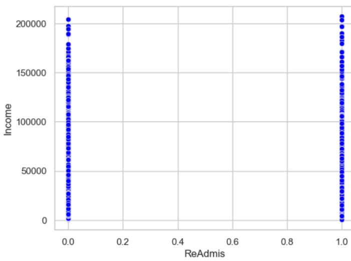


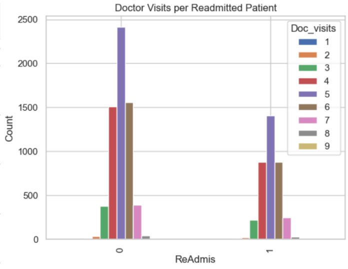
- Because this data has already been cleaned and although there appear to be variables not evenly distributed, we can create boxplots to look for outliers. Even though the boxplots show that some outliers exist, I assume that since this data has been cleaned, it was determined that the range of these outliers had been reduced enough and any further cleaning or reduction of outliers could compromise the integrity of certain values.

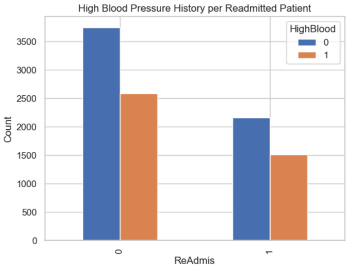




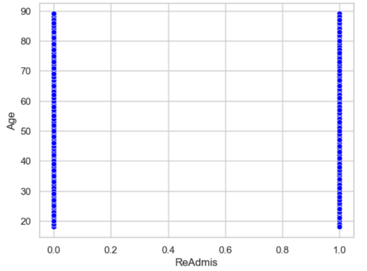
```
#Scatterplot bivariate relationship for Income
sns.scatterplot(x=mdf['ReAdmis'], y=mdf['Income'], color='Blue')
plt.show();
```



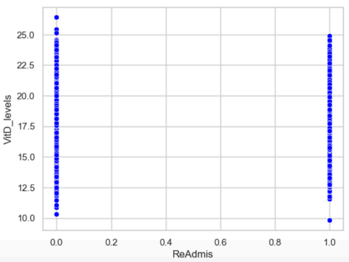




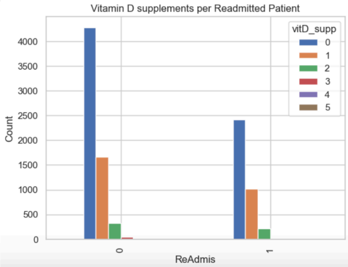
```
#Scatterplot bivariate relationship for age
sns.scatterplot(x=mdf['ReAdmis'], y=mdf['Age'], color='Blue')
plt.show();
```



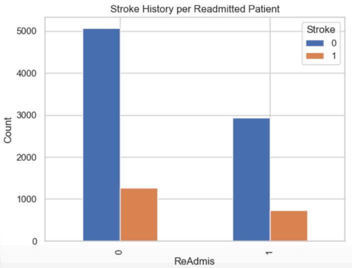
```
#Scatterplot bivariate relationship for VitD_levels
sns.scatterplot(x=mdf['ReAdmis'], y=mdf['VitD_levels'], color='Blue')
plt.show();
```

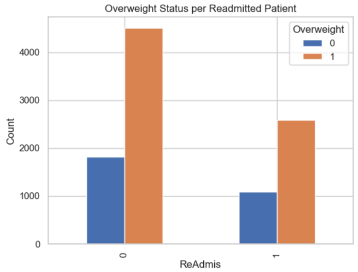


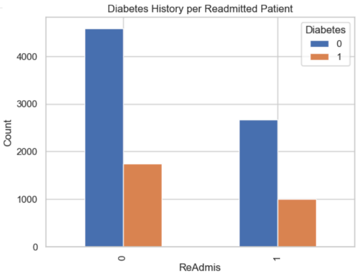
```
#Bar chart relationship for vitD supplements
crosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['vitD_supp'])
crosstab.plot(kind='bar')
plt.ylabel('Count')
plt.title('Vitamin D supplements per Readmitted Patient')
plt.show()
```

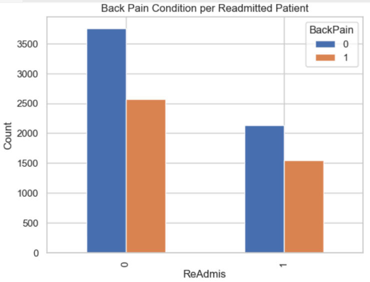


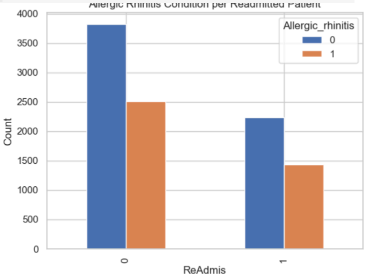
```
#Bar chart relationship for Stroke
crosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['Stroke'])
crosstab.plot(kind='bar')
plt.ylabel('Count')
plt.title('Stroke History per Readmitted Patient')
plt.show()
```



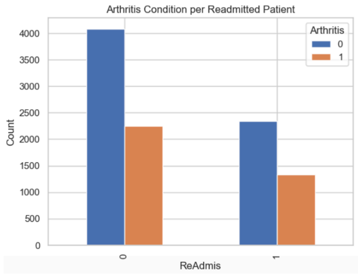




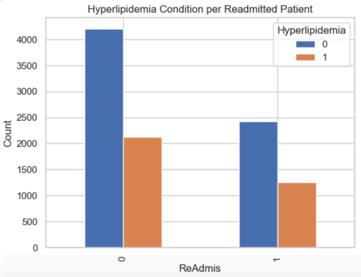




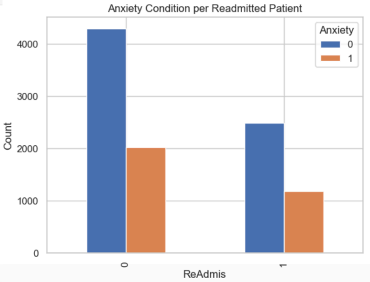
```
#Bar chart relationship for Arthritis
crosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['Arthritis'])
crosstab.plot(kind='bar')
plt.ylabel('Count')
plt.title('Arthritis Condition per Readmitted Patient')
plt.show()
```



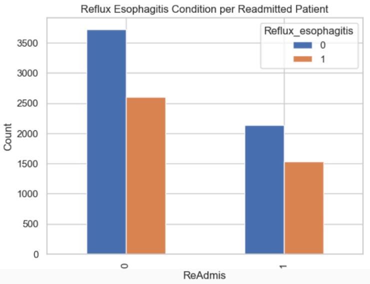
```
#Bar chart relationship for Hyperlipidemia
:rosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['Hyperlipidemia'])
:rosstab.plot(kind='bar')
)lt.ylabel('Count')
)lt.title('Hyperlipidemia Condition per Readmitted Patient')
)lt.show()
```

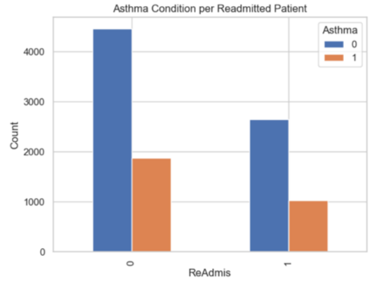


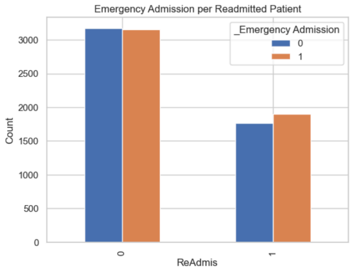
```
#Bar chart relationship for Anxiety
:rosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['Anxiety'])
:rosstab.plot(kind='bar')
)lt.ylabel('Count')
)lt.title('Anxiety Condition per Readmitted Patient')
)lt.show()
```

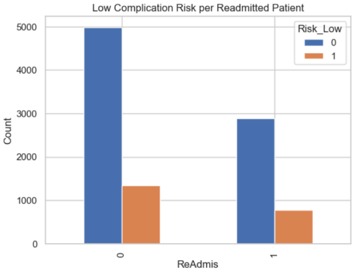


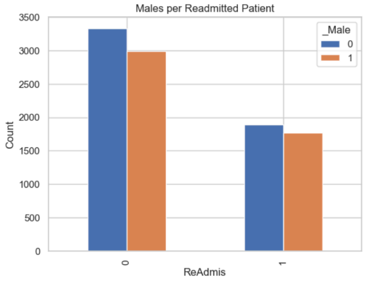
```
#Bar chart relationship for Reflux Esophagitis
:rosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['Reflux_esophagitis
:rosstab.plot(kind='bar')
)lt.ylabel('Count')
)lt.title('Reflux Esophagitis Condition per Readmitted Patient')
)lt.show()
```



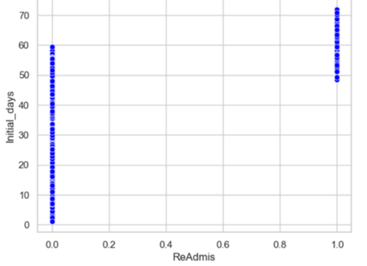




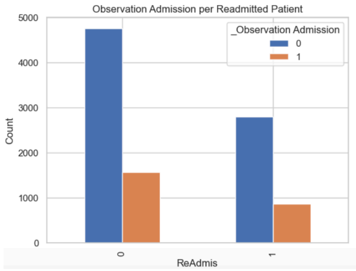




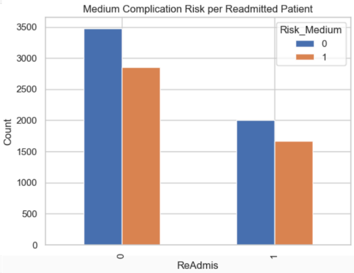
```
#Scatterplot bivariate relationship for initial days
sns.scatterplot(x=mdf['ReAdmis'], y=mdf['Initial_days'], color='Blue')
)lt.show();
```



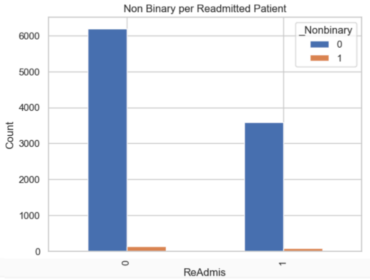
```
#Bar chart relationship for Initial admin observation
crosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['_Observation Admis
crosstab.plot(kind='bar')
plt.ylabel('Count')
plt.title('Observation Admission per Readmitted Patient')
plt.show()
```



```
#Bar chart relationship for complication risk medium
:rosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['Risk_Medium'])
:rosstab.plot(kind='bar')
)lt.ylabel('Count')
)lt.title('Medium Complication Risk per Readmitted Patient')
)lt.show()
```



```
#Bar chart relationship for gender of non-binary
:rosstab= pd.crosstab(index=mdf['ReAdmis'], columns=mdf['_Nonbinary'])
:rosstab.plot(kind='bar')
)lt.ylabel('Count')
)lt.title('Non Binary per Readmitted Patient')
)lt.show()
```

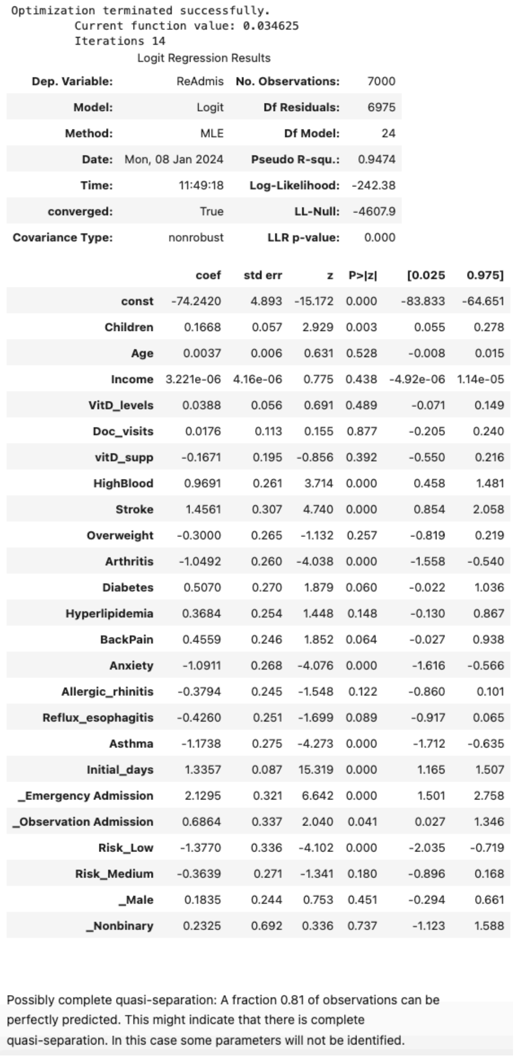


Part IV: Model Comparison and Analysis

D. Compare an initial and reduced logistic regression model

1. Initial Model

- An initial logistic model is run on the independent/predictor variables. These are compared against the dependent variable to determine a relationship and answer our research question. The logistic regression model can be seen below.
- The initial regression model has a pseudo R-squared value of 0.9474. When looking at pseudo R-squared values, an analyst is looking for a value that is closer to one. With the value being 0.9474, the model indicates a strong relationship between the predictor variables and the dependent variable. The LLR p-value is also within a range of likeness to an analyst, at 0.00, where p-values at or below 0.05 indicate a statistical significance.



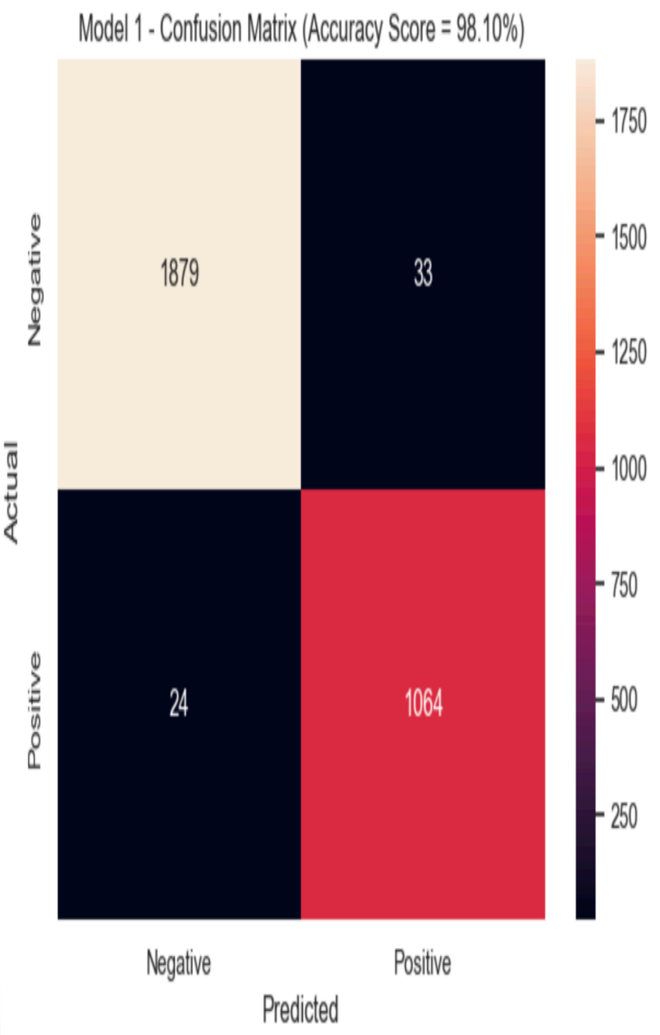
```
Optimization terminated successfully.
         Current function value: 0.034625
         Iterations 14
```

Logit Regression Results

Dep. Variable:	ReAdmis	**No. Observations:**	7000
Model:	Logit	**Df Residuals:**	6975
Method:	MLE	**Df Model:**	24
Date:	Mon, 08 Jan 2024	**Pseudo R-squ.:**	0.9474
Time:	11:49:18	**Log-Likelihood:**	-242.38
converged:	True	**LL-Null:**	-4607.9
Covariance Type:	nonrobust	**LLR p-value:**	0.000

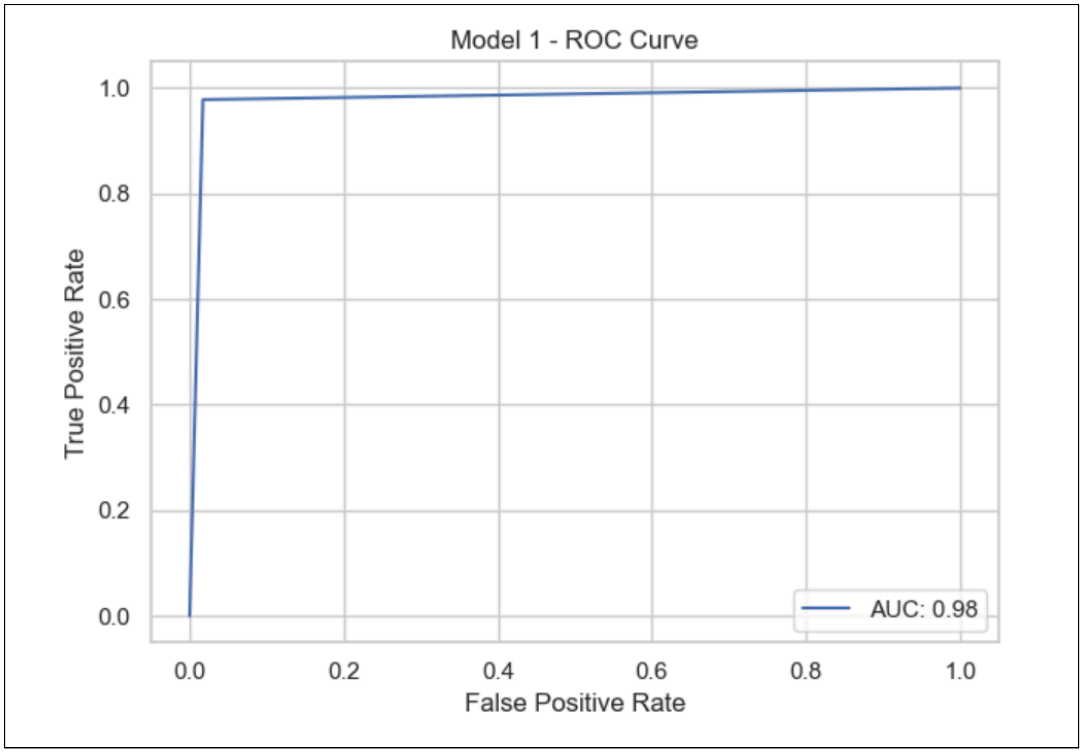
	coef	std err	z	P>|z|	[0.025	0.975]
const	-74.2420	4.893	-15.172	0.000	-83.833	-64.651
Children	0.1668	0.057	2.929	0.003	0.055	0.278
Age	0.0037	0.006	0.631	0.528	-0.008	0.015
Income	3.221e-06	4.16e-06	0.775	0.438	-4.92e-06	1.14e-05
VitD_levels	0.0388	0.056	0.691	0.489	-0.071	0.149
Doc_visits	0.0176	0.113	0.155	0.877	-0.205	0.240
vitD_supp	-0.1671	0.195	-0.856	0.392	-0.550	0.216
HighBlood	0.9691	0.261	3.714	0.000	0.458	1.481
Stroke	1.4561	0.307	4.740	0.000	0.854	2.058
Overweight	-0.3000	0.265	-1.132	0.257	-0.819	0.219
Arthritis	-1.0492	0.260	-4.038	0.000	-1.558	-0.540
Diabetes	0.5070	0.270	1.879	0.060	-0.022	1.036
Hyperlipidemia	0.3684	0.254	1.448	0.148	-0.130	0.867
BackPain	0.4559	0.246	1.852	0.064	-0.027	0.938
Anxiety	-1.0911	0.268	-4.076	0.000	-1.616	-0.566
Allergic_rhinitis	-0.3794	0.245	-1.548	0.122	-0.860	0.101
Reflux_esophagitis	-0.4260	0.251	-1.699	0.089	-0.917	0.065
Asthma	-1.1738	0.275	-4.273	0.000	-1.712	-0.635
Initial_days	1.3357	0.087	15.319	0.000	1.165	1.507
_Emergency Admission	2.1295	0.321	6.642	0.000	1.501	2.758
_Observation Admission	0.6864	0.337	2.040	0.041	0.027	1.346
Risk_Low	-1.3770	0.336	-4.102	0.000	-2.035	-0.719
Risk_Medium	-0.3639	0.271	-1.341	0.180	-0.896	0.168
_Male	0.1835	0.244	0.753	0.451	-0.294	0.661
_Nonbinary	0.2325	0.692	0.336	0.737	-1.123	1.588

Possibly complete quasi-separation: A fraction 0.81 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.



- In addition to the logistic model, an initial confusion matrix heatmap, (shown above) was created to obtain an accuracy score for predicting readmission based on the predictor variables. The confusion matrix heatmap helps an analyst determine the correct and incorrect number of predictions from the performance of the logistic model. A confusion matrix heatmap consists of the true negatives (TN) in the top left corner, the false positives (FP) in the top right corner, the false negatives (FN) in the bottom left corner, and the true positives (TP) in the bottom right corner. Our accurate predictions are 1879 true negatives and 1064 true positives. Our inaccurate predications are 33 false positives and 24 false negatives. Our accuracy score for the initial model appears to show the model can accurately predict readmission 98 percent of the time, however. To really ensure that model is efficient in prediction, we should look into other metrics of the model's performance. Accuracy is determined by using the formula TN+TP/TN+FP+TP+FN.
- After determining accuracy, I created a classification report using the recall, specificity, and precision to further examine the model's performance. Recall tells an analyst how well the model predicts positives out of all positive values, with the formula being TP/TP+FN. Specificity tells an analyst how well the model predicts negative values out of all negatives, with the formula being TN/TN+FP. Precision tells an analyst how well the positive predictions perform based on all positive predictions that were made, the formula being TP/TP+FP.
- For the initial model, the classification report is as follows:
  Recall: 97.79%
  Specificity: 98.27%
  Precision: 96.99%
  Looking at the results from the classification report, the initial model appears to perform very well overall.

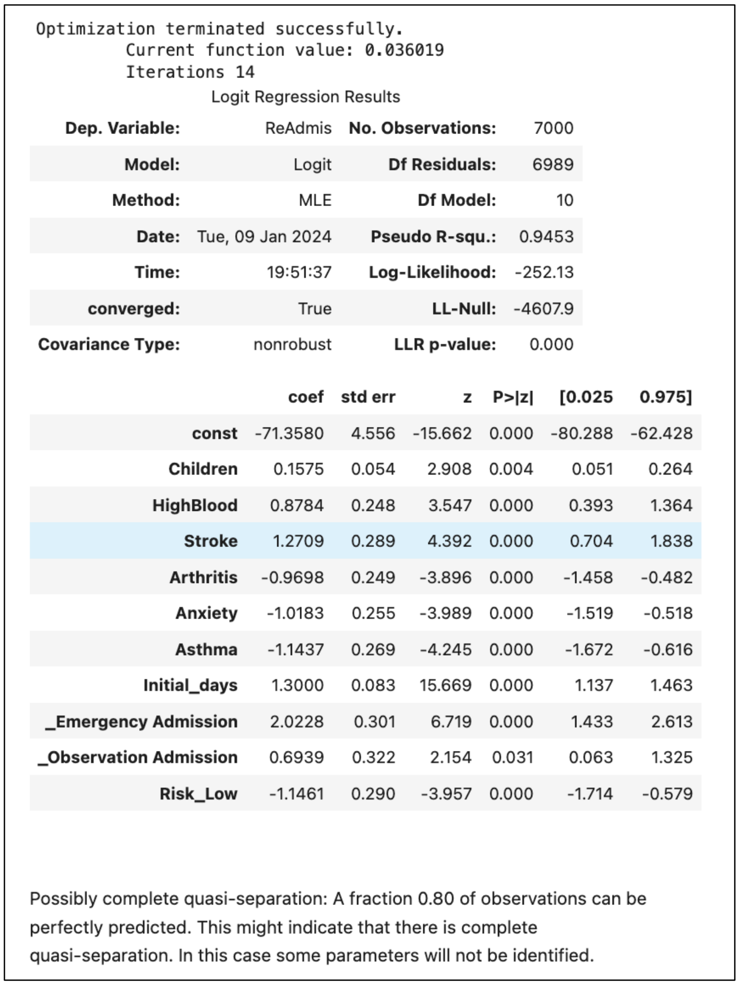
- I also created an ROC curve and looked at the AUC value which also helps determine model performance. An ROC curve plots true positives against false positives. An AUC score is the area under the curve and goes from 0-1, or 0% to 100%. The higher the AUC score the better. In this case, the initial models ROC curve had an AUC score of 98%. Again it appears that the initial logistic regression model performs very well. Below is a screen shot of my ROC curve for justifying the initial model.



2. Model Reduction Method and Justification

- Following the construction of my initial model and the confusion matrix, I used a backwards stepwise elimination method to reduce my initial model, using p-values to determine which variables should be removed. After constructing the initial logistic model, I was able to look and see which variables p-values were over the 0.05 threshold, indicating a lack of statistical significance.
- Looking at my initial model, it appears that age, income, VitD_levels, doc_visits, vitD_supp, Overweight, diabetes, hyperlipidemia, back pain, allergic_rhinitis, reflux_esophagitis, risk_medium, _male, and _nonbinary all has p-values that exceed the 0.05 threshold. Thus the backward stepwise elimination feature selection process should exclude them from the reduced model.

3. Reduced Model



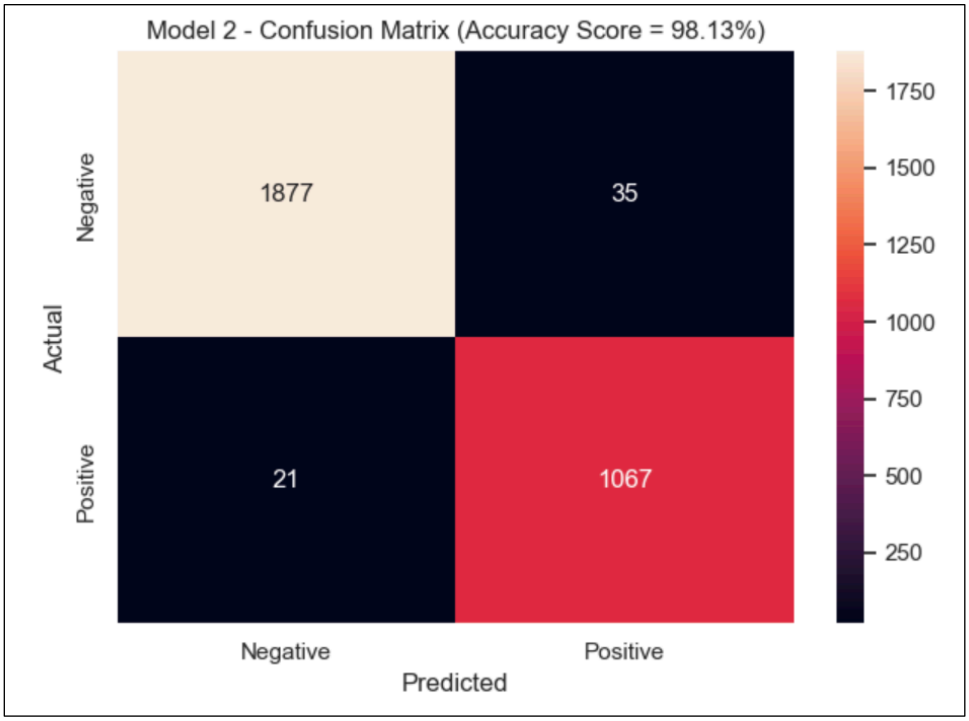
```
Optimization terminated successfully.
         Current function value: 0.036019
         Iterations 14
```

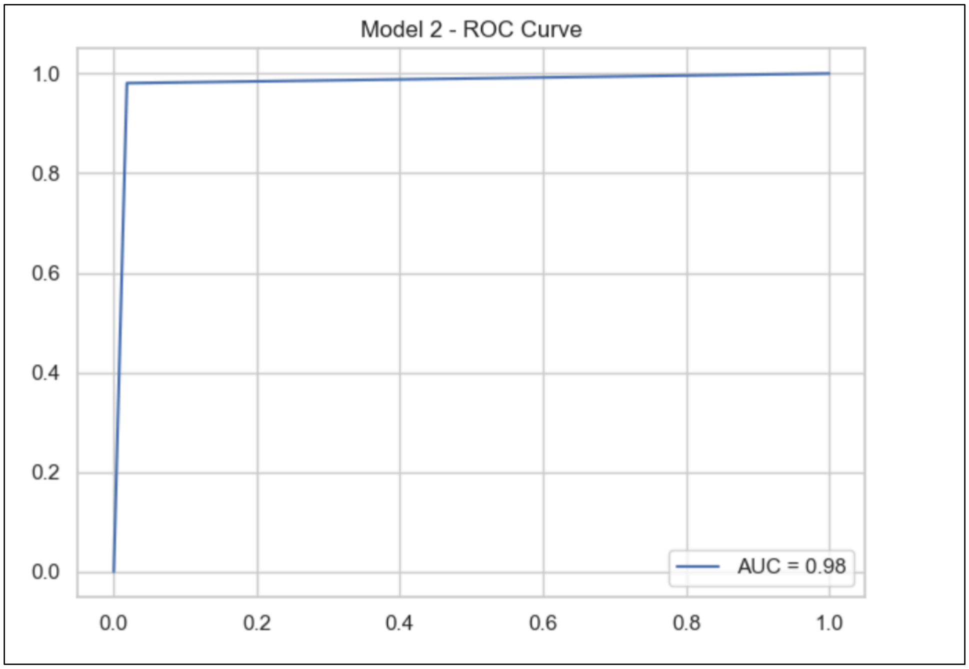
Logit Regression Results

Dep. Variable:	ReAdmis	**No. Observations:**	7000
Model:	Logit	**Df Residuals:**	6989
Method:	MLE	**Df Model:**	10
Date:	Tue, 09 Jan 2024	**Pseudo R-squ.:**	0.9453
Time:	19:51:37	**Log-Likelihood:**	-252.13
converged:	True	**LL-Null:**	-4607.9
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-71.3580	4.556	-15.662	0.000	-80.288	-62.428
Children	0.1575	0.054	2.908	0.004	0.051	0.264
HighBlood	0.8784	0.248	3.547	0.000	0.393	1.364
Stroke	1.2709	0.289	4.392	0.000	0.704	1.838
Arthritis	-0.9698	0.249	-3.896	0.000	-1.458	-0.482
Anxiety	-1.0183	0.255	-3.989	0.000	-1.519	-0.518
Asthma	-1.1437	0.269	-4.245	0.000	-1.672	-0.616
Initial_days	1.3000	0.083	15.669	0.000	1.137	1.463
_Emergency Admission	2.0228	0.301	6.719	0.000	1.433	2.613
_Observation Admission	0.6939	0.322	2.154	0.031	0.063	1.325
Risk_Low	-1.1461	0.290	-3.957	0.000	-1.714	-0.579

Possibly complete quasi-separation: A fraction 0.80 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.


Model 2 - Confusion Matrix (Accuracy Score = 98.13%)
Actual
Negative
Positive
1877
35
21
1067
Negative
Positive
Predicted
1750
1500
1250
1000
750
500
250


Model 2 - ROC Curve
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.98

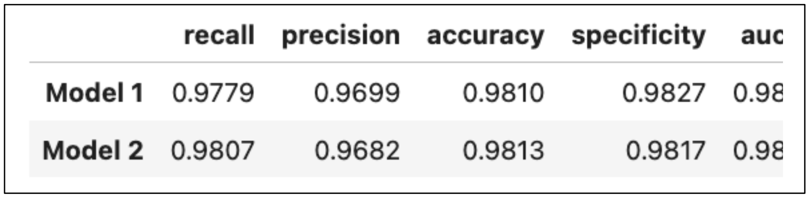
E. Model Comparison

1. Compare Initial and Reduced Model

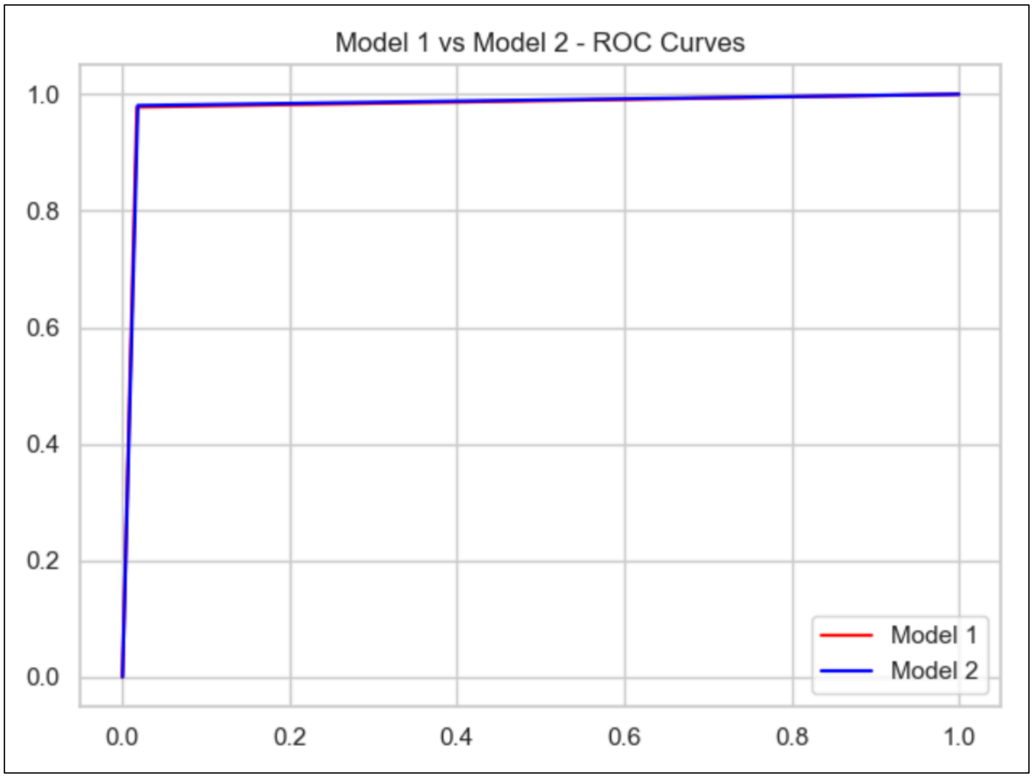
- During the initial logistic regression model, I observed the p-values for each of the predictor variables, the model accuracy via a confusion matrix, the recall, specificity, and precision metrics via a classification report, and an ROC curve with an AUC value. Evaluating all of these metrics allowed me to implement a backward stepwise elimination process and construct a reduced logistic regression model. In addition, another evaluation metric used to compare the two models is the LLR p-value, where both models' values stood at 0.00 and indicate statistical significance between the predictor and target variables.
- Comparing the initial and reduced model, I used these evaluation metrics. A confusion matrix heat map and ROC for the reduced model can be seen above. For the initial model, accuracy was calculated to be 98.10% versus the reduced model at 98.13%, just a touch better. Recall for the initial model was calculated at 97.79% versus 98.07% for the reduced model, again performing a bit better in the reduced model. Precision for the initial model was calculated at 96.99% versus 96.82% in the reduced model, the initial model's precision may have been higher due to there being more variables. Specificity for the initial model came to 98.27% versus 98.17% in the reduced model, again more variables initially may have contributed to the initial model performing better here. Finally, the AUC value in the ROC curve for the initial and reduced model were both 98%. Overall, it appears that the reduced model performs better, even though both models perform relatively the same, the reduced model is easier to explain. The comparison of both model performances can be seen below, along with an ROC/AUC comparison.

2. Executable Code

- Please see code attached as a file.



	recall	precision	accuracy	specificity	auc
Model 1	0.9779	0.9699	0.9810	0.9827	0.98
Model 2	0.9807	0.9682	0.9813	0.9817	0.98



Part V: Data Summary and Implications

F. Results of Analysis

1. Regression Equation of Reduced Model

- ReAdmis= (Children* 0.1575)+(HighBlood* 0.8784)+(Stroke* 1.2709)+(Arthritis* -0.9698)+(Anxiety*-1.0183)+(Asthma* -1.1437)+(Initial_days* 1.3)+(_Emergency Admission* 2.0228)+(_Observation Admission* 0.6939)+(Risk_Low* -1.1461) -71.3580 const

2. Interpretation of Coefficients

- With interpreting coefficients, you either get a positive or negative relationship to the dependent variable. My dependent variable is ReAdmis. My predictor variables coefficients are:
  Children 0.1575, HighBlood 0.8784, Stroke 1.2709, Arthritis -0.9698, Anxiety -1.0183, Asthma -1.1437, Initial_days 1.3, _Emergency Admission 2.0228, _Observation Admission 0.6939, and Risk_Low -1.1461. To elaborate, for every one unit change of Initial_days, Readmission will increase by 1.3 days, whereas for every one unit of change in Observation Admission, readmission will increase 0.6939 admissions.

3. Statistical and Practical Significance of Reduced Model

- Even though the initial model performed really well, we still had to drop some variables because the p-values indicated a lack of statistical significance. In the reduced model, we retained a total of 10 independent variables according to p-values. Our goal was to determine if health conditions had a correlation to readmission rates, however. We can see that in the reduced model, five of the variables retained are not health conditions, rather they are observations and parameters of admittance to the hospital initially. This is not to say that those variables do not have a relationship to health conditions themselves. For example, maybe during the initial days a patient was admitted, they had a hospital acquired infection that required a longer stay, that infection continued even after discharge and required hospital readmission. Therefore an analyst would need to consider the statistical significance or potential relationship of non-health related variables to actual medical conditions/ causes of readmission.

- In the reduced model, our retained variables all had p-values at or below the 0.05 threshold. This tells an analyst that there is a statistical significance between the independent variables retained and the dependent variable. In other words, we were able to answer our research question that yes, some health conditions are correlated to the potential of a hospital readmission. Those health conditions most likely to be related to a readmission are high blood pressure, a stroke history, arthritis, anxiety, and asthma. Therefore, a patient with these health conditions is more like to be readmitted to the hospital than patients without these health conditions.
- For this research question, it can be said that there is a practical significance for the reduced model. That means that this model has meaningful insight to a hospital. Knowing that certain patients, with specific medical conditions, are at higher risk of hospital readmission, can put administrators and doctors on high alert to be more detailed in patient assessment and initial treatments. This is especially so during a patient's initial admission, with a goal to prevent hospital readmission. As we mentioned in part one, the goals of this analysis are to identify these patients, reduce hospital readmissions, and prevent fines from the Center for

Medicare and Medicaid Services. Excessive hospital readmission rates can have negative implications for a hospital, this includes deterring patients from seeking medical help at that hospital, impacting revenue because of fines from the Center for Medicare and Medicaid Services, and give a bad image of the hospital to community members.

4. **Limitations of the Data Analysis**

- An analyst may need to run more than just one regression model to find an answer to the organizational question. This regression model consists of predictions and changing variables can improve the overall results. In addition, some medical conditions can be resolved with basic lifestyle changes, such as diet and exercise, if there was a significant decrease in a health condition within the community, this could change the overall outcome of the model. Another limitation of this analysis is that during specific times of the year, hospital admission rates tend to increase, especially in the winter months. So this research question could have a large degree of variation depending on the time of year.
- Because the initial model results required the removal of many predictor variables, we would need to reobserve the datapoints and potentially select more variables for the regression model. This could potentially improve overall significance to the dependent variable and lead the analyst to finding stronger correlation results. There is also a great deal of medical conditions not listed in the data analysis, rather it appears just some of the most common medical conditions are inputted in the data frame. An analyst could request further surveys on medical conditions of patients and all diagnosis to the data frame, which could give a stronger insight into overall health conditions and

likelihood of being admitted and readmitted to the hospital.

- Lastly, a recommended course of action an analyst could take, is to ask the hospital to reframe the research question. Suggesting to not only look at readmission in relation to overall health but specific demographics that emphasize a patient's overall health, such as living in poverty which may contribute to medical conditions.

Part VI: Demonstration

G: A link to a Panopto video will be included and uploaded with the submission. This will demonstrate the execution of my code and elaborate on discussion of my models.

H: Sources

Upadhyay, S., Stephenson, A. L., & Smith, D. G. (2019). Readmission Rates and Their Impact on Hospital Financial Performance: A Study of Washington Hospitals. Inquiry : a journal of medical care organization, provision, and financing, 56, 46958019860386. https://doi.org/10.1177/0046958019860386

Ali Awan, A. (2022, July 27). *Understanding Logistic Regression in Python*. Data Camp. Retrieved January 10, 2024, from https://app.datacamp.com/workspace/w/11b0dd9d-49be-429d-992e-45f73b3c23fb