Predictive Modeling – D208

Task 2

Western Governor's University

Performance Assessment

Matthew Morgan

Student ID: 010471280

11/28/2022

Part I: Research Question
A1. What variables lead to patient readmissions?

A2. Hospitals are penalized by an external organization for excessive readmissions. To help prevent the hospital from being penalized we need to identify factors that lead to patient readmissions. Once those factors are identified we can hopefully find ways to reduce patient readmissions.

Part II: Method Justification
B1. Logistic regression predicts whether something is True or False, instead of something continuous. Therefore, we are using a categorical variable for logistic regression. Logistic regression will fit an s-shaped line to the predictions to produce predictions. This line allows us to predict the probability of a prediction and use the likelihood of that prediction to classify it within the categorical variable. "The logistic regression model is based on different assumptions than linear regression:

- It is based on the Bernoulli distribution because the dependent variable is binary.
- The predict values are restricted to a range of nominal values like 'Yes' and 'No', not Small, Medium, Large.
- It predicts the probability of particular outcomes rather than the outcome itself.
- It is the logarithm of the odds of achieving 1." (Sewell, 2022)

B2. I am using Python as it's the language I am most comfortable with and its versatility. Python has many packages and libraries available to run logistic regression easily and quickly. Because of its versatility, you can run the whole ETL pipeline in one python script.

B3. Logistic regression is used specifically for categorical variables and can be used in conjunction with other categorical or continuous variables to make predictions. Because we are using a categorical variable as our dependent variable in this task, we want to just logistic regression and build a model that can predict future patient readmissions.

Part III: Data Preparation
C1.

- Import medical_clean.csv into Jupyter Notebook
- Build boxplots to check for outliers
- Convert Yes/No into quantitative data
- Use pd.get_dummies to convert categorical variables into quantitative data
- Remove redundant columns (Marital_Divorced, Gender_Female to reduce possibilities of multicollinearity
- Rename columns from pd.get_dummies by replacing spaces with underscores
- Run univariate stats script to calculate how many rows, missing values, unique values, data type, Mean, Mode, Min, Median, Max, Standard Deviation, Skew, Kurtosis for each numeric column.
- Export cleaned data set

C2. The target categorical variable I chose for this task was ReAdmis which has been converted to ReAdmis_numeric. Because it is categorical we can use it for logistic regression. The predictor variables I chose were the following; Initial_days, vitD_supp, Children, Income, Full_meals_eaten, Additional_charges, TotalCharge, VitD_levels, Age, Doc_visits, HighBlood_numeric, Stroke_numeric, Arthritis_numeric, Diabetes_numeric, Hyperlipidemia_numeric, BackPain_numeric, Allergic_rhinitis_numeric, Reflux_esophagitis_numeric, Asthma_numeric, Overweight_numeric, Anxiety_numeric, Marital_Married, Marital_Never_Married, Marital_Separated, Marital_Widowed, Services_Blood_Work, Services_CT_Scan, Services_Intravenous, Services_MRI, Gender_Male, Gender_Nonbinary, Initial_admin_Elective_Admission, Initial_admin_Emergency_Admission, Initial_admin_Observation_Admission, Complication_risk_High, Complication_risk_Low, and Complication_risk_Medium

I included screenshots of summary statistics below:

Summary statistics

```
                                     Count  Missing  Unique    Dtype  Numeric  \
Gender_Nonbinary                     10000        0       2    uint8     True   
Services_MRI                         10000        0       2    uint8     True   
Services_CT_Scan                     10000        0       2    uint8     True   
Population                           10000        0    5951    int64     True   
vitD_supp                            10000        0       6    int64     True   
Marital_Never_Married                10000        0       2    uint8     True   
Marital_Separated                    10000        0       2    uint8     True   
Stroke_numeric                       10000        0       2    int64     True   
Marital_Married                      10000        0       2    uint8     True   
Marital_Widowed                      10000        0       2    uint8     True   
Children                             10000        0      11    int64     True   
Income                               10000        0    9993  float64     True   
Complication_risk_Low                10000        0       2    uint8     True   
Initial_admin_Observation_Admission  10000        0       2    uint8     True   
Initial_admin_Elective_Admission     10000        0       2    uint8     True   
Soft_drink_numeric                   10000        0       2    int64     True   
Diabetes_numeric                     10000        0       2    int64     True   
Full_meals_eaten                     10000        0       8    int64     True   
Asthma_numeric                       10000        0       2    int64     True   
Additional_charges                   10000        0    9418  float64     True   
Services_Intravenous                 10000        0       2    uint8     True   
Anxiety_numeric                      10000        0       2    int64     True   
Complication_risk_High               10000        0       2    uint8     True   
Hyperlipidemia_numeric               10000        0       2    int64     True   
Arthritis_numeric                    10000        0       2    int64     True   
ReAdmis_numeric                      10000        0       2    int64     True   
Allergic_rhinitis_numeric            10000        0       2    int64     True   
HighBlood_numeric                    10000        0       2    int64     True   
BackPain_numeric                     10000        0       2    int64     True   
Reflux_esophagitis_numeric           10000        0       2    int64     True   
Complication_risk_Medium             10000        0       2    uint8     True   
Gender_Male                          10000        0       2    uint8     True   
Initial_days                         10000        0    9997  float64     True   
TotalCharge                          10000        0    9997  float64     True   
VitD_levels                          10000        0    9976  float64     True   
Age                                  10000        0      72    int64     True   
Doc_visits                           10000        0       9    int64     True   
Initial_admin_Emergency_Admission    10000        0       2    uint8     True   
Services_Blood_Work                  10000        0       2    uint8     True   
Overweight_numeric                   10000        0       2    int64     True   
```

	Mean	Mode	Min \
Gender_Nonbinary	0.021400	0.00000	0.000000
Services_MRI	0.038000	0.00000	0.000000
Services_CT_Scan	0.122500	0.00000	0.000000
Population	9965.253800	0.00000	0.000000
vitD_supp	0.398900	0.00000	0.000000
Marital_Never_Married	0.198400	0.00000	0.000000
Marital_Separated	0.198700	0.00000	0.000000
Stroke_numeric	0.199300	0.00000	0.000000
Marital_Married	0.202300	0.00000	0.000000
Marital_Widowed	0.204500	0.00000	0.000000
Children	2.097200	0.00000	0.000000
Income	40490.495160	14572.40000	154.080000
Complication_risk_Low	0.212500	0.00000	0.000000
Initial_admin_Observation_Admission	0.243600	0.00000	0.000000
Initial_admin_Elective_Admission	0.250400	0.00000	0.000000
Soft_drink_numeric	0.257500	0.00000	0.000000
Diabetes_numeric	0.273800	0.00000	0.000000
Full_meals_eaten	1.001400	0.00000	0.000000
Asthma_numeric	0.289300	0.00000	0.000000
Additional_charges	12934.528587	3883.66416	3125.703000
Services_Intravenous	0.313000	0.00000	0.000000
Anxiety_numeric	0.321500	0.00000	0.000000
Complication_risk_High	0.335800	0.00000	0.000000
Hyperlipidemia_numeric	0.337200	0.00000	0.000000
Arthritis_numeric	0.357400	0.00000	0.000000
ReAdmis_numeric	0.366900	0.00000	0.000000
Allergic_rhinitis_numeric	0.394100	0.00000	0.000000
HighBlood_numeric	0.409000	0.00000	0.000000
BackPain_numeric	0.411400	0.00000	0.000000
Reflux_esophagitis_numeric	0.413500	0.00000	0.000000
Complication_risk_Medium	0.451700	0.00000	0.000000
Gender_Male	0.476800	0.00000	0.000000
Initial_days	34.455299	63.54432	1.001981
TotalCharge	5312.172769	7555.45200	1938.312067
VitD_levels	17.964262	15.26009	9.806483
Age	53.511700	47.00000	18.000000
Doc_visits	5.012200	5.00000	1.000000
Initial_admin_Emergency_Admission	0.506000	1.00000	0.000000
Services_Blood_Work	0.526500	1.00000	0.000000
Overweight_numeric	0.709400	1.00000	0.000000

	Median	Max \
Gender_Nonbinary	0.000000	1.000000
Services_MRI	0.000000	1.000000
Services_CT_Scan	0.000000	1.000000
Population	2769.000000	122814.000000
vitD_supp	0.000000	5.000000
Marital_Never_Married	0.000000	1.000000
Marital_Separated	0.000000	1.000000
Stroke_numeric	0.000000	1.000000
Marital_Married	0.000000	1.000000
Marital_Widowed	0.000000	1.000000
Children	1.000000	10.000000
Income	33768.420000	207249.100000
Complication_risk_Low	0.000000	1.000000
Initial_admin_Observation_Admission	0.000000	1.000000
Initial_admin_Elective_Admission	0.000000	1.000000
Soft_drink_numeric	0.000000	1.000000
Diabetes_numeric	0.000000	1.000000
Full_meals_eaten	1.000000	7.000000
Asthma_numeric	0.000000	1.000000
Additional_charges	11573.977735	30566.070000
Services_Intravenous	0.000000	1.000000
Anxiety_numeric	0.000000	1.000000
Complication_risk_High	0.000000	1.000000
Hyperlipidemia_numeric	0.000000	1.000000
Arthritis_numeric	0.000000	1.000000
ReAdmis_numeric	0.000000	1.000000
Allergic_rhinitis_numeric	0.000000	1.000000
HighBlood_numeric	0.000000	1.000000
BackPain_numeric	0.000000	1.000000
Reflux_esophagitis_numeric	0.000000	1.000000
Complication_risk_Medium	0.000000	1.000000
Gender_Male	0.000000	1.000000
Initial_days	35.836244	71.981490
TotalCharge	5213.952000	9180.728000
VitD_levels	17.951122	26.394449
Age	53.000000	89.000000
Doc_visits	5.000000	9.000000
Initial_admin_Emergency_Admission	1.000000	1.000000
Services_Blood_Work	1.000000	1.000000
Overweight_numeric	1.000000	1.000000

	Std	Skew	Kurt
Gender_Nonbinary	0.144721	6.615434	41.772323
Services_MRI	0.191206	4.833456	21.366572
Services_CT_Scan	0.327879	2.303141	3.305119
Population	14824.758614	2.229959	5.880913
vitD_supp	0.628505	1.550205	2.330763
Marital_Never_Married	0.398815	1.512784	0.288572
Marital_Separated	0.399042	1.510420	0.281425
Stroke_numeric	0.399494	1.505705	0.267202
Marital_Married	0.401735	1.482369	0.197456
Marital_Widowed	0.403356	1.465500	0.147720
Children	2.163659	1.448013	2.076321
Income	28521.153293	1.405899	2.745690
Complication_risk_Low	0.409097	1.405815	-0.023688
Initial_admin_Observation_Admission	0.429276	1.194810	-0.572544
Initial_admin_Elective_Admission	0.433265	1.152412	-0.672081
Soft_drink_numeric	0.437279	1.109354	-0.769488
Diabetes_numeric	0.445930	1.014712	-0.970553
Full_meals_eaten	1.008117	1.009461	1.042727
Asthma_numeric	0.453460	0.929485	-1.136285
Additional_charges	6542.601544	0.831842	-0.142684
Services_Intravenous	0.463738	0.806652	-1.349583
Anxiety_numeric	0.467076	0.764483	-1.415849
Complication_risk_High	0.472293	0.695470	-1.516625
Hyperlipidemia_numeric	0.472777	0.688834	-1.525813
Arthritis_numeric	0.479258	0.595206	-1.646059
ReAdmis_numeric	0.481983	0.552412	-1.695180
Allergic_rhinitis_numeric	0.488681	0.433498	-1.812442
HighBlood_numeric	0.491674	0.370238	-1.863296
BackPain_numeric	0.492112	0.360153	-1.870664
Reflux_esophagitis_numeric	0.492486	0.351350	-1.876929
Complication_risk_Medium	0.497687	0.194137	-1.962703
Gender_Male	0.499486	0.092914	-1.991765
Initial_days	26.309341	0.070286	-1.754525
TotalCharge	2180.393838	0.069661	-1.668267
VitD_levels	2.017231	0.032435	-0.022112
Age	20.638538	0.005117	-1.189527
Doc_visits	1.045734	-0.018563	0.025999
Initial_admin_Emergency_Admission	0.499989	-0.024005	-1.999824
Services_Blood_Work	0.499322	-0.106165	-1.989127
Overweight_numeric	0.454062	-0.922526	-1.149176

The summary stats show us that the dataset has many continuous variables, due to the mean/max/mode being outside of 0 and 1. Which allowed me to go back and change the Yes/No columns into 0s and 1s, and also use pd.get_dummies to one-hot encode other categorical columns.

The summary statistics overall show us that our average patient has 2 children (with a standard deviation of 2.16), has an income of $40k/yr (with a standard deviation of $28,521), eats 1 full meal a day while in the hospital (with a standard deviation of 1), receives $12,934 in additional charges (with a standard deviation of $6,542), spends 34 days on their initial stay in the hospital (with a standard deviation of 26 days), receives $5,312 in total charges (with a standard deviation of $2180), has Vitamin D levels of 17.96 ng/mL upon admission (with a standard deviation of 2), is 53 years old (with a standard deviation of 21) , and is visited by their doctor 5 times during their stay (with a standard deviation of 1).

C3. To prepare the data for analysis I re-expressed some categorical variables on my own, and used pd.get_dummies to automate one-hot encoding of others. Code snippets are below.

Re-expression of categorical variables
#Data Wrangling; turn categorical values into quantitative data df['ReAdmis_numeric'] = df['ReAdmis']

```
dict_ReAdmis = {"ReAdmis_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_ReAdmis, inplace=True)


df['Soft_drink_numeric'] = df['Soft_drink']

dict_Soft_drink = {"Soft_drink_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_Soft_drink, inplace=True)


df['HighBlood_numeric'] = df['HighBlood']

dict_HighBlood = {"HighBlood_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_HighBlood, inplace=True)


df['Stroke_numeric'] = df['Stroke']

dict_stroke = {"Stroke_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_stroke, inplace=True)


df['Arthritis_numeric'] = df['Arthritis']

dict_arthritis = {"Arthritis_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_arthritis, inplace=True)


df['Diabetes_numeric'] = df['Diabetes']

dict_diabetes = {"Diabetes_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_diabetes, inplace=True)


df['Hyperlipidemia_numeric'] = df['Hyperlipidemia']
```

```
dict_hyperlipidemia = {"Hyperlipidemia_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_hyperlipidemia, inplace=True)


df['BackPain_numeric'] = df['BackPain']

dict_backpain = {"BackPain_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_backpain, inplace=True)


df['Allergic_rhinitis_numeric'] = df['Allergic_rhinitis']

dict_allergies = {"Allergic_rhinitis_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_allergies, inplace=True)


df['Reflux_esophagitis_numeric'] = df['Reflux_esophagitis']

dict_reflux = {"Reflux_esophagitis_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_reflux, inplace=True)


df['Asthma_numeric'] = df['Asthma']

dict_asthma = {"Asthma_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_asthma, inplace=True)


df['Overweight_numeric'] = df['Overweight']

dict_Overweight = {"Overweight_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_Overweight, inplace=True)


df['Anxiety_numeric'] = df['Anxiety']
```

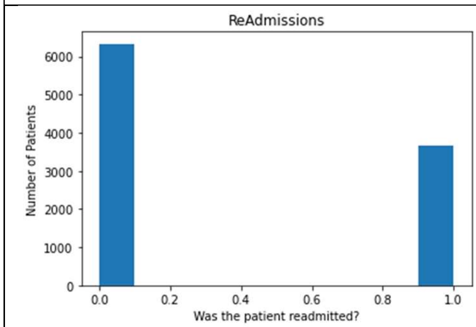
```
dict_Anxiety = {"Anxiety_numeric": {"No": 0, "Yes": 1}}

df.replace(dict_Anxiety, inplace=True)
```

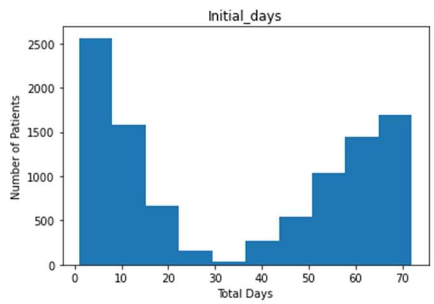
Pd.get_dummies one-hot encoding
df = pd.get_dummies(df, columns=["Marital", "Services", "Gender", "Initial_admin", "Complication_risk"])

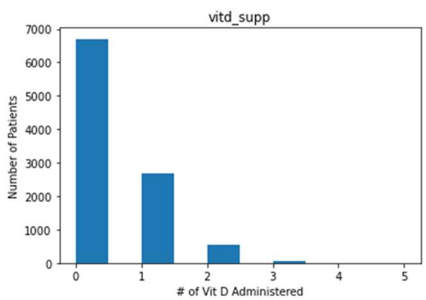
I used both methods for the practice and to get experience with both.

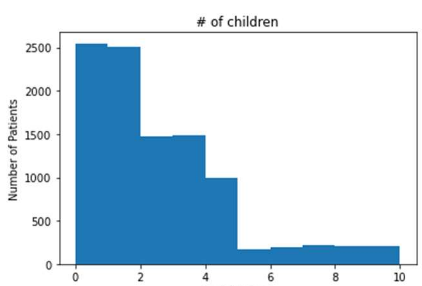
C4. Univariate and Bivariate visualizations are below:

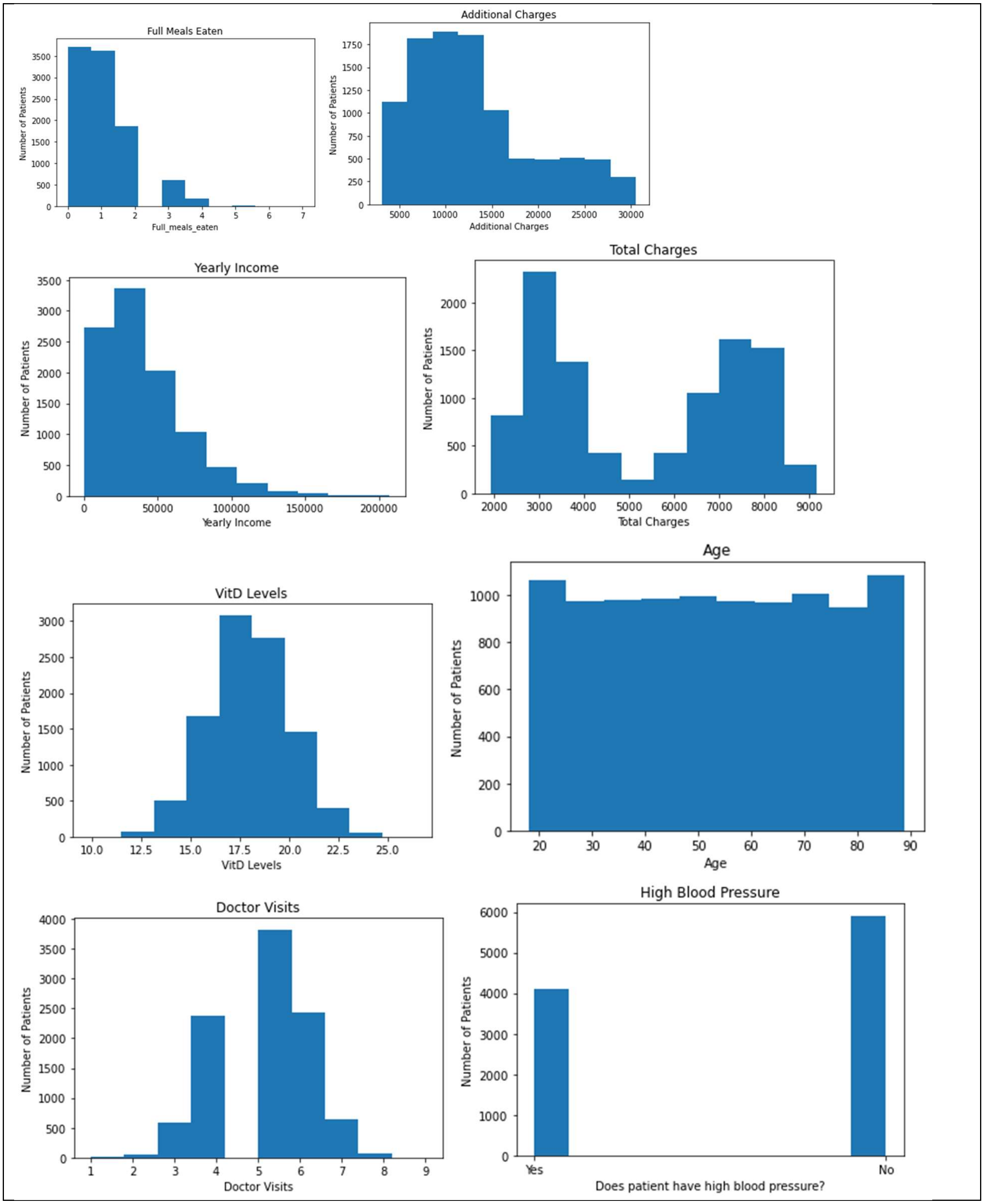
Univariate Visualizations

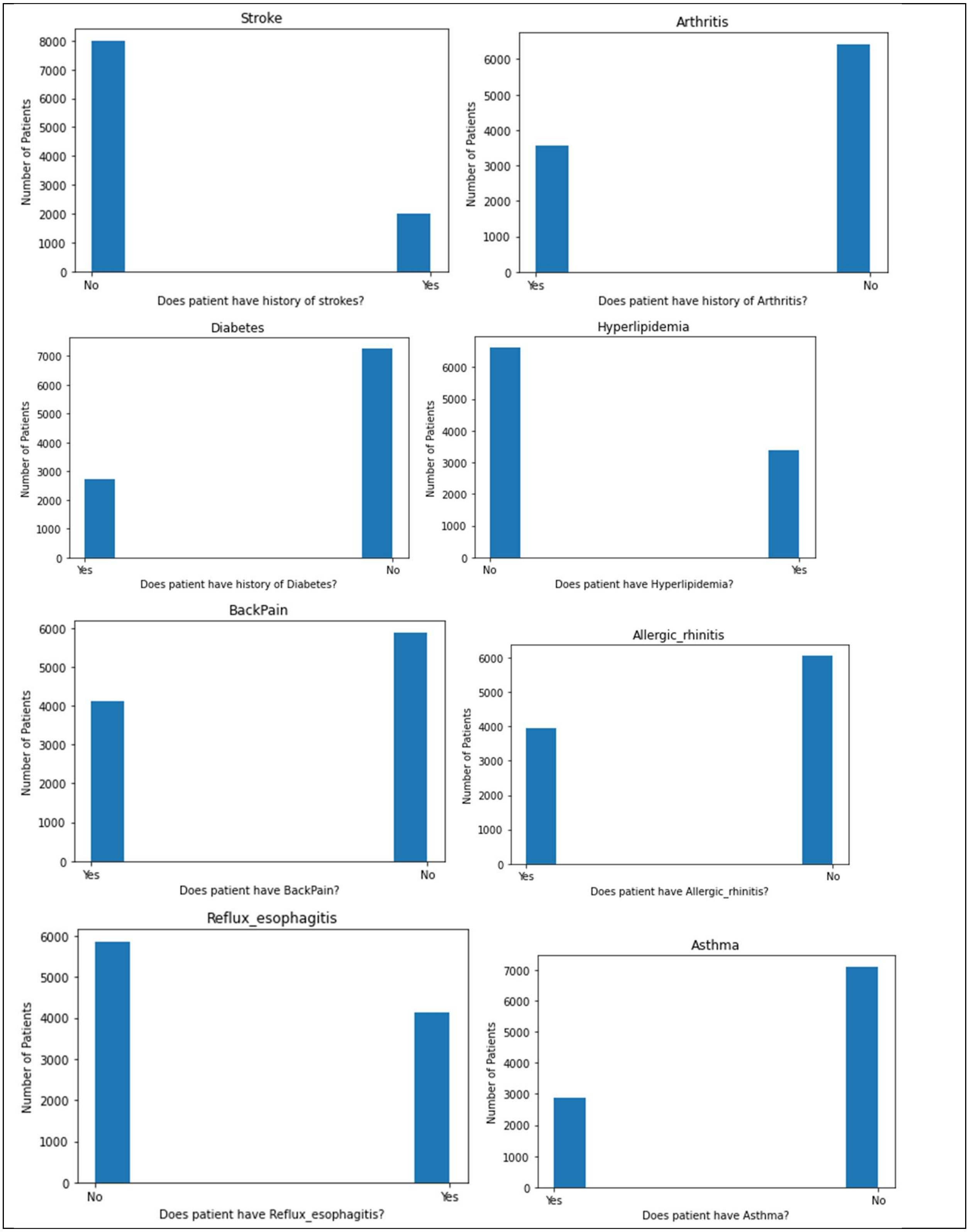


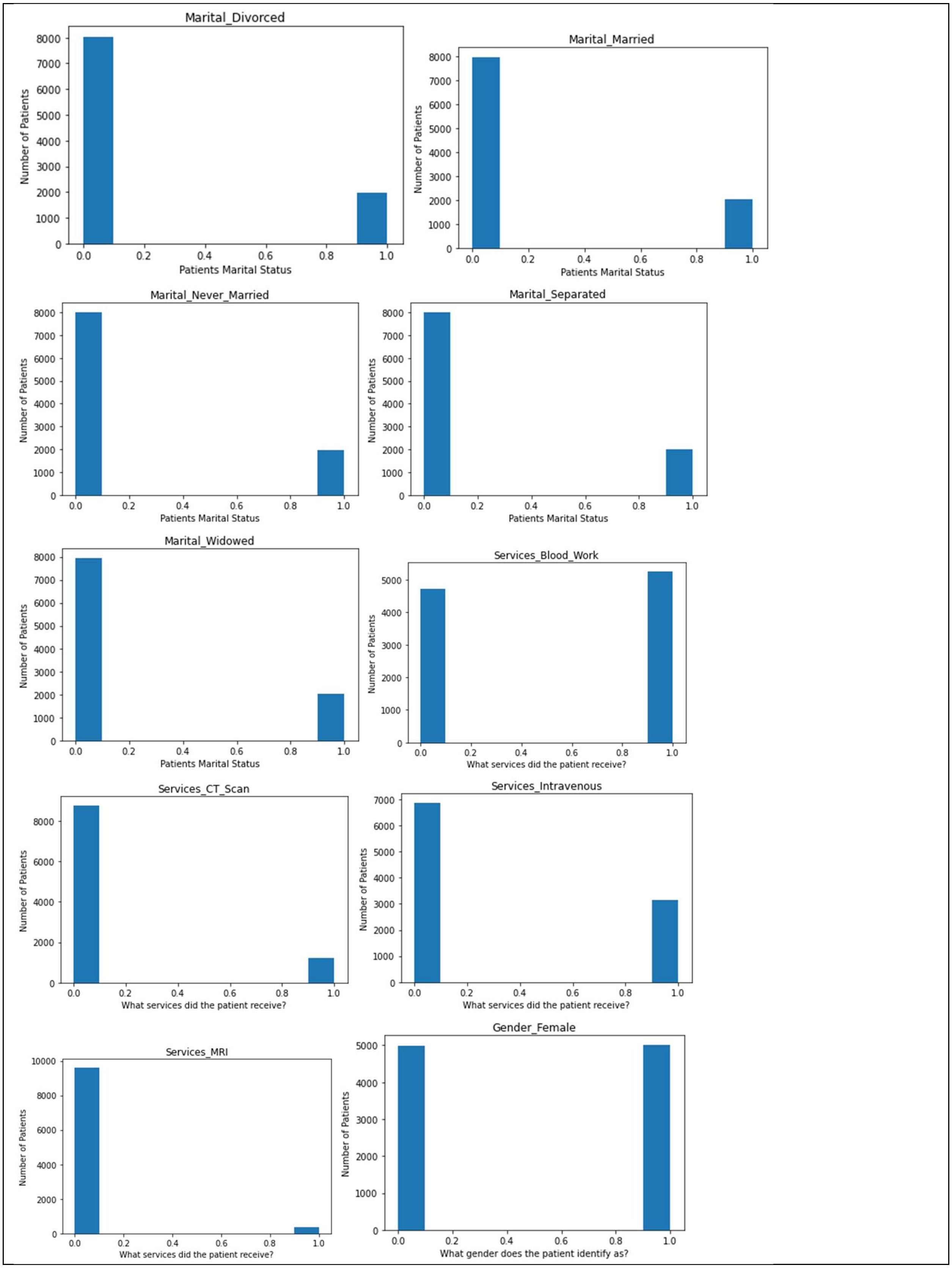


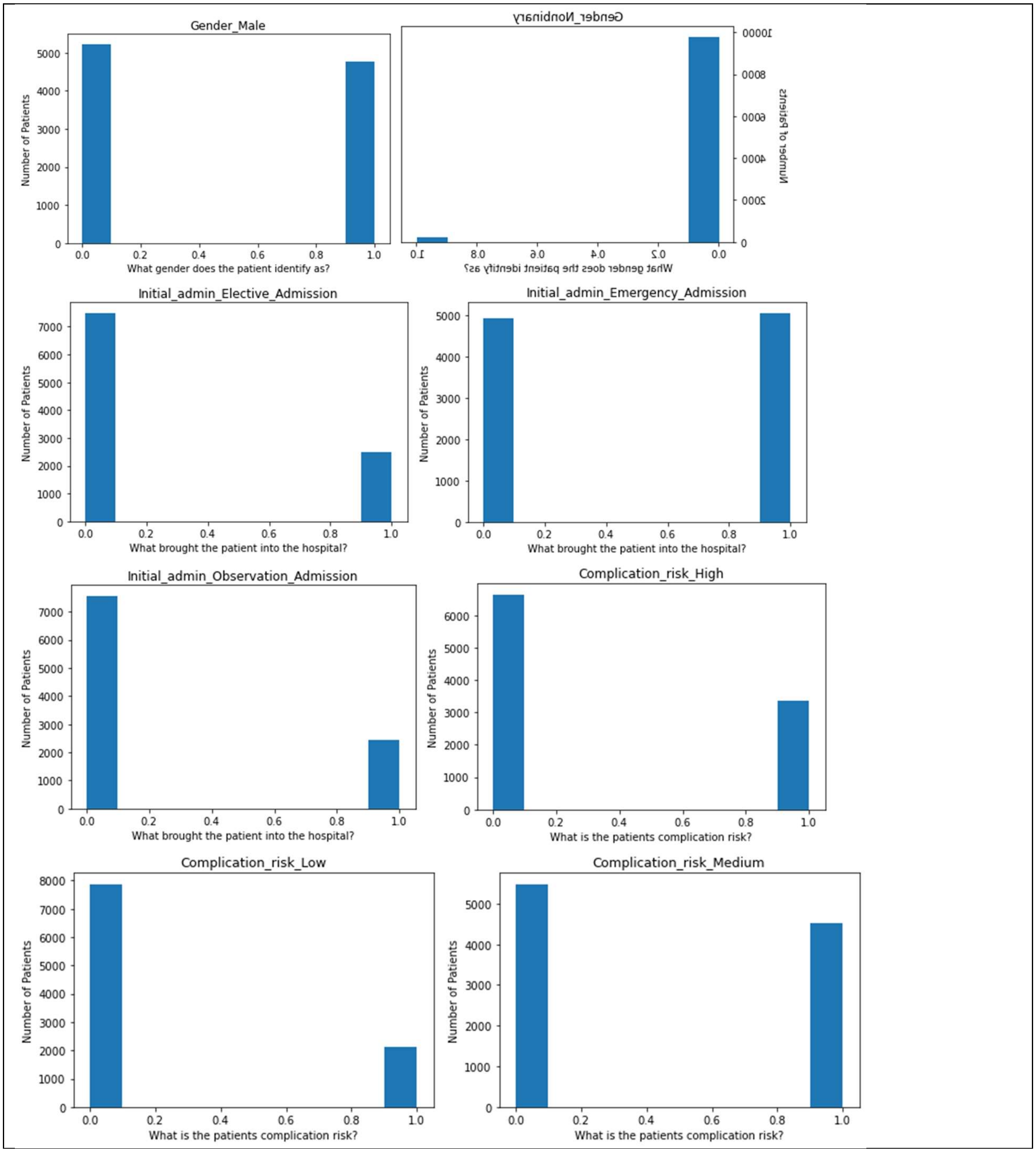




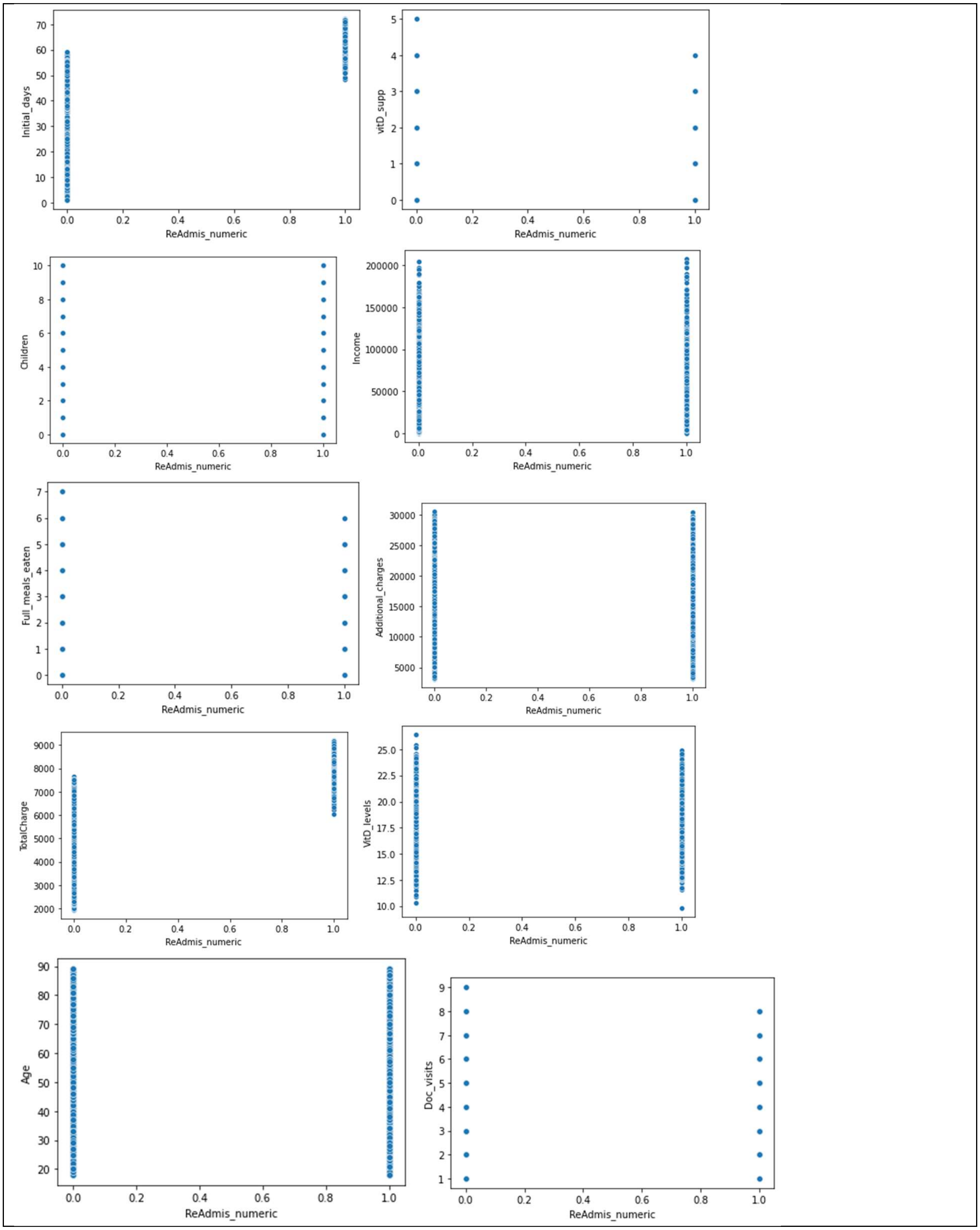

Full Meals Eaten
Number of Patients
0 500 1000 1500 2000 2500 3000 3500
0 1 2 3 4 5 6 7
Full_meals_eaten
Additional Charges
Number of Patients
0 250 500 750 1000 1250 1500 1750
5000 10000 15000 20000 25000 30000
Additional Charges
Yearly Income
Number of Patients
0 500 1000 1500 2000 2500 3000 3500
0 50000 100000 150000 200000
Yearly Income
Total Charges
Number of Patients
0 500 1000 1500 2000
2000 3000 4000 5000 6000 7000 8000 9000
Total Charges
VitD Levels
Number of Patients
0 500 1000 1500 2000 2500 3000
10.0 12.5 15.0 17.5 20.0 22.5 25.0
VitD Levels
Age
Number of Patients
0 200 400 600 800 1000
20 30 40 50 60 70 80 90
Age
Doctor Visits
Number of Patients
0 500 1000 1500 2000 2500 3000 3500 4000
1 2 3 4 5 6 7 8 9
Doctor Visits
High Blood Pressure
Number of Patients
0 1000 2000 3000 4000 5000 6000
Yes
No
Does patient have high blood pressure?

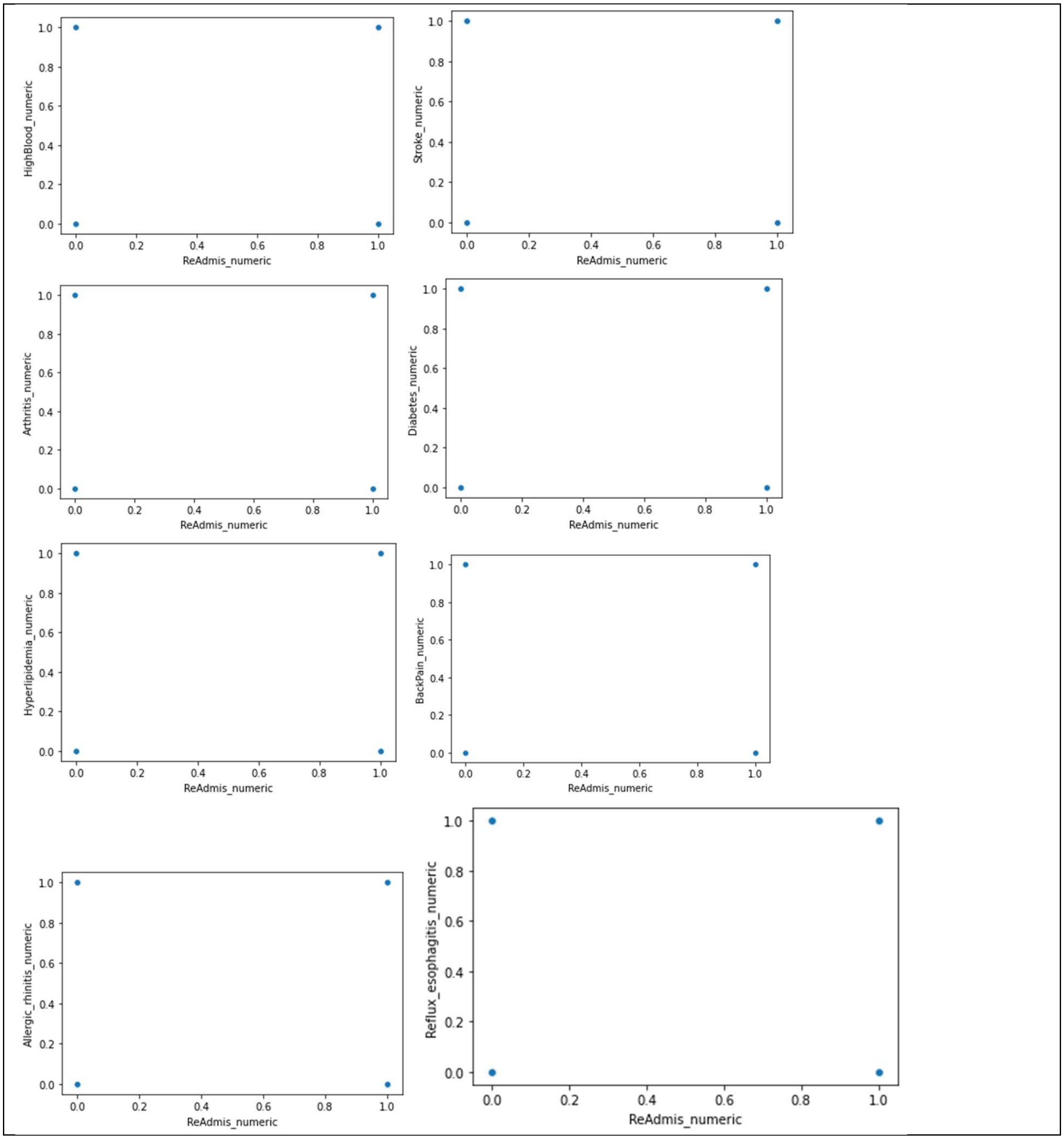

Stroke
Number of Patients
0
1000
2000
3000
4000
5000
6000
7000
8000
No
Yes
Does patient have history of strokes?
Arthritis
Number of Patients
0
1000
2000
3000
4000
5000
6000
Yes
No
Does patient have history of Arthritis?
Diabetes
Number of Patients
0
1000
2000
3000
4000
5000
6000
7000
Yes
No
Does patient have history of Diabetes?
Hyperlipidemia
Number of Patients
0
1000
2000
3000
4000
5000
6000
No
Yes
Does patient have Hyperlipidemia?
BackPain
Number of Patients
0
1000
2000
3000
4000
5000
6000
Yes
No
Does patient have BackPain?
Allergic_rhinitis
Number of Patients
0
1000
2000
3000
4000
5000
6000
Yes
No
Does patient have Allergic_rhinitis?
Reflux_esophagitis
Number of Patients
0
1000
2000
3000
4000
5000
6000
No
Yes
Does patient have Reflux_esophagitis?
Asthma
Number of Patients
0
1000
2000
3000
4000
5000
6000
7000
Yes
No
Does patient have Asthma?

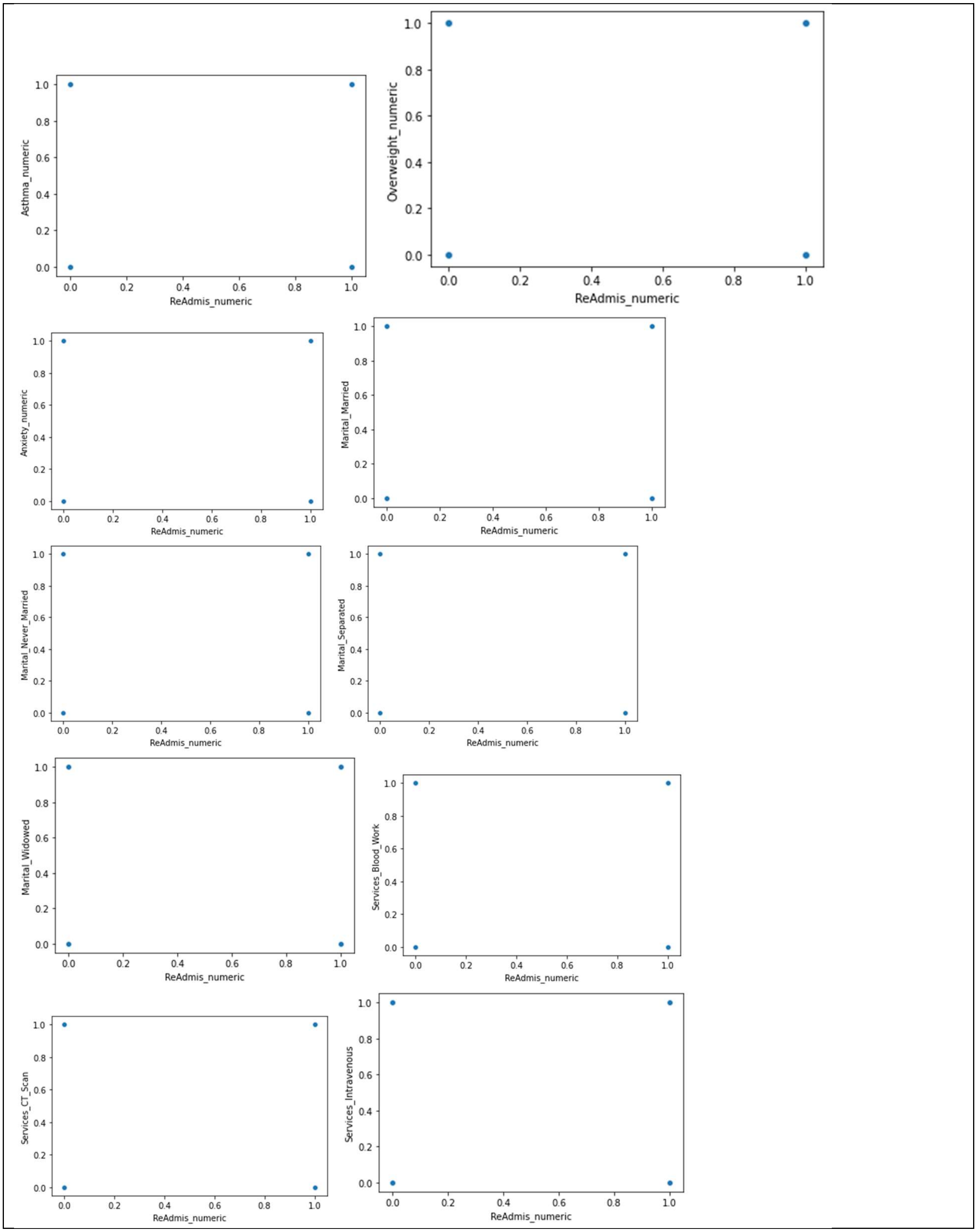

Marital_Divorced
Number of Patients
Patients Marital Status
Marital_Married
Number of Patients
Patients Marital Status
Marital_Never_Married
Number of Patients
Patients Marital Status
Marital_Separated
Number of Patients
Patients Marital Status
Marital_Widowed
Number of Patients
Patients Marital Status
Services_Blood_Work
Number of Patients
What services did the patient receive?
Services_CT_Scan
Number of Patients
What services did the patient receive?
Services_Intravenous
Number of Patients
What services did the patient receive?
Services_MRI
Number of Patients
What services did the patient receive?
Gender_Female
Number of Patients
What gender does the patient identify as?



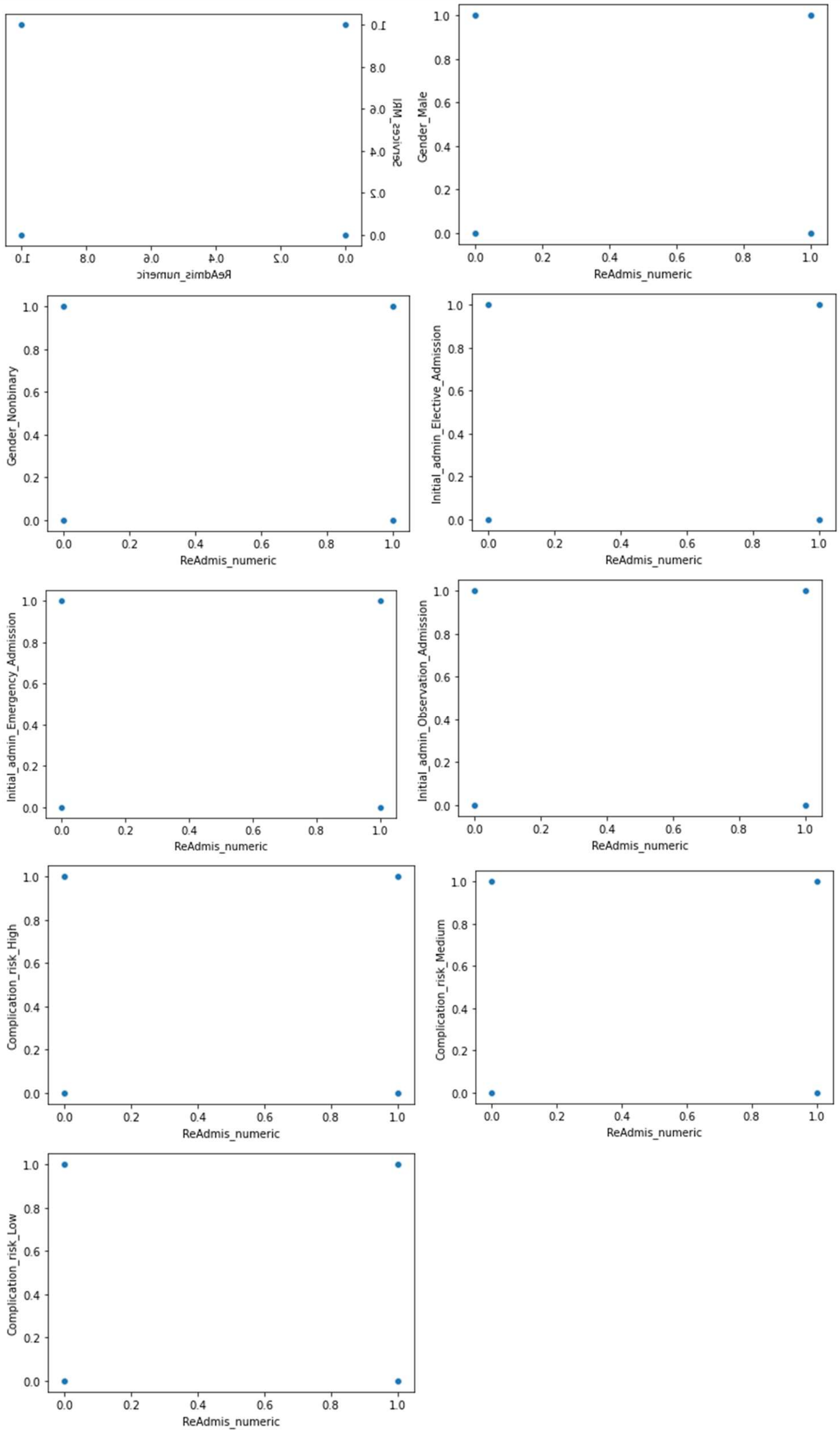
Bivariate Visualizations


Initial_days
ReAdmis_numeric
vitD_supp
ReAdmis_numeric
Children
ReAdmis_numeric
Income
ReAdmis_numeric
Full_meals_eaten
ReAdmis_numeric
Additional_charges
ReAdmis_numeric
TotalCharge
ReAdmis_numeric
VitD_levels
ReAdmis_numeric
Age
ReAdmis_numeric
Doc_visits
ReAdmis_numeric


HighBlood_numeric
ReAdmis_numeric
Stroke_numeric
ReAdmis_numeric
Arthritis_numeric
ReAdmis_numeric
Diabetes_numeric
ReAdmis_numeric
Hyperlipidemia_numeric
ReAdmis_numeric
BackPain_numeric
ReAdmis_numeric
Allergic_rhinitis_numeric
ReAdmis_numeric
Reflux_esophagitis_numeric
ReAdmis_numeric



Asthma_numeric
ReAdmis_numeric
Overweight_numeric
ReAdmis_numeric
Anxiety_numeric
ReAdmis_numeric
Marital_Married
ReAdmis_numeric
Marital_Never_Married
ReAdmis_numeric
Marital_Separated
ReAdmis_numeric
Marital_Widowed
ReAdmis_numeric
Services_Blood_Work
ReAdmis_numeric
Services_CT_Scan
ReAdmis_numeric
Services_Intravenous
ReAdmis_numeric



ReAdmis_numeric
IRM_Services
Gender_Male
Gender_Nonbinary
Initial_admin_Elective_Admission
Initial_admin_Emergency_Admission
Initial_admin_Observation_Admission
Complication_risk_High
Complication_risk_Medium
Complication_risk_Low

C5. Prepared data set uploaded as part of submission.

```
df.to_csv(r'C:\Users\mmorg\Desktop\D208 Assessment Files\Cleaned208data.csv')
```

Part IV: Model Comparison and Analysis

D1. Initial regression model with variables identified in C2

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.032914
         Iterations: 35
Intercept                              -78.391168
Initial_days                            -1.143612
vitD_supp                               -0.105465
Children                                 0.088933
Income                                   0.000002
Full_meals_eaten                         0.048269
Additional_charges                       0.000047
TotalCharge                              0.032097
VitD_levels                              0.029913
Age                                     -0.008115
Doc_visits                               0.006554
HighBlood_numeric                       -3.125420
Stroke_numeric                           1.651497
Arthritis_numeric                       -3.669106
Diabetes_numeric                        -1.908568
Hyperlipidemia_numeric                  -2.757763
BackPain_numeric                        -2.508272
Allergic_rhinitis_numeric               -2.268662
Reflux_esophagitis_numeric              -2.342286
Asthma_numeric                          -1.389130
Overweight_numeric                      -0.286489
Anxiety_numeric                         -3.817203
Marital_Married                          0.268609
Marital_Never_Married                    0.356522
Marital_Separated                       -0.127193
Marital_Widowed                          0.136171
Services_Blood_Work                    -20.676842
Services_CT_Scan                       -19.061885
Services_Intravenous                   -20.683935
Services_MRI                           -17.968342
Gender_Male                              0.169142
Gender_Nonbinary                         0.365800
Initial_admin_Elective_Admission       -21.750667
Initial_admin_Emergency_Admission      -35.680140
Initial_admin_Observation_Admission    -20.960246
Complication_risk_High                 -34.251935
Complication_risk_Low                  -22.802367
Complication_risk_Medium               -21.336813
dtype: float64
```

Logit Regression Results

Dep. Variable:	ReAdmis_numeric	**No. Observations:**	10000
Model:	Logit	**Df Residuals:**	9965
Method:	MLE	**Df Model:**	34
Date:	Thu, 24 Nov 2022	**Pseudo R-squ.:**	0.9499
Time:	21:07:42	**Log-Likelihood:**	-329.14
converged:	False	**LL-Null:**	-6572.9
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-78.3912	nan	nan	nan	nan	nan
Initial_days	-1.1436	1.5e+04	-7.62e-05	1.000	-2.94e+04	2.94e+04
vitD_supp	-0.1055	0.167	-0.632	0.528	-0.433	0.222
Children	0.0889	0.046	1.913	0.056	-0.002	0.180
Income	1.572e-06	3.62e-06	0.434	0.664	-5.53e-06	8.67e-06
Full_meals_eaten	0.0483	0.103	0.469	0.639	-0.153	0.250
Additional_charges	4.664e-05	6.25e-05	0.746	0.455	-7.58e-05	0.000
TotalCharge	0.0321	183.195	0.000	1.000	-359.024	359.088
VitD_levels	0.0299	0.049	0.613	0.540	-0.066	0.126
Age	-0.0081	0.015	-0.549	0.583	-0.037	0.021
Doc_visits	0.0066	0.098	0.067	0.947	-0.186	0.200
HighBlood_numeric	-3.1254	2.06e+04	-0.000	1.000	-4.03e+04	4.03e+04
Stroke_numeric	1.6515	0.272	6.069	0.000	1.118	2.185
Arthritis_numeric	-3.6691	1.32e+04	-0.000	1.000	-2.58e+04	2.58e+04
Diabetes_numeric	-1.9086	1.38e+04	-0.000	1.000	-2.7e+04	2.7e+04
Hyperlipidemia_numeric	-2.7578	1.72e+04	-0.000	1.000	-3.38e+04	3.37e+04
BackPain_numeric	-2.5083	1.56e+04	-0.000	1.000	-3.06e+04	3.06e+04
Allergic_rhinitis_numeric	-2.2687	1.11e+04	-0.000	1.000	-2.18e+04	2.17e+04
Reflux_esophagitis_numeric	-2.3423	1.09e+04	-0.000	1.000	-2.14e+04	2.14e+04
Asthma_numeric	-1.3891	0.237	-5.852	0.000	-1.854	-0.924
Overweight_numeric	-0.2865	0.229	-1.250	0.211	-0.736	0.163
Anxiety_numeric	-3.8172	1.58e+04	-0.000	1.000	-3.09e+04	3.09e+04
Marital_Married	0.2686	0.331	0.811	0.418	-0.381	0.918
Marital_Never_Married	0.3565	0.338	1.055	0.292	-0.306	1.019
Marital_Separated	-0.1272	0.344	-0.369	0.712	-0.802	0.548
Marital_Widowed	0.1362	0.333	0.409	0.682	-0.516	0.788
Services_Blood_Work	-20.6768	4.07e+06	-5.08e-06	1.000	-7.98e+06	7.98e+06

Services_CT_Scan	-19.0619	4.07e+06	-4.68e-06	1.000	-7.98e+06	7.98e+06
Services_Intravenous	-20.6839	4.07e+06	-5.08e-06	1.000	-7.98e+06	7.98e+06
Services_MRI	-17.9683	4.07e+06	-4.41e-06	1.000	-7.99e+06	7.99e+06
Gender_Male	0.1691	0.210	0.807	0.420	-0.242	0.580
Gender_Nonbinary	0.3658	0.714	0.512	0.608	-1.033	1.765
Initial_admin_Elective_Admission	-21.7507	nan	nan	nan	nan	nan
Initial_admin_Emergency_Admission	-35.6801	nan	nan	nan	nan	nan
Initial_admin_Observation_Admission	-20.9602	nan	nan	nan	nan	nan
Complication_risk_High	-34.2519	2.15e+06	-1.59e-05	1.000	-4.22e+06	4.22e+06
Complication_risk_Low	-22.8024	2.34e+06	-9.75e-06	1.000	-4.59e+06	4.59e+06
Complication_risk_Medium	-21.3368	2.34e+06	-9.12e-06	1.000	-4.59e+06	4.59e+06

Possibly complete quasi-separation: A fraction 0.82 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. The log-likelihood value is -329.14 and the Pseudo R-Squared is 0.9499. These numbers will be used later to compare them to the reduced model. The LLR p-value is 0.000 which tells us that the model is "useful" in predicting the values of the response variable.

To reduce the model I am going to run a VIF and create a new model with a reduced amount of variables. The VIF will help me determine the multicollinearity of the variables used in the initial model. I can also calculate the AIC of the models and that can help me to determine what variables I want to continue with in building a model to predict patient readmissions.

Results of VIF from initial model below:

	feature	VIF
0	Initial_days	2880.163153
1	vitD_supp	1.003676
2	Children	1.003506
3	Income	1.002683
4	Full_meals_eaten	1.004107
5	Additional_charges	16.303881
6	TotalCharge	2944.078834
7	VitD_levels	1.003914
8	Age	9.273563
9	Doc_visits	1.003377
10	HighBlood_numeric	9.711378
11	Stroke_numeric	1.010014
12	Arthritis_numeric	1.760819
13	Diabetes_numeric	1.696512
14	Hyperlipidemia_numeric	2.198017
15	BackPain_numeric	2.112077
16	Allergic_rhinitis_numeric	1.551215
17	Reflux_esophagitis_numeric	1.527188
18	Asthma_numeric	1.003104
19	Marital_Married	1.627325
20	Marital_Never_Married	1.618488
21	Marital_Separated	1.617166
22	Marital_Widowed	1.630512
23	Services_Blood_Work	inf
24	Services_CT_Scan	inf
25	Services_Intravenous	inf
26	Services_MRI	inf
27	Gender_Male	1.026146
28	Gender_Nonbinary	1.023726
29	Initial_admin_Elective_Admission	inf
30	Initial_admin_Emergency_Admission	inf
31	Initial_admin_Observation_Admission	inf
32	Complication_risk_High	inf
33	Complication_risk_Low	inf
34	Complication_risk_Medium	inf

D3. Results of reduced logistic regression model

```
Optimization terminated successfully.
         Current function value: 0.045687
         Iterations 13
Intercept             -57.974877
Initial_days            1.066472
Children                0.069580
Stroke_numeric          1.274963
Asthma_numeric         -0.943442
Overweight_numeric     -0.109219
dtype: float64
```

Logit Regression Results

Dep. Variable:	ReAdmis_numeric	**No. Observations:**	10000
Model:	Logit	**Df Residuals:**	9994
Method:	MLE	**Df Model:**	5
Date:	Thu, 24 Nov 2022	**Pseudo R-squ.:**	0.9305
Time:	21:28:45	**Log-Likelihood:**	-456.87
converged:	True	**LL-Null:**	-6572.9
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-57.9749	2.752	-21.064	0.000	-63.369	-52.580
Initial_days	1.0665	0.051	21.092	0.000	0.967	1.166
Children	0.0696	0.038	1.833	0.067	-0.005	0.144
Stroke_numeric	1.2750	0.223	5.713	0.000	0.838	1.712
Asthma_numeric	-0.9434	0.190	-4.972	0.000	-1.315	-0.572
Overweight_numeric	-0.1092	0.186	-0.588	0.557	-0.473	0.255

Possibly complete quasi-separation: A fraction 0.75 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

E. Analyze the data set.

E1. For the initial model I chose to use all the variables that related to patient demographic and health condition. In the reduced model I removed variables that had high p-values and were redundant based on have multicollinearity with other variables whether identified through running the VIF or looking at our bivariate visualizations. This biggest example of redundant data that could be removed just from looking at the visualizations is TotalCharges because that data is essentially based on Initial_days already.

When comparing our initial model and reduced model we want to use a few key stats from our logit summary.

The initial model gives us a log-likelihood value of -329.14 and Pseudo R-Squared value of 0.9499.

The reduced model gives us a log-likelihood value of -457.04 and Pseudo R-Squared value of 0.9305.

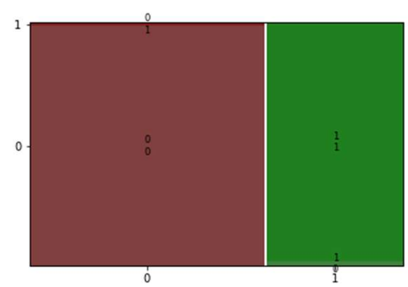
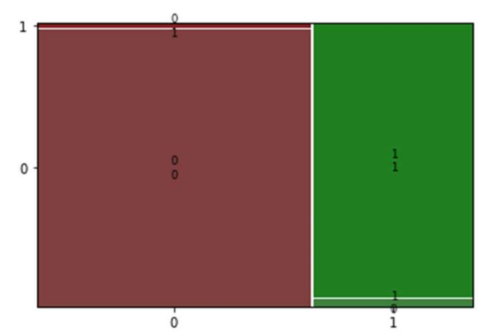
Both models have a LLR p-value of 0.000.

Based on comparing the log-likelihood and Pseudo R-Squared values, our initial model is a better fit for making predictions. This is also backed up by the AIC scores and confusion matrix from each model which are included below. The initial model scores as a "better" model according to both AIC and the accuracy calculation from the confusion matrix.

E2.

Initial Model	Reduced Model
AIC Score and Code: #Calculating AIC of Initial Model from sklearn.linear_model import LinearRegression import statsmodels.api as sm #define response variable y = df['ReAdmis_numeric'] #define predictor variables x = df[['Initial_days', 'vitD_supp', 'Children', 'Income', 'Full_meals_eaten', 'Additional_charges', 'TotalCharge', 'VitD_levels', 'Age', 'Doc_visits', 'HighBlood_numeric', 'Stroke_numeric', 'Arthritis_numeric', 'Diabetes_numeric', 'Hyperlipidemia_numeric', 'BackPain_numeric', 'Allergic_rhinitis_numeric', 'Reflux_esophagitis_numeric', 'Asthma_numeric', 'Marital_Married', 'Marital_Never_Married', 'Marital_Separated', 'Marital_Widowed', 'Services_Blood_Work', 'Services_CT_Scan', 'Services_Intravenous', 'Services_MRI',	AIC Score and Code: #Calculating AIC of Reduced Model #1 #define response variable y = df['ReAdmis_numeric'] #define predictor variables x = df[['Initial_days', 'Children', 'Stroke_numeric', 'Asthma_numeric', 'Overweight_numeric']] #add constant to predictor variables x = sm.add_constant(x) #fit regression model model = sm.OLS(y, x).fit()

'Gender_Male', 'Gender_Nonbinary', 'Initial_admin_Elective_Admission', 'Initial_admin_Emergency_Admission', 'Initial_admin_Observation_Admission', 'Complication_risk_High', 'Complication_risk_Low', 'Complication_risk_Medium']] #add constant to predictor variables x = sm.add_constant(x) #fit regression model model = sm.OLS(y, x).fit() #view AIC of model print(model.aic) **Score:** `897.5899676679401` `(Statology, 2021)`	#view AIC of model print(model.aic) **Score:** `918.392404996368` `(Statology, 2021)`
Confusion Matrix and Code: #Confusion Matrix for Initial Model conf_matrix = mdl_readmis_vs_variables.pred_table() print(conf_matrix) from statsmodels.graphics.mosaicplot import mosaic mosaic(conf_matrix) #Calculating accuracy: the proportion of correct predictions	Confusion Matrix and Code: #Confusion Matrix for Reduced Model conf_matrix = mdl_readmis_vs_variables1.pred_table() print(conf_matrix) from statsmodels.graphics.mosaicplot import mosaic mosaic(conf_matrix) #Calculating accuracy: the proportion of correct predictions

TN = conf_matrix[0,0] TP = conf_matrix[1,1] FN = conf_matrix[1,0] FP = conf_matrix[0,1] acc = (TN + TP) / (TN + TP + FN + FP) print('Accuracy:', acc) #Sensitivity: proportion of true positives sens = TP / (FN + TP) print('Sensitivity:', sens) #Specificity: proportion of true negatives spec = TN / (TN + FP) print('Specificity:', spec) [[6261. 70.] [60. 3609.]] Accuracy: 0.987 Sensitivity: 0.9836467702371219 Specificity: 0.9889432948981204 	TN = conf_matrix[0,0] TP = conf_matrix[1,1] FN = conf_matrix[1,0] FP = conf_matrix[0,1] acc = (TN + TP) / (TN + TP + FN + FP) print('Accuracy:', acc) #Sensitivity: proportion of true positives sens = TP / (FN + TP) print('Sensitivity:', sens) #Specificity: proportion of true negatives spec = TN / (TN + FP) print('Specificity:', spec) [[6228. 103.] [103. 3566.]] Accuracy: 0.9794 Sensitivity: 0.9719269555737258 Specificity: 0.9837308482072342 

Reduced Model Residual Standard Error (Techhelpnotes, 2022)

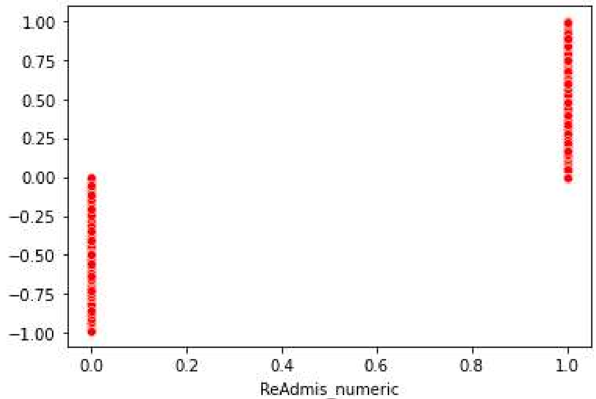
```
#Residual Standard Error for reduced model
np.sqrt(mdl_readmis_vs_variables1.scale)
```

```
1.0
```

Reduced Model Residual Plot

```
#Residual plot for reduced model
df['intercept'] = 1
residuals = df['ReAdmis_numeric'] - mdl_readmis_vs_variables1.predict(df[['Initial_days', 'Children',
sns.scatterplot(x=df['ReAdmis_numeric'], y=residuals, color='red')
```

```
<AxesSubplot:xlabel='ReAdmis_numeric'>
```



E3.

Initial Model
#Initial Logistic Regression Model mdl_readmis_vs_variables = logit("ReAdmis_numeric ~ Initial_days + vitD_supp + Children + Income + Full_meals_eaten + Additional_charges + TotalCharge + VitD_levels + Age + Doc_visits + HighBlood_numeric + Stroke_numeric + Arthritis_numeric + Diabetes_numeric + Hyperlipidemia_numeric + BackPain_numeric + Allergic_rhinitis_numeric + Reflux_esophagitis_numeric + Asthma_numeric + Overweight_numeric + Anxiety_numeric + Marital_Married + Marital_Never_Married + Marital_Separated + Marital_Widowed + Services_Blood_Work + Services_CT_Scan + Services_Intravenous + Services_MRI + Gender_Male + Gender_Nonbinary + Initial_admin_Elective_Admission + Initial_admin_Emergency_Admission + Initial_admin_Observation_Admission + Complication_risk_High + Complication_risk_Low + Complication_risk_Medium", data=df).fit() print(mdl_readmis_vs_variables.params) mdl_readmis_vs_variables.summary()

VIF to reduce model
#Variable Selection # Checking for the VIF values of the variables. from statsmodels.stats.outliers_influence import variance_inflation_factor

```
X = df[['Initial_days', 'vitD_supp', 'Children', 'Income', 'Full_meals_eaten', 'Additional_charges',
'TotalCharge', 'VitD_levels', 'Age', 'Doc_visits', 'HighBlood_numeric', 'Stroke_numeric',
'Arthritis_numeric', 'Diabetes_numeric', 'Hyperlipidemia_numeric', 'BackPain_numeric',
'Allergic_rhinitis_numeric', 'Reflux_esophagitis_numeric', 'Asthma_numeric', 'Marital_Married',
'Marital_Never_Married', 'Marital_Separated', 'Marital_Widowed', 'Services_Blood_Work',
'Services_CT_Scan', 'Services_Intravenous', 'Services_MRI', 'Gender_Male', 'Gender_Nonbinary',
'Initial_admin_Elective_Admission', 'Initial_admin_Emergency_Admission',
'Initial_admin_Observation_Admission', 'Complication_risk_High', 'Complication_risk_Low',
'Complication_risk_Medium']]




# VIF dataframe

vif_data = pd.DataFrame()

vif_data["feature"] = X.columns



# calculating VIF for each feature

vif_data["VIF"] = [variance_inflation_factor(X.values, i)

                          for i in range(len(X.columns))]



print(vif_data)
```

(GeeksforGeeks, 2019)

Reduced Model

```
#Reduced Model removing complication risk, initial admin, services, demographics, and charges
columns due to VIF being high, redundancy, and high p-value

mdl_readmis_vs_variables1 = logit("ReAdmis_numeric ~ Initial_days + Children + Stroke_numeric +
Asthma_numeric", data=df).fit()

print(mdl_readmis_vs_variables1.params)

mdl_readmis_vs_variables1.summary()
```

Part V: Data Summary and Implications
F1. My linear regression equation:

Y = -57.9744 + 1.0651 (Initial_days) + 0.0698 (Children) + 1.2735 (Stroke_numeric) + -0.9440 (Asthma_numeric)

This line means for every 1 unit of:

Initial_days, ReAdmis_numeric will increase 1.0651 units
Children ReAdmis_numeric will increase 0.0698 units
Stroke_numeric ReAdmis_numeric will increase 1.2735 units
Asthma_numeric ReAdmis_numeric will decrease 0.9440 units

As stated previously, when comparing our initial model and reduced model we want to use a few key stats from our logit summary.

The initial model gives us a log-likelihood value of -329.14 and Pseudo R-Squared value of 0.9499.

The reduced model gives us a log-likelihood value of -457.04 and Pseudo R-Squared value of 0.9305.

Both models have a LLR p-value of 0.000.

Based on comparing the log-likelihood and Pseudo R-Squared values, our initial model is a better fit for making predictions. This is also backed up by the AIC scores and confusion matrix from each model which are included below. The initial model scores as a "better" model according to both AIC and the accuracy calculation from the confusion matrix.

Based on these key stats, I would say that these models are not practically significant, though the initial model could be statistically significant.

Neither model should be used to make predictions about what patients will be readmitted. Both models are very limited in what they can do. I believe this mainly has to do with the dataset. More data and different data needs to be captured.

F2. Based on my results, there really isn't a course of action to be recommended. The initial model is too robust and complex, and no predictions can really be made. The reduced model is less accurate than the initial model so we wouldn't want to use that for predictions either. The course of action would be to start back at square 1, reduce our model with a different method, and re-evaluate what data we capture moving forward.

Part VI: Demonstration
G. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=57dfea46-a1e1-499f-beb9-af5800362f65

H.

Techhelpnotes, 2022 https://techhelpnotes.com/residual-standard-error-of-a-regression-in-python/

GeeksforGeeks, 2019 https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

Statology, 2021 https://www.statology.org/aic-in-python/

I. Sewell, William. (2022). *D208 Predictive Modeling Webinar Episode 3* [Slide 17]