Raquel Ocasio

D208 – Predictive Modeling, Task 2

October 11, 2023

Western Governors University

Part A1

Which variables influence the probability of a patient being readmitted within a month of release?

Part A2

The goal of the data analysis is to determine if the probability of the patient being readmitted within a month of release is influenced by other variables in the dataset.

Part B1

Logistic regression relies on four key assumptions: first, a linear relationship between predictor variables and the log-odds of the dependent variable; second, independence of observations, meaning that one observation's outcome does not influence another's; third, minimal or no multicollinearity among independent variables to prevent high correlation issues; and finally, a reasonably large sample size, typically with at least 10-20 observations per predictor variable to ensure the reliability of statistical inference. Adherence to these assumptions is crucial for accurate model estimation and interpretation.

Part B2

Two benefits of using R for logistic regression analysis are its open-source nature and extensive package ecosystem. Being open-source, R is freely available, making it accessible to a broad user base and eliminating licensing costs. Moreover, R has a large collection of packages tailored for statistical modeling, including logistic regression, allowing users to tap into a wealth of specialized functions and tools for data preprocessing, model building, and result interpretation. These advantages enhance the efficiency and versatility of logistic regression analysis in R.

Part B3

The target variable for this analysis is categorical. Logistic regression is appropriate for this analysis because it can help to understand the relationship between a categorical response variable and one or more explanatory variables that are continuous and/or categorical.

Part C1

The goals of the data cleaning process are to detect and treat duplicate values, missing values, and outlier values. The unique values for categorical variables also need to be detected to check for inconsistency in presentation of the data.

Duplicate values are detected using the sum(duplicated()) functions. No duplicate values were detected.

Missing values are detected using the colSums(is.na()) functions. No missing values were detected.

Outliers for quantitative variables are detected using a function that uses the mean(), sd(), and sum() functions to calculate the z-score and count how many z-scores have a value greater than three or less than negative three. Seven variables were found to have outliers.

For categorical variables, unique values are detected using the unique() function. None of the categorical variables had inconsistent presentation of the data.

Part C2

Summary statistics for dependent variable

	Column Labels		
	No	Yes	Grand Total
Count of ReAdmis	6331	3669	10000

Summary statistics for categorical independent variables

	Column Labels			
	Female	Male	Nonbinary	Grand Total
Count of Gender	5018	4768	214	10000

	Column Labels		
	No	Yes	Grand Total
Count of HighBlood	5910	4090	10000

	Column Labels		
	No	Yes	Grand Total
Count of Stroke	8007	1993	10000

	Column Labels		
	No	Yes	Grand Total
Count of Overweight	2906	7094	10000

	Column Labels		
	No	Yes	Grand Total
Count of Arthritis	6426	3574	10000

	Column Labels		
	No	Yes	Grand Total
Count of Diabetes	7262	2738	10000

	Column Labels		
	No	Yes	Grand Total
Count of Hyperlipidemia	6628	3372	10000

	Column Labels		
	No	Yes	Grand Total
Count of BackPain	5886	4114	10000

	Column Labels		
	No	Yes	Grand Total
Count of Anxiety	6785	3215	10000

	Column Labels		
	No	Yes	Grand Total
Count of Allergic_rhinitis	6059	3941	10000

	Column Labels		
	No	Yes	Grand Total
Count of Reflux_esophagitis	5865	4135	10000

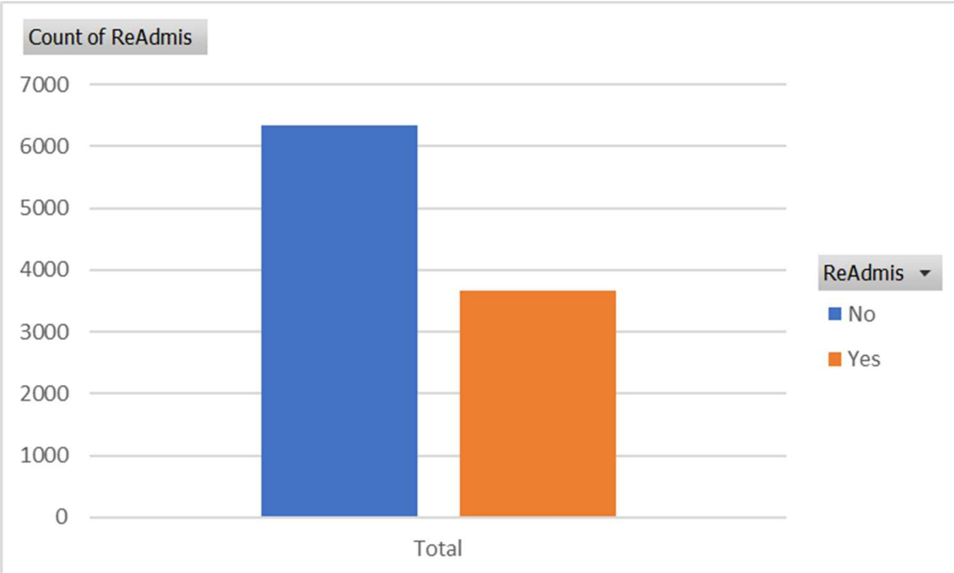
	Column Labels		
	No	Yes	Grand Total
Count of Asthma	7107	2893	10000

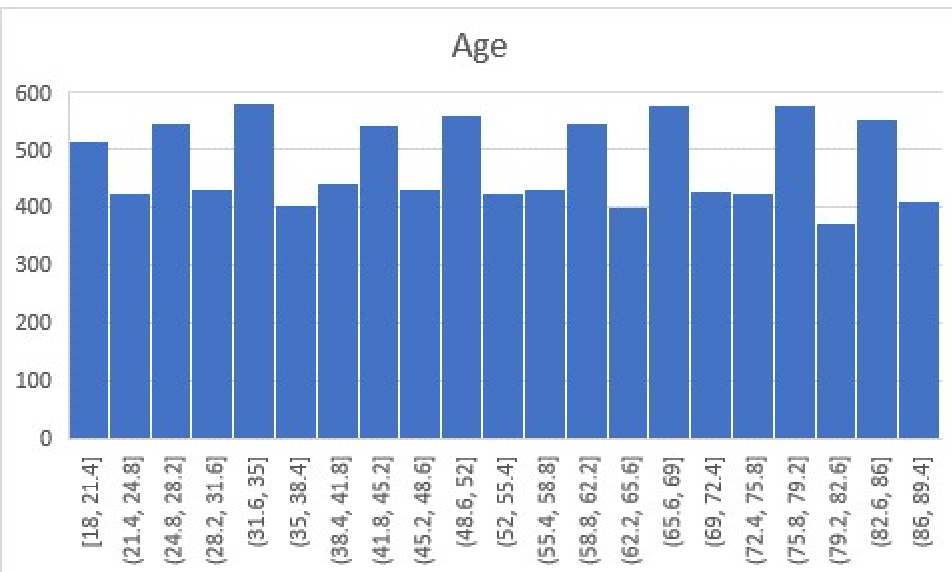
Summary statistics for quantitative independent variables

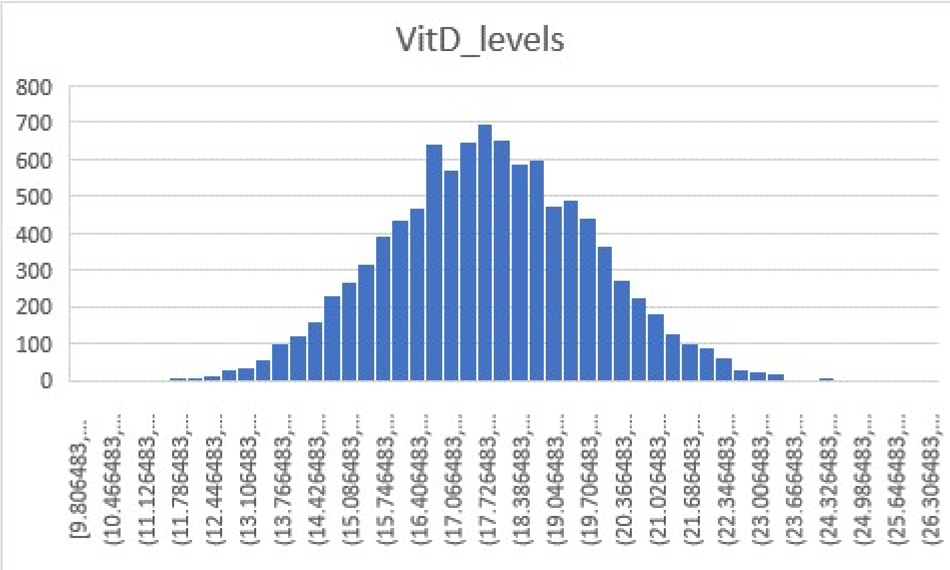
```
> # Summary statistics for independent quantitative variables
> summary(medical$Children)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   1.000   2.097   3.000  10.000 
> summary(medical$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   36.00   53.00   53.51   71.00   89.00 
> summary(medical$Income)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
   154.1  19598.8  33768.4  40490.5  54296.4 207249.1 
> summary(medical$VitD_levels)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  9.806  16.626  17.951  17.964  19.348  26.394 
> summary(medical$Doc_visits)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   4.000   5.000   5.012   6.000   9.000 
> summary(medical$Full_meals_eaten)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   1.000   1.001   2.000   7.000 
> summary(medical$vitD_supp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3989  1.0000  5.0000 
> summary(medical$Initial_days)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.002   7.896  35.836  34.455  61.161  71.981 
> summary(medical$TotalCharge)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1938    3179    5214    5312    7460    9181 
> summary(medical$Additional_charges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3126    7986   11574   12935   15626   30566 
```

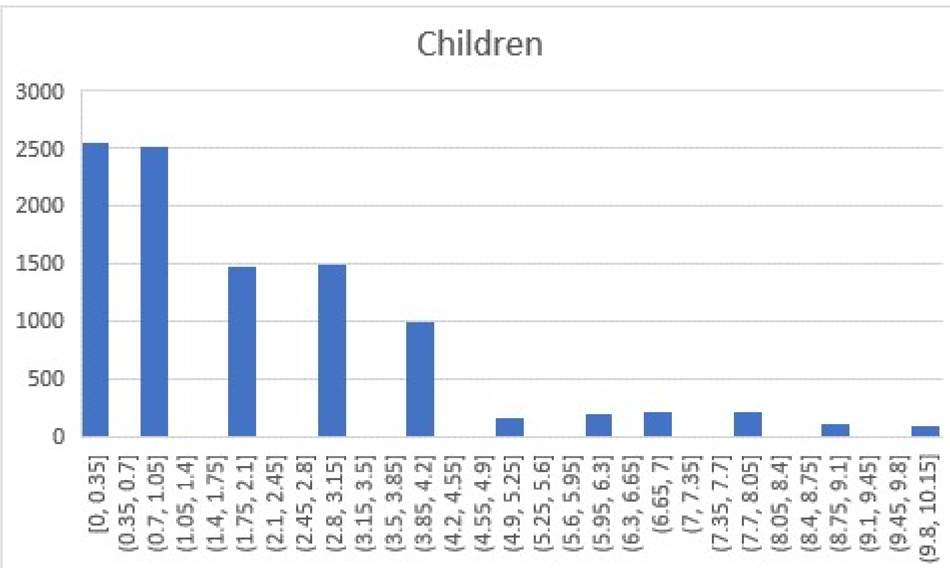
Part C3

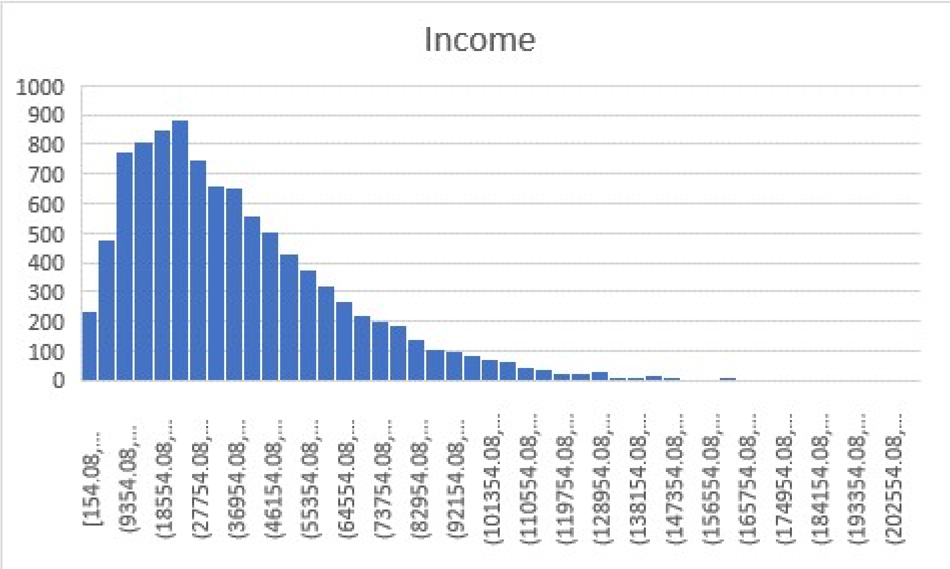
Univariate visualizations (Paula, 2020b)



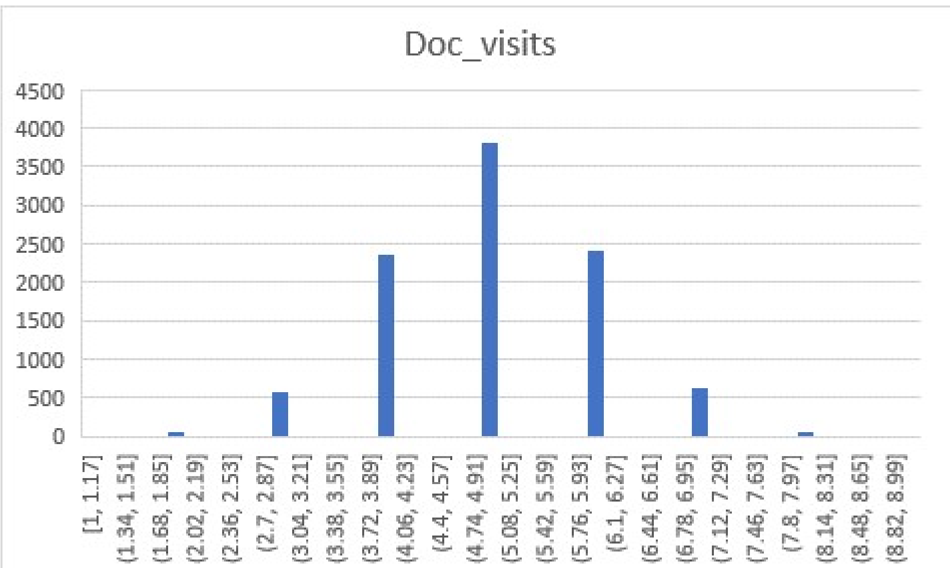



VitD_levels
800
700
600
500
400
300
200
100
0
[9.806483,...
(10.466483,...
(11.126483,...
(11.786483,...
(12.446483,...
(13.106483,...
(13.766483,...
(14.426483,...
(15.086483,...
(15.746483,...
(16.406483,...
(17.066483,...
(17.726483,...
(18.386483,...
(19.046483,...
(19.706483,...
(20.366483,...
(21.026483,...
(21.686483,...
(22.346483,...
(23.006483,...
(23.666483,...
(24.326483,...
(24.986483,...
(25.646483,...
(26.306483,...


Children
3000
2500
2000
1500
1000
500
0
[0, 0.35]
(0.35, 0.7]
(0.7, 1.05]
(1.05, 1.4]
(1.4, 1.75]
(1.75, 2.1]
(2.1, 2.45]
(2.45, 2.8]
(2.8, 3.15]
(3.15, 3.5]
(3.5, 3.85]
(3.85, 4.2]
(4.2, 4.55]
(4.55, 4.9]
(4.9, 5.25]
(5.25, 5.6]
(5.6, 5.95]
(5.95, 6.3]
(6.3, 6.65]
(6.65, 7]
(7, 7.35]
(7.35, 7.7]
(7.7, 8.05]
(8.05, 8.4]
(8.4, 8.75]
(8.75, 9.1]
(9.1, 9.45]
(9.45, 9.8]
(9.8, 10.15]

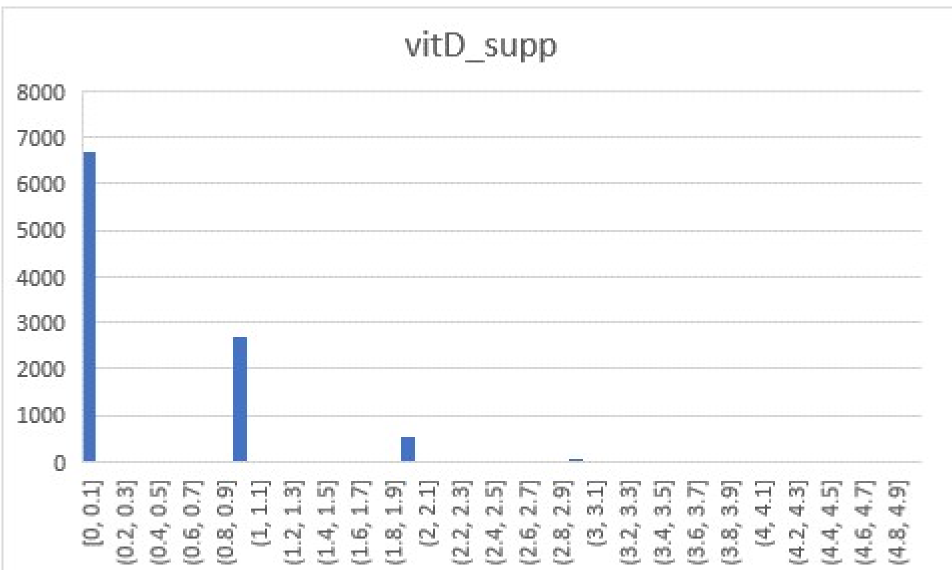


Income
1000
900
800
700
600
500
400
300
200
100
0
[154.08,...
(9354.08,...
(18554.08,...
(27754.08,...
(36954.08,...
(46154.08,...
(55354.08,...
(64554.08,...
(73754.08,...
(82954.08,...
(92154.08,...
(101354.08,...
(110554.08,...
(119754.08,...
(128954.08,...
(138154.08,...
(147354.08,...
(156554.08,...
(165754.08,...
(174954.08,...
(184154.08,...
(193354.08,...
(202554.08,...

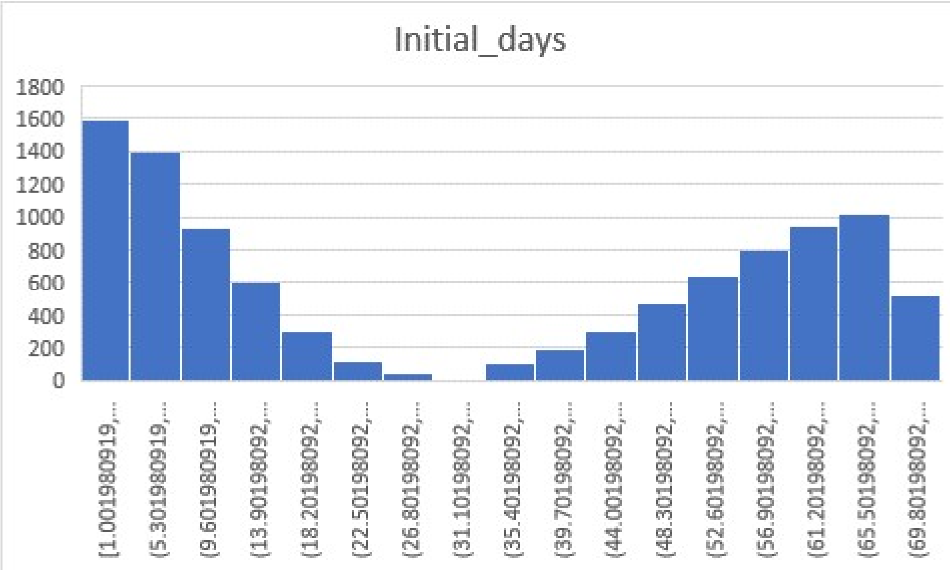


Doc_visits
4500
4000
3500
3000
2500
2000
1500
1000
500
0
[1, 1.17]
(1.34, 1.51]
(1.68, 1.85]
(2.02, 2.19]
(2.36, 2.53]
(2.7, 2.87]
(3.04, 3.21]
(3.38, 3.55]
(3.72, 3.89]
(4.06, 4.23]
(4.4, 4.57]
(4.74, 4.91]
(5.08, 5.25]
(5.42, 5.59]
(5.76, 5.93]
(6.1, 6.27]
(6.44, 6.61]
(6.78, 6.95]
(7.12, 7.29]
(7.46, 7.63]
(7.8, 7.97]
(8.14, 8.31]
(8.48, 8.65]
(8.82, 8.99]

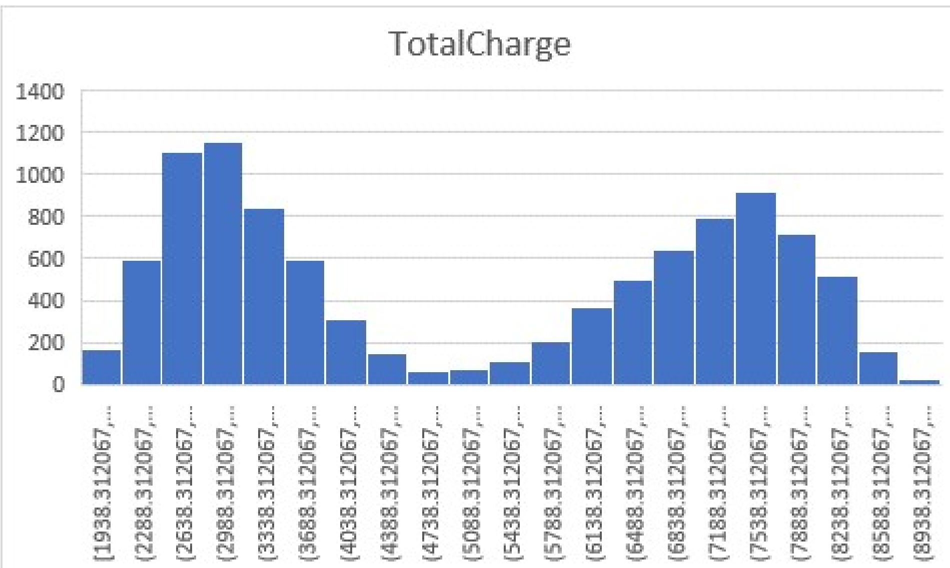
Full_meals_eaten
4000
3500
3000
2500
2000
1500
1000
500
0
[0, 0.16]
(0.32, 0.48]
(0.64, 0.8]
(0.96, 1.12]
(1.28, 1.44]
(1.6, 1.76]
(1.92, 2.08]
(2.24, 2.4]
(2.56, 2.72]
(2.88, 3.04]
(3.2, 3.36]
(3.52, 3.68]
(3.84, 4]
(4.16, 4.32]
(4.48, 4.64]
(4.8, 4.96]
(5.12, 5.28]
(5.44, 5.6]
(5.76, 5.92]
(6.08, 6.24]
(6.4, 6.56]
(6.72, 6.88]



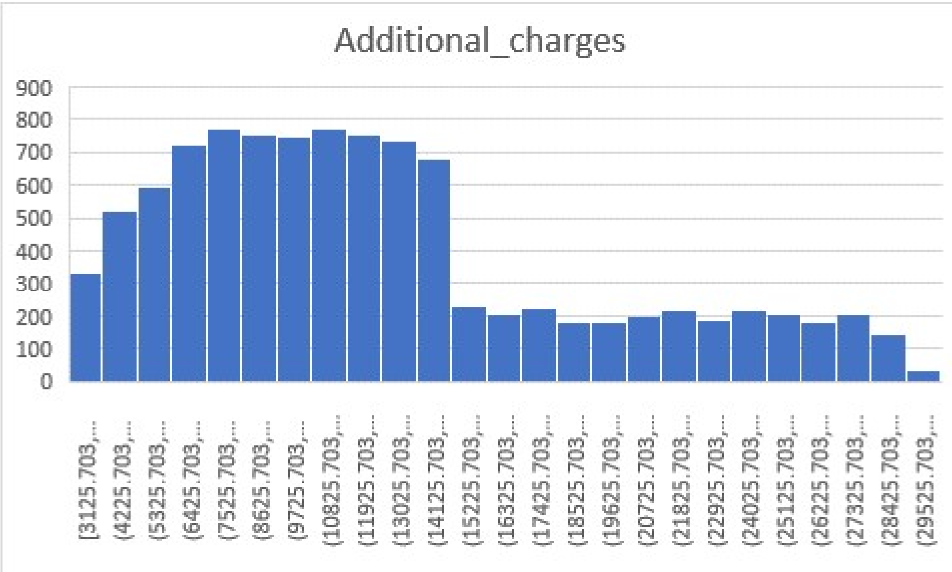
vitD_supp
8000
7000
6000
5000
4000
3000
2000
1000
0
[0, 0.1]
(0.2, 0.3]
(0.4, 0.5]
(0.6, 0.7]
(0.8, 0.9]
(1, 1.1]
(1.2, 1.3]
(1.4, 1.5]
(1.6, 1.7]
(1.8, 1.9]
(2, 2.1]
(2.2, 2.3]
(2.4, 2.5]
(2.6, 2.7]
(2.8, 2.9]
(3, 3.1]
(3.2, 3.3]
(3.4, 3.5]
(3.6, 3.7]
(3.8, 3.9]
(4, 4.1]
(4.2, 4.3]
(4.4, 4.5]
(4.6, 4.7]
(4.8, 4.9]



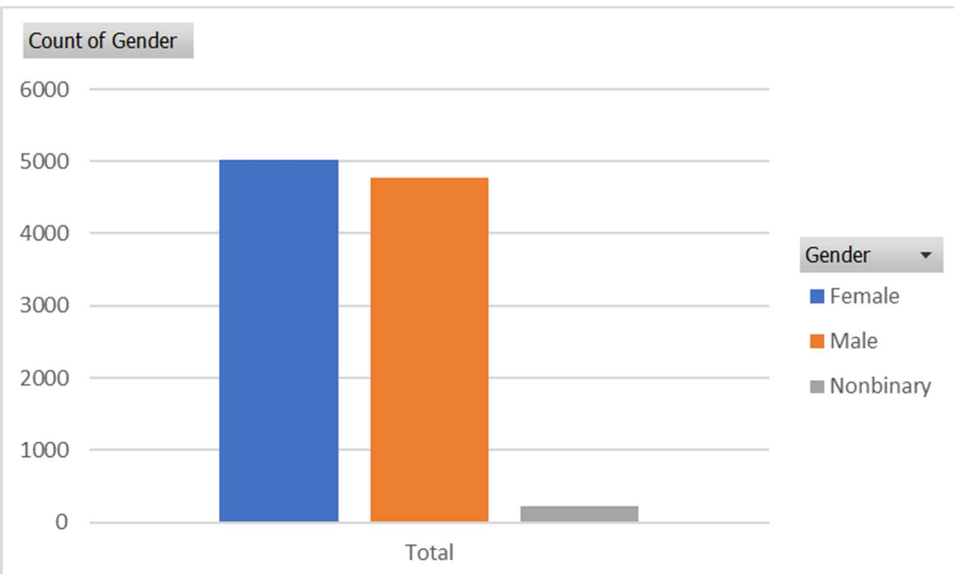
Initial_days
1800
1600
1400
1200
1000
800
600
400
200
0
[1.001980919,...
(5.301980919,...
(9.601980919,...
(13.90198092,...
(18.20198092,...
(22.50198092,...
(26.80198092,...
(31.10198092,...
(35.40198092,...
(39.70198092,...
(44.00198092,...
(48.30198092,...
(52.60198092,...
(56.90198092,...
(61.20198092,...
(65.50198092,...
(69.80198092,...



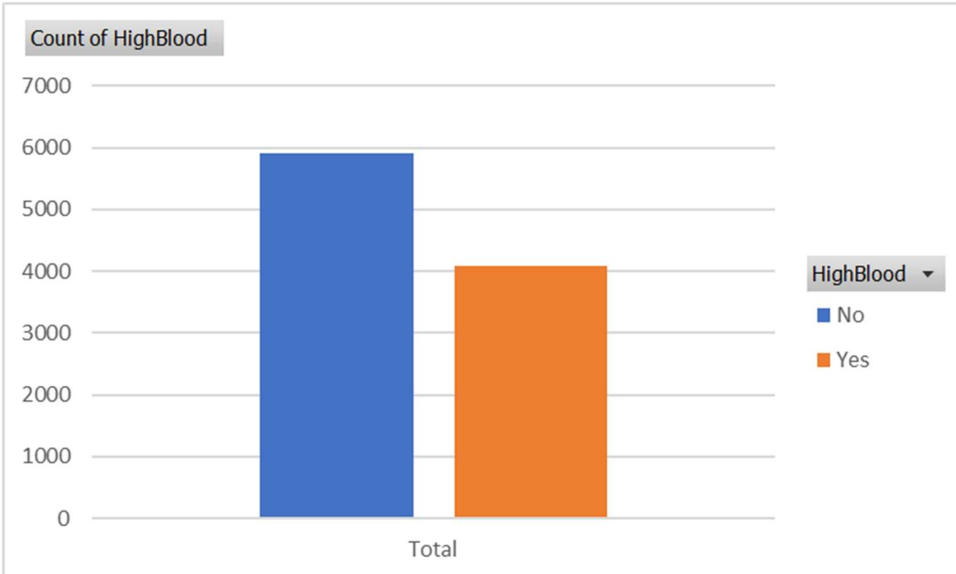
TotalCharge
1400
1200
1000
800
600
400
200
0
[1938.312067,...
(2288.312067,...
(2638.312067,...
(2988.312067,...
(3338.312067,...
(3688.312067,...
(4038.312067,...
(4388.312067,...
(4738.312067,...
(5088.312067,...
(5438.312067,...
(5788.312067,...
(6138.312067,...
(6488.312067,...
(6838.312067,...
(7188.312067,...
(7538.312067,...
(7888.312067,...
(8238.312067,...
(8588.312067,...
(8938.312067,...

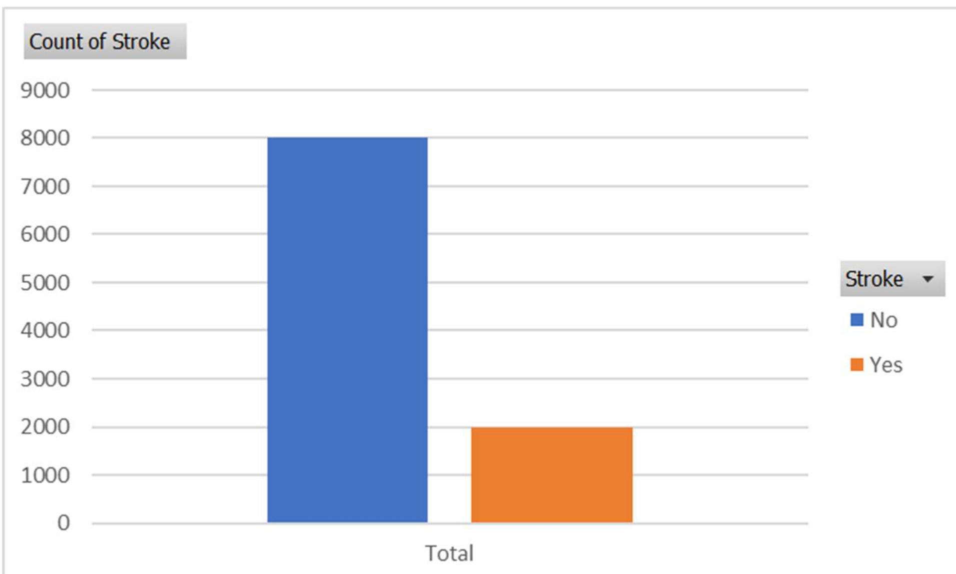


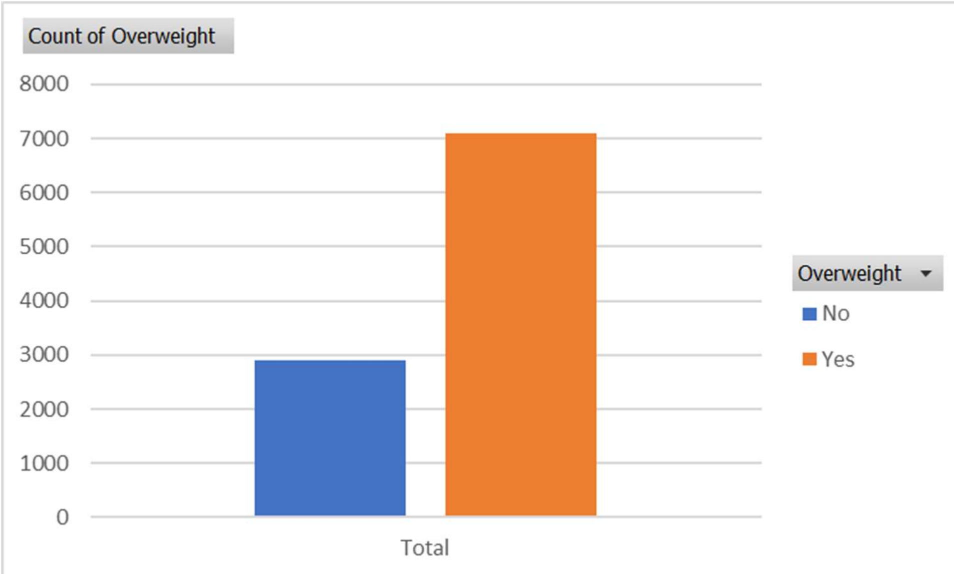
Additional_charges
900
800
700
600
500
400
300
200
100
0
[3125.703,...
(4225.703,...
(5325.703,...
(6425.703,...
(7525.703,...
(8625.703,...
(9725.703,...
(10825.703,...
(11925.703,...
(13025.703,...
(14125.703,...
(15225.703,...
(16325.703,...
(17425.703,...
(18525.703,...
(19625.703,...
(20725.703,...
(21825.703,...
(22925.703,...
(24025.703,...
(25125.703,...
(26225.703,...
(27325.703,...
(28425.703,...
(29525.703,...

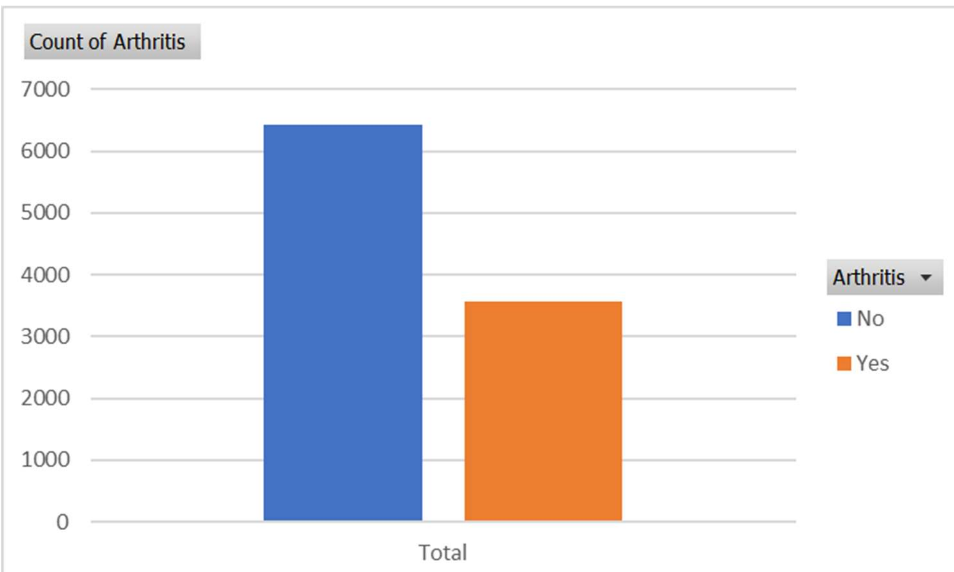


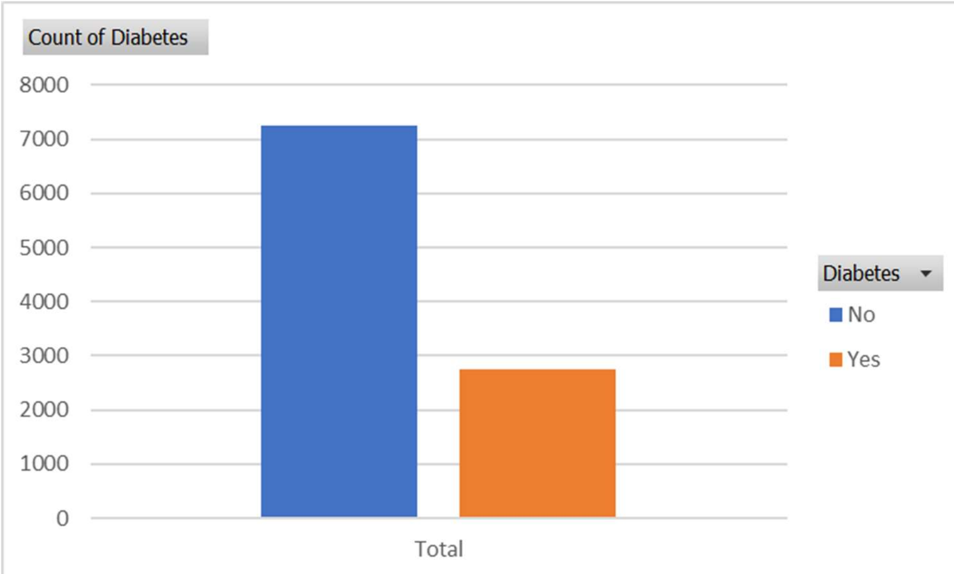
Count of Gender
6000
5000
4000
3000
2000
1000
0
Total
Gender
Female
Male
Nonbinary

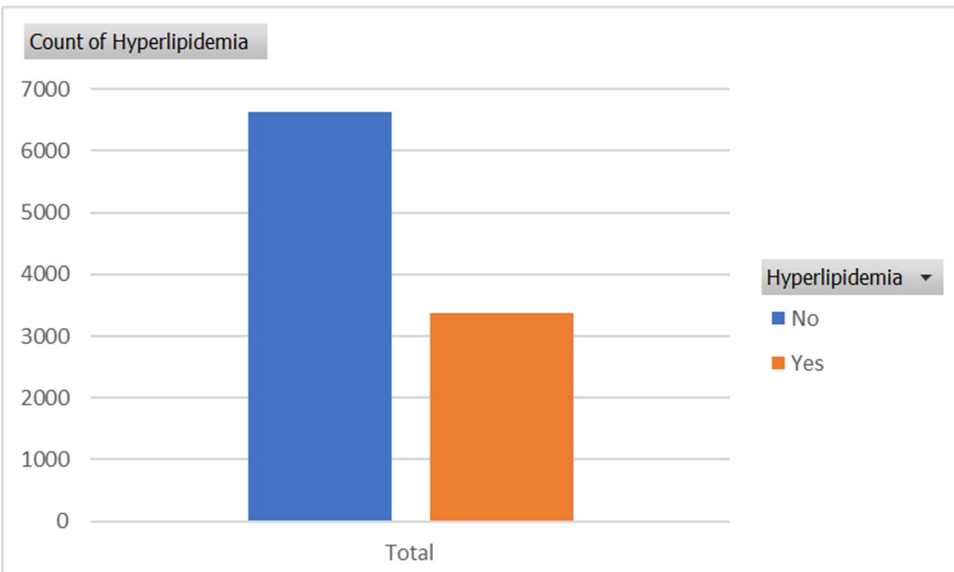

Count of HighBlood
7000
6000
5000
4000
3000
2000
1000
0
Total
HighBlood
No
Yes

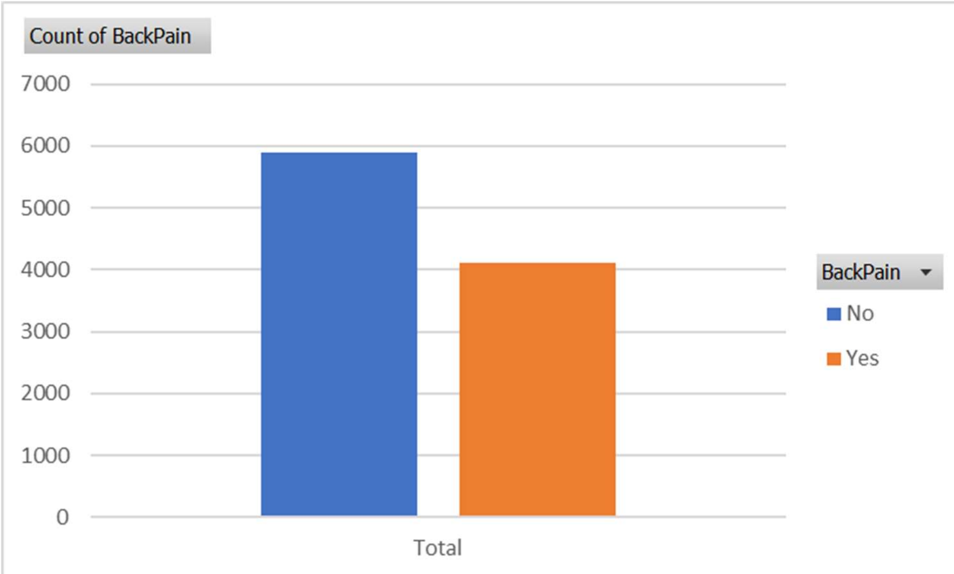

Count of Stroke
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Total
Stroke
No
Yes

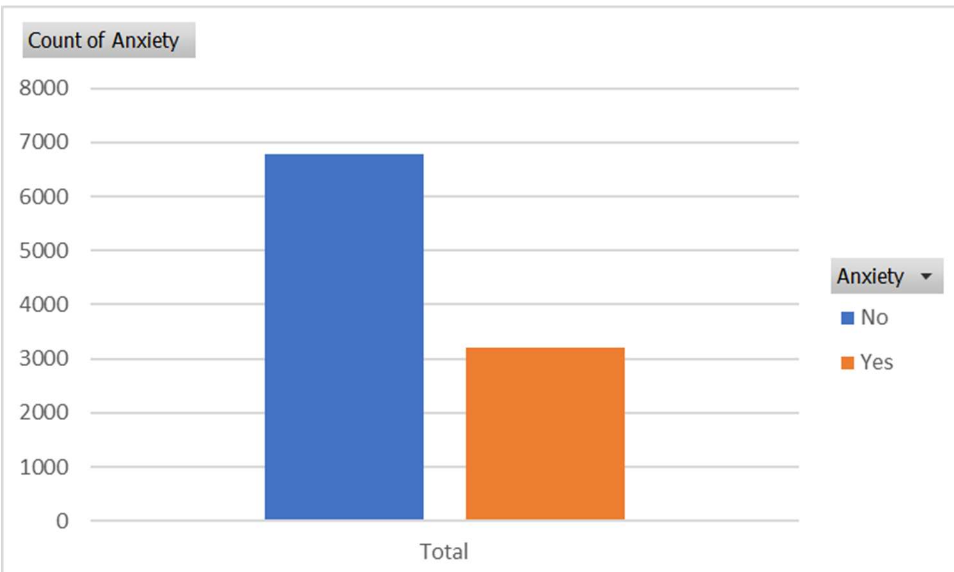

Count of Overweight
8000
7000
6000
5000
4000
3000
2000
1000
0
Total
Overweight
No
Yes

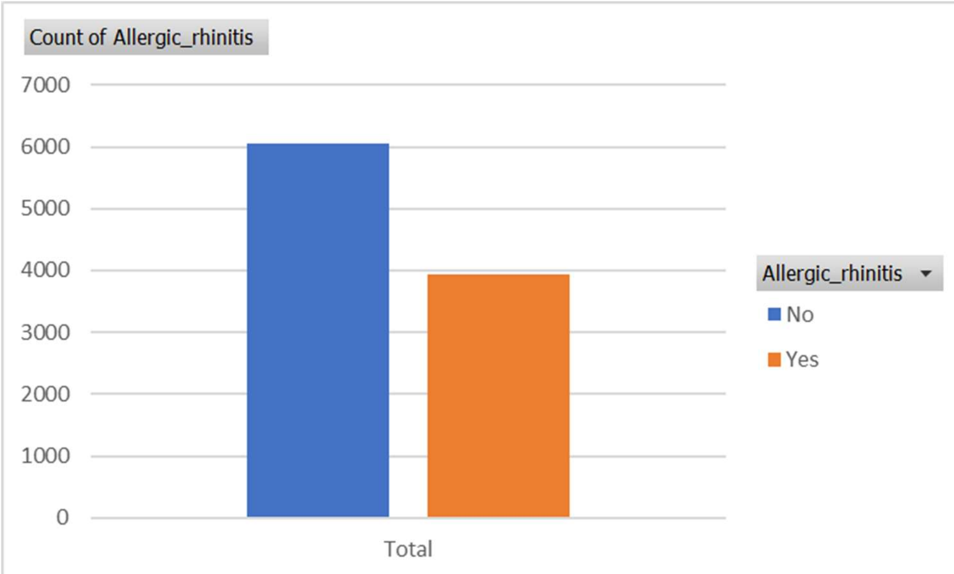

Count of Arthritis
7000
6000
5000
4000
3000
2000
1000
0
Total
Arthritis
No
Yes

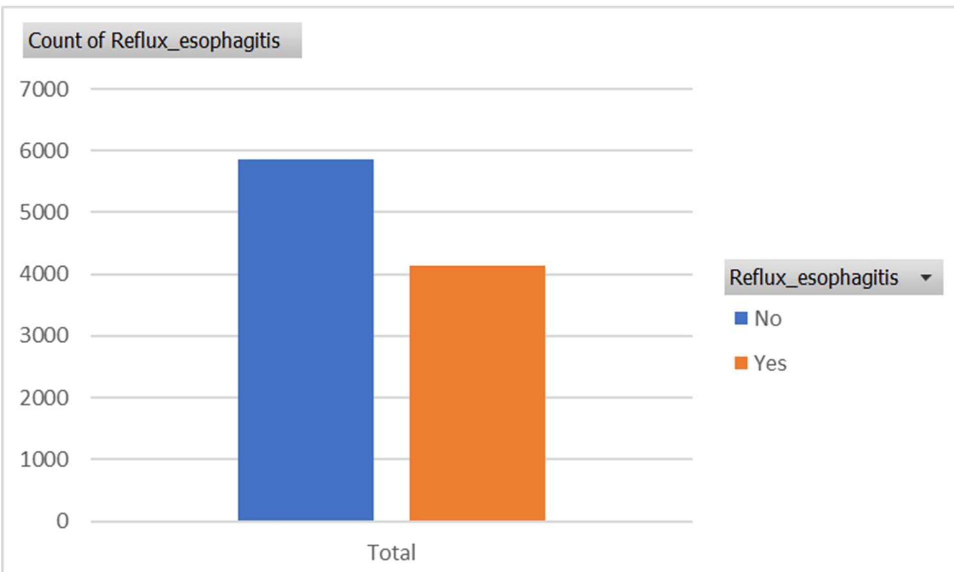

Count of Diabetes
8000
7000
6000
5000
4000
3000
2000
1000
0
Total
Diabetes
No
Yes

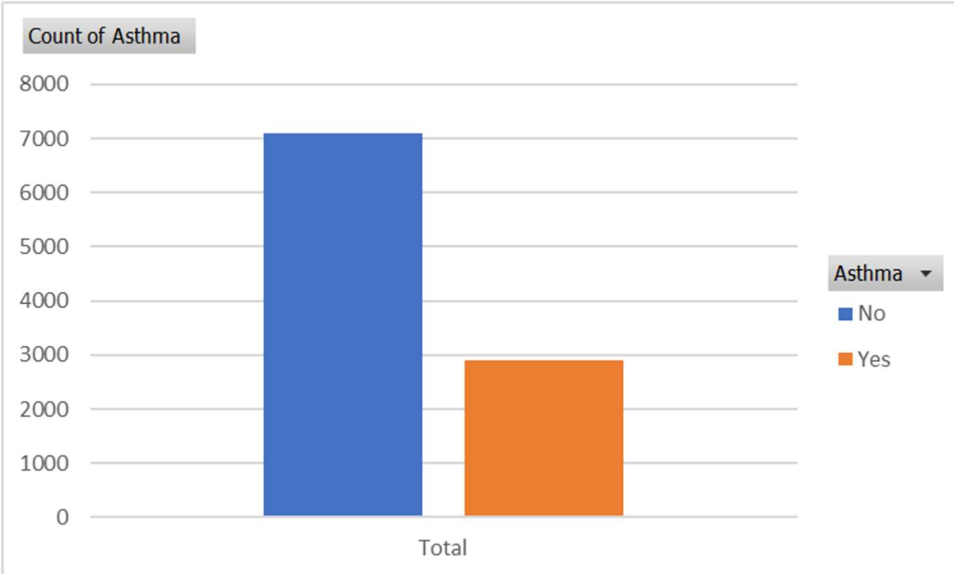

Count of Hyperlipidemia
7000
6000
5000
4000
3000
2000
1000
0
Total
Hyperlipidemia
No
Yes


Count of BackPain
7000
6000
5000
4000
3000
2000
1000
0
Total
BackPain
No
Yes

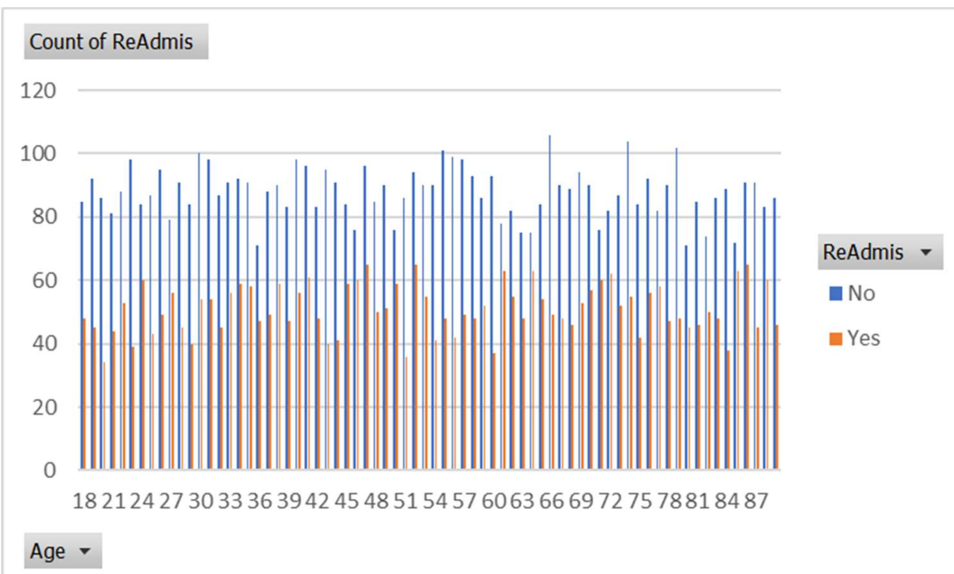

Count of Anxiety
8000
7000
6000
5000
4000
3000
2000
1000
0
Total
Anxiety
No
Yes

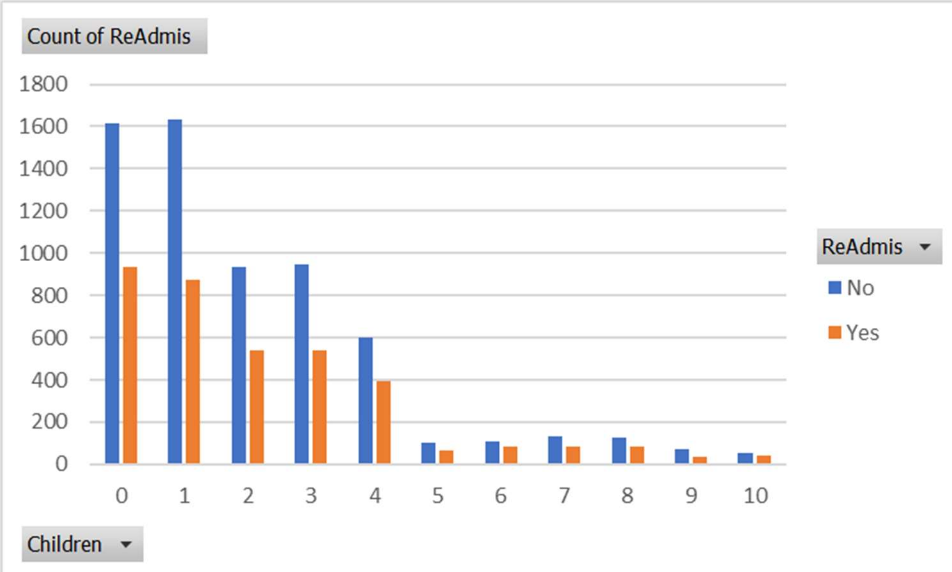

Count of Allergic_rhinitis
7000
6000
5000
4000
3000
2000
1000
0
Total
Allergic_rhinitis
No
Yes

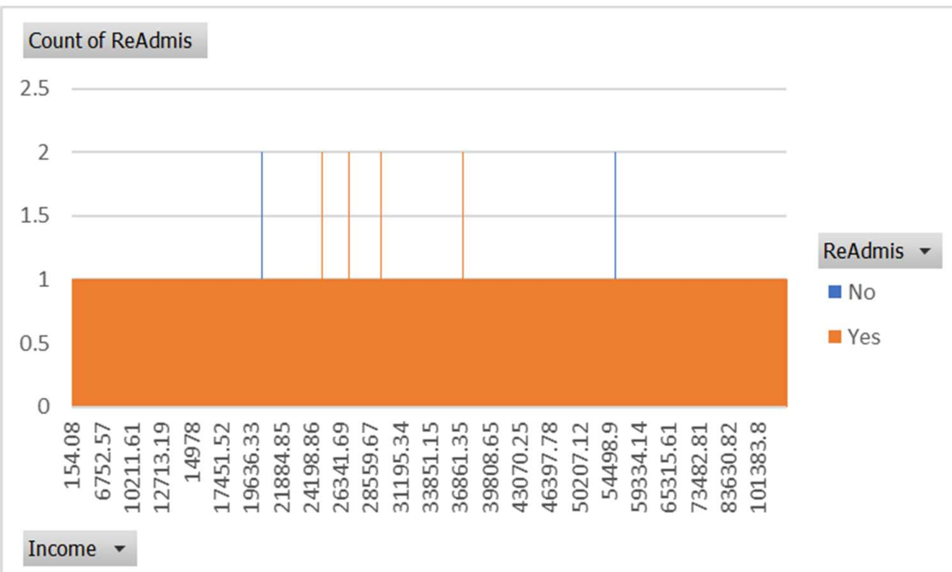

Count of Reflux_esophagitis
7000
6000
5000
4000
3000
2000
1000
0
Total
Reflux_esophagitis
No
Yes

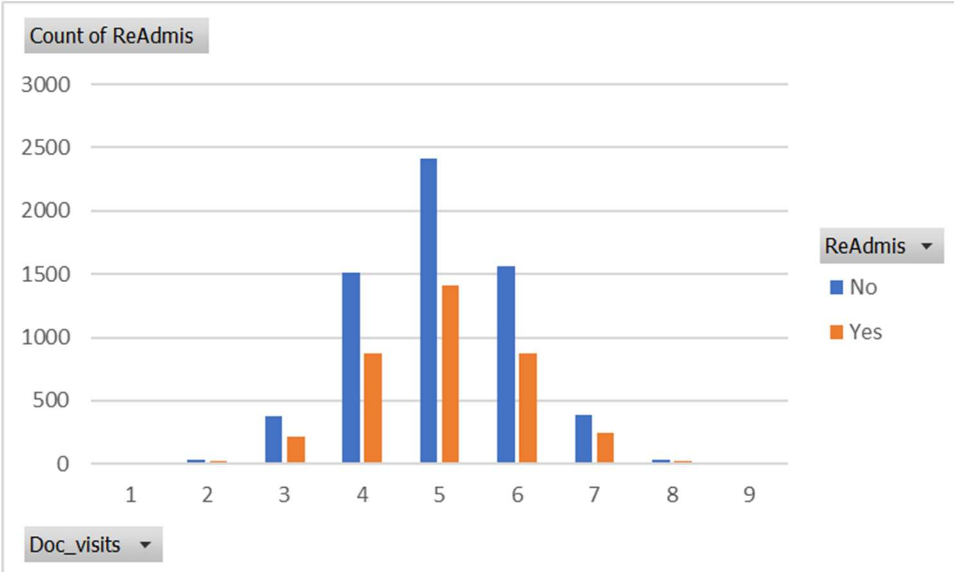


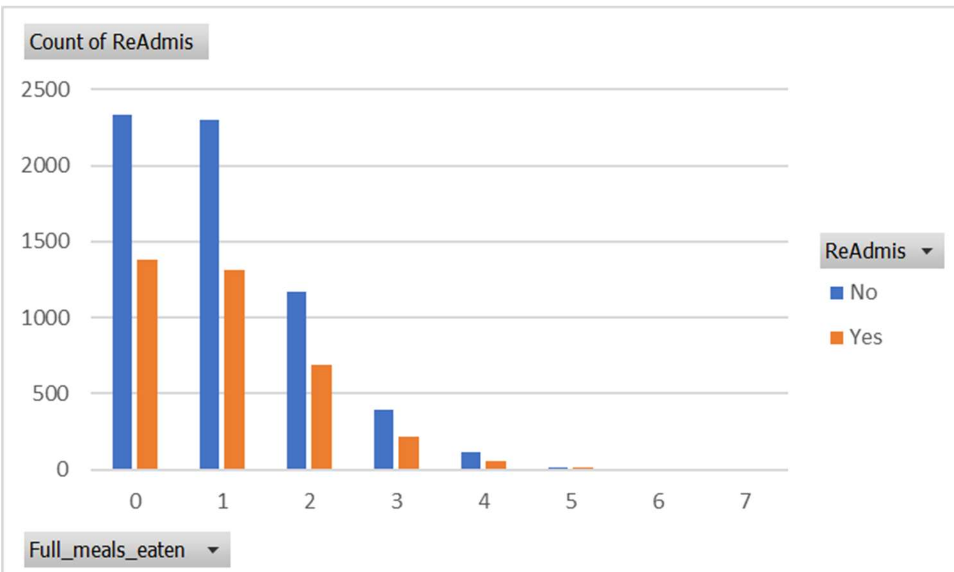
Bivariate visualizations (Paula, 2020b)

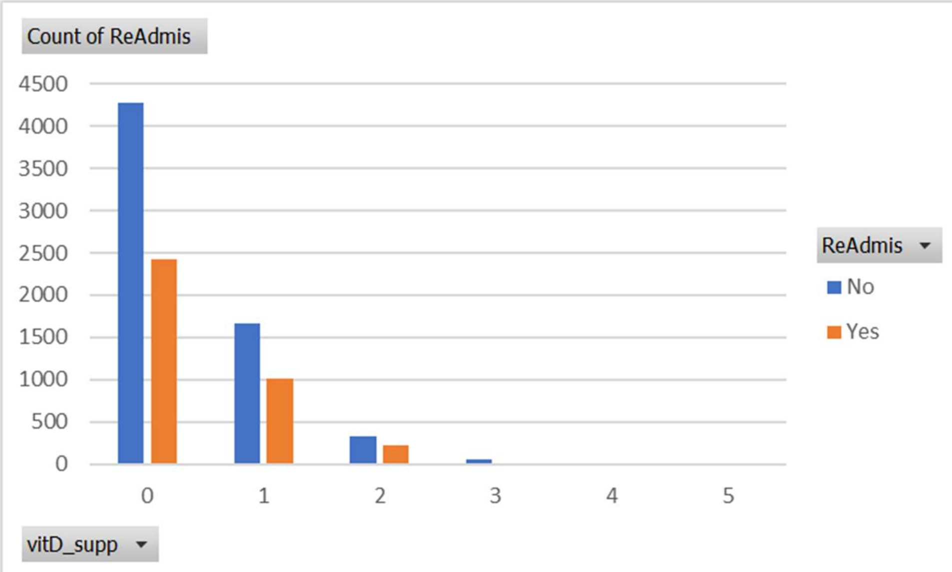


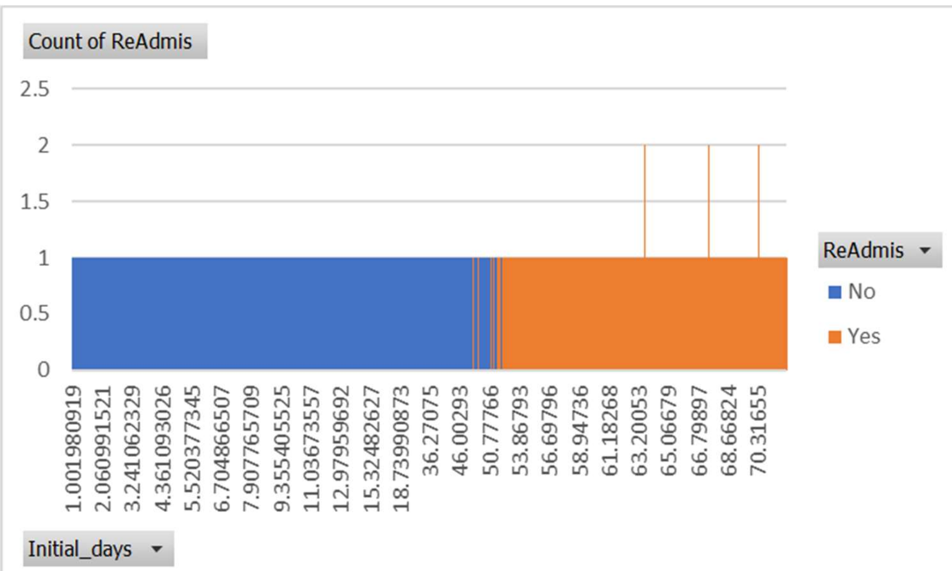

Count of ReAdmis
1800
1600
1400
1200
1000
800
600
400
200
0
0
1
2
3
4
5
6
7
8
9
10
Children
ReAdmis
No
Yes

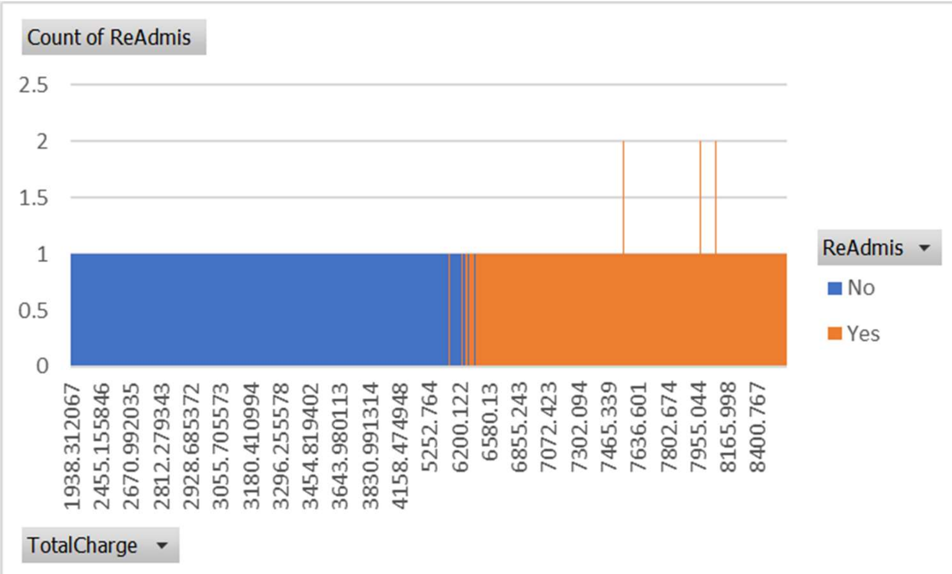

Count of ReAdmis
2.5
2
1.5
1
0.5
0
154.08
6752.57
10211.61
12713.19
14978
17451.52
19636.33
21884.85
24198.86
26341.69
28559.67
31195.34
33851.15
36861.35
39808.65
43070.25
46397.78
50207.12
54498.9
59334.14
65315.61
73482.81
83630.82
101383.8
Income
ReAdmis
No
Yes

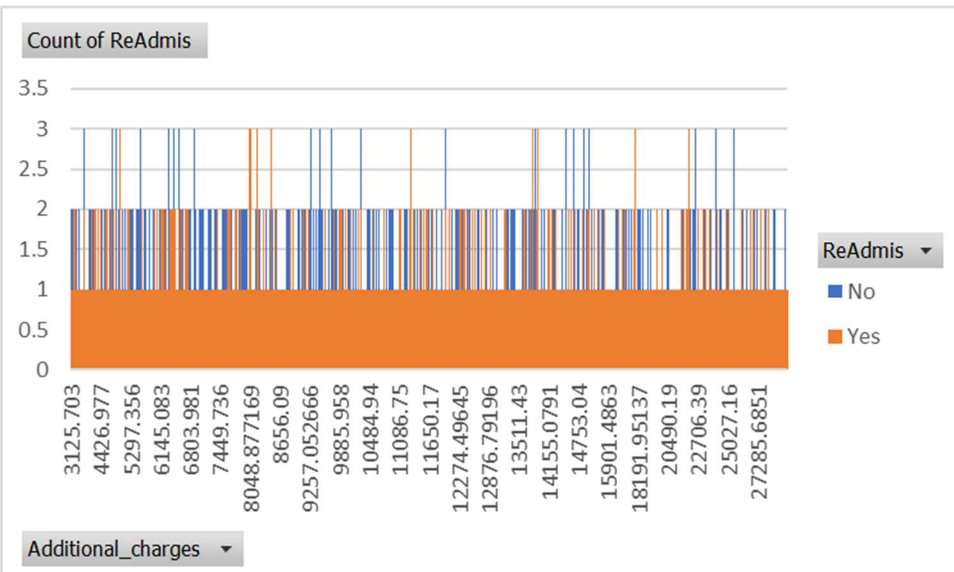

Count of ReAdmis
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
Doc_visits
ReAdmis
No
Yes

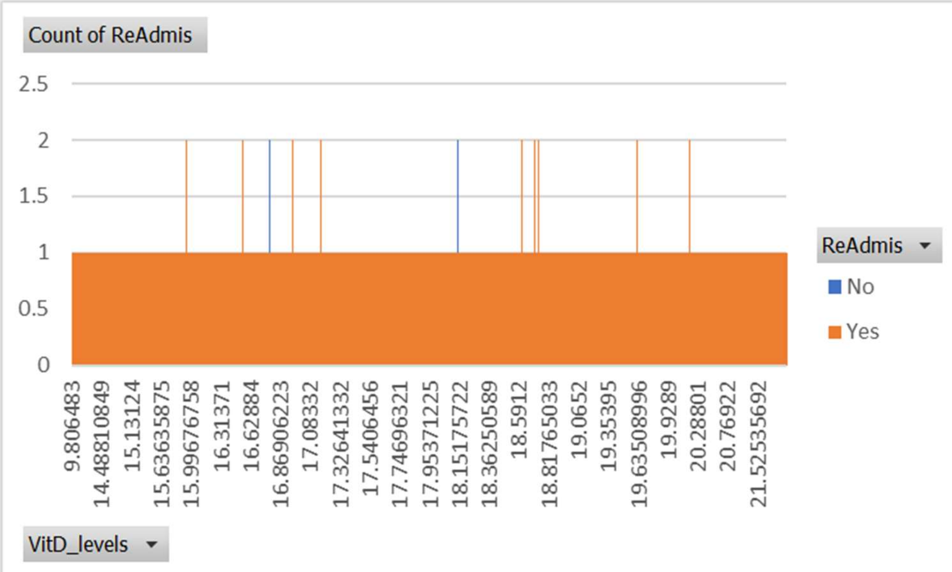

Count of ReAdmis
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
Full_meals_eaten
ReAdmis
No
Yes

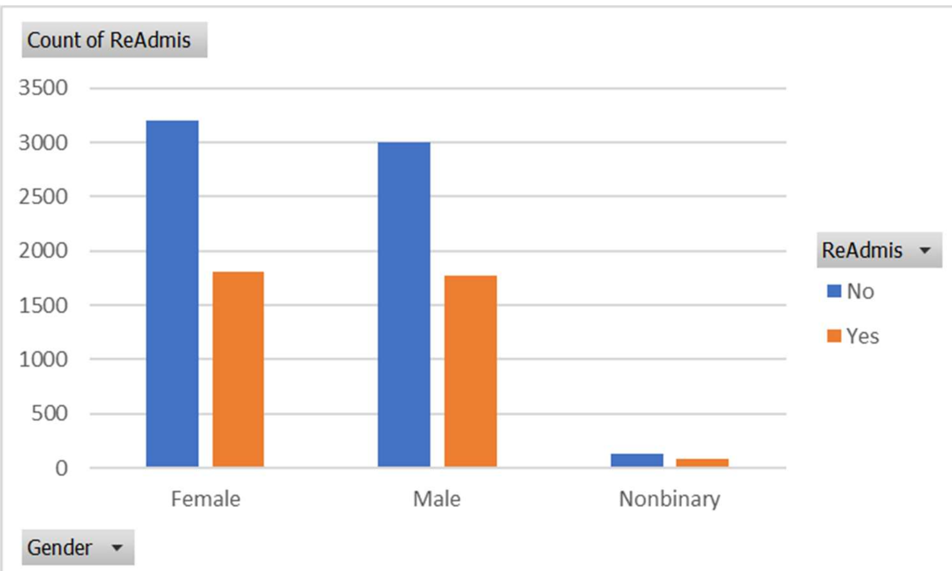

Count of ReAdmis
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
vitD_supp
ReAdmis
No
Yes

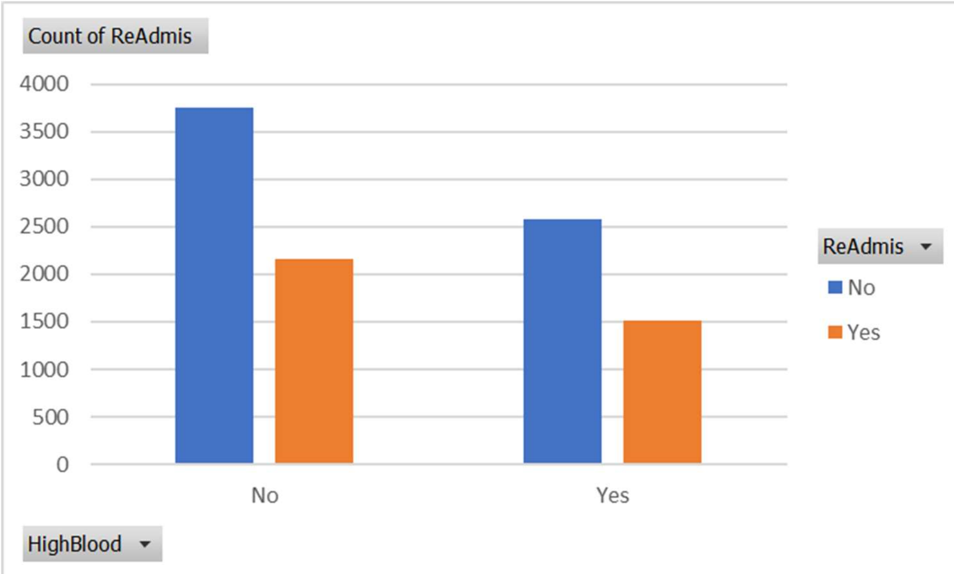

Count of ReAdmis
2.5
2
1.5
1
0.5
0
1.001980919
2.060991521
3.241062329
4.361093026
5.520377345
6.704866507
7.907765709
9.355405525
11.03673557
12.97959692
15.32482627
18.73990873
36.27075
46.00293
50.77766
53.86793
56.69796
58.94736
61.18268
63.20053
65.06679
66.79897
68.66824
70.31655
Initial_days
ReAdmis
No
Yes


Count of ReAdmis
2.5
2
1.5
1
0.5
0
ReAdmis
No
Yes
1938.312067
2455.155846
2670.992035
2812.279343
2928.685372
3055.705573
3180.410994
3296.255578
3454.819402
3643.980113
3830.991314
4158.474948
5252.764
6200.122
6580.13
6855.243
7072.423
7302.094
7465.339
7636.601
7802.674
7955.044
8165.998
8400.767
TotalCharge

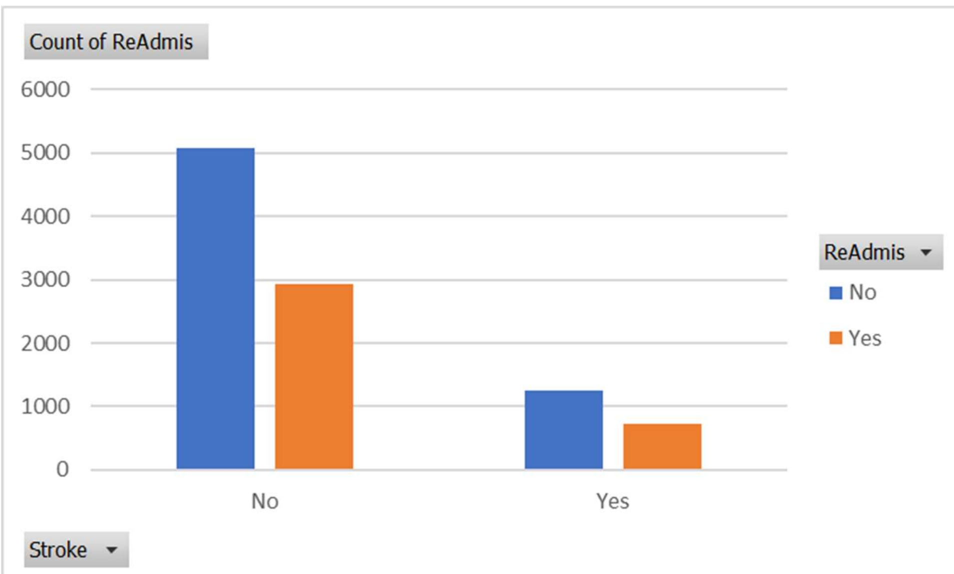

Count of ReAdmis
3.5
3
2.5
2
1.5
1
0.5
0
ReAdmis
No
Yes
3125.703
4426.977
5297.356
6145.083
6803.981
7449.736
8048.877169
8656.09
9257.052666
9885.958
10484.94
11086.75
11650.17
12274.49645
12876.79196
13511.43
14155.0791
14753.04
15901.4863
18191.95137
20490.19
22706.39
25027.16
27285.6851
Additional_charges


Count of ReAdmis
2.5
2
1.5
1
0.5
0
ReAdmis
No
Yes
9.806483
14.48810849
15.13124
15.63635875
15.99676758
16.31371
16.62884
16.86906223
17.08332
17.32641332
17.5406456
17.74696321
17.95371225
18.15175722
18.36250589
18.5912
18.81765033
19.0652
19.35395
19.63508996
19.9289
20.28801
20.76922
21.52535692
VitD_levels

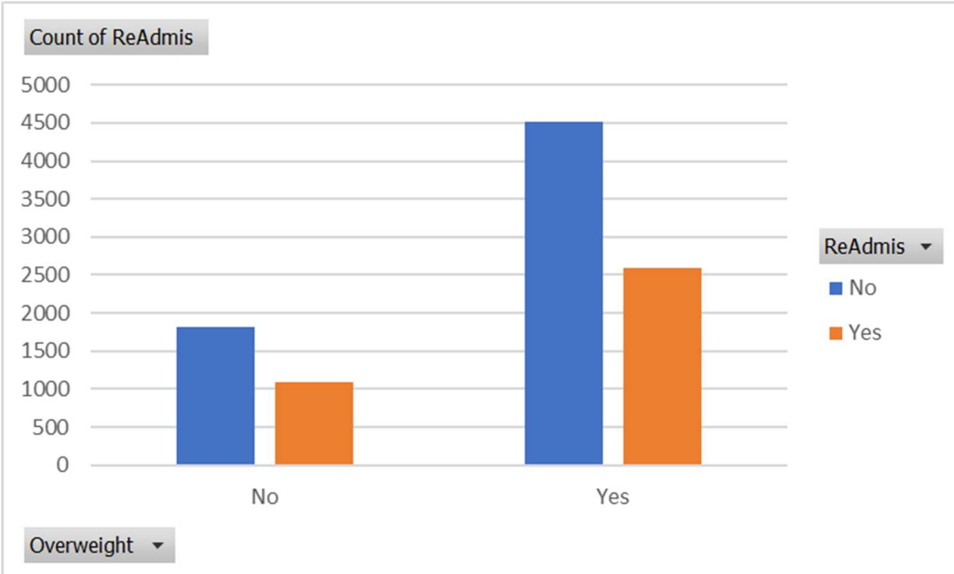

Count of ReAdmis
3500
3000
2500
2000
1500
1000
500
0
Female
Male
Nonbinary
ReAdmis
No
Yes
Gender

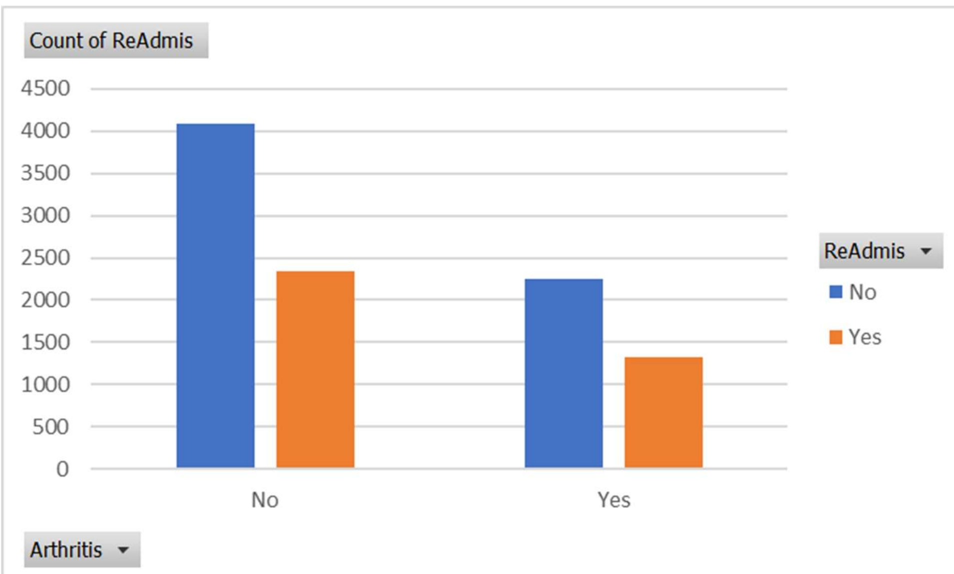


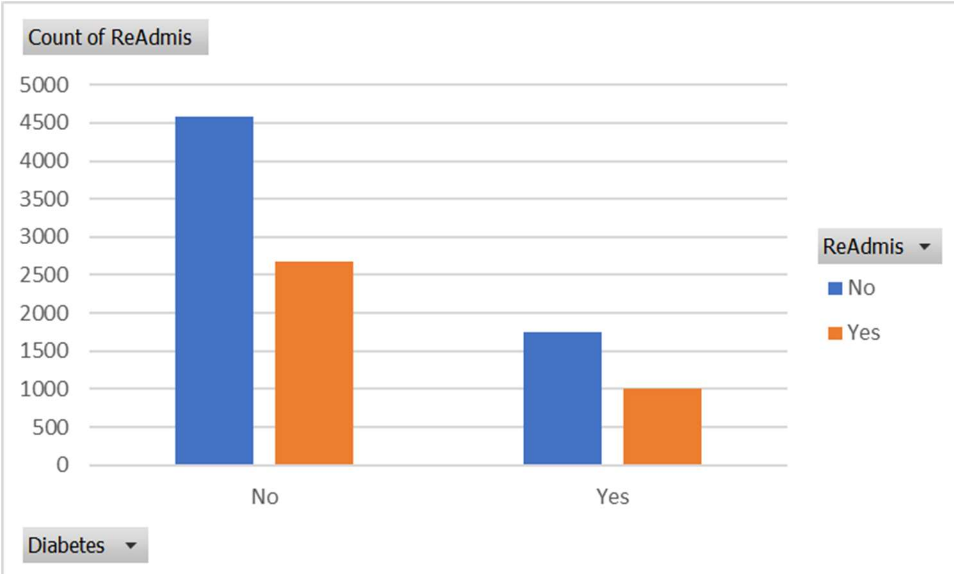
Count of ReAdmis
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
HighBlood
ReAdmis
No
Yes



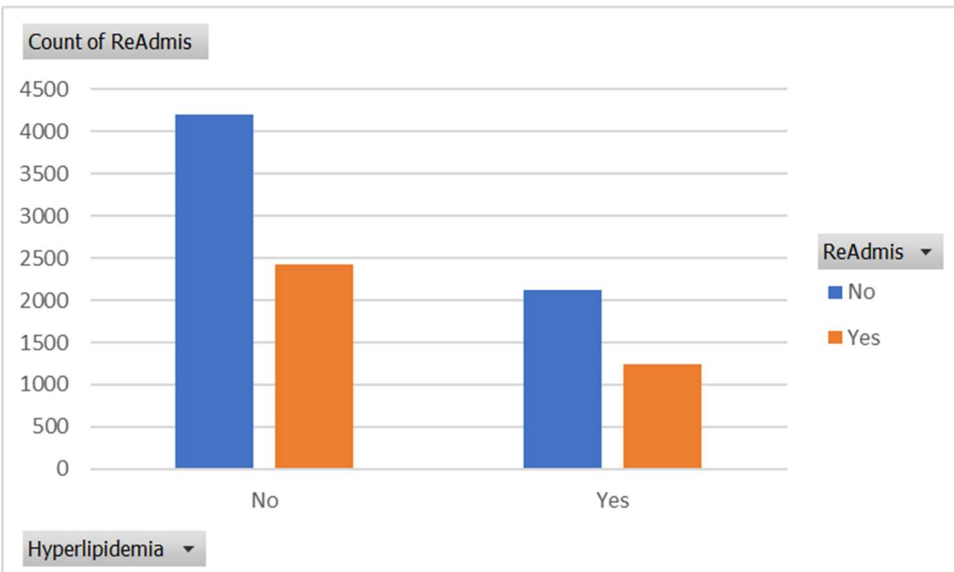
Count of ReAdmis
6000
5000
4000
3000
2000
1000
0
No
Yes
Stroke
ReAdmis
No
Yes


Count of ReAdmis
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
ReAdmis
No
Yes
Overweight

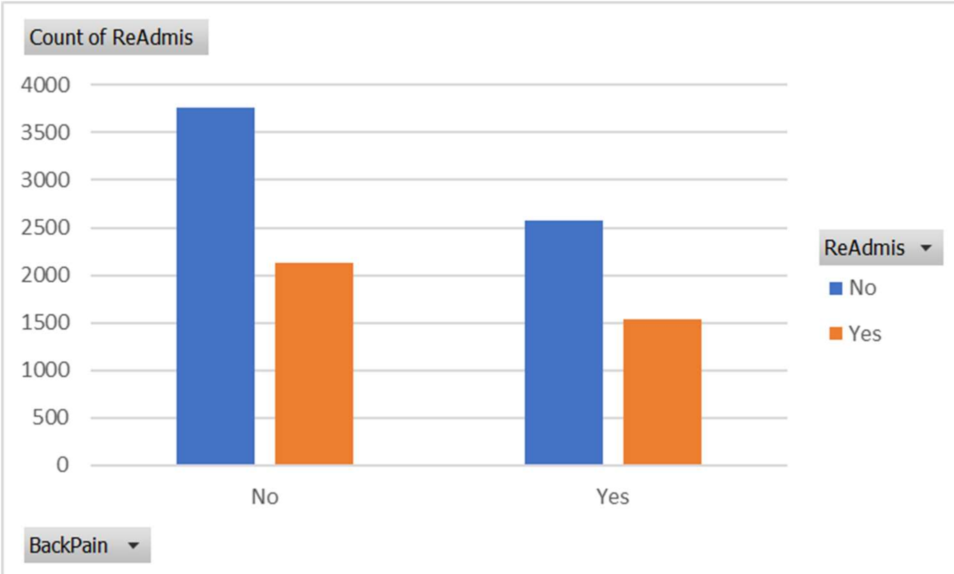

Count of ReAdmis
4500
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
ReAdmis
No
Yes
Arthritis

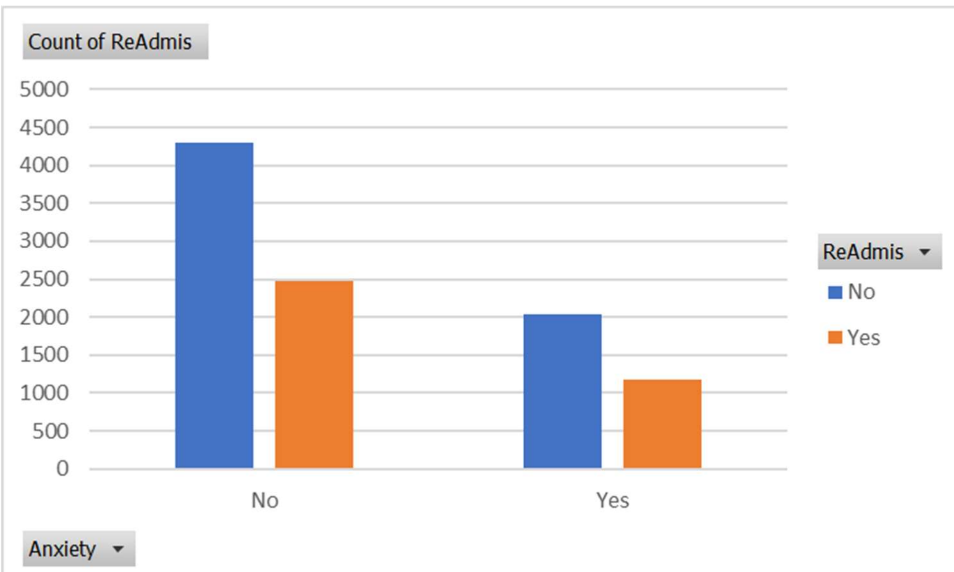


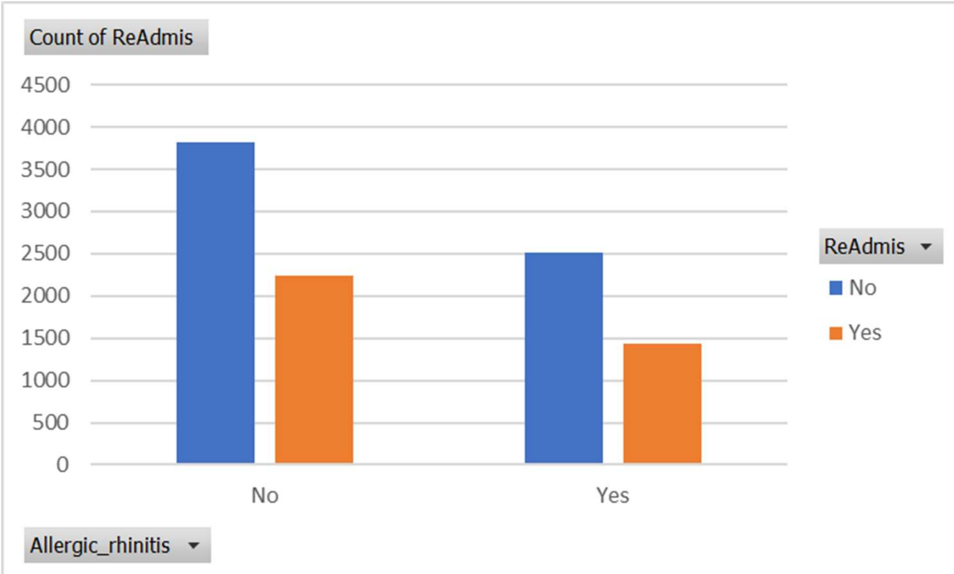
Count of ReAdmis
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
ReAdmis
No
Yes
Diabetes



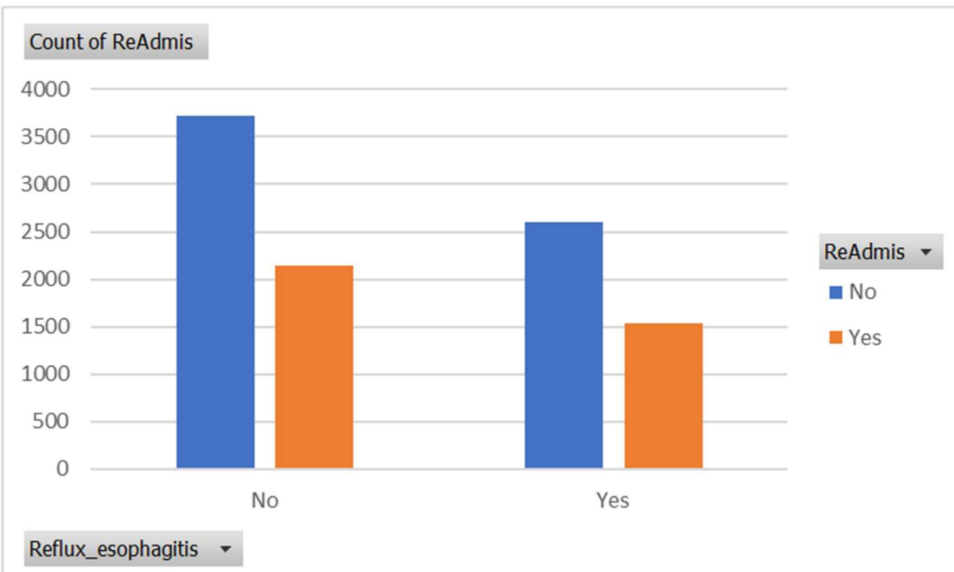
Count of ReAdmis
4500
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
ReAdmis
No
Yes
Hyperlipidemia


Count of ReAdmis
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
ReAdmis
No
Yes
BackPain

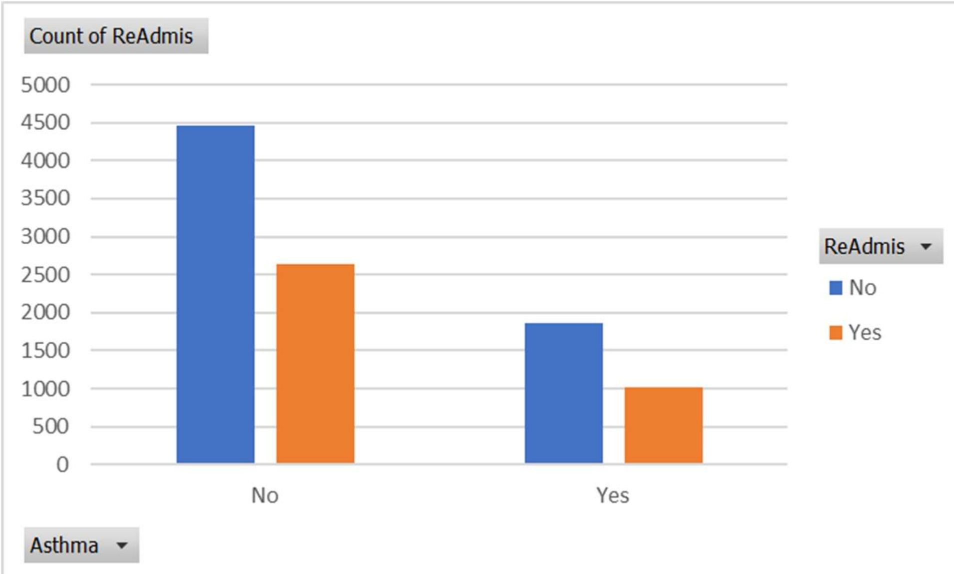

Count of ReAdmis
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
ReAdmis
No
Yes
Anxiety



Count of ReAdmis
4500
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
ReAdmis
No
Yes
Allergic_rhinitis



Count of ReAdmis
4000
3500
3000
2500
2000
1500
1000
500
0
No
Yes
ReAdmis
No
Yes
Reflux_esophagitis



Part C4

Categorical variables to be used in the logistic regression model will need to be re-expressed. The Gender variable was re-expressed using the one-hot encoding method with the dummyVars() function. The remaining categorical variables were re-expressed using the label method with the lapply() and revalue() functions. See attached code. (Zach, 2021a)

Part C5

See attached file.

Part D1

The initial logistic regression model is:

ln(p^/(1-p^)) = (-7.647e+01 - 1.845e-02(Age) - 5.412e-01(GenderFemale) - 4.460e-01(GenderMale) + 4.118e-02(VitD_levels) - 1.946e-01(HighBlood) + 1.523e+00(Stroke) - 2.556e-01(Overweight) - 1.339e+00(Arthritis) + 2.082e-01(Diabetes) + 4.686e-02(Hyperlipidemia) + 1.043e-01(BackPain) - 1.117e+00(Anxiety) - 4.792e-01(Allergic_rhinitis) - 4.915e-01(Reflux_esophagitis) -1.135e+00(Asthma) + 6.808e-02(Children) + 5.854e-07(Income) + 7.721e-03(Doc_visits) + 5.661e-02(Full_meals_eaten) - 1.115e-01(vitD_supp) + 1.070e+00(Initial_days) + 2.765e-03(TotalCharge) + 8.316e-05(Additional_charges)

```
> summary(logres_initial)

Call:
glm(formula = ReAdmis ~ Age + GenderFemale + GenderMale + GenderNonbinary +
    VitD_levels + HighBlood + Stroke + Overweight + Arthritis +
    Diabetes + Hyperlipidemia + BackPain + Anxiety + Allergic_rhinitis +
    Reflux_esophagitis + Asthma + Children + Income + Doc_visits +
    Full_meals_eaten + vitD_supp + Initial_days + TotalCharge +
    Additional_charges, family = binomial, data = medical_encoded)

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.647e+01  4.175e+00 -18.315  < 2e-16 ***
Age                -1.845e-02  1.367e-02  -1.349   0.1772
GenderFemale       -5.412e-01  6.509e-01  -0.831   0.4057
GenderMale         -4.460e-01  6.520e-01  -0.684   0.4940
GenderNonbinary            NA         NA      NA       NA
VitD_levels         4.118e-02  4.528e-02   0.909   0.3631
HighBlood          -1.946e-01  5.344e-01  -0.364   0.7157
Stroke              1.523e+00  2.516e-01   6.052 1.43e-09 ***
Overweight         -2.556e-01  2.118e-01  -1.207   0.2275
Arthritis          -1.339e+00  2.136e-01  -6.268 3.65e-10 ***
Diabetes            2.082e-01  2.146e-01   0.970   0.3321
Hyperlipidemia      4.686e-02  2.044e-01   0.229   0.8186
BackPain            1.043e-01  1.935e-01   0.539   0.5900
Anxiety            -1.117e+00  2.133e-01  -5.239 1.62e-07 ***
Allergic_rhinitis  -4.792e-01  1.973e-01  -2.429   0.0151 *
Reflux_esophagitis -4.915e-01  2.002e-01  -2.455   0.0141 *
Asthma             -1.135e+00  2.143e-01  -5.297 1.18e-07 ***
Children            6.808e-02  4.257e-02   1.599   0.1097
Income              5.854e-07  3.363e-06   0.174   0.8618
Doc_visits          7.721e-03  8.855e-02   0.087   0.9305
Full_meals_eaten    5.661e-02  9.495e-02   0.596   0.5510
vitD_supp          -1.115e-01  1.497e-01  -0.745   0.4564
Initial_days        1.070e+00  6.221e-02  17.204  < 2e-16 ***

TotalCharge         2.765e-03  3.292e-04   8.399  < 2e-16 ***
Additional_charges  8.316e-05  5.828e-05   1.427   0.1536
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13145.7  on 9999  degrees of freedom
Residual deviance:   751.8  on 9976  degrees of freedom
AIC: 799.8

Number of Fisher Scoring iterations: 12
```

Part D2

Backward Stepwise Elimination was used as a feature selection procedure to reduce the initial model. This procedure allowed for first evaluating all the possible explanatory variables, and then improving the performance of the model by removing least significant features based on their p-value. The cutoff p-value of 0.05 was used to determine whether an independent variable was statistically significant. This allowed for the model to be evaluated at multiple steps by removing variables with a p-value greater than 0.05, one at a time, until an acceptable model was achieved.

Part D3

The reduced logistic regression model is:

ln(p^/(1-p^)) = (-7.556e+01 – 1.398e-02(Age) + 1.497e+00(Stroke) – 1.323e+00(Arthritis)
– 1.086e+00(Anxiety) – 4.823e-01(Allergic_rhinits) – 4.738e-01(Reflux_esophagitis)
– 1.139e+00(Asthma) + 1.054e+00(Initial_days) + 2.815e-03(TotalCharge)
+ 5.976e-05(Additional_charges)

```
> summary(logres_final)

Call:
glm(formula = ReAdmis ~ Age + Stroke + Arthritis + Anxiety +
    Allergic_rhinitis + Reflux_esophagitis + Asthma + Initial_days +
    TotalCharge + Additional_charges, family = binomial, data = medical_encoded)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.556e+01  3.973e+00 -19.018  < 2e-16 ***
Age                -1.398e-02  6.625e-03  -2.110  0.03486 *
Stroke              1.497e+00  2.449e-01   6.111 9.88e-10 ***
Arthritis          -1.323e+00  2.102e-01  -6.293 3.11e-10 ***
Anxiety            -1.086e+00  2.086e-01  -5.206 1.93e-07 ***
Allergic_rhinitis  -4.823e-01  1.934e-01  -2.494  0.01263 *
Reflux_esophagitis -4.738e-01  1.950e-01  -2.430  0.01511 *
Asthma             -1.139e+00  2.124e-01  -5.360 8.31e-08 ***
Initial_days        1.054e+00  6.064e-02  17.385  < 2e-16 ***
TotalCharge         2.815e-03  3.152e-04   8.928  < 2e-16 ***
Additional_charges  5.976e-05  2.144e-05   2.787  0.00531 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13145.70  on 9999  degrees of freedom
Residual deviance:   760.89  on 9989  degrees of freedom
AIC: 782.89

Number of Fisher Scoring iterations: 12
```

Part E1

The initial and reduced regression models were evaluated using the AIC value. The AIC for the initial model is 799.8, while the AIC for the reduced model is 782.89. The AIC for the reduced model is lower than the AIC for the initial model, implying that the reduced model is a better fit for the data.

Part E2

```
> # Use model to predict probability of readmission
> predicted <- as.numeric (predict(logres_final, medical_encoded, type="respons
e"))
> predicted <-ifelse(predicted > 0.5,1,0)
> predicted <- as.factor(predicted)
> str(predicted)
 Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
> predicted
   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [36] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [71] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [106] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [141] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [176] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [211] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [246] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [281] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [316] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [351] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [386] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [421] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [456] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [491] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [526] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [561] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [596] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [631] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [666] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [701] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [736] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [771] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [806] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [841] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [876] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [911] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [946] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [981] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [ reached getOption("max.print") -- omitted 9000 entries ]
Levels: 0 1
```

```
> unique(predicted)
[1] 0 1
Levels: 0 1
```

```
> # Convert values from "Yes" and "No" to 1's and 0's
> medical_encoded$ReAdmis <- as.factor(medical_encoded$ReAdmis)
> medical_encoded$ReAdmis
   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [36] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [71] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [106] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [141] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [176] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [211] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [246] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [281] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [316] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [351] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [386] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [421] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [456] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [491] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [526] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [561] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [596] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [631] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [666] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [701] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [736] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [771] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [806] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [841] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [876] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [911] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [946] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [981] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [ reached getOption("max.print") -- omitted 9000 entries ]
Levels: 0 1

> unique(medical_encoded$ReAdmis)
[1] 0 1
Levels: 0 1
```

```
> # Create confusion matrix
> confusionMatrix(medical_encoded$ReAdmis,predicted)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6250   81
         1   75 3594

               Accuracy : 0.9844
                 95% CI : (0.9818, 0.9867)
    No Information Rate : 0.6325
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9664

 Mcnemar's Test P-Value : 0.6889

            Sensitivity : 0.9881
            Specificity : 0.9780
         Pos Pred Value : 0.9872
         Neg Pred Value : 0.9796
             Prevalence : 0.6325
         Detection Rate : 0.6250
   Detection Prevalence : 0.6331
      Balanced Accuracy : 0.9831

       'Positive' Class : 0
```

Part E3

See attached code.

Part F1

The regression equation for the reduced model is:

ln(p^/(1-p^)) = (-7.556e+01 – 1.398e-02(Age) + 1.497e+00(Stroke) – 1.323e+00(Arthritis) – 1.086e+00(Anxiety) – 4.823e-01(Allergic_rhinits) – 4.738e-01(Reflux_esophagitis) – 1.139e+00(Asthma) + 1.054e+00(Initial_days) + 2.815e-03(TotalCharge) + 5.976e-05(Additional_charges)

Interpretation of the coefficients is detailed in the following table.

Keeping all things constant,

A one unit increase in	*changes the log odds of ReAdmis by*
Age	–1.398e-02
Stroke	+1.497e+00
Arthritis	–1.323e+00
Anxiety	–1.086e+00
Allergic_rhinitis	–4.823e-01
Reflux_esophagitis	–4.738e-01
Asthma	–1.139e+00
Initial_days	+1.054e+00
TotalCharge	+2.815e-03
Additional_charges	+5.976e-05

In terms of statistical significance, Initial_days and TotalCharge are the most significant variables (having p-values less than 2e-16), followed by Stroke, Arthritis, Anxiety, and Asthma. Allergic_rhinitis, Reflux_esophagitis, and Additional_charges are also statistically significant but less so than the previously mentioned variables.

In terms of practical significance, considering the magnitude of the coefficient for each variable, those variables with the larger absolute value coefficients (Additional_charges, Allergic_rhinitis, and Reflux_esophagitis) indicate a stronger effect on the log-odds of the patient being readmitted within a month of release.

The data analysis is limited by the initial selection of explanatory variables. Variables that were not included could have produced a more accurate model.

Part F2

Based on these results, my recommendation is to use this model to predict the likelihood of a patient being readmitted within a month of release. This information can guide treatment or intervention decisions.

Part G

The demonstration can be viewed at https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4304a653-99cc-40ca-b905-b09e0150a5f2

Part H

Paula. (2020b). Tips for analyzing categorical data in Excel. The Excel Club. https://theexcelclub.com/tips-for-analyzing-categorical-data-in-excel/

Zach. (2021). How to convert categorical variables to numeric in R. Statology. https://www.statology.org/convert-categorical-variable-to-numeric-r/

Zach. (2021a). How to perform One-Hot encoding in R. Statology. https://www.statology.org/one-hot-encoding-in-r/

Part I

None used.