D208: Predictive Modeling

Performance Assessment

Task 1

Logan Rosemeyer

Western Governor's University

D208: Predictive Modeling

Dr. Eric Straw

January 2, 2023

Contents

A1: Research Question

My research question is: What factors influence a patient's income?

A2: Objectives and Goals

The goal of the data is to see what factors influence a patient's income. This will allow us to see if different diseases affect different levels of income, or even if we treat patients differently based on their income level.

B1: Summary of Assumptions

There are a few assumptions of a multiple regression model. The first is that there is a linear relationship between the dependent variable and independent variables. The next is that the independent variables are not correlated with each other. This would create multicollinearity. Residuals should also be normally distributed with a mean of 0.

B2: Tool Benefits

I chose to use R for all phases of my analysis. The reason for this is because many function I will need to use are already built into R as opposed to Python. I had to run a linear regression, and R already has a stats package with the lm function loaded into it. Another reason that I used R is because it creates better graphics and data visualization. Although this task is about multiple linear regression, I had to make a numerous amount of univariate and bivariate visualizations. This is much easier in R rather than Python.

B3: Appropriate Technique

Multiple regression is the appropriate technique to analyze a patient's income. Income is a continuous variable, so multiple linear regression is the appropriate technique.

C1: Data Goals

My first goal is to make a data frame of variables that I might use. This includes most of the variables except the last eight (survey responses), unique IDs, and some location information. After I did this, my second goal was to make sure my data would give me accurate information. I looked for duplicates and null values. After that, I looked for any outliers in the data that I would need to treat. My final goal was to re-express categorical variables to be able to use them in my model.

C2: Summary Statistics

The population variable had a median of 2769 and a mean of 9965.2, which is skewed right. The area attributes are uniformly distributed, with all values appearing around 3350 times. The children variable had a median of 1 and a mean of 2, which is skewed right. The income variable had a median of 33768 and a mean of 40490, which is skewed right. The marital variable is uniformly distributed, with all values appearing around 2000 times. The gender variable has a similar amount of female and male values, and around 200 nonbinary. There are 6331 patients that were no readmitted within a month of first admission. 3669 had been readmitted within a month. Vitamin D levels had a median of 17.95 and mean of 17.96, with a normal distribution. Doctor visits had a median and mean of 5, with a normal

distribution. Full meals eaten had a mean and median of 1, with a skewed right distribution. Vitamin D supplements had a median of 0 and a mean of 0.4, with a skewed right distribution. 7425 patients did not drink soft drinks 3 times a day, while 2575 did. There were 5060 emergency admissions, 2504 elective emissions, and 2436 observation admissions. 5910 patients did not have high blood pressure, and 4090 did. 8007 patients did not have a stroke, while 1993 did. 2125 patients had a low complication risk, 4517 had a medium complication risk, and 3358 had a high complication risk. 7094 patients were overweight, and 2906 were not. 6426 patients did not have arthritis, while 3574 did. 7262 patients did not have diabetes, and 2738 did. 6628 patients did not have hyperlipidemia, and 3372 did. 5886 patients had back pain, while 4114 did. 6785 patients did not have anxiety, while 3215 did. 6059 patients did not have allergic rhinitis, while 3941 did. 5865 patients did not have reflux esophagitis, and 4135 did. 7107 patients did not have asthma, and 2893 did. The median initial days hospitalized was 35.8 with a mean of 34.5 and was bimodally distributed. The total charge median was 5214 with a mean of 5312 with a bimodal distribution. The additional charges had a median of 11574 and a mean of 12935 with a slightly skewed right distribution.

C3: Steps to Prepare the Data

The first thing I did was to narrow down the columns I might use for the data.

```
mlr_medical_data <- medical %>%
      select(-CaseOrder:-Lng, -Item1:-Item8, -TimeZone, -Job, -
      Services)
```

After this I checked for duplicates and null values.

```
sum(duplicated(mlr_medical_data))

colSums(is.na(mlr_medical_data))
```

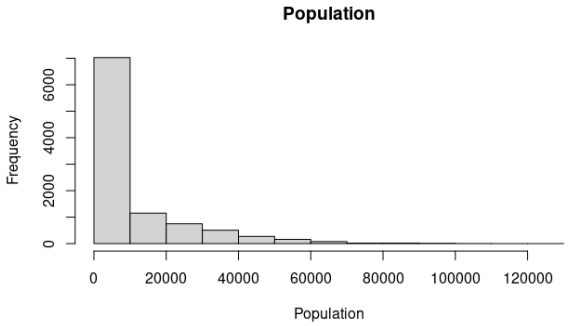
Then I checked for outliers of numeric values.

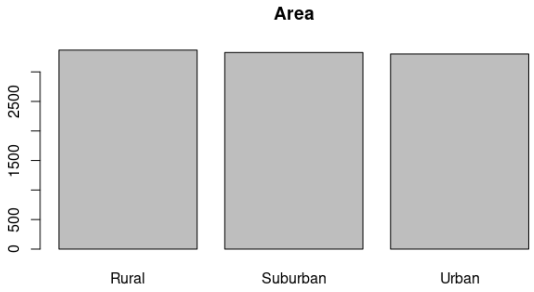
```
mlr_medical_data_zscore <- scale(mlr_medical_data %>%
select(Population, Children, Age, Income, VitD_levels,
Doc_visits, Full_meals_eaten, vitD_supp, Initial_days,
TotalCharge, Additional_charges))


head(mlr_medical_data_zscore)


mlr_medical_data_zscore <- as.data.frame(mlr_medical_data_zscore)

hist.data.frame(mlr_medical_data_zscore)
```

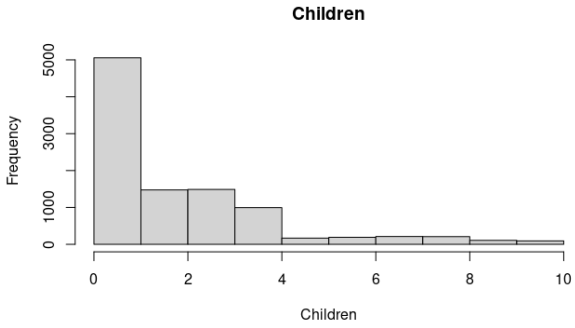
The last thing I did was to create dummy columns for the categorical variables.

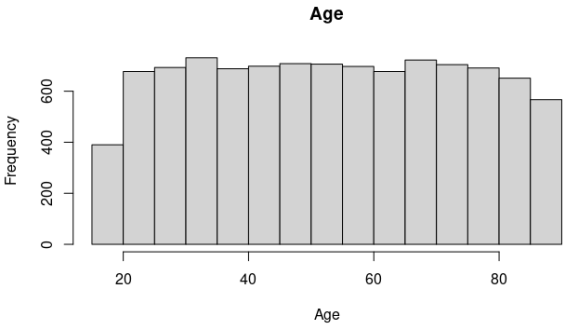
```
dummy_cols(mlr_medical_data, remove_first_dummy = T)
```

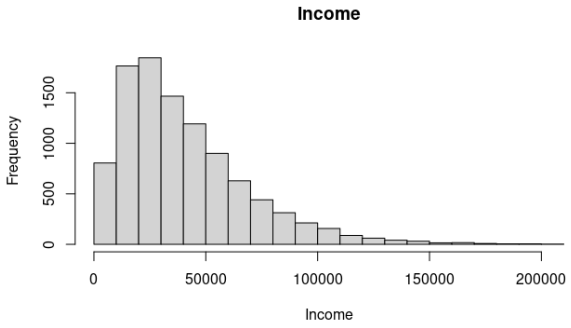
C4: Visualizations

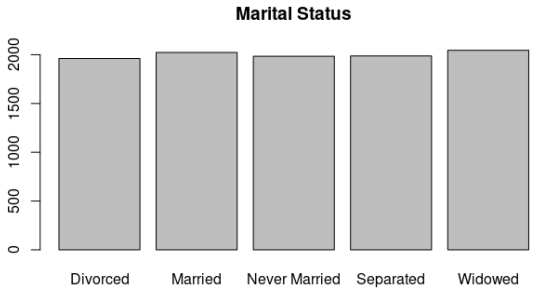


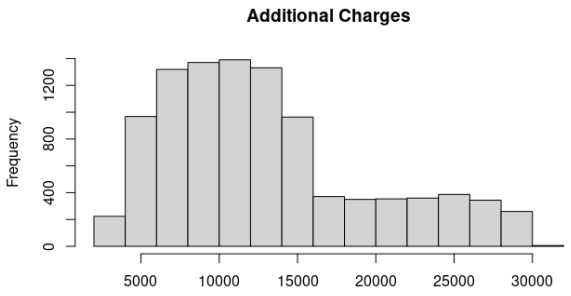


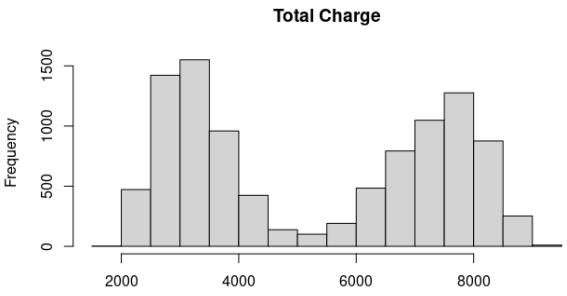


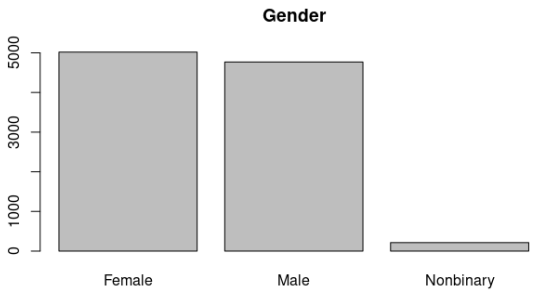


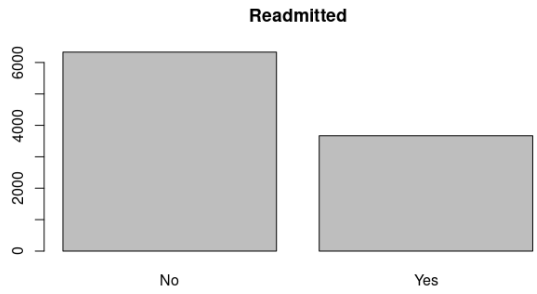


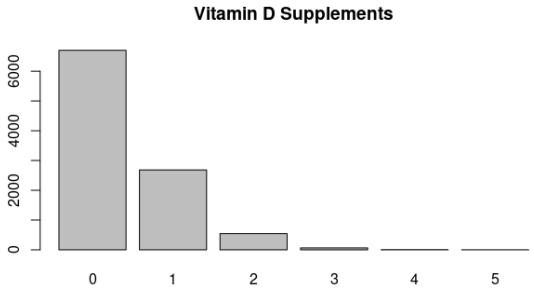


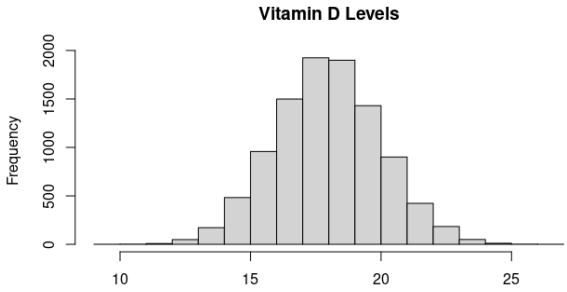


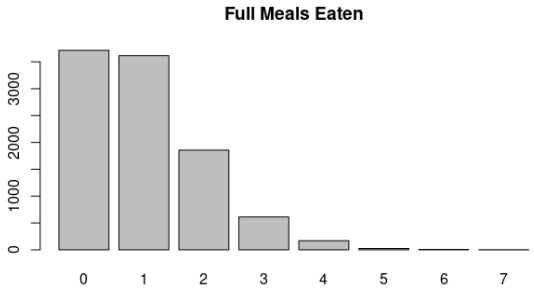


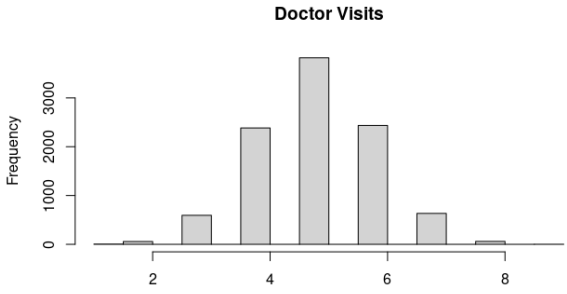

Gender
0
1000
3000
5000
Female
Male
Nonbinary

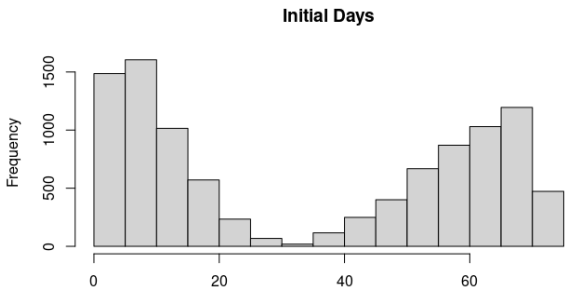

Readmitted
0
2000
4000
6000
No
Yes

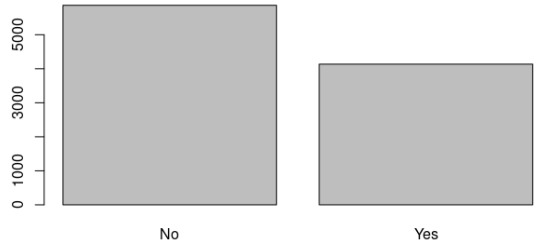

Vitamin D Supplements
0
2000
4000
6000
0
1
2
3
4
5

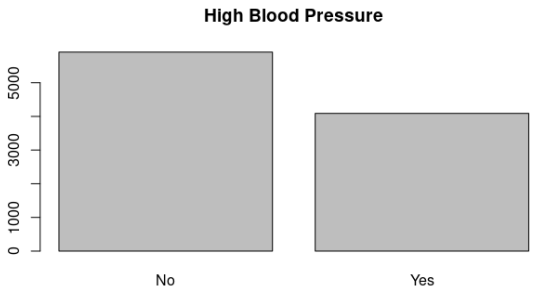

Vitamin D Levels
Frequency
0
500
1000
1500
2000
10
15
20
25


Full Meals Eaten
0
1000
2000
3000
0
1
2
3
4
5
6
7

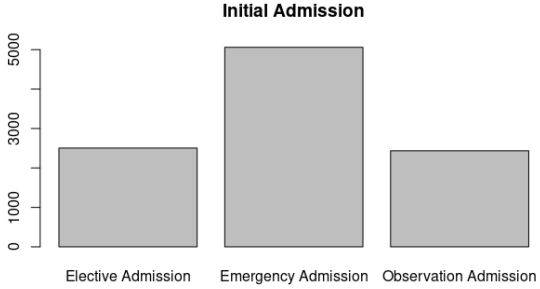

Doctor Visits
Frequency
0
1000
2000
3000
2
4
6
8

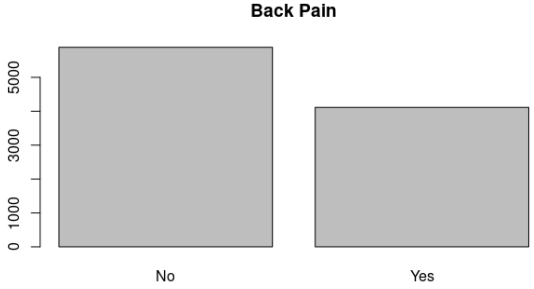

Initial Days
Frequency
0
500
1000
1500
0
20
40
60

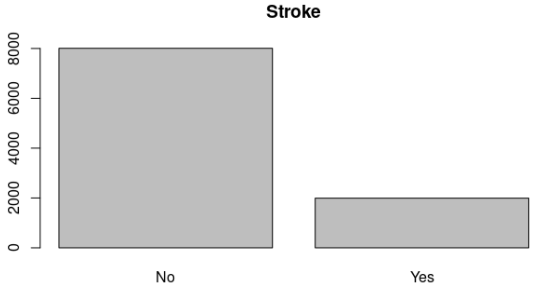

Reflux Esophagitis
0
1000
3000
5000
No
Yes

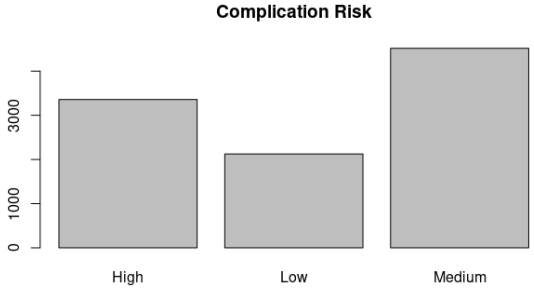

High Blood Pressure
0
1000
3000
5000
No
Yes

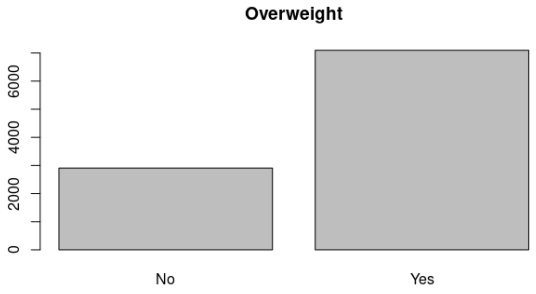

Soft Drink
0
2000
4000
6000
No
Yes

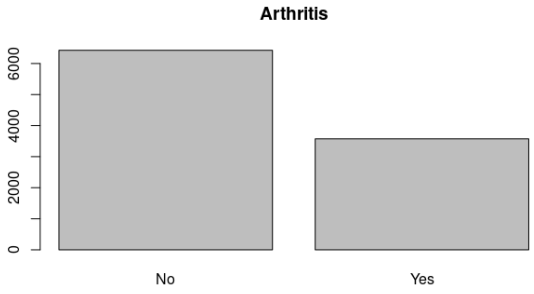

Initial Admission
0
1000
3000
5000
Elective Admission
Emergency Admission
Observation Admission

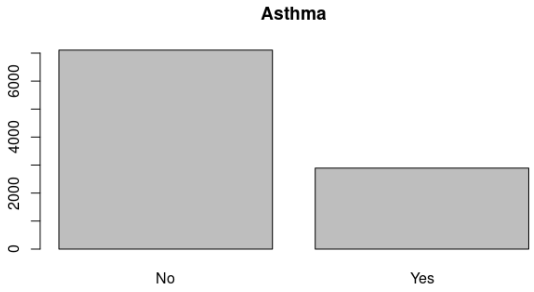

Back Pain
0
1000
3000
5000
No
Yes

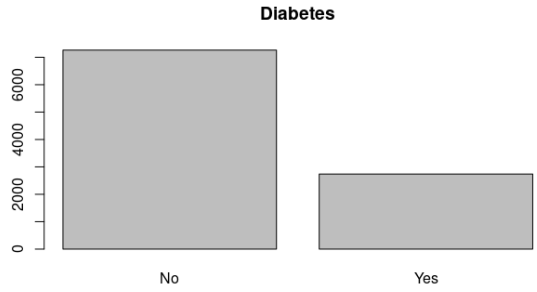

Stroke
0
2000
4000
6000
8000
No
Yes

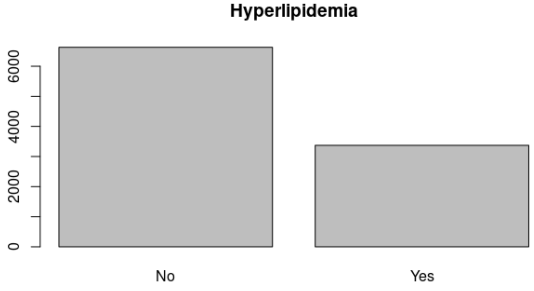

Complication Risk
0
1000
3000
High
Low
Medium


Overweight
0
2000
4000
6000
No
Yes

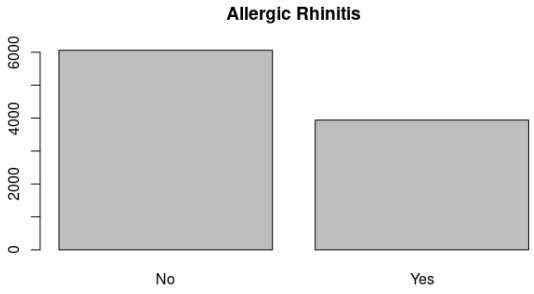

Arthritis
0
2000
4000
6000
No
Yes

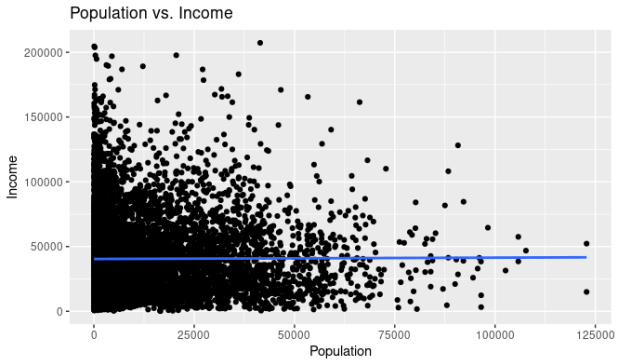

Asthma
0
2000
4000
6000
No
Yes

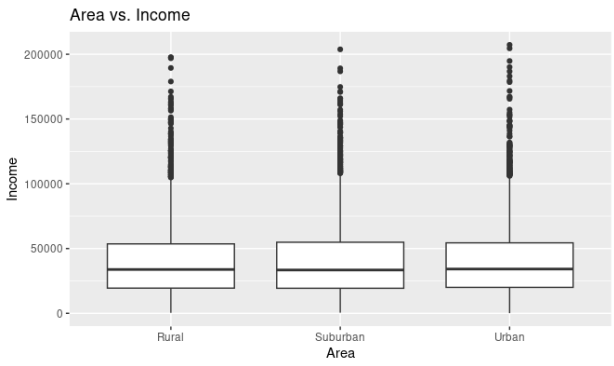

Diabetes
0
2000
4000
6000
No
Yes

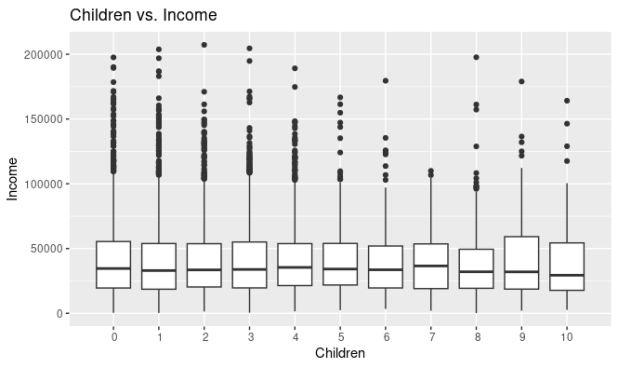

Hyperlipidemia
0
2000
4000
6000
No
Yes

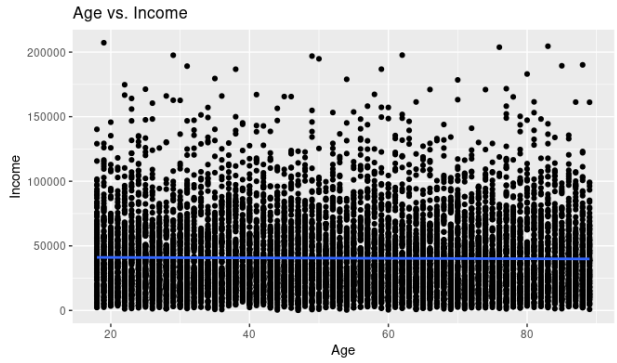

Anxiety
0
2000
4000
6000
No
Yes

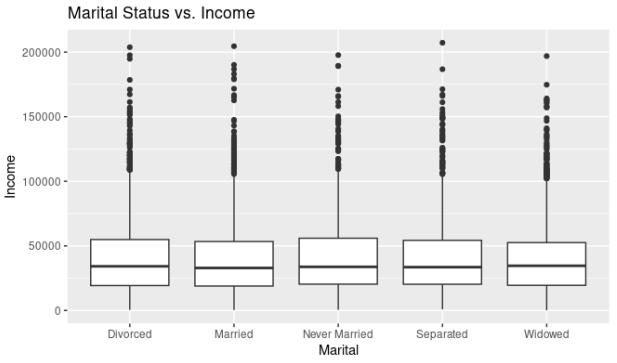

Allergic Rhinitis
0
2000
4000
6000
No
Yes

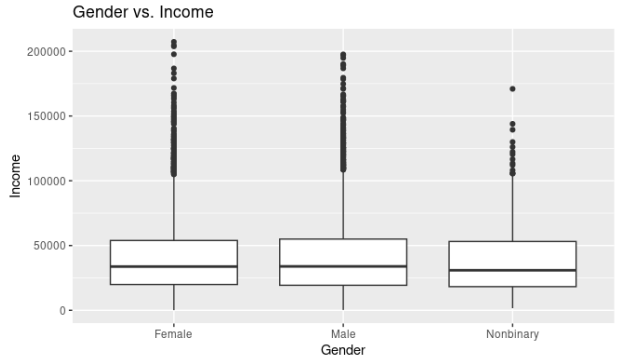

Population vs. Income
Income
0
50000
100000
150000
200000
0
25000
50000
75000
100000
125000
Population

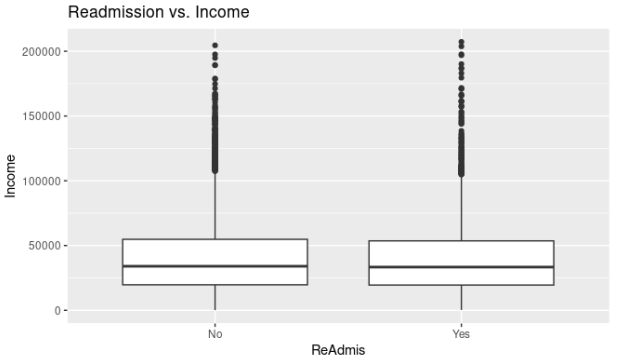

Area vs. Income
Income
0
50000
100000
150000
200000
Rural
Suburban
Urban
Area

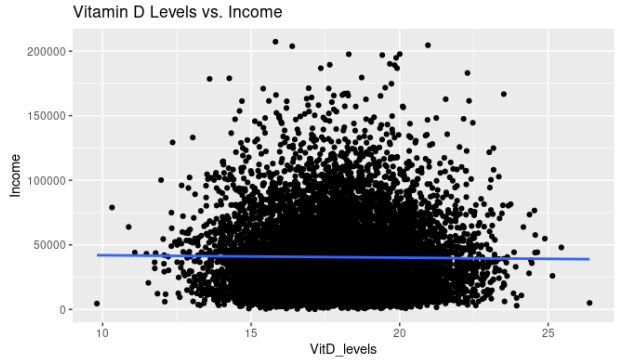

Children vs. Income
Income
0
50000
100000
150000
200000
0
1
2
3
4
5
6
7
8
9
10
Children

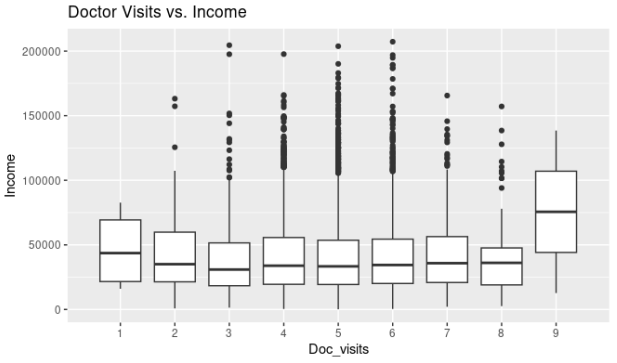

Age vs. Income
Income
200000
150000
100000
50000
0
20
40
60
80
Age

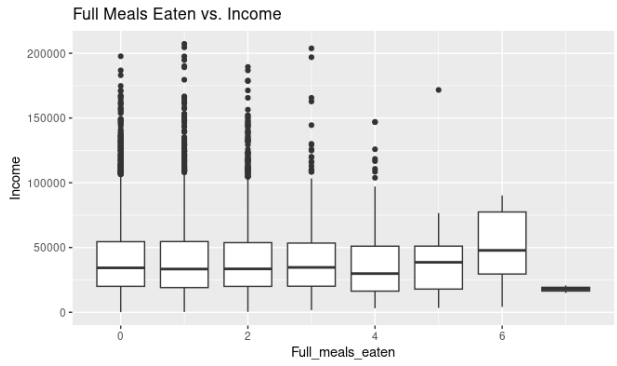

Marital Status vs. Income
Income
200000
150000
100000
50000
0
Divorced
Married
Never Married
Separated
Widowed
Marital

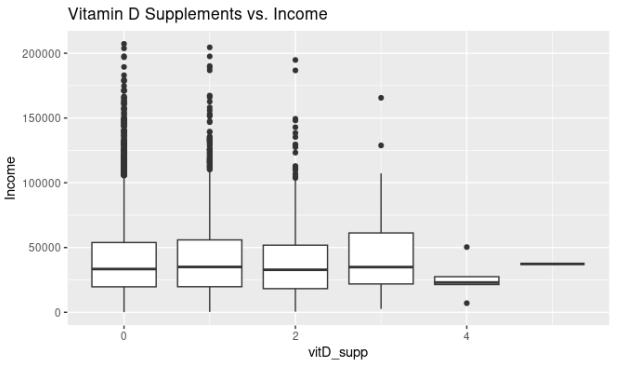

Gender vs. Income
Income
200000
150000
100000
50000
0
Female
Male
Nonbinary
Gender

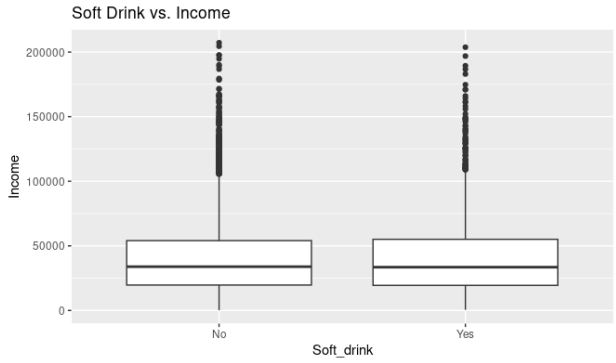

Readmission vs. Income
Income
200000
150000
100000
50000
0
No
Yes
ReAdmis

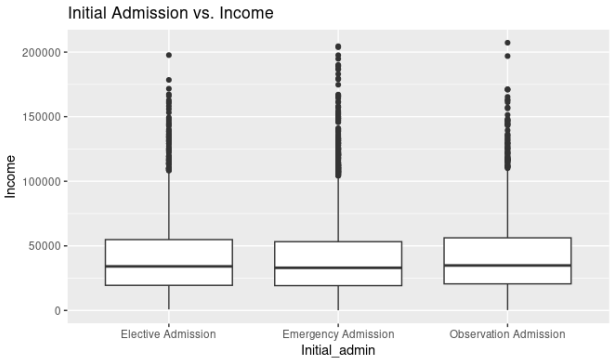

Vitamin D Levels vs. Income
Income
200000
150000
100000
50000
0
10
15
20
25
VitD_levels

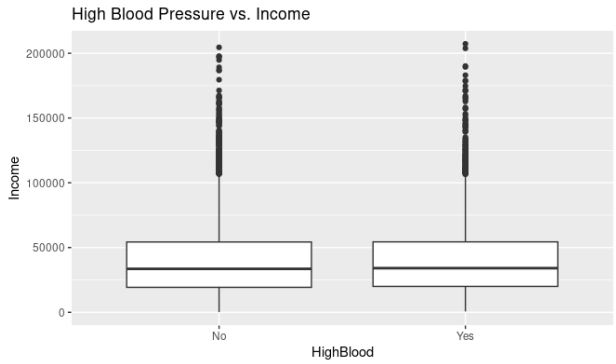

Doctor Visits vs. Income
Income
200000
150000
100000
50000
0
1
2
3
4
5
6
7
8
9
Doc_visits

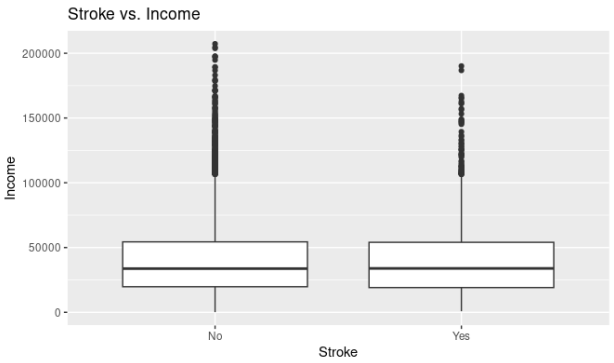

Full Meals Eaten vs. Income
Income
200000
150000
100000
50000
0
0
2
4
6
Full_meals_eaten

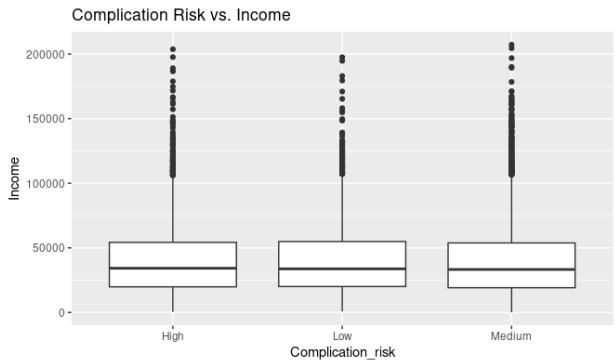

Vitamin D Supplements vs. Income
Income
200000
150000
100000
50000
0
0
2
4
vitD_supp

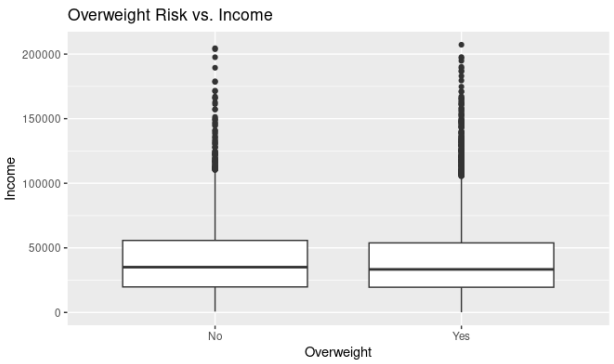

Soft Drink vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Soft_drink

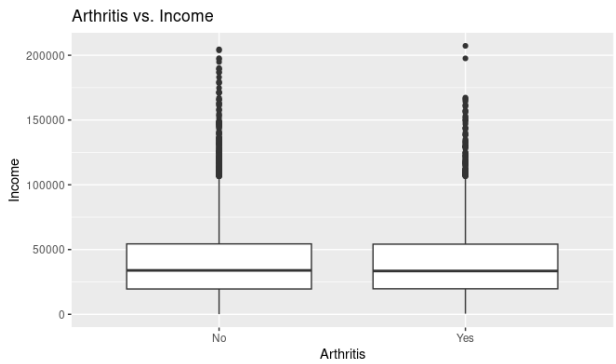

Initial Admission vs. Income
Income
200000
150000
100000
50000
0
Elective Admission
Emergency Admission
Observation Admission
Initial_admin

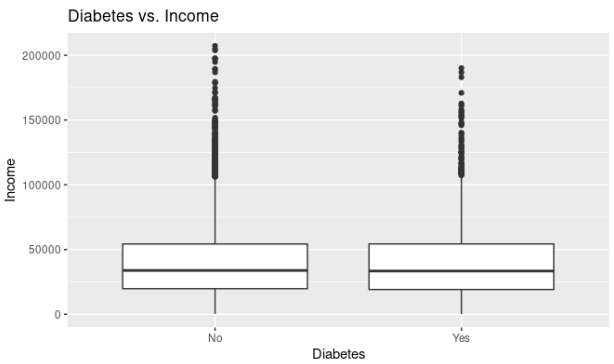

High Blood Pressure vs. Income
Income
200000
150000
100000
50000
0
No
Yes
HighBlood

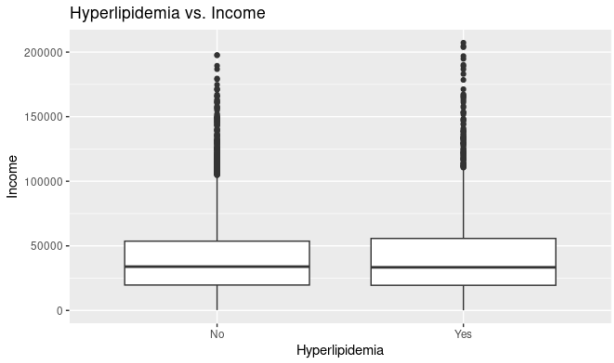

Stroke vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Stroke

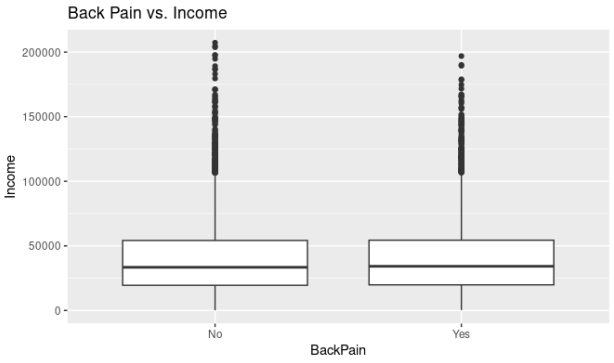

Complication Risk vs. Income
Income
200000
150000
100000
50000
0
High
Low
Medium
Complication_risk

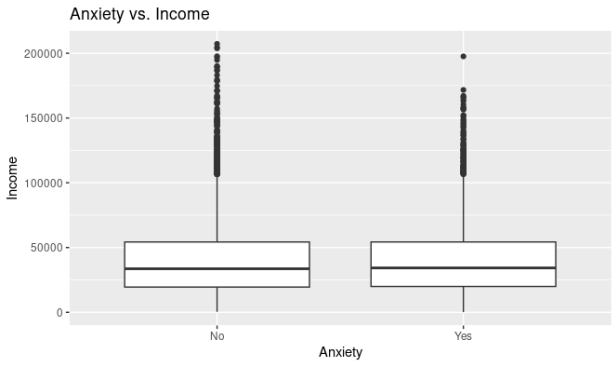

Overweight Risk vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Overweight

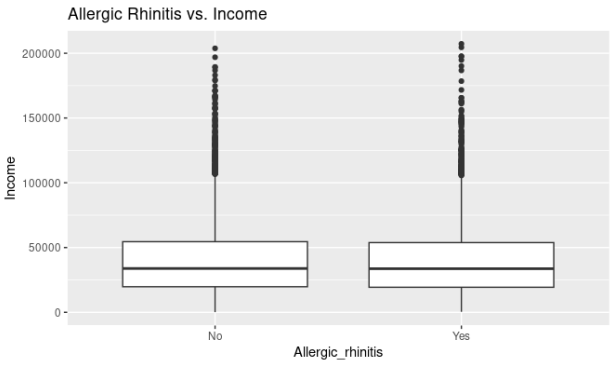

Arthritis vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Arthritis

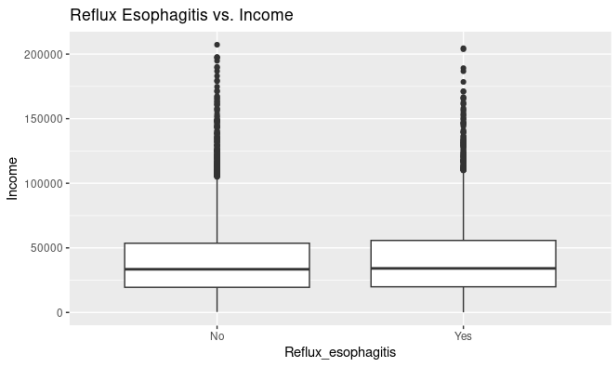

Diabetes vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Diabetes

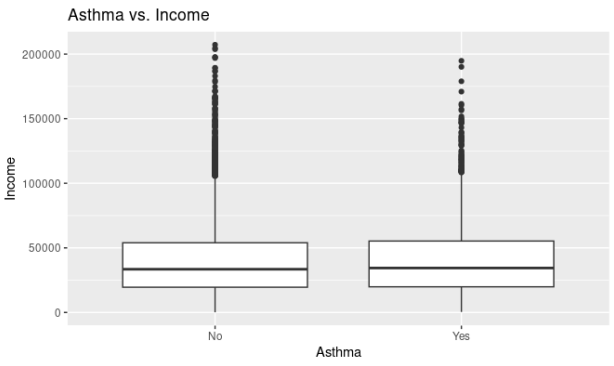

Hyperlipidemia vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Hyperlipidemia

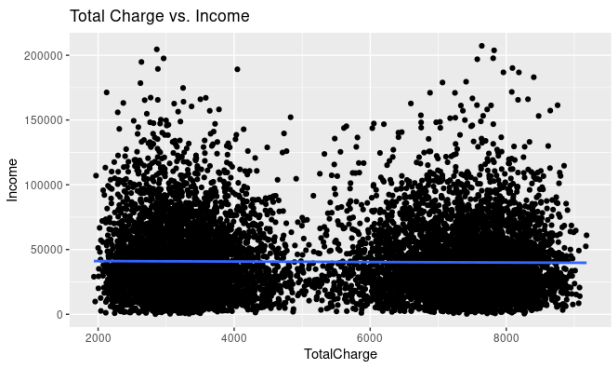

Back Pain vs. Income
Income
200000
150000
100000
50000
0
No
Yes
BackPain

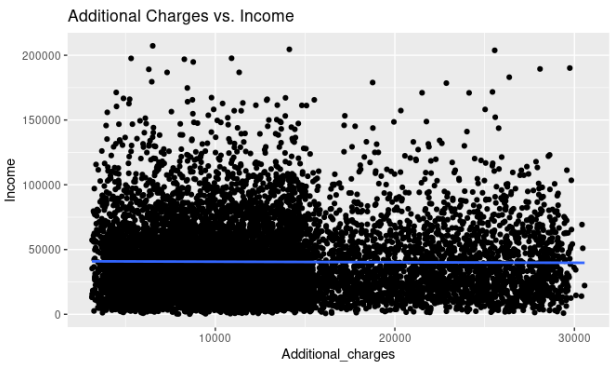

Anxiety vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Anxiety

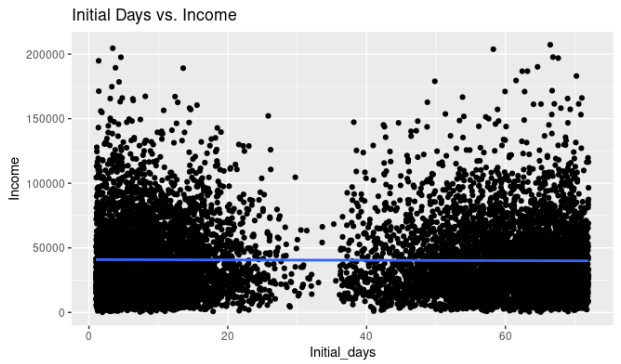

Allergic Rhinitis vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Allergic_rhinitis


Reflux Esophagitis vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Reflux_esophagitis


Asthma vs. Income
Income
200000
150000
100000
50000
0
No
Yes
Asthma


Total Charge vs. Income
Income
200000
150000
100000
50000
0
2000
4000
6000
8000
TotalCharge


Additional Charges vs. Income
Income
200000
150000
100000
50000
0
10000
20000
30000
Additional_charges



C5: Prepared Data Set

Prepared data set is attached in the submission.

D1: Initial Model

My initial model had a target variable of income and explanatory variables of population, area, children, age, marital, gender, readmission, vitamin D levels, doctor visits, full meals eaten, vitamin D supplements, soft drinks, initial admission reason, high blood pressure, stroke, complication risk, overweight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, reflux esophagitis, asthma, total charge, and additional charges. I was initially going to include initial days in my model, but it is so highly correlated to total charge that it created multicollinearity issues to the point that I couldn't even look at the vif of the model. That is why initial days is included in C2 but has not been included in my initial model.

```
income_mlr <- lm(Income ~ ., data=mlr_medical_data)

summary(income_mlr)

vif(income_mlr)


income_mlr <- lm(Income ~ . -Initial_days, data=mlr_medical_data)
```

```
Call:
lm(formula = Income ~ . - Initial_days, data = mlr_medical_data)

Residuals:
   Min     1Q Median     3Q    Max
-42912 -20602  -6572  13831 165843

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.417e+04  3.434e+03  12.862   <2e-16 ***
Population                       8.863e-03  1.928e-02   0.460   0.6458
AreaSuburban                     5.491e+02  6.982e+02   0.787   0.4316
AreaUrban                        7.125e+02  7.001e+02   1.018   0.3089
Children                         1.053e+02  1.321e+02   0.797   0.4253
Age                             -2.509e+01  4.210e+01  -0.596   0.5513
MaritalMarried                  -1.368e+03  9.058e+02  -1.510   0.1310
MaritalNever Married            -5.276e+00  9.102e+02  -0.006   0.9954
MaritalSeparated                -9.662e+02  9.091e+02  -1.063   0.2879
MaritalWidowed                  -1.195e+03  9.031e+02  -1.323   0.1858
GenderMale                       1.995e+02  5.785e+02   0.345   0.7302
GenderNonbinary                  4.440e+02  1.994e+03   0.223   0.8238
ReAdmisYes                      -2.745e+01  1.130e+03  -0.024   0.9806
VitD_levels                     -1.745e+02  1.418e+02  -1.231   0.2185
Doc_visits                       3.802e+02  2.733e+02   1.391   0.1642
Full_meals_eaten                -3.166e+02  2.836e+02  -1.116   0.2643
vitD_supp                        6.183e+01  4.547e+02   0.136   0.8919
Soft_drinkYes                    3.005e+02  6.536e+02   0.460   0.6457
Initial_adminEmergency Admission -7.212e+02  7.102e+02  -1.016   0.3098
Initial_adminObservation Admission 9.449e+02  8.139e+02   1.161   0.2457
HighBloodYes                    -3.779e+02  1.628e+03  -0.232   0.8164
StrokeYes                        1.343e+02  7.175e+02   0.187   0.8515
Complication_riskLow             1.530e+02  8.015e+02   0.191   0.8486
Complication_riskMedium         -4.001e+02  6.640e+02  -0.603   0.5468
OverweightYes                   -1.196e+03  6.292e+02  -1.900   0.0574 .
ArthritisYes                    -2.578e+02  5.974e+02  -0.432   0.6661
DiabetesYes                     -6.893e+02  6.414e+02  -1.075   0.2825
HyperlipidemiaYes                5.685e+02  6.048e+02   0.940   0.3472
BackPainYes                      5.784e+02  5.820e+02   0.994   0.3204
AnxietyYes                       9.189e+00  6.126e+02   0.015   0.9880
Allergic_rhinitisYes            -2.363e+01  5.852e+02  -0.040   0.9678
Reflux_esophagitisYes            9.216e+02  5.805e+02   1.588   0.1124
AsthmaYes                        4.279e+02  6.302e+02   0.679   0.4971
TotalCharge                     -1.710e-01  2.529e-01  -0.676   0.4989
Additional_charges               4.035e-02  1.761e-01   0.229   0.8188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28530 on 9965 degrees of freedom
Multiple R-squared:  0.003105,  Adjusted R-squared:  -0.0002961
F-statistic: 0.9129 on 34 and 9965 DF,  p-value: 0.6134
```

D2: Justification of Model Reduction

I decided to use the regsubsets function in the leaps library, along with P-values in the summary of the model, to reduce the variables used in the model. The regsubsets function performs best subset selection, using residual sum of squares to select which variables are best (Astaraky, n.d.). The variables selected by this function were then confirmed by having relatively small P-values. I used F-statistic and P-value to evaluate my model. My initial model had an F-statistic of less than 1, and a P-value of 0.61. This indicates that my initial model is not even close to being statistically significant. After I reduced the model to vitamin D levels, doctor visits, initial admission reason, overweight, reflux esophagitis, and total charge, my model had an F-statistic of 2.478 on 7 and 9992 degrees of freedom, and a P-value of 0.015. Both of these metrics tell me that my model is statistically significant. (Yashwanth, Oct 7, 2020)

```
income_mlr <- lm(Income ~ . -Initial_days, data=mlr_medical_data)

vif(income_mlr)

summary(income_mlr)

anova(income_mlr)

incomefit <- regsubsets(Income ~ . -Initial_days,
data=mlr_medical_data)
```

```
summary(incomefit)
```

```
reduced_income_mlr <- lm(Income ~ VitD_levels + Doc_visits +
Initial_admin + Overweight + Reflux_esophagitis + TotalCharge,
data=mlr_medical_data)
```

```
summary(reduced_income_mlr)
```

D3: Reduced Multiple Regression Model

My model includes the categorical variables initial admission, overweight, and reflux esophagitis. It also includes the continuous variables vitamin D levels, doctor visits, and total charge.

```
reduced_income_mlr <- lm(Income ~ VitD_levels + Doc_visits +
Initial_admin + Overweight + Reflux_esophagitis + TotalCharge,
data=mlr_medical_data)
```

```
summary(reduced_income_mlr)
```

```
Call:
lm(formula = Income ~ VitD_levels + Doc_visits + Initial_admin +
    Overweight + Reflux_esophagitis + TotalCharge, data = mlr_medical_data)

Residuals:
   Min     1Q Median     3Q    Max
-42521 -20738  -6647  13934 166093

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     43214.2932  3036.8173  14.230   <2e-16 ***
VitD_levels                      -174.2159   141.4042  -1.232   0.2180
Doc_visits                        372.7413   272.7160   1.367   0.1717
Initial_adminEmergency Admission -700.3165   699.2942  -1.001   0.3166
Initial_adminObservation Admission  928.1383   811.5036   1.144   0.2528
OverweightYes                   -1194.9474   628.0386  -1.903   0.0571 .
Reflux_esophagitisYes             962.1363   579.1962   1.661   0.0967 .
TotalCharge                        -0.1665     0.1316  -1.265   0.2058
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28510 on 9992 degrees of freedom
Multiple R-squared:  0.001733,  Adjusted R-squared:  0.001033
F-statistic: 2.478 on 7 and 9992 DF,  p-value: 0.01538
```

E1: Model Comparison

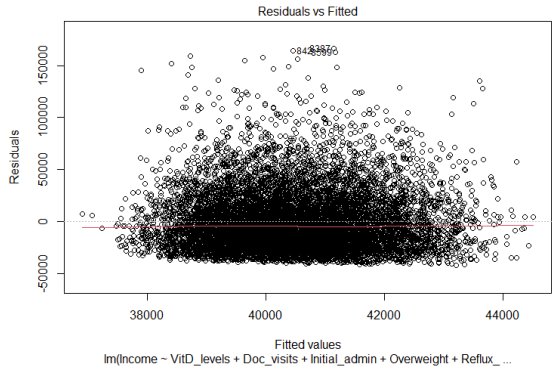
I decided to narrow my model down using regsubsets. Unfortunately, when I reduced my model, the residual standard error and R^2 values didn't really change. There was slight improvement, but not as much as I would've liked to see. The RSE improved from 28530 to 28510. The multiple R^2 went from .003 to .001. The adjusted R^2 went from -0.0003 to 0.001. There was improvement on the F-statistic and P-value. The F-statistic improved from 0.91 to 2.48. The P-value improved from 0.613 to 0.015.

E2: Output and Calculations

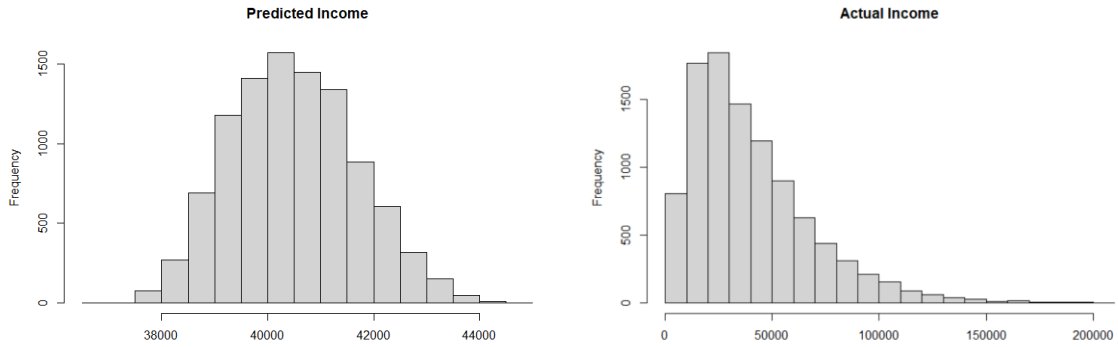
```
Call:
lm(formula = Income ~ . - Initial_days, data = mlr_medical_data)

Residuals:
   Min     1Q Median     3Q    Max
-42912 -20602  -6572  13831 165843

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.417e+04  3.434e+03  12.862   <2e-16 ***
Population                       8.863e-03  1.928e-02   0.460   0.6458
AreaSuburban                     5.491e+02  6.982e+02   0.787   0.4316
AreaUrban                        7.125e+02  7.001e+02   1.018   0.3089
Children                         1.053e+02  1.321e+02   0.797   0.4253
Age                             -2.509e+01  4.210e+01  -0.596   0.5513
MaritalMarried                  -1.368e+03  9.058e+02  -1.510   0.1310
MaritalNever Married            -5.276e+00  9.102e+02  -0.006   0.9954
MaritalSeparated                -9.662e+02  9.091e+02  -1.063   0.2879
MaritalWidowed                  -1.195e+03  9.031e+02  -1.323   0.1858
GenderMale                       1.995e+02  5.785e+02   0.345   0.7302
GenderNonbinary                  4.440e+02  1.994e+03   0.223   0.8238
ReAdmisYes                      -2.745e+01  1.130e+03  -0.024   0.9806
VitD_levels                     -1.745e+02  1.418e+02  -1.231   0.2185
Doc_visits                       3.802e+02  2.733e+02   1.391   0.1642
Full_meals_eaten                -3.166e+02  2.836e+02  -1.116   0.2643
vitD_supp                        6.183e+01  4.547e+02   0.136   0.8919
Soft_drinkYes                    3.005e+02  6.536e+02   0.460   0.6457
Initial_adminEmergency Admission -7.212e+02  7.102e+02  -1.016   0.3098
Initial_adminObservation Admission 9.449e+02  8.139e+02   1.161   0.2457
HighBloodYes                    -3.779e+02  1.628e+03  -0.232   0.8164
StrokeYes                        1.343e+02  7.175e+02   0.187   0.8515
Complication_riskLow             1.530e+02  8.015e+02   0.191   0.8486
Complication_riskMedium         -4.001e+02  6.640e+02  -0.603   0.5468
OverweightYes                   -1.196e+03  6.292e+02  -1.900   0.0574 .
ArthritisYes                    -2.578e+02  5.974e+02  -0.432   0.6661
DiabetesYes                     -6.893e+02  6.414e+02  -1.075   0.2825
HyperlipidemiaYes                5.685e+02  6.048e+02   0.940   0.3472
BackPainYes                      5.784e+02  5.820e+02   0.994   0.3204
AnxietyYes                       9.189e+00  6.126e+02   0.015   0.9880
Allergic_rhinitisYes            -2.363e+01  5.852e+02  -0.040   0.9678
Reflux_esophagitisYes            9.216e+02  5.805e+02   1.588   0.1124
AsthmaYes                        4.279e+02  6.302e+02   0.679   0.4971
TotalCharge                     -1.710e-01  2.529e-01  -0.676   0.4989
Additional_charges               4.035e-02  1.761e-01   0.229   0.8188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28530 on 9965 degrees of freedom
Multiple R-squared:  0.003105,  Adjusted R-squared:  -0.0002961
F-statistic: 0.9129 on 34 and 9965 DF,  p-value: 0.6134
```

```
Call:
lm(formula = Income ~ VitD_levels + Doc_visits + Initial_admin +
    Overweight + Reflux_esophagitis + TotalCharge, data = mlr_medical_data)

Residuals:
   Min     1Q Median     3Q    Max
-42521 -20738  -6647  13934 166093

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    43214.2932  3036.8173  14.230   <2e-16 ***
VitD_levels                     -174.2159   141.4042  -1.232   0.2180
Doc_visits                       372.7413   272.7160   1.367   0.1717
Initial_adminEmergency Admission -700.3165   699.2942  -1.001   0.3166
Initial_adminObservation Admission  928.1383   811.5036   1.144   0.2528
OverweightYes                  -1194.9474   628.0386  -1.903   0.0571 .
Reflux_esophagitisYes            962.1363   579.1962   1.661   0.0967 .
TotalCharge                       -0.1665     0.1316  -1.265   0.2058
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28510 on 9992 degrees of freedom
Multiple R-squared:  0.001733,  Adjusted R-squared:  0.001033
F-statistic: 2.478 on 7 and 9992 DF,  p-value: 0.01538
```

Residuals are scattered around the y=0 line. There are some values with higher residuals than I would like to see.



Here is a histogram of predicted income based on my model. The model predicts income centered around just over 40000 and no values over 50000. This is an issue because, based off of the actual income data, there are a significant amount of observations over 50000.



E3: Code

```
reduced_income_mlr <- lm(Income ~ VitD_levels + Doc_visits +
Initial_admin + Overweight + Reflux_esophagitis + TotalCharge,
data=mlr_medical_data)

summary(reduced_income_mlr)

anova(reduced_income_mlr)

hist(reduced_income_mlr$residuals)

plot(reduced_income_mlr)
```

F1: Results

My regressions equation is: $Income = -174.22(Vit\ D\ Levels) + 372.74(Doctor\ Visits) - 700.32(Emergency\ Initial\ Admission) + 928.14(Observation\ Initial\ Admission) - 1194.95(Overweight) + 962.14(Reflux\ Esophagitis) - 0.17(Total\ Charge) + 43214.29$

My model found that as vitamin D levels increase one unit, income generally decreases by $174.22. As doctor visits increase one unit, income generally increases by $372.74. If a patient had an emergency initial admission, their income generally decreased by $700.32. If a patient had an observational initial admission, their income generally increased by $928.14. If a patient is overweight, their income generally decreased by $1194.95. If a patient has reflux esophagitis, their income generally increased by $962.14.

I do not think this model is strong enough to accurately predict someone's actual income. None of the variables had a P-value below 0.05 so I can't say for sure that any are statistically significant.

F2: Recommendations

I think the next step for the organization would be to look into how income is related to being overweight, as well as how it relates to reflux esophagitis. In my final model, these variables were both significant at the 0.1 level. This tells me that there could be some relationship yet. While doing this, I would recommend is splitting income into groups. Predicting which group a patient falls into might be more accurate, therefore more useful, than trying to predict the patient's actual income.

G: Panopto Demonstration

Panopto video link is shown below. It is also attached in the submission.

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=56871750-7472-446a-98b9-af8400ebc3f8

H: Sources of Third-Party Code

Predict.lm: Predict method for Linear Model Fits. RDocumentation. Retrieved January 7, 2023. Rdocumentation.org/packages/stats/versions/3.6.2/topics/predict.lm

Schork, Joachim. *Plot Predicted vs. Actual Values in R (2 examples).* Statistics Globe. Retrieved January 7, 2023. Statisticsglobe.com/plot-predicted-vs-actual-values-in-r#example-1-draw-predicted-vs-observed-using-base-r.

Hist.data.frame: Histograms for Variables in a Data Frame. Rdrr.io. Retrieved January 7, 2023. https://rdrr.io/cran/Hmisc/man/hist.data.frame.html

I: Sources

Astaraky, Davood. *Linear Model Selection and Regularization – Subset Selection Methods.* RPubs by RStudio. Retrieved January 7, 2023. Rpubs.com/davoodastaraky/subset.

Yashwanth, NVS. *Evaluation metrics & Model Selection in Linear Regression.* Medium. October 7, 2020. Retrieved January 7, 2023. Towardsdatascience.com/evaluation-metrics-model-selection-in-linear-regression-73c7573208be