



JANUARY 5, 2023

PERFORMANCE ASSESSMENT

D208 TASK ONE

ROBERT PATTON
WESTERN GOVERNORS' UNIVERSITY
Rpatt33@wgu.edu




Table of Contents

Part I: Research Question	2
<i>A. Purpose of Data Analysis</i>	2
1. Research Question	2
2. Goals of Data Analysis	2
Part II: Method Justification	2
<i>B. Describe Multiple Linear Regression</i>	2
1. Four MLR Assumptions	2
2. Benefits of Python	2-3
3. MLR as a Tool	3
Part III: Data Preparation	3
<i>C. Summarize the Data Preparation Process</i>	3
1. Goals and Techniques	3-4
2. Summary Statistics	4-6
3. Univariate and Bivariate Visuals	6-9
Part IV: Model Comparison and Analysis	10
<i>D. Compare an Initial and Reduced Linear Regression Model</i>	10
1. Initial Model	10
2. Model Reduction Method/Justification	11-12
3. Reduced Model	13
<i>E. Model Comparison</i>	14
1. Initial and Reduced Model Comparison	14
2. Multiple Linear Regression	14-15
3. Executable Code	15
Part V: Data Summary and Implications	15
<i>F. Regression Equation, Coefficients, Etc</i>	15
1. Reduced Model Regression Equation	15
2. Interpretation of Coefficients	15
3. Statistical and Practical Significance of Reduced Model	15-16
4. Disadvantages of Methods Used/Recommendations	16-17
Part VI: Demonstration	17
<i>G. Panopto Video</i>	17
Sources	17

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:

1. Research Question:

Is there a correlation between patient income and overall health? That is, does income correlate to more or less health conditions?

2. Goals of Data Analysis:

The goals for this data analysis project are to determine if patient income is correlated to overall health. My overall goal is to determine whether or not someone who makes more money is healthier than someone who makes less money. My research could then further take me into determining why lower income would contribute to more medical conditions. For example, are there environmental factors a patient is exposed to, due to lower income, and having to live near waste plants instead of on a beach.

Part II: Method Justification

B. Describe multiple linear regression methods:

1. Four Multiple Linear Regression Assumptions:

- There is a linear relationship between a dependent variable and varying independent variables. In multiple linear regression, two or more independent variables are used to determine or predict an outcome of a dependent variable.
- Independent variables are not highly correlated, called multicollinearity. This is something an analyst attempts to prevent as multicollinearity leads to data skewedness and can impair results for determining prediction of a dependent variable.
- The more independent variables included in a model will improve the explanatory power for a predictive outcome.
- Residuals from the model should have a normal distribution. Residuals are the differences between the true value and predicted value of a regression model.

2. Benefits of Using Python:

- Python is a great program coding language that is user friendly. It is a common tool used by a majority of data analysts who deal with large

datasets, allowing them to work in an interactive development environment that makes analytic workflow processes easier to manage. I chose to use Python because it will be a program that I will use when I complete the MSDA program and begin my new career. The Python language is a great tool because it contains packages that allow a user to access specifically designed libraries for statistical analysis, visualization of data, and other analytic tools that can provide stakeholders with insight into a research question.

3. Multiple Linear Regression as a Tool:

- Multiple linear regression can be a very effective tool for a data analyst. Using multiple linear regression allows a data analyst to assess the strength of a relationship between a dependent variable and multiple independent variables, also called predictor variables. In addition, multiple linear regression can help us determine how significant each predictor variable is to a dependent variable, thus allowing an analyst to eliminate those that have little to no correlation.

Part III: Data Preparation

C. Summarize the data preparation process:

1. Data Preparation Goals and Techniques

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

Data preparation techniques are as follows:

- Import all necessary and required Python libraries for statistical information and data visualization.
- Read in our CSV file.
- Examine our data types, structure, and variables.
- Detect and eliminate any duplicates.
- Detect and address any null/missing values by using the replace method with mean, median, or mode.

- Rename any variables to improve the clarity of the data frame.
- Examine the variables as univariate and bivariate visualizations.
- Detect any outliers in the data variables.
- Drop any variables I feel are irrelevant to my research question.
- Transform categorical data into numerical data types with dummy codes.
- *See code attached with submission*

2. Summary Statistics

- The dataset originally consisted of 10000 rows and 50 columns. After cleaning the data and selecting the variables for linear regression, the dataset consisted of 10000 rows and 22 columns.
- For this performance assessment, my dependent variable is income, a float or continuous data type consisting of a numerical value with decimals. It is being used to determine whether or not a patient's income is correlated to their overall health.
- My independent variables, Lat, Lng, Children, age, VitD_levels, Doc_visits, vitD_supp, Initial_admin, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic_rhinitis, Reflux_esophagitis, Asthma, and Initial_days, are observations taken from the data that I feel may contribute to defining a patient's overall health. Lat, Lng, VitD_levels, and Initial_days are also float/continuous data types. Children, age, Doc_visits, and vitD_supp, are all integer data types, whole numbers. The rest of the variables are all object data types, meaning their values are organized into a category such as yes/no for medical conditions or admission level type for Initial_admin. These variables will be used to determine if there is a correlation between overall health and income.
- After preparing all the data as described in the steps above, I then obtained my summary of statistics for non-categorical/quantitative variable types. Reviewing dataset summary statistics is valuable to an analyst because it can provide information on a few things. It can help determine where the

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

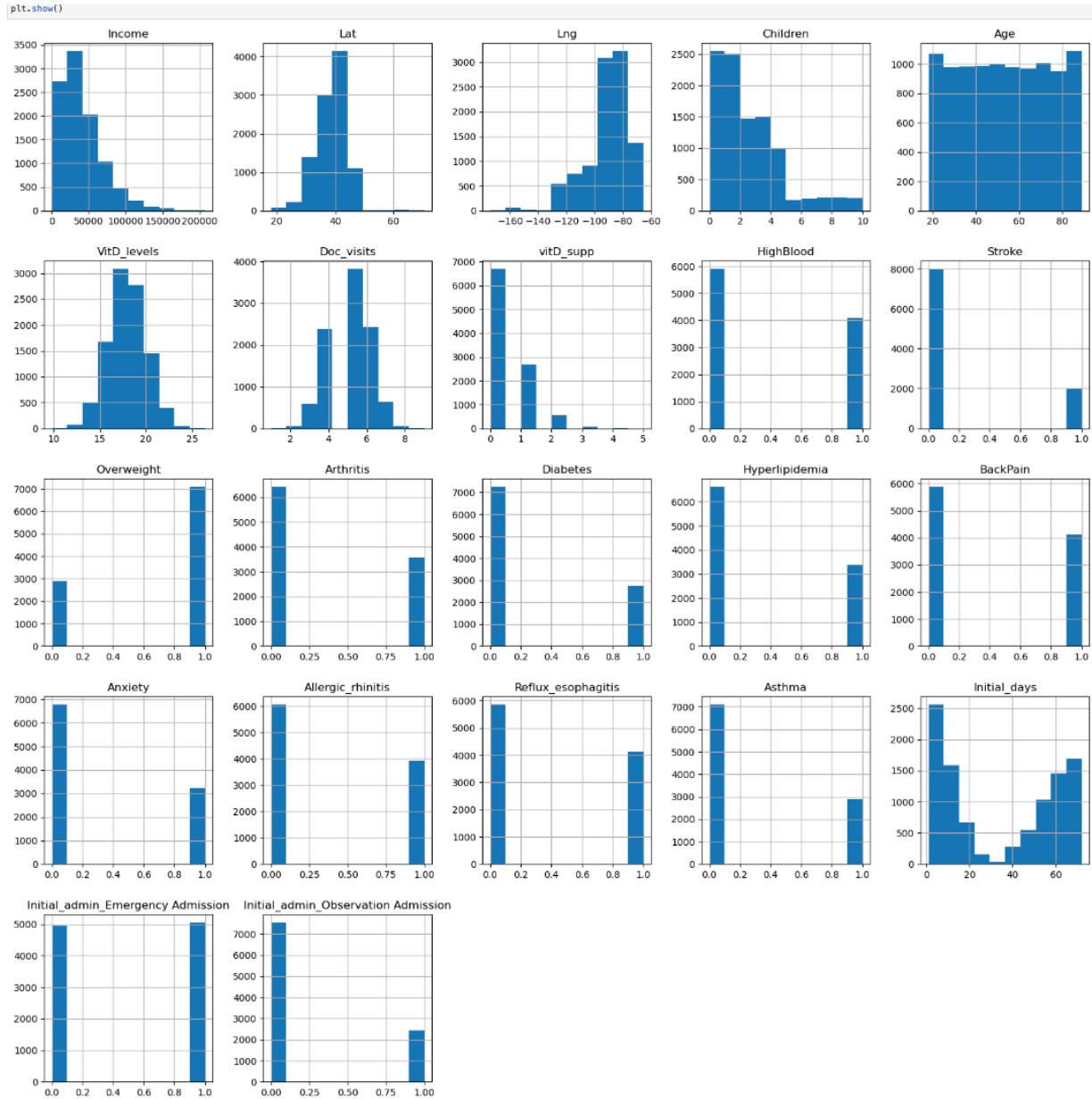
```
#Get summary statistics
print(mdf.describe())
```

	Income	Lat	Lng	Children	Age	\
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	
mean	40490.495160	38.751099	-91.243080	2.097200	53.511700	
std	28521.153293	5.403085	15.205998	2.163659	20.638538	
min	154.080000	17.967190	-174.209700	0.000000	18.000000	
25%	19598.775000	35.255120	-97.352982	0.000000	36.000000	
50%	33768.420000	39.419355	-88.397230	1.000000	53.000000	
75%	54296.402500	42.044175	-80.438050	3.000000	71.000000	
max	207249.100000	70.560990	-65.290170	10.000000	89.000000	
	VitD_levels	Doc_visits	vitD_supp	HighBlood	Stroke	\
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	
mean	17.964262	5.012200	0.398900	0.409000	0.199300	
std	2.017231	1.045734	0.628505	0.491674	0.399494	
min	9.806483	1.000000	0.000000	0.000000	0.000000	
25%	16.626439	4.000000	0.000000	0.000000	0.000000	
50%	17.951122	5.000000	0.000000	0.000000	0.000000	
75%	19.347963	6.000000	1.000000	1.000000	0.000000	
max	26.394449	9.000000	5.000000	1.000000	1.000000	
	Diabetes	Hyperlipidemia	BackPain	Anxiety	\	
count	10000.000000	10000.000000	10000.000000	10000.000000		
mean	0.27380	0.337200	0.411400	0.321500		
std	0.44593	0.472777	0.492112	0.467076		
min	0.00000	0.000000	0.000000	0.000000		
25%	0.00000	0.000000	0.000000	0.000000		
50%	0.00000	0.000000	0.000000	0.000000		
75%	1.00000	1.000000	1.000000	1.000000		
max	1.00000	1.000000	1.000000	1.000000		
	Allergic_rhinitis	Reflux_esophagitis	Asthma	Initial_days	\	
count	10000.000000	10000.000000	10000.000000	10000.000000		
mean	0.394100	0.413500	0.28930	34.455299		
std	0.488681	0.492486	0.45346	26.309341		
min	0.000000	0.000000	0.00000	1.001981		
25%	0.000000	0.000000	0.00000	7.896215		
50%	0.000000	0.000000	0.00000	35.836244		
75%	1.000000	1.000000	1.00000	61.161020		
max	1.000000	1.000000	1.00000	71.981490		
	Initial_admin_Emergency Admission	Initial_admin_Observation Admission				
count	10000.000000	10000.000000				
mean	0.506000					
std	0.499989					
min	0.000000					
25%	0.000000					
50%	1.000000					
75%	1.000000					
max	1.000000					

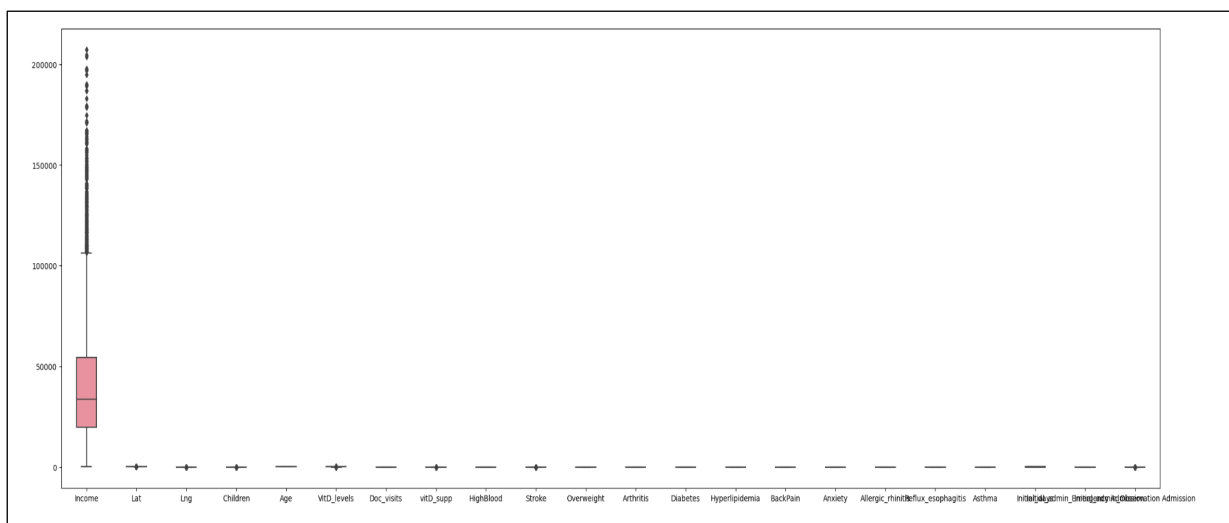
- Using the `.value_counts()` function, I can obtain the exact number of values for each response in the categorical/qualitative variables.
 1. Initial_admin has three values, emergency admission with 5060 responses, observation admission with 2436 responses, and elective admission with 2504 responses.
 2. HighBlood had 5910 no response and 4090 yes responses.
 3. Stroke had 8007 no responses and 1993 yes responses.
 4. Overweight had 2906 no responses and 7094 yes responses.
 5. Arthritis had 6426 no responses and 3574 yes responses.
 6. Diabetes had 7262 no responses and 2738 yes responses.
 7. Hyperlipidemia had 6628 no responses and 3372 yes responses.
 8. BackPain had 5886 no responses and 4114 yes responses.
 9. Anxiety had 6785 no responses and 3215 yes responses.
 10. Allergic_rhinitis had 6059 no responses and 3941 yes responses.
 11. Reflux_esophagitis had 5865 no responses and 4135 yes responses.
 12. Asthma had 7107 no responses and 2893 yes responses.

3. Generate Univariate and Bivariate Visualizations

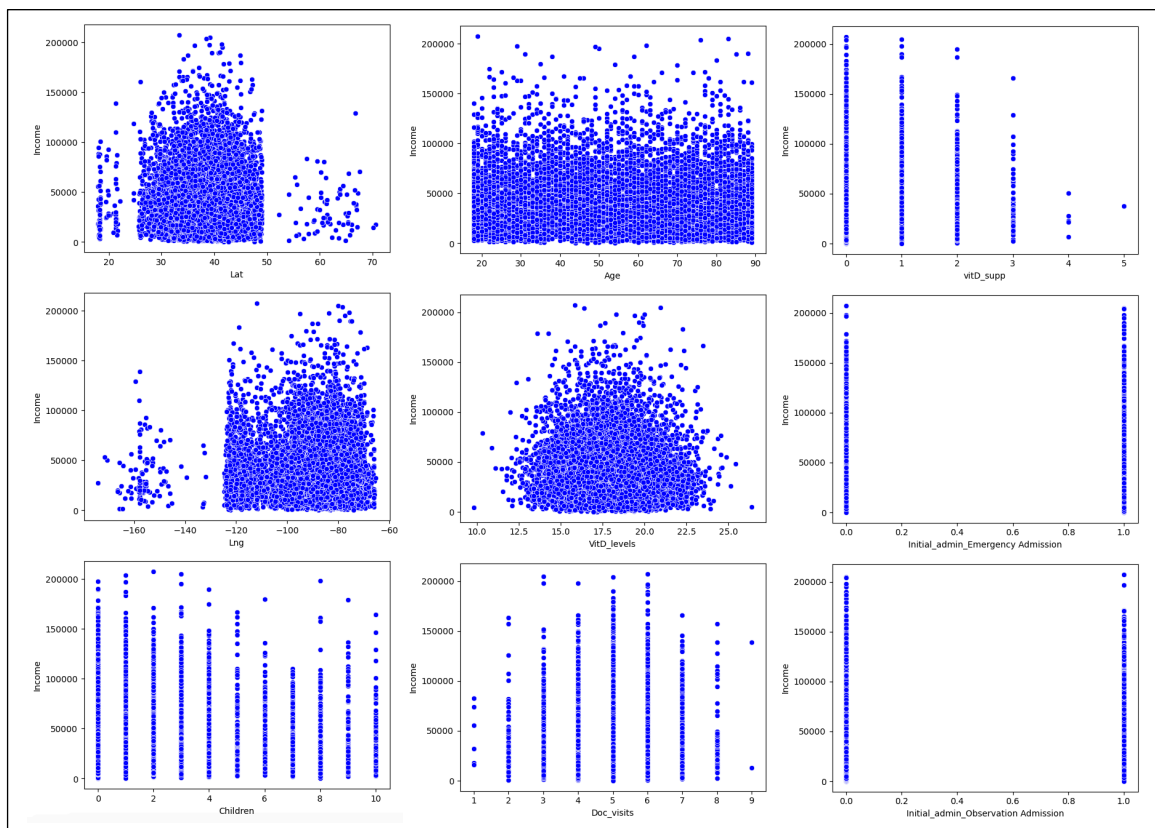
- Both univariate and bivariate visualizations can help an analyst determine which variables will be used for multiple linear regression. They enable an analyst to see the types of distributions for each data type and provide insight into how the dependent variable is compared to the potential predictor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

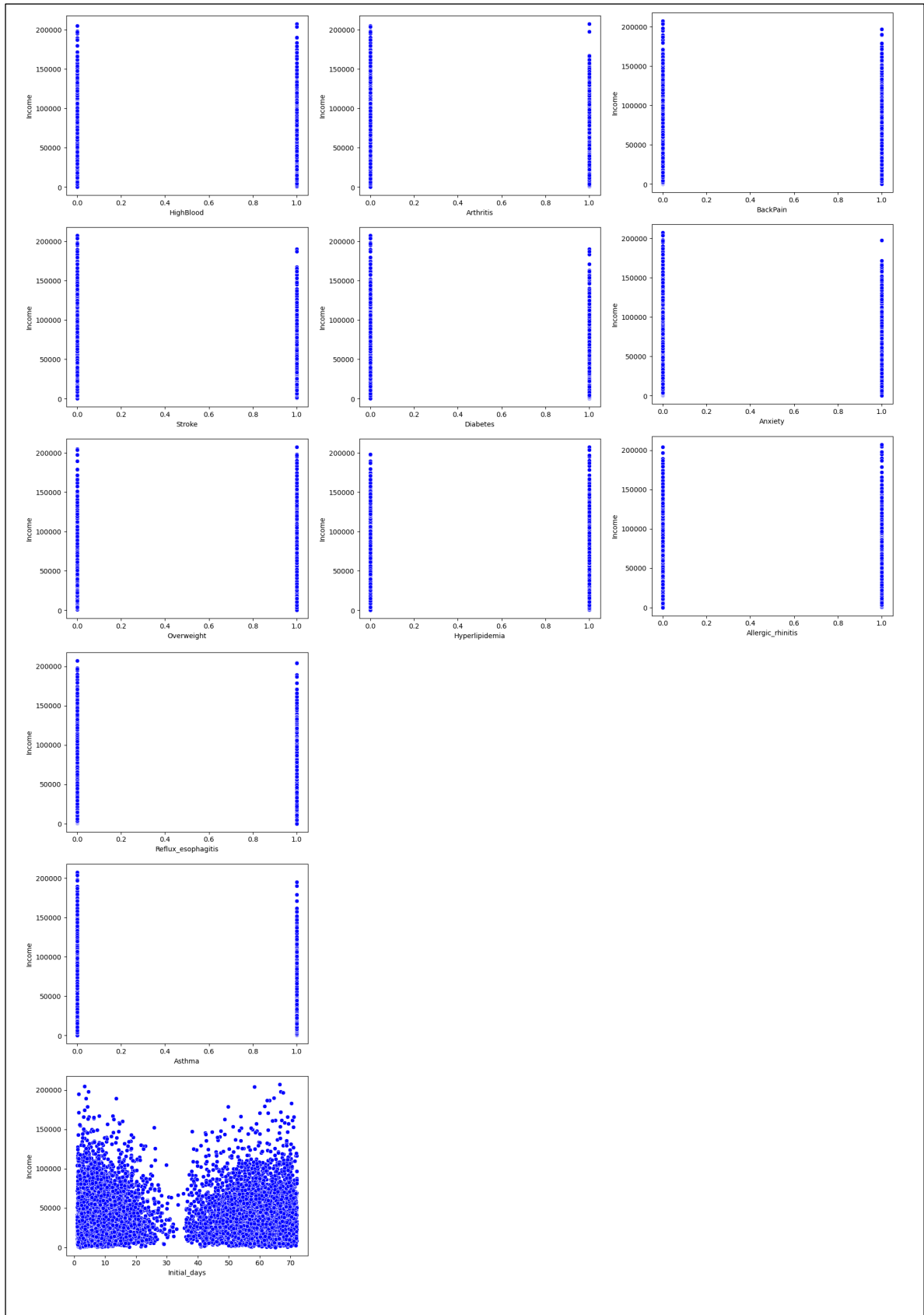


- Because this data has already been cleaned and although there appear to be variables not evenly distributed, we can create boxplots to look for outliers. Even though the boxplots show that some outliers exist, I assume that since this data has been cleaned, it was determined that the range of these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



- Bivariate visualizations were created using scatterplots and show the relationships between predictor variables and the dependent variable.





Part IV: Model Comparison and Analysis

D. Compare an initial and reduced linear regression model

1. Initial Model

- An initial regression model is run on the independent/predictor variables. These are compared against the dependent variable to determine correlation. The ordinary least-squared results can be seen below.
- The initial regression model has an R squared value of just .003, meaning that only 0.3% of the model can explain variation. This infers that none of the predictor variables influence income. The condition number is also very large and indicates a high multicollinearity relationship among the variables.

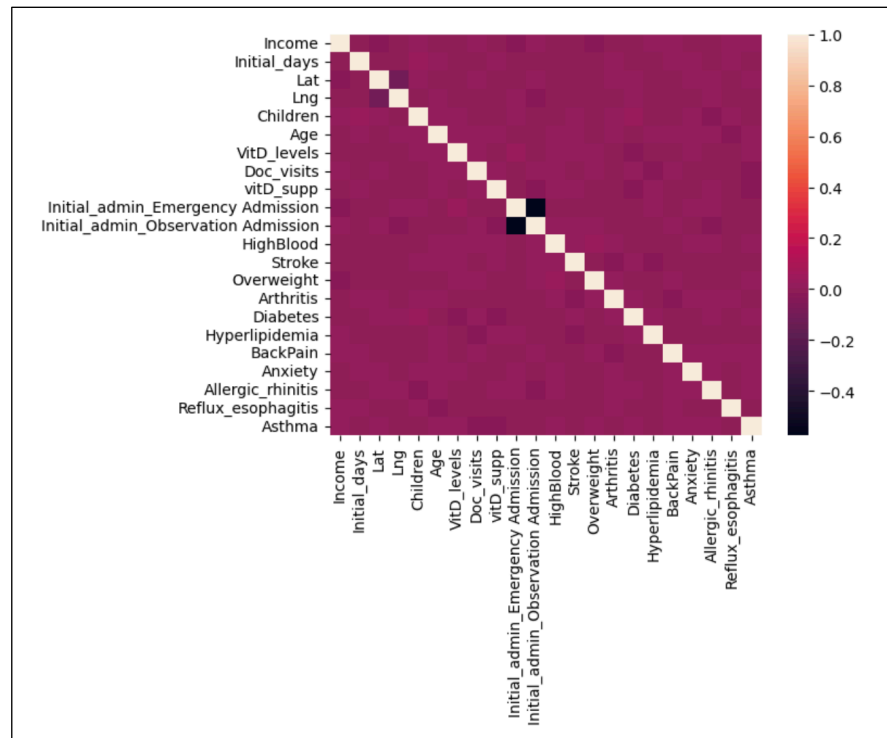
OLS Regression Results						
Dep. Variable:	Income	R-squared:	0.003			
Model:	OLS	Adj. R-squared:	0.001			
Method:	Least Squares	F-statistic:	1.327			
Date:	Sun, 24 Dec 2023	Prob (F-statistic):	0.144			
Time:	08:11:15	Log-Likelihood:	-1.1676e+05			
No. Observations:	10000	AIC:	2.336e+05			
Df Residuals:	9978	BIC:	2.337e+05			
Df Model:	21					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Lat	-109.4112	53.146	-2.059	0.040	-213.588	-5.234
Lng	-16.5179	18.886	-0.875	0.382	-53.538	20.503
Children	105.1884	131.935	0.797	0.425	-153.431	363.807
Age	-16.8670	13.832	-1.219	0.223	-43.980	10.246
VitD_levels	-175.1708	141.526	-1.238	0.216	-452.590	102.248
Doc_visits	391.5805	272.982	1.434	0.151	-143.519	926.680
vitD_supp	68.0852	454.201	0.150	0.881	-822.240	958.411
Initial_admin_Emergency Admission	-827.2409	697.404	-1.186	0.236	-2194.294	539.812
Initial_admin_Observation Admission	906.9932	812.532	1.116	0.264	-685.733	2499.719
HighBlood	-65.8454	580.339	-0.113	0.910	-1203.428	1071.737
Stroke	150.9052	714.190	0.211	0.833	-1249.050	1550.861
Overweight	-1218.1354	628.687	-1.938	0.053	-2450.488	14.218
Arthritis	-298.6226	595.619	-0.501	0.616	-1466.155	868.910
Diabetes	-672.1686	640.397	-1.050	0.294	-1927.475	583.138
Hyperlipidemia	559.9664	603.802	0.927	0.354	-623.608	1743.541
BackPain	560.0075	580.094	0.965	0.334	-577.094	1697.109
Anxiety	8.2032	610.961	0.013	0.989	-1189.404	1205.810
Allergic_rhinitis	-3.0460	584.098	-0.005	0.996	-1147.997	1141.905
Reflux_esophagitis	915.8540	579.500	1.580	0.114	-220.082	2051.790
Asthma	405.9979	629.497	0.645	0.519	-827.943	1639.938
Initial_days	-14.3918	10.852	-1.326	0.185	-35.664	6.881
const	4.599e+04	4018.999	11.443	0.000	3.81e+04	5.39e+04
Omnibus:	2561.374	Durbin-Watson:	1.985			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6430.603			
Skew:	1.402	Prob(JB):	0.00			
Kurtosis:	5.751	Cond. No.	1.70e+03			

Notes:

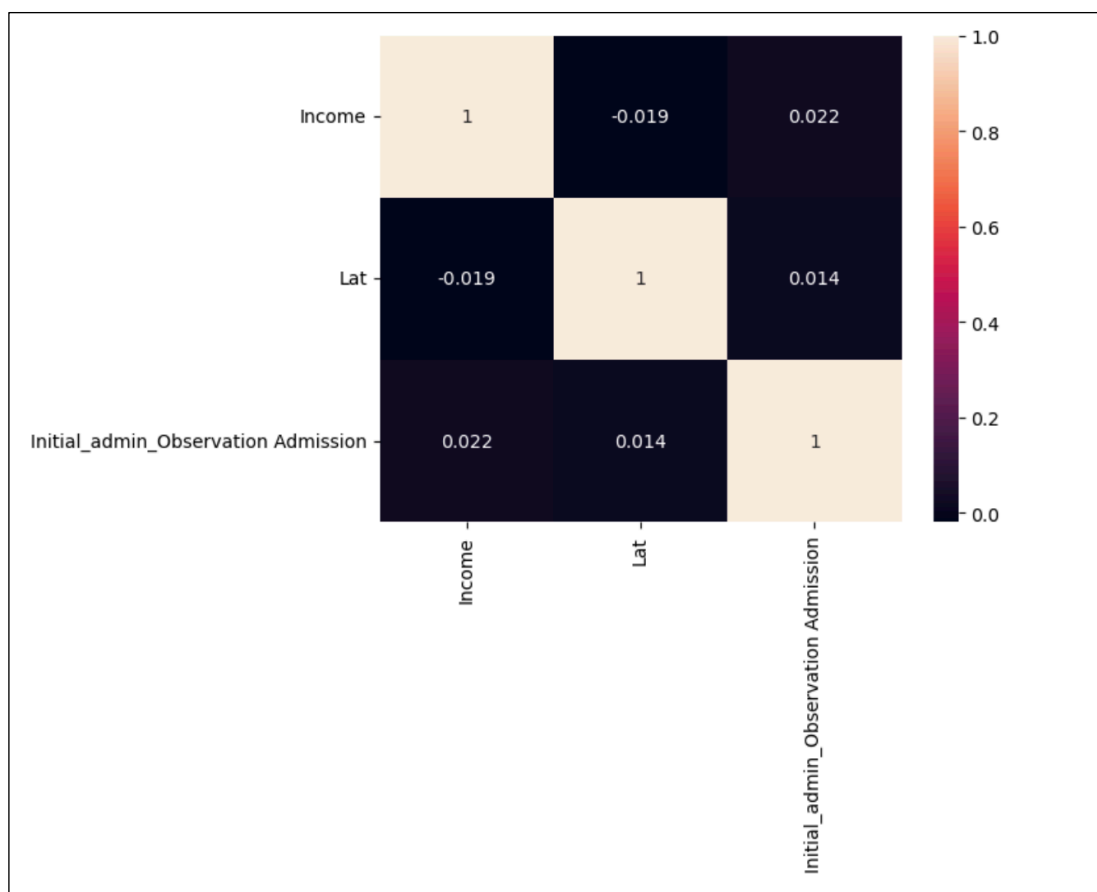
- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
 [2] The condition number is large, 1.7e+03. This might indicate that there are strong multicollinearity or other numerical problems.

2. Model Reduction Method and Justification

- Following the construction of a heatmap, I used a backwards stepwise elimination method to reduce my initial model, using p-values to determine which variables should be removed. The heatmap helps give a more detailed insight and visualization of which predictor variables may be correlated to the dependent variable. In addition, using the heatmap will allow an analyst to see where variables with high multicollinearity may be present.



- Using a heatmap, we can look to see which variables correlate to income. Squares that are lighter in color represent a higher correlation versus that of the darker squares. We can see by looking at the heatmap above that there is no correlation between the independent variables and income. Heatmaps can help an analyst determine which variables are not strong predictors and exclude them from a reduced model.
- Looking at the heatmap and the initial regression model, the only variables with a p-value of 0.05 or less are Lat and Initial_admin_Observation Admission. Therefore, we can confidently retain those variables for the reduced model. A reduced heatmap can be seen below:



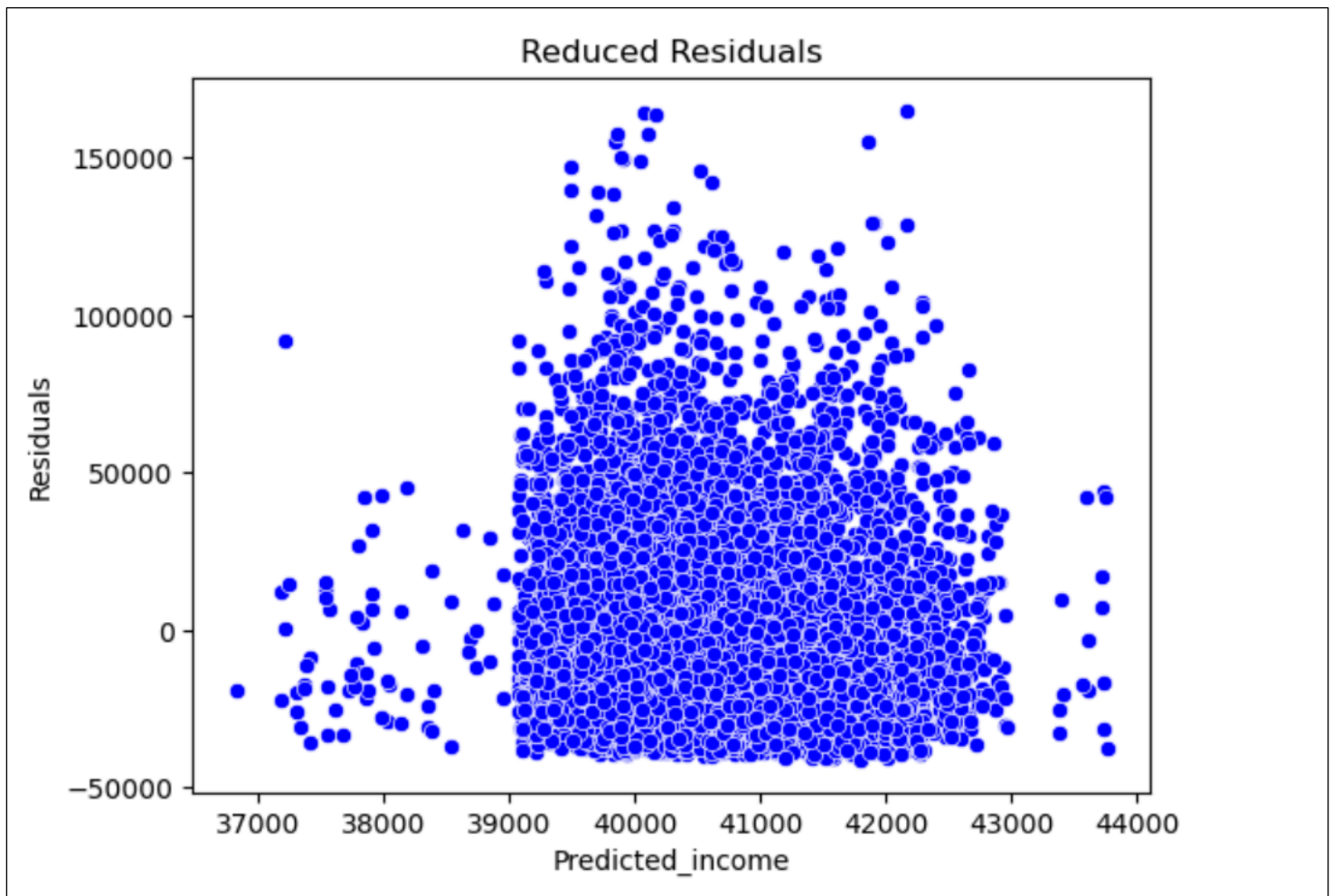
E. Model Comparison

1. Compare Initial and Reduced Regression Model

- During the initial regression model, I observed the p-values for each of the predictor variables. This allowed me to implement a backward stepwise elimination process and construct a reduced regression model. The evaluation metric used to compare both models, the r-squared values, can be seen in the initial and reduced OLS regression models. My initial r-squared value began at 0.003 and the reduced model ended at 0.001, an indication that initially only 3 percent of the variation in data could be explained, followed by only 1 percent of the variation in the reduced model.

2. Multiple Linear Regression (Task 1)

- The residual plot for the reduced regression model can be seen below.



- As seen in the residual plot above, we can tell that heteroscedasticity is present and there is not a normal distribution pattern. It appears that in this scatterplot, there are three different groups plotted.
- The model's residual standard of error was calculated and came out to 28511.623588158807

3. Provide Executable Code

- A copy of the code is attached to the submission.

Part V: Data Summary and Implications

F. Regression Equation, Coefficients, Etc.

1. Provide a Regression Equation for Reduced Model

- Provided is the regression equation for the variables in the reduced model.
- $\hat{Y} = 4.416e+04 - 103.8909(\text{Lat}) + 1475.1594(\text{Initial_admin_Observation Admission})$

2. Provide an Interpretation of the Coefficients

- With interpreting coefficients, you either get a positive or negative relationship to the dependent variable. My dependent variable is income, and my predictor variables are Lat with a negative coefficient of -103.890 and Initial_admin_Observation Admission with a positive coefficient of 1475.1594. To elaborate, for every one unit change of Lat, income will decrease by 103.89 dollars, whereas for every one unit of change in Observation Admission, income will increase 1475.1594 dollars.

3. Provide a Discussion Regarding Statistical Significance and Practical Significance of the Reduced Model

- In the reduced model, the only variables retained were Lat and Initial_admin_Observation Admission. The p-value associated with this predictor variables were 0.049 for Lat and 0.026 for Initial_admin_Observation Admission. This shows a statistical significance between income and latitude and Income and Initial_admin_Observation Admission, however. Because many of

the other selected predictor variables had to be removed from the final reduced model, it signifies that there was no statistical significance of income to overall health. This means that if a patient had a higher income and less health conditions than that of a patient with more health conditions, it could be so by chance.

- For this research question, it can be deemed that there is no practical significance for this model. That is, this model has no meaningful use for the real world or to an analyst for finding correlation of health to income.

4. Disadvantages of Methods Used/ Recommendations

- The disadvantages of running this regression model are that it only considers one dependent variable, meaning an analyst may want to use other predictor variables to find a correlation to income. In addition, it can also be said that a different dependent variable can be selected and ran with the same predictor variables to improve r-squared values and explanation of variance.
- Moreover, an analyst may need to run more than just one linear regression model to find an answer to an organizational question. This regression model consists of predictions and changing variables can improve the overall results.
- Because there is such a low confidence of relation between the predictor and dependent variables, we would need to reobserve the datapoints and potentially select more variables for the regression model. This could potentially improve over significance to the dependent variable and lead the analyst to finding stronger correlation results. In addition, there are also a great deal of medical conditions not listed in the data analysis, rather it appears just some of the most common medical conditions are inputted in the data frame. An analyst could request further surveys on medical conditions of patients and all diagnosis to the data frame.

- An analyst could also ask the hospital to reframe the research question, suggesting to not only look at income in relation to overall health but specific demographics that emphasize a patient's income, such as living in poverty which may contribute to medical conditions.

Part VI: Demonstration

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

Sources

No third-party or web sources used.