1. Describe the purpose of this data analysis
   1. What factors contribute the most to the patient diagnosed with high blood pressure?
   2. The goal of the data analysis is to identify the key factors that most significantly contribute to patients being diagnosed with high blood pressure. This involves examining various potential predictors and determining their impact on the likelihood of a high blood pressure diagnosis. The insights gained from this analysis can help in understanding the underlying causes, improving patient risk assessments, and developing targeted interventions for prevention and management.
2. Describe Logistic Regression Methods
   1. The four assumptions of the multiple logistic regression are linearity, no outliers, independence, and no multicollinearity. Linearity implies that the multiple logistic regression fits a logistic curve to binary data, which can be interpreted as the probability of each outcome based on the values of the independent variables. It assumes a linear relationship between the natural log of the odds or probabilities and the predictor variables. No outliers indicate that the variables for the multiple logistic regression must have no outliers due to the high sensitivity of unusual large or small values. Independence implies that each of the data points or observed values are independent. No multicollinearity occurs when two or more independent variables are highly correlated with each other. This condition makes the regression coefficients and their statistical significance unstable and less reliable, although it does not necessarily impact the overall fit of the model. (Multiple)
   2. Two key benefits of using Python in various phases of analysis are the functionalities provided by Pandas, and the graphic capabilities offered by Seaborn and Matplotlib. Pandas enables users to handle large datasets, including loading, reading, and cleaning the data efficiently. Seaborn and Matplotlib offer diverse graphic functionalities that enhance data visualization.
   3. Multiple logistic regression is an appropriate technique for analyzing the research question because the multiple logistic regression is designed to handle categorical dependent variables, making it suitable for analyzing complication risk levels category. It allows the simultaneous examination of multiple independent variables such as age, gender, medical history etc. to determine their dividual and combined effects on the complication risk. Multiple logistic regression can also capture non-linear relationships between the independent variables and the log-odds of the outcome, offering more flexibility in modeling complex relationships.
3. Summarize the data preparation process for logistic regression
   1. The goal of this data cleaning process is to convert all Boolean columns into numerical columns using one-hot encoding and change the selected features' data types to float64 for logit calculations in logistic regression. I created another data frame with dependent (HighBlood) and independent variables (Age, Gender, VitD\_levels, Initial\_admin, Stroke, ReAdmis, Asthma, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Services, Soft\_drink, Initial\_days, TotalCharge, Doc\_visits) and started cleaning the dataset from there. These independent variables were chosen because they are related to the illness, which can be associated with the high blood pressure. They are a mixture of numeric, nominal, and Boolean variables. There were no missing values in 19 columns when checked with the info function. The get\_dummies function was conducted with drop\_first=True to avoid multicollinearity. However, after one-hot encoding, four column headers (Initial\_admin\_Emergency Admission, Initial\_admin\_Observation Admission, and Services\_CT Scan) were created with spaces, so they were replaced with underscores (\_).  
      The columns that are not included in the linear regression model are CaseOrder, Customer\_id, Interaction, UID, City, State, Country, Zip, Lat, Lng, Population, Area, TimeZone, Job, Children, Income, Marital, Full\_meals\_eaten, vitD\_supp, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Additional\_charges, Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8.
   2. The dependent variable y is “HighBlood\_Yes”, and the following are the independent variables x: Age, Gender, VitD\_levels, Initial\_admin, Stroke, ReAdmis, Asthma, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Services, Soft\_drink, Initial\_days, TotalCharge, Doc\_visits.

A screenshot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generatedA screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

* 1. Following screenshots are the univariate of variables

A group of blue and white bars

Description automatically generated

A group of blue rectangular bars

Description automatically generated with medium confidence

A white background with black text

Description automatically generated

Following screenshots are the bivariate of variables with the dependent variable, HighBlood.

A screenshot of a computer screen

Description automatically generated

A group of white squares with blue dots

Description automatically generated

A group of white squares with blue dots

Description automatically generated

* 1. One-hot encoding was used enable the inclusion of categorical variables in models that require numerical input. It can also enhance model performance by supplying more detailed information about the categorical variables. Additionally, it helps to prevent issues related to ordinality, which can arise when a categorical variable has a natural order such as small, medium, and large (One).   
     One-hot encoding was applied to identify categorical variables like Gender, which categories such as Male and Nonbinary (excluding Female as the reference category). Each categorical column was converted into binary columns, allowing the model to analyze each category in greater depth. This enables a detailed examination of which factors most significantly contribute the length of the hospital stay.   
     One-hot encoding was used with get\_dummies with dropping the first column of each categorical columns, and updated the column name with space to underscore(\_).   
     All the data types were converted to float64 from into64 and bool to have all numeric values in the dataset for the logistic regression modeling. MinMaxScaler was also used to set the minimum as 0 and the maximum as 1 to normalize the values in the entire data set.
  2. ‘medical\_one\_hot.csv’ is submitted.

1. Compare an initial and reduced logistic regression model
   1. Initial Logistic Regression Model

First, the logit method from statsmodels.formula.api was conducted to get the initial intercept and the coefficients on each independent variables.   
A screenshot of a computer

Description automatically generated

A table with numbers and text

Description automatically generated  
Next, y was set to ‘HighBlood\_yes’, X as mentioned predictors, and set split the dataset into x\_train, X\_test, y\_train, and y\_test with the test size 0.3 and random state 21.   
The initial logistic regression score on X\_train, and y\_train is 0.6218571428571429, and the score on X\_test, and y\_test is 0. 6386666666667.

* 1. Utilized RFE (Recursive Feature Elimination) with LogisticRegression() to reduce the features down from 23 to 8. Selected features are Initial\_days, TotalCharge, Initial\_admin\_Emergency\_Admission, Complication\_risk\_Low, Complication-risk\_Medium, Diabetes\_Yes, Hyperlipidemia\_Yes, and BackPain\_Yes.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

* 1. Following screenshots are the logit function of the reduced logistic regression model with selected features.

A screenshot of a computer

Description automatically generated

A screenshot of a white background with numbers

Description automatically generated

‘HighBlood\_yes’ was set to y, selected features as X, and set split the dataset into x\_train, X\_test, y\_train, and y\_test with the test size 0.3 and random state 21.   
The reduced logistic regression score on X\_train, and y\_train is 0.6235714285714286, and the score on X\_test, and y\_test is 0.649666666666666.

1. Analyze the data set using the reduced logistic regression model
   1. Compare initial & reduced logistic regression model
      1. A table with numbers and letters

         Description automatically generated

Based on the screenshot above, the logistic regression function can be concluded as ln(p̂/(1- p̂)) = -13.8612 - (139.0871 \* Initial\_days) + (173.0383 \* TotalCharge) – (12.2648 \* Initial\_admin\_Emergency\_Admission) + (9.6931 \* Complication\_risk\_Low) + (9.8687 \* Complication\_risk\_Medium) – (1.8242 \* Diabetes\_Yes) - (2.2184 \* Hyperlipidemia\_Yes) – (2.0333 \* BackPain\_Yes)

* + 1. The following can be used to conclude:  
       Keeping all things constant, for one unit decrease in Initial\_days, the changes log odds of CVD by 139.0871, for one use increase in TotalCharge, the changes log odds of CVD by 139.0871, for the one unit decrease in Initial\_admin\_Emergency\_Admission, the changes log odds of CVD by 12.2648, for one unit increase in Complication\_risk\_Low, the changes log odds of CVD by 9.6931, for one unit increase in Complication\_risk\_Medium, the changes log odds of CVD by 9.8687, for one unit decrease in Diabetes\_Yes, the changes log odds of CVD by 1.8242, for one unit decrease in Hyperlipidemia\_Yes, the changes log odds of CVD by 2.2184, for one unit decrease in BackPain\_Yes, the changes log odds of CVD by 2.0333.
    2. This reduced logistic regression model is not statistically significant because there are almost no different in the logistic regression score between initial and reduced in both train and test (0.6219 to 0.6236 for training set and 0.6387 to 0.65 for the test set).   
       This is not practically significance either due to the accuracy of the reduced logistic regression model is only 65%.
    3. The initial regression model started with 19 independent variables out of 49 available in the dataset. It’s possible that some important columns were omitted and some unnecessary columns were included, which could have affected the results. Additionally, the dataset does not indicate whether the patient died during their hospitalization, was transferred to another hospital for further treatment, or was sent home. Eating habits or life style would’ve better fitted in the dataset. The total income should not have been collected or not listed in the data. This might indicate that the lower income personnel would have been treated poorly or not received necessary treatment due to their financial issue.
    4. To answer the research question, Initial\_days, TotalCharge, Initial\_admin\_Emergency\_Admission, Complication\_risk\_Low, Complication-risk\_Medium, Diabetes\_Yes, Hyperlipidemia\_Yes, and BackPain\_Yes are affecting to determine the patient with the high blood pressure.   
       The organization can develop an risk stratification and monitoring protocol to categorize patients by their complication risks (low, medium, and high) to regularly monitor the patients for high blood pressure and intervene early to prevent escalation. Emergency admission protocols can be adopted to ensure the high blood pressure screening is part of the initial assessment. Include the back pain management as part of the overall care plan as a chronic pain can contribute to increased blood pressure. Offer pain management strategies such as physical therapy, medication, and lifestyle modifications may reduce in the blood pressure.
    5. Pseudo R-squared: utilized when the outcome variable is nominal or ordinal, making the traditional R-squared coefficient of determination unsuitable for measuring goodness of fit. These values are applied in contexts where a likelihood function is used to fit a model. The pseudo R-square value was decreased from 0.4659 to 0.3716 which indicates that the reduced logistic regression model fits less than the initial model even after dropping the predictors that are not statistically significant using RFE function with LogisticRegression().  
       Log-Likelihood: the probability that observed values of the dependent variable can be predicted from observed independent variable values. The log-likelihood value was decreased from -3612.8 to -4251.0 which indicates that the reduced logistic regression model fits less than the initial model even after dropping the predictors that are not statistically significant using RFE function with Logistic Regression().
  1. Output and all calculations of the analysis performed
     1. Confusion Matrix  
        Following is a confusion matrix of an initial logistic regression model  
        A chart of different colored squares

        Description automatically generated

Following is a confusion matrix of an reduced logistic regression model  
A chart of different colored squares

Description automatically generated

* + 1. Accuracy Calculation

Following is the screenshot of the accuracy of the initial logistic regression model  
A screenshot of a computer

Description automatically generated

Following is the screenshot of the accuracy of the reduced logistic regression model

A screenshot of a computer

Description automatically generated

* 1. Submitted “Logistic Regression Modeling – PA2.ipynb”.

1. Summarize the findings and assumptions
   1. Results of analysis
      1. Again, this is the equation for the reduced logistic regression model:   
         ln(p̂/(1- p̂)) = -13.8612 - (139.0871 \* Initial\_days) + (173.0383 \* TotalCharge) – (12.2648 \* Initial\_admin\_Emergency\_Admission) + (9.6931 \* Complication\_risk\_Low) + (9.8687 \* Complication\_risk\_Medium) – (1.8242 \* Diabetes\_Yes) - (2.2184 \* Hyperlipidemia\_Yes) – (2.0333 \* BackPain\_Yes)
      2. Out of 8 variables that affects to determine the patient diagnosed with high blood pressure, 3 would increase and the other 5 would decrease the log odds of CVD.
      3. Based on the result, this model is not statistically and practically significance due to the reduced model having lower accuracy and almost no difference than the initial model.
      4. The limitation of this analysis was there were limits on the number of feature selections to start with for the initial model. It would’ve been better if the dataset did not have the income column and added the information if the patient’s daily habits.
   2. Recommend a course of action
      1. My analysis based on the given dataset wouldn’t mean significant due the dataset not having all the factors that need to be considered for issue in the real life. It also would have been a greater dataset if it also had variety of treatments from different departments for how long or how many times.
2. Panopto Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=f541ce58-96aa-46ba-93e4-b19d015110f5>
3. Web sources used to acquire data or segments of third-party code to support the application

Hackers Realm. *Recursive Feature Elimination (RFE) | Feature Selection | Python*. YouTube. (June 9, 2022). https://www.youtube.com/watch?v=vxdVKbAv6as.

*Machine Learning – Confusion Matrix*. W3schools. (n.d.). <https://www.w3schools.com/python/python_ml_confusion_matrix.asp>.

*Multiple logistic regression*. StatsTest.com. (2020, May 18). https://www.statstest.com/multiple-logistic-regression/

RegenerativeToday. *Step by Step Tutorial on Logistic Regression in Python | sklearn | Jupyter Notebook*. YouTube. (May 18, 2022). https://www.youtube.com/watch?v=bSXIbCZNBw0.

Villazon, Andrew. *Logistic Regression in Python with statsmodels*. Andrew Villazon. (November 14, 2021). https://www.andrewvillazon.com/logistic-regression-python-statsmodels/.

1. Acknowledge sources, using in-text citations and references

Multiple logistic regression. StatsTest.com. (2020, May 18). <https://www.statstest.com/multiple-logistic-regression/>.

*One Hot Encoding in Machine Learning.* GeeksforGeeks. (March 21, 2024). <https://www.geeksforgeeks.org/ml-one-hot-encoding>.