1. Describe the purpose of this data analysis
   1. What factors contribute the most to the length of the hospital stay?
   2. Employ multiple linear regression analysis to ascertain the most significant factors influencing the length of hospital stays. This approach can elucidate the reason behind patients’ varying lengths of hospitalization, the duration of specific treatment, and the impact of patient conditions such as age, gender, and location. (Simple)
2. Describe multiple linear regression methods
   1. The four assumptions of the multiple linear regression are linearity, independence, homoscedasticity, and normality. Linearity implies that the relationship between the independent variable X and the mean of the dependent variable Y is linear. Homoscedasticity indicates that the variance of the residuals remains constant for any value of X. Independence means that the observations are independent of each other. Normality asserts that the dependent variable Y is normally distributed for any fixed value of X. ()
   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 linear regression is an appropriate technique for analyzing the research question because the multiple linear regression quantifies the relationship between the dependent variable (length of hospital stay) and multiple independent variables (factors such as age, gender, medical condition, etc). It provides coefficients that indicate the strength and direction of each factor’s influence on the hospital stay duration. Multiple linear regression can also provide p-values for each predictor, helping to determine which factors are statistically significant contributors to the length of hospital stay.
3. Summarize the data preparation process for multiple linear regression analysis
   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 OLS calculations in linear regression. I created another data frame with dependent (Initial\_days) and independent variables (Age, Income, Gender, VitD\_levels, Doc\_visits, Initial\_admin, HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Services, and TotalCharge) 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 days of hospitalization, and represent the patient (such as age and gender). They are a mixture of numeric, nominal, and Boolean variables. There were no missing values in 17 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, 3 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, Marital, ReAdmis, Full\_meals\_eaten, vitD\_supp, Soft\_drink, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, Additional\_charges, Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8.
   2. The dependent variable y is “initial\_days”, and the following are the independent variables x: Age, Income, Gender, VitD\_levels, Doc\_visits, Initial\_admin, HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Services, and TotalCharge.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

* 1. Following screenshots are the univariate of variables

A group of blue and white graphs

Description automatically generated with medium confidence

A group of blue rectangular bars

Description automatically generated

A close-up of a graph

Description automatically generated

Following screenshots are the bivariate of variables with the dependent variable, Initial\_days

A group of blue squares

Description automatically generated

A group of graphs with numbers

Description automatically generated with medium confidence

A group of graphs with numbers

Description automatically generated with medium confidence

* 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(\_).
  2. ‘medical\_one\_hot.csv’ is submitted.

1. Compare an initial and a reduced linear regression model
   1. See the following screenshots

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

* 1. After applying one-hot encoding, the data types of each column were checked, which revealed int64, bool, and float64 types. The columns that were into64 and bool were converted to float64.   
     The initial model started with 21 independent variables.   
     The VIF (variance inflation factor) method was applied to check for the multicollinearity and remove the columns with the VIF score greater than 10. It was found that the VIF for VitD\_levels and Docs\_visits were over 10. Those two columns were removed and conducted the VIF function again to ensure no other columns had a VIF over 10.   
     I used MinMaxScaler to standardize the variables which will convert the values as minimum to 0 and maximum to 1.  
     OLS regression method was conducted again, this time removing columns based on their p-values. Columns with p-values greater than 0.05 (Age, Services\_MRI, Services\_Intravenous, Services\_CT\_Scan, Income, Gender\_Nonbinary, Stroke\_Yes, Overweight\_Yes, Gender\_Male, and Initial\_admin\_Observation\_Admission) were removed one by one.
  2. Following is the screenshots of reduced linear regression result.

A screenshot of a computer screen

Description automatically generated

A white and black text

Description automatically generated with medium confidence

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

         Description automatically generated

Based on the screenshot above, the linear regression function can be concluded as Initial\_days = -0.0832 + (1.2439 \* TotalCharge) – (0.0882 \* Initial\_admin\_Emergency\_Admission) – (0.0195 Blood\_Yes) + (0.0710 \* Complication\_risk\_Low) + (0.0712 \* Complication\_risk\_Medium) – (0.0128 \* Arthritis\_Yes) – (0.0129 \* Diabetes\_Yes) – (0.0159 \* Hyperlipidemia\_Yes) – (0.0149 \* BackPain\_Yes)

* + 1. The following can be used to conclude:

Keeping all things constant, 1 unit increase in total charge is associated with 1.2439 increase in days hospitalization, a patient with emergency admission as initial admission is associated with 0.0882 decrease in days hospitalization, patient with high blood pressure is associated with 0.0195 decrease in days hospitalization, a patient with low complication risk is associated with 0.0710 increase in days hospitalization, a patient with medium complication risk is associated with 0.0712 increase in days hospitalization, a patient with arthritis is associated with 0.0128 decrease in days hospitalization, a patient with diabetes is associated with 0.0129 decrease in days hospitalization, a patient with hyperlipidemia is associated with 0.0159 decrease in days hospitalization, and a patient with back pain is associated with 0.0149 decrease in days hospitalization.

* + 1. This reduced linear regression model is statistically significant because all the columns with their p-values equal to or less than 0.05 remained and the rest was removed. This means that the null hypothesis should be rejected. Also removed 2 columns that were multicollinearity with VIF scores over 10.
    2. The initial regression model started with 16 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. 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.  
       Using the linear regression model with the categorical-heavy variables were not a good decision because linear regression is fit the best for the continuous values to predict the dependent variable.
    3. To answer the research question, TotalCharge, Initial\_admin\_Emergency\_Admission, HighBlood\_Yes, Complication\_risk\_Low, Complication\_risk\_Medium, Arthritis\_Yes, Diabetes\_Yes, Hyperlipidemia\_Yes, and BackPain\_Yes were contributing the most to determine the length of the hospitalization of the patient.  
       The organization can develop a comprehensive care plans that address multiple conditions simultaneously. Implement early intervention strategies and continuous monitoring for patients with high-risk conditions like diabetes, hyperlipidemia, and back pain. This can help in preventing complications and reducing the length of stay. Also creating a specialized units or teams for managing chronic conditions such as arthritis, diabetes, and backpain can lead to better patient outcomes and shorter hospital stays.
    4. R-squared: shows how well the data fits the model. R-squared was 0.999 for both initial and reduced models. Even after reducing the model by dropping the predictors that are not statistically significant, the R-squared remained the same.  
       Adjusted R-squared: shows the modified version of R-squared that measures the goodness of fit in linear regression models. Adjusted R-squared was also 0.999 for both initial and reduced models which means the adjusted R-squared remained the same after dropping non statistically significant predictors.
  1. Residual Plot & Residual Standard Error
     1. Residual Plot  
        A group of graphs showing different types of data

        Description automatically generated with medium confidenceA screenshot of a graph

        Description automatically generated  
        A group of graphs with blue dots

        Description automatically generated  
        A group of graphs with text

        Description automatically generated with medium confidenceA group of graphs with different colored lines

        Description automatically generated with medium confidenceA group of graphs with blue dots

        Description automatically generated  
        A group of graphs with red and blue dots

        Description automatically generated  
        A graph of a graph of a graph

        Description automatically generated with medium confidence  
        A group of graphs showing different values

        Description automatically generated with medium confidence
     2. Model’s residual standard error  
        The initial linear regression model’s residual standard error was 0.7044857037541815.   
        The reduced linear regression model’s residual standard error was 0.00992684848177141.
  2. Submitted ‘Saemi Ramirez D208 Linear Regression Modeling PA1.ipynb’.

1. Summarize the findings and assumptions
   1. Results of analysis
      1. Again, this is the equation for the reduced linear regression model:  
         **Initial\_days** = -0.0832 + (1.2439 \* TotalCharge) – (0.0882 \* Initial\_admin\_Emergency\_Admission) – (0.0195 Blood\_Yes) + (0.0710 \* Complication\_risk\_Low) + (0.0712 \* Complication\_risk\_Medium) – (0.0128 \* Arthritis\_Yes) – (0.0129 \* Diabetes\_Yes) – (0.0159 \* Hyperlipidemia\_Yes) – (0.0149 \* BackPain\_Yes)
      2. Out of 9 variables that affects the patients’ length of the hospitalization, 3 would increase and the other 6 would decrease the hospitalization duration.
      3. Based on the result, the organization of the hospitals can develop a comprehensive care plans, implement early intervention strategies and continuous monitoring, and create a special unit or team for the chronic conditions management.
      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 died during their hospitalization.
   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=90138a93-a1ab-4585-91ce-b19a002af6b2>
3. Web sources used to acquire data or segments of third-party code to support the application

Arham, Muhammad. *Pandas: How to One-Hot Encode Data.* KDnuggets. (July 24, 2023). https://www.kdnuggets.com/2023/07/pandas-onehot-encode-data.html.  
  
*How to Create a Residual Plot in Python*. GeeksforGeeks. (February 21, 2022). https://www.geeksforgeeks.org/how-to-create-a-residual-plot-in-python/.

Keith, Mark. *Python: MLR, OLS, Standardization, normalization*. YouTube. (October 11, 2021). https://www.youtube.com/watch?v=QH\_elD\_JKuc&t=205s.

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

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

*Simple Linear Regression*. Correlation and Regression with R. (n.d.). https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5\_Correlation-Regression/R5\_Correlation-Regression4.html#:~:text=Linearity%3A%20The%20relationship%20between%20X,X%2C%20Y%20is%20normally%20distributed.