**Multiple Linear Regression: Initial Hospital Stay Duration**

Heidi L. Petersen

College of Information Technology, Western Governors University

D208: Predictive Modeling

Dr. William Sewell

June 21, 2021

**Objectives of Analysis**

**The main objective of this analysis will be to create a linear equation to predict the number of days a patient will initially be admitted to the hospital. The target variable will be Initial\_days, a continuous variable.**

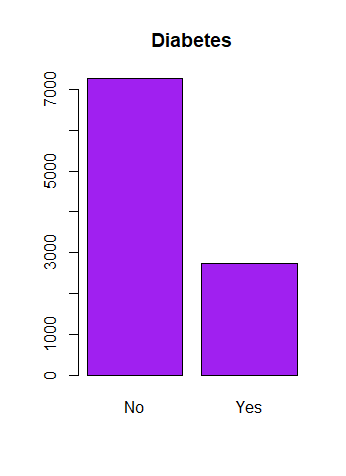
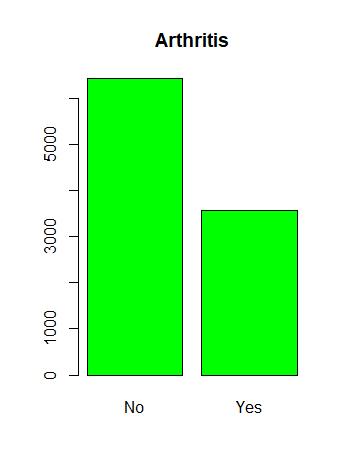
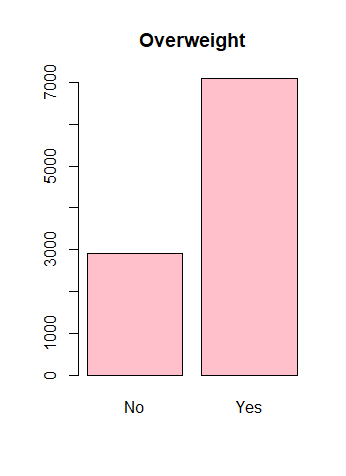
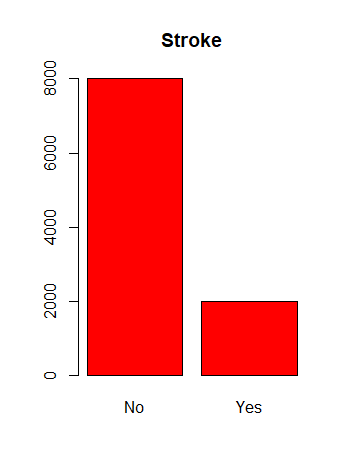
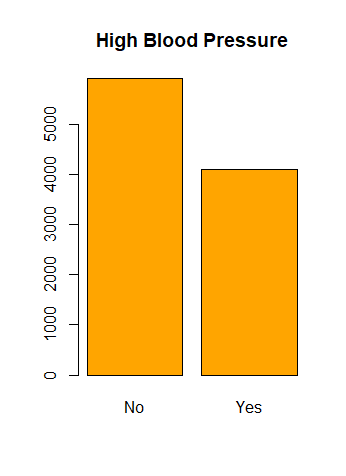
**The analysis will use multiple linear regression. Linear regression is an appropriate method for this analysis because it is able to forecast trends in data that fulfill certain assumptions (which this data set does, as will be demonstrated). This technique uses many predicator variables to predict the value of a specific response variable, which is what type of variables contained in the data set.**

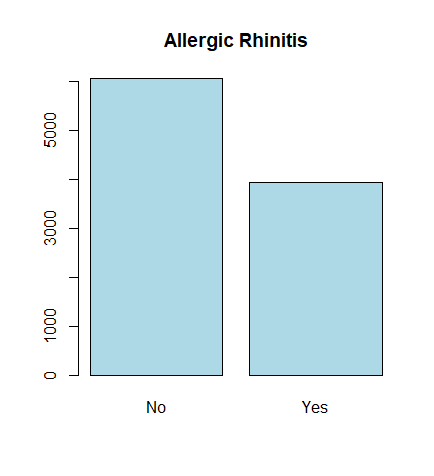
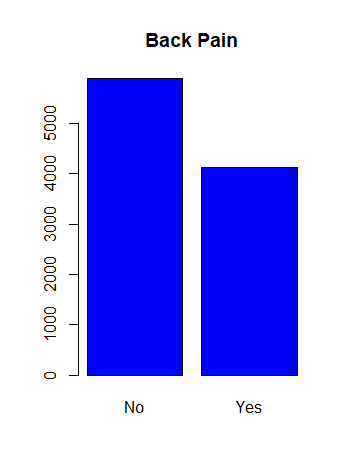
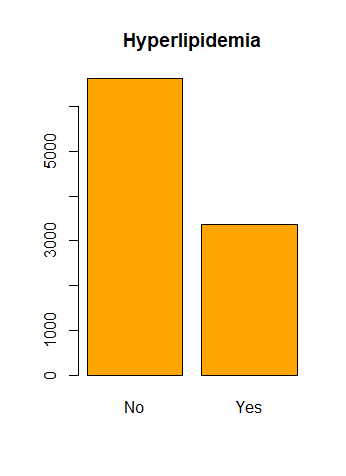
**Multiple linear regression requires a few assumptions: linearity (there exists a linear relationship between the dependent and independent variables), independence (the independent variables do not have a high correlation with each other), normality of residuals (the residuals are normally distributed), and homoscedasticity (the residuals in the model have constant variance).**

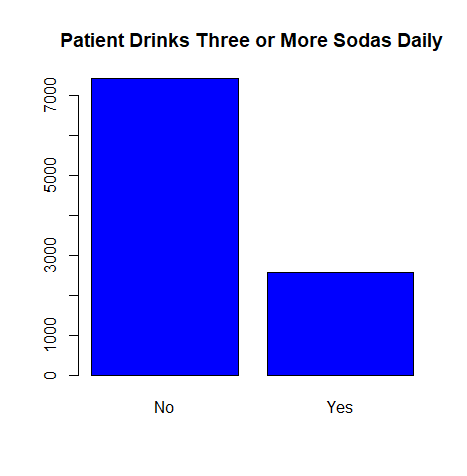
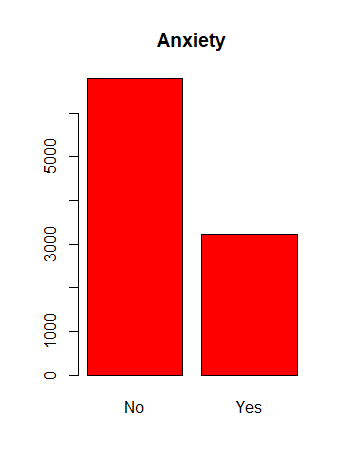
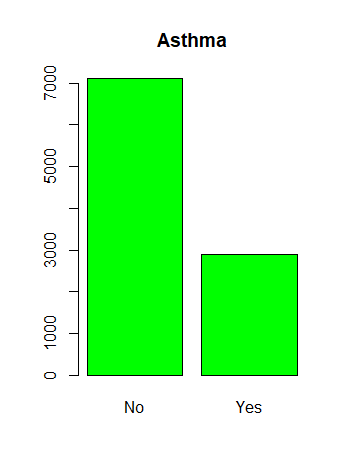
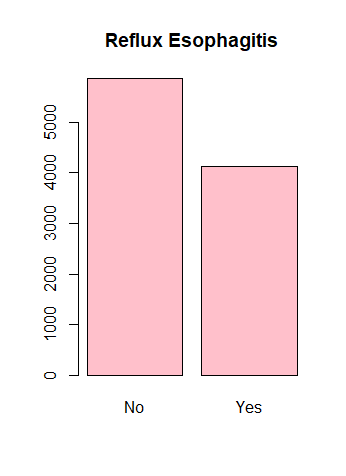
I chose R to process this data. R is a good resource for statistical analysis. It has the features needed to achieve the goals of this particular analysis quickly and efficiently. I mostly chose R for its ease. SAS would have been my choice if the data set had been much larger, and I needed to worry about RAM. Python is more of a generalized program where R is specifically designed for statistics.

**Univariate Analysis**

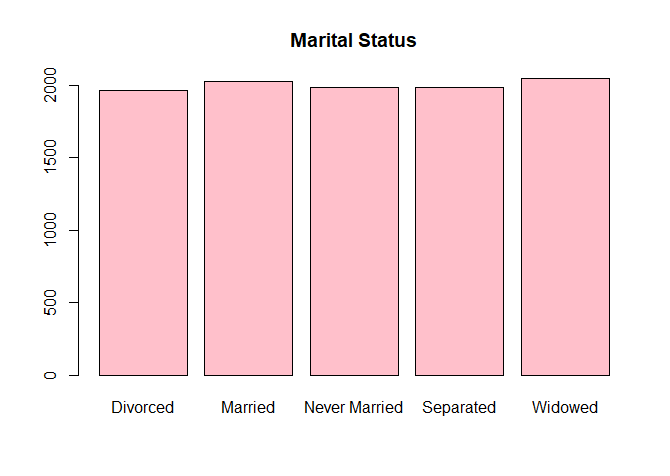
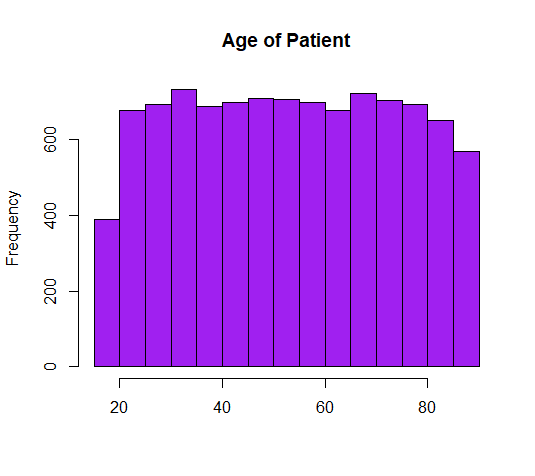
Included in the data, was information on whether each patient had several different conditions: high blood pressure, stroke, overweight, arthritis, diabetes, hyperlipidemia, chronic back pain, allergic rhinitis, reflux esophagitis, asthma, anxiety, and excessive soda intake. The only condition that was more prevalent than not was being overweight.

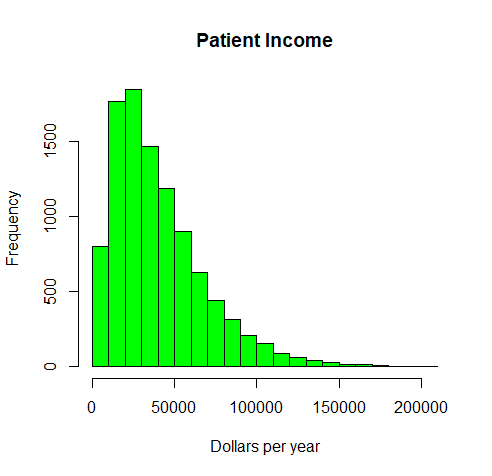
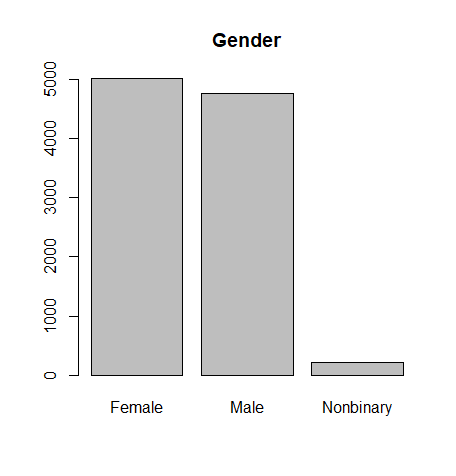


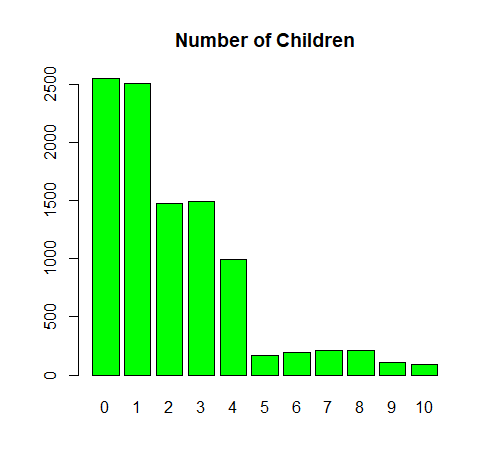
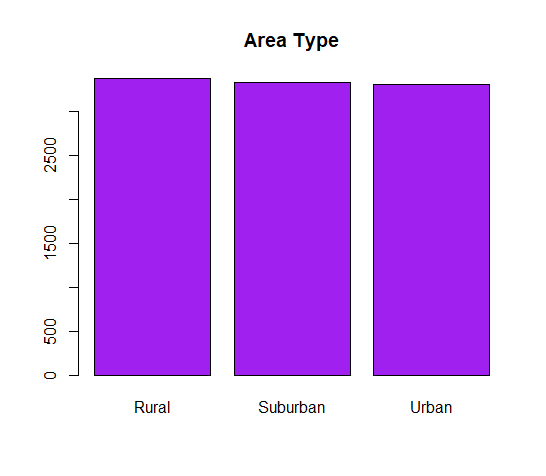




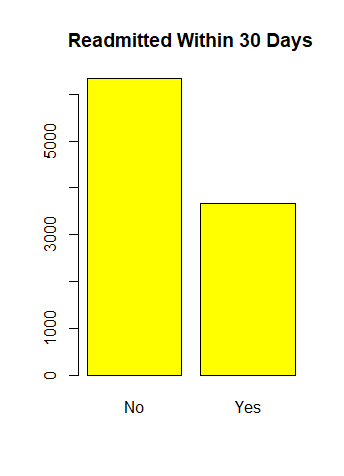
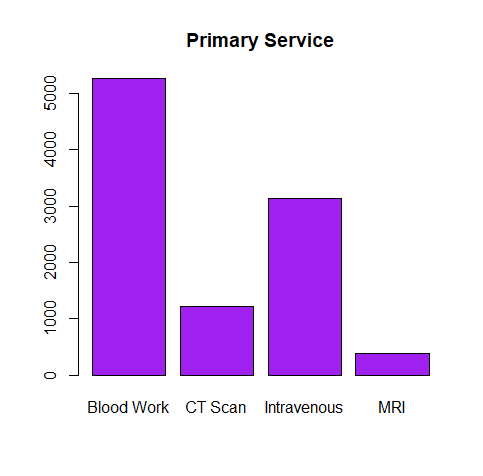
The data also included demographic information on each patient. These categories included age, marital status, gender, income, area, and number of children.





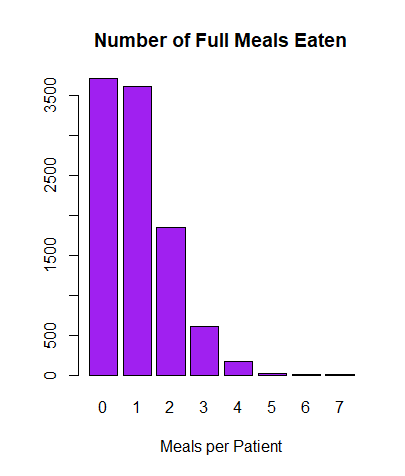
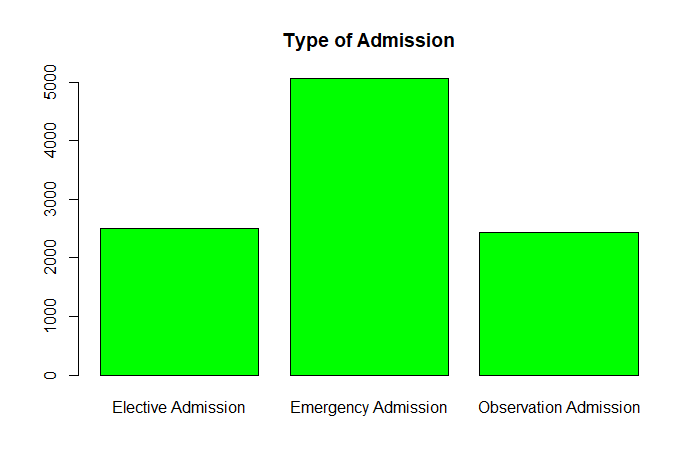


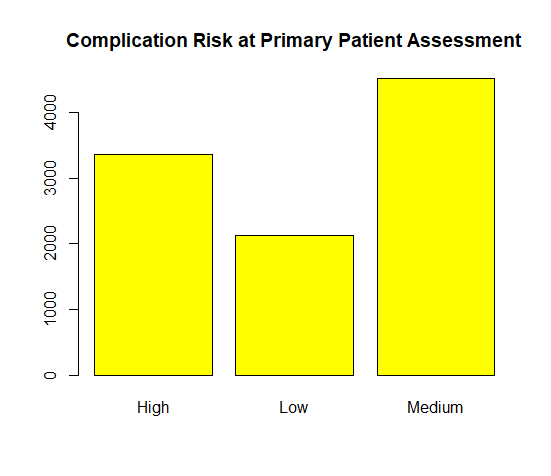
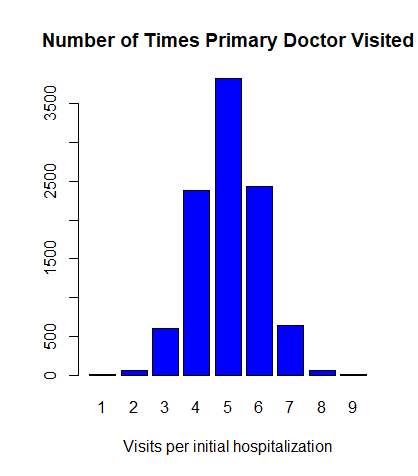
The data includes information about hospitalization: the primary service the patient received, whether they were readmitted within 30 days of initial hospitalization, how many days the initial stay consisted of, type of admission, number of full meals the patient ate, the number of times the primary physician visited during the initial stay, complication risk as assessed by a primary assessment, and the number of times a patient was given a Vitamin D supplement.



Chart, histogram

Description automatically generated



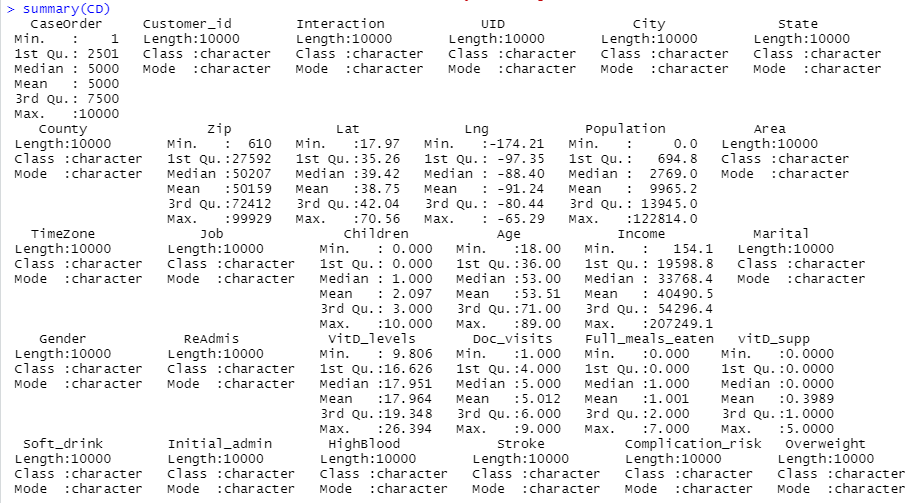
Chart, histogram

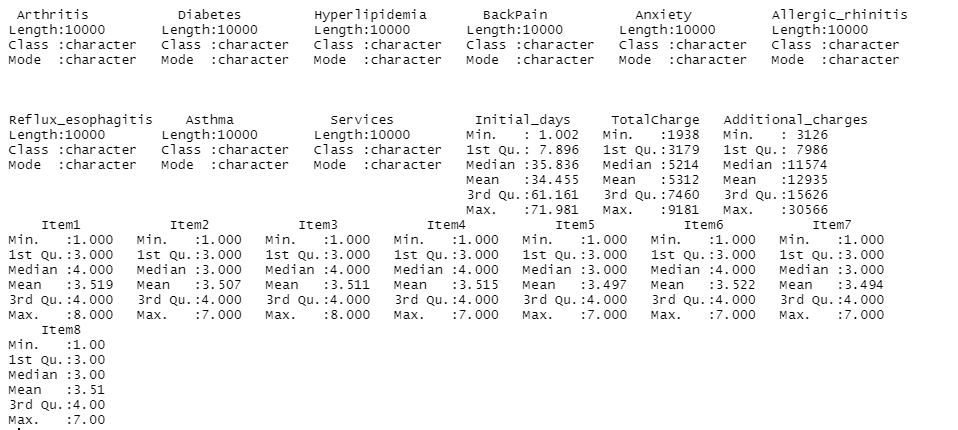
Description automatically generated

**Cleaning the Data Set:**

The goals of data preparation are to drop irrelevant data, check for missing or unexpected values, transform variables, and do an initial overview of all the variables in the data set.

I imported the data set into R and changed the data into a data frame. I summarized the data frame to get an idea of what I would be working with. There are 10,000 observations and 50 variables contained in the set.

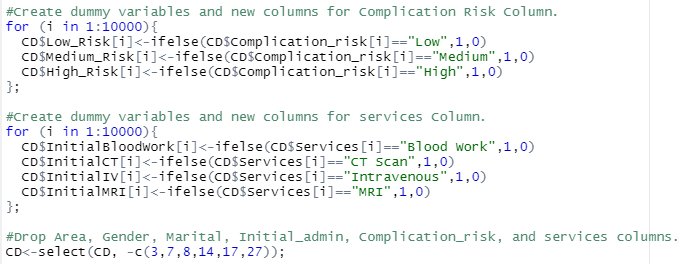




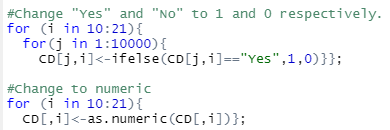
I first dropped CaseOrder, Customer\_id, Interatction, and UID because they are not relevant to the analysis. City, State, County, Lat, Lng, and TimeZone can be captured with the zip code variable so I dropped those as well. I dropped the Job column because that info is captured in Income, and it could lead to overfitting by including specific occupations. In addition, I dropped ReAdmis because I am only interested in the initial hospitalization. This left me with 38 variables.



After dropping the columns, I needed to create dummy variables from the categorical variables and remove those columns.

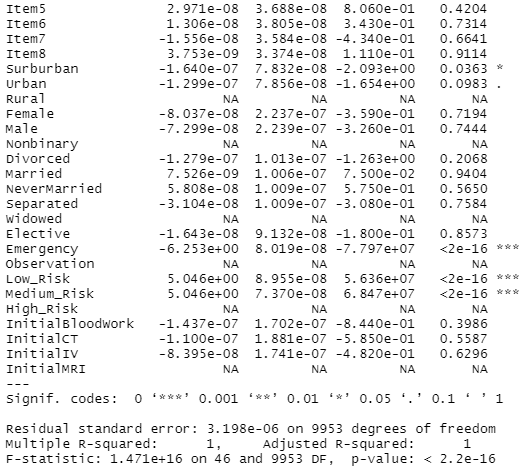
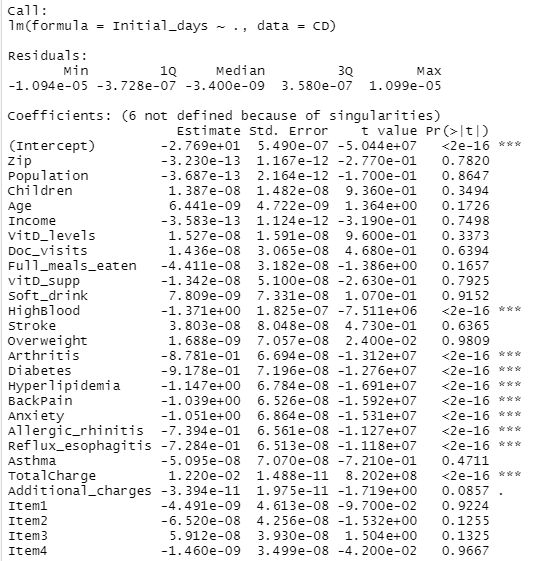


I then changed any “yes” or “no” values to 1 and 0. I also changed any values that were not already numeric into numeric values.



**Gross Multiple Linear Regression Model**

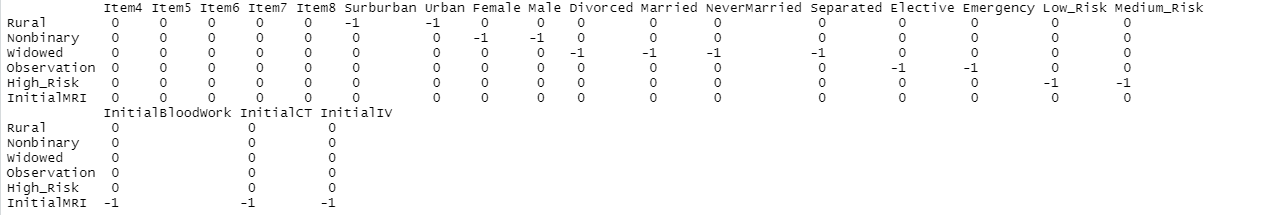
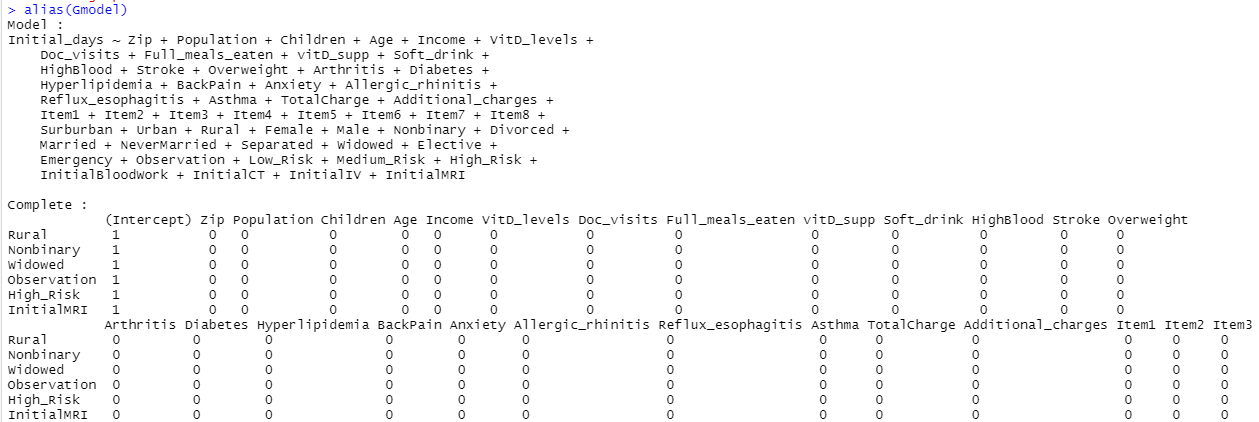
After I had the cleaned data, I need to create my initial model with those variables. After creating the data, I could see that there were a few variables that were linearly related to each other, as well as having a large number of independednt variables. I needed to create a linear regression model that fulfilled the assumptions.



**Reduced Linear Regression Model**

To create a reduced linear regression model, I needed to drop variables with no predictive value and those that were highly correlated with others, run a primary component analysis to detemine the most important factors, repeat regression to reduce variables until R2 was acceptably close to adjusted R2.

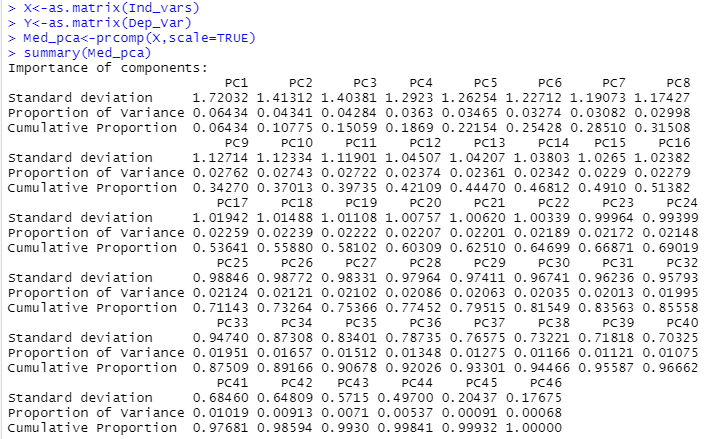
The initial model, gave the following information:



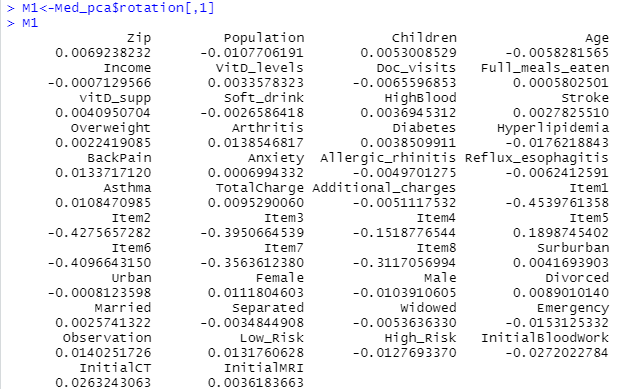
First, I needed to drop variables that were linearly related to others. I dropped medium risk, elective admission, nonbinary gender, never married marital status, rural area, and intravenuous initial service.

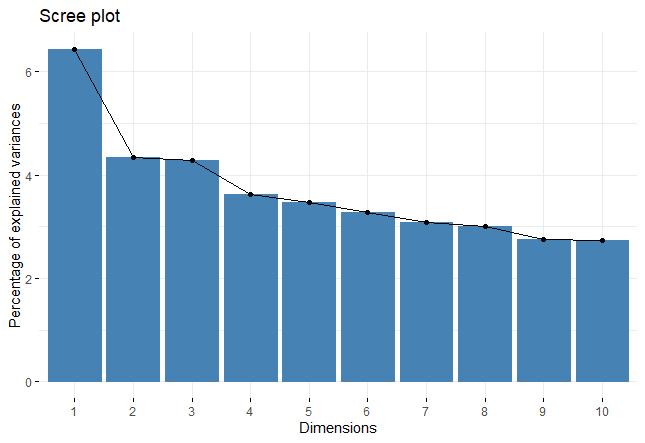


Next, I ran a principal component analysis to determine the most important variables.

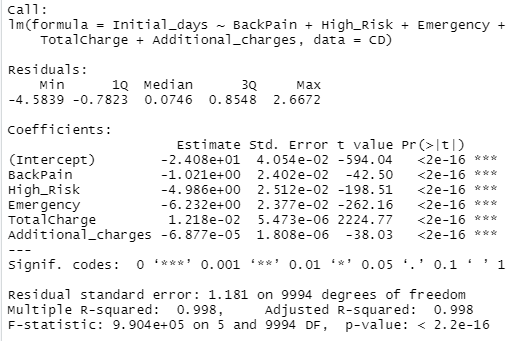
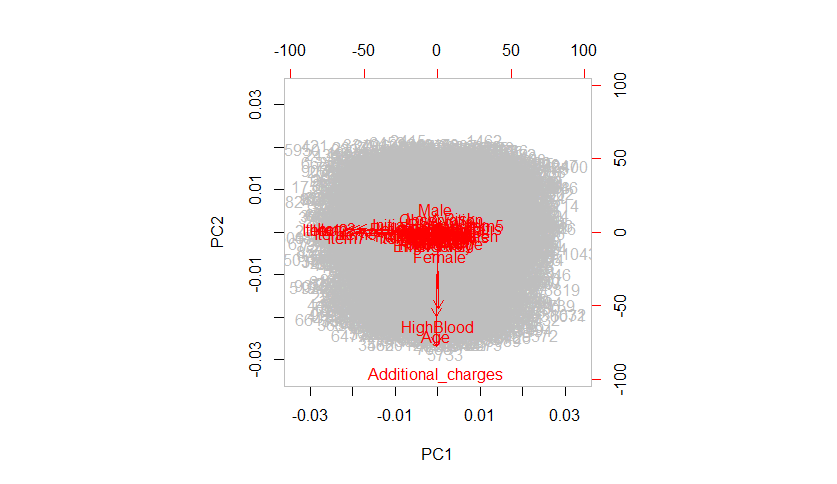
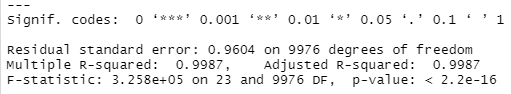
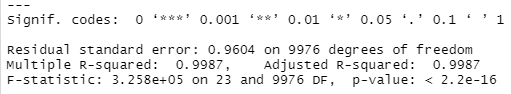
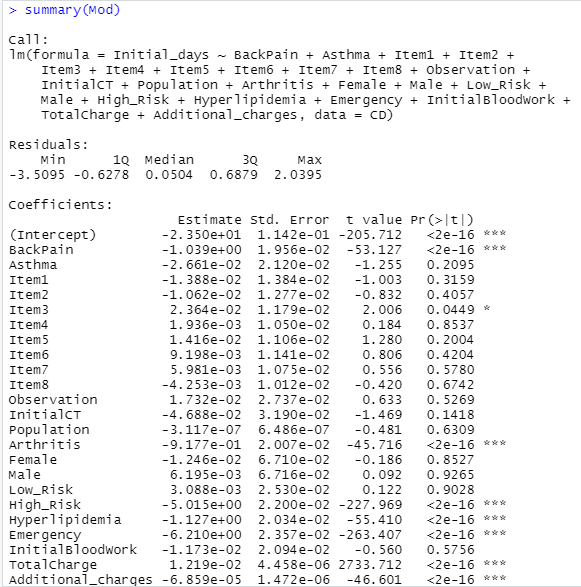


22 of the primary components had eigenvalues above 1 but it takes 29 to get a cumulative variance around 0.8. To see which variables correspond to the primary components.





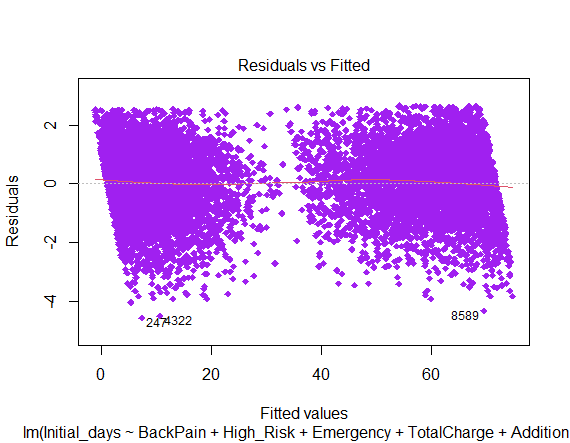
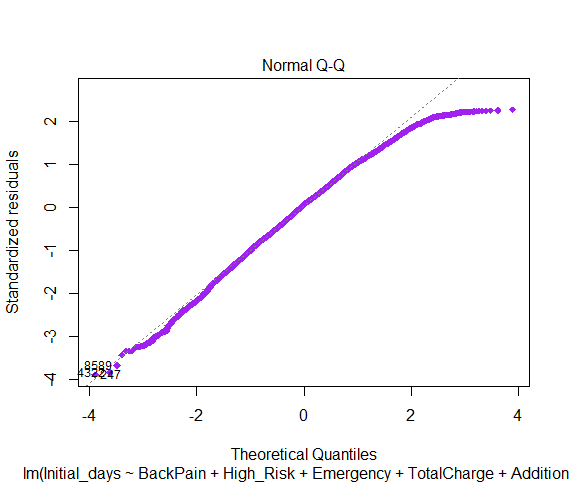
I created a model with those 22 variables and then created a biplot to see correlation between variables. Those variables that were highly correlated were dropped. That left me with 5 independent variables that I created my final reduced equation from: back pain, high risk, emergency admit, total charge, and additional charges.



**Initial\_days** ≈ **-0.2408 – 1.021 X1 – 4.986 X2 – 6.343 X3 + 0.01218 X4 – 0.00006877 X5**

(Where X1=Back Pain, X2=High Risk, X3=Emergency Admit, X4=Total Charge, X5=Additional Charges)

The graphs of the residuals showed that the residulas were varied and were approximately normal. This final reduced equation met the assumptions and had aceptable R2 .

**Conclusion**

This analysis can help provide an estimate for how many days a patient will be hospitalized. If a patient is requiring a higher dollar amount average per day, the expected initial stay will be longer. There are conditions that have greater significance than others on hospital stay: high blood pressure, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, and reflux esophagitis. There are a number of limitations of this analysis. The average daily charges value will be constantly changing. One really expensive medical procedure and the average will be altered significantly. This model would be most useful for a patient who has already been in the hospital quite a few days and has undergone tests to get a diagnosis and treatment plan so that the average charges could be more accurate. At that point, it could help with an idea of how much longer the patient can expect to be in the hospital.