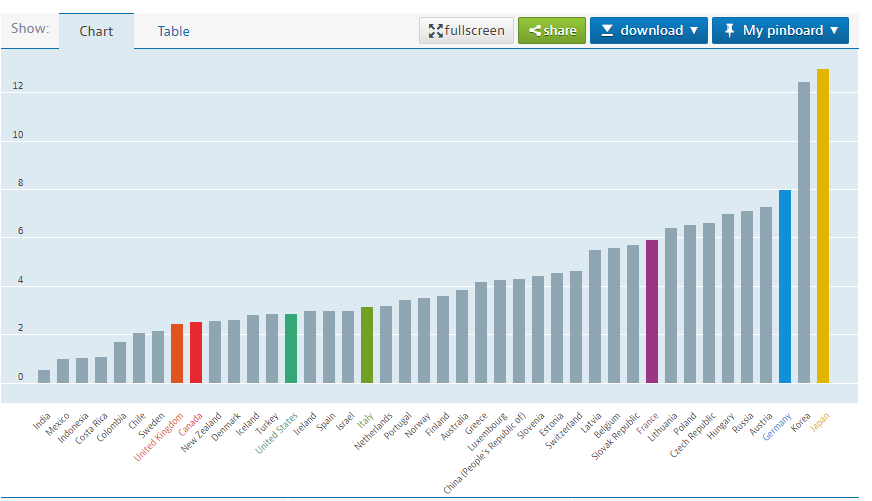
1. Abstract and Introduction

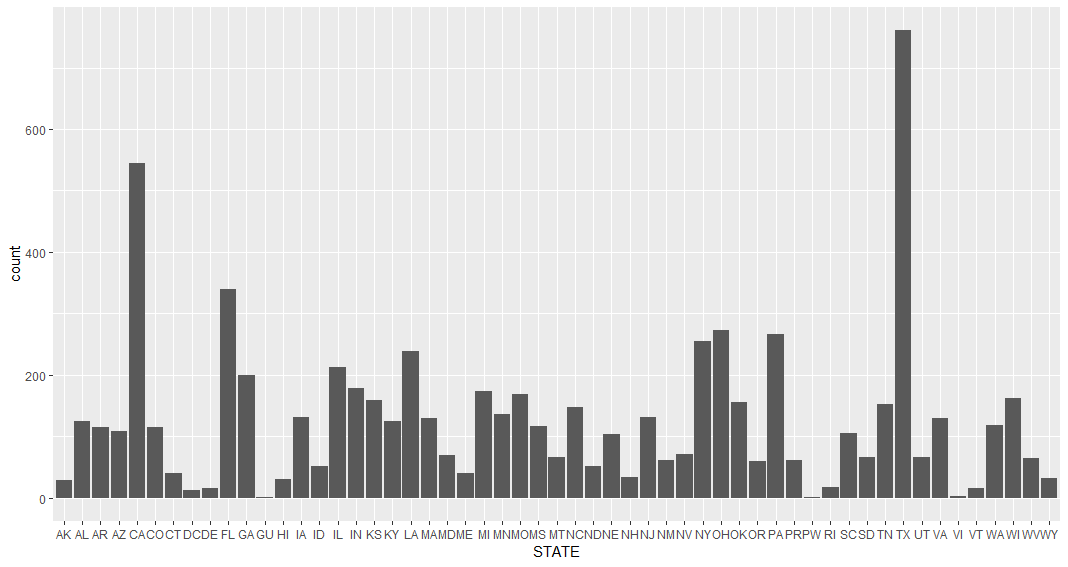
Data Analytics is something that has come to the forefront of the business world in recent years. Collecting data and analyzing it has proved beneficial for all sorts things including the health sector. One of the things that I have always been curious about is how people decide where hospitals should be built and how big they should be made. This motivated me to look at hospitals in the United States to see how they are spread out, how many people they serve and if there were any trends or correlations between area population and number of hospitals. Using hospital data for the US I wanted to investigate if there were counties and states in the US that were underrepresented in terms of numbers of hospitals compared to population size. My initial thoughts or hypothesis is that we see a high density of hospitals on the US coasts and the Southern border with density decreasing as you approach central US and with the lowest density being in the mid North. This would be states like South and North Dakota and Montana. Another interesting thing to see would be to see if areas in the United States could add variety to the types of hospitals that they have. This could meaning adding more psychiatric hospitals or acute care hospitals as an example. Creating an even distribution of hospitals would allow a wider population have access to the care that they will need.

1. Data Description and Exploratory Data Analytics

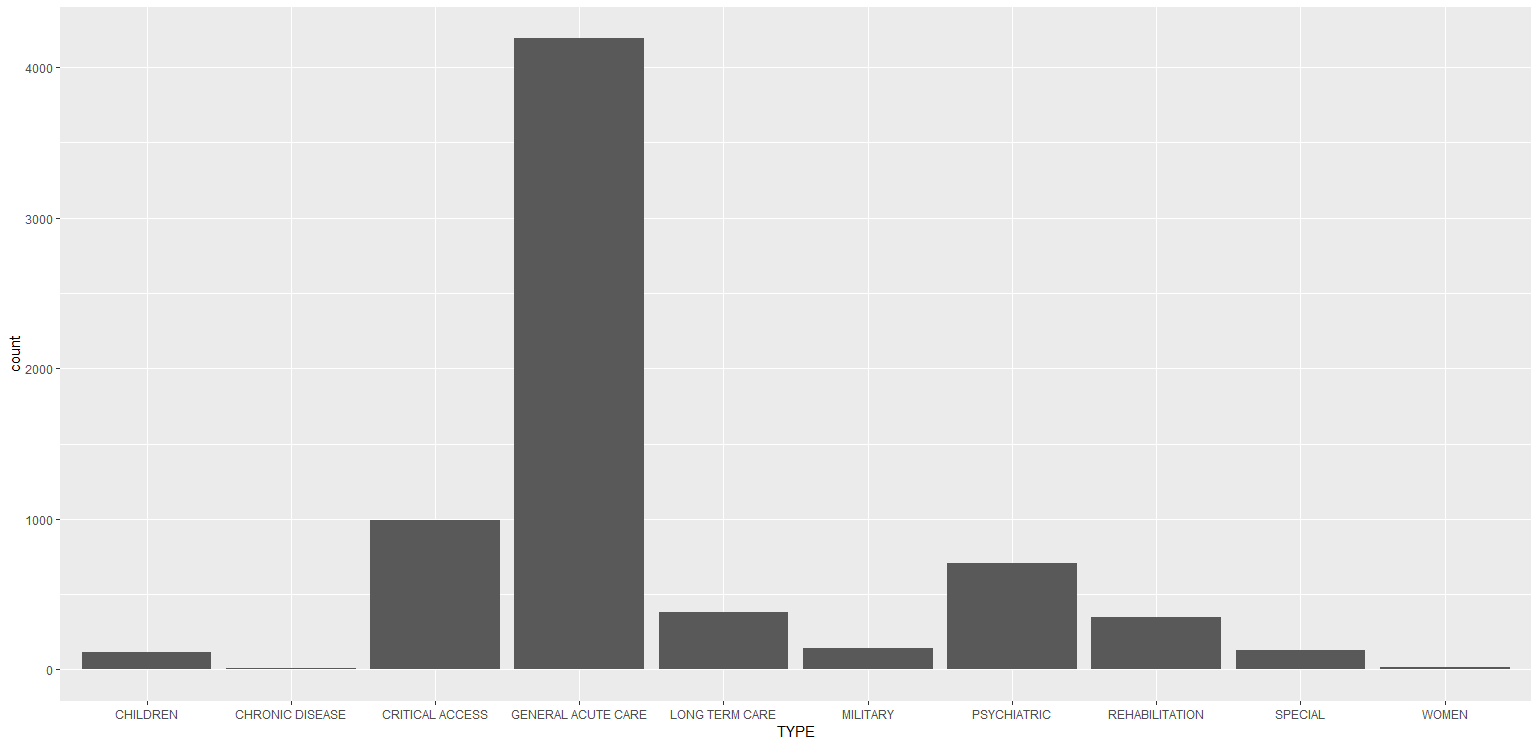
When looking for the data that would help me I knew I would have to find a dataset with hospitals and the data surrounding them. For the hospital dataset the things that I was looking for was a dataset that provided hospital location, number of patients that a hospital could hold(aka number of beds in the hospital), what type of service the hospitals specialized in, and potentially the number of staff the hospital employed. This would allow me to observe how and if hospitals in a certain region could properly supply care to the population of that region. Another thing that I looked for in the dataset was that it came from a reliable source. This meant that the source could show where and how the data for the dataset was collected. If it didn’t collect the data personally it could at least show the sources where the data was collected from. With this criteria Professor Thilanka provided the Homeland Infrastructure Foundation-Level Data Hospital Dataset. This dataset provides hospital data for all 50 states and even some US territories such as Puerto Rico and Guam. This dataset provided supplied data about hospitals such as number of beds, state, county, and hospital type. There are 10 different types of hospitals ranging from children to military. The dataset came in a csv file that could be downloaded from the internet. The hospital dataset fit perfectly into the criteria that I had but was missing the amount of staff, something that I believed I could do without. Other than the hospital dataset I needed to find a population dataset that was divided up by county. This dataset had to have all the counties in the United States and the most up to date population numbers for those counties. The source for it had to have similar criteria as the hospital dataset. The source had to have a reliable collection method along with being from an official group. When thinking about population numbers I recognized that the best place to start looking was the US Census. When it comes to population numbers in the United States the US Census Bureau was always the one collecting data and the source for most population statistics. In order to find the most up to date county population data I accessed the US Census Bureau which had county population totals for years 2010 to 2019. The populations were estimates as a census in the United States is conducted every 10 years with the last census taken in 2010. Even though the data is an estimation it will be the closest to the correct data than any other source. Lastly I had to find if and what the global hospital beds per population indicator was. This would allow me to compare the results in the United States to the global average or the average of other countries. This data would have to contain hospital beds per population for many different countries in the most recent year. When looking for the data I was able to find it at OECD.org (Organization for Economic Co-operation and Development) a organization that works to build policies worldwide, setting international standards using data and analysis. Below leading nations hospital beds per 1000 population can be seen. 

1. Analysis and Exploratory Data Analytics

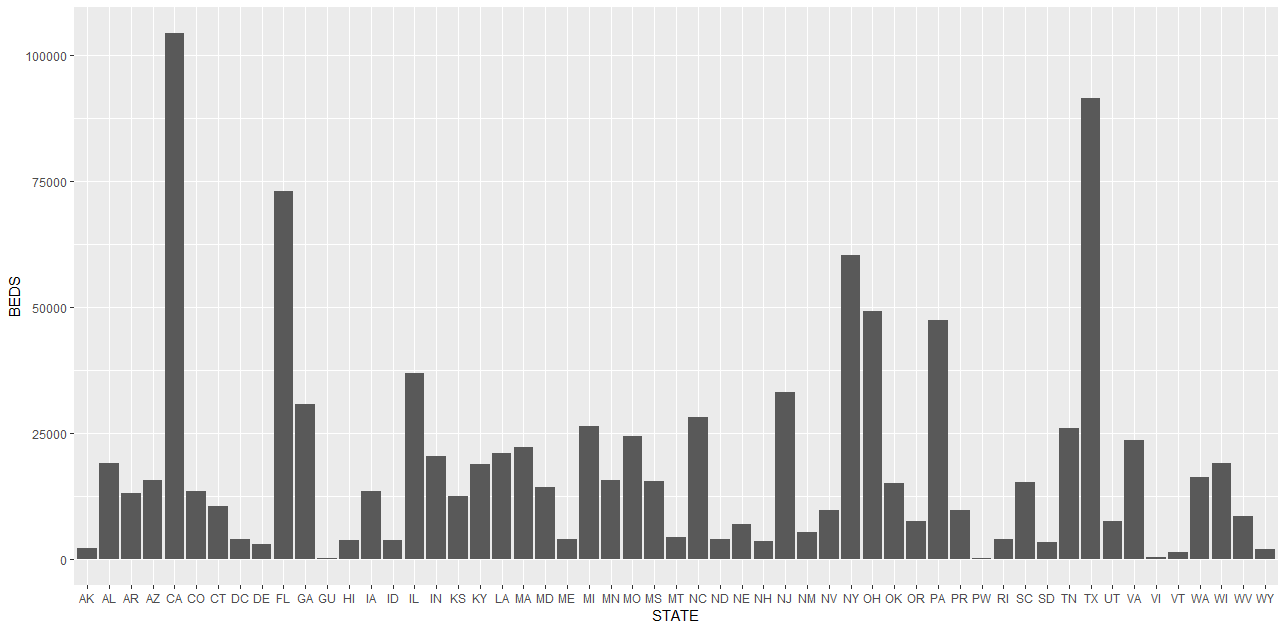
Before being able to start working on the data I had to perform transformations that changed the two datasets that I would be working with. The first dataset that I transformed was the Hospital dataset. The first thing that I noticed when looking over the data was that there was a lot of data that I wasn’t going to need, things such as zip codes, source website, and helipad information. I decided to reduce the dataset to a smaller version to help with how quick models could be created, so I created a subset that included ObjectID, State, Type, County, Country, Beds and Status. After this I also had to remove hospitals that were no longer opened, status was closed, and Hospitals that didn’t have data for Hospital beds. This gave me a dataset that would allow me to run my models in a more efficient manner with only the specific data I needed. For the County dataset I had to do a couple extra transformations in order to prepare it for EDA and model development. As I was only interested in Counties and their population in 2019, the most recent year, I removed all the other years (2010-18) and some descriptions at the bottom of the data. This allowed me to have a dataset with counties in the first column, states in the second and their population in the third column. With the datasets transformed into easier to work with datasets I started to perform EDA to observe more in depth the variables.



The bar graph above represents the number of hospitals per state. One of the generalizations that I had was that there would be larger amounts of hospitals on the coasts and in the south as those have the highest concentrations of people in the United States. When looking at this graph we see that Texas and California, the two biggest states, have the largest number of hospitals. This is followed by Florida, which supports the argument that the higher the population the higher the number of hospitals. This is because those 3 states have the three largest populations in the United States. We can then see that states with smaller populations tend to have lower numbers of hospitals.



When looking at all the different types of hospitals in the United States we can see that most of the hospitals are General Acute Care hospitals. As most hospitals don’t serve just a certain section of medicine but try to solve most medical issues this makes sense. These General Acute Care hospitals tend to take patients and try to treat them before sending them to more particular hospitals for issues that specific hospitals serve better.

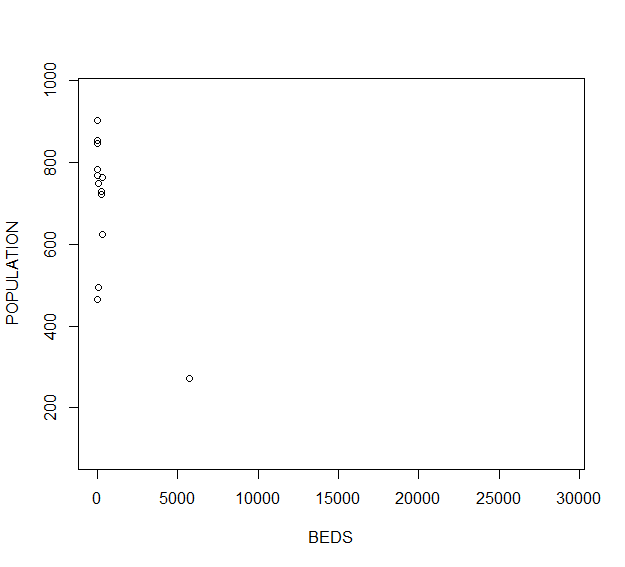


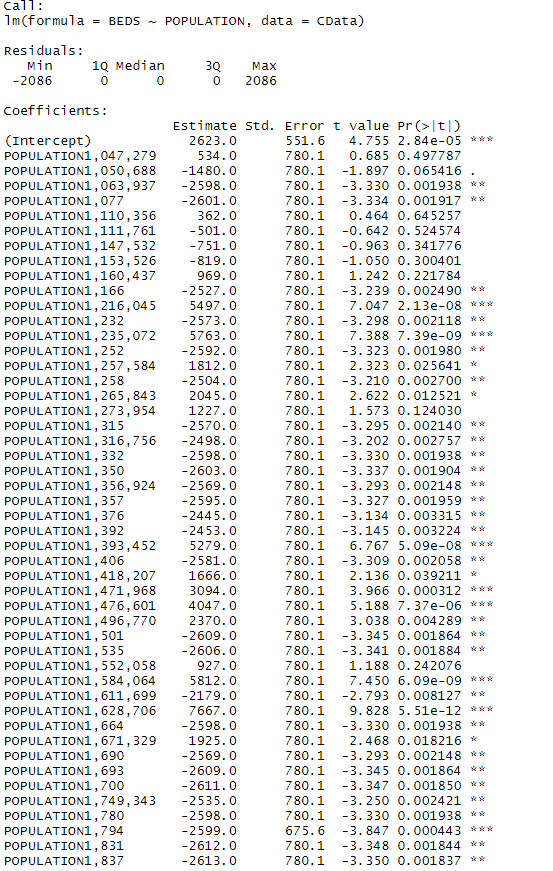
Next I wanted to take a look at the number of beds per state so I created a bar plot for beds per state and this follows the trend that was seen in hospital per state. The more hospitals there are in a state the more beds they will have. One thing that will be interesting to look at as well will be if some of the smaller states have more beds per hospital or if beds per hospital averages tend to stay the same between all the different states.

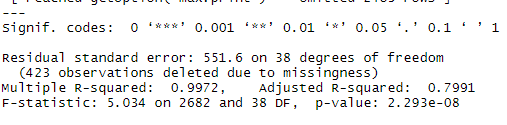
When looking at the data I came across a couple of potential sources of error that I might have to deal with. One of those might be the location of counties, this could skew data and show that certain counties don’t have the same hospital concentration when there could be a hospital in the county over. Another bias that could arise from data I didn’t use is the average distance from hospital to hospital. If hospitals are close to each other they tend to serve the same population, which doesn’t have the same impact as hospitals which are evenly spread out through a county or state.

1. MODEL DEVELOPMENT and APPLICATION OF MODELS

The first model that I decided to develop was a linear regression model. I wanted to be able to see the correlation between Population of a county and the number of beds that were available in hospitals. One of the things that helped me decide to use linear regression was when looking at the bar graphs from the EDA. I saw that there seemed to be a trend where larger population states had higher number of hospitals and beds. This led me to assume that there was a linear regression model that might fit the data. We can then look at the summary for the results of the linear regression model.







When looking at the summary statistics for the linear regression model we can notice some interesting things. The first thing that we can look at is the Residual standard error. We can see that it is high at 551.6 and this is supported with the R-squared statistics. This means that the linear regression model did not fit well. This is also supported by the F-Statistic being relatively high, meaning that there is a weak to no correlation between beds and population. We also see that the p-value is small causing us to reject that there is a correlation. When looking at the regression model we see that all the statistical indicators are pointing to no correlation between population and beds.

For the next model that I developed I thought that there could be a classification model to be created using the helipad as the classification. Thinking that there could be a correlation between the numbers of bed a hospital has, county population and if it has a helipad. Since there are only two classes for the Helipad, yes or no, I decided to create a KNN classification model to test my hypothesis.

I realized that I made a mistake when creating the data and didn’t leave enough to create a second model. I tried to create a Svm model and a KNN classification model but both didn’t seem to work and kept throwing errors as I was trying to create the models. With these problems I ran out of time and decided to work on the conclusion instead.

1. CONCLUSION

For a conclusion I came to the conclusion that there isn’t a real correlation between hospitals, beds and population of the area where they are located. It seems that there is a loose correlation where there are more hospitals in areas where the population is higher but it doesn’t seem to be correlated to the county that the hospital is located in. I had predicted that this was closely related but I mistook the power of money in the hospital industry. Hospitals are large money makers, especially private hospitals, which can lead to more hospitals being overcrowded in certain areas rather than others.