IST 652: Scripting for Data Science

Hospital Locality

Professor: Dr. Landowski

Student Name: Lan Tran

Table of Contents:

Introduction:

The start of 2020 brought forth the first case of COVID-19 to the U.S. In the following months, confirmed cases climbed tremendously. In an effort to curve the affect toward a manageable level, strict implementation of social distancing and shelter in place were put into effect. Even with the push to slow the spread of the virus, outbreaks still push through across the country. America's hospital and health system have stepped up during these harsh times by ramping up testing efforts, establishing testing tents, adding general intensive care unit (ICU) bed capacity, and developing COVID-19 units to isolate and treat patients with disease while safeguarding the health of other patients and hospital staffs.

Across the globe, hospital systems were put to their maximum capacity. Turning away those who show symptoms but not severe enough to warrant services in the already over capacitated hospital setting. With spotlights on the healthcare system and whether or not it is able to handle pandemic type of events, many head to social platform to voice their unrest. In an attempt to further expand the understanding of the hospital system. For the final project, data sets from various sources are gathered and analyzed for the purpose of understanding hospital locality and whether there are certain attributes that constitute where a hospital is placed in the United States.

Background: Based on the Center for Medicare & Medicaid Services (CMS) definition: A hospital is an institution primarily engaged in providing, by or under the supervision of physicians, inpatient diagnostic and therapeutic services or rehabilitation services. Critical access hospitals are certified under separate standards. Psychiatric hospitals are subject to additional regulations beyond basic hospital conditions of participation. The State Survey Agency evaluates and certifies each participating hospital as a whole for compliance with the Medicare requirements and certifies it as a single provider institution.

Business Questions:

1. How accessible are the hospital for those who are of low income or elderly?
2. Are there specific factors/attributes that contribute to the hospital locality?
3. Are the hospital beds sufficient for the population at hand? If not, what is gap between actual and benchmark?
4. What are the current hospital ratings and what specifications comprise of the ratings?

Pre-Processing:

Data Definition

Hospital General (HIFLD) Dataset:

Notes:

- Data Frame = hosp_df
- Updated as of 12/8/2020
- Source: https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals/data

#	FIELD NAME	DATA TYPE	DATA DESCRIPTION
0	OBJECTID	INDEX	Unique ID
1	NAME	OBJECT	Hospital Name
2	CITY	OBJECT	Hospital City Name
3	STATE	OBJECT	Hospital State Two-letter abbreviation including Washington D.C. , US territories of Puerto Rico, Guam, American Samoa, Northern Mariana
4	TYPE	OBJECT	Type of Hospital: General Acute Care, Critical Access, Psychiatric, Long Term Care, Rehabilitation, Military, Special, Children, Women,
5	STATUS	OBJECT	Open or Closed
6	LATITUDE	FLOAT64	Hospital Latitude coordinates
7	LONGITUDE	FLOAT64	Hospital Longitude coordinates
8	NAICS_CODE	INT64	North american Industry Classification System:
9	NAICS_DESC	OBJECT	Hospital Description base on the North american Industry
10	ST_FIPS	INT64	Two-letter alphabetic codes : Federal Information Processing Standard State Code
11	BEDS	INT64	Number of hospital beds available
12	TRAUMA	OBJECT	Different level of trauma units available in the hospital and whether
13	HELIPAD	OBJECT	Helipad availability = N or Y

Hospital CMS Rating:

Notes:

- Data Frame = rating_df
- Updated as of 12/10/2020
- source : https://data.cms.gov/provider-data/dataset/xubh-q36u
- Star rating description : https://data.cms.gov/provider-data/topics/hospitals/overall-hospital-quality-star-rating

#	FIELD NAME	DATA TYPE	DATA DESCRIPTION
0	Facility ID	INDEX	Unique ID
1	Facility Name	OBJECT	Hospital Name
2	Address	OBJECT	Hospital Address
3	City	OBJECT	Hospital City
4	State	OBJECT	Hospital State Two-letter abbreviation including Washington D.C. , US territories of Puerto Rico, Guam, American Samoa, Northern Mariana Island, Palau, And Virgin Islands.
5	ZIP Code	INT64	Hospital Zip Code
6	County Name	OBJECT	Hospital County Name
7	Hospital Type	OBJECT	Type of Hospital : Acute care Hospital, Critical Access Hospital,
8	Hospital Ownership	OBJECT	Ownership: Department of Defense, Government (federal,
9	Emergency Services	OBJECT	ER services : Yes or No?
10	Meets criteria for promoting interoperability of EHRs	OBJECT	Meets criteria for promoting interoperability of Electronic Health Records (Yes or blank = No)
11	Hospital overall rating	Original = OBJECT Converted = Float	Hospital Rating (1 through 5 or not available) base on 7 groups of measures: 1) Mortality (22% weight) 2) Safety or Care (22% weight) 3) Readmission (22% weight) 4) Patient Experience (22% weight) 5) Effectiveness of Care (4% weight) 6) Timeliness of Care (4% weight) 7) Efficient Use of Medical Imaging (4% weight)
12	Mortality national comparison	OBJECT	Mortality national comparison
13	Safety of care national comparison	OBJECT	Safety of care national comparison
14	Readmission national comparison	OBJECT	Readmission national comparison
15	Patient experience national comparison	OBJECT	Patient experience national comparison
16	Effectiveness of care national comparison	OBJECT	Effectiveness of care national comparison
17	Timeliness of care national comparison	OBJECT	Timeliness of care national comparison
18	Efficient use of medical imaging national comparison	OBJECT	Efficient use of medical imaging national comparison

Medicaid & CHIP enrollment:

Notes:

- Data Frame = medicaid_df
- Updated as of 9/1/2020
- source : https://data.medicaid.gov/Enrollment/2020-09-Preliminary-applications-eligibility-deter/s6d2-cpd9/data

#	FIELD NAME	DATA TYPE	DATA DESCRIPTION
0	State Abbreviation	OBJECT	State Abbreviation
1	State Name	OBJECT	State Name
2	State Expanded Medicaid	OBJECT	State Expanded Medicaid (Y or N)
3	New Applications Submitted to Medicaid and CHIP Agencies	FLOAT64	New Applications Submitted to Medicaid and CHIP Agencies
4	Total Applications for Financial Assistance Submitted at State Level	FLOAT64	Total Applications for Financial Assistance Submitted at State Level
5	Medicaid and CHIP Child Enrollment	FLOAT64	Medicaid and CHIP Child Enrollment
6	Latitude	FLOAT64	Latitude
7	Longitude	FLOAT64	Longitude
8	Total Medicaid Enrollment	INT64	Total Medicaid Enrollment

Median Income by States:

Notes:

- Data Frame = income_df
- Updated as of 9/1/2020
- source : https://www2.census.gov/programs-surveys/cps/tables/time-series/historical-income-households/h08.xlsx

#	FIELD NAME	DATA TYPE	DATA DESCRIPTION
0	State	OBJECT	State Full Name including United States
1	Code	OBJECT	State Abbreviation
2	Median income	INT64	Median Income

#	FIELD NAME	Data Source	Aggregate Function
0	State	All source	Group by
1	RATING	rating_df : Hospital overall rating	Average
2	CHILDREN	hosp_df : [TYPE]	categorical - Count
3	CHRONIC DISEASE	hosp_df : [TYPE]	categorical - Count
4	CRITICAL ACCESS	hosp_df : [TYPE]	categorical - Count
5	GENERAL ACUTE CARE	hosp_df : [TYPE]	categorical - Count
6	LONG TERM CARE	hosp_df : [TYPE]	categorical - Count
7	MILITARY	hosp_df : [TYPE]	categorical - Count
8	PSYCHIATRIC	hosp_df : [TYPE]	categorical - Count
9	REHABILITATION	hosp_df : [TYPE]	categorical - Count
10	SPECIAL	hosp_df : [TYPE]	categorical - Count
11	WOMEN	hosp_df : [TYPE]	categorical - Count
12	BEDS	hosp_df : [BEDS]	sum
13	MEDICAID	medicaid_df : Total Medicaid Enrollment	sum
14	MEDIAN INCOME	income_df : median income	lookup
15	2020 POPULATION	demographic_df : POPESTIMATE2019	sum
16	HOSP_CNT	hosp_df : OBJECTID	count
17	M_0-10	Census Demographic data	sum (hospital Count) for Male ages 0-10
18	M_11-20	Census Demographic data	sum (hospital Count) for Male ages 11-20
19	M_21-30	Census Demographic data	sum (hospital Count) for Male ages 20-30
20	M_30-40	Census Demographic data	sum (hospital Count) for Male ages 30-40
21	M_40-50	Census Demographic data	sum (hospital Count) for Male ages 40-50
22	M_50-60	Census Demographic data	sum (hospital Count) for Male ages 50-60
23	M_60-65	Census Demographic data	sum (hospital Count) for Male ages 60-65
24	M_65+	Census Demographic data	sum (hospital Count) for Male ages 65 and older
25	F_0-10	Census Demographic data	sum (hospital Count) for Female ages 0-10
26	F_11-20	Census Demographic data	sum (hospital Count) for Female ages 11-20
27	F_21-30	Census Demographic data	sum (hospital Count) for Female ages 20-30
28	F_30-40	Census Demographic data	sum (hospital Count) for Female ages 30-40
29	F_40-50	Census Demographic data	sum (hospital Count) for Female ages 40-50
30	F_50-60	Census Demographic data	sum (hospital Count) for Female ages 50-60
31	F_60-65	Census Demographic data	sum (hospital Count) for Female ages 60-65
32	F_65+	Census Demographic data	sum (hospital Count) for Female ages 65 and older
33	beds/k	Derived : combined_df	(beds/2020 Population)*1000

Data Loading:

Preparing the data:

Before data can be loaded into the python environment, one must understand the data that is being provided. For the final project, there are a total of five different data source which contains sizes (columns and rows). As with many collected data, not all are relevant and are then omitted form the imported fields. In order to be able to compare the different dataset to one another, there should be a commonality amongst the datasets. For this instance, the "State's name" are the common fields of the five different sources. Since the States name are not uniform between the different dataset, a derived "States" abbreviation column is added to the datasets with full name.

- Determines the relevant fields name to be imported.
- Find out the year for which the data are representing:
  - Median income = most relevant information is 2019 data.
  - Hospital = ranges from 2012 through 2020, updated as of 12/8/2020. Since hospital are not easily built and demolish, this should be okay to use.
  - The rest of the dataset presents 2020 data.
- Create a consolidated file with "states" as a point of reference/lookup value. Since the dataset is quite small, it is easier to create the consolidated file using excel with lookups and sumifs functions.

Importing the data:

The various datasets are imported using pandas with visualization using seaborn, matplotlib, and plotly. To use Plotly for mapping of the US states, credentials need to be obtained before one can fully use Plotly for visualization.

```
#-------------------------------------#
# import packages:
#-------------------------------------#
import os
import pandas as pd
import numpy as np
import seaborn as sns


import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
from matplotlib import pyplot
import matplotlib
plt.style.use('ggplot')
from matplotlib.pyplot import figure

import chart_studio.plotly as py
import chart_studio.tools as tls
import plotly.graph_objects as go
```

```
#-------------------------------------#
# inline plotting:
#-------------------------------------#

# for plotting inline for plotly you need to download
# conda install -c plotly chart-studio and get credentials & api key to display
# https://plotly.com/python/renderers/

tls.set_credentials_file(username='Latran', api_key='7j4HLSxBeKDzdm2Bt5jW')

#to plot within jupyter notebook
%matplotlib inline
```

Data Cleaning/Transformation:

After loading the data into pandas Data Frame, look through the dataset to make sure all is loaded correctly. Convert, transform the data so you can aggregate, visualize and compare the datasets. Below are things that were done to clean/transform the data. This step is more fluid as you begin to explore the data, you might find out later on during the visualization/aggregation process that certain fields need to be converted in order to use specific functions.

- Change any header fields that are too long or not intuitively comprehendible from the rating and Medicaid data frames.
- There are two fields that have missing rows, which is okay as those fields are not being used for analytical purposes at the moment.
    - New Applications Submitted to Medicaid and CHIP Agencies - 4%
    - Meets criteria for promoting interoperability of EHRs - 29%
- Using the d.types to determine the datatypes related to the fields imported. Convert the rating field to a float as it is a continuous attribute and should be numeric instead of an object.
- Create a new column called “beds/k” to capture the beds per 1,000 inhabitants as a benchmark set out by the World Health Organization (WHO) as a way to determine the number of beds needed for every 1,000 people. As set out by the WHO, 3 beds per 1,000 should be a benchmark, which I’ve created a new column to house this benchmark to create a bar/line graph overlay.

Analysis:

This portion of the project is to slice and dice the dataset in different ways to see if there are patterns that can be recognizable and whether or not any conclusion can be drawn from it. This is the fun part of the project and often time the most time consuming as one can be consumed by the data and start to go down the rabbit hold of infinite possibilities. Understanding that this is one of my downfalls, there are many occasions where I had to look back at the question proposed to reel me back to reality. This is a constant struggle of course, but one that I try to polish through many more project such as this. In the following sections, several exploratory data analysis techniques will be shown.

EDA: Variable Visualization

Using the seaborn "seaborn.pairplot" on the combined data frame to see if there are any paired variables that shows any interesting relationship. As this function pairs the variables and graph it, similar to the correlation matrix, it is better to trim down the data Frame.

Findings:

- Median income compared to the rest of the other variable trend a little different.
- Ratings are in the range of 3-4 with not much variability. However, the Critical access and Median income are not as bunched up to the lower left quartile.

EDA: Data Exploration

Using "groupby" function so see the categorical groupings by count() or summation if it is a numeric variable.

Hospital Ownership	Facility Name
Department of Defense	35
Government - Federal	47
Government - Hospital District or Authority	536
Government - Local	420
Government - State	210
Physician	72
Proprietary	1043
Tribal	9
Voluntary non-profit - Church	321
Voluntary non-profit - Other	409
Voluntary non-profit - Private	2222

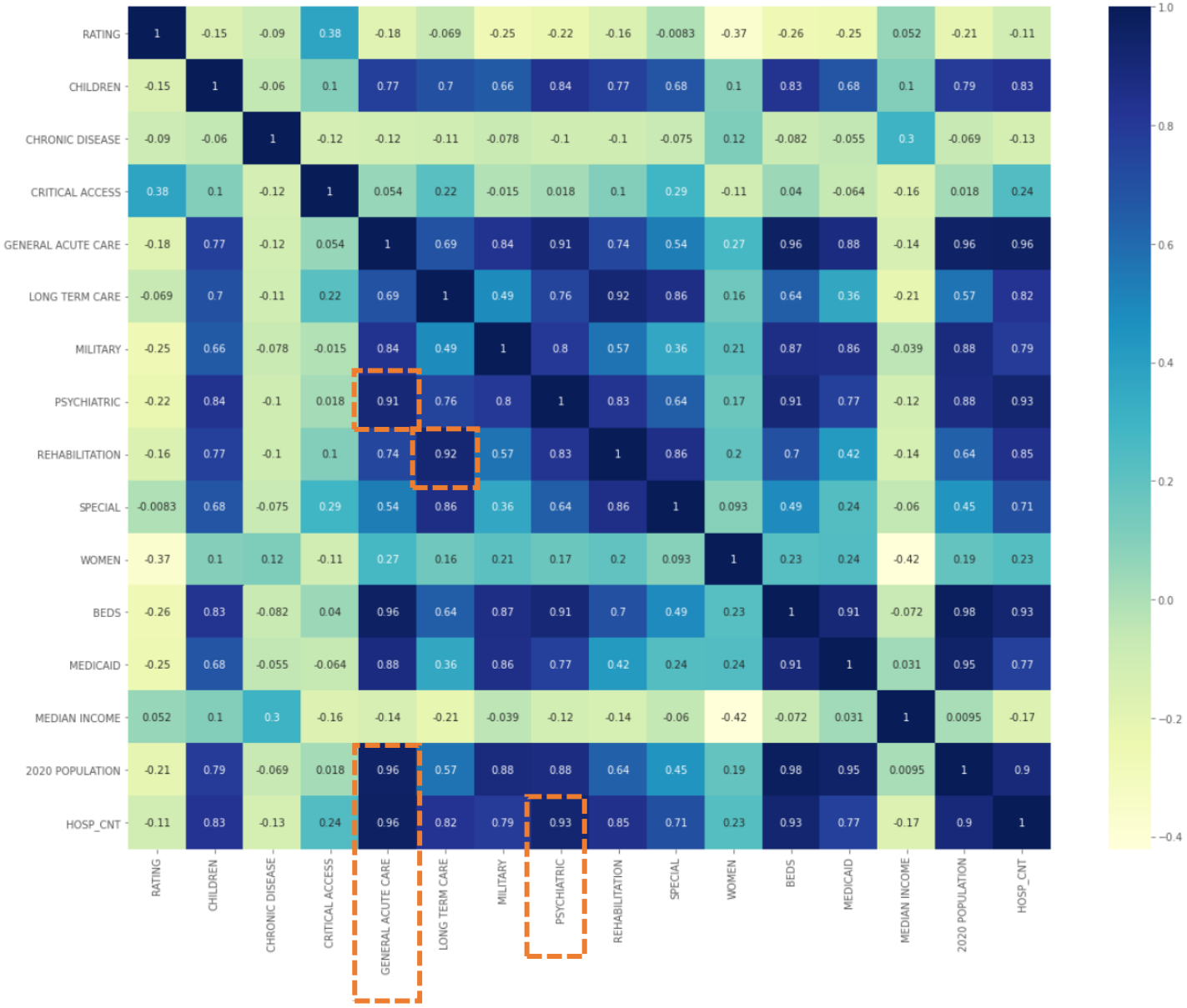
One of CMS Rating calls for timeliness and having a helipad would in turn help with that score.

Findings:

- 28% of the hospital has a trauma unit. Of the hospital that has a trauma unit, 86% have a helipad in place.

HELIPAD	OBJECTID
N	3419
Y	4177

EDA: Data Correlation



Findings:

- I wanted to look at correlation scores that are above 0.90 which shows General Acute care to population. This makes sense as the most people who goes into a hospital are more for acute type of services as oppose to more specialize treatments.
- Psychiatric is another correlation with hospital count as this specialize in mental health needs and illness using medication, psychotherapy and behavioral therapies. Some of these hospitals we can think of Alcohol/drug/substance abuse type of facilities.

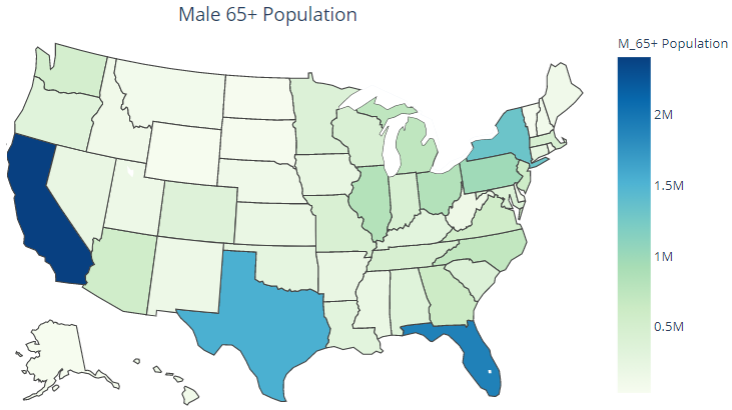
- Psychiatric correlates with General Acute care since most of the referral would come from either a patient self-admitting to a psychiatric facility or from a general hospital following some service rendered in the more often general acute hospital setting.
- Rehabilitation in correlation with Long term care make sense as most rehab takes a bit of time to overcome.

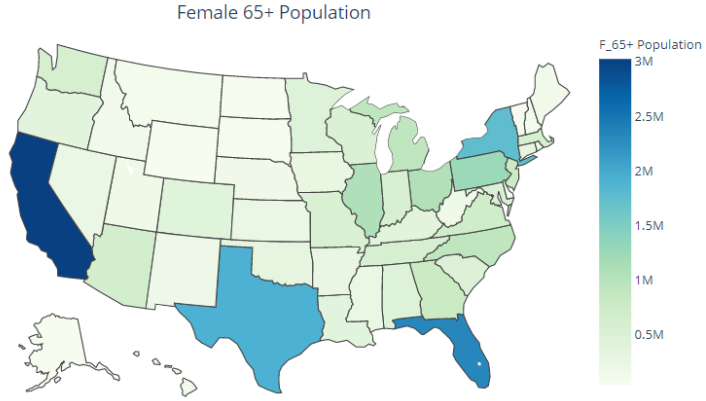
EDA: Data Visualization

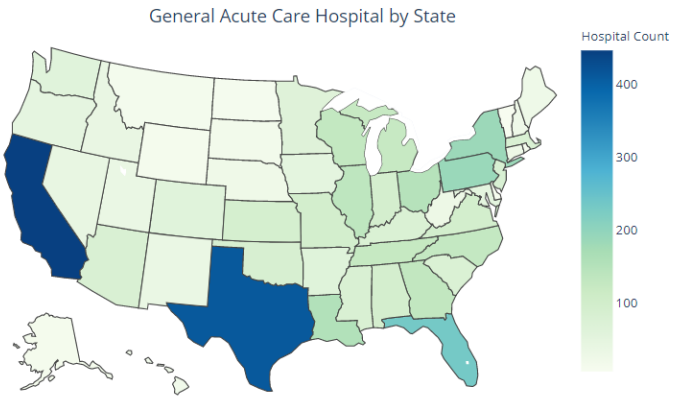
Another way to find patterns within your data set is to visualize it. Unless you work with tables and numbers all day, it is harder to find patterns within tables than with images. Python with its numerous packages are able to create images easily. The image below is created using Plotly. In actuality, the original thought was to use geopanda as it seemed to be easier to map, but due to limitation on time, plotly was use instead.

Q: How accessible are the hospital for those who are of low income or elderly?

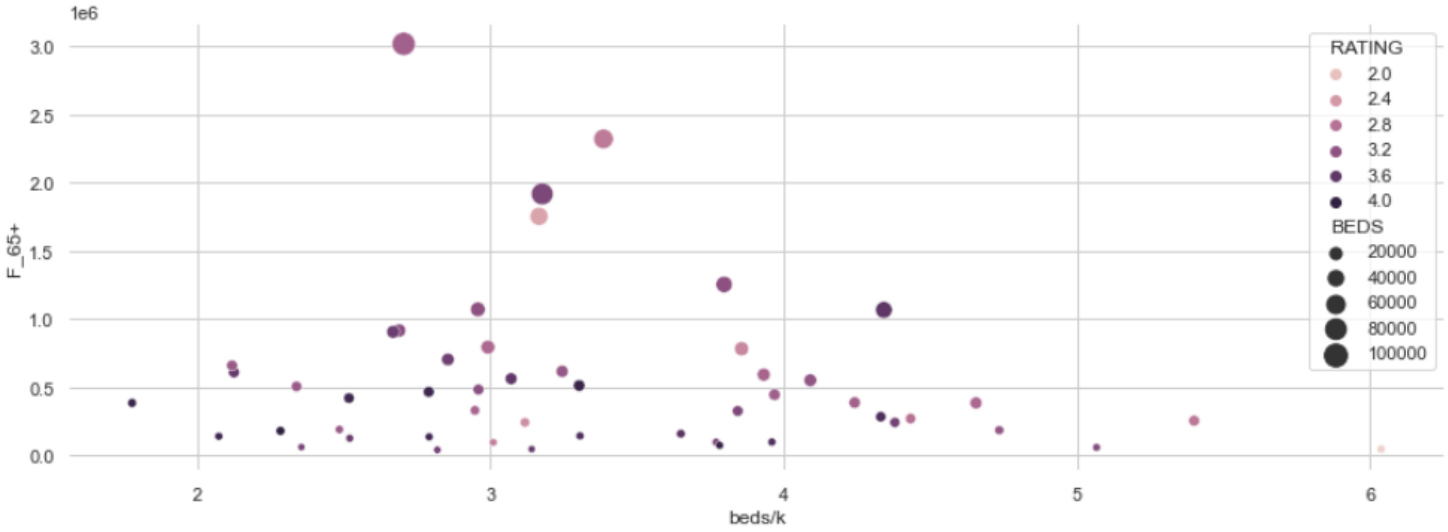
From the density of where the older population >65 resides, hospital in those states states looks to be more closely align. However, one can argue that the distribution of hospital in Florida and New York for the older groups are more lacking according to the heat map below.

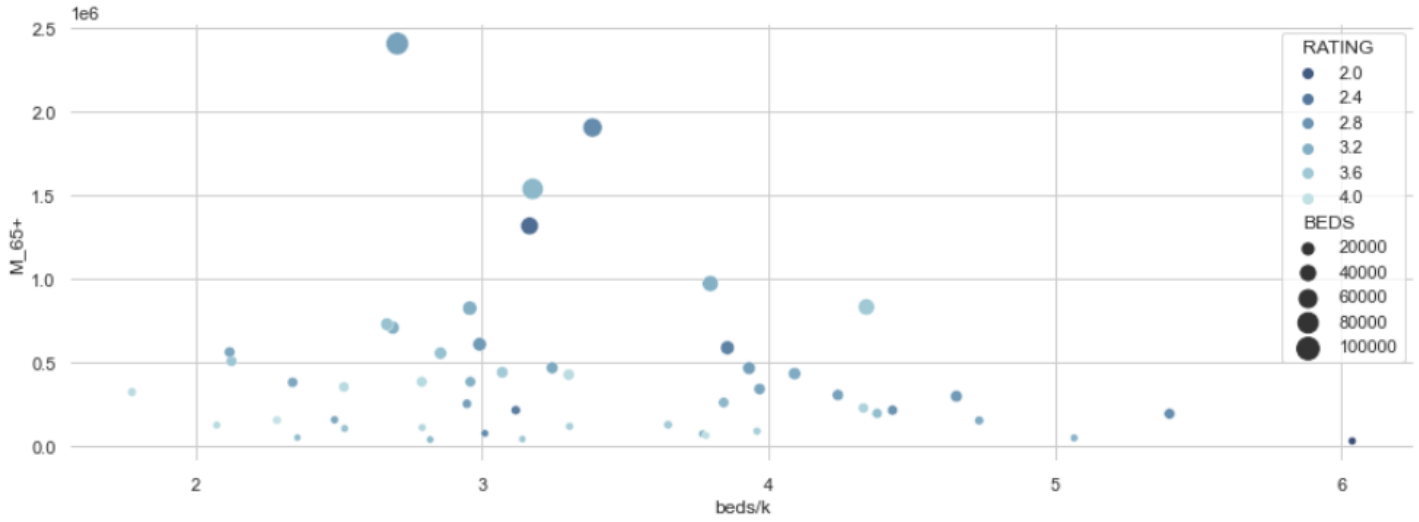






```
<AxesSubplot:xlabel='beds/k', ylabel='F_65+'>
```





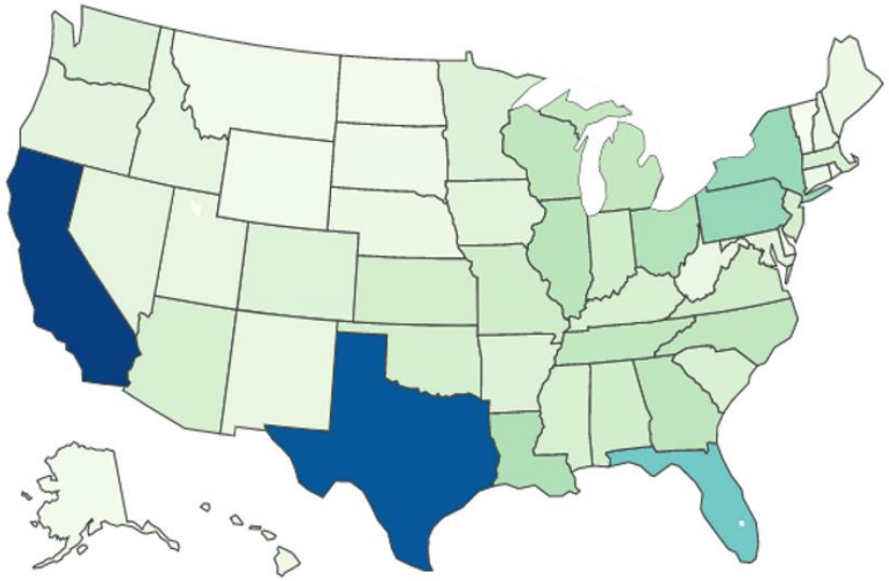
As the heatmap only shows so much, I've decided to create a scatter plot to determine the beds/k for Male and Female ages 65+. It does show that as the population grows, the more beds are available which is a comfort to see.

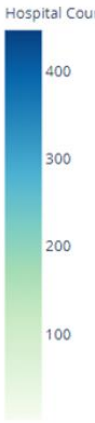
Q: Are there specific factors/attributes that contributes to hospital locality?

As hospitals are broken out by different classification or types, groupings may vary from one source to another. According to the Homeland Security dataset used for this assignment, below are hospital counts base on types. From the look of the heat map, it looks like hospital is correlated to the population density and needs base on the hospital types.

HOSPITAL TYPE : GENERAL ACUTE CARE

Provide short-term patient care

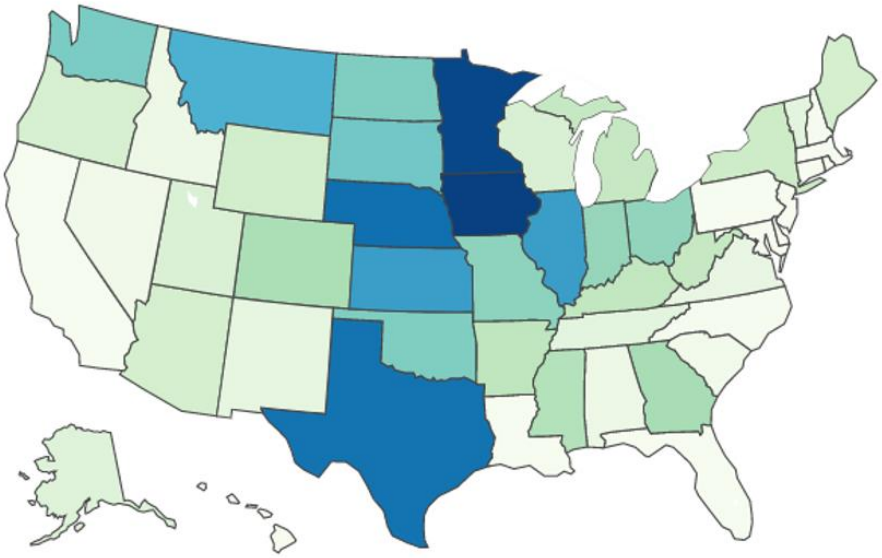


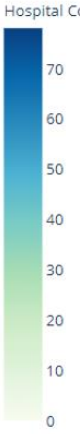


General Acute	State
1	CA
2	TX
3	FL
4	PA
5	NY
6	LA
7	OH
8	IL
9	GA
10	WI

HOSPITAL TYPE : CRITICAL ACCESS

Small facilities that give limited outpatient and inpatient hospital services to people in rural areas. These type of hospital normally have fewer than 25 beds

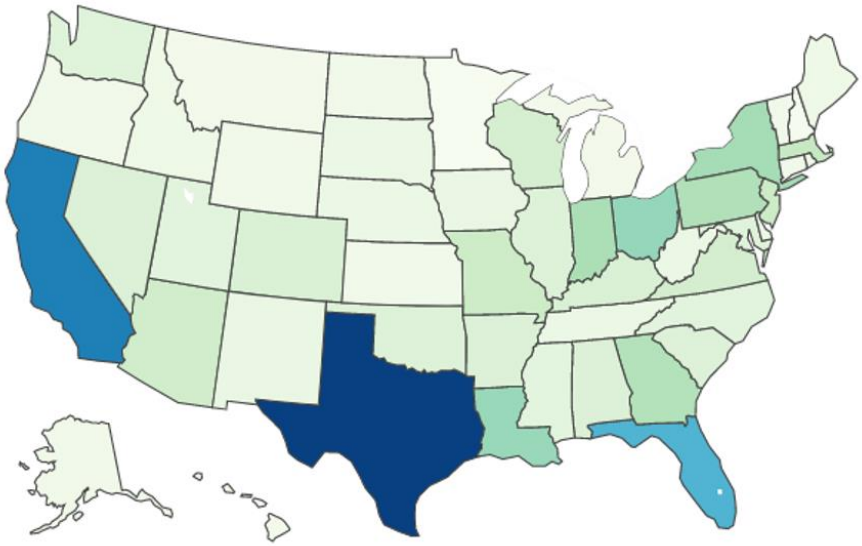


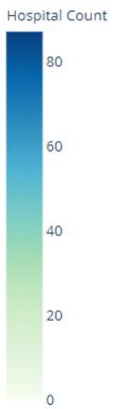


Critical Access	State
1	IA
2	MN
3	NE
4	TX
5	IL
6	KS
7	MT
8	SD
9	WA
10	ND

SPECIALTY HOSPITAL TYPE : PSYCHIATRIC

specialize in mental health needs and illness using medication, psychotherapy and behavioral therapies. Some focus on short-term, emergent treatment, while some guide patients through long-term stabilization.





Psychiatric	State
1	TX
2	CA
3	FL
4	OH
5	LA
6	NY
7	IN
8	PA
9	GA
10	MA

OTHER SPECIALTY HOSPITAL

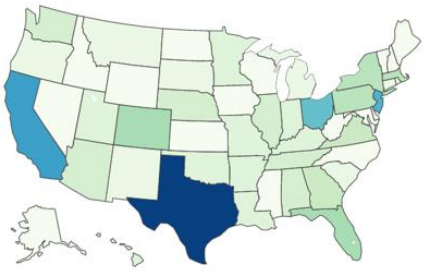
Long Term Care

Rehabilitation

Military

Special

Children

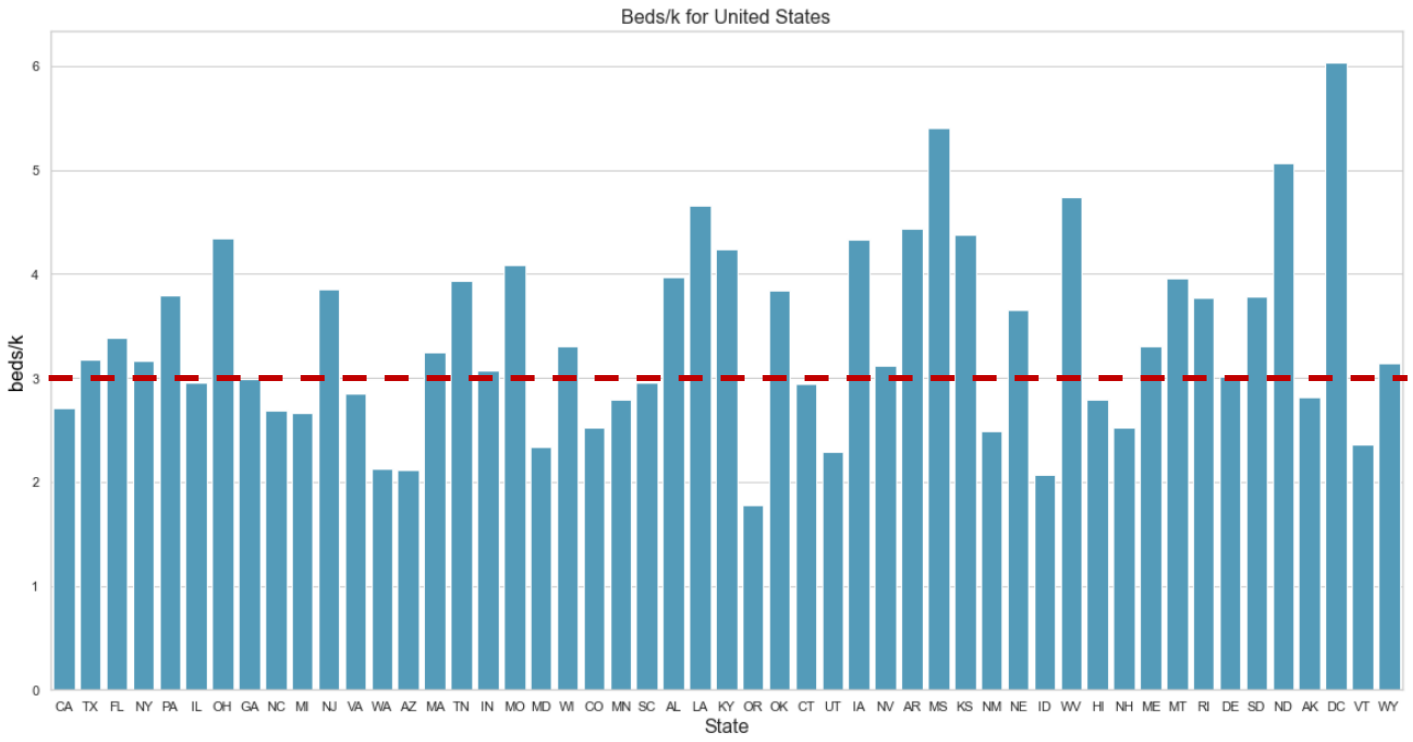


Women

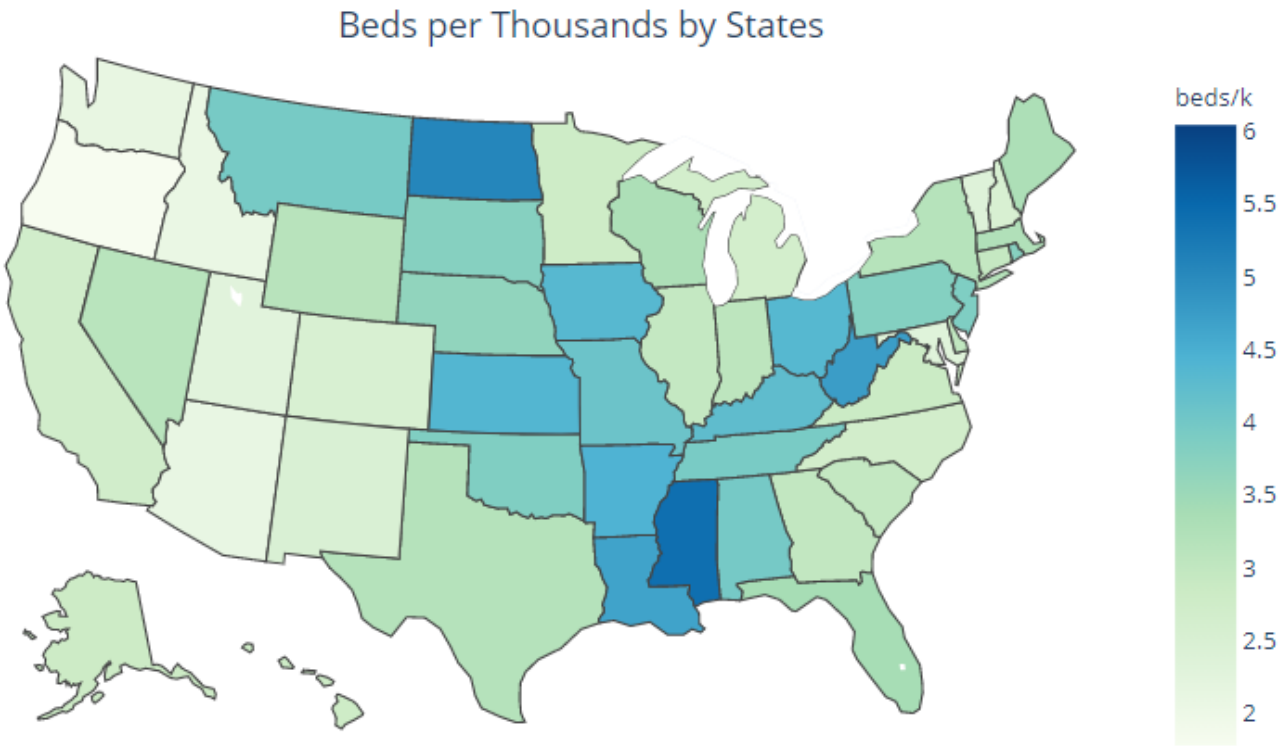
Q: Are the hospital beds sufficient for the population at hand? If not, what is gap between actual and benchmark?

According to the World Health Organization, hospital beds are used to indicate the availability of inpatient services. There is no global norm for the density of the hospital beds in relation to the total population. However, there is a consensus that 3 hospital beds for every 1,000 inhabitants is the sufficient amount designated by W.H.O.

Using Census data along with the Homeland security dataset for population and bed counts, currently the United States average beds per thousand lives is approximately 3. The Minimum beds per thousand is 2 coming from Oregon and the highest is 6.04 for DC (if you count that) , or 5.40 for Mississippi. California, with its high population has lower beds per thousand than Texas, Florida and New York interestingly enough.



From the heat map below, it looks to be that the middle to eastern coast of the United States has higher beds per thousand than the coastal states.

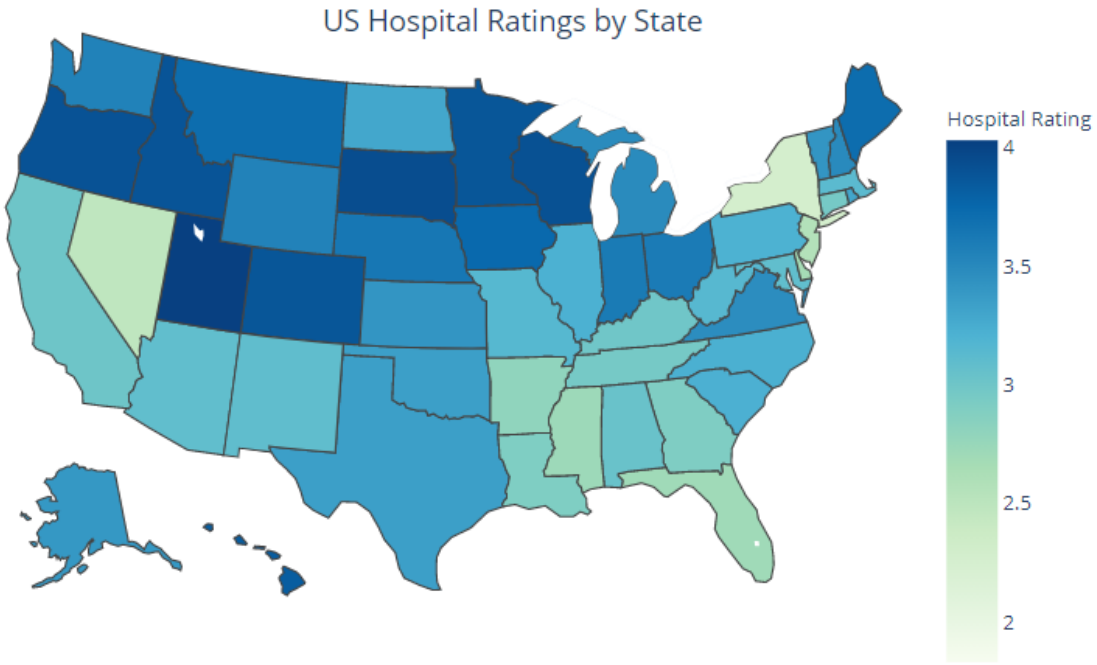


Q: What are the current hospital ratings and what specifications comprise of the ratings?

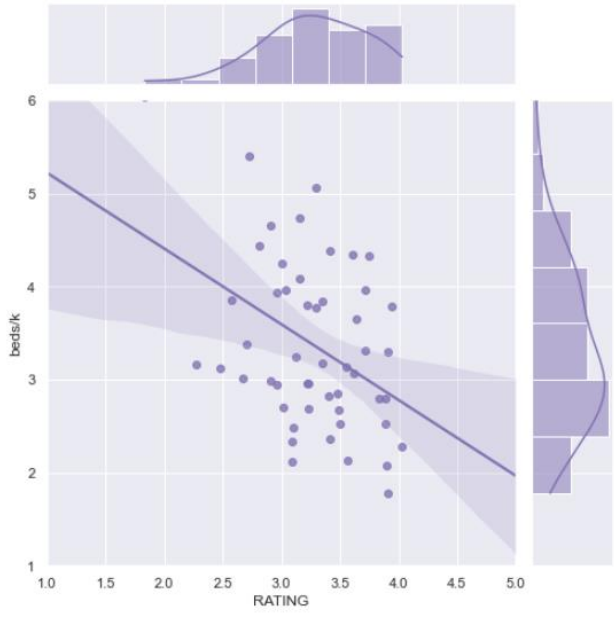
According to the Center for Medicare and Medicaid Services (CMS), hospitals are rated on a scale of 1 through 5 or not available with 5 being the highest rating. There are a total of 7 measures that the ratings are based on with a different weight to the measures.

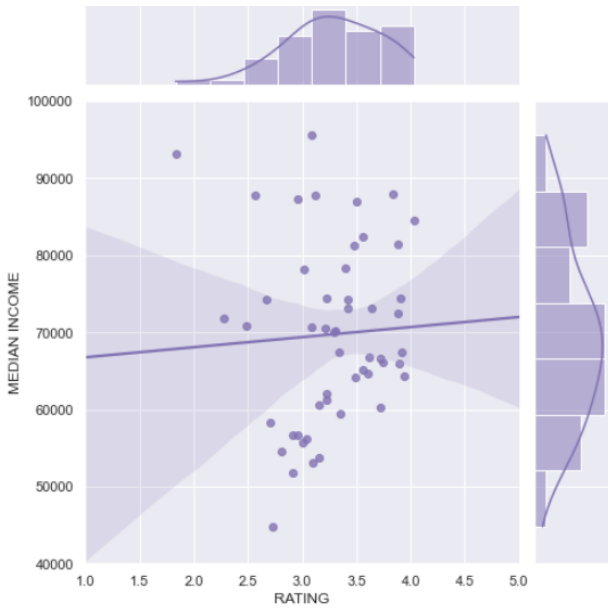
1) Mortality (22% weight)
2) Safety or Care (22% weight)
3) Readmission (22% weight)
4) Patient Experience (22% weight)
5) Effectiveness of Care (4% weight)
6) Timeliness of Care (4% weight)
7) Efficient Use of Medical Imaging (4% weight)

It is hard to tell of the relationship to ratings base on the map of the United States. The only thing that can be drawn are that most of the upper United States have higher hospital ratings as oppose to the coastal states.



With the jointplot showing the histogram of the rating variable compared to income and bed days using regression line to draw out the relations. It looks to have a negative correlation as you have more beds and the opposite with median income. However, looking at the ends of the regression lines, there is a large gap in error rate.





Conclusion:

Coming back full circle to the question at hand: How accessible are the hospitals? are there specifications relating to the location of the hospitals? are the hospital beds sufficient for the population at hand? What are the current hospital ratings and what specifications comprise of the ratings?

From the various slicing and dicing of the dataset, I cannot say that the conclusions drawn out from the exploratory data analysis are one hundred percent accurate as further drilling of the data would be required to further boost the confidence level of the analysis. With that said, the accessibility of the hospital for the United States seemed adequate enough compared to the standard 3 beds per thousand population. United States for 2020 are above at 3.2 beds for every thousand lives. Hospitals are currently placed where the needs are most prominent base on the type of services needed. For the totality of the United States, the rating system base on CMS standard is a little above a 3 out of 5-star rating measurement. Although the pandemic has caused quite a few worries on whether or not the United States health system can support the people needing the services. Based upon the data used for this assignment, the United States health system in regard to hospital is sufficient base on regular benchmark. However, the pandemic is more of an extreme case to which different benchmarks might be more adequate for. Building of hospitals are not something that can be done on a whim and having a hospital for precaution seemed a little extreme. In light of this, having an action plan to expand healthcare in case of emergency seemed to be more reasonable.