Examining the Homeless Population of the United States:

Does Each State Have the Means to Support their Homeless?

By Andrew Watkins

Introduction:

In this study, I will determine if the United States is able to care for the homeless population. I will evaluate if the facilities, or Continuum of Care (CoC), a program under the Department of Housing and Urban Development which aims to end homelessness, have enough beds to manage their homeless community. These facilities include regional or local planning bodies that coordinate housing and service funding for homeless families and individuals. This will also allow me to inspect how these facilities have changed over the years, thus enabling me to evaluate their performances.

I will be using a variety of techniques to gather the necessary information from the Department of Housing and Urban Development (HUD) and prepare it for the study. With the cleaned and prepared data, I will perform some exploratory data analysis (EDA) to better understand the data along with its trends using some visualization methods. I will also be using some machine learning techniques to predict how they will perform in the future.

The primary audience for this study is the same organization that provides the dataset, HUD. Other Non-Governmental Organizations (NGOs), also known as nonprofits, can benefit from this study as well. The results of the study will help these organizations understand how to manage their resources to better tackle the homeless issue locally and nationwide.

Dataset:

When conducting this study, I utilized two different datasets. The data I used came from HUD and was collected between 2007 and 2017. The datasets are the Point-in-Time estimates of the homeless population (PIT) and the accompanying Housing Inventory Count (HIC). These datasets needed to be examined, cleaned, and prepared to be useable in this study.

The datasets were provided as a spreadsheet, so the first thing I had to do was load them. I started with the HIC dataset, and checked the name of all the sheets provided to ensure they were labeled by year. I examined the first and last years (2007 and 2017) by checking the head and the information of the sheets. There was a difference in the amount of columns between 2007 and 2017, which was mostly likely due to HUD reporting on new aspects throughout the years. I double checked this and indeed a new column or even program was added almost every year. With a better understanding of how the dataset was structured, I started creating the DataFrame by designing a loop that goes through all the sheets and concatenates them together, making sure to keep all the columns. In addition, I added the year as a new column to be used later. Using the newly created year column and the state column, I created a multi-level index for the DataFrame.

I proceeded to examine the freshly created HIC DataFrame. Each program[1] has a 'Total Beds' column and additional columns on how these beds are distributed[2]. In this study I was not looking at how the beds are distributed, but if the beds are enough to accomodate the homeless community. Therefore, I deleted all the columns that did not contain any useful information. Some titles of the programs have changed the way they are written over the years; in the first few years one

[1] All Programs: Total Year-Round Beds (ES),Total Year-Round Beds (TH), Total Year-Round Beds (SH), Total Year-Round Beds (PSH), Total Year-Round Beds (RRH), Total Year-Round Beds (DEM), Total Year-Round Beds (OPH)

[2] Amount of beds designated for veterans, children, women, etc.

of the programs was written as “Total Year-Round ES Beds,” but in the last few years it has changed to "Total Year-Round Beds (ES).” These columns were merged into one -- I kept the way it was written in the later years to maintain consistency with future data. The final step for the DataFrame was to fill all NaN/null values with zero.

Next, I began to handle the PIT dataset by again loading and examining the spreadsheets. Just like the HIC spreadsheet, this one was labeled by year -- but it contains an extra sheet named ‘Changes.’ This sheet has the percentage change in homeless population in each state by year; I created a separate DataFrame, as this might be useful later on. When checking the head of the other sheets, I noticed that the columns have the year in the name -- which poses a problem when I try to concatenate the sheets. I renamed all the columns to the same name without the year which enabled me to concatenate all the sheets into one DataFrame. Once I was done, I noticed that one of the indexes was wrong. I checked the tail of the DataFrame, which showed that there were extra indexes written as notes in the last two years of the original dataset. I eliminated the indexes and the row associated with them and filled all the NaN/null values with zero.

At this point I had 3 DataFrames, but I wanted to merge them into one. I only needed two columns from the PIT DataFrame: ‘Total Homeless’ and ‘Number of CoCs,’ so I added them to the HIC DataFrame. The way the Changes DataFrame was structured caused difficulties when adding it to the other DataFrame, so I used the ‘Total Homeless’ column to calculate the next two columns. ‘Change to Date’ is the percentage change of the population since 2007 (i.e. 2007 - 2010, 2011, etc) and ‘Change from Last Year’ is the percentage change in population from the year before (i.e. 2009-2010, 2010-2011, etc). With these new columns, I was able determine if a state is prepared to manage its

homeless population. I created a new boolean column called 'Able,' this column will display ‘True’ if the 'Change to Date' and the ‘Change from last year' column is 0.00 or lower. The cleaning and preparation of the data was now finished, as seen below in the sample of the DataFrame (Fig 1.1):

year	State	Total Year-Round Beds (ES)	Total Year-Round Beds (TH)	Total Year-Round Beds (SH)	Total Year-Round Beds (PSH)	Total Year-Round Beds (RRH)	Total Year-Round Beds (DEM)	Total Year-Round Beds (OPH)	Total Homeless	Number of CoCs	Total Beds	Change to date	Change from last year	Able
2017	VT	649.0	308.0	4.0	609.0	703.0	0.0	0.0	1225.0	2.0	2273.0	0.183575	0.096688	False
	WA	8428.0	5815.0	45.0	11504.0	5906.0	0.0	3026.0	21112.0	7.0	34724.0	-0.096967	0.013684	False
	WI	3422.0	2012.0	91.0	3265.0	1367.0	0.0	460.0	5027.0	4.0	10617.0	-0.109950	-0.115743	True
	WV	1338.0	306.0	13.0	1239.0	262.0	0.0	330.0	1309.0	4.0	3488.0	-0.456621	-0.056236	True
	WY	361.0	337.0	0.0	285.0	25.0	0.0	0.0	873.0	1.0	1008.0	0.625698	0.018670	False

Fig 1.1: The last 5 entries in our DataFrame

Exploratory Data Analysis:

Because the data was prepared, I moved on to exploratory data analysis (EDA). The goal of EDA is to get a better understanding of the data, which I accomplished through statistical analysis and visualization techniques. While there were several plots and techniques I used to visualize the data, I only published the most relevant to this study. To begin, I plotted the homeless population (Fig 2.1):

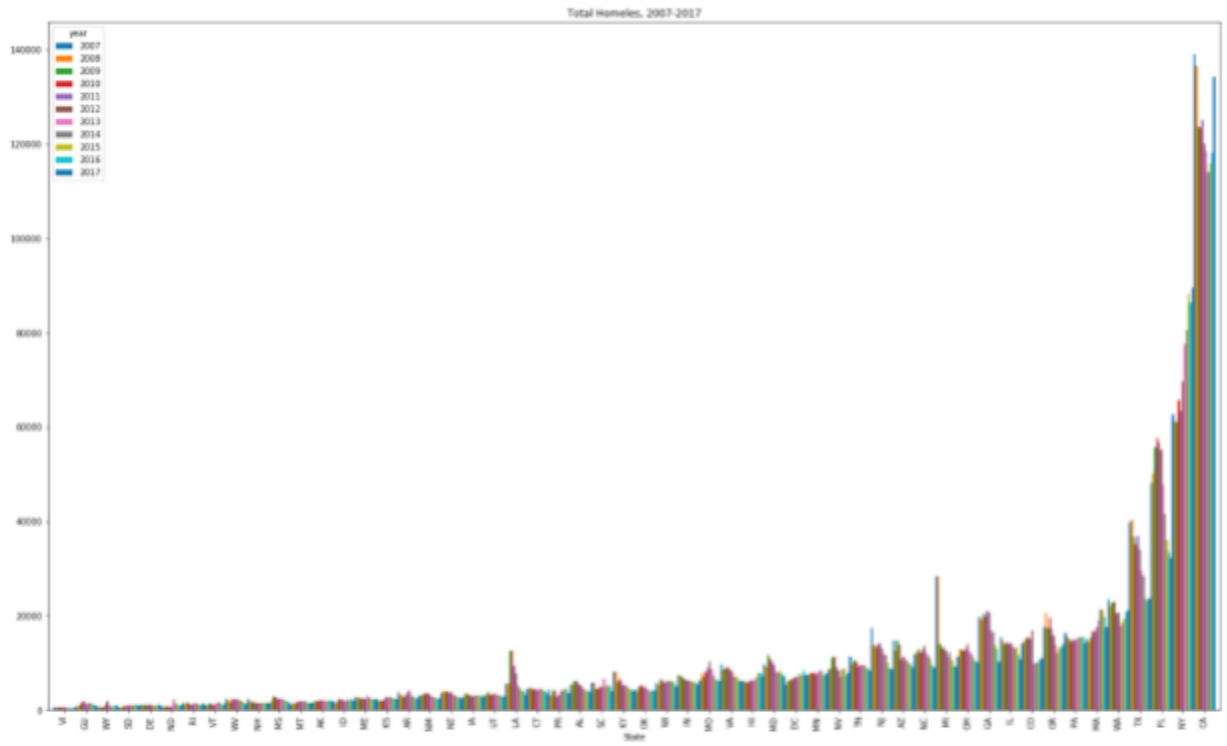

Fig 2.1: Total Homeless per state, 2007-2017

This plot shows that most of the US states from 2007-2017 have a homeless population of 20,000 or less with only four states[3] above 40,000. These four states could be possible outliers that may need to be deleted before creating my machine learning models -- an issue I address later. To confirm the numbers I observed in the plot, I calculated some basic statistics of the DataFrame (Fig 2.2):

[3] Texas, Florida, New York and California

	Total Homeless	Total Beds	Number of CoCs	Change to date	Change from last year
count	594.000000	594.000000	594.000000	594.000000	594.000000
mean	11168.698653	13517.013468	7.907407	0.021145	-0.004291
std	19863.448043	19900.731143	8.546856	0.313976	0.160595
min	337.000000	149.000000	1.000000	-0.680120	-0.504213
25%	2247.250000	3091.000000	2.000000	-0.147263	-0.071907
50%	5526.000000	7559.500000	4.000000	0.000000	-0.004449
75%	11650.750000	14648.250000	10.000000	0.121574	0.036916
max	138986.000000	135102.000000	43.000000	2.376164	2.007267

Fig 2.2: Basic stats of the DataFrame

The average homeless population is 11,168 -- well below 20,000 initially thought. The standard deviation is higher than the average at 19,863; the outliers are definitely skewing this statistic. The states do not appear to have a huge issue with the homeless community, because the average homeless population is not incredibly high. However, just looking at one of the columns is not enough to make this assessment. In order to determine if a state can tend to the homeless community, I needed to examine the amount of beds it has versus the number of homeless people and the ratio between them.

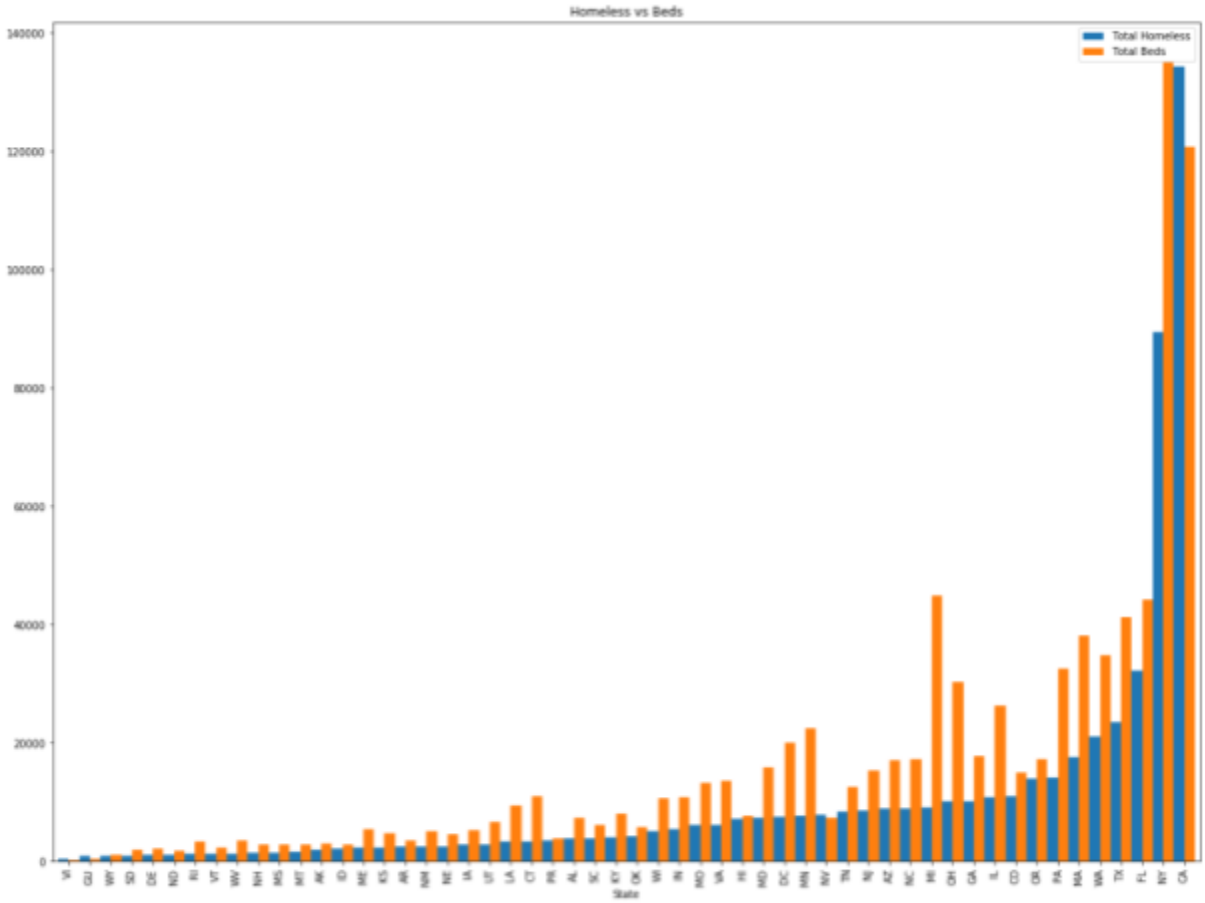

Fig 2.3: Total Homeless Population vs Total Beds, 2017

The Total Homeless Population vs Total Beds plot (Fig 2.3) is for 2017 only, and it shows that most states have more beds than homeless. The other plot displays the ratio of homeless to beds (Fig 2.4), and shows that most states are below a one-to-one ratio. However, the bulk of the states are between a one-to-one and one-to-two ratio (0.50 or 1 bed for every 2 homeless person). These two plots contradict each other -- one claims the states have more than enough beds, while the other shows that the ratio between bed to homeless is less than one-to-one. So what is causing this? Since the Total Homeless Population vs Total Beds plot is only from 2017, the states may have increased their number of beds throughout the years to compete with rising homeless populations. If this is

the case, does it indicate the states have the means of managing their homeless population?

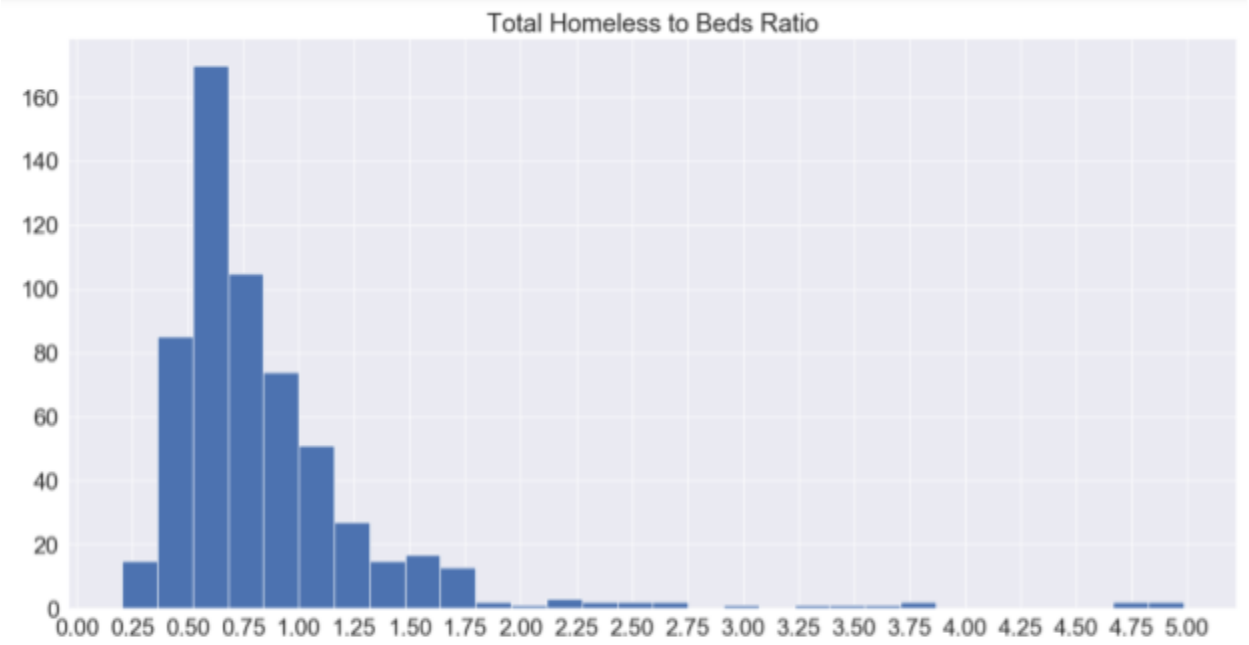

Fig 2.4: Distribution of Total Homeless to Beds Ratio

The DataFrame contained two columns that were incredibly useful to help answer these questions: ‘Change to Date’ and ‘Change from Last Year.’ These two columns contained the percentage change in population, which was more indicative of a state being able to tend to its homeless community. If the homeless population of a state is dropping each year and has been steadily dropping since 2007, then I believe that is an suggests that the state has all the means to help eradicate their homeless population. To see if a state is enabled, I plotted the percentage changes since 2007 (Fig 2.5):

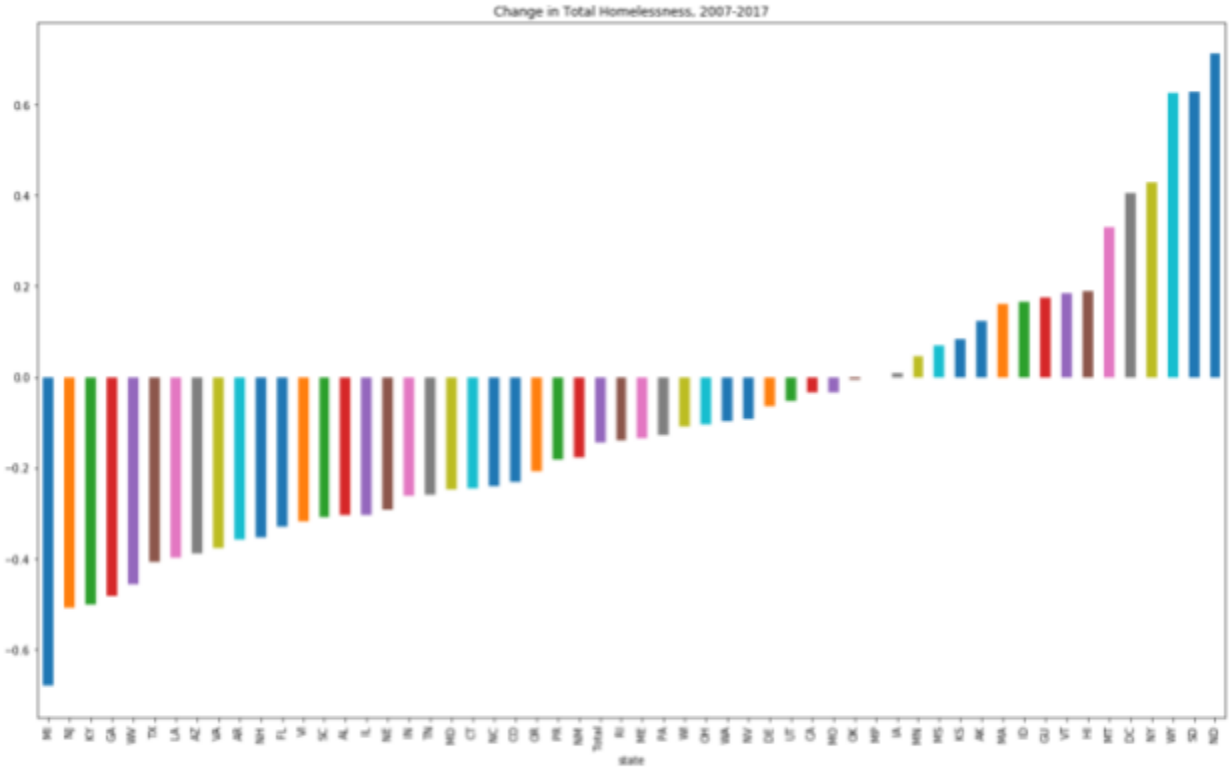

Fig 2.5: Change in Total Homelessness, 2007-2017

The most important takeaway from this plot is that overall homelessness is going down in the U.S. as a whole and per state. Nonetheless, this is an overarching view of the change in population and does not look at the changes year to year. I needed plot the distribution of these two columns to get a better understanding of the changes in population.

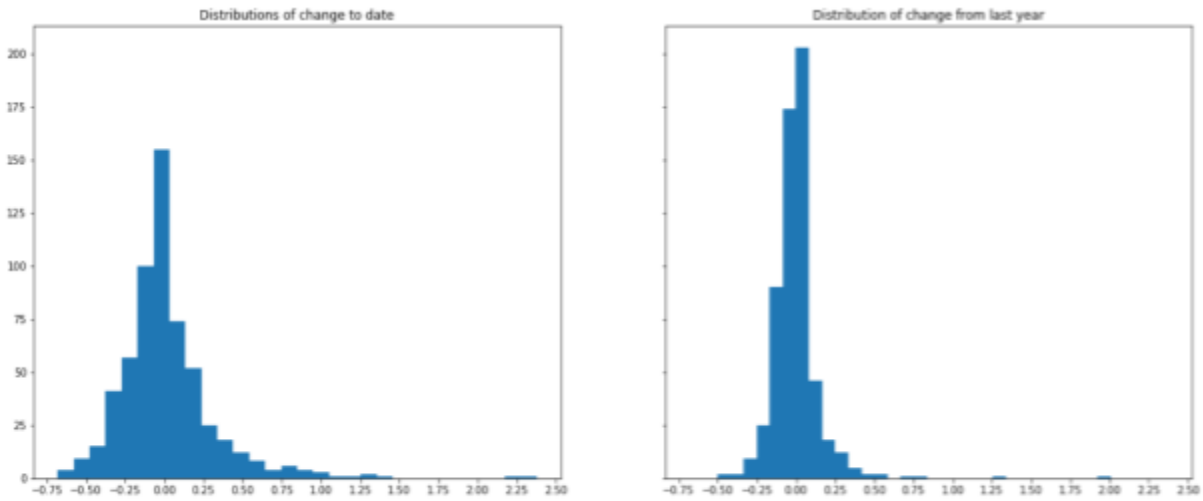

Fig 2.5: Distribution of changes in population

These two plots signify an interesting result, because one has a much wider distribution than the other. The distribution of changes in population since 2007 to date varies drastically, which means the changes in population vastly fluctuates. Yet, when compared to the changes from last year (i.e. 2009-2010) one may note that the distribution is much tighter and centers just above zero percent. This means on a year to year basis the homeless population within states increases by a small amount, which then most likely drops the year after. This trend suggests states are playing catch up, or being reactionary to the changes in the homeless population, instead of getting ahead of the problem.

As mentioned before, when preparing the data I created a new column also known as ‘Able,’ that was calculated using two other columns[4]. This column determines whether a state was able to handle the population if the 'Change to Date' and the ‘Change from Last Year' were 0.00 or lower. About 35.9%[5] or 213 out of 594 were actually able to manage their homeless population across the years (Fig 2.6). The concentration of homelessness in the top 4 most populated

[4] ‘Change to date’ and ‘Change from last year’
[5] 54 States and Territories times 11 years (54 * 11 = 594)

states[6] coupled with the decline in homelessness suggests people migrate to these states. Even though homelessness has gone down overall in the U.S., it seems that the states and homelessness are in a constant back-and-forth routine -- as the homelessness population increases, so do the number of beds; yet the root cause is not being addressed. While there could be multiple different factors contributing to the root cause of homelessness in the U.S., they are unfortunately out of the scope of this study.

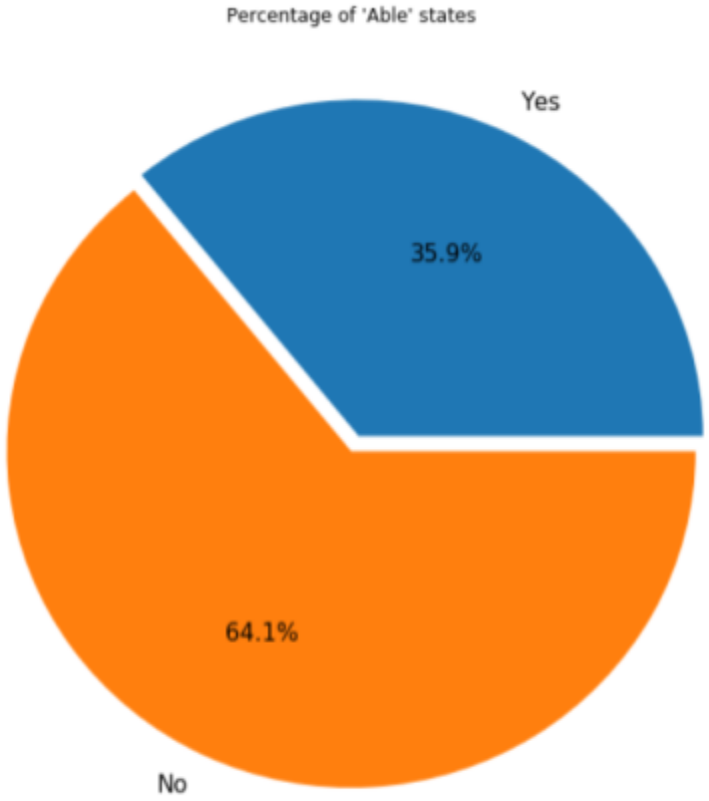

Fig 2.6: Percentage of 'Able' states

[6] California, Texas, New York and Florida

Machine Learning:

In this section, I wanted to apply both linear and logistic regression to the data. The first thing I needed to do was examine how the columns relate to each other to know which columns I needed to use in my linear regression models (Fig 3.1).

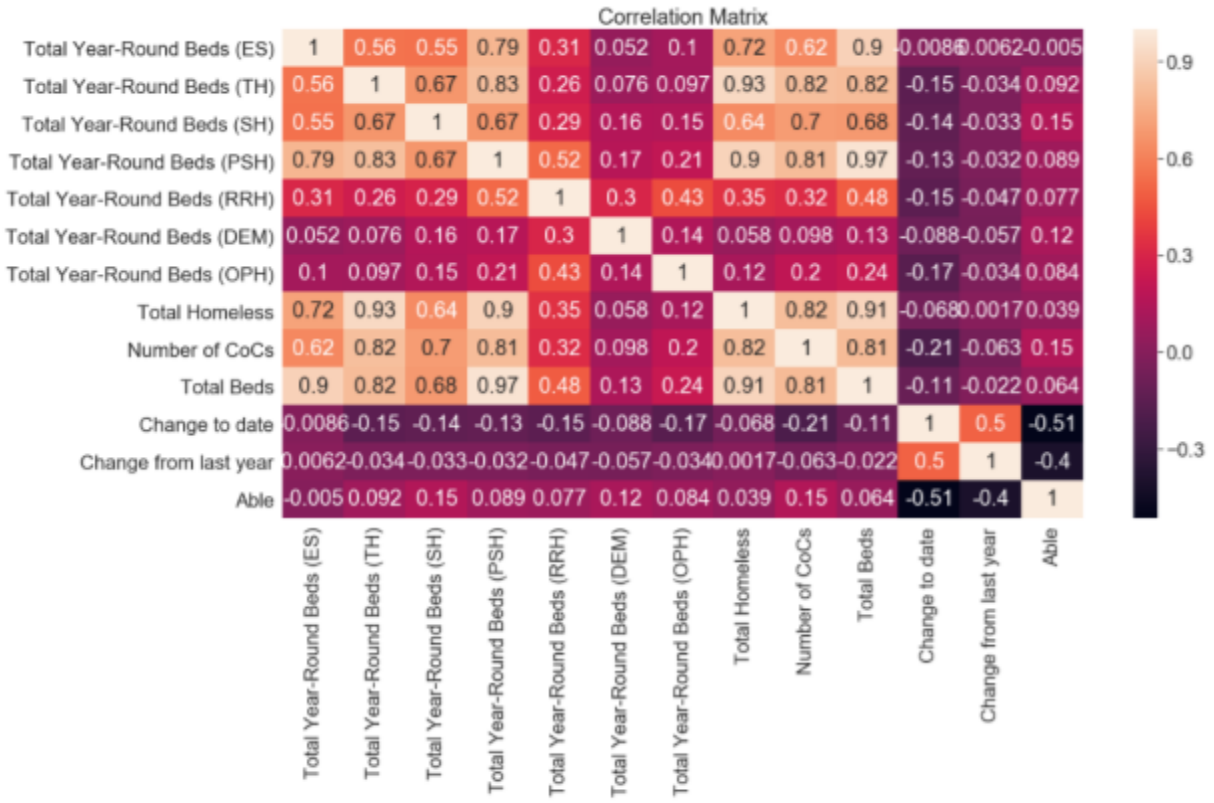

Fig 3.1: Correlation Matrix

The matrix displays ‘Total Homeless’ and ‘Total Beds’ as highly correlated. There are other fields with a higher correlation, such as the distribution of beds related to ‘Total Beds.’ Since I was not doing any multi-class classification, I disregarded these correlations. I moved forward and created a linear regression plot (Fig 3.2):

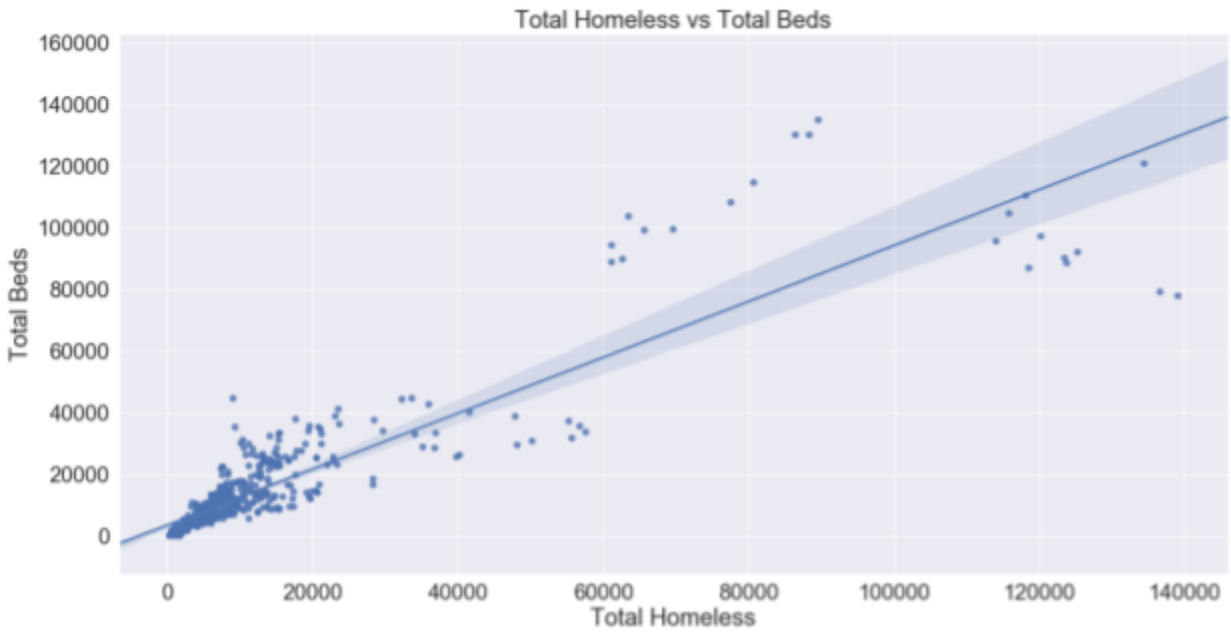

Fig 3.2: Linear Regression

The linear regression shows some heavy outliers in the data. These data points could be interfering with the model, so I removed them. This created gaps in the data, which had the potential to make the model weaker. The first step to remove the outliers was to identify them using an influence plot (Fig 3.3). The plot clearly shows that CA and NY are the two major outliers that need to be removed from the data. I created a second DataFrame to store the dataset with the outliers removed. Finally, I plotted a linear regression again to see the effects of removing the outliers (Fig 3.4).

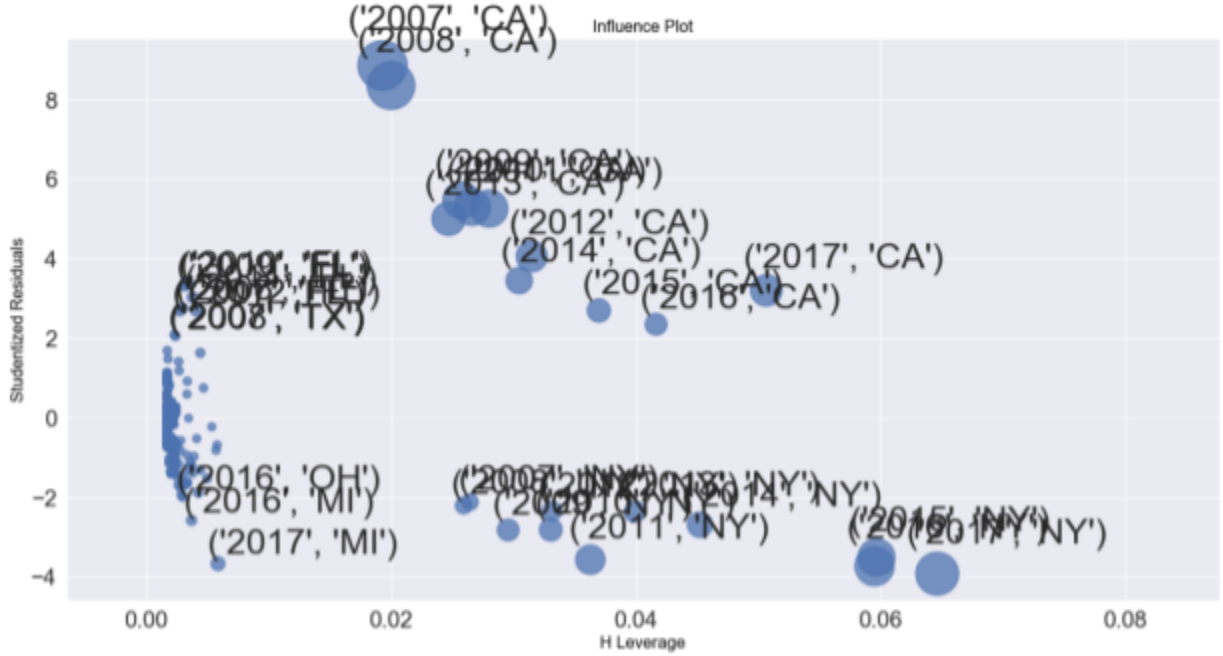

Fig 3.3: Influence Plot

Fig 3.4: Linear Regression w/o outliers

The linear regression has a much narrower range at the end, which is promising; thus I created all my models with both datasets[7]. Before continuing, I needed to decide what models I wanted to use and how to evaluate them. The linear models I used were Linear Regression, Ordinary least squares (OLS), Decision Tree, KNN and Random Forests. Further, I used R squared and Root Mean Squared Error to evaluate them. I placed all the scores in a new DataFrame to compare them thoroughly (Fig 3.5).

	Linear Regression	Linear Regression Min	OLS	OLS Min	Decission Tree	Decission Tree Min	KNN	KNN Min	Random Forest	Random Forest Min
R square	0.894323	0.692234	0.819979	0.665533	0.756775	0.542302	0.000000	0.00000	0.000000	0.008696
RMSE	5202.403699	5247.726461	0.000000	0.000000	7908.784209	6399.562940	7861.498409	6399.56294	8008.179527	6399.562940

Fig 3.5: Linear Regression Scores

All the models that used the datasets without the outliers performed worse than their counterparts. My initial hypothesis was correct: removing the outliers created gaps in the data, hindering the performance of the models. Additionally, some models performed poorly as they had a R^2 score of zero, where the desire is to have them as close to 1 as possible. The best performing model was a simple Linear Regression, with a score of 0.89, and the Ordinary Least Square as a close second with 0.82. These two models worked well in predicting the amount of beds a state would have at any point given its homeless population.

In the logistic regression section, I wanted to classify the states. I used the threshold I determined earlier: a column called 'Able.' With logistic regression, I decided not to use the dataset with the outliers removed as it performed poorly across the board with linear regression. I determined I wanted to use seven different logistic regression models: Logistic Regression, KNN, K-Means, Decision Tree, SVM, Naives Bayes and Random Forests. I evaluated the models

[7] The one with the data intact and the other with the outliers removed.

using Accuracy, Confusion Matrix, Classification Report, Area Under the Curve (AUC) and AUC with a 5-fold cross-validation. I placed all the scores (except Classification Report) in a DataFrame to compare them (Fig 3.6):

	Logistic Regression	KNN	K-Means	Descision Tree	SVM	Naive Bayes	Random Forest
Accuracy	0.831933	0.638655	0.680672	1	1	0.630252	0.983193
AUC	0.921378	0.65627	0	1	0.513158	0.654646	0.999838
AUC cross val	[0.8278465720326186, 0.960220318237454, 0.9124...	[0.5377529447296889, 0.7423500611995104, 0.727...	0	[0.9534883720930232, 1.0, 1.0, 1.0, 1.0]	[0.5, 0.5, 0.5, 0.5, 0.5]	[0.6324373301117487, 0.6814565483476132, 0.734...	[0.9978858350951374, 1.0, 1.0, 1.0, 1.0]
Confusion Matrix	[[72, 9], [11, 27]]	[[67, 14], [29, 9]]	[[81, 0], [38, 0]]	[[81, 0], [0, 38]]	[[81, 0], [0, 38]]	[[72, 9], [35, 3]]	[[81, 0], [2, 36]]

Fig 3.6: Logistic Regression Scores

Almost all of models performed incredibly well. Two of the models had perfect accuracy scores: Decision Tree and SVM. Yet, SVM had a poor AUC score. Furthermore, K-Means, KNN, and Naive Bayes did unsatisfactorily in comparison, with an accuracy score of around 60%.

Conclusion:

This study was created to try to determine if the United States is able to care for its homeless population. On the surface, it appears that the states do have the resources to manage the homeless population, but the truth is not so straightforward. There is a huge concentration of homelessness in the top 4 most populated states -- while homelessness has gone down overall in the U.S., it is likely that the concentration of homelessness is moving to more populated states. I created multiple machine learning models that are able to predict how many beds a state needs to accommodate the homeless. However, HUD is just reactionary to the increase in homelessness. As homelessness increases, so do the number of beds provided, yet the underlying problem is not being addressed. So long as the underlying problem is not addressed, or a different method is

implemented to eradicate homelessness, the states with the highest populations will always be in need of more beds.

Another important issue that was not covered in this study is how the distribution of beds can help reduce the homeless population. The distribution of beds varies widely by state; some states, such as New York, have a high amount of beds that are mostly concentrated in the Emergency Services program.