Capstone Project 1: Milestone Report

Examining the Homeless Population of the United States:

By Andrew Watkins

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

The primary audience for this study is actually the same organization that provides the dataset, the Department of Housing and Urban Development (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.

To do this study I need to use two different datasets. The datasets I will be using come from HUD and were 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). So I need to examine, clean and prepare both datasets.

The datasets are provided as an excel document so the first thing I have to do is load the HIC excel and check the sheats. The sheets are labeled by the year. So

I examine the first and last year (2007 and 2017) by checking the head and the info of the sheets. Now that I have a better idea of how the dataset is structured I can start creating the DataFrame. I do this by creating a loop that goes through all the sheets and concatenates them together. Also during this process I add the year as a new column and use it with the state column to create a multi level index. I now examine the DataFrame, and notice that I have columns that are the sum of other columns, this represents duplicate data so I drop these columns and fill all NaN/null values with 0.

I now proceed to work on the PIT dataset. As before I load the excel and examine the sheets. Just as the HIC excel this one is also 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 so I create a seperate DataFrame as this might be useful later on. Now, when checking the head of the other sheets I noticed that the columns have the year in the name, this poses a problem when I try to concatenate the sheets. I proceed to rename all the columns with the same name without the year and then I concatenate all the sheets into one DataFrame. Once im done I notice that one of the indexes is wrong so I check the tail of the DataFrame. Indeed there are extra indexes which were notes in the last two years original datasets excel document. I eliminate the indexes and the row associated with them and fill all the NaN/null values with 0.

At this point in the study I do not look for or deal with any outliers because I need to examine and understand the data better before I can consider them outliers and if they would distort the data or results.

I proceed to do some exploratory data analysis (EDA) by visualizing the data. I begin plotting data for the HIC dataset, after plotting the dataset as is I can see that there are plenty of columns that are not needed. I proceed to create a new

DataFrame with only the necessary columns[1]. I proceed to re-plot many the previously mentioned plots with the new smaller DataFrame. The data starts to plot a lot nicer and I can make sense of the data.

I now move on and continue to plot the PIT dataset, as with the previous dataset I notice that there are columns here that we are not going to be using so I proceed to clean this dataset further. The dataset now only contains 2 columns[2]. From here I merge both datasets into one since they were shrunk down to the point that it doesn't make sense to work on them separately. In addition to merging the two DF I create a new columns that adds the total amount of beds from the HIC dataset as 'Total Beds'. I use this columns to plot along side with the 'Total Homeless' column. Finally, I plot the last dataset, the PIT change of the homeless population from 2007 to 2017.

We can see that overall the homeless population has gone down across the US. However, it has increase in some states such as California and New York, this is most likely due to the population density in those states. The number of CoCs has not change much in each state and the amount is not proportionate to the amount of homeless people each state has. We have also noticed that having more beds in a state does not mean that it will lower the homeless population. New York state has the highest amount of beds and it has the second largest homeless population in the US. New York by far has the most amount of beds for Emergency Services but very little in the other programs offered. This leads me to believe that the distribution of the beds in the programs are crucial to help lower the homeless population.

---

[1] 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] Total Homeless and Number of CoCs

The next step is to do some inferential statistics. I do this by using the pandas describe method on the DataFrame to view some basic stats such as the mean, standard deviation, max and min values. Then I create and visualize with a heatmap a correlation matrix. Using the heatmap I can see that Total Beds and Total Homeless are highly correlated, which is not surprising. I then use a scatter plot to visualize this correlation. It shows that it has some very big outliers. I proceed to inspect them with an influence plot. The plot shows that CA and NY are the two outliers. I eliminate CA and NY from the DataFrame and create a regression plot. With the new DataFrame you can see that points are spread out more evenly in the graph then before and thus creates a regression line that is more representative of all the States. I also check the distributions, before and after the removal of CA and NY, of the columns and can there is a clear distinction in the distribution.

It is important to take into account all states when doing this study but it is hard to ignore how CA and NY skew the data. I have to take careful consideration how I am going to manage these states when creating my machine learning model. I will first try to create the model with all the states and evaluate the performance of this model. Then I will proceed to remove CA and NY from the DataFrame and recreate the models. I can then compare which one had the better performance.