Homeless Data Analysis - Insights

Sudha Subramanian (Sparkfish)

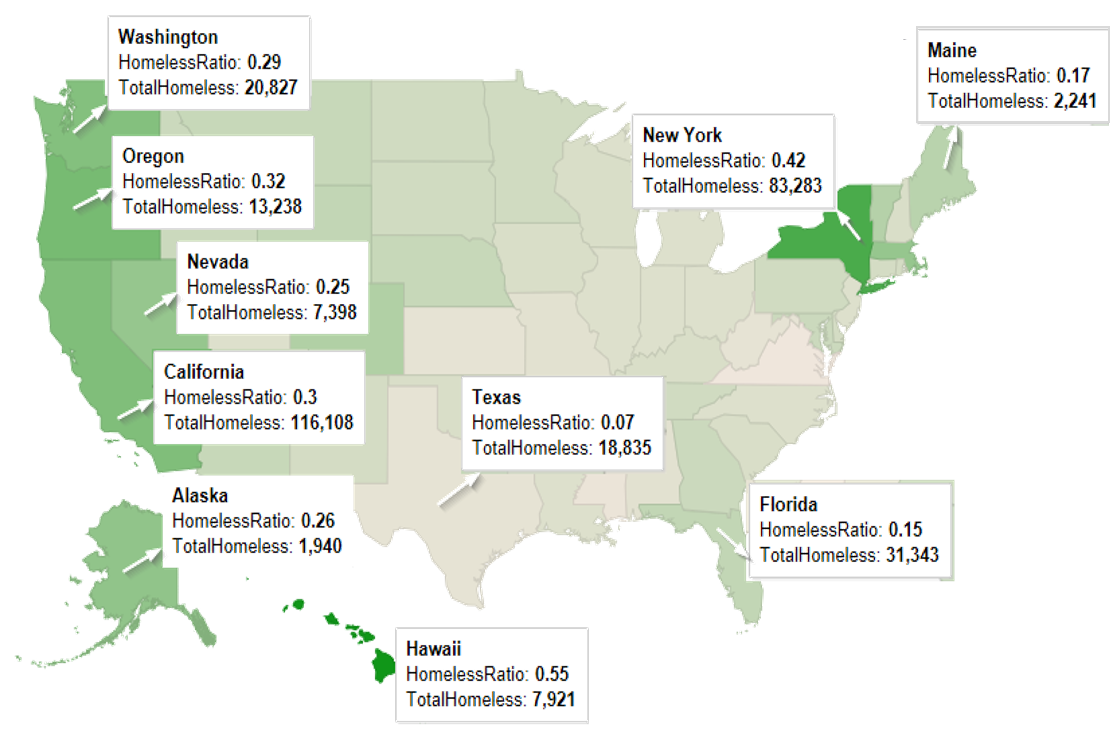
Oct 28, 2017

## Homelessness in United States - A Study

This is an analysis done primarily using R on a public dataset on homelessness across United States, downloaded from HUD website. For the scripts / datasets, please visit <https://github.com/susub31/EARL-Boston2017-Presentation>.

When we think of Hawaii, we tend to think it is one of the most beautiful vacation spots and a must-see place in your lifetime. And, needless to say that we know it is one of the most expensive places too. But, did you know that in the year 2016, 7900+ were reported as homeless in the state of Hawaii? Well, the actual count of homeless people in Hawaii may not be anywhere near what it is in California or New York. Considering the percentage of people who are homeless relative to the population in the state, Hawaii tops the list. Yes, Hawaii has the highest homeless ratio (number of homeless people relative to the State's population) in the country.

### Homeless Ration Map



## Homeless Counts

“Point-in-time” surveys are taken periodically to track counts of homeless people periodically. This information is very useful to understand spread of homelessness across all States. It also feeds information to the Department of Housing and Urban Development to determine funding for cities and other initiatives that need to be taken to reduce homelessness.

### Datasets for this analysis:

* Homeless Data 2016 (from HUD)
* US States (for map)

### Data Preparation:

Data Preparatory work has already been done on the original dataset to include lat/lon details across for all counties / cities. The saved information is used for this analysis.

### Insights:

Initial analysis done on the dataset shows different categories of homeless people such as counts of adults, veterans and youth. This is further categorized as sheltered, unsheltered, critically homeless etc., to name a few. The following section of code groups the counts by the state so that we can get overall homeless counts in each state across US. We build this dataset and categorize them as 'Low', 'Medium' and 'High' based on the homeless counts.

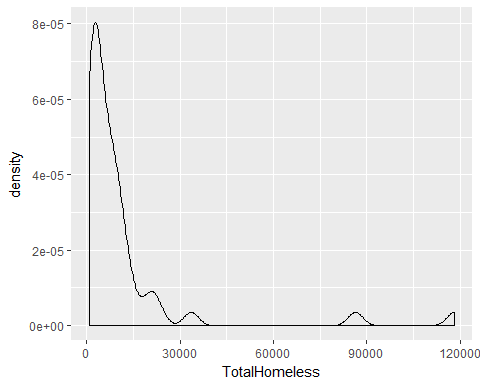
# Read the dataset that holds Homeless Data for 2016  
hdata <- read.csv("../Datasets/HomelessData2016.csv", stringsAsFactors = FALSE)  
  
# Read the dataset that holds the extracted values of Lat and Lon for the CoCs  
modhdata <- read.csv("../Datasets/HomelessData2016\_With\_LatLon.csv", stringsAsFactors = FALSE)  
modhdata$Total.Homeless..2016 = NULL  
modhdata$CoC.Name=NULL  
  
# Merge the datasets and extract the desired columns; rename the columns  
alldata <- merge(hdata, modhdata, by.x ="CoC.Number", by.y="CoC.Number")  
alldata <- select(alldata, CoC.Number, CoC.Name, cityname, lat, lon, Total.Homeless..2016,   
 Homeless.Veterans..2016, Homeless.Unaccompanied.Youth..Under.25...2016) %>%  
 rename(TotalHomeless = Total.Homeless..2016, HomelessVeterans = Homeless.Veterans..2016,   
 HomelessYouth = Homeless.Unaccompanied.Youth..Under.25...2016)  
  
# The following conversion is required to get the numeric value for the homeless counts  
alldata$TotalHomeless <- as.numeric(gsub(",","", alldata$TotalHomeless))  
alldata$HomelessVeterans <- as.numeric(gsub(",", "", alldata$HomelessVeterans))  
alldata$HomelessYouth <- as.numeric(gsub(",", "", alldata$HomelessYouth))  
  
# Extract the lat/lon coordinates pertaining to mainland US  
# http://en.wikipedia.org/wiki/Extreme\_points\_of\_the\_United\_States#Westernmost  
top = 49.3457868 # north lat  
bottom = 24.7433195 # south lat  
left = -124.7844079 # west long  
right = -66.9513812 # east long  
  
#select for states in continous United States  
contusdata <- alldata %>%   
 filter(lat > 24 & lat < 50) %>%   
 filter(lon < -66 & lon > -124)  
  
#Include a column to capture "State" from CoCNumber  
contusdata$State = substr(contusdata$CoC.Number,1,2)  
  
# Load US States dataset for generating the map  
us <- map\_data("state")  
usstates <- read.csv("../Datasets/StateNames.csv")  
usstates <- usstates %>%  
 add\_rownames("region") %>%  
 mutate(region=tolower(StateName))  
  
# Merge US States and Cost of Living by cities datasets for color-coding the US Map  
usstates$StateName=NULL  
usstates <- merge(usstates, contusdata, by.x="State", by.y="State")  
  
# Group by and get totals by state in each category  
CountsByState <- usstates %>%   
 group\_by(State, region) %>%   
 summarize(TotalHomeless = sum(TotalHomeless), HomelessVeterans = sum(HomelessVeterans), HomelessYouth = sum(HomelessYouth))   
  
CountsByState$Count.Category <- ifelse(CountsByState$TotalHomeless > 20000, "HIGH", ifelse(CountsByState$TotalHomeless > 10000, "MEDIUM", "LOW"))

### Density Map of Homeless Counts

This shows the density of homelessness across the states. As can be seen from the graph below, there are very few states with very high homeless counts - there is one state with counts around 80K (New York) and one which is over 110K (California). Categorization is done as follows:

* States with homeless counts above 20K are categorized as "HIGH"
* States with homeless counts above 10K and below 20K are categorized as "MEDIUM"
* States with homeless counts below 10K are categorized as "LOW"

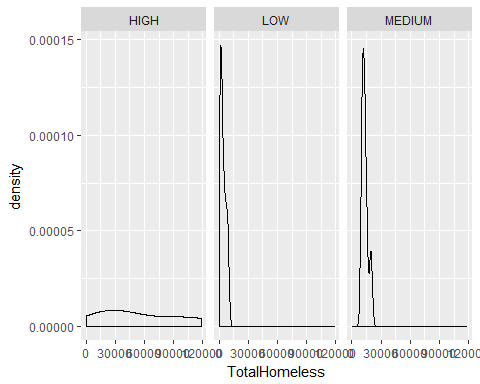
ggplot(CountsByState, aes(x=TotalHomeless)) +  
 geom\_density()



### Density Map of Homeless Counts - by Category

This graph shows a faceted view specific to each category to give an understanding of the number of states and the distribution in each of these three categories. As you can see, there are very few states in the 'high' category.

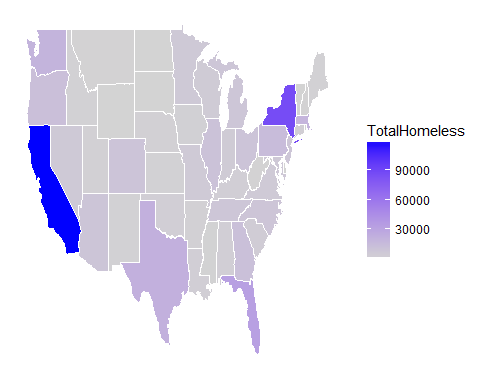
ggplot(CountsByState, aes(x=TotalHomeless)) +  
 geom\_density() +  
 facet\_wrap(~ Count.Category)



### Color-coded map of Homeless Counts

The homeless counts in the dataset is used to color code the US map to get an understanding of the extent of homelessness across different states. This will be used as the base map on which the counts in individual categories (Adults, Veterans, Youth) are plotted. This paints extent of homelessness across different states. ### Inference Here is what we can infer from this map: \* California and New York showcase the highest in that order in terms of counts of homeless people \* There are quite a few states in the 'medium' and 'low' categories

#Plot a blank US Map  
BlankUSMap <- ggplot()  
BlankUSMap <- BlankUSMap + geom\_map(data=us, map=us, aes(x=long, y=lat, map\_id=region),   
 fill="white", color="black")  
  
#US Map, color coded based on the total homeless counts in the state  
HomelessMap <- BlankUSMap + geom\_map(data=CountsByState, map=us,   
 aes(fill=TotalHomeless, map\_id=region), color="#ffffff", size=0.15)   
  
HomelessMap <- HomelessMap + scale\_fill\_continuous(low='lightgrey', high='blue', guide='colorbar')  
HomelessMap <- HomelessMap + labs(x=NULL, y=NULL) +   
 theme(panel.border = element\_blank()) +   
 theme(panel.background = element\_blank()) +  
 theme(axis.ticks = element\_blank()) +   
 theme(axis.text = element\_blank())   
  
HomelessMap



### Homeless Counts (Adults) - Top 100

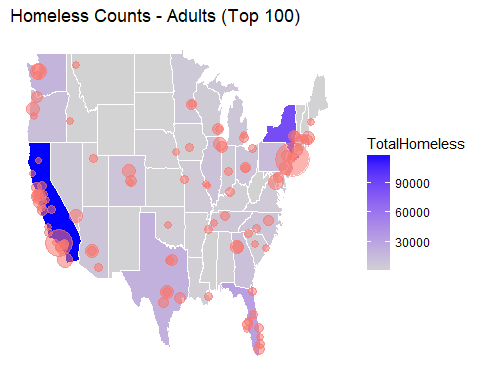
Plotting the count of homeless people (adults) for the top 100 continuums (CoC) with high counts of homelessness, paints the following picture, where size of the bubble indicates high count of homelessness.

### Inference

Overlaying the base map with points that correspond to size of homeless counts helps us understand where the counts are relatively higher than other places.

We can infer the following: \* NY and CA have the continuums with very high homeless counts \* The number of continuums (24 out of 100) in California featuing in this notorious top 100 list shows that homelessness is spread across the entire state (from San Diego to San Francisco) \* Likewise, Florida also has 9 continuums in the top 100 listgit

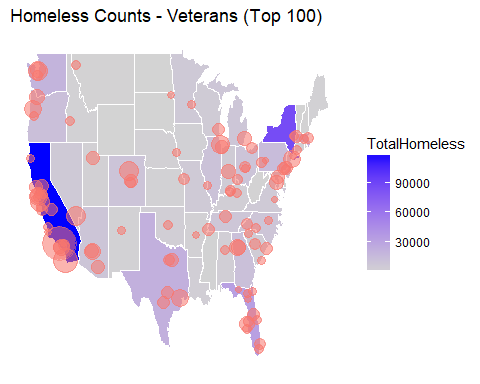
# select top 100 based on homeless counts - adults  
hdatatop100 <- tbl\_df(subset(alldata, !((substr(alldata$CoC.Number,1,2) =="HI")) & !((substr(alldata$CoC.Number,1,2) =="PR")) )) %>%  
 top\_n(100, TotalHomeless)  
  
Top100HomelessAdultsMap <- HomelessMap +  
 geom\_point(aes(x=lon, y=lat, size=TotalHomeless, colour="red", alpha=0.8), data=hdatatop100) +   
 ggtitle("Homeless Counts - Adults (Top 100)") +  
 scale\_size\_continuous(name="Homeless Counts", range = c(2,12), guide = FALSE) +  
 scale\_alpha(guide=FALSE) +  
 scale\_colour\_discrete(guide=FALSE)   
   
Top100HomelessAdultsMap



### Homeless Counts (Veterans) - Top 100

#### Plotting the count of homeless people (veterans) for the continuums (CoC) with high counts of homelessness, paints the following picture, where size of the bubble indicates high count of homelessness.

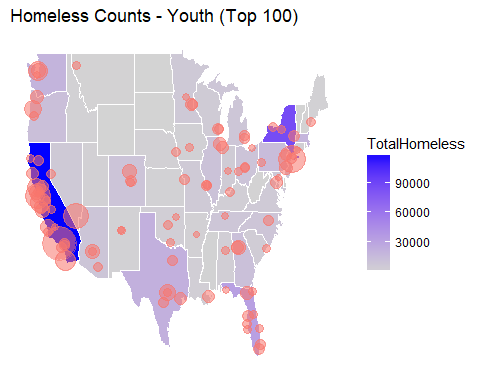
# select top 100 based on homeless counts - veterans  
hvetsdatatop100 <- tbl\_df(subset(alldata, !((substr(alldata$CoC.Number,1,2) =="HI")) & !((substr(alldata$CoC.Number,1,2) =="PR")) )) %>%  
 top\_n(100, HomelessVeterans)  
  
Top100HomelessVeteransMap <- HomelessMap +  
 geom\_point(aes(x=lon, y=lat, size=HomelessVeterans, colour="red", alpha=0.8), data=hvetsdatatop100) +   
 ggtitle("Homeless Counts - Veterans (Top 100)") +  
 scale\_size\_continuous(name="Homeless Counts", range = c(2,12), guide = FALSE) +  
 scale\_alpha(guide=FALSE) +  
 scale\_colour\_discrete(guide=FALSE)   
  
Top100HomelessVeteransMap



### Homeless Counts (Youth) - Top 100

#### Plotting the count of homeless people (Youth) for the continuums (CoC) with high counts of homelessness, paints the following picture, where size of the bubble indicates high count of homelessness.

hyouthdatatop100 <- tbl\_df(subset(alldata, !((substr(alldata$CoC.Number,1,2) =="HI")) & !((substr(alldata$CoC.Number,1,2) =="AK")) & !((substr(alldata$CoC.Number,1,2) =="PR")) )) %>%  
 top\_n(100, HomelessYouth)  
  
Top100HomelessYouthMap <- HomelessMap +  
 geom\_point(aes(x=lon, y=lat, size=HomelessYouth, colour="red", alpha=0.8), data=hyouthdatatop100) +   
 ggtitle("Homeless Counts - Youth (Top 100)") +  
 scale\_size\_continuous(name="Homeless Counts", range = c(2,12), guide = FALSE) +  
 scale\_alpha(guide=FALSE) +  
 scale\_colour\_discrete(guide=FALSE)   
  
Top100HomelessYouthMap



### Grants vs. Homeless Counts - An analysis

#### Read ESG Grants Dataset (with geocodes for mapping); exclude mapping for Hawaii, Puerto Rico so that the focus is on mainland United States only. Extract the data pertaining to top 'n' (in this example, top 25 are selected) for both the homeless counts dataset and grants dataset for performing a comparative analysis. As can be seen from the map, there is disparity between where we see top homelessness and where top grants are allocated.

#### For the base map, the cost of living index values are used to color-code the US map. Thus, it is possible to see how the homelessness correlates to the cost of living index values

#Read Grants dataset  
GrantsDS <- read.csv("../Datasets/ESGGrantDSWithGeoCodes.csv")   
GrantsDStop35 <- tbl\_df(GrantsDS) %>%   
 top\_n(35, TOT\_AMT)  
GrantsDStop35 <- subset(GrantsDStop35, !(STUSAB == "PR"))  
  
AllGrantsDStop35 <- group\_by(GrantsDS, round(lat, 2)) %>% mutate(sum(TOT\_AMT))  
  
AllHomelessTop35 <- tbl\_df(subset(alldata, !((substr(alldata$CoC.Number,1,2) =="HI")) & !((substr(alldata$CoC.Number,1,2) =="PR")) )) %>%  
 top\_n(25, TotalHomeless)  
GrantsDStop35 <- tbl\_df(subset(GrantsDS, !((substr(GrantsDS$STUSAB,1,2) =="HI")) & !((substr(GrantsDS$STUSAB,1,2) =="PR")) )) %>%   
 top\_n(25, TOT\_AMT)  
  
  
us <- map\_data("state")  
usstates <- read.csv("../Datasets/StateNames.csv")  
usstates <- usstates %>%  
 add\_rownames("region") %>%  
 mutate(region=tolower(StateName))  
  
COLByCities <- read.csv("../Datasets/COL\_ByCities.csv")  
COLByCities$UrbanArea <- gsub(", ", " ", COLByCities$UrbanArea)  
COLByCities$State <- COLByCities$UrbanArea  
COLByCities$State <- substrRight(COLByCities$State, 2)  
  
usstates$StateName=NULL  
usstates <- merge(usstates, COLByCities, by.x="State", by.y="State")  
  
COLByCities <- group\_by(COLByCities, State)  
COLByCities <- summarise(COLByCities, COLStateAvg=mean(COLIndex), HousingIndex=mean(Housing))  
  
COLMap <- BlankUSMap + geom\_map(data=usstates, map=us,   
 aes(fill=Housing, map\_id=region), color="#ffffff", size=0.15)   
  
COLMap <- COLMap + scale\_fill\_continuous(low='lightgray', high='lightblue', guide='colorbar')  
  
COLMap <- COLMap + labs(x=NULL, y=NULL) +   
 theme(panel.border = element\_blank()) +   
 theme(panel.background = element\_blank()) +  
 theme(axis.ticks = element\_blank()) +   
 theme(axis.text = element\_blank())  
  
GrantsAndCountsmap <- COLMap +   
 geom\_point(aes(x=lon, y=lat, size=TOT\_AMT, colour="Total Grants", alpha=0.8), data=GrantsDStop35) +   
 geom\_point(aes(x=lon, y=lat, size=TotalHomeless, colour="Homeless Counts", alpha=0.8), data=AllHomelessTop35) +   
 labs(x='Longitude', y='Latitude') +  
 ggtitle("Homeless Counts vs. Grants") +  
 scale\_alpha(guide=FALSE) +  
 theme(legend.position="none")  
  
GrantsAndCountsmap <- GrantsAndCountsmap + labs(x=NULL, y=NULL) +   
 theme(panel.border = element\_blank()) +   
 theme(panel.background = element\_blank()) +   
 theme(axis.ticks = element\_blank()) +   
 theme(axis.text = element\_blank()) +  
 theme(legend.position = "none")  
  
GrantsAndCountsmap

