Homelessness in NYC

Elijah Hall and Dayana Delgado

April 4, 2018

#Libraries  
if (!require(tidyverse)) install.packages("tidyverse")

## Loading required package: tidyverse

## Loading tidyverse: ggplot2  
## Loading tidyverse: tibble  
## Loading tidyverse: tidyr  
## Loading tidyverse: readr  
## Loading tidyverse: purrr  
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats  
## lag(): dplyr, stats

library(tidyverse)  
if (!require(lubridate)) install.packages("lubridate")

## Loading required package: lubridate

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(lubridate)  
if (!require(data.table)) install.packages("data.table")

## Loading required package: data.table

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':  
##   
## hour, isoweek, mday, minute, month, quarter, second, wday,  
## week, yday, year

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

library(data.table)  
if (!require(xts)) install.packages("xts")

## Loading required package: xts

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'xts'

## The following objects are masked from 'package:data.table':  
##   
## first, last

## The following objects are masked from 'package:dplyr':  
##   
## first, last

library(xts)  
if (!require(reshape2)) install.packages("reshape2")

## Loading required package: reshape2

##   
## Attaching package: 'reshape2'

## The following objects are masked from 'package:data.table':  
##   
## dcast, melt

## The following object is masked from 'package:tidyr':  
##   
## smiths

library(reshape2)  
if (!require(rmarkdown)) install.packages("rmarkdown")

## Loading required package: rmarkdown

library(rmarkdown)

## *Part I*

**Introduction** New York City (NYC) is one of the largest and most expensive cities to live in. These two factors naturally increase the likelihood of homelessness amongst the population. Media report on the trends of homeless usually throughout times of policy changes or the holiday seasons. They have repeatedly called the increase in homeless population a crisis and epidemic (Murphy, 2017). They quote point in time estimates that suggest almost a 20% growth rate.

**Problem Statement** Homelessness is said to be increasing, but the numbers constantly quoted in news and media are point in time comparisons based on the number of people in homeless shelters. There are numerous organizations that give conflicting estimates on the real homeless population and its growing rate calling it a crisis. To summarize, we want to know what is the homeless population and how has it been growing. Our research will help describe what the total population of homeless in and out of shelters is and how are they changing from year to year.

**Data Collection** We collected observational data from four sources. (1) The Department of Homeless Services (DHS) publishes numerous data sets. We collected daily count of homeless in shelters using the data set “DHS Daily Report”. There are 12 variable, but only 2 are of interest to answer our question, “Date of Census” and “Total Individuals in Shelter”. We keep the other variables to help us understand the total values in our analysis (NYC Open Data, 2017). (2) We also collected temperature data from the National Center for Environmental Information. Since some homeless sleep outside it is likely that most will transition to shelters when the temperature is coldest (Ncdc.noaa.gov, 2017). (3) Homeless Outreach Population Estimate (HOPE) conducts annual counts to estimate the homeless population on the streets. They usually do this near the beginning of the year. The counts are estimates from actual counts where they use statistical analysis to estimate the total street homeless. The data was aggregated into a .csv from their annual reports (NYC Homeless Outreach Population Estimate (HOPE), 2017). (4) We also collected NYC population data from the official website of New York State (State of New York, 2017).

**Data processing** We imported the data into R to clean the data and merge them into two separate data frames. The first data frame is called “newdata” which merges the DHS observations and the temperature data since both are daily measurements. We removed duplicate data and outliers that were too extreme to be accurate counts. The second data frame is called “Yearlydata” and merges NYC population data and HOPE count data since they are annual measurements. We added a separate “Year” column as an index since the dates of measurements are not the same.

**Methodology** For our analysis we identified the peaks of the homeless population in shelters and valleys of the temperature data. Since the homeless population peaks identified too many values we identified the nearest neighbor to the valleys of the temperature data. We took an average population of 15 days before and after the identified date to try to eliminate point in time estimation errors. The dates we found were not the same, but close. We calculated the average date to be February 4th. This is the average time that the temperature is the coldests and the homeless counts the highest in the shelters. For 2015 the data, February 3rd and 4th were removed due to either NA’s or being outliers. Therefore we used 2015-02-05 as the value for our estimations in 2015. We also did not have NYC population estimates for 2017 or 2018 so we used linear regression to estimate those values. We also didn’t have total homeless percent growth for 2014 since we didn’t have values for total homeless in shelters in 2013 or 2018. We used linear regression to estimate the missing values and were able to then produce a growth estimate for 2014 and 2018. The final calculations are to adjust for NYC population growth. We do this by subtracting the NYC population growth rate from the homeless population growth rate. Our final data frame is called “estimates” where we include our adjusted growth rates.

### Data

The "path\_" objects are file paths or url links to the data

## Parsed with column specification:  
## cols(  
## `Date of Census` = col\_character(),  
## `Total Adults in Shelter` = col\_integer(),  
## `Total Children in Shelter` = col\_integer(),  
## `Total Individuals in Shelter` = col\_integer(),  
## `Single Adult Men in Shelter` = col\_integer(),  
## `Single Adult Women in Shelter` = col\_integer(),  
## `Total Single Adults in Shelter` = col\_integer(),  
## `Families with Children in Shelter` = col\_integer(),  
## `Adults in Families with Children in Shelter` = col\_integer(),  
## `Children in Families with Children in Shelter` = col\_integer(),  
## `Total Individuals in Families with Children in Shelter` = col\_integer(),  
## `Adult Families in Shelter` = col\_integer(),  
## `Individuals in Adult Families in Shelter` = col\_integer()  
## )

## Parsed with column specification:  
## cols(  
## STATION = col\_character(),  
## NAME = col\_character(),  
## DATE = col\_date(format = ""),  
## TAVG = col\_character(),  
## TMAX = col\_integer(),  
## TMIN = col\_integer(),  
## TOBS = col\_integer()  
## )

## Parsed with column specification:  
## cols(  
## Date = col\_date(format = ""),  
## Surface = col\_integer(),  
## Subways = col\_integer(),  
## `Total Unsheltered` = col\_integer()  
## )

## Parsed with column specification:  
## cols(  
## `FIPS Code` = col\_integer(),  
## Geography = col\_character(),  
## Year = col\_integer(),  
## `Program Type` = col\_character(),  
## Population = col\_integer()  
## )

We collected data from four sources that needed to be cleaned.

#### 1. DHS Daily Report dataset:

1. About this Dataset: Provided by the Department of Homeless Services and available at the NYC Open Data webpage: <https://data.cityofnewyork.us/Social-Services/DHS-Homeless-Shelter-Census/3pjg-ncn9/data> Created on August 22, 2013 and Last updated on October 6th, 2017. 12 Columns and 1479 Rows
2. Variables:
3. Date of Census
4. Total Adults in Shelter: The number of single adults, individuals in adult families, and adults in families with children in shelter as of the date of census
5. Total Children in Shelter: The number of children in families with children in shelter as of the date of census.
6. Total Individuals in Shelter: The number of single adults, individuals in adult families, and adults and children in families with children in shelter as of the date of census
7. Single Adult Men in Shelter vi.Single Adult Women in Shelter
8. Total Single Adults in Shelter: The number of single adult men and women in shelter as of the date of census
9. Families with children in shelter
10. Adults in Families with Children in Shelter
11. Total Individuals in Families with Children in Shelter: The number of adults and children in families with children in shelter as of the date of census
12. Adult Families in Shelter
13. Individuals in Adult Families in Shelter

#### 2. Daily Temperature

1. About this Dataset: Provided by the National Center for Environmental Information Available at <https://www.ncdc.noaa.gov/data-access/quick-links> Dataset details the temperature by day. We will be merging the minimum temperature into our homeless population data table. 2 Columns and 18426 Rows
2. Variables:
3. Date: The date of the temperature record
4. TMIN: Minimum temperature registered for the day

#### 3. NYC Population Estimates

1. About this Dataset: Provided by the Official Website of the New York State Available at <https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k/data> 3 Columns and 3276 Rows
2. Variables:
3. Year: The year for which the population is calculated
4. Population: number of residents
5. Geography: geographic area name

#### 4. NYC Homeless Outreach Population Estimate (HOPE) counts

1. About this Dataset:

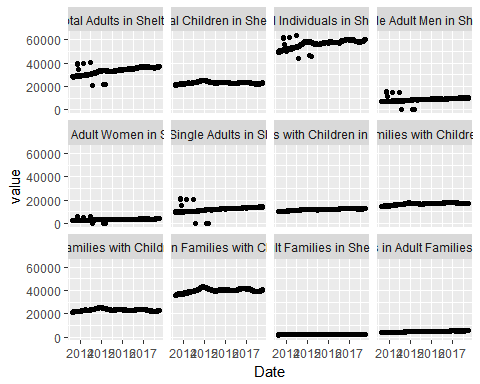
Provided by the NYC Department of Homeless Services Available at <http://www1.nyc.gov/assets/dhs/downloads/pdf/hope-2017-results.pdf> This point-in-time survey estimates the number of individuals living on city streets, parks, and in other public spaces throughout the five boroughs. Thousands of volunteers fanned out to complete the count. 4 Columns and 5 Rows

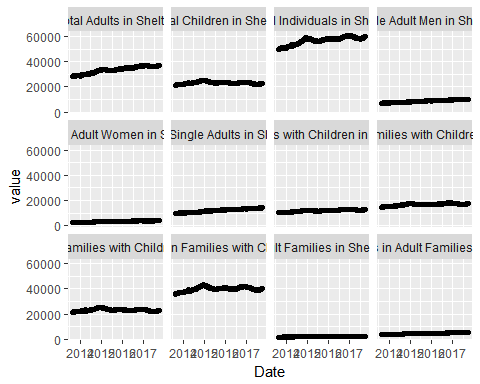
1. Variables:
2. Date: Date the HOPE count as conducted
3. Total Unsheltered Individuals: homeless people sleeping in public places such as streets, parks, and subways on a single winter night
4. Surface Areas: total homeless found in surface areas of all boroughs (Manhattan, Bronx, Brooklyn, Staten Island, Queens)
5. Subways: total homeless found in subways of all boroughs (Manhattan, Bronx, Brooklyn, Staten Island, Queens)

### Cleaning homeless data, "data"

## Warning: package 'bindrcpp' was built under R version 3.4.1

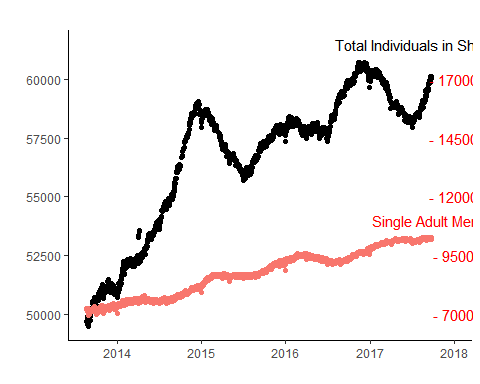
I want to know how each variable changes over time, maybe certain categories are more influencial than others.

 There are obvious outliers in Single Adults male and female. I will remove those that obviously do not belong since there was little to no metat data it is difficult to understand why these are here. There are a few outliers that are reasonable to assume a possibility of actually seeing the reported numbers. Some reasonable considerations are the some facilities misreported numbers, certain facilities were added for a short time to the reporting system, or simple data misentry. So we removed the obviuos outliers

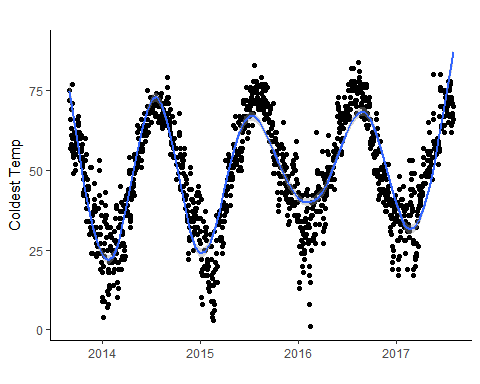
 Some groups of homeless are growing faster than others and contributing more to the problem than others. The numbers seen could also be a influenced by priority of service since families might be given services faster or more frequently on average, especially if children are present.

#look at variance by category  
vars<- sort(apply(nonas.nodupdata[,-1], 2, sd), decreasing = T)  
pr.vars<- vars^2/vars[1]^2  
pr.vars

## Total Individuals in Shelter   
## 1.000000000   
## Total Adults in Shelter   
## 0.804883026   
## Total Individuals in Families with Children in Shelter   
## 0.280067738   
## Total Single Adults in Shelter   
## 0.228072741   
## Single Adult Men in Shelter   
## 0.111658554   
## Adults in Families with Children in Shelter   
## 0.093342486   
## Total Children in Shelter   
## 0.084953313   
## Children in Families with Children in Shelter   
## 0.084952646   
## Families with Children in Shelter   
## 0.070254827   
## Individuals in Adult Families in Shelter   
## 0.025018300   
## Single Adult Women in Shelter   
## 0.021460265   
## Adult Families in Shelter   
## 0.005840627

As you can see about 80% of variance is from Adults and a large portion is from Single Men. This would seem to point to two potential concerns. This group is either more prone to homelessness or priroity of service is not being given equaly as mentioned before.  The graph also shows that the total change in Single Adult Men accounts for about 25% of total growth in homeless.

## `geom\_smooth()` using method = 'gam'

 There are obvious cycles as we would expect with seasonal changes. If we look at the stable portion of the data we see a meaningful negative correlation of cor(t\_new$Total Individuals in Shelter, t\_new$TMIN). The pattern is enough to suggest a peak demand for homeless shelter to be in the winter. This is not new information. However, it gives reason to use these time periods to predict and analyze point to point growth as it is likely the best and most accurate count of homeless within NYC.

The HOPE counts are conducted in late January and early February witch appear to be the local minimums of the temperature graph. Their reasoning is likely based on the same intuition that if it is dangerously cold outside, then the homeless will seek shleter.

### Cleaning NYC population data "NYCpop"

### Cleaning HOPE count data "Hope"

The two tables are created as a result of the cleaning, "newdata" and "Yearlydata". "newdata" is for daily count data that will be used for estimating capacity needs of DHS shelters and evaluating at what time periods max capacity tend to be reached. "Yearlydata" reports point to point estimates fo populations and will be used to help give estimates of future population growth

View(Yearlydata)

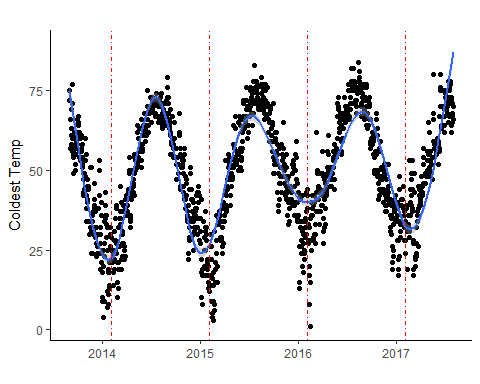
## *Part II*

Next lets visualize the suspected pattern of temperature and population fluctuation in DHS shelters. This is helpfull in matching population growth with peak seasonal needs for available beds.

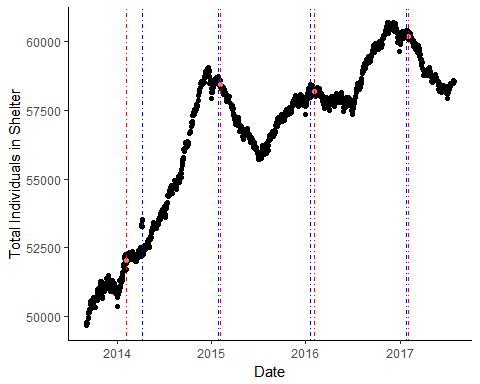
The points of highest demand closest to the coldest time of year are newdata$Date[min\_avg].

#see how avg dates match smoothed valleys  
ggplot(newdata, mapping = aes(x=Date, y=newdata$TMIN))+  
 geom\_point() +  
 geom\_smooth()+  
 geom\_vline(xintercept = newdata$Date[min\_avg], linetype=4, color = "red")+  
 mytheme+   
 labs(title ="",x="",y="Coldest Temp")

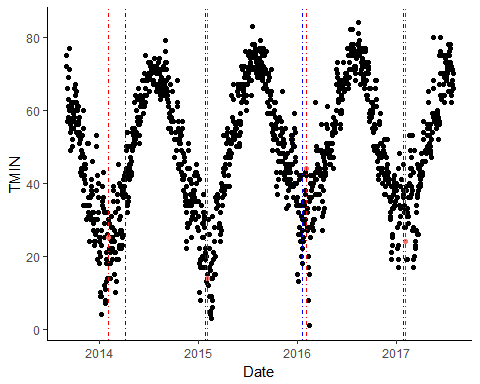
## `geom\_smooth()` using method = 'gam'

 The average dates appear to match relatively well with the local minimums. Next lets look at how the peaks of homeless demand match with these dates. Additionaly I don't want to just take the reported count on the specific day due to high variance. Instead I want to take an average over a 30 day window, +/- 15 days.

#show the mean homeless data on the daily pop data as an overlay  
ggplot(newdata, mapping =aes(x=Date, y=`Total Individuals in Shelter`)) +  
 geom\_point() +  
 geom\_point(peaks, mapping =aes(x=Date, y=mean\_peaks, color= "red")) +  
 mytheme+  
 geom\_vline(xintercept = newdata$Date[min\_avg], linetype=4, color = "red") + #minimum temperature  
 geom\_vline(xintercept = newdata$Date[pop\_pk], linetype=4, color = "blue")#peaks of homeless population



#Show the valley points on temperature data  
temp\_min <- newdata[min\_avg,]  
  
ggplot(newdata, mapping= aes( x= Date, y= TMIN))+  
 geom\_point()+  
 geom\_point(temp\_min, mapping=aes(x=Date,y=TMIN, color="red"))+  
 mytheme+  
 geom\_vline(xintercept = newdata$Date[min\_avg], linetype=4, color = "red")+  
 geom\_vline(xintercept = newdata$Date[pop\_pk], linetype=4, color = "blue")

 We can see that the dates for demand peaks match very well with tempurature valleys with the exception of 2014. 2014 is in an unusual growth period with some outliers pulling it to the right.

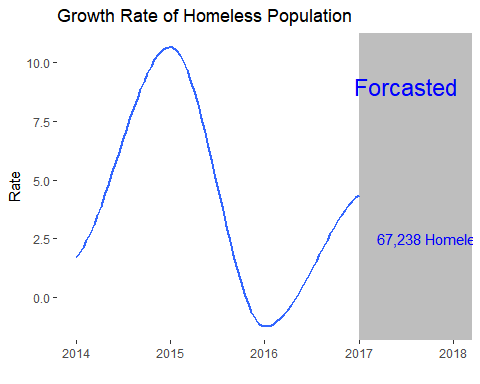
After cleaning I used the point to point estimates of growth to project 2018 populations and growth.

estimates

## Years\_estimates unsheltered\_estimates sheltered\_estimates Total\_Homeless  
## 1 2014 3357 52052.52 55409.52  
## 2 2015 3182 58432.90 61614.90  
## 3 2016 2794 58208.71 61002.71  
## 4 2017 3892 60192.06 64084.06  
## 5 2018 NA NA Inf  
## homeless\_perc\_growth NY\_perc\_growth adjusted\_growth  
## 1 2.2788779 0.5880705 1.690807  
## 2 11.1991361 0.5254019 10.673734  
## 3 -0.9935803 0.2485880 -1.242168  
## 4 5.0511770 0.6839897 4.367187  
## 5 Inf 0.4274039 Inf

#Visualize adjusted growth rates by year  
ggplot(estimates, mapping = aes(x=Years\_estimates, adjusted\_growth))+  
 geom\_rect(data=NULL,aes(xmin=2017,xmax=Inf,ymin=-Inf,ymax=Inf),  
 fill="grey")+  
 geom\_rect(data=NULL,aes(xmin=-Inf,xmax=2017,ymin=-Inf,ymax=Inf),  
 fill="white")+  
 geom\_smooth()+  
 annotate("text", label = c("Forcasted", "67,238 Homeless"), x = c(2017.5, 2017.8), y = c(9, 2.5), size = c(6, 4), colour = "blue")+ labs(title ="Growth Rate of Homeless Population",  
 x = "",  
 y = "Rate")

## `geom\_smooth()` using method = 'loess'



The forcasted growth rate is relatively naive as it doesn't consider political issues such as policy changes or election influences that might impact DHS.

write.csv(estimates, file = "estimates.csv")

## Conclusion

**Results** The total homeless population is increasing, but at a fluctuating rate that is difficult to predict. It appears that 2013 and 2014 may have been the years where the population grew in "crisis like" rates that levels off after 2015. The total population of homeless is a summation of street homeless and those in shelters. If capacity is always assumed to be at 100%, then we can also say that capacity fluctuates with the homeless population. The cost of emergency homeless shelters is higher than regular contracts. Therefore, if DHS can accurately plan for increased capacity, then they can reduce costs and use taxpayer money more responsibly. Growth rates appear to be stabilizing in recent years and NYC should plan to serve an estimated 67,238 homeless individuals this february.

One consideration to that needs more data to evaluate is that policy changes lead to the high growth rates seen before 2015. The data is just not available to be able to test that hypothesis. The analysis done here is recommended as a framework to track the homeless population annually as well as a way to test future policy changes. Knowing trends before and after policy implementation can help determine causal effects of policy and guide more effective political strategies to handle the growing problem that is homelessness in NYC

**Limitations** Our data spans 5 years, but since we are using annual estimates the total number of observations is 5 with 3 being supported by earlier estimations using linear modeling in at least one of the values used to calculate the adjusted growth rate. The total population of homeless is a number that is difficult to quantify accurately and relies on questionable data and collection techniques. Data was only available for the end of 2013 to September 2017. To have better more accurate results we need data from at least 2008. Our results only represent the trends in the data we analysed and should not be used to definitively describe the homeless population. The best statement we can make is that the population is at least as big as we observed.

Since data was not collected on homeless population giving unique identifiers it is impossible to determine when people are entering or exiting the homeless shelter system. This would give a better visual on the dynamic nature of the problem. It is also possible that past policies gave aid to homeless and transitioned them into permanent housing which would remove them from the Daily Counts in the shelters only to have them re enter at a later date. If this happened in 2013 and 2014, it might explain the high growth rate. [The Coalition for the Homeless](http://www.coalitionforthehomeless.org/basic-facts-about-homelessness-new-york-city/) count “more than 129,803 different homeless” individuals in fiscal year 2017, although meta data is limited to their data. Although this gives evidence to the fact that there are a large number of people each year moving through the system and just because the proportion seen each day is rising at an alarming rate doesn’t mean the total population is. This difference could be explained through policy changes or migration habits. Either way the data to answer more specific questions on about the homeless population is not easily found if it is available.

This project was completed for the Data Frameworks and Method course (APAN5200) in the [MS Applied Analytics](http://sps.columbia.edu/applied-analytics/master-of-science-in-applied-analytics) program at Columbia University on December 8, 2018 and edited on April 4, 2018.