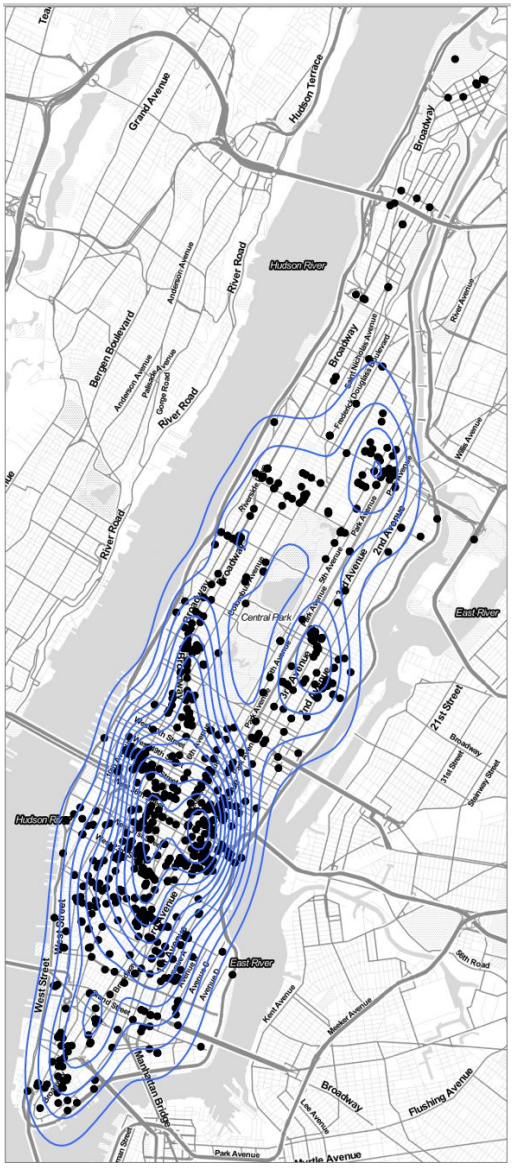
EXPLORATORY DATA ANALYSIS AND VISUALIZATION
Team: MASALA NOODLES

Data-Driven Analysis to assist Homeless People in New York City

"The homeless people's suffering belongs to amusement of our political order under a game over the right of marginalised group being transformed into citizens for merely punishment and humiliation. The Public Space Protection Orders is a penalty over one's condition suffering – it is a fine over the disempowered for being disempowered. This act allows power to fragment the homeless into sub-humans punishable for the state of utter misery."
— Bruno De Oliveira



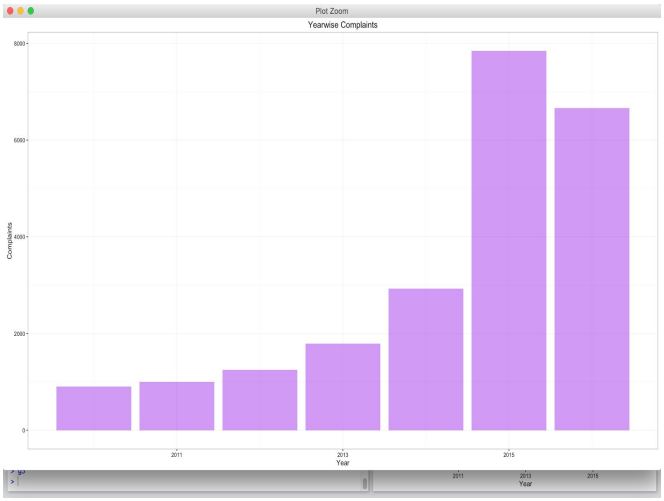
The above image is a density plot of homeless complaints in manhattan area of New York City in the year 2015.

Homelessness has been one of the most pressing problems in the world and in particular, New York city. One of the recent reports suggested that the homelessness has reached its peak in NYC this time since great depression. We felt that this is a serious issue and wanted to analyze it in more depth using 311 and other datasets that are at our disposal. Since it is a multi-layered problem, we want to analyze it and provide humane solutions to resolve this grave issue. To start with, we perform analysis of the correlation of number of homelessness related complaints with different factors such as economic conditions, health conditions, subway stations, police stations, and weather. We also performed overall analysis of number of homeless complaints with other complaint types in the city on a particular day. Instead of restricting ourselves to the analysis, we move a step ahead by proposing few predictive models. In one model, we wanted to predict the

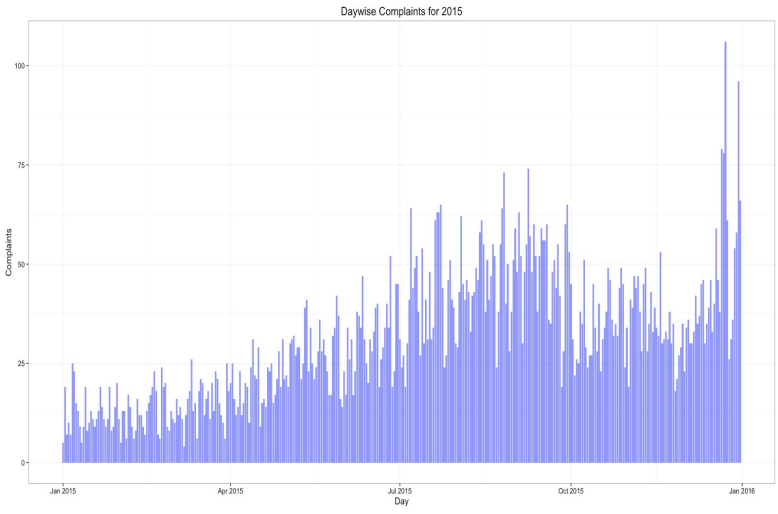
expected number of complaints on a given day using available data. We have come up with a model to analyze the resolution times of homeless complaints in search of any valuable insights. We have also performed cost-to-actual ratio of several areas of NYC using several metrics to build a predictive model. The aim of this analysis is to identify under-priced locations that can be potentially converted to facilities for homeless people. We present our findings in the following sections. We present our analysis of homeless complaints in the first section. In the sections that follow, we present our models and report the scores of each model. We conclude the report by presenting our analysis and directions for further analysis that we think will be a step in the right direction towards mitigating the homeless complaints in a humane way.

Analysis of homeless complaints

Distribution with time:

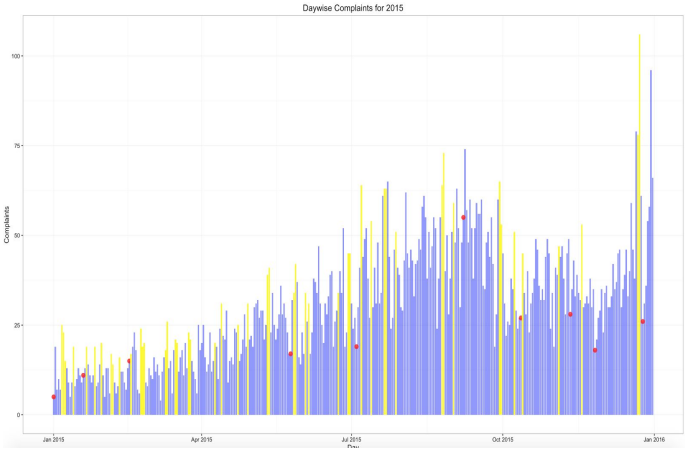


In the plots beside, we map number of homeless complaints from the year 2010. We can see that the number of complaints is increasing steadily with every year. The complaints in first 5 months of the current year is already coming close to total number of complaints received in 2015 . The reasons could be the actual increase in the number of homeless people, fall in the number of available jobs in NYC and/or increasing awareness of the 311 service for homeless complaints.



In the next plot, we look at the number of complaints in the year 2015. The complaints peak during the end of summer months and then we see a fall in the average number of complaints. There are few steep peaks at the end of the year, around the holiday season.

Distribution with Holidays:



Let's look at the holidays in the year 2015 and how the homeless complaints were. In the plot to the right, all the blue bars indicate each day of the year in 2015. The red dots indicate the holidays. These are the list of holidays provided by the state of New York. We built an algorithm to find the abnormal peaks in number of complaints i.e. to point out the days when a sudden change was observed. The yellow bars indicate these sudden peaks, and when correlated with the holidays of the year (red dots), we see that the peaks are observed few days before and after the public holiday but **not on the day** of the public holiday. This can be clearly seen in the case of Christmas Day, Thanksgiving Day, July 4th, etc. There is more than *70% change* in the number of complaints closer to the holidays. The reasons that we can think off are increasing amount of pedestrian activity during the holiday season, people could be more observant or vice-versa homeless people could be more adamant and less sober.

CartoDB visualization:

We first present a CartoDB animation of the homeless complaints onto the New York City map from 2010 onwards. The animation can be found at the link

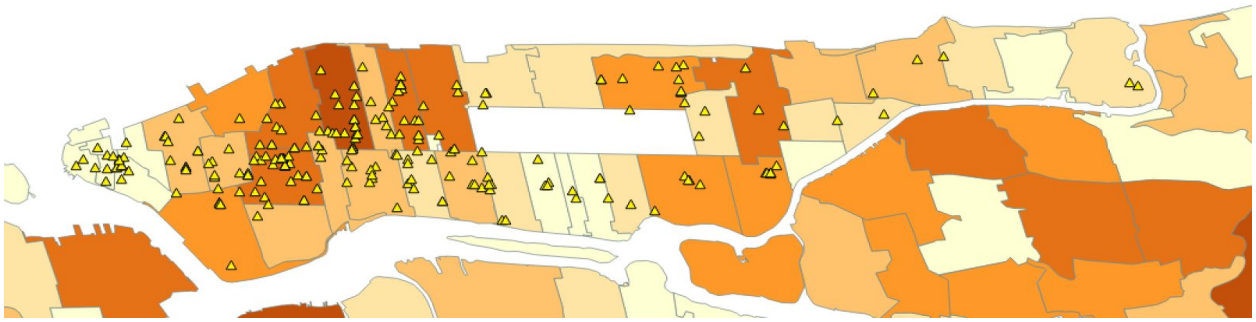
CartoDB Animation

We can see that although the complaints are spread across the city but there are a lot of incidents in Manhattan. One direct intuition from this is that homeless people tend to move in the areas of financial centres/ public places/ tourist attractions with an expectation of making more money through alms. From the plots, we have also observed that the complaints on the map are close to the subway lines. In the next few parts of the report we'll look at the different factors affecting homeless complaints.

For the sake of simplicity we will look at the influence of various factors - average income of an area, crime rate, number of subway stations, police stations in a particular zip code zone and their effect on homeless complaints. We have also restricted ourselves to look at complaints only in manhattan area and in the months of January and August to factor in the climate's influence. Every sub-section starts with a map of

the complaints with respect to the topic of interest in the month of January above the text and the month of August below the text.

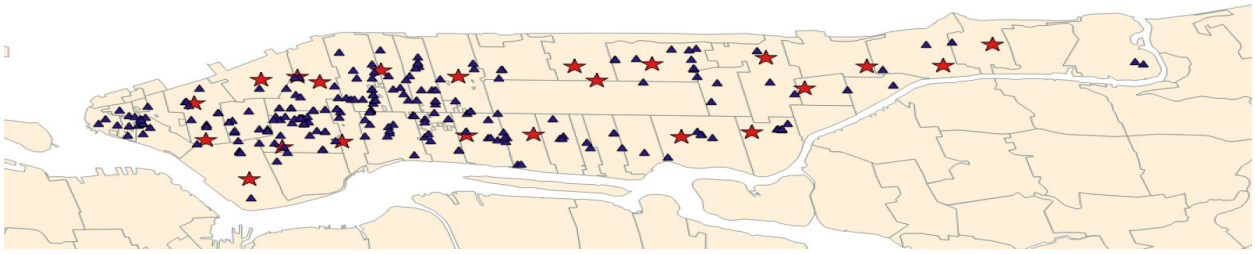
Distribution with Crime Rate:



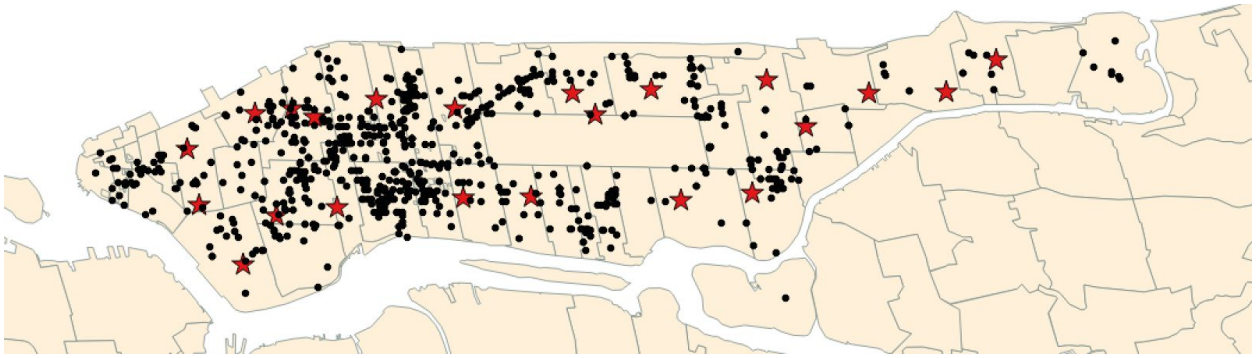
The above plot is for the month of January and the one below is for August 2015. All the zip code zones of Manhattan are color coded with their Crime Rate. Higher the intensity of the color, higher the crime rate in the zip code zone. We do not find any direct correlation of number of complaints with the crime rate which does not give a compelling evidence to break the myth of the crime rate affecting the homeless encampments.

However, there are few areas where higher the crime rate, higher the number of homeless complaints. And, since homeless complaints are unaffected by crime rate, we cannot generalize the effect of crime rate on the number of homeless complaints. There is some correlation of complaints in Manhattan and crime rate which can be used as a feature for predicting the homeless complaints.

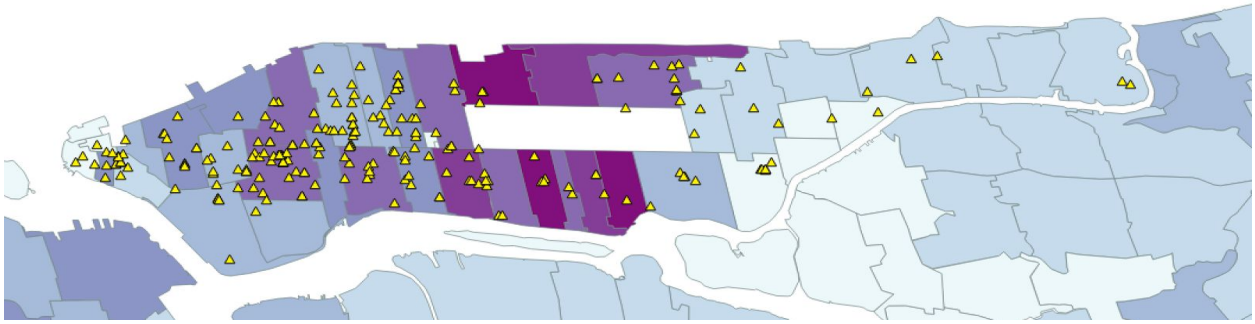
Distribution with Police Stations:



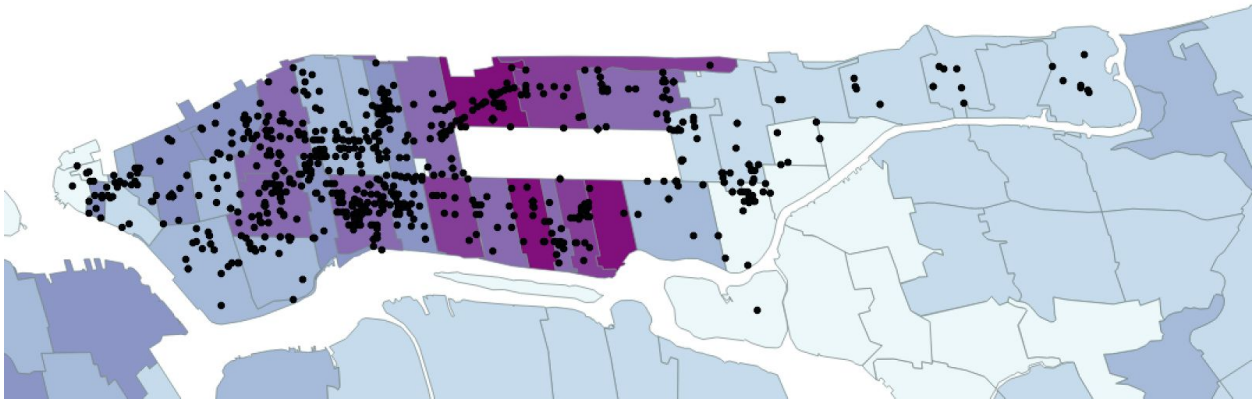
We found the data for Manhattan NYPD Stations in address format. After converting the data to geo-coordinates, we can the see the red stars on the two maps. These are the Manhattan NYPD Stations. The above plot is for the month of January and the one below is for August 2015. We can see a trend that the homeless complaints are seen at some distance from the NYPD stations. As the homeless complaints come under NYPD department, this can be expected. The distance to police station can be an important predictor to analyze the future occurrence of homeless complaints.



Distribution with income:



New York City also releases the data for the average income for the households of a zip code zone. In the plots, it is no surprise that the wealthier regions in Manhattan are the Upper and Lower West and East sides. The above plot is for the month of January and the one below is for August 2015. There is not a very strong correlation, as most of the complaints are centered around the Midtown area which isn't really the wealthiest of the zones but at a more granular level we can see that more clusters fall into the purplish zones denoting income on the higher. We will look into income as a factor more when we do the predictive modeling part to see how important a factor it is in reality.

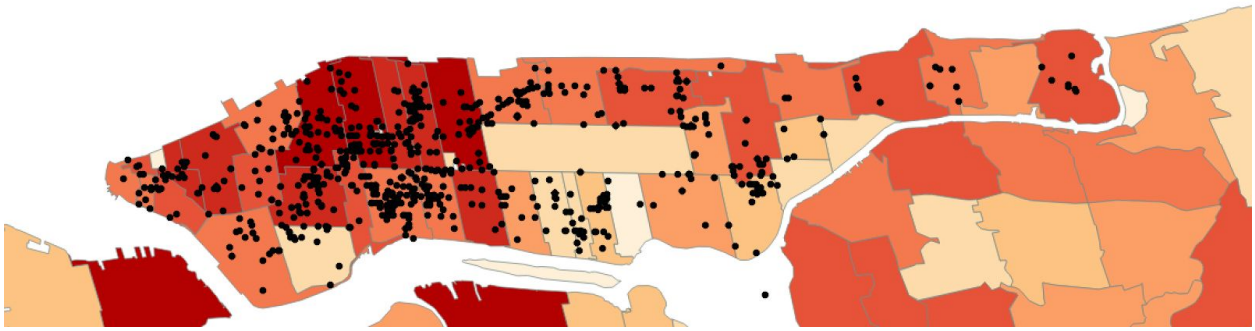


Distribution with Population:

We generally think, more the population of an area, more the people on the streets, and more the possibility of earning alms by the homeless. The above plot is for the month of January and the one below is for August 2015. Higher the intensity of the color, higher the resident population of the area. The plot does not indicate any direct correlation in the two factors.

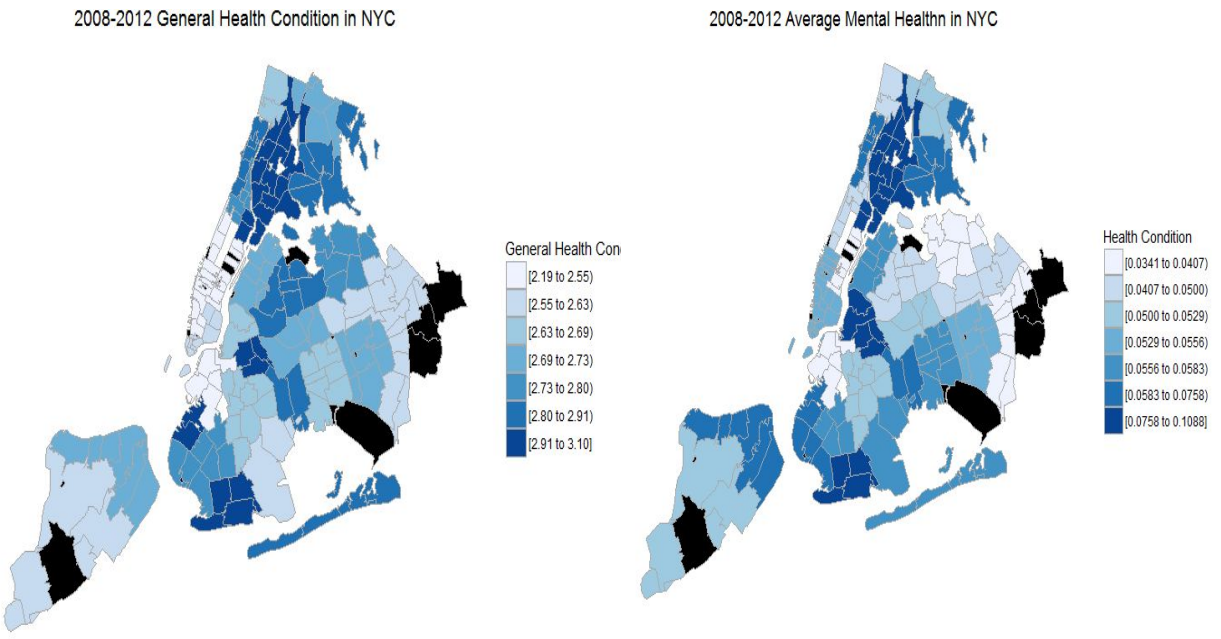
Distribution with Subway stations:

As we saw in the CartoDB map, the homeless were found around the subway lines. We obtained the data for subway stations, which was point data and aggregated it into zip code zones. Higher the intensity of the color, more the number of subway entrances/stations in the area. It is pretty evident that higher the number of subway stations in the area, more is the number of homeless complains. We know that subway stations indicate the footfall of an area. From this, we can deduce that more the stations, more the tourists and people walking around in the area and more the possibility of earning alms by the homeless.



Distribution with mental health:

According to *coalitionforthehomeless.org*, there are research studies showing that the large majority of homeless people in New York are suffering severe health problems and living under mental illness as well. In the spirit of finding empirical evidence from the historical data, we analyze and visualize the general health condition as well as mental health condition in the zip-code level. The data is collected from the New York City Community Health Survey (CHS), which is a yearly telephone survey with an annual sample of about 8500 individual adults from five boroughs of New York City (Manhattan, Brooklyn, Queens, Bronx, and Staten Island). The respondents are asked to give a grade on the general health condition from 1 to 5, which lower score indicating a better health. In addition, respondents are supposed to answer a list of questions regarding the stress level and mental health. A final score (0 or 1) indicating whether the subject is suffer mental health issues is computed based on the answers to those questions. Similarly, lower score indicates a better mental situation.



In the above plots, we plotted the average general and mental health condition from 2008 to 2012 respectively. One can find similar pattern in both graphs. Darker areas are associated with worse health condition zip-code zones. In both plots, the Harlem-Bronx area and the area on the boundary of Queens and Brooklyn are the zones that attracts

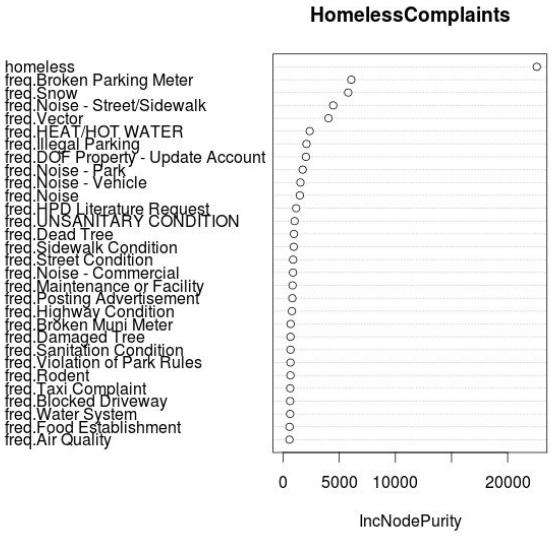
the attention. Both clusters suffer from severe health issues and mental illness. Combines with previous plots on the distribution of homeless complaint counts, one can find that the clusters overlaps a lot with the health condition maps here, which provides evidence that health condition distribution are somehow related with the homeless distribution.

Prediction Models

As mentioned in the introduction, in this section we look at 3 prediction models that we have designed and built. The first model looks at the number of homeless complaints for each day and predicts the *expected number of complaints*. This model helps the government departments to be prepared for the complaints on a given day. In the second model, we analyze resolution times for homeless complaints and come up with a model to predict the *resolution time of the complaint*. The main motivation for this model is to identify areas where the resolution times are less so that homeless people can have a temporary stay. We finally build a model based to identify *under-priced areas in NYC* that can be potential areas to build help centres for homeless people.

Number of complaints:

We look at number of complaints of different complaint types on a previous day to come up with an expected number of complaints for a given day. We fit a linear regression model and observe the R-squared value to determine the goodness of fit. We also fit a random forest to identify the important predictors for the number of homeless complaints.



The R-squared value for the linear regression model is **0.8353**. This shows that the model fits very well with the observed variance in the number of complaints. Using the same dataset, we build a random forest to draw a variable importance plot that is shown in the figure beside. We can see that the number of complaints in the current day is a very strong predictor for the expected number of complaints on the next day.

Effect of weather on number of complaints:

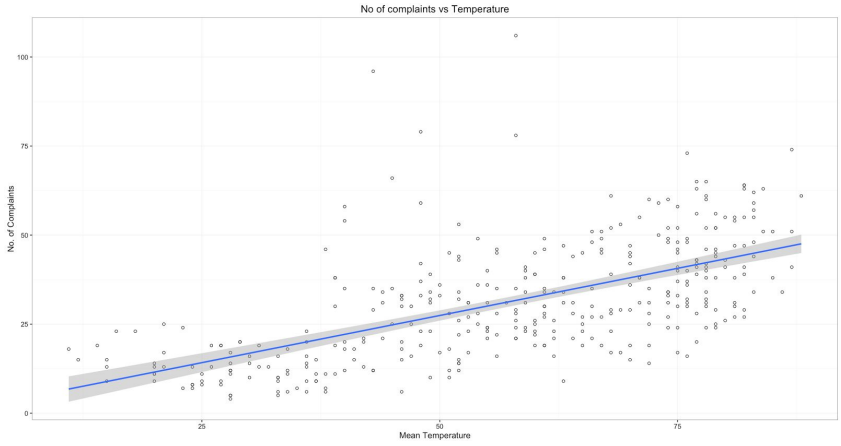
We scraped https://www.wunderground.com/ to collect weather information (mean temperature, minimum temperature, maximum temperature) on each day of the year 2015. We then combined the weather data with the total number of complaints on each day of the year 2015 to fit a linear regression model.

The results and the general fit of the linear regression model are as below:

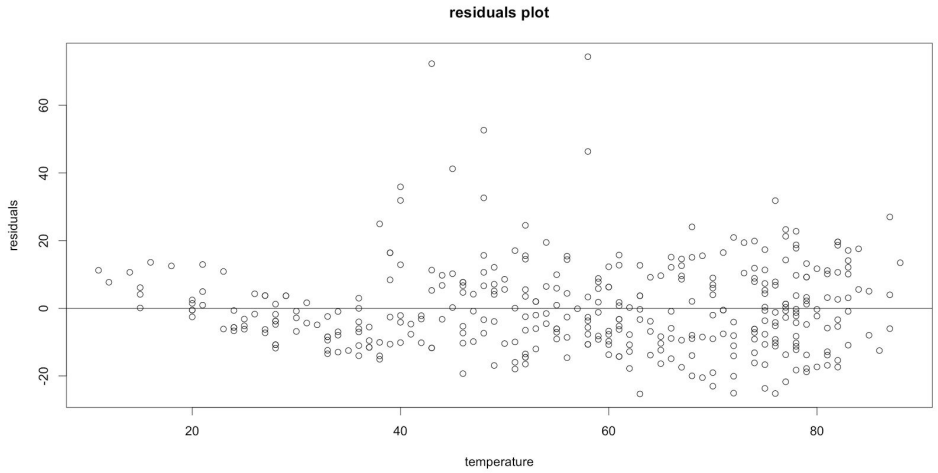
```
# Call:
#   lm(formula = total ~ temp, data = day_wise)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -25.311  -9.018  -2.018   7.334  74.337
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.94913    2.17995   0.435    0.664
# temp         0.52956    0.03622  14.619   <2e-16 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 13.15 on 363 degrees of freedom
# Multiple R-squared:  0.3706,     Adjusted R-squared:  0.3688
# F-statistic: 213.7 on 1 and 363 DF,  p-value: < 2.2e-16
```

We note that **p-value: < 2.2e-16** is very small and hence we can ignore the null hypothesis that expected number of complaints do not depend on weather at all.

The correlation value was ~0.4 between mean temperature and number of complaints per day. The regression line is plotted in the figure below:



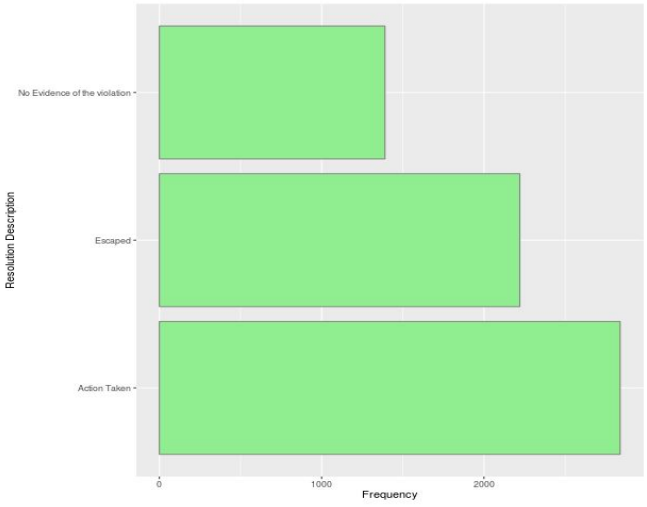
Below is a plot of the residuals varying with temperature.



After adding weather to the complaint types, we have fitted a new linear regression model to predict expected number of complaints on a given day using the number of complaints of each complaint type and the mean temperature on the previous day. This model performed better than the model without temperature with the R-squared value of **0.8851.**

These model help to predict the expected number of complaints on a given day which will certainly help the government departments to come up appropriate measures to handle the complaints.

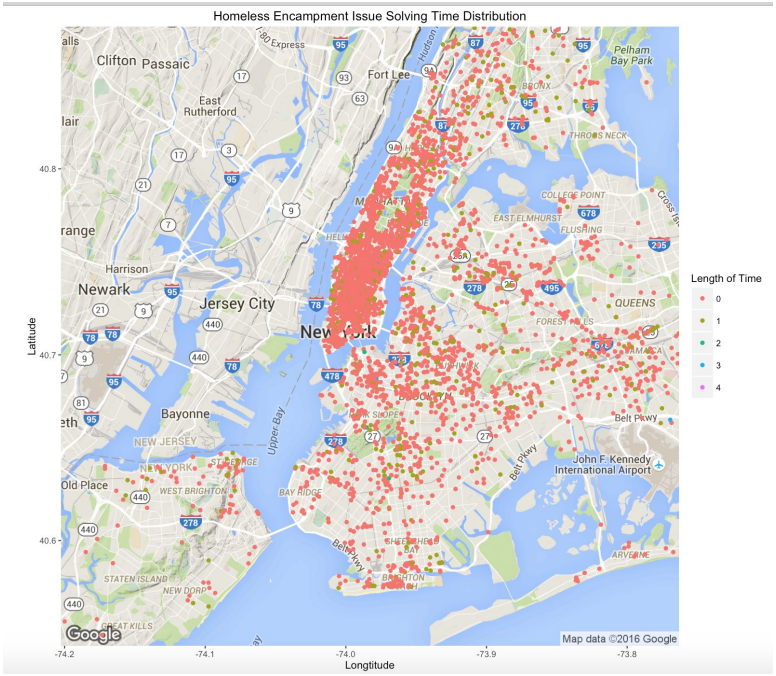
Resolution Times for Homeless Complaints:



We have observed that homeless complaints are closed very quickly with most of the complaints closed on the same day. There are predominantly 3 types of resolution descriptions for the homeless complaints as shown in the figure to the right, with most of the complaints being resolved with some action taken. As the homeless complaints come under the department of NYPD, this is also one of our main motivation to understand the types of

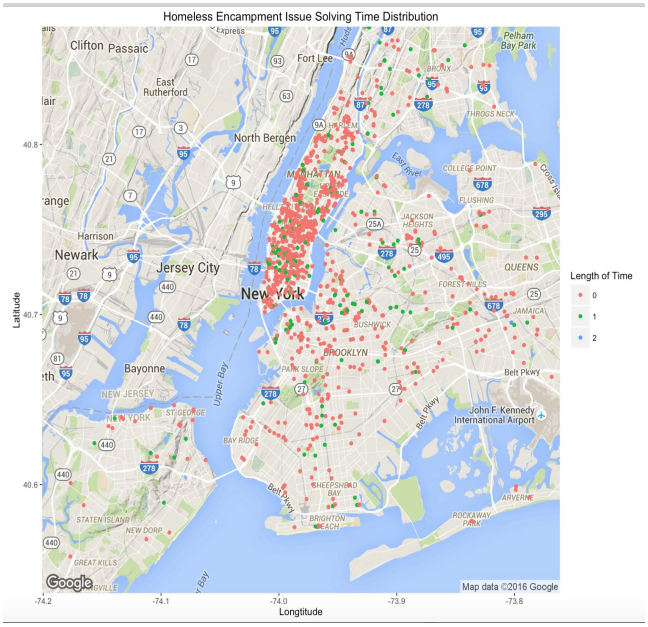
complaints and provide humane solutions.

We built a model to predict the resolution time of homeless complaints. Here we focused our analysis on “Homeless Encampment Issue”. First we plot a map to see the resolution time distribution for “Homeless Encampment Issue”.



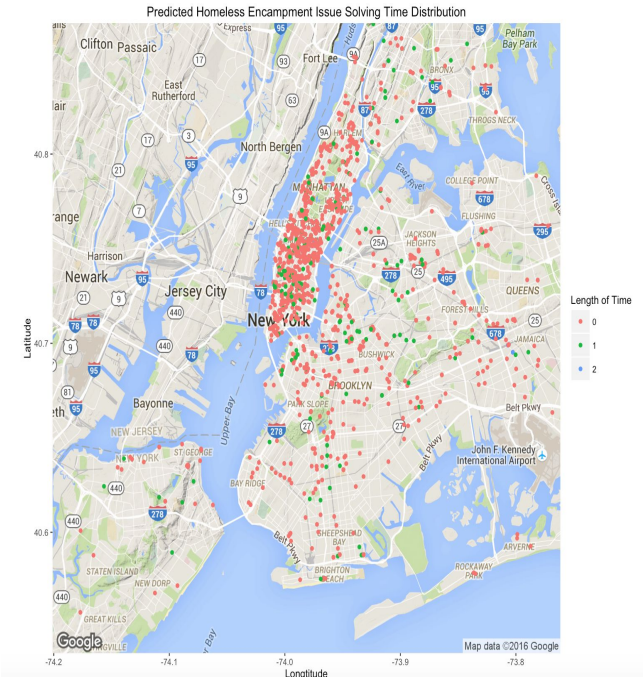
Homeless people (HP) prefer staying at a "place" for longer time. We can use this plot to help homeless people find such a place that is less vulnerable to NYPD. It is obvious that HP should build their encampment at the places that are not red in the graph since they will be "kicked out" at the same day! It is possible that HP can find a place in downtown area that has 1-day "time".

Next we build a model using random forest algorithm to predict the resolution time. Here is the details of the model:




```
##         OOB estimate of  error rate: 12.26%
##               Confusion matrix:
## 0 1 2 3 4 class.error
## 0 5196 0 0 0 0       0
## 1  712 0 0 0 0       1
## 2    5 0 0 0 0       1
## 3    8 0 0 0 0       1
## 4    1 0 0 0 0       1
##                                MeanDecreaseGini
## Address.Type.                          0.25609753
## Address.Type.ADDRESS                   0.92669095
## Address.Type.BLOCKFACE                 0.67528660
## Address.Type.INTERSECTION              1.05878878
## Address.Type.LATLONG                   0.00000000
## Address.Type.PLACENAME                 0.61788033
## Borough.BRONX                          3.63351055
## Borough.BROOKLYN                       1.26237690
## Borough.MANHATTAN                      4.10548209
## Borough.QUEENS                         0.86104820
## Borough.STATEN.ISLAND                  1.19250659
## Location.Type.Bridge                   0.14573281
```

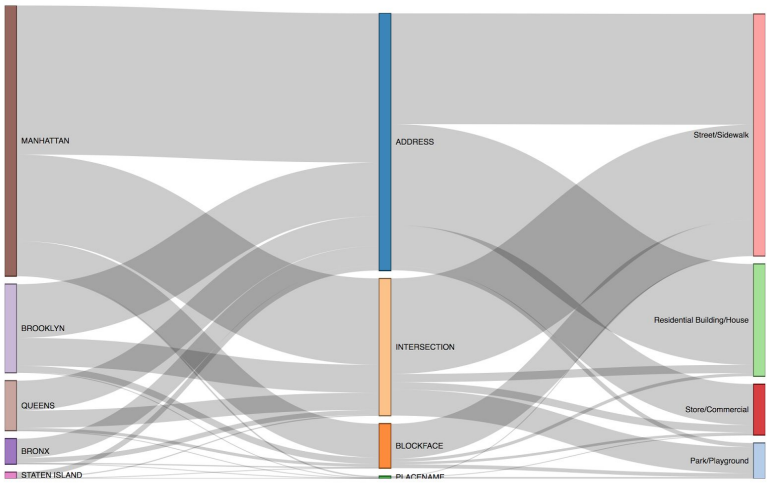
```
## Location.Type.Highway                          0.22011861
## Location.Type.Park.Playground                  0.93422292
## Location.Type.Residential.Building.House       1.78301346
## Location.Type.Roadway.Tunnel                   0.03695538
## Location.Type.Store.Commercial                 1.33736323
## Location.Type.Street.Sidewalk                  1.01798239
```



The OOB estimate of error rate is 12.26% (pretty decent), which is unbiased for the test set with the same size as the training set. From the importance level, we can see that the factor 'Borough' has the biggest impact on the resolution time (Manhattan and Bronx have biggest MeanDecreaseGini). 'Addresss type' and 'location type' also have impact on resolution time especially the 'ADDRESS' type under 'Addresss type' and 'Residential.Building.House' type under 'location type' have high power on predicting our reFsponse variable (We also tested that Agency.Name have no impoact. It has 3 levels).

From the visualization perspective, we used a test set (not used in training) to see how big the difference is between original result and predicted result of this test set on a map. From the maps (the above two maps on the right) we can see that they look pretty similar.

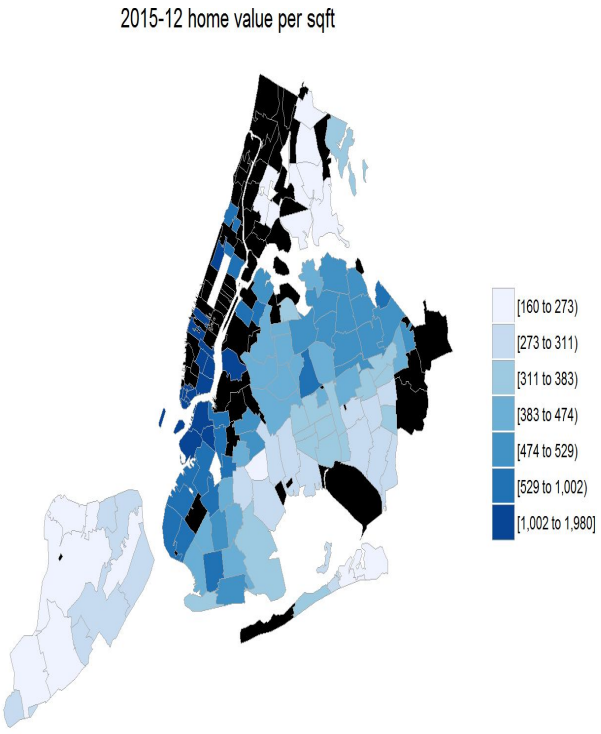
Finally, since we found that "Borough", "Address type" and "location type" are three most important factors, we plotted a sankey plot to show how these complaints flow between these three factors. Here is the plot:



New Homeless Centers Location Targets:

In this part, we assume that we can build new homeless centers to help the homeless people with a set of services such as free/cheap food/clothing storage and delivery, mental treatment etc. To find the best location, we want to find a location with lower home value because it is better to reduce the cost of this non-profit organization. However, sometimes the cheap areas are also disturbed by crimes and other complaints. In this section, we are trying to build a home value regression model to take into consideration the living condition in each area and predict the value that each area worths.

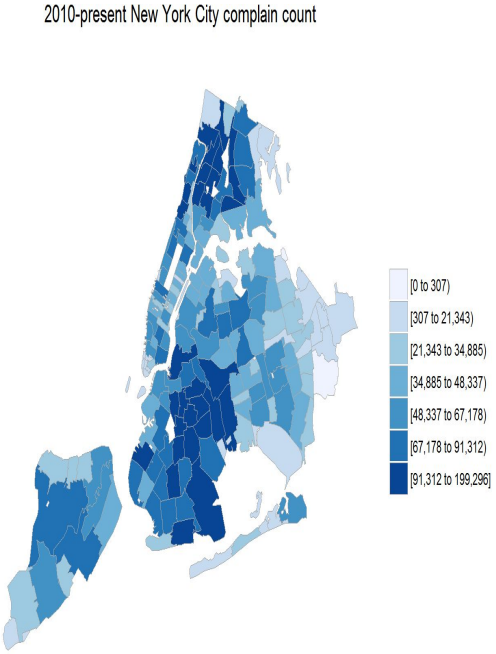
The plot beside shows the home value (price per square feet) distribution in New York City, and the black zones are missing. (Data source: Zillow data). As we can see, many areas in Queens, Brooklyn, Bronx and Staten Island are not so high. Those maybe our potential candidates for the location of the center.



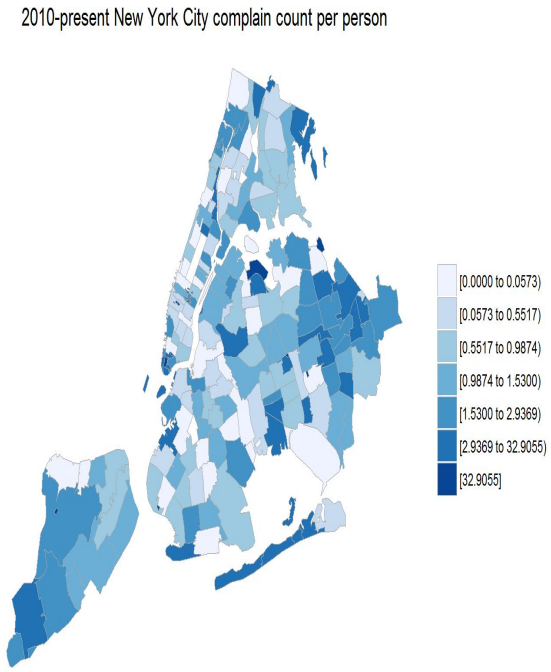
Variables taken for consideration

Complaints: We regard the number of complaints as a negative factor to the price because no one wants to pay for a room with a lot of living problems on a high price. We basically want to see if this factor is significantly influencing the prices. The following s plot shows the distribution of the total number of the top 5 complaints in each area.

So similar to our analysis, we will use the population-standardized data for other numerical variables. The data source for population is the 2010 US Census. The two plots can be seen to the right of the text and in the next page.



Income: The income of a given area is highly correlated with the home value on the assumption that people will choose their home locations according to their income level. We can also verify this by doing a correlation test on the given data. As we can see, they are highly correlated and hence we can use this as a predictor.




```
##
##  Pearson's product-moment correlation
##
## data:  pprice$price and pprice$avg_income
## t = 8.5426, df = 121, p-value = 4.641e-14
## alternative hypothesis: true correlation is not equal to
0
## 95 percent confidence interval:
## 0.4894743 0.7129744
## sample estimates:
## cor
## 0.6133588
```

Others: We also want to include the crime rate and transportation condition (using Metro station numbers as proxy) to predict the price.

In this part, we use a linear regression model with stepwise variable selection to fit the data, the equation is

$$price=\beta_0+\beta_1 complains+\beta_2 income+\beta_3 crimes+\beta_4 transportation$$

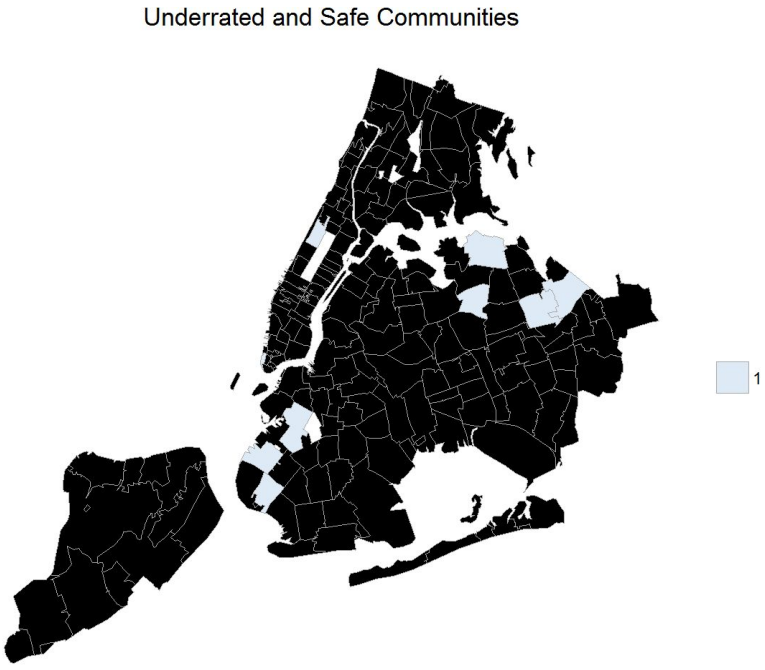
The model fitting result is

```
## Call:
## lm(formula = price ~ avg_income + avg_metro + crime_rate, data = pprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -675.89 -183.14  -68.81   79.83  991.03
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.783e+02  4.460e+01   6.239 6.97e-09 ***
## avg_income  2.745e+00  6.117e-01   4.488 1.67e-05 ***
## avg_metro   8.125e+04  3.954e+04   2.055  0.04208 *
## crime_rate  1.695e+04  4.538e+03   3.734  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275.7 on 119 degrees of freedom
## Multiple R-squared:  0.5049, Adjusted R-squared:  0.4924
## F-statistic: 40.45 on 3 and 119 DF,  p-value: < 2.2e-16
```

The final model is significant but eliminates the effect of the complaints. And then we use this model to predict the actual value of the areas and select the areas whose prices are underrated by more than 15 percent with crime rate and number of complains below median. The regions are in the list below and their locations are shown in the following plot.

```
## [1] "11355" "11362" "11364" "11357" "11228" "11220" "11215" "10280" "10025"
```



The regions that are marked white in the above plot can be potential places w.r.t cost analysis to build centres to help homeless people.

Conclusion

With the help of 311 dataset and external datasets, we performed a comprehensive analysis on the complaints related to homelessness. Without restricting ourselves to analysis, we went a step ahead to build models that can be helpful in predicting the expected number of complaints on a given day, complaint resolution times, and cost analysis for various zip codes in NYC to suggest potential areas to build help centres for homeless people. We hope that this analysis is useful and will provide the necessary insights to take steps in the right direction. However, we do think that there is a lot of work to be done to come up with a more humane solution to resolve the problem of homelessness and it doesn't end with building a help centre but a comprehensive approach towards education, employment, and awareness. The appendix reports few facts that we came across in the news related to homelessness in New York City.

Appendix

Some facts about homeless people of NYC according to media reports

New York City Homelessness: The Basic Facts

- In recent years, homelessness in New York City has reached the highest levels since the Great Depression of the 1930s.
- In February 2016, there were 60,144 homeless people, including 14,654 homeless families with 23,424 homeless children, sleeping each night in the New York City municipal shelter system. Families comprise just over three-quarters of the homeless shelter population.
- Over the course of the last City fiscal year (FY 2015), more than 109,000 different homeless men, women, and children slept in the New York City municipal shelter system. This includes over 42,000 different homeless New York City children.
- The number of homeless New Yorkers sleeping each night in municipal shelters is now 91 percent higher than it was ten years ago.
- Research shows that the primary cause of homelessness, particularly among families, is lack of affordable housing. Surveys of homeless families have identified the following major immediate, triggering causes of homelessness: eviction; doubled-up or severely overcrowded housing; domestic violence; job loss; and hazardous housing conditions.

- Research shows that, compared to homeless families, homeless single adults have much higher rates of serious mental illness, addiction disorders, and other severe health problems.
- Each night thousands of unsheltered homeless people sleep on New York City streets, in the subway system, and in other public spaces. There is no accurate measurement of New York City's unsheltered homeless population, and recent City surveys significantly underestimate the number of unsheltered homeless New Yorkers.
- Studies show that the large majority of street homeless New Yorkers are people living with mental illness or other severe health problems.
- African-American and Latino New Yorkers are disproportionately affected by homelessness. Approximately 58 percent of New York City homeless shelter residents are African-American, 31 percent are Latino, 7 percent are white, less than 1 percent are Asian-American, and 3 percent are of unknown race/ethnicity.

Team members: (from left to right)
Xuyan Xiao, Shenghan Yu, Rohit Bharadwaj Gernapudi, Arushi Arora, Vishal Juneja, Jiannan Zhang