Team 8 Group Write Up
Group Members: Jarrod Daniels, Thomas Guaetta, Corey Kozlovski

Research Questions

When working on this project, our group came up with a number of research questions that we wished to investigate in order to gain better insights on the state of homelessness in the city of Boston and the surrounding Massachusetts Area. Specifically, the questions we tasked ourselves with trying to answering included:

- Is Massachusetts homelessness data in 2018 different from the rest of the US population?
- Is Boston's CoC data different from all non-Boston Mass CoC data?
- Can we predict the proportion of homeless for a given CoC based upon city and year?

Background & Significance

According to the U.S Department of Housing and Urban Development (HUD), an average of 553,000 people in America experienced homelessness on any given night in 2018. To add onto this concerning statistic, according to endhomelessness.org, in 2017 in the United States alone, more than 40 million people were living in poverty and 18.5 million were living in deep poverty which leaves them at risk of becoming homeless. One commonly stated fact is that homelessness is decreasing in the United States as seen in (**Figure 1**). However if you look at major cities around the United States, you can see that the number of overall homelessness is actually increasing (**Figure 2**). When looking at Massachusetts overall, we can also see that homelessness has been on the rise as well (**Figure 3**). Specifically in 2018, it was reported that there was a total homeless population of 20,068 (this means that 29 out of every 10,000 people are experiencing) homelessness. This was a significant 14.2% jump from 2017 and a 20.6% increase from 2010. On top of this, when looking at the distribution of homelessness in MA, it is clear the Boston holds the majority of the state's homeless population (**Figure 4)**. With homeless on the rise in Massachusetts and the fact that all of our group members have lived in the Boston area for the majority of our lives, we wanted the aim of our project to be on our home state with a special focus on the city of Boston.

Methods Used to Obtain & Analyze Data

For this project our group utilized two datasets. The first was the 2007-2018 Point-in-Time Estimates by CoC. This data set contains a counts people who experienced homelessness on a single night in January, taken by various CoC's across the United Sates. For context, CoC's are known as Continuums of Care which are local planning bodies that coordinate housing and services funding for homeless individuals and families. Furthermore the data goes further and categorizes those experiencing homeless based on various factors and demographics (such as Sheltered and unsheltered, veterans, youth etc.). The second data-set our group used in our study was the 2007 - 2018 Point-in-Time Estimates by State which holds similar data as the CoC data but contains overall state numbers rather than those of individual CoC's. This data was taken from the Housing and Urban Development database.

The main methods we used to analyze our data included hypothesis testing, T-tests (one & two tailed tests), and a multiple linear regression model. To start analyzing the data we first had to start by normalizing it in order to better compare different data values, since the raw values wouldn't work well for a comparison. To do this we used a proportion of the number of homeless people in each state/CoC rather than the raw number (Overall homelessness divided by Total Population). From there, we moved on to testing. While running the hypothesis tests, we decided to use an alpha value of 0.05, since that indicates a 5% risk of concluding that a difference exists when there is no actual difference; In this case, we were able to take that risk. The first test we ran was a hypothesis test on the overall proportion of homelessness in MA against all other states. Then, we decided to take a more focused look at how Boston's CoC compared to the rest of the MA's CoC's using another hypothesis test.

Finally, we moved on to creating a multiple linear regression model. To do this, we used historical data from four major city CoC's (Boston, NYC, San Francisco, and Seattle) provided by our data sets. However before we could construct the model we once again had to prepare it. To do so, we began by subtracting all of the yearly data from 2006 (starting at 2007 will greatly skew the intercepts of a model) and from there, took the logarithm of the difference to account for the fact that the human population tends to grow in an exponential like fashion over time. From there, we were able to fit the model using the proportion of the population that is homeless as our ŷ, city as $X_{1,}$ and year as X_2. Once the model was fit, we then created the appropriate graphs to check the assumptions of linearity.

Results of Analysis

The first questions we wanted to tackle was to see if MA's overall proportion of homelessness was in line with the rest of the states. To answer that question we used a Two-Tailed hypothesis; (H0: Mu = 0.0029075 & Ha: Mu != 0.0029075), Null = All non-MA states have the same overall homelessness as MA and Alternate Hypothesis = All non-MA states have a different overall homelessness as MA. Using the p-value of 9.535515e-08 and t-value of 6.23183. With the p-value being lower than the alpha we reject the null hypothesis about how MA's overall homelessness as not the same as the rest of the country.

After running the test and rejecting the null, we ran a one tailed hypothesis test, with a H0: Mu >= 0.0029075 and Ha: Mu < 0.0029075, in order to see if MA's overall proportion of homelessness was greater than or less than the rest of the state's proportion of homelessness. We calculated a T-Statistic and a p-value to further expand on our analysis and got the values of 6.170431 and 6.388e-08 respectively. With the p-value being less than our alpha value of .05, we are able to reject the null hypothesis that all non-MA states have a greater than or equal to overall homelessness as MA. We have statistically significant evidence to conclude that non-MA states have a lower overall homelessness proportion than MA itself.

Our second test was concerned with the question "Is Boston's CoC data different from all non-Boston Mass CoC data? Comparing Boston's CoC other large cities in the US, is the data different?" To best answer this question, we ran a hypothesis test comparing Non-Boston CoC's against Boston's CoC. First, we ran a two-tailed test with Null Hypothesis: All non-Boston CoC's have the same overall homelessness as Boston's CoC and an Alternate Hypothesis: All non-Boston CoC's have a different overall homelessness than Boston's CoC's. After running the test, we received a t-value of 6.20904 & a p-value of 4.524516e-05. With the p-value being much lower than our alpha of 0.05, we reject the null hypothesis that the average proportion of overall homelessness in MA equals that of Boston.

Afterwards, we expanded on this by running a one-tailed hypothesis test in order to figure out where Boston's CoC lies in comparison to the rest of Massachusetts' CoC's. For this portion, our null hypothesis was H0: Mu >= 0.008908942 and Ha: Mu < 0.008908942. We then calculated the T-Statistic and our p value, which were 6.20904 and 2.262258e-05 respectively. Once again, our p value was less than our alpha value of .05, so we are able to reject the null hypothesis that All non-Boston CoC's in MA have a greater than or equal to overall homelessness as Boston. We have statistically significant evidence to suggest that the non-Boston CoC's have a lower average homelessness ratio than Boston's CoC, which tells us that Boston has the highest homelessness ratio in comparison to the rest of MA.

For the result of the multiple-linear regression model, we first needed to check that the assumptions of linear regression held. For the first assumption we checked linearity by creating a scatterplot of residuals for our year variable (**Figure 5a**). We can see that for the most part, the residuals appear to be to be randomly distributed across the line however taking a closer look we can see that there does appear to an arc in the data. This could possibly be accounted for based upon the fact that the data was taken over time rather than collected at once. Since our city variable is categorical, we can assume that it is necessarily linear and therefore passed. Next we had to check for constant variance which could be done by creating a plot of residuals against predicted values (**Figure 5b**). Upon examining this we can see that there is a large gap missing in the data which is most likely the result of the amount of data the model is built off. As a result it is difficult to determine whether or not the test passes. Moving on to the third assumption of Independent observations, we can also assume this to pass as the data from one City's CoC should have been taken independently of another and should not have impacted each other. Finally Moving on to the final assumption of normality of residuals, we constructed a normal probability plot (**Figure 5c**). As we can see the data appears to follow the line for the most part however it is important to mention that the data appears to veer off towards the tails. Taking into account all of these assumptions, we can look at the output of our model (**Figure 5d**). Looking at Boston (intercept) specifically, we can see that when X=0, Boston is expected to have a homelessness proportion of 0.00784 and that as year increases by one (2007 as our starting year), that value is expected to increase on average by 0.00063. On top of this, the r value for the model is far below our chosen alpha value, meaning that there is a significant relationship between the predictor variables and the predicted outcome. On top of this, the R squared value was 0.95 which is significantly high meaning the predictors account for about 95% of all variance for the response variable.

Discussion

Based on our research, we concluded that Massachusetts does on have a higher proportion of homelessness than the average of the United States. This could possibly be explained by the fact that Massachusetts contains Boston, which is a major hub city not only in the country but the world as a whole. To add onto this, Boston clearly has a higher proportion of homelessness than other areas of Massachusetts. As for the multiple linear regression models, some of the assumptions checked did not appear to clearly pass and a result we must be careful when interpreting its results. However looking at the model we can see that it is predicted that homelessness in the city will increase as time goes on.
In the future, it would be interesting to see how the model changes if we included more data and more predictor variables. It would also be interesting to conduct further studies on other parts of Massachusetts that are known to have high poverty rates such as Springfield and Lynn. It would also be insightful in future studies to look at demographics within the homeless population such as veterans and unaccompanied youth in order find trends and seek possible solutions. As for limitations with the project, our data-sets only went back as far as 2007, meaning or studies on historical data could only go back 11 years. On top of this, given that we only had a certain amount of time to complete the study; we were unable to investigate more research questions on the homeless data. As the future approaches, homelessness in MA will continue to be prevalent and as such we should all work to help fight it with the support of many local organizations such as the Pine Street Inn, Boston Health Care for the homeless, Massachusetts Housing & Shelter Alliance, and many more.

References

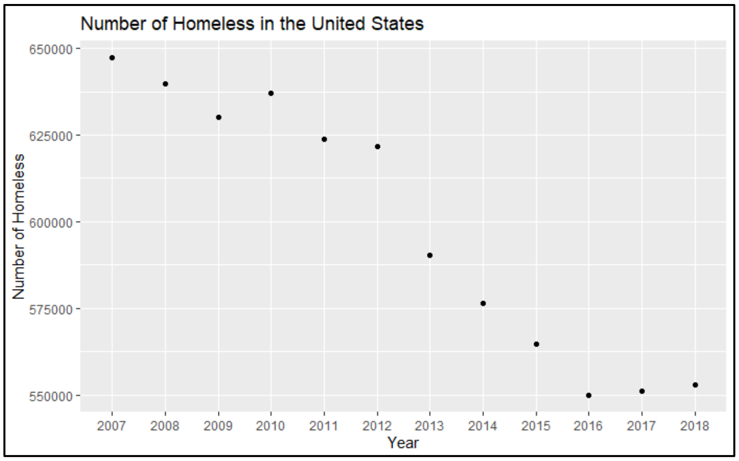


Figure 1. Graph of overall homeless in the US from 2007 to 2018

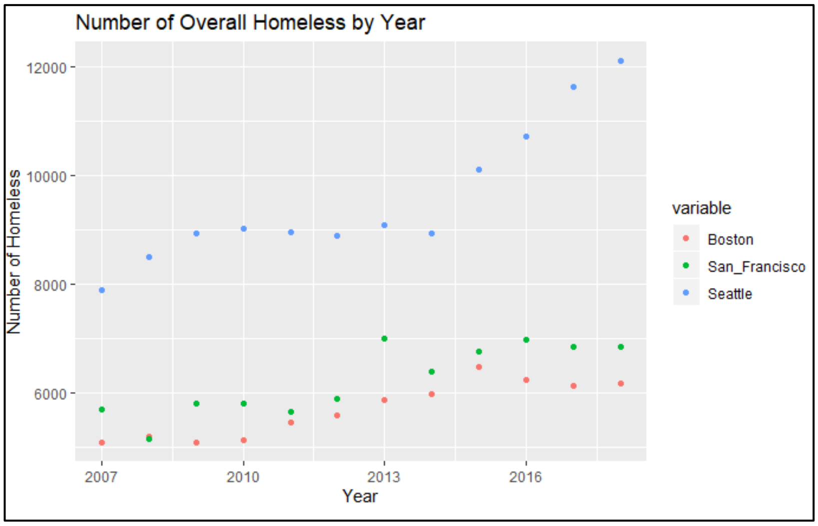


Figure 2. Graph of the overall homeless in three major US cities from 2007 to 2018

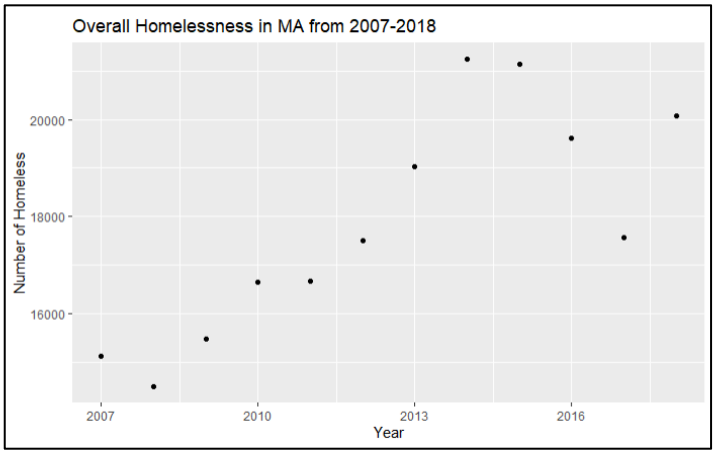


Figure 3. Graph of the overall homeless in Massachusetts in from 2007 to 2018

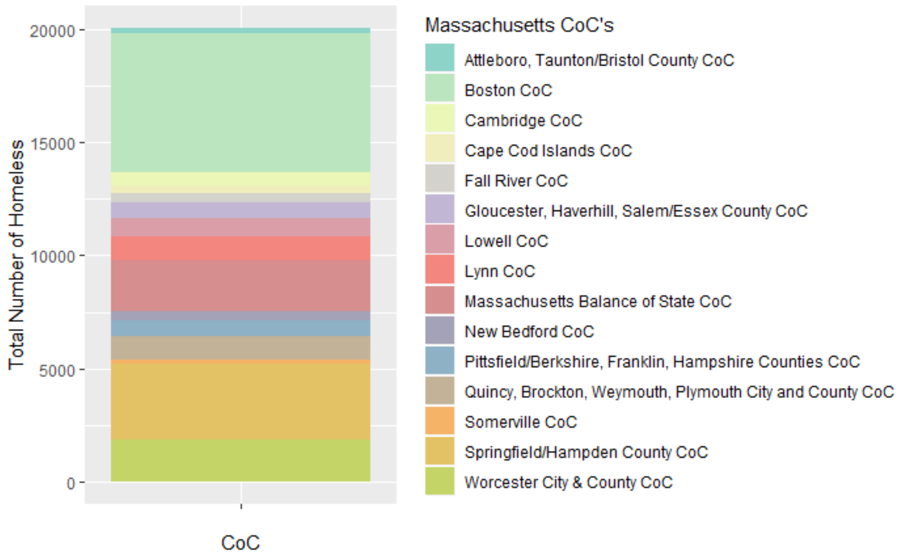


Figure 4. Distribution of Massachusetts Homeless Population in 2018 by Continuum of Care

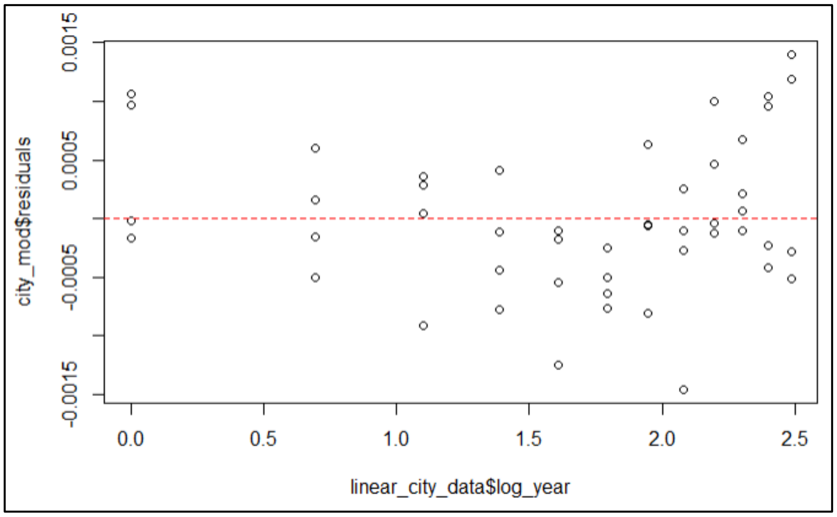


Figure 5a. Plot of residuals for year for multiple linear regression models

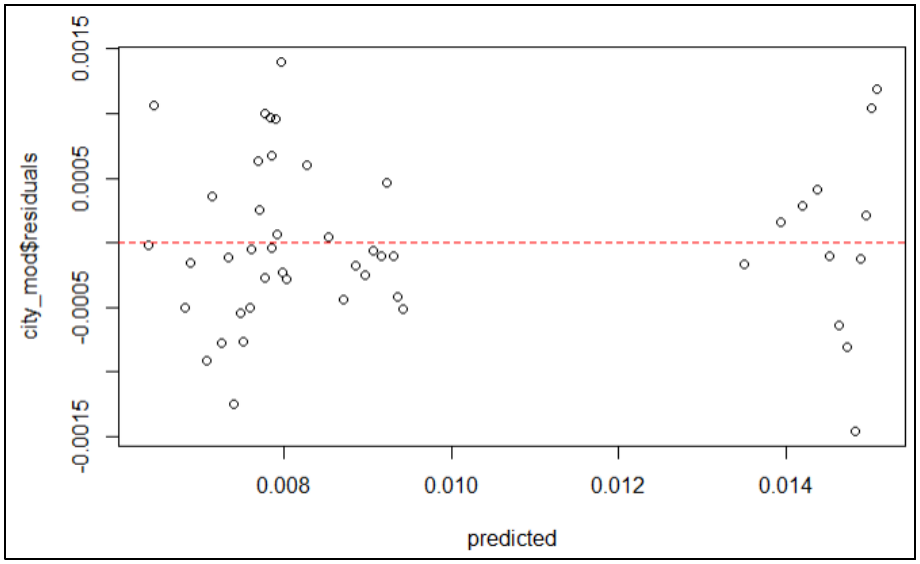


Figure 5b. Plot of residuals vs predicted values for multiple linear regression models

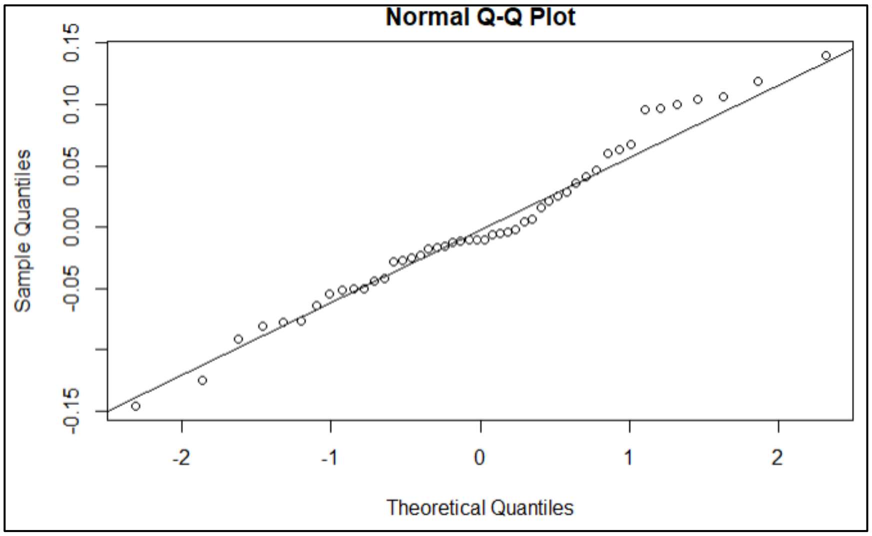
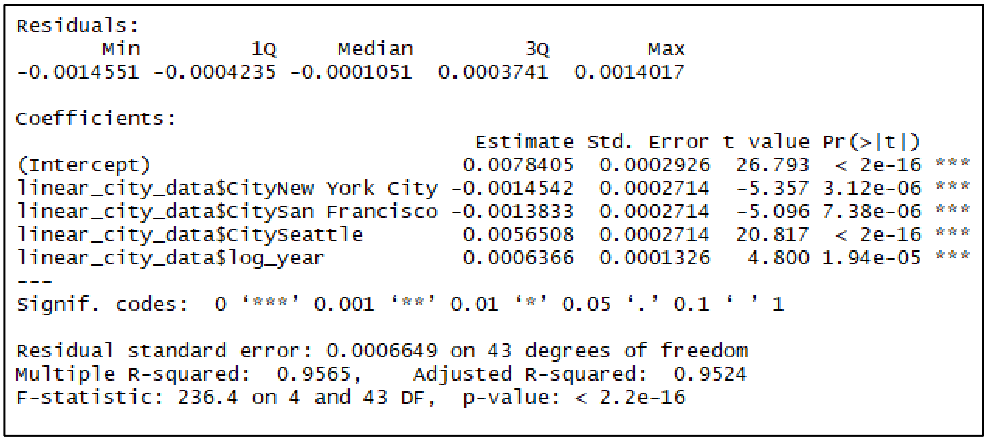


Figure 5c. Normal probability plot for multiple linear regression model



```
Residuals:
       Min         1Q     Median         3Q        Max 
-0.0014551 -0.0004235 -0.0001051  0.0003741  0.0014017 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         0.0078405  0.0002926  26.793  < 2e-16 ***
linear_city_data$CityNew York City -0.0014542  0.0002714  -5.357 3.12e-06 ***
linear_city_data$CitySan Francisco -0.0013833  0.0002714  -5.096 7.38e-06 ***
linear_city_data$CitySeattle        0.0056508  0.0002714  20.817  < 2e-16 ***
linear_city_data$log_year           0.0006366  0.0001326   4.800 1.94e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0006649 on 43 degrees of freedom
Multiple R-squared:  0.9565,	Adjusted R-squared:  0.9524 
F-statistic: 236.4 on 4 and 43 DF,  p-value: < 2.2e-16
```

Figure 5d. R output of multiple linear regression model fitting

Documents/Articles:

The U.S. Department of Housing and Urban Development. “The 2018 Annual Homeless Assessment Report (AHAR) to Congress.” 2018, files.hudexchange.info/resources/documents/2018-AHAR-Part-1.pdf.

Swasey, Benjamin. “The State's Homeless Population Jumped 14 Percent Over The Year.” *WBUR News*, WBUR, 18 Dec. 2018, https://www.wbur.org/news/2018/12/18/homelessness-massachusetts-hud-report.