Spring 2020

Analysis of the Homeless Population of Three Major Cities

Katherine Ferrara

Analysis of the Homeless Population of Three Major Cities

Katherine Ferrara
Dr. Kokkotos
16 December 2019

Executive Summary:

Homelessness within the United States of America has become a national epidemic. The Department of Housing and Urban Development (HUD) estimates that in 2018 there were 552,830 people experiencing homelessness on a given night within the United States. Therefore, 17 in every 10,000 people in the United States were homeless. In 2018, 35% of those who were homeless were unsheltered.

The 2019 American Statistical Association (ASA) annual data challenge called for high school and undergraduate students to analyze the HUD's 2018 Point-in-Time Estimate of Homelessness in the U.S. data set. The challenge wanted students to use statistics and data visualization to find the most effective ways to reduce homelessness in one of the three cities: Los Angeles, New York City, or Seattle. This project is an extension of the original ASA data challenge.

This project aims to analyze the homeless population for New York City, NY, Los Angeles, CA, and Seattle, WA in 2018. This project also examines how each of these cities is attempting to manage its homeless population. Finally, this project will investigate if there are demographic differences between these three cities that exacerbate the level of homelessness.

After analyzing data provided by the HUD and other external data sources such as the US Bureau of Labor Statistics, several conclusions were reached. Of the three cities studied, in 2018, New York City had the largest homeless population, followed by Los Angeles, then Seattle. New York City also had the highest proportion of sheltered homeless individuals compared to both Los Angeles and Seattle. New York City's shelters primarily was in the form of temporary housing, specifically emergency shelter. It was also found that Los Angeles and Seattle had the highest percentages of unsheltered chronically homeless and veterans.

In 2018 New York City's average rent and vacancy rate were well above the national average. This indicated that even though New York City had available apartments, the rent was too high for individuals to afford to live there. Another contributing factor to New York City's large homeless population was the high number of drug overdoses.

Los Angeles was also below the national average for the percentage of high school and college graduates for 2018, which might account for the high unemployment rate of Los Angeles. Los Angeles also had the highest percentage of rent as a fraction of income compared to New York City, Seattle, and the national average.

Seattle had the lowest homeless population out of the three cities studied. Seattle also had the highest number of high school and college graduates, as well as an unemployment rate below the national average. Education is a contributing factor that allowed Seattle's homeless population to remain lower than in other cities in 2018.

Data:

Each year the HUD conducts a study to analyze the homelessness population within the United States. The HUD provides data on 398 Continuums of Care (CoC), which divide the

entire United States into geographic regions. The CoCs are categorized as major cities, other largely urban, largely suburban, and largely rural based on the geographic data published by the Department of Education's National Center for Educational Statistics.

The HUD determines the total homeless population in each CoC by counting those who are homeless for 10 consecutive days during the month of January. This allows the HUD to create a Point-in-Time (PIT) estimate for the total number of homeless individuals in each CoC. The HUD provides counts for the homeless population on a variety of categories such as sheltered homeless, unsheltered homeless, homeless veterans, and homeless youth. The HUD also counts the total number of beds that are available for homeless individuals within each CoC, such as emergency shelters, transitional housing programs, safe havens, rapid rehousing, and permanent housing programs.

The terms below are defined by the HUD to better understand the homeless population:

- Individual refers to a person who is not part of a family with children
- People in Families with Children is defined as a collection of people with at least one adult and one child under 18 years old
- Chronically homeless individuals refers to an individual with a disability who has been homeless for over one year.
- Chronically Homeless people in families refers to a family where the head of household has a disability and has been homeless for at least one year.
- Veterans refers to people who have served on active duty in the armed forces
- Unsheltered homeless refers to people who spend nights in public or private locations not intended for sleeping (such as street, cars, or parks…)
- Sheltered homelessness refers to people who are staying in emergency shelters, transitional housing or safe havens.
- Emergency shelter (ES) refers to a facility that provides temporary shelter for a homeless person.
- Permanent Supportive Housing (PSH) provides housing assistance and other supportive services on a long-term basis to formerly homeless disabled people.
- Rapid rehousing (RRH) provides temporary housing to homeless individuals and quickly move them into permanent housing
- Safe havens (SH) are temporary shelters for hard to serve individuals
- Transitional housing (TH) provides temporary housing for homeless individuals for up to 24 months
- Temporary housing refers to a short-term housing option such as emergency shelter, transitional housing, and rapid rehousing

Approach:

The original data set included 122 variables each related to the homeless population for each of the 398 CoCs. In order to control each of the predictor variables, the data set was

normalized. Each numeric column was standardized so each variable had a mean of zero and a standard deviation of one.

Since the data set had so many variables, it was important to check for multicollinearity, which occurs when three or more variables are correlated. This was done by using eigen system analysis. Eigen system compares all variables against each other which makes it more effective than a correlation matrix which only shows pairwise relationships. An eigenvalue was computed for each variable within the data set. If all eigenvalues are similar in magnitude, there is not a significant amount of multicollinearity. If the eigen values vary greatly in magnitude, then multicollinearity is present (Mack, 2016). In the case of the HUD data set, the eigen values ranged from $-4.37 * 10^{-15}$ to 4.38. This indicated a significant amount of multicollinearity within the data set.

To address the multicollinearity, the variance inflation factor (VIF) was computed. The VIF is calculated based on a linear regression model of the data set. The VIF is computed for each variable. The VIF for each variable is the ratio of the variance of the regression coefficient when the model is fit with all variables divided by the variance of the regression coefficient when the model is fit with only one variable. If the VIF is between 1 and 5, there is little to no collinearity. A VIF over 5 indicates possible multicollinearity, while a VIF over 10 demonstrates significant multicollinearity that needs to be addressed (James, Witten, Hastie, & Tibshirani, 2015, p.101).

The VIF for each variable in the HUD data set was calculated and all variables with a VIF of over 5,000 were removed from the data set. After the number of variables was reduced, a new linear regression model was created with the new reduced data set. Next, the VIF for each variable was calculated again and variables with a high VIF were removed again. This process was repeated until the VIF for all variables was under 7. This successfully removed the multicollinearity from the data set and left 19 of the original 122 variables in the data set that were uncorrelated.

Next, a principal component analysis (PCA) was run. This process allows a large number of variables to be summarized into a new variable that compiles multiple variables into one. PCA computes each principal component as a linear combination of all variables. For example, the first principal component can be defined as

$$Z = \Phi_{11} X_1 + \Phi_{21} X_2 + ... + \Phi_{p1} X_p$$

where X_p is each variable, and Φ_{p1} is the loading or coefficient for each variable. These loadings define the first principal component.

The second principal component vectors can be written as

$$Z_{i2} = \Phi_{12} X_{i1} + \Phi_{22} X_{i2} + ... + \Phi_{p2} X_{ip}$$

where X_{ip} is each variable, and Φ_{p2} is the loading for each variable. The linear combination of the loadings for each variable in the second principal component has maximal variance compared to the loadings of the first principal component vector. This allows the first principal

component to be orthogonal to the second principal component (James, Witten, Hastie, & Tibshirani, 2015, p.375).

Since there are *p* total principal components, where *p* is the number of variables in the data set, it is important to determine how many of these principal components are needed to provide an accurate representation of the data. The proportion of variance explained can be calculated to show the variance explained by each principal component. The proportion of variance explained by all principal components always sum to one (James, Witten, Hastie, & Tibshirani, 2015, p.382). A scree plot, shown in Figure 1, can be used to determine the fewest number of principal components that are needed to account for the majority of the variance. Based on Figure 1, 47% of the variance is explained by the first principal component and 11% of the variance is explained by the second principal component. The third principal component only explains 7% of the total variance. Each of the remaining principal components accounts for less than 7% of the variance. This means that after the second principal component vector, very little information can be learned from each principal component. Therefore, only the first two principal components will be analyzed.

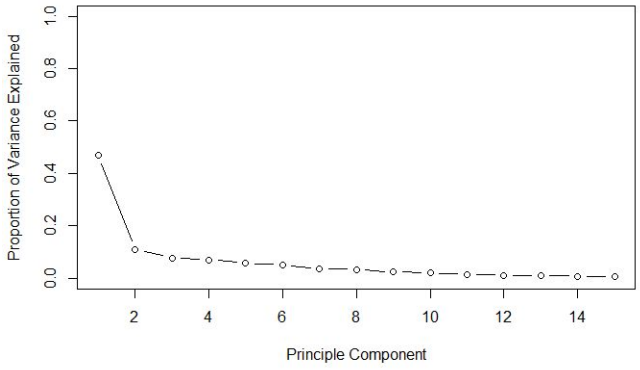

Figure 1

To understand the first two principal components in the context of the HUD data set, the first two principal components were graphed against each other, shown in Figure 2. The red arrows indicate the first two principal component vectors. For example, CoC category is located at (0.17, -0.13). This is because the loading on the first principal component for CoC category is 0.17 and the loading on the second principal component for CoC category is -0.13.

From this graphic, the first two principal components can be defined. The first principal component put the greatest weight on the variables related to temporary housing available to the homeless population. A few of these variables included total TH, ES, and RRH for veterans and youth. This also indicated that all the variables corresponding with temporary housing are related to each other.

The second principal component can be defined as the total unsheltered population. The second principal component put the majority of its weight on the variables: unsheltered homeless

and unsheltered homeless family households. This also indicated that the unsheltered homeless population was not correlated with the temporary housing available since the second principal component has the maximum variance to the first principal component.

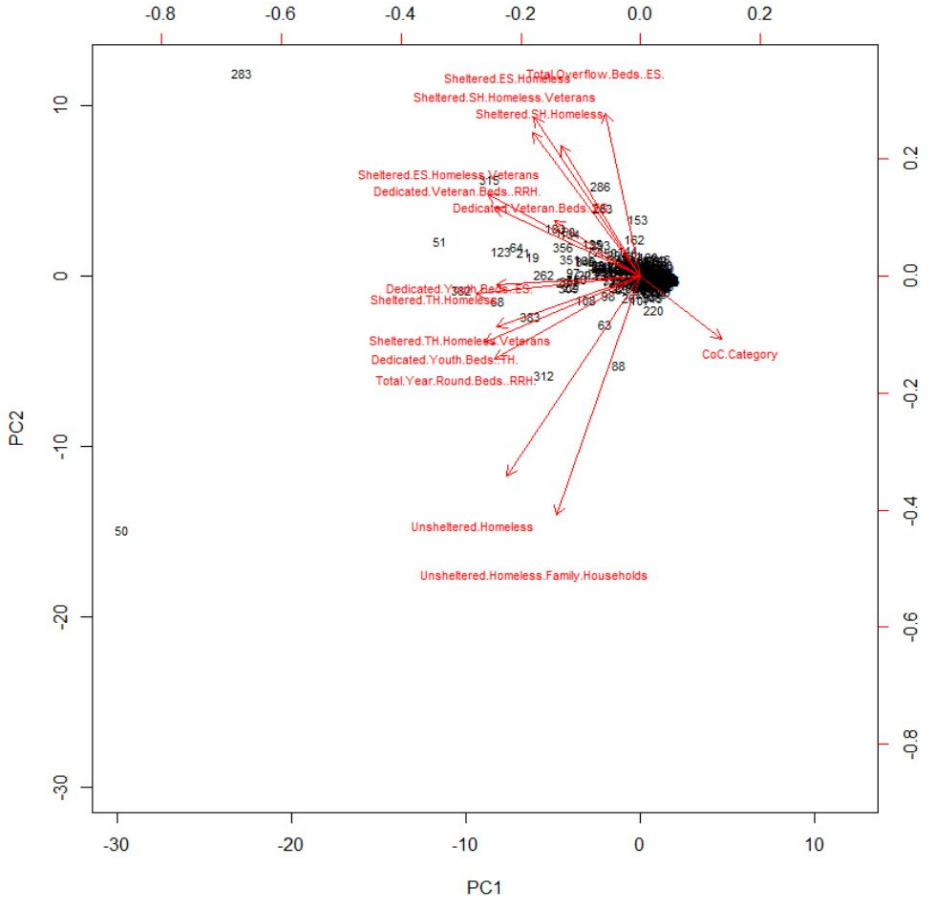

Figure 2

Figure 2 also shows the differences between each CoC according to the two principal components. This is represented by the black numbers which correspond to each one of the CoCs. There is a large cluster of numbers towards the center of the graph, which indicates that the majority of all CoCs behaved very similarly with respect to their homeless population.

It is also seen that CoC 283, which corresponds to New York City, and CoC 50, which corresponds to Los Angeles, are vastly different from the CoCs that are clustered at the center of figure 2. Figure 2 was then simplified by removing the vectors of each predictor variable and all CoCs other than New York City, Los Angeles, and Seattle, WA, which was located near the large cluster of CoCs in figure 2. This new graph is shown in Figure 3.

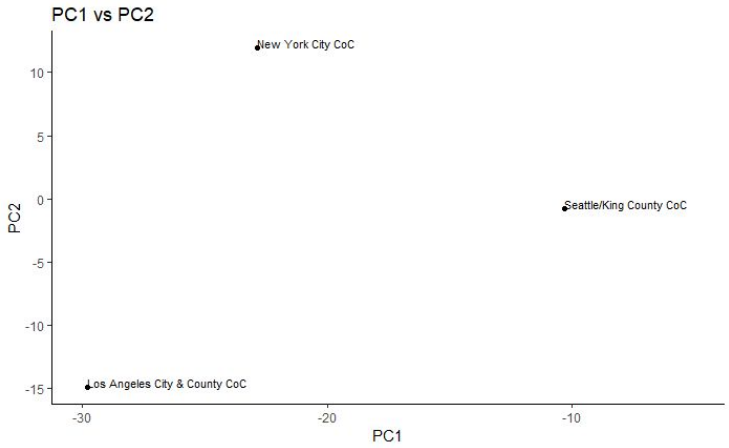

Figure 3

Figure 3 shows how different these three cities behave related to the first two principal components of temporary housing and unsheltered homeless. To identify what creates these differences three objectives were studied. First, the total homeless sheltered and unsheltered of each city was analyzed. Second, how each city attempts to manage its homeless population was investigated. Third, the demographic differences that may contribute to homelessness was determined. Each of these three issues are addressed in the detailed findings section below.

Detailed Findings:

Since the second principal component represents the unsheltered homeless population, it is important to determine the differences of the unsheltered homeless for each city. A proportion test that compared the total number of unsheltered individuals over the total homeless population for the three cities, had a p-value of $< 2.2 * 10^{-16}$. This small p-value shows that the three cities had extremely different levels of unsheltered homeless. Figure 4 shows the total homeless population for each city based on those who are sheltered and unsheltered. Figure 4 shows that even though New York City had the largest homeless population, it had the highest proportion of sheltered homeless. Los Angeles had a very small proportion of sheltered homeless compared to unsheltered homeless. This explains the differences in the second principal component score shown in Figure 3. Los Angeles had a score of -14.9, while New York City had a score of 11.9. Seattle was in the middle with a score of -0.77 and a better proportion of sheltered homeless compared to Los Angeles.

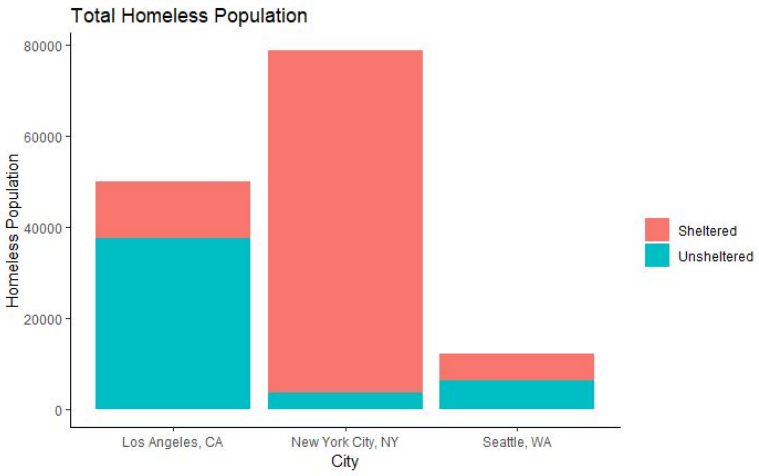

Figure 4

The data provided by the HUC categorizes the total homeless population into four categories: chronically homeless, homeless families, homeless individuals, and homeless veterans. Figure 5 below shows the percent of each category that is unsheltered for each city. The p-values for a proportion test for the number of individuals in each category listed above, related to the total homeless population for each city was $< 2.2 * 10^{-16}$. These small p-values show that the three cities did not share a similar distribution for each category.

As expected, New York City had the lowest percent of unsheltered homeless for all categories. Chronically homeless individuals had the highest percentages of being unsheltered, where 85% of chronically homeless individuals were unsheltered in Los Angeles. Unsheltered homeless veterans were a problem for both Los Angeles and Seattle, where 75% and 57% of all homeless veterans were unsheltered respectively. This is compared to New York City, where 1.3% of homeless veterans were unsheltered. Of the four homeless categories in all three cities, homeless families had the lowest percentage of being unsheltered, especially New York which had zero unsheltered homeless families.

Figure 5

Now that it is understood how many homeless individuals were unsheltered and which category they fall into, it is important to analyze the amount of temporary and permanent housing was provided by each city. A proportion test was run for the proportion of permanent housing related to the total number of homeless individuals. The p-value for this proportion test was $3.29 * 10^{-105}$. This indicated that the cities do not share similar proportions of permanent shelter based on their total homeless population. Temporary housing is defined as the total of ES, TH, and RRH. Figure 6 displays the total homeless population in red and the total temporary housing beds in blue. Even though New York City had the largest homeless population, it provided enough temporary housing to those who were homeless. Los Angeles and Seattle both did not provide enough temporary housing for individuals who were homeless, with Los Angeles providing the fewest number of shelters.

Figure 6

Figure 6 also shows the amount of permanent supportive housing available within each city. New York City, again, provided the most permanent supportive housing. Los Angeles provided an equal amount of permanent housing as temporary housing.

Next, the proportion of each type of temporary housing was determined. Figure 7 categorizes the total temporary housing provided by each city into emergency shelter, transitional housing, and rapid rehousing. The majority of New York City's temporary housing was emergency shelters. New York City also provided less rapid rehousing and less transitional housing than both Los Angeles and Seattle, yet New York City still had the fewest number of unsheltered homeless.

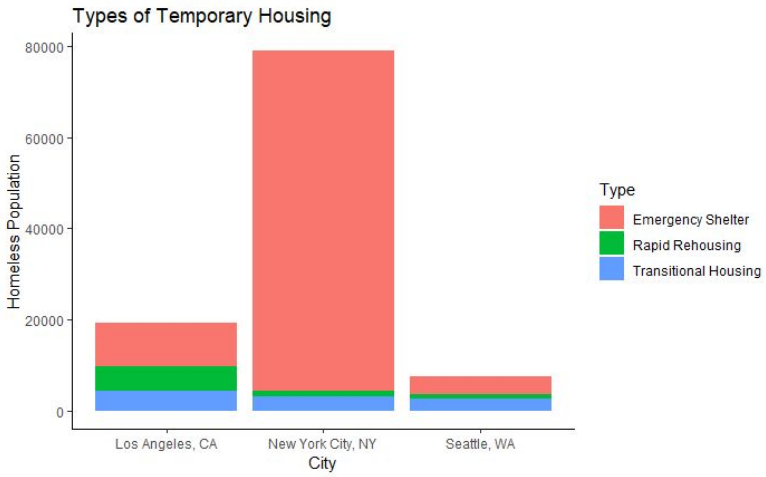

Figure 7

Last, the demographics of each city was analyzed to understand if the demographics of each city contributed to the differences in homelessness. Figure 8 shows the percentage of each race of the total population for each city. Three proportion test was done to compare the total number of Caucasian, African American, and Hispanic individuals to the total population of the three cities. These tests had a p-value of $2.71 * 10^{-6}$, 0.000420, and 1.23*10^{-8}5 respectively. This shows that the race of all three cities differed from each other. Seattle had the highest percentage of Caucasians at 59%. New York City had almost equal percentages of Caucasians, African-Americans, and Hispanics at 32%, 24%, and 29% respectively. The majority of Los Angeles's population was made up of Hispanics or Latinos at 48%.

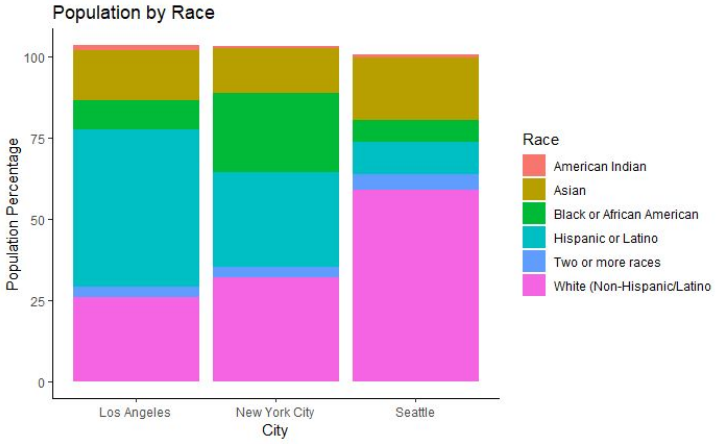

Figure 8

Next, CPI, minimum wage, unemployment, and drug overdoses were compared for each city, which is shown in Figure 9. The CPI or consumer price index indicates the average cost of a basket of consumer goods and services. All three cities have basically identical CPI averages to the US average. In 2018, the minimum wage in New York City was the highest at $13, while both Los Angeles and Seattle had a minimum wage of $11. All of these minimum wages were above the national average of $7.25.

A proportion test was run to compare the unemployment percentage of the three cities and the national average. This test had a p-value of 0.974, which indicated that the unemployment was similar for locations. That being found, Los Angeles did have the highest unemployment percentage at 4.7%, which is above the national average of 4%. Seattle's unemployment rate was below the national average at 3.4%. Also, New York City overwhelmingly had the greatest number of drug overdoses at 1,444 individuals compared to Los Angeles and Seattle which both were under 500.

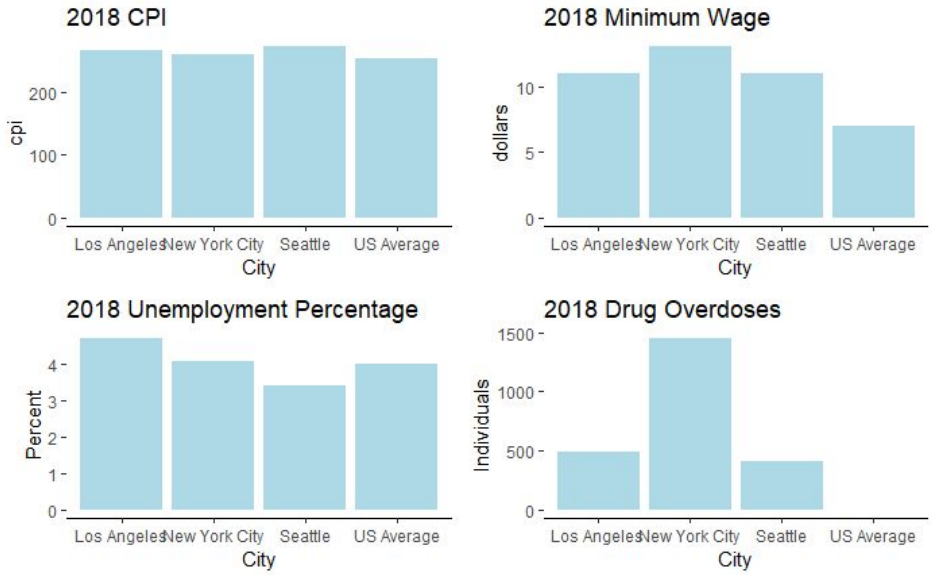

Figure 9

It is also important to analyze the cost of housing for each city. Figure 10 below shows the average rent, vacancy rate, rent as a fraction of income, and percentage of renters for the three cities and the US average. The average rent for the three cities was at least $500 higher than the national average of $1,012, with New York City having the highest average rent at $1,601. The vacancy rate of each of the three cities was also below the national average of 6.18%, with Los Angeles having the lowest vacancy rate at 3.3%. The average rent as a fraction of income for the US and Seattle was both at 20%, which is lower than the Los Angeles which was at 25.88%.

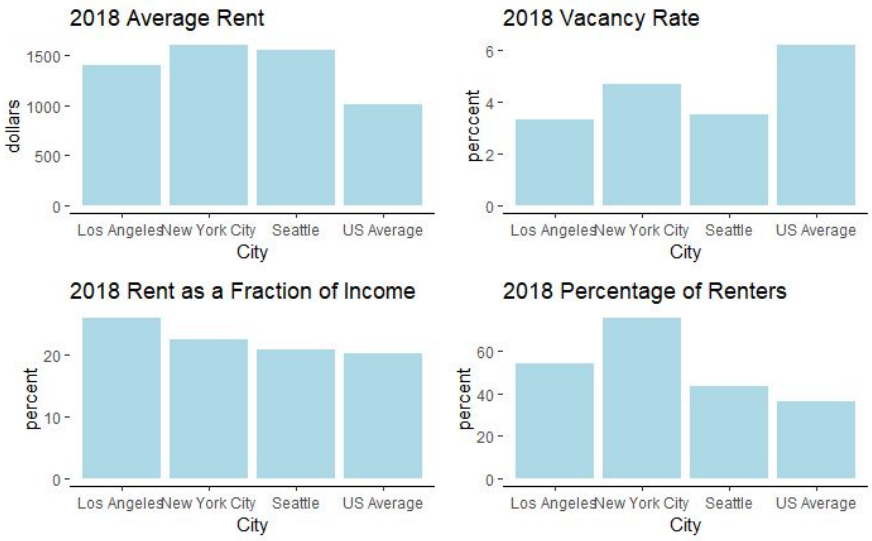

Figure 10

The last demographic analyzed was education, shown in Figure 11. A proportion test that compared the proportion of high school graduates to the entire population of each city was done. The p-value for this test was 0.0125, which indicated that the proportion of high school graduates differed for each city. A similar test was done to compare the proportion of college graduates with the total population. This test had a p-value of 0.0174, which also indicated that the proportion of college graduates differed for each city. Seattle had a higher percentage of high school graduates and bachelor's degrees at 92.7% and 50.3% respectively compared to the other two cities and US average which was 88% and 32% respectively. Los Angeles was the furthest below the US average at 78.2% for high school graduates and 31.2% for bachelor's degree or higher.

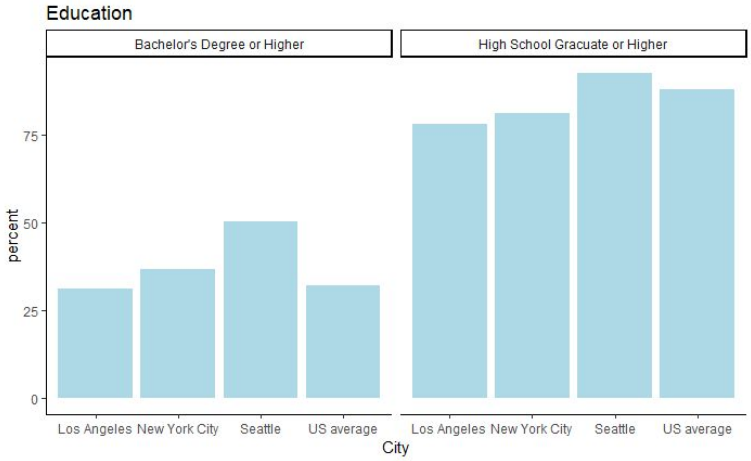

Figure 11

Conclusion:

The figures above provide a better understanding of the differences in homelessness for New York City, Los Angeles, and Seattle. New York City had the highest homeless population, but of the three cities, it was the only city that provided enough temporary housing for its homeless population. Although the majority of New York City's homeless population in 2018 was sheltered, it was mostly in the form of temporary housing, specifically emergency shelter. This provides relief to individuals and families for a few nights but is not a long-term solution. New York City did provide permanent shelter in 2018 but it was not nearly enough beds to sustain the entire homeless population. To fix this issue, New York City could work to provide more permanent supportive housing for homeless individuals, especially for those who are chronically homeless.

Between the three cities, Los Angeles had the highest proportion of unsheltered homeless. This may be due to the fact that the weather in Los Angeles is typically moderate year-round. New York City and Seattle might have felt a greater need to provide more temporary housing as the weather becomes colder. Since Los Angeles never reaches dangerously cold temperatures, they may not feel as great a need to provide temporary housing. Los Angeles should work to create more housing options for the homeless population even if it is just temporary, especially for those who are chronically homeless and veterans. Los Angeles could follow New York City's program and increase emergency shelters.

The rent for all three of these cities in 2018 was well above the national average, especially New York City. This was a contributing factor to New York City's large homeless population. New York City had the highest vacancy rates yet still had the highest homeless population. Even though New York City's minimum wage was higher than the national average by almost $5 it was not enough for individuals to afford the rent for the available apartments. If New York City wants to reduce the homeless population, the rent for the available apartment needs to be lowered to an affordable range or minimum wage needs to be raised so individuals can afford housing.

Another major contributing factor to New York City's large homeless population in 2018 was the large number of drug overdoses. Those who are addicted to drugs may have trouble holding down a consistent job, resulting in sporadic pay and a higher likelihood of homelessness. New York City should implement more drug intervention programs for citizens of all ages to prevent further drug addiction and overdoses.

Of the three cities, Seattle had the lowest homeless population, the most educated individuals, and the lowest unemployment rate. The more educated an individual is, the more likely they will be able to find a job that allows them to provide for themselves. New York City and Los Angeles might try offering educational classes to homeless individuals. This will help those who are homeless build their skill set and be more desirable to employers.

The findings above were similar to the results of the winners of the ASA fall data challenge. Three undergraduates from Willamette University who focused on the homeless of

Los Angeles had comparable findings. For example, they found that there was a much larger proportion of homeless individuals compared to unsheltered families. Just like the findings above, they determined that the homeless problem is due to the fact that many individuals are still not sheltered. They created Figure 12 below to show the relationship between those who are sheltered and how many beds are available. This graphic shows that in 2018 there were 1,689 beds that were unused. This group's greatest suggestion was to make sure all temporary housing available is being utilized by those who are homeless. This could be done by moving the homeless shelters closer to the where the unsheltered homeless are staying. Similarly they also recommended the creation of more programs geared at helping those who are homeless find jobs (El-Askari, Gomez & Gandy, 2019) .

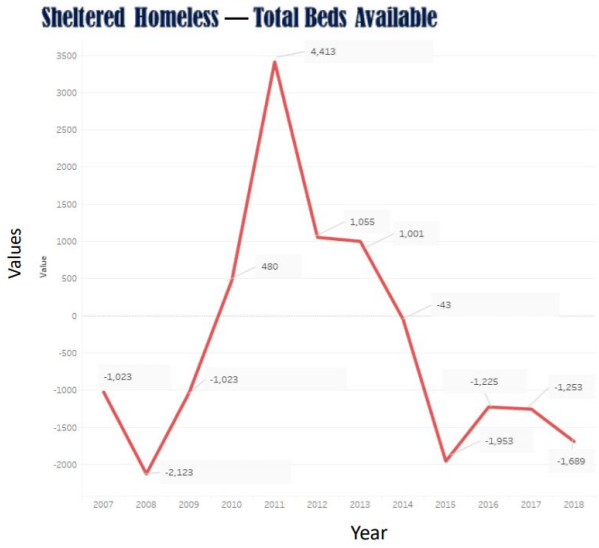

Figure 12

Three undergraduates from Loyola Marymount University studied the homelessness of Seattle. Consistent with the findings above, they found that Seattle had the lowest number of homeless individuals compared to the three cities and that the median rent was a driving factor of the increasing levels of homelessness. They created Figure 13 to show where the unsheltered homeless were staying from 2008-2016 in Seattle. They found that the majority of those who were homeless were living out of their cars/trucks. These students thought Seattle should address the issue of affordable housing and create safer overnight parking for those who are homeless (Martinez & Mauro, 2019).

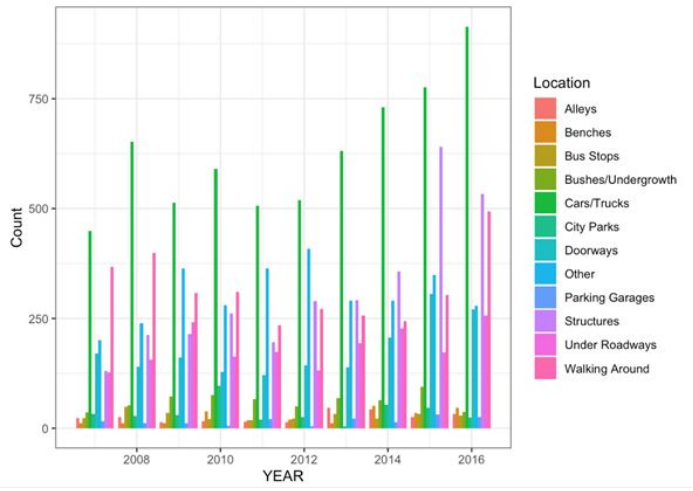

Figure 13

Overall, homelessness within the United States has become a major issue within the last few years. The suggestions above are only based on the data analyzed. To eradicate homelessness, it will require major social and political changes. No individual deserves to be homeless and governments should be working tirelessly in order to provide that to their citizens.

Limitations:

This project analyzed the homelessness of only three major cities during 2018. For a more substantial understanding of the homelessness problem, more cities could be analyzed over a greater time period. This would provide a better understanding of the trends of homelessness populations, demographics, and the economy over time.

The data provided by the HUD was also limited. All demographic and economic data included from this project were provided by outside sources. The racial demographics analyzed also provided very little insight into the homelessness population since it was for the entire population of each city, not only the homeless population. The HUD should work to include more demographic data into their study such as race, gender, age, and education of those who are homeless. This additional information would provide more insight as to which groups are in the most need of aid.

The HUD could also provide more information as to what the most common sources of temporary and permanent housing are such as schools, government organizations, and churches. The HUD could also provide information on how many organizations offer other forms of support such as soup kitchens and thrift shops. If this information was provided it may provide a better message as to what type of temporary and permanent housing are most beneficial and what other sources of aid are needed.

Sources:

Assefa, S. (2019). Population & Demographics. *Office of Planning & Community Development.* Retrieved from https://www.seattle.gov/opcd/population-and-demographics.

CBS New York. (2019) NYC Drug Overdose Stats Down, Fentanyl Still Top Killer. Retrieved from https://newyork.cbslocal.com/2019/08/26/nyc-drug-overdoses/.

CPI Inflation Calculator. (2018). 2018 CPI and Inflation Rate for the United States. Retrieved from https://cpiinflationcalculator.com/2018-cpi-and-inflation-rate-for-the-united-states/

El-Askari, L., Gomez, M., & Gandy, G. (2019) 2019 Fall Data Challenge Homelessness in Seattle. *Willamette University.*

Engebreth. B. (2018). Residential Rent and Rental Statistics. *Department of Numbers.* Retrieved from https://www.deptofnumbers.com/rent/washington/king-county/

Gerfinkel, M. (2018). Minimum Wage Ordinance. *Office of Labor Standards.* Retrieved from http://www.seattle.gov/laborstandards/ordinances/minimum-wage.

Henry, M., Mahathey, A., Morrill, T., Robinson, A., Shivji, A., & Watt, R. (2018). The 2018 Annual Homeless Assessment Report (AHAR) to Congress. T*he U.S. Department of Housing and Urban Development.* Retrieved from https://files.hudexchange.info/resources/documents/2018-AHAR-Part-1.pdf.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2015). *An Introduction to Statistical Learning.* New York: Springer.

Kunkler, A. (2019). Overdose Deaths Continue to rise locally and Nationally. *Seattle Weekly.* Retrieved from https://www.seattleweekly.com/news/overdose-deaths-continue-to-rise-locally-and-nationally-2/

Liu, C., Milton, J. & McIntosh, A. (2016). Comparing More Than 2 Proportions. *Boston University School of Public Health.* Retrieved from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R6_CategoricalDataAnalysis/R6_CategoricalDataAnalysis6.html.

Mack, C. [Chris Mack]. (29 October 2016). *Lecture52 (Data2Decision) Detecting Multicollinearity in R.* Retrieved from https://www.youtube.com/watch?v=QruEcbgfhzo.

Martinez, E., & Mauro, J. (2019). Combatting Homelessness in LA County. *Loyola Marymount University.* Retrieved from https://thisisstatistics.org/wp-content/uploads/2019/11/Undergrad_JackandElena_Presentation.pdf.

New York Department of Labor. (2019). Minimum Wage. Retrieved from https://www.labor.ny.gov/workerprotection/laborstandards/workprot/minwage.shtm

United States Census Bureau. (2018). *Quick-facts.* Retrieved from https://www.census.gov/quickfacts/fact/table/kingcountywashington,newyorkcitynewyork,losangelescountycalifornia#.

U.S. Department of Labor, Bureau of Labor Statistics. (2019). *Seattle-Tacoma-Bellevue consumer price index.* Retrieved from https://www.bls.gov/regions/west/data/consumerpriceindex_seattle_table.pdf.