A look into Homelessness Data in America (Final Project Step3)

Janine Par

2022-06-04

Introduction

Homelessness is one of the social problems our nation is facing now. The pandemic, inflation, and other economic events have been contributing factors to the number of homeless Americans, which continues to increase significantly in an alarming rate. McKinney–Vento Homeless Assistance Act of 1987 defines homelessness as people lacking a fixed, regular, and adequate nighttime residence. Adults, children, families of different age and race struggle with homelessness. According to the U.S. Department of Housing and Urban Development (AHAR) about 18 of every 10,000 people in United State experience homelessness in United State. California, New York, and Florida are the top states with homeless populations however the state with the lowest homeless population is seeing an increase as well. Even before the Pandemic, Chicago sees a two percent increase over 2019. State and Federal government is allocating and spending billions of dollars to fund initiatives that aim to end and reduce homelessness.

I have started this project by identifying data requirements to analyze and understand homelessness. I have researched public information on the homeless population per state and found that counting the homeless population is a challenge because of the lack of a clear definition of homelessness and homeless people are often reluctant to be interviewed. Different approaches are identified and used to collect population data, but each method has a technical limitation. For this project, I have used the Point-in-Time (PIT) count, a count of sheltered and unsheltered people experiencing homelessness on a single night in January. It does not seem reflective of the accurate yearly scale of homelessness. Still, it's a snapshot of a number of people experiencing homelessness at a point in time, which may impact the quality of the result for this data analysis.

As I consider the homeless population as the outcome variable, I looked for factors which are commonly thought to predict the homelessness count. I have considered the following and run a correlation analysis in an attempt to measure the relationship with the homelessness population.

1. US Population by State
2. Unemployment by State
3. Poverty by State
4. Percent of Adult with Depression by State
5. Violence and Crime Rate
6. Available Public support like number of bed available to shelter homelessness is effective in decreasing homeless

7. Weather
8. Housing Rental Rate

Problem Statement

Measure the relationship of multiple variables that can impact the homeless population and identify if available public support (total number of available beds) can somehow show effectiveness when decreasing homelessness.

Approach to Address Problem Statement

1. Defining the topic of the research in which I have identified homelessness and defining variables that impacts homelessness in the United States.
2. Start data gathering by identifying the type of data to collect and sources of information.
3. Data Preparation and Cleansing. I have collected public information in CVS and excel format. I have analyzed and understand the raw data and implemented R functions to:

- Extract relevant information
- Reorganize the data in accordance with my intended final dataset structure format
- Merge multiple data frame by a common data elements: state and year
- Create new variables such as the Ratio of Homelessness to State population and the Mean value of housing rentals.

4. Run r code to plot the variable and see relationship with the variables in a graph.
5. Run r code and perform correlation analysis

```
##                     Total_PIT_Homeless TOTAL_YEAR_BED Population
## Total_PIT_Homeless          1.00000000     0.86319454  0.8288174
## TOTAL_YEAR_BED              0.86319454     1.00000000  0.6912695
## Population                  0.82881736     0.69126949  1.0000000
## homeless_pop_ratio          0.54894997     0.56297639  0.2113302
## Employment                  0.83719056     0.72309910  0.9969207
## Poverty                     0.79478819     0.66108419  0.9870546
## DepressPCT                 -0.29404455    -0.29741200 -0.3186320
## CrimeViolence               0.82056869     0.64050841  0.9758060
## ave_f                       0.13404419     0.04819492  0.3531001
## ave_rent                   -0.04076237    -0.04094572 -0.1333686
##                     homeless_pop_ratio Employment    Poverty DepressPCT
## Total_PIT_Homeless           0.5489500  0.8371906  0.7947882 -0.2940446
## TOTAL_YEAR_BED               0.5629764  0.7230991  0.6610842 -0.2974120
## Population                   0.2113302  0.9969207  0.9870546 -0.3186320
## homeless_pop_ratio           1.0000000  0.2300254  0.1742797 -0.3118304
## Employment                   0.2300254  1.0000000  0.9768423 -0.3308227
## Poverty                      0.1742797  0.9768423  1.0000000 -0.2814486
## DepressPCT                  -0.3118304 -0.3308227 -0.2814486  1.0000000
## CrimeViolence                0.2007415  0.9647042  0.9732991 -0.2723869
## ave_f                       -0.1169984  0.3168972  0.4245043 -0.1346470
## ave_rent                     0.2460467 -0.1240782 -0.1728088 -0.3469375
##                     CrimeViolence       ave_f    ave_rent
## Total_PIT_Homeless     0.8205687  0.13404419 -0.04076237
## TOTAL_YEAR_BED         0.6405084  0.04819492 -0.04094572
## Population             0.9758060  0.35310010 -0.13336865
```

```
## homeless_pop_ratio     0.2007415 -0.11699837  0.24604674
## Employment             0.9647042  0.31689717 -0.12407822
## Poverty                0.9732991  0.42450432 -0.17280876
## DepressPCT            -0.2723869 -0.13464696 -0.34693746
## CrimeViolence          1.0000000  0.38632189 -0.17530011
## ave_f                  0.3863219  1.00000000 -0.08025326
## ave_rent              -0.1753001 -0.08025326  1.00000000
```

```
##
##  Pearson's product-moment correlation
##
## data:  homeless_df_topstate$Total_PIT_Homeless and homeless_df_topstate$Population
## t = 20.275, df = 202, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7678540 0.8595366
## sample estimates:
##       cor
## 0.8188518
```

```
##
##  Pearson's product-moment correlation
##
## data:  homeless_df_topstate$Total_PIT_Homeless and homeless_df_topstate$TOTAL_YEAR_BED
## t = 23.212, df = 202, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8103965 0.8863577
## sample estimates:
##       cor
## 0.8528267
```

```
##
##  Pearson's product-moment correlation
##
## data:  homeless_df_topstate$Total_PIT_Homeless and homeless_df_topstate$Employment
## t = 18.435, df = 151, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7758521 0.8752109
## sample estimates:
##       cor
## 0.8320909
```

```
##
##  Pearson's product-moment correlation
##
## data:  homeless_df_topstate$Total_PIT_Homeless and homeless_df_topstate$Poverty
## t = 13.004, df = 100, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7072868 0.8553354
## sample estimates:
```

```
##      cor
## 0.792723
```

```
##
##  Pearson's product-moment correlation
##
## data:  homeless_df_topstate$Total_PIT_Homeless and homeless_df_topstate$DepressPCT
## t = -3.7992, df = 150, p-value = 0.0002104
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4349609 -0.1438610
## sample estimates:
##        cor
## -0.2962761
```

```
##
##  Pearson's product-moment correlation
##
## data:  homeless_df_topstate$Total_PIT_Homeless and homeless_df_topstate$CrimeViolence
## t = 19.811, df = 202, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7599928 0.8545247
## sample estimates:
##       cor
## 0.8125345
```

```
##
##  Pearson's product-moment correlation
##
## data:  homeless_df_topstate$Total_PIT_Homeless and homeless_df_topstate$ave_f
## t = 1.8891, df = 198, p-value = 0.06034
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.005787825  0.266874318
## sample estimates:
##     cor
## 0.13306
```

```
##
##  Pearson's product-moment correlation
##
## data:  homeless_df_topstate$Total_PIT_Homeless and homeless_df_topstate$ave_rent
## t = -0.31065, df = 202, p-value = 0.7564
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1587464  0.1158673
## sample estimates:
##        cor
## -0.0218517
```

6. Attempt to implement a simple linear regression for the total available beds

```
##
## Call:
## lm(formula = Total_PIT_Homeless ~ TOTAL_YEAR_BED, data = homeless_df_topstate)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37503  -1573   -595   -115  84611
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -207.6346   974.0477  -0.213    0.831
## TOTAL_YEAR_BED    1.4483     0.0624  23.212   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12110 on 202 degrees of freedom
## Multiple R-squared:  0.7273, Adjusted R-squared:  0.726
## F-statistic: 538.8 on 1 and 202 DF,  p-value: < 2.2e-16
```

Analysis.

The following are the pearson correlation coefficient result that shows positive association between these variables and the PIT Homeless population which means that increase on these variables have impact to increasing homeless population

1. Population with r= 0.8188518
2. Unemployment Rate= 0.8320909
3. Poverty = 0.792723
4. Crime and Violence = 0.8125345
5. Average weather temperature =0.13306

Interestingly, these variables resulted to pearson correlation coefficient that shows negative association between these variables and the PIT Homeless population which means that increase on these variables have impact to decreasing homeless population which I think result is unexpected as these are known factors that impacts homeless population in a state. I think it may be related to the quality and coverage of data collected for this project which I will address in the limitation section:

5. Average Rent = -0.0218517
6. Percent of Adult with Depression by State -0.2962761

Implications.

In this project, I have used public downloadable US Fact information and analyzed the correlation of these factors to homelessness though some results are not as expected, and maybe due to the quality of the data downloaded, this exercise shows how data science and technology can be used to guide the government in making informed decisions that can use to fight the homelessness crisis.

This study implicates that a solid process and methodology must be applied to collect, transform and study the data to understand what model best fits into building a good data-driven solutions for solving homelessness similar to these recommendations:

- Linear regression model to estimate the relationship of multiple predictors to the homeless population such as increasing city population, high rate of health and substance use problems, unemployment rate, housing cost, etc.
- Logistic regression for predicting what solution may work or not work for known cases of homelessness in the city.
- k-Nearest-Neighbors (k-NN) model for predicting whether a person is at risk of homelessness or becoming a chronic homeless. The prediction can be based on the person's demographic, health, and administrative information.

Limitations.

1. I'm limited to downloadable public information from the internet on homelessness. This analysis can be improved if more detailed raw information on homelessness is captured and available such as:

- Homeless demographic information
- HUD Administrative information on the sheltered homeless person
- Population on homelessness by state and year. The only available for me to use is a snapshot point in time.

2. US Fact information

- Scope and Data Granularity where some data are only available for a specific year or aggregated at the country level as supposed to state level.
- Inconsistent formatting will require more time to transform like state can either be name, code, or spelled differently.
- Variables that I have chosen for this project are continuous though they came with different scales, which could be a problem when comparing (i.e., Depression is PCT while others are numerical count or captured rating)
- Missing and Duplicate Data

Correlation analysis of the homeless population with Depression pct and Average rental rate has resulted in a negative correlation coefficient, which was unexpected for me as these two variables are known factors to the growth of the homelessness population. This shows that data accuracy is very important because when data is insufficient, it can lead to faulty predictions and bad outcomes. The closer the data to reality or normalcy, our analysis and prediction will result to higher accuracy.

Concluding Remarks

This project has allowed me to apply what I have learned from this course and have a hands-on experience using a real-life use case:

1. Apply R packages, libraries, and functions to extract and transform data from multiple sources to build datasets needed for the analysis
2. Practice step by step methodology from data gathering, transformation and statistical test and analysis, such as correlation analysis and simple linear regression
3. Present results using different techniques such as plotting diagrams and building R Markdown

References

2021 ahar: Part 1 - pit estimates of homelessness in the U.S. 2021 AHAR: Part 1 - PIT Estimates of Homelessness in the U.S. | HUD USER. (n.d.). Retrieved May 15, 2022, from https://www.huduser.gov/portal/datasets/ahar/2021-ahar-part-1-pit-estimates-of-homelessness-in-the-us.html

USAFacts. (n.d.). Homeless population. USAFacts. Retrieved May 15, 2022, from https://usafacts.org/data/topics/people-society/poverty/public-housing/homeless-population/?utm_source=bing&utm_medium=cpc&utm_campaign=ND-StatsData&msclkid=a229785546631d9d16451ad4d1a50df6

COC housing inventory count reports. HUD Exchange. (n.d.). Retrieved May 15, 2022, from https://www.hudexchange.info/programs/coc/coc-housing-inventory-count-reports/

Publisher data.cityofnewyork.us. (2022, May 12). DHS Daily Report. DHS Daily Report - CKAN. Retrieved May 15, 2022, from https://catalog.data.gov/dataset/dhs-daily-report

System performance measures. HUD Exchange. (n.d.). Retrieved May 15, 2022, from https://www.hudexchange.info/programs/coc/system-performance-measures/#data

https://www.ncbi.nlm.nih.gov/books/NBK218229/