Drug Overdose and Student Homelessness: How the Habits of Parents Affect the Lives of Students

Aiden Manuel
Department of Computer Science
University of New Brunswick

December 2023

Abstract

This paper serves as a basic data analysis of the correlation between accidental deaths due to drug use and the homelessness rates of students from grades PK-12. The data used is from the state of Connecticut, and statistical analysis has shown a fairly certain correlation that the deaths of ages <40 by region and the relative homelessness of students experienced in the corresponding regions. The aim of this research is to provide a concrete way to estimate how many students in a school district or city are experiencing homelessness, as this statistic tends to be under reported or hard to record.

Background

The idea behind the research is that homelessness in students is a problem which gets much more common with economic recession, as was widely seen in the United States back in 2009-10 [3]. Since the Covid-19 Pandemic put everything to a halt, and housing prices and global inflation have been escalating, some believe that I may soon enter into another global recession [1]. If what we've seen in the past is any indicator, I should expect that this will result in more students from grades PK-12 experiencing homelessness around the globe.

The aim of this paper is to test the correlation between a potential factor contributing heavily to student homelessness (overdose due to drug usage) and, if a correlation is found, construct a model which could hopefully predict the proportion of homeless students in a region using that factor as a predictor variable. The utility of such a predictive model would be to enable the focusing of state resources to the school districts with more students in need, without necessarily needing to rely on students providing accurate census data.

Hypothesis

The hypothesis going into this research is that deaths due to accidental drug use per capita will be positively correlated with the proportion of students in a region experiencing homelessness. It is no secret that drug usage is tied heavily to socio-economic factors [2], however the exact relationship I am proposing here is slightly different. I seek to show that by specifically looking for parents (people of ages 25-40) who have died due to drug overdose, I will be able to predict how many students are suffering from homelessness.

The reasoning behind this is that from the ages 4-18, students are very dependent on their parents, at least in the United States and Canada. So, I would expect that if the parents of those students were to pass away, then the students would be more likely to be left homeless.

Data

The data was sourced from publicly available data sets on the state of Connecticut. It was devoid of all personally identifiable information, serving as a completely anonymous amalgamation of census date based on the demographics of students by district and the basic info on the persons who overdosed. This study only goes into the age demographics and the regions of residence.

Any racial or gender based demographics were not considered for the analyses. The temporal focus was on the 2021-2022 school year.

Student Attendance Demographics

T data on student attendance was sourced from the United States government's open dataset website (found here). The data was collected by the local government of Connecticut on a state level, and organizes student attendance by the following groups: Students experiencing homelessness, Students with disabilities, Students who qualify for free/reduced lunch, English learners, All high needs students, Non-high needs students and Students by race/ethnicity (Hispanic/Latino of any race, Black or African American, White, All other races).

For the sake of this analysis, the relevant group was Students experiencing homelessness.

Accidental Drug Related Deaths

The data on accidental drug related deaths was also sourced from Data.gov (found here). This dataset was the result of an investigation by the Office of the Chief Medical Examiner into the cause and location of drug related deaths. It has data on the type of drug(s) at fault; the city of residence, injury and death; the age of the victim; the sex of the victim; time and date of death and more.

For the sake of this study, I are only considering the age of the victim (25 - 40), the time of their death (Sep. 1st 2021 - Jun. 30th 2022) and their city of residence. Other factors are not being taken into account.

Population of Connecticut

The data for city populations in Connecticut was found online by a company called Cubit (found here). They source their demographic data from the American Community Survey's 2022 5-year estimates, which are also used by the US Census Bureau. Their website offers a plethora of data on the city-by-city demographics in Connecticut, although much of this data is locked behind a paywall. Thankfully, for the purposes of this study only population data was needed, which was complementary on behalf of Cubit.

Data Analysis

The process of data analysis for this study was as follows. (1) The datasets had to be filtered to find the relevant data, as was described above. (2) Then each dataset was plotted individually, so that it would become clear if there was any visible correlation. Since the goal of this analysis was to see if the number of deaths due to accidental drug use per capita was a good predictor of the proportion of homeless students in a school district, (3) the now filtered datasets were plotted on a scatterplot, with drug deaths as the independent variable. (4) Finally a linear regression analysis was performed to generate a p-value, and inform whether or not to reject the null hypothesis. All data analysis was performed in Jupyter Notebook using Python 3.11.7.

1. Filtering the Data

Attendance Data

The first dataset to filter was the attendance data. First, the dataset was imported into python using a pandas dataframe. The relevant columns to extract were "District Name", "Student Group" and "2021-2022 student count - year to date". The reason for selecting student count instead of attendance rate is because the study is focused on how many students are experiencing homelessness, not necessarily how much they will be attending school. From there, the relevant rows were selected from the Student Group column: "All Students" and "Students experiencing homelessness". This data was then sorted alphabetically based on the district name, and merged into one dataframe with columns "District Name", "ALL STUDENTS" and "HOMELESS STUDENTS". The final dataframe can be seen in Table 1.

District Name	Total	Homeless
Bridgeport	18482	123
Bristol	7439	65
Danbury	11783	25
Hartford	16371	107
Middletown	4346	35
New Britain	9299	81
New Haven	18028	234
New London	3055	184
Norwalk	12356	33
Stamford	15986	57
Stratford	6724	24
Waterbury	17806	245

Table 1

Worth noting is that only 12 school districts in Connecticut actually recorded data on homeless students. This only accounts for 6% of districts in the state. Because of this, our functional dataset is very small, and this will be discussed later on.

Drug Death Data

The second dataset needed to be filtered as well. The dataset was imported into a pandas dataframe, and the Date column was converted to a more pandas friendly datetime format. Two masks were created: One to filter for the dates during the 2021-2022 school year, and one to filter for the ages between 25 and 40. After applying these masks, the desired column was selected: "Residence City".

Based on this column, I filtered for the 12 cities which show up in Table 1, and then created a new dataframe. This dataframe has 2 columns: "City Names" and "Deaths per Capita". To calculate the deaths per capita in each relevant city, I added up the number of occurrences of a given city, then divided that by the total population of that city (retrieved from the population database) and multiplied by 1000. The output of this calculation was then placed into the Deaths per Capita column under the row of its relevant city, and I did this for each of the 12 cities.

At the end of this, I had a dataframe which included each of the 12 cities from Table 1 with the corresponding number of deaths due to accidental drug use per 1000 persons in those cities. This is shown in Table 2.

City Names	Deaths per Capita
Bridgeport	0.094258
Bristol	0.164655
Danbury	0.092811
Hartford	0.139846
Middletown	0.106013
New Britain	0.135095
New Haven	0.186743
New London	0.398767
Norwalk	0.055053
Stamford	0.044504
Stratford	0.095493
Waterbury	0.193351

Table 2

2. Plotting the Data

To begin plotting the data, there was one last key step. In the Student Population Dataframe, I had to calculate the proportion of students in the school population who were experiencing Homelessness. The plain amount would vary drastically depending on the student population of a district, and so calculating the percentage of homelessness among students would effectively normalize the data.

To do this, I simply added a new column titled "Proportion Homeless", which was calculated by dividing each row in the Homeless column by its respective row in the Total column. To convert this to a percentage, it was multiplied by 100.

Once this was done. The plotting of the data could begin. Since there were only 12 cities in question, and all the data was discrete, a bar graph made the most sense for this initial phase. The results of plotting can be seen in Figure 1 and Figure 2.

Proportion of Homeless Students by District (2021-2022)

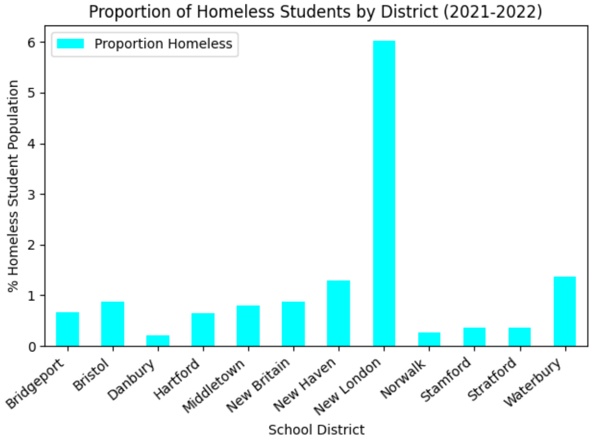

Figure 1

Number of Deaths due to Accidental Drug usage by City of Residence in the State of Connecticut (Sept 2021 - June 2022)

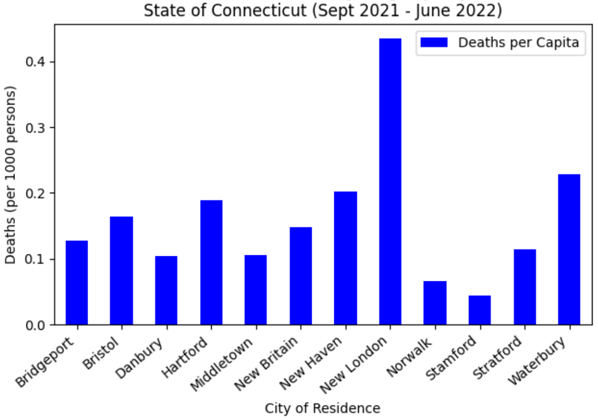

Figure 2

Initial analysis of this data seems to indicate that there is some merit to the hypothesis. It appears as though the cities which have high rates of accidental drug overdose (Figure 1) do in fact correspond to the cities which experience higher rates of student homelessness (Figure 2). They even share the same outlier, as New London is irregularly high in both cases.

There are some important things to note in the journey to get these results. Initially I wasn't filtering the drug related deaths by age, and this lead to a significantly less similar result (Shown in Figure 3). Additionally, the importance of using the data from the city of residence instead of the city of death is significant. When plotting the deaths by city of death, as seen in Figure 4, we see that it looks slightly less correlated.

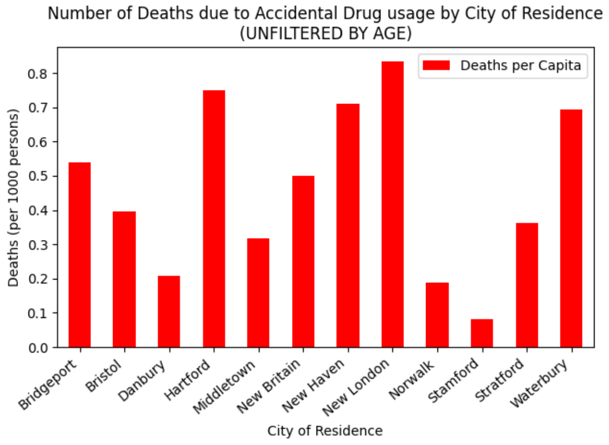

Figure 3

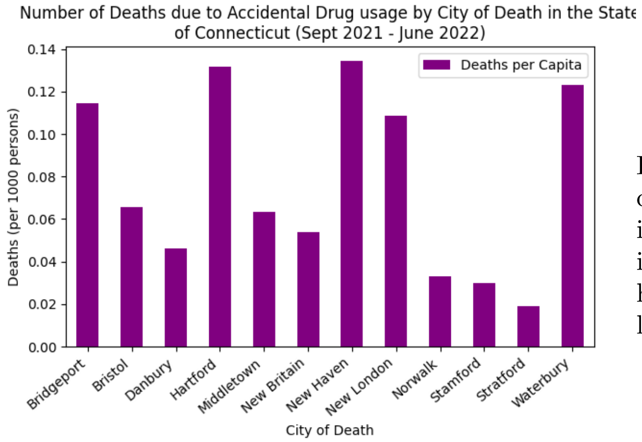

Figure 4

Both of these necessary filtering steps lead me to believe that it is in fact the parents who are dying due to drug overdose which results in the homelessness of students, something which makes intuitive sense. The age is of course important for this, but also the city of residence vs death gives a hint of it. If a parent goes to another city to do drugs, and dies in that city, their child(ren) will not be afflicted with homelessness in the city of death, but rather the city where they are living. This leads me to believe my intuitions were in fact true, and we will see more results reaffirming this in the coming models.

3. Scatterplot

To confirm the strength of the correlation between the two variables, it was necessary to plot them on a scatterplot, and perform some descriptive statistics. I chose deaths due to drug use per capita as my independent variable, and the percentage of homeless students as my dependent variable. Using this, the plot in Figure 5 was generated.

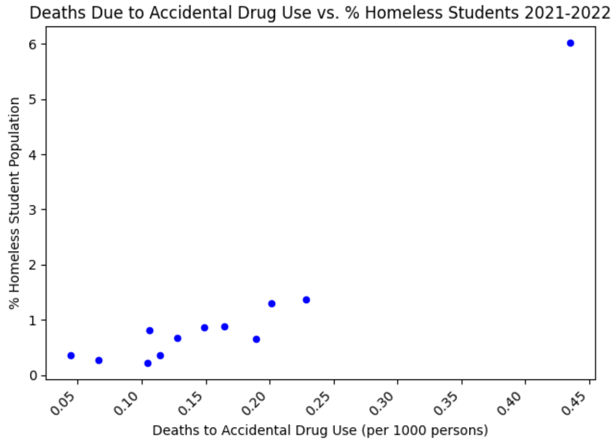

Figure 5

In this plot, there does indeed appear to be a positive correlation, with the high deaths per capita resulting in a higher percentage of homeless students. To verify the statistical significance of these results, I next had to choose a regression model, and use the seaborn library in Python to generate descriptive statistics.

Regression Analysis

Due to the large gap in data between all cities and New London, I was unsure of whether I should perform a linear regression or a quadratic regression. The trend looked somewhat parabolic, but the lack of data meant that it was impossible to say for sure. To keep things simple, I will only be discussing the linear regression here, although the quadratic regression did show promising results.

In Figure 7, we can see the results of the linear regression of the scatterplot (using the seaborn package in Python).

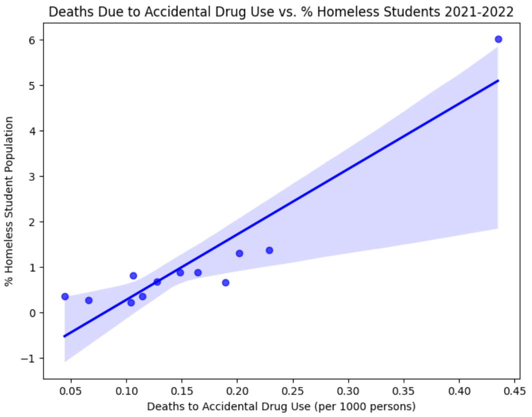

Figure 7

This regression line looks very good, as all the values are well within a reasonable deviation, and the slope is clearly positive. The P-value generated from the regression analysis was 0.00003, which is quite low. Additionally, the R-squared value was 0.89, indicating a very high degree of correlation. These results are enough to conclude that my results are statistically significant, and allow me to reject the null hypothesis.

Discussion and Conclusions

From the data I was able to analyze, I think that it is fair to say my results are statistically significant. That being said, I do not believe that there is enough data to conclusively say whether or not my results could be applied to a wider array of data. Because of the very small sample size (only 12 cities), it is impossible to say whether or not my results are a universal phenomenon or not. The extremely low P-value leads me to believe there may be some merit to my findings, but simply put, more data is required.

Going beyond what was discussed in this paper, I also tried a similar analysis on previous years to verify my findings. Although the population data for those years wasn't accurate (which affects my calculation of the deaths per capita), I found similar results, which further leads me to believe that there is indeed a strong correlation here.

All additional graphs that I generated for previous years, as well as my Python code and datasets, can be found in the github repository made for this project. In conclusion, I think the main problem with creating this model is the problem that it was aiming to solve: It is difficult to measure the homelessness rates in students. Since it relies heavily on data from surveys, and thus the honesty of students, and this makes it hard to ever confirm the data collected.

Ultimately, the analysis I performed here was extremely interesting, and certainly a good starting point. In the future, I would like to be able to access some more data, and perhaps even build a reliable predictive model. I think there are also a lot more avenues to explore, such as economic status, racial discrimination, or even sexual orientation. Drug use is one factor in a much larger picture, and certainly not the root cause of these issues.

References

[1] Justin Damien Guénette, M Ayhan Kose, and Naotaka Sugawara. “Is a Global Recession Imminent?” In: *Available at SSRN* (2022).

[2] Gene M Heyman, Nico McVicar, and Hiram Brownell. “Evidence that social-economic factors play an important role in drug overdose deaths”. In: *International Journal of Drug Policy* 74 (2019), pp. 274–284.

[3] Peter M Miller. “A critical analysis of the research on student homelessness”. In: *Review of educational Research* 81.3 (2011), pp. 308–337.