Coronavirus EDA Memo

Data Science I (STAT 301-1)

Derex Wangmang

Contents

Data Sources

I plan to use multiple datasets to examine the coronavirus spread and testing in Chicago neighborhoods.

Sources for my study may include:

- COVID-19 Cases, Tests, and Deaths by ZIP Code

The first dataset provides information on COVID-19 statistics separated by zip code (City of Chicago, 2020). These statistics include zip code, week number, weekly cases, weekly tests, percent tested positive weekly, and weekly deaths. Given that the data is tidy, I expect to combine the results of the dataset with other data, joining them by a unique id, the zip code. This becomes my frame of reference for any future data I use, so that all other datasets will include the zip code.

- COVID-19 Testing Sites

The second dataset provides locations of testing sites in Chicago (City of Chicago, 2020). It contains variables including facility, phone, address, website, and location. The most important one for me would be the address. I plan to parse through the address column, extracting the zip codes, and sum up the counts for each unique zip code, providing information about how many testing sites exist in Chicago per zip code.

- Public Health Statistics - Diabetes

The third dataset provides information about the number of hospital discharges for those with diabetes between 2000 and 2011 (Illinois Department of Public Health, 2012). The dataset contains variables include the zip code, the number of hospializations in each year, and the crude rate in each year. This could provide further information about the pre-existing conditions that exist among different populations.

- Public Health Statistics - Asthma

The fourth dataset provides information about the number of hospital discharges for those with asthma between 2000 and 2011 (Illinois Department of Public Health, 2012). The dataset contains variables include the zip code, the number of hospializations in each year, and the crude rate in each year. This could provide further information about the pre-existing conditions that exist among different populations.

- Obtaining Median Income Data for Zip Codes

The above source is an informative guide to accessing the median income data for different zip codes ("How to obtain," 2020). I have not explored the census data released by the US government, but I believe it should be tidy. That might explain the difference in hospital visits per zip code; the income may be correlated with the zip code for each company.

Why This Data

I chose this data because I read that minorities are disproportionately affected by the pandemic. Given Chicago's diverse populations in terms of socioeconomic status and ethnicity, I view Chicago as a microcosm of the United States. This data may reveal answers to some of my questions:

- How can we quantify the disparities of the pandemic on populations in Chicago?
- How do pre-existing conditions affect the rates of hospitalization?
- How does socioeconomic status affect the rate of hospitalization?

Potential Data Issues

For the most part, the data is tidy, separated by zip codes.

One potential issue is that the Public Health Statistics sometimes include aggregate of zip codes, rather than each zip code by itself. When interacting with that data, I may have to assume that the aggregate data is reflective of each individual zip code, which is a flawed assumption since the aggregate average may be affected by outliers.

Another potential issue may include outdated information. For example, the Public Health Statistics are from 2000 - 2011, which may not be accurate ~10 years later due to rising populations or migration.

Furthermore, the two pre-existing conditions I chose are asthma and diabetes. There are many more pre-existing conditions that are not accounted for. A person can have multiple pre-existing conditions as well. Given these external factors, the relationship between these two pre-existing conditions may be influenced by confounding variables.

Additionally, I would like to explore this dataset. However, given that the dataset is separated by community areas rather than zip codes, it may be more difficult to parse that data and assign it accurately. If I were to use that data, I would have to account for the proportion of the community within each zip code and calculate a weighted average for each statistic per zip code.

Citations

City of Chicago. (2020, October 15). *COVID-19 Cases, Tests, and Deaths by ZIP Code.* Chicago Data Portal. https://data.cityofchicago.org/Health-Human-Services/COVID-19-Cases-Tests-and-Deaths-by-ZIP-Code/yhhz-zm2v.

City of Chicago. (2020, September 10). *COVID-19 Testing Sites. Chicago Data Portal.* https://data.cityofchicago.org/Health-Human-Services/COVID-19-Testing-Sites/thdn-3grx.

How to obtain median income data for zip codes. Reddit. (2020). https://www.reddit.com/r/datasets/comments/hixfeo/how_to_obtain_median_income_data_for_zip_codes/.

Illinois Department of Public Health (IDPH). (2012, August 6). *Public Health Statistics - Diabetes hospitalizations in Chicago, 2000 - 2011.* Chicago Data Portal. https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Diabetes-hospitalizations/vekt-28b5.

Illinois Department of Public Health (IDPH). (2012, September 17). *Public Health Statistics - Asthma hospitalizations in Chicago, by year, 2000 - 2011.* Chicago Data Portal. https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Asthma-hospitalizations-i/vazh-t57q.