Elliot Druga, Tessa Saporito, and Janna Serrao

Professor Williamson

DS 2002

6 December 2024

Final Project Reflection

For our group's final project, we chose data on public health, specifically national death rates from two different causes: suicide and drug overdoses. We thought this would be fruitful to gather insights on two salient public health crises. Looking for patterns and trends among these datasets could prove useful for resource allocation and targeted preventative measures. Analyzing both datasets in tandem could reveal patterns or intersections among demographics. For example, if a single demographic experiences both causes of death at a higher rate, it is likely that there are underlying factors or shared vulnerabilities contributing to this heightened risk. Such insights can inform the implementation of comprehensive support systems, such as community mental health programs, substance use prevention initiatives, and socioeconomic policies that mitigate stressors.

Each dataset had similar features. For example, they were both disaggregated by sex, race, hispanic origin, and age. There were only two noticeable differences between the datasets. The first was that the drug overdose dataset further disaggregated the data by type of drug for each year, while the suicide dataset did not include method or means of suicide. Because our data analysis was conducted as comparisons between the two datasets on different factors, we chose to remove this variation in our data cleaning process by only keeping data with 'All drug overdose deaths.' Secondly, the suicide dataset stretched back to 1950, while the first available year in the drug overdose dataset was not until 1998. Because we conducted a time series

analysis, we removed the data in the suicide dataset that came from before 1998. We additionally removed the FLAG column in both datasets because it did not have any useful information for our purposes.

Building the rest of our ETL pipeline required first developing a clear plan for implementation. For this, our flowchart allowed us to clearly visualize the ETL steps, ensuring a structured approach to data extraction, transformation, and loading. After cleaning the data to address inconsistencies and redundancies, we established connections with SQL databases for efficient querying and with cloud storage solutions for secure and scalable data backup. This integration ensured that the pipeline could reliably process and store the data for analysis while maintaining accessibility and durability.

We conducted two analyses. The first was an analysis comparing drug overdoses and suicide rates by sex over time. While building the plot itself was relatively easy, we encountered an issue where the graph had two lines for each year and sex. We eventually noticed that this was because each dataset included both age-adjusted and crude death rates. To overcome this, we filtered the data to only include the age-adjusted death rate and created a new graph. We then conducted an analysis on differences in death rates across the lifespan by creating a histogram that displayed suicide and drug overdose death rates for each age group in 2017, the most recent year available in the dataset. This analysis required more trial and error compared to the first, as it involved merging and reconciling the datasets. Specifically, we concatenated the overdose and suicide data while adding a distinguishing column for cause of death, filtered out non-age-specific data, and ensured that both datasets included only age groups present in both sources. The code we developed for each of these analyses could be adjusted or turned into a function to give insights on other variables, such as race or hispanic origin.

Overall, our team worked well together. Each of us consistently contributed, but at times the expectations for each person were ambiguous or unclear. If we were to conduct a future project together, it would be useful to allocate tasks first such that each team member has clear responsibilities and expectations. Through the project, we learned how to work as a team as well as gained skills in automating the process of extracting data, improving its quality and usability, and storing it.