Katrina Garcia and Tamer El Aridi

**ETL Project Reflection**

Throughout the implementation of our ETL pipeline project, we encountered several challenges that tested our problem-solving skills and pushed us to adapt our approach. One of the most significant hurdles we faced was the initial attempt to work with the FBI crime API. The complexities of authentication, understanding the API structure, and figuring out the correct keys proved to be more challenging than we initially anticipated. After considerable effort and frustration, we made the decision to pivot to the CDC death rate API, which ultimately proved to be more accessible and easier to work with for our purposes.

Surprisingly, some aspects of the project were easier than we expected. Once we had settled on our data sources (the CDC API and the college dataset CSV), the process of cleaning and merging the data was relatively straightforward. Pandas proved to be an incredibly powerful and intuitive tool for data manipulation, making tasks like handling missing values and performing basic transformations quite manageable. On the other hand, some aspects proved more challenging than anticipated. Implementing user interactivity within the ETL pipeline was trickier than we initially thought. Balancing user input with robust error handling and ensuring the pipeline could handle various scenarios (like filtering for specific states or choosing output formats) required careful consideration and multiple iterations. Another unexpected challenge was dealing with the intricacies of data types and inconsistencies across our datasets. For instance, handling "PS" (Privacy Suppressed) and "NA" values in the college dataset required us to think carefully about how to treat these in our analysis without skewing results.

Despite these challenges, we believe that the utility we've created could be immensely useful for future data projects. The ability to quickly ingest, clean, and merge data from different sources (both local files and APIs) provides a solid foundation for various analyses. The flexibility to filter data and choose output formats makes it adaptable to different project requirements. Moreover, the experience of building this pipeline has given us valuable insights into the ETL process, API interactions, and data cleaning techniques. These skills are transferable to a wide range of data science projects and will undoubtedly prove useful in our future endeavors.

In conclusion, while we faced several hurdles during implementation, particularly with API selection and user interactivity, the project has been an invaluable learning experience. It has not only enhanced our technical skills but also improved our ability to adapt to challenges in data-driven projects. The resulting ETL utility, with its flexibility and robustness, will serve as a useful tool and a stepping stone for more complex data analysis projects in the future.