Nathan Tran

For this project, I decided to build an ETL pipeline to compare the relationship between violent crime rates and housing prices across all U.S. states. To do this, I combined a local CSV file from Redfin containing data of housing prices in every U.S. state with public API data from the FBI Crime Data API. Prior to beginning to develop the project, I wanted to compare the relationship between these two variables across all U.S. counties as I believe it would provide better insight into the relationship, as counties can vary a lot when it comes to both variables. For example, a county containing a large urban city would contrast a lot and provide a lot of insight when compared to a county that is more rural. However, as I did more research on the available free public APIs, most of them did not support data on a county by county basis, making it very difficult to find suitable data. I also considered comparing metropolitan areas as the FBI Crime Data API supported data from specific agencies. However, this meant that I had to manually map every metropolitan area to a series of police agencies to associate the two data sets, which is very difficult as metropolitan areas can be very extensive and consist of multiple police agencies, making the task very difficult and extensive. Therefore, I had to settle on comparing states as a whole instead as all of these APIs supported it, but this caused me to compromise more insightful data. When I did this, I ran into another issue where the Redfin CSV file associated data with the state name, while the API data from the FBI Crime Data API associated data with the state abbreviation. This meant that in the code I had to manually map every state to its abbreviation to link the two data sets, which was a tedious task. Another difficult thing I had to do was using the FBI Crime Data API within the code, as it was pretty confusing even given the documentation. The data type of both data sets were also different from each other, as the FBI Crime Data API returned data in json format, which was different from the csv file from Redfin. Once I implemented both, the rest of the project was pretty smooth sailing, especially with the

use of Pandas. Pandas made it very easy to group data by state and compute other statistics such as average median sale prices, state rankings by crime and housing prices, and correlation. I wanted to integrate a scatterplot to better visualize the data, and this was not as difficult as I thought it would be as the data was already formatted and laid out for me. For the future, this ETL pipeline could be adapted to include other public datasets for information such as unemployment, education, or environmental statistics. These variables can provide more context behind the data, allowing for a more comprehensive understanding of these societal metrics. The code could also be modified so it can be used for more specific regional selections such as counties and cities once the proper data is acquired, allowing for more detailed insight.