Taib Diallo, Alexis Hernandez, Madelyn Baker

GT Data Science and Analytics Bootcamp

June 11, 2022

ETL Technical Report

To begin, both of our datasets examined state data during the years between 1991 and 2018. One data set looked at the minimum wage and compared it to the federal minimum wage. While, the other dataset looked at hate crimes that have occured in each state and the number of victims in each of those states. The aim for this project was to determine if there was any correlation between the state's minimum wage and hate crimes that occured in that year and state.

Extraction

During the extraction process, we used the csv files from kaggle. We read both of those datasets into our jupyter notebook. Once we read the csv file, we imported it into a pandas DataFrame in order to prep the data to be transformed. For the minimum wage dataset, we used an encoding so that we could see the DataFrame in our jupyter notebook.

Transformation

After the extraction process was complete, we started to clean the data using pandas. For example, we renamed the columns in the hate crime dataset so that we could improve readability. The columns we changed were "STATE_NAME", "DATA_NAME", and "VICTIM_COUNT" to "State", "Year", and "Victim_Count". We also performed a groupby function on the hate crime dataset so that we could count the total number of victims in each state per year. Afterward, we performed a loc function on the minimum wage dataset to narrow down our

column only to years between 1991 and 2018. We used the loc function to match the years contained in the hate crime dataset. Next, for both of the datasets, we created new data frames with only select columns that we needed for our analysis. Those include the year, state, and victim count from the hate crime dataset. And the state, year, federal minimum wage, and state minimum from the minimum wage dataset. It is important to note that we included both the state and federal minimum wage because some states did not have a minimum wage so we used the federal wage in its place. Lastly, we reset the index in the minimum wage dataset and merged the two DataFrames on the state and year columns with an outer join. It is important to note that we used an outer join instead of an inner join, because some of the states did not have any hate crime victims in certain years. Therefore, if we used an inner join, those rows would not be present in our DataFrame.

Load

To load these DataFrames into postgres, we first created a connection to the database using a connection string. Then, we went over to pgAdmin where we created each of our tables in postgresQL. We verified our tables were created by using the function "engine.table_names( )" in our jupyter notebook. Next, we connected our postgres database table to the pandas DataFrames. Finally, in order to verify that the connection was successful, we queried each database table to make sure the output was similar to what we created above. In summary, we chose these datasets because we wanted to examine the effects of a state's minimum wage on the occurrence of hate crimes in the area. We hypothesize that the lower the state's minimum wage is, the higher the amount of victims there will be.