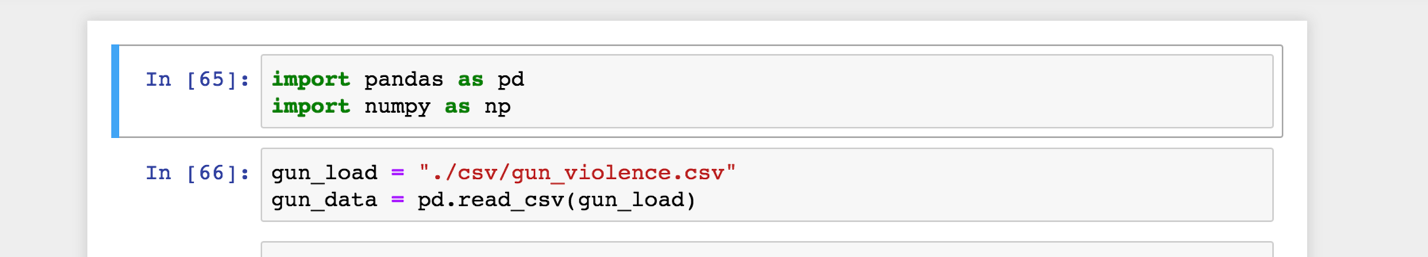
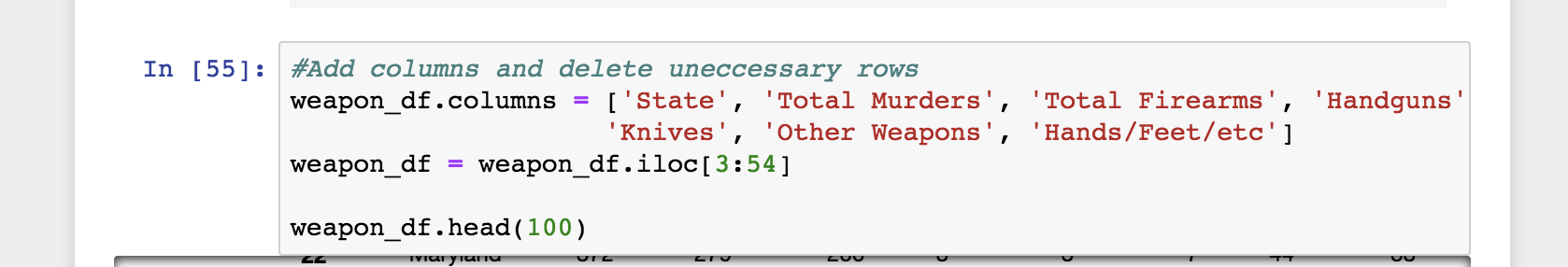
ETL Project

Kevin Menz & Conor Healy

For the ETL Project (Extract, transform and load) we decided to explore data based on gun statistics in the United States for the year 2015. Our data was focused on the murder aspect of gun statistics, and included data from several sources. In order to achieve our desired database, we accessed data from the FBI government database, web scraped from Wikipedia, and utilized a Kaggle database based on gun violence incidents. When accessing the FBI database, we uploaded a csv file from FBI.gov website. Our second form of data came from web scrapping from Wikipedia. After web scraping the Wikipedia page through the pandas web scraping function, we uploaded a table to our notebook. The final set of data was from Kaggle.com. This data included murder statistics from 2013-2018, and in order to access this data we uploaded a csv to pandas to manipulate into a data frame.

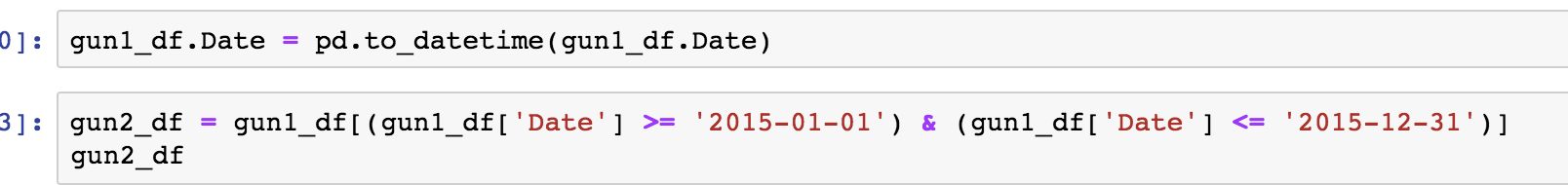




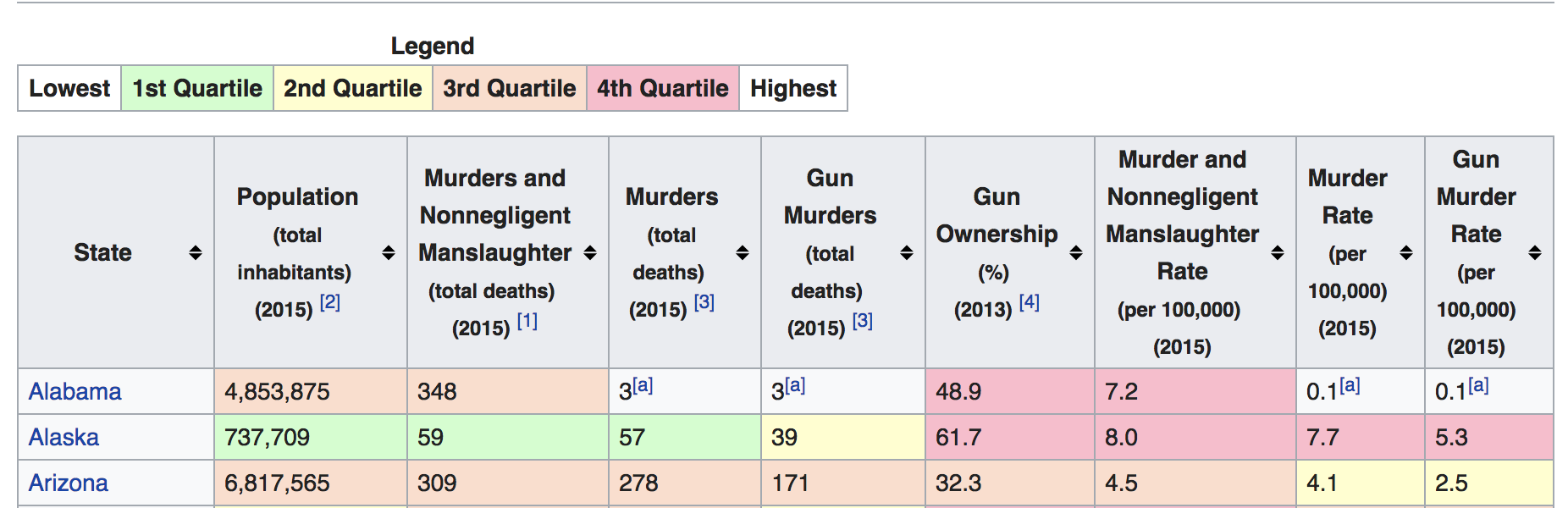
The second aspect of our ETL Project, transform, is where our data frames took shape. Our FBI csv was uploaded into a pandas data frame. To initially clean our data, we dropped extraneous rows and added in specific column headers by using the iloc and columns function. 

Our next step, we dealt with records or values that had additional characters attached. In order to achieve this, we used a lambda function and an rstrip function to strip away any unnecessary characters which could have caused problems with our uploading. We set the index to “State” to better visualize our table and uploaded it to a csv and that was ready to upload to SQL.

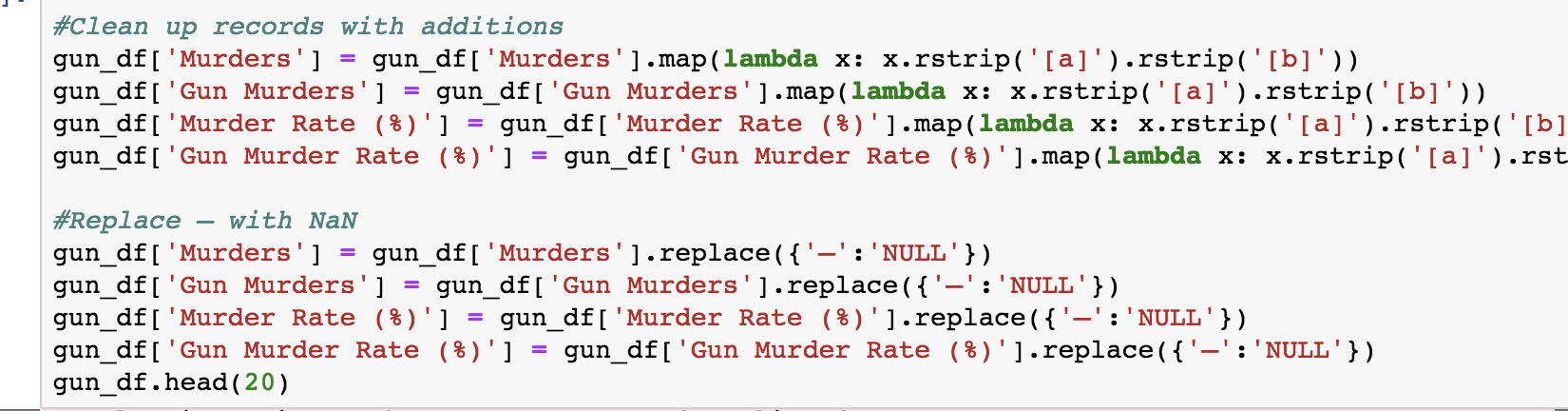
The second form of data we utilized was the Kaggle.com csv that included several forms of violent gun incidents in the United States. After uploading our csv file, we set headers and dropped any unnecessary rows by creating a new data frame. To better visualize our data and to match it to our other tables, we then set the index to “State”. In order to filter our table to include only incidents from the year 2015, we had to convert the “Date” values into date time, and then filter for only dates within the year 2015. To achieve this, we created a new data frame, which only included dates greater than or equal to ‘2015-01-01’ and dates less than ‘2015-12-31’. After this our table was saved to csv and ready to convert into SQL.



The last form of data, pandas web scrapping from Wikipedia, uploaded an initial table which outputted data as a list format, and our first step was to transform this into a table.



We continued to clean the data by naming certain columns and striping away any unnecessary rows, as well as setting our index to” State”. Next, we used the lambda function again to strip away any unnecessary characters in our data as well as using the rstrip function. Our last step was to utilize the .replace function to replace any symbols associated with empty values and replace it will ‘NULL’.



After loading all of our different data and producing three final data frames which we transformed into csv files, we utilized the import wizard on SQL, and converted our individual csv into three separate SQL tables. We had decided to use MySQL because our data was produced in tabular format, and MySQL is perfectly suited for our 3 tables as opposed to a NoSQL database, which is more suited to data in a JSON format.

With that, we had successfully taken data from multiple sources and manipulated them in a way that allowed us to upload them to a SQL database ready for querying.