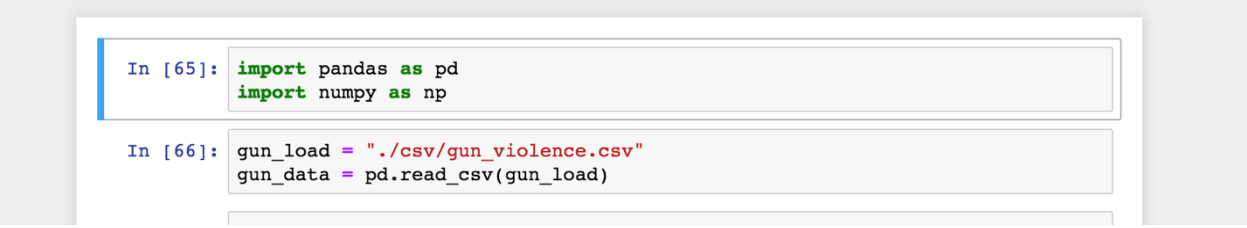
# ETL Project

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.

```
In [55]: #Add columns and delete uneccessary rows
         weapon_df.columns = ['State', 'Total Murders', 'Total Firearms', 'Handguns'
                              'Knives', 'Other Weapons', 'Hands/Feet/etc']
         weapon_df = weapon_df.iloc[3:54]

         weapon_df.head(100)
```

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.

```
0]: gun1_df.Date = pd.to_datetime(gun1_df.Date)

3]: gun2_df = gun1_df[(gun1_df['Date'] >= '2015-01-01') & (gun1_df['Date'] <= '2015-12-31')]
    gun2_df
```

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.

Legend

| Lowest | 1st Quartile | 2nd Quartile | 3rd Quartile | 4th Quartile | Highest |
|---|---|---|---|---|---|

| State | Population (total inhabitants) (2015) [2] | Murders and Nonnegligent Manslaughter (total deaths) (2015) [1] | Murders (total deaths) (2015) [3] | Gun Murders (total deaths) (2015) [3] | Gun Ownership (%) (2013) [4] | Murder and Nonnegligent Manslaughter Rate (per 100,000) (2015) | Murder Rate (per 100,000) (2015) | Gun Murder Rate (per 100,000) (2015) |
|---|---|---|---|---|---|---|---|---|
| Alabama | 4,853,875 | 348 | 3[a] | 3[a] | 48.9 | 7.2 | 0.1[a] | 0.1[a] |
| Alaska | 737,709 | 59 | 57 | 39 | 61.7 | 8.0 | 7.7 | 5.3 |
| Arizona | 6,817,565 | 309 | 278 | 171 | 32.3 | 4.5 | 4.1 | 2.5 |

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'.

```
#Clean up records with additions
gun_df['Murders'] = gun_df['Murders'].map(lambda x: x.rstrip('[a]').rstrip('[b]'))
gun_df['Gun Murders'] = gun_df['Gun Murders'].map(lambda x: x.rstrip('[a]').rstrip('[b]'))
gun_df['Murder Rate (%)'] = gun_df['Murder Rate (%)'].map(lambda x: x.rstrip('[a]').rstrip('[b]
gun_df['Gun Murder Rate (%)'] = gun_df['Gun Murder Rate (%)'].map(lambda x: x.rstrip('[a]').rst

#Replace – with NaN
gun_df['Murders'] = gun_df['Murders'].replace({'–':'NULL'})
gun_df['Gun Murders'] = gun_df['Gun Murders'].replace({'–':'NULL'})
gun_df['Murder Rate (%)'] = gun_df['Murder Rate (%)'].replace({'–':'NULL'})
gun_df['Gun Murder Rate (%)'] = gun_df['Gun Murder Rate (%)'].replace({'–':'NULL'})
gun_df.head(20)
```

After that, we merged the Wikipedia data frame and the FBI data frame together because they contained similar columns and were both on the state level.

After producing csv files from our data frames we decided to load these data frames into a SQLite database because the format of these files is a tabular format with structured columns and data that is more suited to this rather than a NoSQL database. We then used Flask to create an API that allows users to see and access the database.

```
+ @app.route("/api/v1.0/type_and_rates")
+ def type_and_rates():
+     """Return a list of types of weapons and murder rates per state"""
+     # Query all passengers
+     #results = session.query('type_and_rate').all()
+
+     return pd.DataFrame.to_json(type_and_rate, orient='records')
+
+ @app.route("/api/v1.0/gun_incidents")
+ def incidents():
+     """Return a list of types of weapons and murder rates per state"""
+     # Query all passengers
+     #results = session.query('type_and_rate').all()
+
+     return pd.DataFrame.to_json(gun_incidents, orient='records')
+
```

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 and easy viewing with the API.

Data Sources

- Gun Violence Data, James KO, Kaggle.com

  https://www.kaggle.com/jameslko/gun-violence-data

- Gun Violence in the United States by States, Wikipedia

  https://en.wikipedia.org/wiki/Gun_violence_in_the_United_States_by_state

- 2015 Crime in the United States, FBI

  https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/tables/table-20