**Extract – Transform – Load Project 2**

Group:

David Vanie

Burnetta Wood

Russell Collins

**DATASOURCE**

**Extraction**

After identifying out data sources:

1. https://www.kaggle.com/the-guardian/the-counted?select=2016.csv
2. https://www.kaggle.com/andrewmvd/police-deadly-force-usage-us?select=fatal-police-shootings-data.cs

We proceed to data evaluation by transforming our data in various ways as listed below.

**DATA TRANSFORMATION**

**Data Cleansing**

Our first data source had 16 columns and the second data source had 14 columns. We needed 3 extra columns from the second source.

1. Rename column names
   * After joining both tables we had duplicate columns. To clean these, we renamed the columns and removed the duplicate columns.
2. Drop duplicate data rows
   * We joined both data sources with an inner join which removed the excess rows and provided a subset of data.
3. Combine date columns for year, month and date.
   * Both data sources represented the date differently. To clean this up we combined the year, month, and day in the second data sourced to make one column. We removed all other date columns.
4. Merge two (2) data sources
   * It was necessary to merge both data sources to gain the extra information form the 3 columns for later analysis.
5. Removing and renaming columns with underscores
   * Because of the merging of data sources duplicate columns were created with underscore. We renamed these to reflect better presentation in our databases.
6. Capitalize column names in data frame for adding to database
   * We capitalize the columns names.

**LOAD**

**Loading Data**

Create database and made a secure connection to pgAdmin and stored data from pandas.

**SCHEMA**

Our data source was very similar, so we only needed on table.