**DATA CLEANING/WRANGLING STEPS:**

**Gun\_violence df (shape: 239,677 rows x 28 columns)**

Created month, day, year cols and deleted ‘date’ col

Deleted incident\_url, source\_url, incident\_url\_fields\_missing, sources

Missing Values: There are several columns that are missing 40-80% of their values. I will delete those columns except for ‘gun\_type’ as I’m not sure yet if the data that is supplied will be useful when comparing stats with laws.

Deleted: participant\_relationship, participant\_name, participant\_age, location\_description, gun\_stolen

Duplicates: None found

Mixed-type cols: None found, though several columns were returned as having an ‘empty’ dtype

New shape: 239677 x 20

Discovered that there are 14 values missing from ‘state’ from 2013. I will remove 2013 data from all data sets.

2018 is a partial year so will remove from data sets.

New Shape: 225597 x 20

Created new files for the columns that required delimiting.

New: gun\_type (225597, 401), participant\_age\_group(225597, 104), participant\_gender(225597, 79), participant\_status(225597, 104), participant\_type(225597, 104), incident\_characteristics(225597, 20), type\_status\_combined(225597, 104)

Created new file (incident\_rate) and column ‘inc\_rate’

Shape: (204,3)

Removed from original data set: all of the above and city\_or\_county, address, congressional\_district, state\_house\_district, state\_senate\_district

New Shape: 239677 x 8

**Gun Laws**

Converted tsv file to csv and loaded into Excel

Converted to xlsx file

Checked for missing values (none)

Converted to csv and imported to Jupyter

Shape: 1450 rows x 137 cols

Duplicates: None found

Mixed-type cols: None found

Changed all col names to lowercase to match gun\_violence df

Remove all rows NOT pertaining to 2013-18

New shape: 300 rows x 137 cols

Determined necessary cols to answer 11 gun law questions (in Word)

Copy/pasted unnecessary cols to Excel and formatted list with single quotes for use in Jupyter

Deleted unnecessary cols

Removed 2013 and 2018 data due to incompleteness

New shape: 200 x 43

Created new law grouped cols:

assault\_weapon\_ban

guns\_at\_elemen\_sch

guns\_on\_college\_campus

concealed\_carry\_permit\_req

bg\_check\_required

dealer\_req\_license

conv\_stalker\_may\_own

violent\_offender\_may\_own

dv\_offender\_may\_own

restraining\_order\_may\_own

felons\_may\_own

permit\_req\_purchase

open\_carry\_permit\_req

open\_carry\_legal

registration\_req

Deleted remaining old cols

New shape: 200 rows x 17 cols

Added rows for DC

New shape: 204 x 17

**State Population Data Set**

Removed all columns not including 2013-18

Removed rows pertaining to US territories

Converted to CSV

Checked quality in Excel. No missing values or duplicates

Converted state names to abbreviations

Imported to Jupyter

Shape: 51 rows x 7 columns

Made ‘State’ column lowercase to match

Converted state names to abbreviations to match

Transformed df ahead of merging

* Added ‘year’ and ‘population’ cols
* Removed ‘2013’, ‘2014’, ‘2015’, ‘2016’, ‘2017’, ‘2018’ cols
* New shape: 306 rows x 3 cols

Removed 2013 and 2018 data due to incompleteness

New shape: 204 x 3

**State Party Control**

I created this data set in Excel from PDF tables.

I loaded the data into Excel and then removed all cols and rows

excepting those for the 50 states and the name of the party in control that year

I created a sheet for each year and then compiled them into one worksheet and

renamed the ‘state\_control’ column to represent each year (2013-18)

Original shape: 50 rows x 7 cols

Converted state names to abbreviations

Transformed df to match others

New shape: 300 rows x 3 cols

Added rows for DC

New shape: 306 x 3

Removed data for 2013 and 2018 due to incomplete data

New shape: 204 x 3

**Part 2: Limitations and Ethics**

The state population data are estimates.

The gun violence data set is from 2013-18. A more recent set could not be obtained though I reached out to GVA. There are also many missing values in most columns.

2013 data in the gun violence data set excluded 14 states (+DC)

2018 data consisted of a partial year so both 2013 and 2018 data were removed.

The data may also not include all gun violence incidents involving police officers or other law enforcement agencies.

Gun laws for DC were not provided in the original data set. I researched laws and all amendments that impact the categories and timeframe of the rest of the data and imputed values by hand.

**Part 3: Questions to Explore**

* Which state has the highest rate of incidents involving guns?
* Which state has the lowest?
* Which state has the strictest/least strict gun laws?
* What political party controls these states?
* Is there a type of gun that is used more frequently? Is this a result of existing gun laws?
* What is the age group/gender breakdown of those involved in the incidents?
* How many offenders were arrested/found/convicted?

**Merge Process**

I reordered columns for all dfs to match: col1 = state, col2 = year

Merged party control and laws into df\_pol\_laws (party\_laws.csv)

Shape: 204 x 18

Merged participant\_type and participant\_status into one dataset (type\_status\_combined)

Shape: (225597, 104)

Merged incident\_rate into inc\_ordered (incidents\_final)

Shape: (225597, 10)

**More Cleaning/Wrangling**

Separated violent and non-violent incidents by incident\_id to create 2 new dfs—violent\_incs (209035, 8) and nv\_incs (16562, 8)

Imported inc\_ordered, pop\_ordered, and violent\_incs

Dropped unnecessary cols

New Shape of df\_inc (225597, 3)

Created df\_inc\_rate by merging population col with incs\_ordered: Shape: (225597, 4)

Created inc\_type col, values == ‘N’, ‘NV’: Shape: (225597, 5)

Calculated total incs by state/year and inc\_type and placed in new cols, ‘total\_v\_incs’, ‘total\_nv\_incs’

Dropped ‘incident\_id’ col: Shape: (225597, 6)

Dropped duplicates grouped by state, year and renamed as df\_rate\_final

Shape: (204, 6)

Calculated inc\_rate by state/year and placed in new col, v\_‘inc\_rate’

Determined quantiles to create 5 categories for v\_inc\_rate flag by year

Created v\_inc\_rate\_flag col

Removed unnecessary cols

Shape: (204, 5)

Imported party\_laws and merged with df\_rate\_final (df\_inc\_final)

Shape: (204, 8)

Exported as incidents\_final.csv