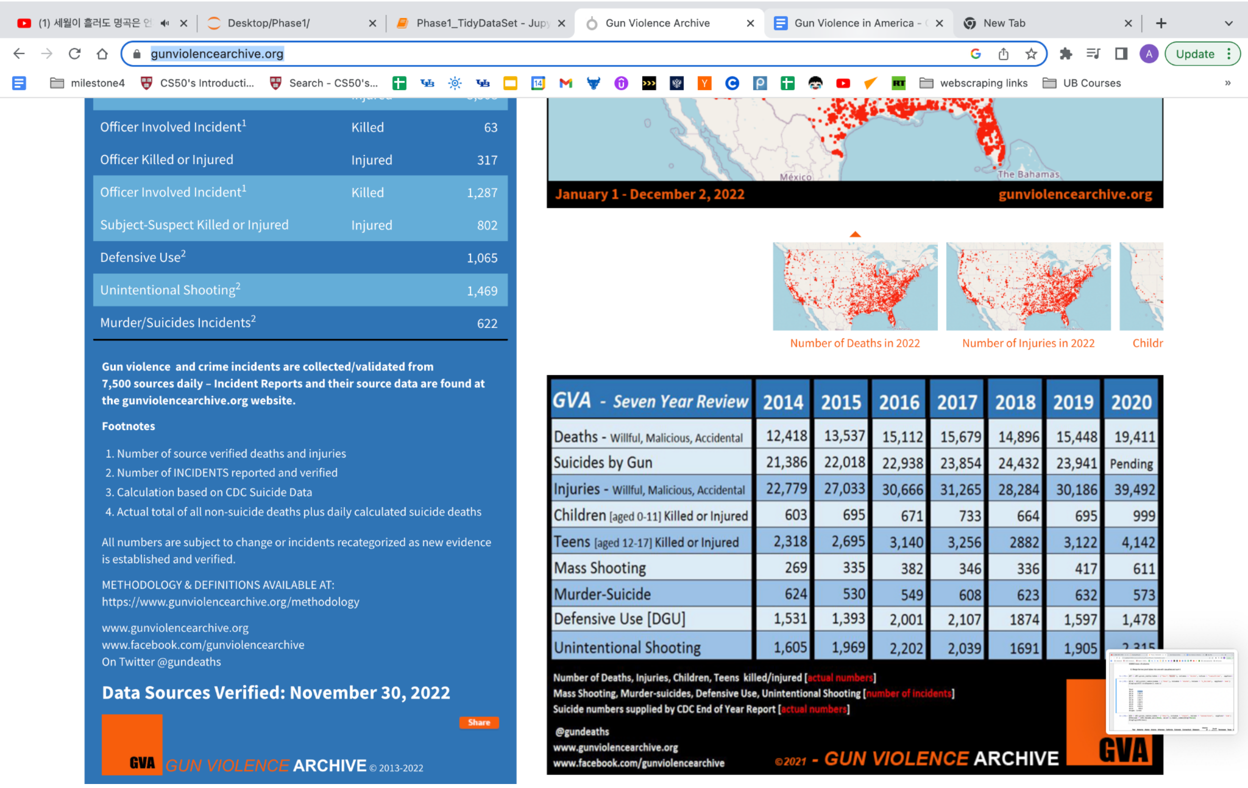
A regression model will be most effective to the user to help them understand the nature of the relationship between population size and gun violence in America. A classifier model may be used in the future to aid in future variable relationships that can be examined (gun culture, population, mass shootings, etc).

In this phase, I will enable a user to upload a data set in the form of a csv file or manually enter data entries and use multiple regressions to evaluate the data. The data expected will be gun victims per state and is structured by year. For now, a full dataset of all 50 states is expected for the model to work as expected.

The user dataset will cover data from the past eight years (2014,2021). If the user wishes to scrape the internet for more years the regression model will only benefit. However, it may be difficult to find accurate reports and one must take more time in order to find good data sets that work with the program models. For example, I have cross examined the data and compared it with other sources and the data that trains our mode for the past eight years appear to be accurate. However, finding good data going back in time is going to be a challenge and must be verified.

A website that displays data similar to ours: Deaths and Injuries (2014-2020)



Officer Involved Incident[1]	Killed	63
Officer Killed or Injured	Injured	317
Officer Involved Incident[1]	Killed	1,287
Subject-Suspect Killed or Injured	Injured	802
Defensive Use[2]		1,065
Unintentional Shooting[2]		1,469
Murder/Suicides Incidents[2]		622

Gun violence and crime incidents are collected/validated from 7,500 sources daily – Incident Reports and their source data are found at the gunviolencearchive.org website.

Footnotes

1. Number of source verified deaths and injuries
2. Number of INCIDENTS reported and verified
3. Calculation based on CDC Suicide Data
4. Actual total of all non-suicide deaths plus daily calculated suicide deaths

All numbers are subject to change or incidents recategorized as new evidence is established and verified.

METHODOLOGY & DEFINITIONS AVAILABLE AT:
https://www.gunviolencearchive.org/methodology

www.gunviolencearchive.org
www.facebook.com/gunviolencearchive
On Twitter @gundeaths

Data Sources Verified: November 30, 2022

GVA GUN VIOLENCE ARCHIVE © 2013-2022

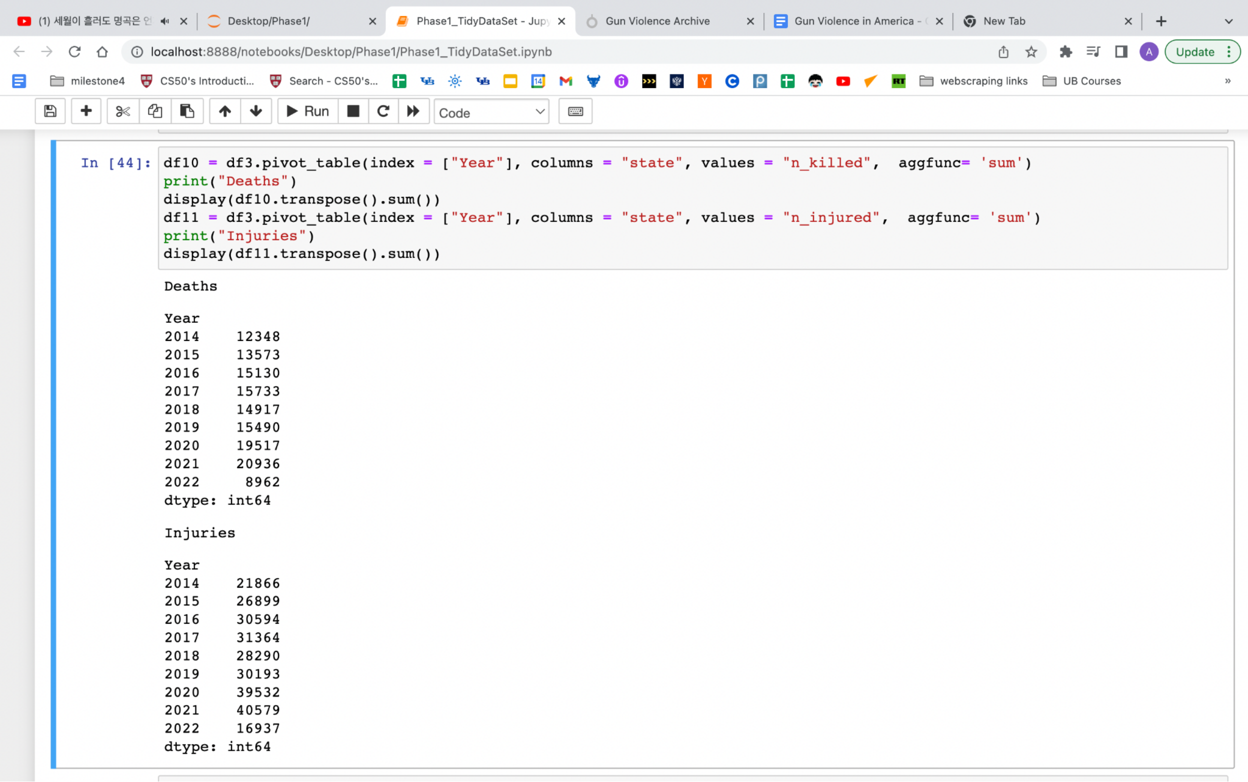
GVA - Seven Year Review	2014	2015	2016	2017	2018	2019	2020
Deaths - Willful, Malicious, Accidental	12,418	13,537	15,112	15,679	14,896	15,448	19,411
Suicides by Gun	21,386	22,018	22,938	23,854	24,432	23,941	Pending
Injuries - Willful, Malicious, Accidental	22,779	27,033	30,666	31,265	28,284	30,186	39,492
Children [aged 0-11] Killed or Injured	603	695	671	733	664	695	999
Teens [aged 12-17] Killed or Injured	2,318	2,695	3,140	3,256	2882	3,122	4,142
Mass Shooting	269	335	382	346	336	417	611
Murder-Suicide	624	530	549	608	623	632	573
Defensive Use [DGU]	1,531	1,393	2,001	2,107	1874	1,597	1,478
Unintentional Shooting	1,605	1,969	2,202	2,039	1691	1,905	2,315

Number of Deaths, Injuries, Children, Teens killed/injured [actual numbers]
Mass Shooting, Murder-suicides, Defensive Use, Unintentional Shooting [number of incidents]
Suicide numbers supplied by CDC End of Year Report [actual numbers]

@gundeaths
www.gunviolencearchive.org
www.facebook.com/gunviolencearchive

©2021 - GUN VIOLENCE ARCHIVE GVA

Program model's data: Deaths and Injuries (2014-2022)



```
In [44]: df10 = df3.pivot_table(index = ["Year"], columns = "state", values = "n_killed",  aggfunc= 'sum')
         print("Deaths")
         display(df10.transpose().sum())
         df11 = df3.pivot_table(index = ["Year"], columns = "state", values = "n_injured",  aggfunc= 'sum')
         print("Injuries")
         display(df11.transpose().sum())
```

```
Deaths

Year
2014    12348
2015    13573
2016    15130
2017    15733
2018    14917
2019    15490
2020    19517
2021    20936
2022     8962
dtype: int64

Injuries

Year
2014    21866
2015    26899
2016    30594
2017    31364
2018    28290
2019    30193
2020    39532
2021    40579
2022    16937
dtype: int64
```

My program checks robustness by normalizing the input data. When it does so we see a distinction between gun violence and gun victims in America. Once we take capita into account we notice two things.

1. States with high populations tend to have higher victims but not always. California has a high population but does not actually have high rates of gun violence per capita.

2. The inverse is also true. States with low populations Louisiana appears to be the immediate outlier, it has a total number of victims equal to states with higher populations.

My program does not fine tune the parameters in any statistical way which can enable a more flexible regression model to be used. In other words, since the program considers user data to be considered as structured, and all needed data is present and the regression plot will accurately make predictions.

<u>Several key findings by the actual regression</u>

1. The model with the highest accuracy prediction score is the regression of all 50 states' total victims vs population. However this is misleading.
2. Larger population have a lot more variability
3. Regression does not work well when we take capita into account.

<u>Future iterations of this model may work on the following</u>

1. Building a stronger classifier model to determine which relationship to model with a regression.
2. Our data is small right now. I had a lot of big data but cleaned it up. Since I did not generalize the dataset, I could keep adding more features to it, distinguish between red flags like mass shootings, exc.
3. Design an implementation which would stream step 2. Our program does not have a direct pipeline to showcase results.