A Data Exploration of Gun Violence in the U.S.

SPRING 2021

COHORT 2

TEAM 97

Stephanie Garcia
Stephanie Jung
Emilio Ramirez
Joe Reynolds

DS 4A EMPOWERMENT

correlation one

Table of Contents

1. INTRODUCTION

1.1 Problem Overview

There are so many guns in the United States and many instances of gun violence that have changed all of our daily lives. Awareness of the prevalence of gun violence and the array of weapons used is low. Most people do not know which guns are recreational, defensive, or excessive. Nevertheless, their overwhelming presence in the U.S. has more cost than gain:

"Not only does the industry create jobs, it also generates sizable tax revenues. In the United States, the industry and its employees pay over $6.98 billion in taxes including property, income, and sales-based levies." National Shooting Sports Foundation 2021 U.S. Economic Impact Report

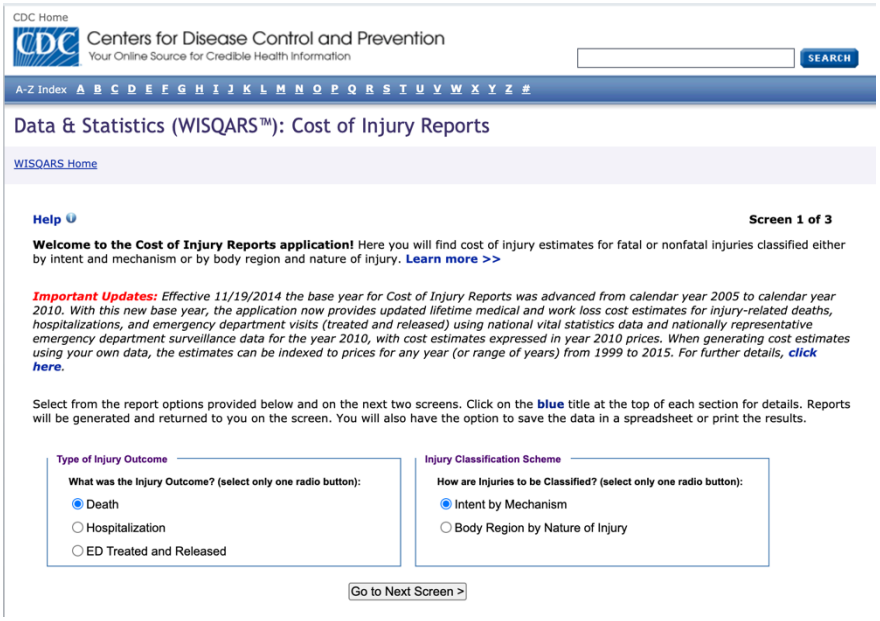
"In an average year, gun violence in America kills nearly 40,000 people, injures more than twice as many, and costs our nation $280 billion. This staggering figure is higher than the entire US Department of Veterans Affairs' annual budget. Without a doubt, the human cost of gun violence—the people who are taken from us and the survivors whose lives are forever altered—is the most devastating"
Economic Cost of Gun Violence Report 2021

ECONOMIC CONTRIBUTION OF ARMS AND AMMUNITION INDUSTRIES, 2020

	Direct			Suppliers			
	Jobs	Wages	Output	Jobs	Wages	Output	Jobs
Alabama	3,227	$ 131,004,200	$ 508,477,900	1,477	$ 94,148,900	$ 332,637,000	1,947
Alaska	677	$ 20,857,000	$ 56,493,100	161	$ 9,391,200	$ 28,438,300	297
Arizona	3,901	$ 272,039,200	$ 911,161,500	2,095	$ 167,389,500	$ 467,407,400	3,509
Arkansas	3,423	$ 136,463,900	$ 839,248,300	1,633	$ 101,124,000	$ 374,049,100	2,057
California	10,010	$ 521,860,900	$ 1,478,666,000	5,318	$ 476,949,600	$ 1,325,131,200	7,868
Colorado	2,966	$ 144,309,800	$ 442,582,800	1,603	$ 122,523,100	$ 328,147,700	2,271
Connecticut	2,146	$ 184,771,200	$ 711,099,200	1,287	$ 136,966,700	$ 357,770,100	1,906
Delaware	189	$ 5,488,400	$ 10,663,300	77	$ 5,274,700	$ 20,991,700	134
District of Columbia	115	$ 6,888,300	$ 10,936,300	33	$ 3,789,400	$ 7,699,300	65

1.2 Rescoping Problem Overview

Initially our team wanted to an easily accessible visual representation of the cost compared to the revenue of the firearms industry in each state. The goal was to provide citizens and industries a clear data tool to better understand the fiscal impact of guns in the United States. The tool was proposed as a map that when you hover over each state you see the values and a horizontal bar in two colors that are proportional to the revenue and costs of guns in the state.



The data exploration for this project idea began with revenue data generated in the National Shooting Sports Foundation 2021 U.S. Economic Impact Report. We were using the CDC's Web-based Injury Statistics Query and Reporting System (WISQARS) injury costs to calculate the costs of injuries in the United States from a variety of different causes, including, but not limited to, firearms.

After reviewing with our mentors and TA's, the group decided to "let the data lead the way" and work with a Gun Violence database we found after some cursory research for gun related databases.

1.3 Transitioning to New Scoped Project

With the introduction of our new starting dataset, the goal of the project became exploring the database using the various data science techniques we learned to gain a nuanced understanding of gun violence in the U.S.

1.4 Team 97 Project Description

This project aims to find factors that impact gun violence, specifically whether gun purchases, laws, and political affiliations in certain states can predict or have an effect on gun violence.

2. DATA: COLLECTION & CLEANING

2.1 Gun Violence Archive Database (GVA)

The data was downloaded from Gun Violence Archive's website. From the organization's description:

> Gun Violence Archive (GVA) is a not for profit corporation formed in 2013 to provide free online public access to accurate information about gun-related violence in the United States. GVA will collect and check for accuracy, comprehensive information about gun-related violence in the U.S. and then post and disseminate it online.

All credits for the data go to Gun Violence Archive.

Field	Type	Description	Required?
incident_id	int	gunviolencearchive.org ID for incident	yes
date	str	date of occurrence	yes
state	str	US State	yes
city_or_county	str	US City or county	yes
address	str	address where incident took place	yes
n_killed	int	number of people killed	yes
n_injured	int	number of people injured	yes
incident_url	str	link to gunviolencearchive.org webpage containing details of incident	yes
gun_stolen	dict[int, str]	key: gun ID, value: 'Unknown' or 'Stolen'	no
gun_type	dict[int, str]	key: gun ID, value: description of gun type	no
incident_characteristics	list[str]	list of incident characteristics	no
n_guns_involved	int	number of guns involved	no
participant_age	dict[int, int]	key: participant ID	no
participant_age_group	dict[int, str]	key: participant ID, value: description of age group, e.g. 'Adult 18+'	no
participant_gender	dict[int, str]	key: participant ID, value: 'Male' or 'Female'	no
participant_name	dict[int, str]	key: participant ID	no
participant_relationship	dict[int, str]	key: participant ID, value: relationship of participant to other participants	no
participant_status	dict[int, str]	key: participant ID, value: 'Arrested', 'Killed', 'Injured', or 'Unharmed'	no
participant_type	dict[int, str]	key: participant ID, value: 'Victim' or 'Subject-Suspect'	no

i jamesqo Github - Gun Violence Data: data for all recorded gun violence incidents in the US between January 2013 and March 2018. Size: 147 MB 239677 rows × 29 columns

2.2 National Instant Criminal Background Check (NICS)

This dataset was used as a proxy for firearms sold. When a person tries to buy a firearm, the seller, known as a Federal Firearms Licensee (FFL), contacts NICS electronically or by phone. The prospective buyer fills out the ATF form, and the FFL relays that information to the NICS.

field	type	description
month	object	date
state	object	date
permit	float64	Permit to purchase gun
permit_recheck	float64	Permit check for current permit holders
handgun	float64	Any firearm which has a short stock and is designed to be held and fired by a single hand;
long_gun	float64	A weapon designed to be shot from the shoulder
other	float64	Frames, receivers, other firearms that are not handgun or long_gun
multiple	int64	Multiple gun purchases
totals	int64	handgun + long_gun + other

[ii] FBI NICS Firearm Background Check Data, Size: 1.1 MB 14905 rows × 27 columns

2.3 Census Population Data

Field	**Type**	**Description**
Year	**int**	**Census Year from 2010 to 2019**
state	**object**	**US state name**
Total Population	**int**	**Total Population of state in a specific Year**

[iii] State Population Totals and Components of Change: 2010-2019, Size: 16 KB 51 rows × 10 columns

2.4 Cleaning GVA Database

participant_age	participant_age_group	participant_gender	participant_name
0::20	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...	0::Male||1::Male||3::Male||4::Female	0::Julian Sims
0::20	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...	0::Male	0::Bernard Gillis
31||2::33||3::34||4::33	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...	0::Male||1::Male||2::Male||3::Male||4::Male	0::Damien Bell||1::Desmen Noble||2::Herman Sea...
29||1::33||2::56||3::33	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::A...	0::Female||1::Male||2::Male||3::Male	0::Stacie Philbrook||1::Christopher Ratliffe||...
18||1::46||2::14||3::47	0::Adult 18+||1::Adult 18+||2::Teen 12-17||3::...	0::Female||1::Male||2::Male||3::Female	0::Danielle Imani Jameison||1::Maurice Eugene ...

The initial data cleaning was straightforward with the exception of the Gun Violence Archives. The web-scraping done via [James Qo's GitHub] contained the data aggregated by the unique Incident ID. We had to de-aggregate the data to have all of the information on their own row.

The double colon and double vertical bars made it difficult to split using the methods we learned in the program. We were able to split the data using the code below:

gun_type_split = best_columns_head['gun_type'].str.split(pat="\|\|", expand=True)

"A Regular Expression (RegEx) is a sequence of characters that defines a search pattern. Backlash \ is used to escape various characters including all metacharacters. This makes sure the character is not treated in a special way."[iv]

We were able to split the data using the same code this time splitting by double colons :: and drop the first column which was the index within the data table. This process was done for several columns that we wanted to do analysis and we ended up with a cleaned version (see below).

incident_id	age	age_group	type	gender	status
461105	20	Adult 18+	Victim	Male	Arrested
460726	20	Adult 18+	Victim	Male	Killed
478855	25	Adult 18+	Subject-Suspect	Male	Injured, Unharmed, Arrested
478925	29	Adult 18+	Victim	Female	Killed
478959	18	Adult 18+	Victim	Female	Injured
...	...	...	...	...	...
575663	None	Adult 18+	Subject-Suspect	Male	Unharmed, Arrested
577157	19	Adult 18+	Victim	Male	Injured
577157	19	Adult 18+	Victim	Male	Injured
577157	30	Adult 18+	Victim	Male	Injured
577157	29	Adult 18+	Victim	Female	Injured

385476 rows × 5 columns

2.5 Cleaning NICS Database

To make best use of the NICS data set, we created a Year column to help organize the final selection of results. Although the NICS database included years 2010 – 2020, we removed the tail end years to match our years of GVA data (2013-2018). Removing unwanted columns helped us visualize which columns are most sensible to sum together and include in the "Total" column we needed to create. Adding together the values from our columns of interest gave us a total amount of background checks produced by state per month and year.

```
nics_copy = nics.copy()
nics_copy["month"]=pd.to_datetime(nics_copy['month'])
nics_copy["year_only"] = pd.to_datetime(nics_copy['month'])
nics_copy["year_only"] = nics_copy['year_only'].dt.strftime('%Y')
nics_copy["year_only"] = nics_copy["year_only"].astype(int)
nics_copy = nics_copy[(nics_copy['year_only'] >= 2013) &
                      (nics_copy['year_only'] <= 2020)]
nics_final = nics_copy.drop(["return_to_seller_other","return_to_seller_long_gun","return_to_seller_handgun",
             "totals", "rentals_handgun","rentals_long_gun",'permit_recheck',"admin",
             'prepawn_handgun', 'prepawn_long_gun','prepawn_other','returned_handgun',
                       'returned_long_gun','returned_other'], axis=1)
nics_final["total"]= nics_final.loc[:,['permit', 'handgun', 'long_gun',
       'other', 'multiple', 'redemption_handgun', 'redemption_long_gun',
       'redemption_other', 'private_sale_handgun', 'private_sale_long_gun',
       'private_sale_other']].sum(axis=1)
nics_final.head(5)
```

	month	state	permit	handgun	long_gun	other	multiple	redemption_handgun	redemption_long_gun	redemption_other	private_sale_handgun	pr
275	2020-12-01	Alabama	33421.0	31103.0	28933.0	1855.0	1388	2447.0	1121.0	9.0	30.0	
276	2020-12-01	Alaska	441.0	3658.0	3822.0	430.0	233	123.0	84.0	1.0	9.0	
277	2020-12-01	Arizona	9043.0	25093.0	14562.0	1942.0	1395	1082.0	349.0	4.0	21.0	
278	2020-12-01	Arkansas	3878.0	10987.0	12719.0	587.0	510	1043.0	935.0	2.0	2.0	
279	2020-12-01	California	26034.0	59909.0	39389.0	6932.0	0	450.0	273.0	20.0	523.0	

3. EDA: EXPLORATORY DATA ANALYSIS

3.1 Initial GVA Statistics

We used the describe function on python on the populations of people injured and people killed from the GVA dataset to give us a numerical representation of the skewness, thus giving us an understanding of the data distribution. with the mean number of people killed and injured being 0.25, and 0.49, respectively. Using the Pearson Mode Skewness we can say that the dataset is right skewed, for the median of both populations is 0, which is less than their means.

Statistical Summary of People Injured and Killed in Gun Violence from GVA Dataset

```
In [18]: df['n_injured'].describe()

Out[18]: count    239677.000000
         mean          0.494007
         std           0.729952
         min           0.000000
         25%           0.000000
         50%           0.000000
         75%           1.000000
         max          53.000000
         Name: n_injured, dtype: float64
```

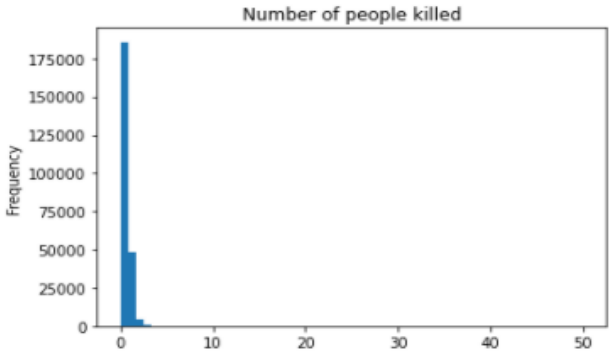
```
In [20]: df['n_injured'].quantile(0.95)

Out[20]: 2.0
```

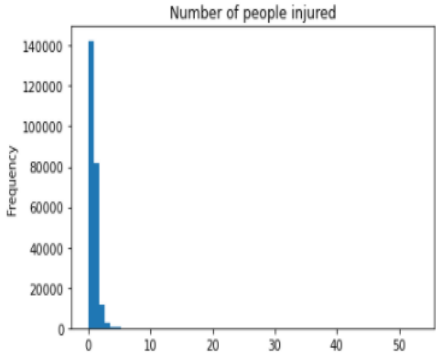
```
In [12]: df['n_killed'].describe()

Out[12]: count    239677.000000
         mean          0.252290
         std           0.521779
         min           0.000000
         25%           0.000000
         50%           0.000000
         75%           0.000000
         max          50.000000
         Name: n_killed, dtype: float64
```

```
In [19]: df['n_killed'].quantile(0.95)

Out[19]: 1.0
```

Given that the median was 0 and the 75th quantile are was 1 person injured and 0 people killed per incident, we wanted to know which quantile would yield more than one incident for injuries, and which quantile would have people that are killed. After trial and error, we found that 2 people are injured per incident at the 95th quantile and one person is killed at the 95th quantile. This shows how far right skewed the gun violence incidents are, although they may be very disruptive, most of the time they don't injure nor kill anyone.

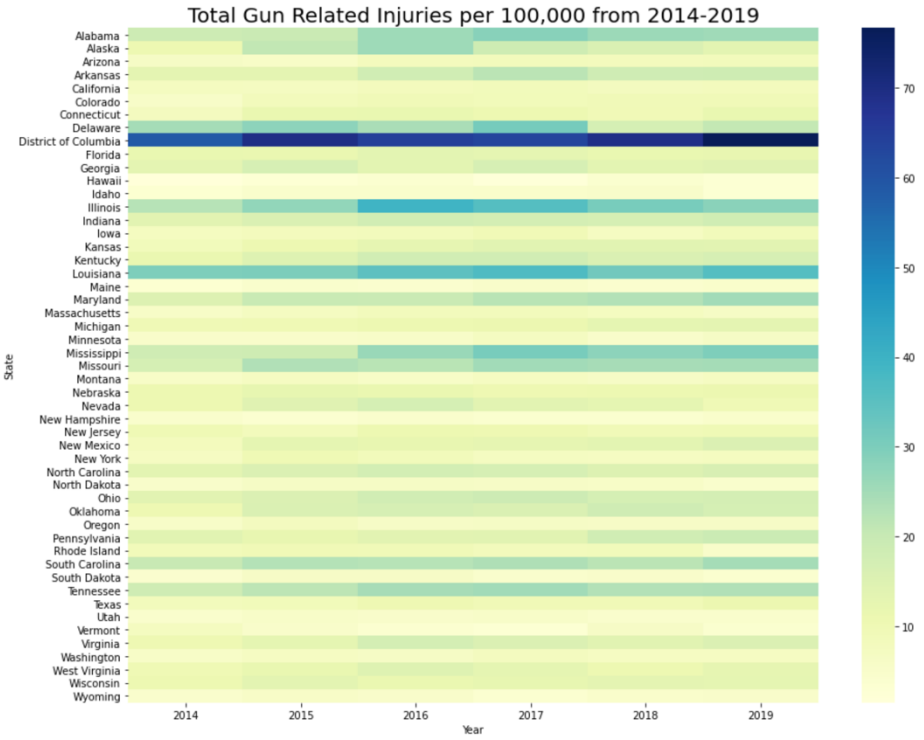
Histograms of People Killed and Injured





Given our understanding of the summary distribution, we wanted to have a visualization of the distribution of shootings, because the way gun violence is portrayed in the media makes it seem that mass shootings are the most important and impactful type of gun violence. However, the histograms and the summary statistics showed that the data is extremely right skewed, which confirm the summary statistics, and they give a good visual representation of how gun violence is very prevalent although it doesn't kill or injure that many people in one isolated incident.

To continue our EDA, we wanted to visualize in a concrete manner the density of gun violence per state across the years. This was important because we understood that if we looked at the raw numbers of incidents per state, heavily populated states would appear disproportionately high in gun violence, whereas we wanted to know how prevalent gun violence was per each state's population. Shown on the heat map, we further explore the top states with incidents later in the report.

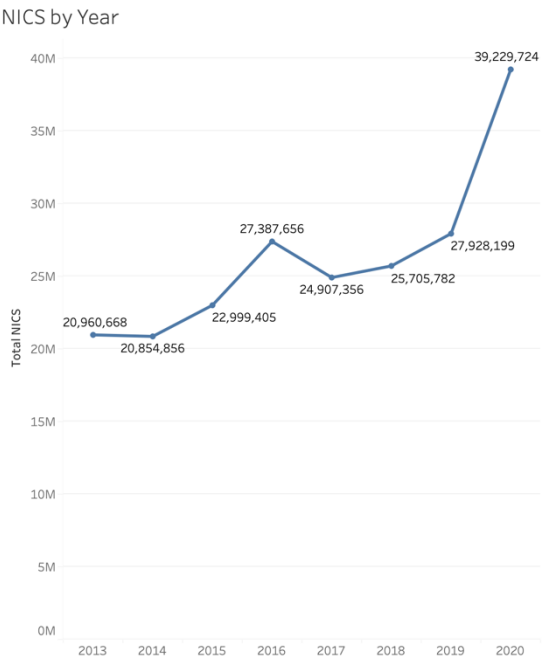


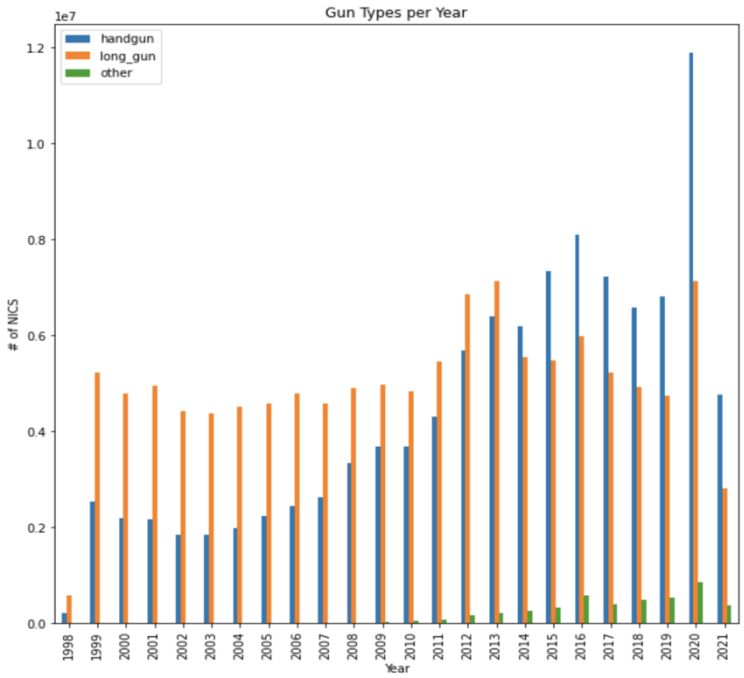
3.2 Initial NICS Statistics

The data and graph show the number of NICS per year from 2013-2020. 2020 had a substantial spike with the most background checks in our dataset with 39,326,079 background checks performed. This reinforces the suggestion that many Americans bought firearms during 2020 because of the national turmoil revolving around the pandemic and the various protests that occurred. It looks like general election years typically have an increase of NICS/gun sales.

Amount of NICS by Year (Graphed 2013-2021)

year_only	
1998	891898
1999	9043747
2000	8427096
2001	8820045
2002	8367069
2003	8402244
2004	8579891
2005	8867055
2006	9993057
2007	11162272
2008	12692985
2009	13984953
2010	14320489
2011	16336732
2012	19463832
2013	20966704
2014	20860241
2015	23026288
2016	27425512
2017	24955919
2018	25769441
2019	28007320
2020	39326079
2021	19020349





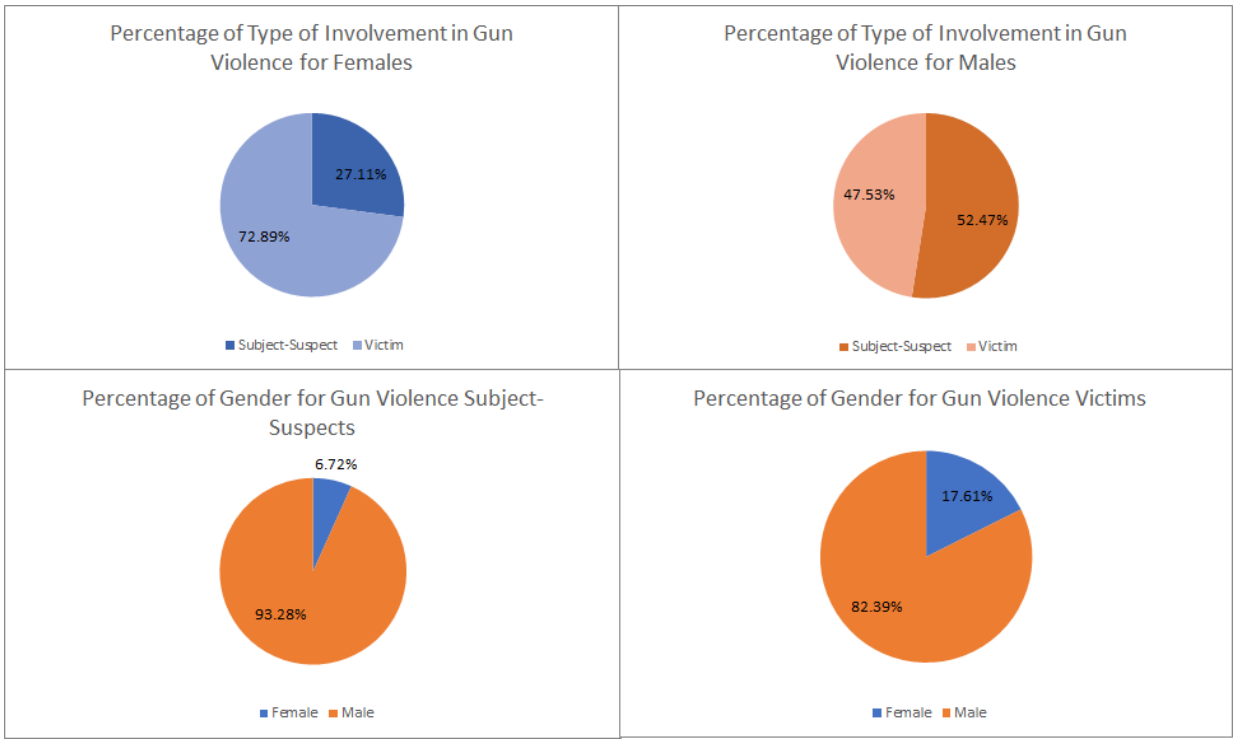
There has been an overall increased amount of NICS done for handguns than long guns. This may be because a lot of people wanted a gun for "self-protection" due to fears of COVID, whereas long guns may be used for more recreational activities (i.e. hunting, sporting). "Other" rises as well indicating stocking up on weapons suspected to be made unavailable or the sheer number of buyers for more common types of guns created a shortage that led to more "Other" sales.

Scale is 0.1=1,000,000

3.3 Insights on Gender in GVA

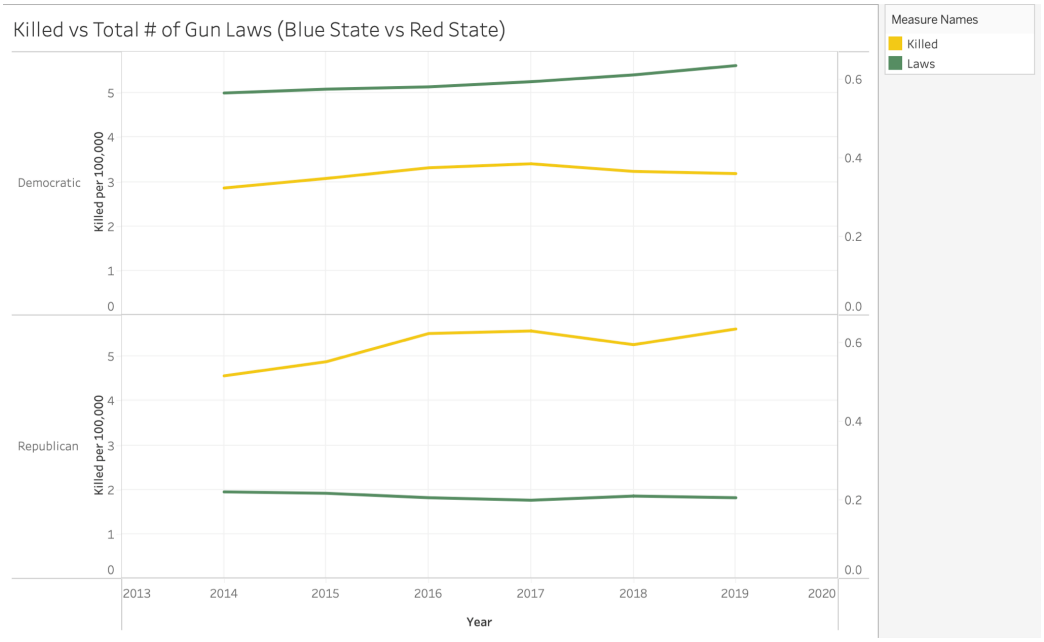
When observing the gun violence incidents between 2013-2018 from the GVA dataset we were particularly interested in the breakdown of gender in gun violence. Although the data is limited for gender (it only includes male or female, thus excluding other gender identities for analysis), the results were nonetheless very interesting.

Pie Charts for Distribution of Involvement Based on Gender and Involvement Type

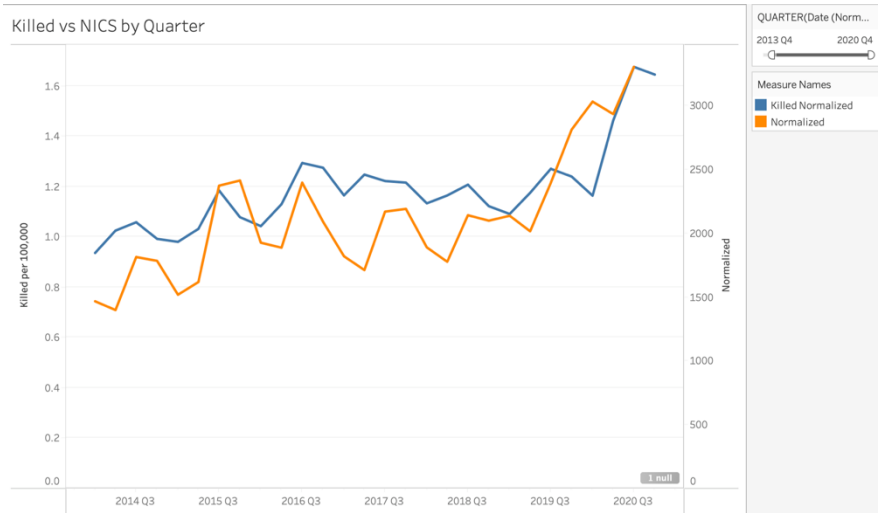


We started by creating two contingency tables that took as variables the genders (Male, Female) of people involved in gun violence, and the type of involvement in gun violence(subject/suspect, victim). The contingency tables were converted into pie charts for ease of visualization as shown above. We observed that females are mostly victims when they are involved in gun violence, as only 27% of females are subject/suspects and 73% of them are victims as shown in the top left pie chart. On the other hand, males have a more even split of 52.5% as Subject-Suspect and 47.5% as victims across their involvement based on the top right chart. However, when it comes to overall involvement, females are only 6.7% of subjects and 17.6% percent of victims per the bottom pie charts. This shows that males are predominantly involved in gun violence, and when it comes to their involvement, they have an even split on both sides of the violence as both perpetrators and victims, but when females are involved in gun violence, they are mostly the victims.

3.4 Insights on NICS

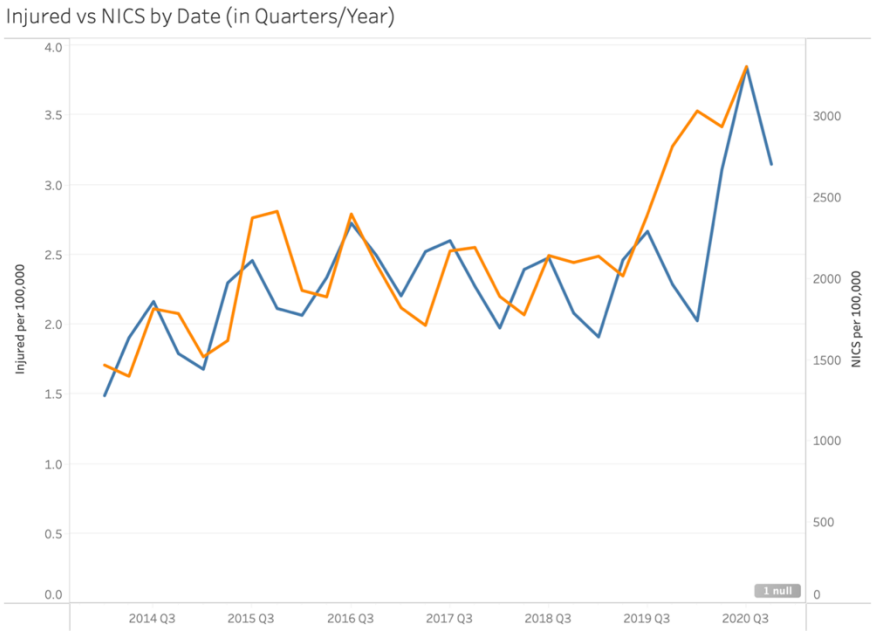


The two multi-line graphs above show the number of people killed per 100,000 (in green) and the number of gun control related laws per 100,000 (in yellow). The graph above represents the democratic states and shows a general increase in gun control related laws and a general decrease in deaths due to gun violence, whereas the republican states show a general decrease in gun control related laws and a general increase in deaths due to gun violence. This shows a slight correlation as laws go up, deaths go down and gives insight that increasing gun control laws may help decrease the overall deaths due to gun violence.



The first image shows the number of people killed per 100,000 with the number of NICS per 100,000. We initially graphed them on the same date, however the correlation was closer when we pulled the NICS data forward by 3 months. This means, as there was an increase in NICS, around 3 months later there was an increase in deaths due to gun violence.

Color code is the same for both graphs.
Left is Killed = Blue, Right is Injured = Blue



The second graph compares the number of people injured per 100,000 compared to NICS per 100,000. This makes sense since injuries are more common than deaths in these gun incidents.

Both graphs show that the more NICS are performed, which we are using as a proxy for gun purchase, the more gun deaths and injuries occur. Since NICS is one of the only steps that could stop someone from purchasing a gun, it would be beneficial to add something right before to decrease the number of NICS or right after when we take into account that NICS is a proxy for gun purchase. We could have a policy that enforces an educational test to pass before allowing the purchase. This would not technically decrease the number of NICS done but could reduce the overall amount of firearm purchases.

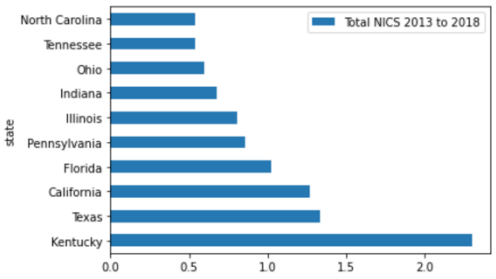
There was a large increase of NICS in 2020 (during the start of COVID-19). Given the above graph and associated data, it looks like general election years typically have an increase of NICS/gun sales.

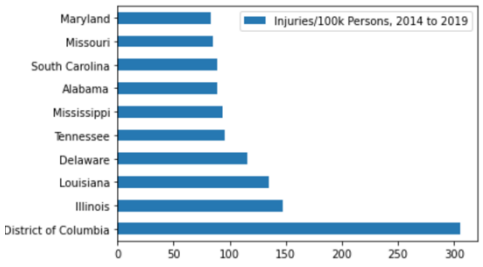
4. DATA VISUALIZATION AND MODELING

4.1 Synthesizing GVA, NICS, and Census

Our data exploration of let us know the profile of our most common suspect/shooter and the impact of NICS in general, but we still wanted to verify which states truly had the most violence alongside their respective amount of NICS.

The first set of queries we needed to visualize were the top 10 states getting permits in order to compare them to states with the most deaths and injuries.





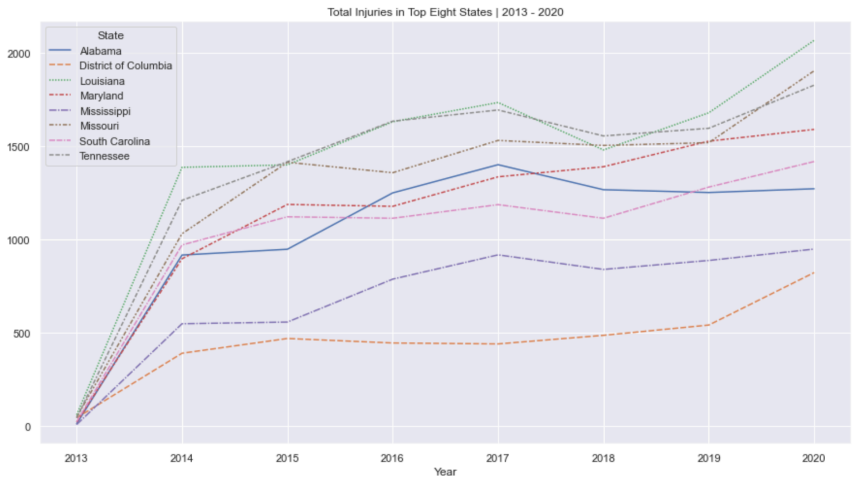
State	Deaths/100k Persons, 2014 to 2019
District of Columbia	103.868776
Louisiana	66.292235
Mississippi	58.865759
Alabama	55.532604
Missouri	52.342464
Alaska	48.946669
South Carolina	47.509600
Maryland	42.344753
Tennessee	40.995678
Arkansas	38.817583

The first hypothesis for this query set was the states with more NICS would have less incidents and states with less NICS would show greater levels of violence. The top ten NICS states only has one state also present in the injury/death top tens: Tennessee.

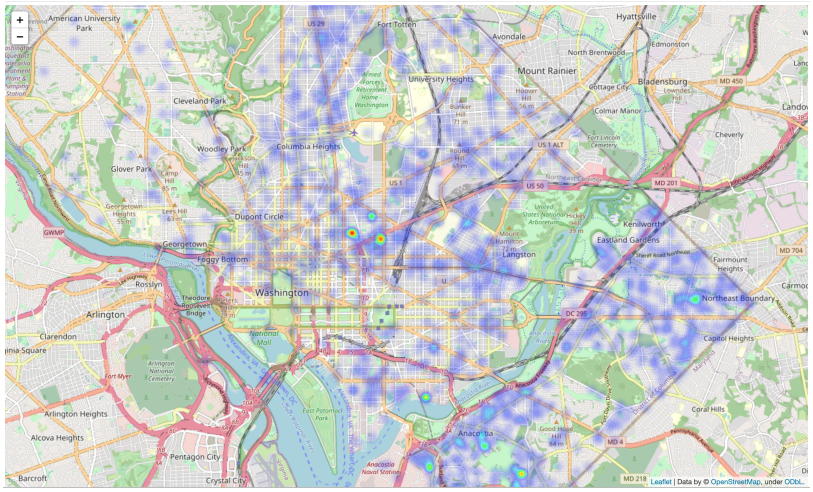
We believe Tennessee exists in this interesting spot because Kentucky, the highest conductor of NICS, is a bordering state; however, Mississippi, Alabama, Missouri, Arkansas are bordering states as well, but on the top ten injury/death lists.

	Deaths/100k Persons, 2014 to 2019	Injuries/100k Persons, 2014 to 2019	Total NICS/100k Persons, 2013 to 2018
District of Columbia	103.868776	305.353794	0.24281
Louisiana	66.292235	134.625142	27.75373
Mississippi	58.865759	93.336866	20.73212
Alabama	55.532604	88.968577	52.47702
Missouri	52.342464	85.364570	45.11536
South Carolina	47.509600	88.818185	28.67616
Maryland	42.344753	82.902749	14.18553
Tennessee	40.995678	95.743517	54.19917

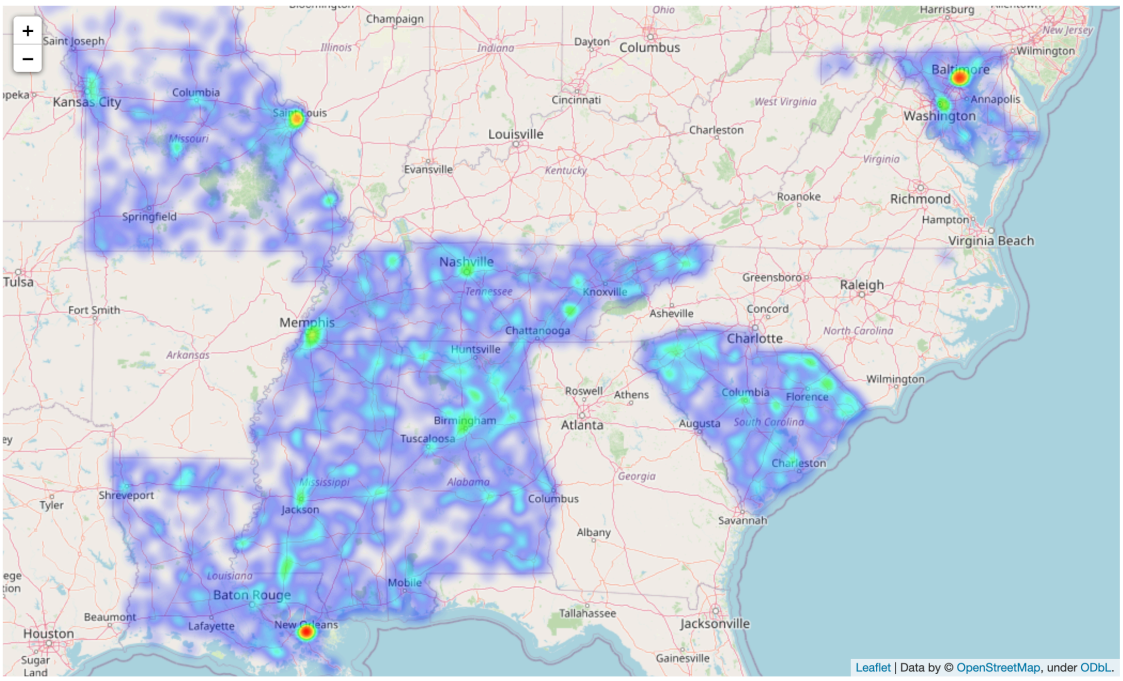
By joining the Total NICS from 2013 to 2018 for each state, we begin to see a relationship between true gun violence levels and the amount of background checks being conducted in those states. Washington DC our deadliest state, also has one of the lowest amounts of NICS. We'd need to check on Washington DC's gun policies to identify why less NICS are being conducted.



Map of Total Gun Related Injuries from 2013-2018 in Washington DC



Map of Total Gun Related Injuries from 2013-2018 in the top Eight Deadly States



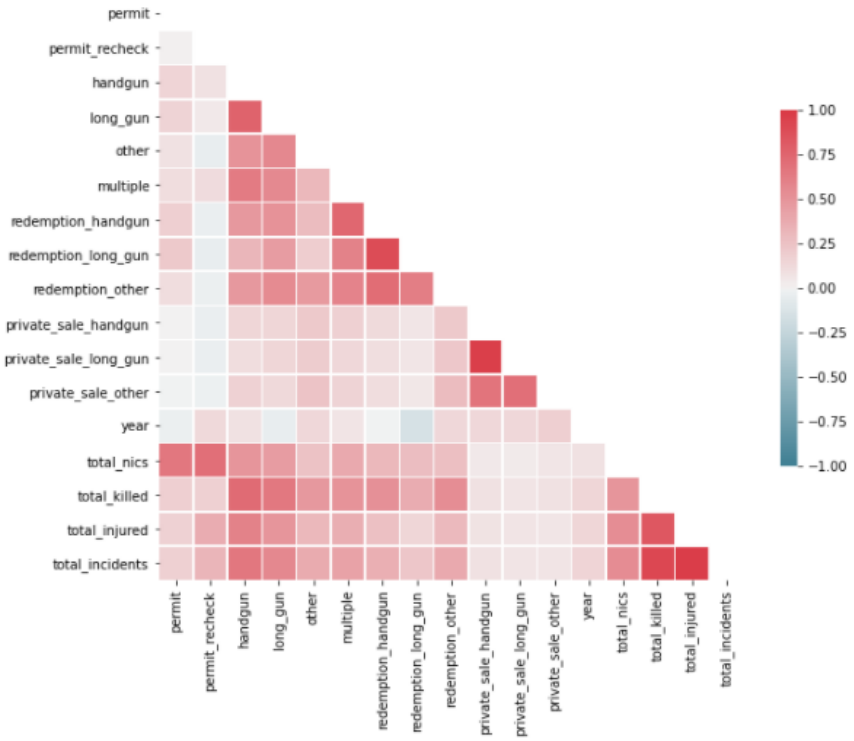
To gain an even more focused view we cross referenced the top ten injuries list against the deaths to resolve with final eight states in the previous table. Using the selected states and pulling the coordinates off the gun incident data in the GVA data set, we were able to map the exact location of all gun related injuries from 2013 to 2018 in the eight most deadly states in the U.S.

By mapping the total gun related injuries in the "top ten" states after being normalized, we were able to see which areas needed immediate policy shifts and the local representation we should prioritize reaching out to deliver to deliver the information.

4.2 Statistical Analysis and Predictive Modeling

We aimed to train a model that would predict the total number of gun violence incidents based on different attributes related to guns. Given the available datasets, we merged the gun incidents information per state from the GVA dataset with the NICSs dataset. Before we could create a model, we wanted to visualize which variables had the strongest correlation with gun violence (in this case, any correlation with total killed, total injured, and the sum of total injured and total killed, called total incidents).

Correlation Matrix for Merged Dataset between GVA and NICS Dataset



In the correlation matrix the stronger a red cell is, the stronger positive correlation is, whereas the stronger a blue cell is the stronger the negative correlation becomes. In this case we see a moderate correlation coefficient of 0.54 for total_nics and total_incidents as well as a moderate correlation of 0.67 between handguns and total_incidents. This gives us a good initial indication of the hypothesis that there is a relationship between the number of guns bought and the number of gun violence incidents. Furthermore, given the information gathered earlier in our EDA we understand that most incidents only involve one person. This leads to a reasonable assumption that a simple gun like a handgun is more likely to be associated with most gun violence incidents given that

handguns are more often bought for personal protection and recreation, whereas long guns like rifles are used for hunting, or semi-automatic weapons may be used for recreation or in very rare instances for mass shootings.

Preliminary Model: total_incidents ~ total_nics

OLS Regression Results

Dep. Variable:	total_incidents	R-squared:	0.300
Model:	OLS	Adj. R-squared:	0.300
Method:	Least Squares	F-statistic:	1909.
Date:	Fri, 06 Aug 2021	Prob (F-statistic):	0.00
Time:	16:45:29	Log-Likelihood:	-25096.
No. Observations:	4451	AIC:	5.020e+04
Df Residuals:	4449	BIC:	5.021e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	41.7212	1.223	34.121	0.000	39.324	44.118
total_nics	0.0007	1.52e-05	43.694	0.000	0.001	0.001

Omnibus:	837.257	Durbin-Watson:	2.195
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3993.999
Skew:	0.829	Prob(JB):	0.00
Kurtosis:	7.334	Cond. No.	9.64e+04

Initially we attempted to build a simple model that would take as input the total number of NICS that were run in order to predict the possible number of people involved. As shown above, this model only has a R-squared of 0.300, and a P value of total_nics of 0.000. This suggests that we needed to develop a more robust model, but the relationship of total_nics and total_incidents is significant at the 0.05 significance level.

When we added states as a categorical variable to the model, our R-squared improved, but the relationship between the total number of incidents per state changes, with some having negative coefficients, which may mean that some states don't have that strong of a relationship with background checks and total incident relative to other states within the normalization curve.

Model of Total Incidents with Total NICS and State as Independent variables

OLS Regression Results

Dep. Variable:	total_incidents	R-squared:	0.759
Model:	OLS	Adj. R-squared:	0.756
Method:	Least Squares	F-statistic:	272.0
Date:	Fri, 06 Aug 2021	Prob (F-statistic):	0.00
Time:	16:51:07	Log-Likelihood:	-22721.
No. Observations:	4451	AIC:	4.555e+04
Df Residuals:	4399	BIC:	4.588e+04
Df Model:	51		
Covariance Type:	nonrobust		

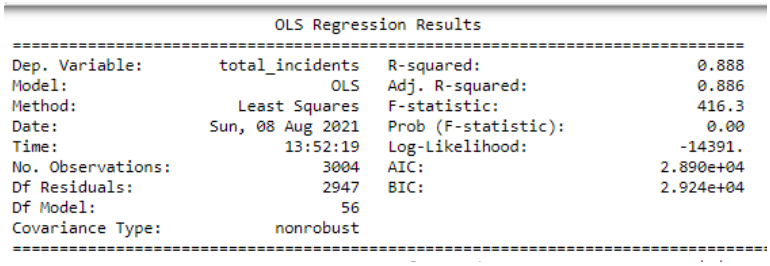
When we integrate the purchases of handguns, long guns, and other type of guns the R-squared increases substantially from our crude model of 0.300 to 0.759 with a decrease of AIC from 45,550 to 44,940, indicating that the model is more robust and we are not running into the risk of overfitting with the added state variable.

We wanted to add further variables to train a more substantial model. We chose permit_recheck as an extra variable as we hypothesized that a person that needs a permit re_check might have a higher propensity to commit a crime. Running this model yielded another boost to the R-squared value of 0.887 while also decreasing the AIC further down to 28,920.

At this point, the R-squared value suggested that the model is somewhat robust, but we attempted to add the variable "multiple" to the model, assuming that people doing multiple

gun purchases would have a propensity towards total incidents, given that owning multiple firearms could increase the risk of losing a gun, having a gun stolen, misplacement of gun within home that may lead to accidents, or even a propensity towards using guns for nefarious means.

Total NICS, States, Handgun, Long Gun, Other, Permit Recheck, and Multiple guns as Independent variables



OLS Regression Results

Dep. Variable:	total_incidents	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.886
Method:	Least Squares	F-statistic:	416.3
Date:	Sun, 08 Aug 2021	Prob (F-statistic):	0.00
Time:	13:52:19	Log-Likelihood:	-14391.
No. Observations:	3004	AIC:	2.890e+04
Df Residuals:	2947	BIC:	2.924e+04
Df Model:	56		
Covariance Type:	nonrobust		

However, the R-squared marginally increased to 0.888 and with a very laggard decrease in AIC of 28,900. This suggests that the model may have reached a limit where further variables may overfit the graph. To test the impact of owning multiple guns, we retrained the model by removing the permit_recheck. In fact, if we compare the model that contains "multiple" with the one that only contains the states and types of guns, we notice that the R-squared and AIC values are the same, suggesting possible collinearity. Therefore, we decided to drop the "multiple variable" and elect to keep the model:

total_incidents~total_nics + C(state) + handgun + long_gun + other + permit_recheck

5. DASHBOARDS

5.1 Use Cases

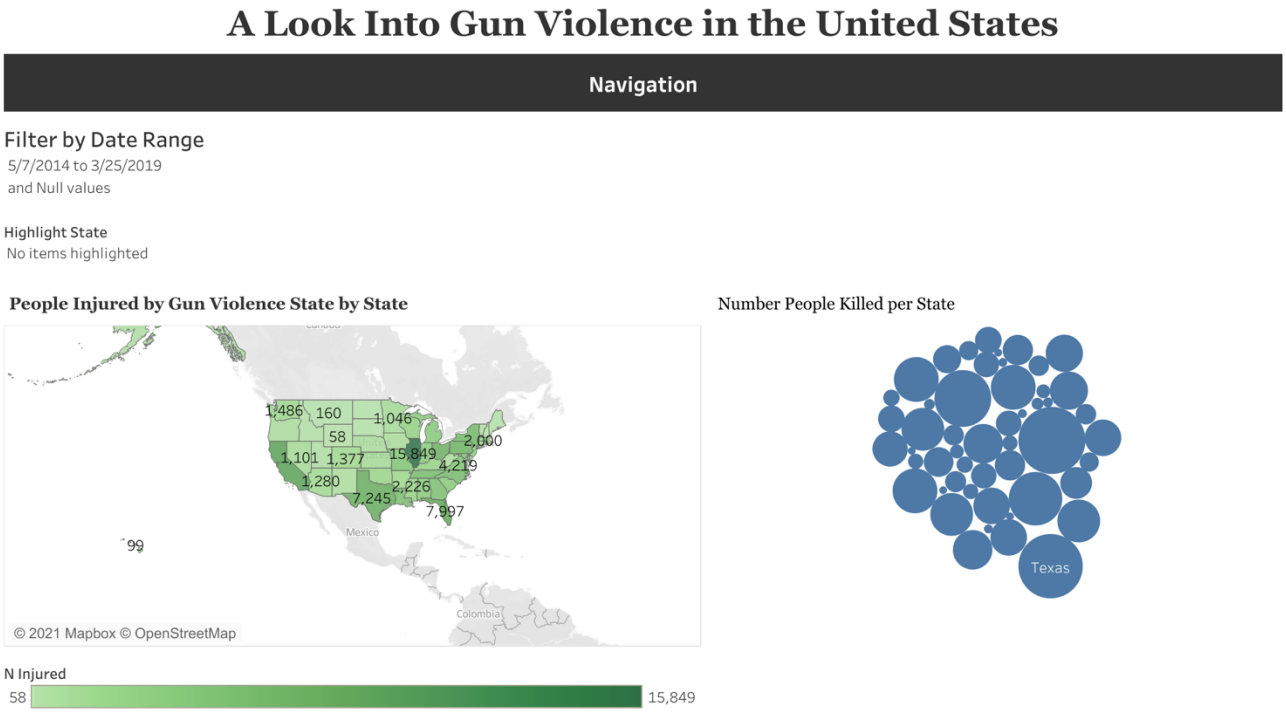
Our goal for the front-end design is to give users the ability to see our data analysis and findings as a dashboard. We wanted to walk users through our process and how we approached not only our exploratory data analysis but our conclusions as well. Too often we see the end results of a data analysis and not the process it took to get to those results. When exploring the topic of gun violence in America, we found it easy to find conclusions and not the process it took data scientist to get there, we would like to change that. Giving end users the ability to walk through what it took to understand the data can help people understand the reality we are facing. In an age of fake news, we find it important for people to understand the importance of facts.

5.2 Data Engineering

We have decided to use Github pages to host our website. The advantages of Github Pages is that it is a free platform that will allow us to host our static website. Our website will not use a backend and will instead use Tableau Public to create our dashboard. This gives us the ability to create a lightweight application with less maintenance and the ability to host using

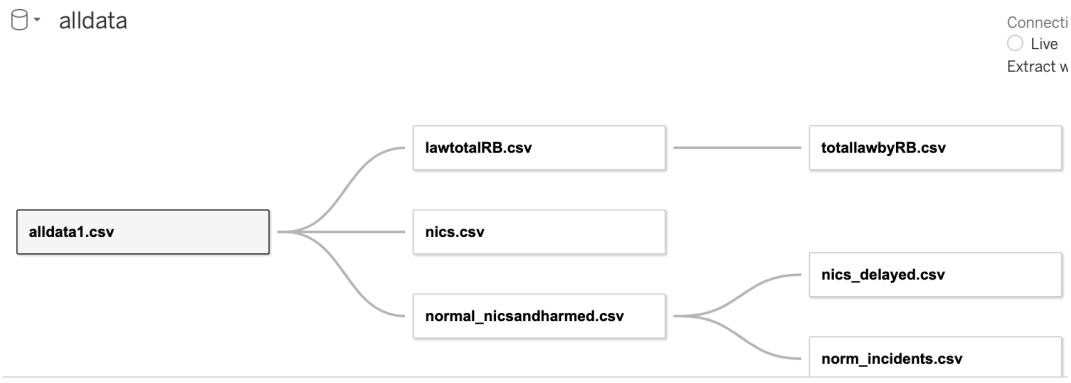
Github Pages. Another advantage of using Github Pages is that if we would like we have the flexibility to use React Framework to add functionality and connectivity to the website. Github Pages doesn't support a backend, but because we are using Tableau Public we wouldn't need one.

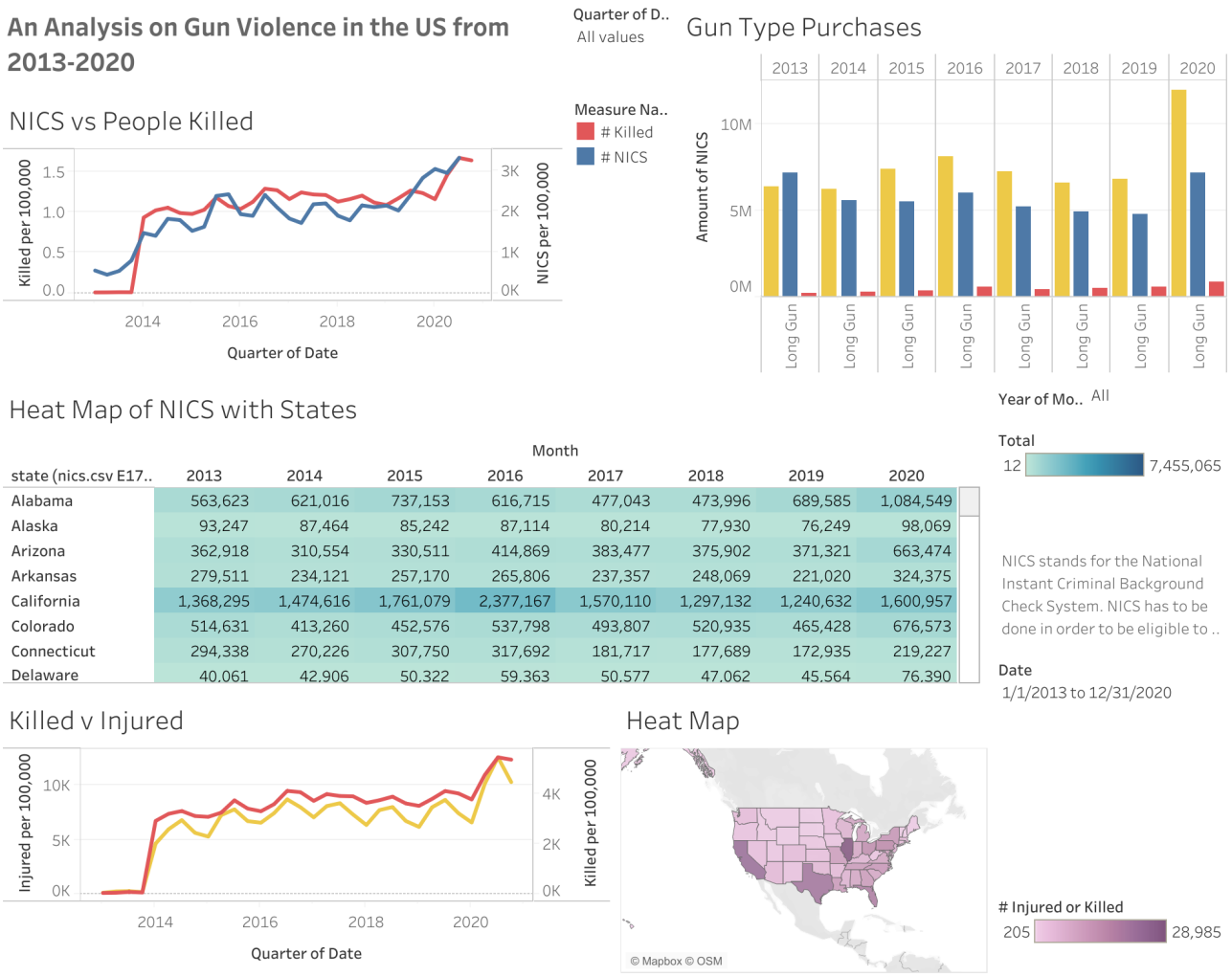
Group Project Github Repo: https://github.com/Steph0088/Team97_DS4A/tree/main



ᵛ Gun Violence in the US Team 97

The image below are the various relationships between the related tables that were used to visualize in Tableau.





[vi]Gun Violence in the US Team 97

6. CONCLUSIONS

6.1 Conclusion

It is important to note that we are using the NICS data as a proxy to gun purchases. NICS does not guarantee that someone purchased a gun but it is a background check that needs to be completed before purchasing a firearm and is valid for only 30 days.

The Gun Violence Archives data was not user-friendly and required us to webscrape. This limited our ability collected the more detailed information they have (like gender, age, etc). We had access to the detailed data only between 2013-2018. We were able to collect the data from 2019-2020 but it did not include all of the demographic information.

The data we collected from Washington D.C. has to be treated as an anomaly in our case. It would be inaccurate to compare a metro to a whole state. If we wanted to include DC in the analysis it would make the most sense to compare only major cities. There's a lot of ways to aggregate the data and trying to incorporate all 50 states throughout the whole process would not have been feasible.

From the EDA we learned that, contrary to popular belief, gun violence incidents are not mostly mass shootings, instead most incidents usually lead to one person injured or dead, and that mass shooting events are very rare.

6.2 Future Work

From the data we observed there are a few next steps we can do to possibly decrease the amount of deaths due to gun violence. The correlation between the number of people killed per 100,000 and NICS done per 100,000 was ~0.5. The NICS data was pulled forward by 3 months as it had a closer correlation to the number of people killed per 100,000. This means that whenever there was an increase in NICS performed, around 3 months later there would be a similar increase in people killed due to gun violence. An increase in gun control policy could deter the amount of people affected by gun violence. An example of a new policy could be having an initial background check that you have to pass before you are even eligible to do the NICS background check. This would have the potential to decrease the number of NICS being done if people fail to pass the initial check which could lead to a decrease in deaths due to gun violence.

There are several next steps that could be incorporated to create a more in-depth analysis. In order to get more detailed data implement/integrate an API that collects all of the detailed information from the Gun Violence aArchives. There was a lot of data that was joined together and there would be more in-depth analysis on at all stages. Integrating the more specific census data would allow for more analysis on the demographics including age, race, gender, income.

7. REFERENCES

Group Project Github Repo: https://github.com/Steph0088/Team97_DS4A/tree/main

[i] GitHub - jamesqo/gun-violence-data: A comprehensive, accessible database that contains records of over 260k US gun violence incidents from January 2013 to March 2018.

[ii] https://github.com/BuzzFeedNews/nics-firearm-background-checks

[iii] https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/totals/nst-est2019-01.xlsx

[iv] https://www.programiz.com/python-programming/regex

[v] https://public.tableau.com/app/profile/team.97/viz/GunViolenceinAmerica_16278708512350/Dashboard1

[vi] https://public.tableau.com/app/profile/team.97/viz/GunViolenceintheUSTeam97/Dashboard12?publish=yes