An Analysis of Gun Violence, 2013-2018

Final Project, May 12, 2021

Abstract

Gun violence has come into the limelight as a serious issue within the past 20 years. It has been branded a national epidemic, and is often framed in emotional terms of being an "unpredictable crisis." Nevertheless, there may be clusters and trends that allow us to better characterize gun violence so that we can allocate resources and direct legislation optimally.

In this spirit, we analyze 260K-entry repository of gun violence data from 2013 to 2018 to answer three main questions:

1. Is gun violence clustered or diffuse?
2. Using year-on-year trends and month-on-month seasonality, can we forecast future incidents with a reasonable degree of efficacy?
3. If this seasonality exists, to what degree does geography and climate bring about monthly fluctuations?

We find significant clustering of gun violence incidents within select "high incident" cities, develop a relatively reliable (though limited) Holt-Winters forecast with additive trend and seasonality for predicting future gun violence-related injury, and find that seasonality is present in gun violence data only in non-tropical climates, though the specific nature of this seasonality cannot be fully characterized by a linear or sinusoidal model. Most importantly, we demonstrate methods for quantitatively characterizing the nationwide gun violence crisis, and provide an array of prescriptive solutions.

Motivation

Gun violence remains one of the United States' most pressing social issues, claiming thousands of lives each year; there have been more domestic gun violence deaths since the late 1960s than there have been American war fatalities *ever*, and yearly counts continue to increase despite a wide array of well-intentioned political initiatives. In response to the inefficacy of recent (and not so recent) policy changes, the *New England Journal of Medicine* has recommended a "public health approach" aimed at examining the crisis from a more scientific, data-driven angle that is "pragmatic rather than dogmatic."[1]

While there is an impressive repertoire of research that attests to the efficacy of this strategy, much of it is focused more on action than on anticipation – in other words, it addresses how to mitigate gun violence in high-casualty areas, but not how to predict *where* and *when* these initiatives will be best employed.[2] The literature that does attempt to forecast gun violence often places a disproportionate focus on mass shooting incidents, which, while tragic, do not represent the bulk of American gun deaths. For instance, while the promising 2021 paper *Forecasting the Severity of Mass Public Shootings in the United States*, published in *Quantitative Criminology*, compares distributions of mass shooting events to well-known mathematical models in shaping its predictions, its use of a mere 156 data points raises questions as to how well its findings fit the ~85,000 American gun-related injuries each year.[3] [4] Finally, another (perhaps misplaced) emphasis found within previous work relates to demographics; while there is substantial literature that focuses on the background (race, gender, socioeconomic status, drug habits, etc...) of gun violence perpetrators, broader trends that characterize gun violence on a macro scale – namely the "when and where" of gun violence incidents. This is what our analysis addresses.

Research Questions

Our analysis is based on the publicly-available data set "Gun Violence Data" by James Ko, published on Kaggle.com.[5] This database encompasses over 260,000 entries, each row representing an individual gun violence episode that occurred This set includes over 260,000 entries, each one representing an individual gun violence episode that occurred between January 1, 2013, and December 31, 2018. Every record has 29 corresponding fields, including date of incident, number of individuals injured, number of individuals killed, longitude & latitude, city or county, congressional district, and number of guns involved. We

[1] https://www.nejm.org/doi/10.1056/NEJMsb1302631
[2] https://www.annualreviews.org/doi/abs/10.1146/annurev-publhealth-031914-122509
[3] https://link.springer.com/article/10.1007/s10940-021-09499-5
[4] https://www.aafp.org/about/policies/all/gun-violence.html
[5] https://www.kaggle.com/jameslko/gun-violence-data

decided to exclude the entries from between January 1, 2013 and December 31, 2013 due to concerns related to data quality and completeness.

Using this information, we divided our analysis into three major portions: clustering of gun violence incidents, predictive episode forecasting, and the role of geography in the seasonality of firearm casualties. More formally, our research questions and corresponding data and methods are as follows:

Research Question	Data Used	Methodology	Practical Implication
Is the proportion of monthly gun-related injuries that occur in "high-casualty" areas related to the number of monthly gun-related injuries nationwide?	Gun violence injuries by month, categorized by whether they occur in a "high-casualty" area or not (2014-2018)	Logistic Regression	Is gun violence clustered or dispersed?
What are the results of a predictive Holt-Winters model with seasonality and additive trend in forecasting gun violence injuries in the United States?	Gun violence injuries by day (2014-2018)	Holt-Winters Forecasting	Can gun violence be forecasted?
Do gun violence injury patterns follow a seasonal trend at latitudes with strong climate fluctuation, and at latitudes without strong climate fluctuation?	Gun violence injures by day, separated into two data frames: those that occurred between 19-32°north, and those that occurred north of 32°(2014-2018)	Linear and Sinusoidal Regression	How do seasonality and climate relate to patterns of gun violence?

Question 1: Is Gun Violence Concentrated or Dispersed?

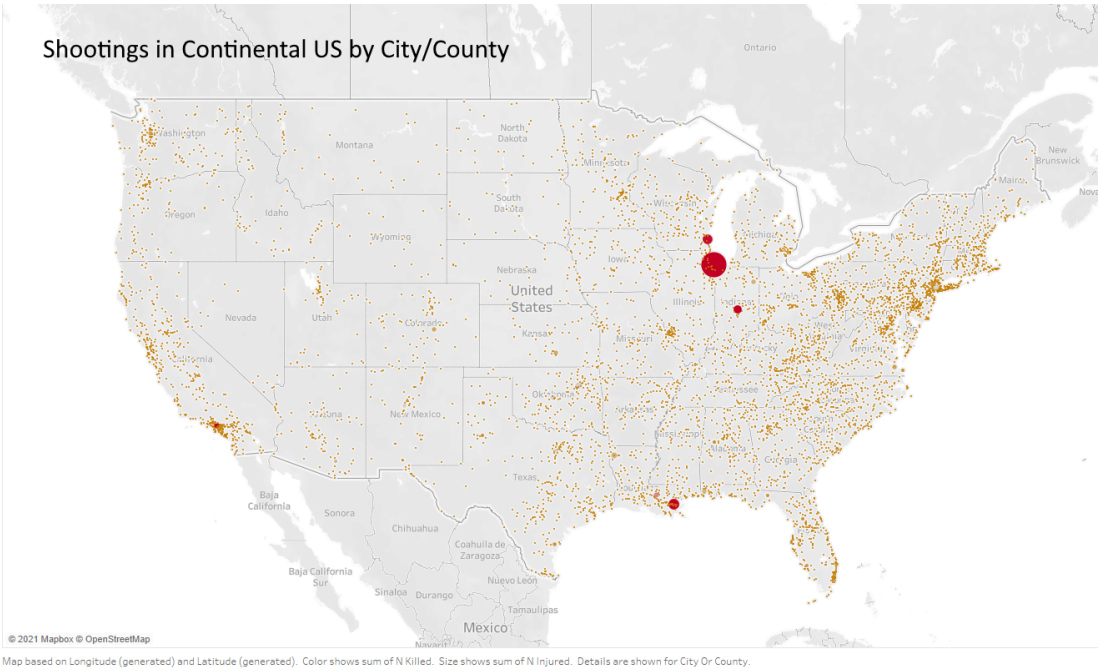


Figure 1: Heatmap of casualties by City/County

Clustering: The Big Picture

As a cursory overview of the dataset will indicate, though shooting incidents continue to victimize thousands of communities throughout all 50 states, the toll of gun violence in the United States is exceptionally uneven. In fact, only nine cities – Chicago, Saint Louis, Baltimore, Philadelphia, Houston, New Orleans, Columbus, Memphis, and Cleveland – represent over 20% of gun violence injuries within the entire dataset. These cities range from 5661 injuries (Chicago) to 669 injuries (Houston).

In economics, the *Pareto Principle* states, "roughly 80% of consequences come from 20% of causes." Upon discovering that approximately 20% of pea plants produced 80% of healthy peapods, Italian economist Vilfredo Pareto found that around 20% of the Italian population owned 80% of Italian land. This idea has diverse implications; research has found that approximately 20% of bugs in computer code are responsible for 80% of system crashes, roughly 80% of consumer complaints are made by 20%

of customers, and around 80% of library book checkouts come from 20% of the collection.[6] At a glance, it appears that far fewer than 20% of communities are responsible for 80% of gun violence incidents, suggesting an "ultra-Pareto" distribution of incidents. This begs the question: If both uniform and Pareto distributions of gun violence throughout the United States prove incorrect, how can we accurately characterize its concentration and clustering?

Logistic Regression Model

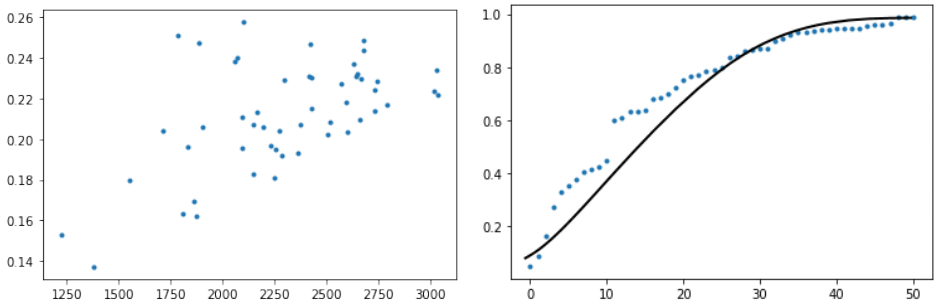
In processing this data, the nine aforementioned municipalities responsible for 20% of gun-related injuries were designated "high-incident" areas. New York City (population ~8.4 million) was broken up into boroughs preceding this evaluation, as its largest borough, Brooklyn, has nearly the same population as the whole of Chicago (~2.7 million). Given New York City's relatively low per-capita gun violence and outlier total population, it would have been misleading to focus on the city as a whole as a "focus cluster" with regards to gun violence. It may be worth mentioning that injuries – rather than deaths or total casualties – were examined, as communities with more overall gun violence display patterns of shootings that are more sporadic than premeditated, and therefore more conducive to injury than death. This can be illustrated by the relatively low "KillProp" (deaths/total casualties) in cities with a higher toll of gun violence (the average "KillProp" throughout the dataset was almost exactly 0.5).

```
casualties.head(11)
```

	city_or_county	Killed	Injured	congressional_district	latitude	longitude	GunsInvolved	Casualties	KillProp
0	Chicago	1186	5661	1	41.7286	-87.6425	6660	6847	0.173215
1	Saint Louis	550	956	1	38.6676	-90.2482	1482	1506	0.365206
2	Baltimore	441	970	7	39.3375	-76.661	2478	1411	0.312544
3	Philadelphia	413	851	2	39.9961	-75.1708	1266	1264	0.326741
4	Houston	549	669	9	29.7201	-95.611	1403	1218	0.450739
5	New Orleans	302	851	2	29.9649	-90.0518	1453	1153	0.261925
6	Columbus	332	720	3	39.9603	-83.0278	1298	1052	0.315589
7	Memphis	279	773	9	35.2045	-89.9872	1320	1052	0.265209
8	Cleveland	295	723	11	41.4592	-81.6133	1407	1018	0.289784
9	Detroit	335	623	13	42.3301	-83.1486	1201	958	0.349687
10	Indianapolis	333	516	7	39.885	-86.1967	1349	849	0.392226

After the data was categorized based on "high-incident" area, it was grouped by month, such that each month had a proportion of total injuries that resulted from "high-incident" areas. If this proportion was more than 0.20, "high-incident" areas were overrepresented in total injuries this month. Months were ranked from 0 (lowest nationwide monthly injuries) to 50 (highest nationwide monthly injuries). A plot of proportion of total injuries from "high-incident" areas on total injuries per month is shown on the left in Figure 2.

As shown above, there was a noticeable positive trend between monthly injuries and proportion of injuries taking place in "high-incident" areas. The most plausible explanation for this may be that exacerbating factors of gun violence nationwide trigger an even *larger* uptick in "high-incident" areas. This data was fit to a logistic regression model below; the trendline is represented by the colored dots, with a logistic fit superimposed for clarity:



Dep. Variable:	HC_prop_binary		**No. Observations:**	51	
Model:	Logit		**Df Residuals:**	49	
Method:	MLE		**Df Model:**	1	
Date:	Mon, 10 May 2021		**Pseudo R-squ.:**	0.2860	
Time:	00:15:19		**Log-Likelihood:**	-21.399	
converged:	True		**LL-Null:**	-29.972	
Covariance Type:	nonrobust		**LLR p-value:**	3.461e-05	
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-7.9102	2.716	-2.912	0.004	-13.234	-2.587
SUM(n_inj)	0.0040	0.001	3.214	0.001	0.002	0.006

Figure 2: Left: Proportion of total injuries from "high-incident" areas vs. total injuries per month; Middle: Logistic trendline; Right: Logistic regression results

As demonstrated above, the regression of monthly overrepresentation of "high-incident" areas among injuries on injury-based severity of month is best fit by a β_0 of -7.9102 and a β_1 of 0.0040, with a McFadden's Pseudo R-Squared (ρ^2) value of 0.2860,

[6] https://www.sciencedirect.com/topics/computer-science/pareto-principle

indicating a strong fit (a common point of confusion is the misinterpretation of McFadden's ρ^2 as R^2; much lower values of ρ^2, as compared to R^2, indicate good fit).[7] In fact, a side-by-side comparison of daily incidents (injuries) over time against the proportion of incidents in "high-incident" areas demonstrates an exceptionally similar trend.

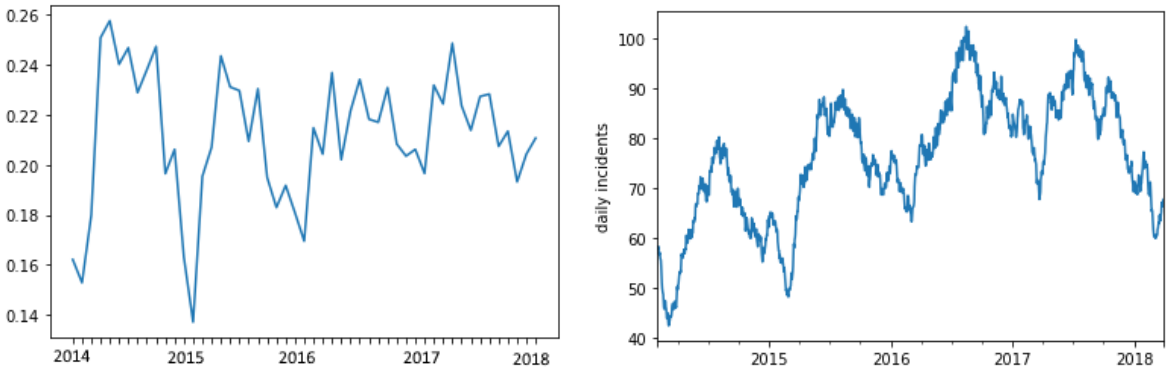


Figure 3: Left: Daily incidents (2014-2018); Right: Projected daily incidents (2018-2022)

Interpretation

Both a raw count of gun violence injuries by community and a logistic model linking clustering of gun violence injuries with severity of month reveal an unmistakable truth: gun violence is not an issue that pervades each community within the United States to the same degree. In other words, it is questionable whether gun violence is empirically a "nationwide epidemic" as pundits and politicians often argue, as opposed to a local "*ultra*-epidemic" in a handful of cities. This finding may lend support to the argument that bulk of staggering national gun violence tallies result most directly from severe, chronic socio-economic issues in a subset of American cities, exacerbated by the presence of firearms.

As such, it is unsurprising that RAND Corporation reports attest to the inefficacy of gun restrictions on gun violence. Their 2020 evidence review found that the evidence is "inconclusive" as to whether bans of sales on assault weapons and licensing and permitting requirements mitigate gun violence, and merely "limited" evidence in support of concealed-carry laws. The evidence behind even seemingly-robust background checks remains only "moderate."[8] Given the extreme degree of clustering among gun violence incidents, policy emphasis on gun restrictions may be misplaced; alternatively, a targeted focus on the social, economic, and cultural causes of gun violence in the specific regions especially plagued by firearm casualty may be more effective. Recognizing and tackling unique, city-specific challenges to safety in nine cities alone, after all, may *substantially* flatten national gun violence "flare-ups", and manage up to 20% of the nationwide problem.

Question Two: Can Gun Violence Be Forecasted?

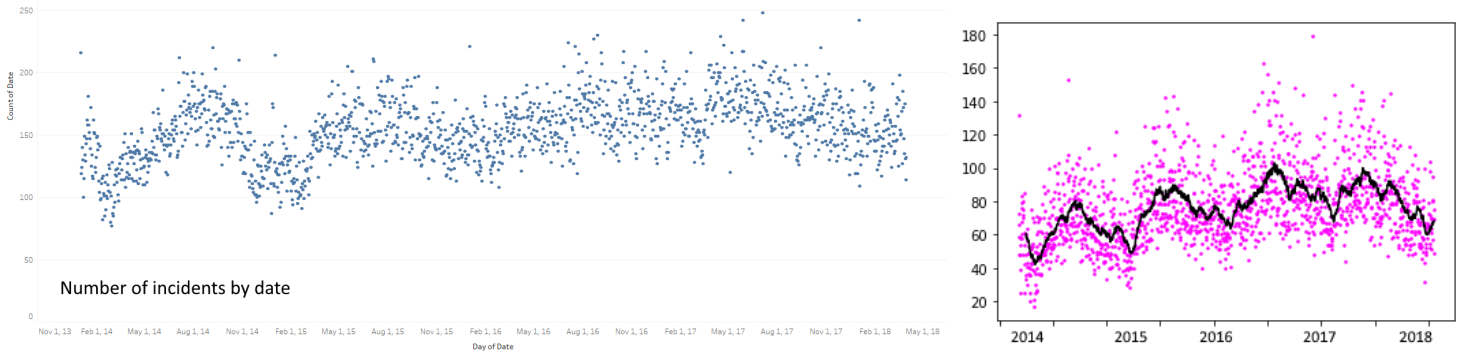


Figure 4: Left: Incidents by date; Right: Shooting injuries, overlaid with rolling average

Predictability: The Big Picture

While gun violence may be more of an endemic than epidemic, the epidemiology analogy holds in another, more surprising way. The groundbreaking 2006 Harvard Engineering paper *Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago* argued that gun violence spreads like a disease, from "carrier" to "carrier". For instance, a member

[7] https://sites.ualberta.ca/~lkgray/uploads/7/3/6/2/7362679/18b_-_logisticregression.pdf
[8] https://www.rand.org/research/gun-policy/key-findings/what-science-tells-us-about-the-effects-of-gun-policies.html

of a specific criminal organization may shoot a rival, who may, in turn, strike another individual in the crossfire during a retaliation attempt. This multi-actor violent outbreak may spread through a given community, and ultimately, a city. This phenomenon, dubbed "social contagion", was used to create an early predictive model of gun violence in Chicago.[9]

Nevertheless, "social contagion" is more pertinent to intra-city spread than it is to categorizing a national phenomenon that is, at its core, the cumulative effect of several intra-city phenomena. The benefits to predicting trends in gun violence, however, are numerous; from allocating deployment of law enforcement and investigation personnel, to mitigating trends from an educational and public health standpoint before their full brunt is bore, to passing gun-related legislation. We attempt to do this on a national scale by examining season-on-season and year-on-year trends, creating a broad predictive forecast through Holt-Winters modelling.

Holt-Winters Seasonal Forecast

A plot of gun-related injuries over time demonstrates two noticeable trends; one subtle, the other prominent. More pronounced is the periodic pattern found within the data. In winter months, firearm injuries are lower, increasing throughout the spring before peaking during the summer, and calming back down again during the fall. In the background, the number of gun-related injuries per year has slowly increased over time. These phenomena are demonstrated by the plot of daily firearm injuries in Figure 4, with each colored dot representing shooting injuries on a given day, and the black trendline indicating a rolling average.

Ideally, patterns within the data render future events more predictable. Using a Holt-Winters forecasting model with additive trend (based on previous year-on-year increase) and seasonality (based on previous month-on-month fluctuation), we developed a predictive forecast for future gun violence injuries in the United States. The Holt-Winters forecast, juxtaposed against data from 2014 through 2018, is below:

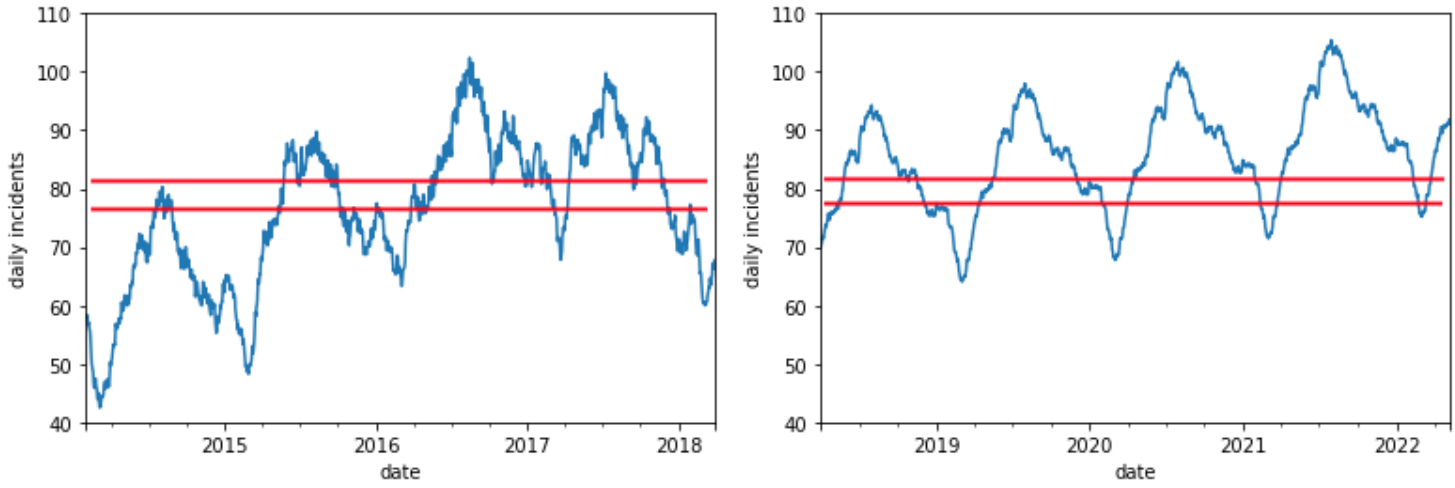


Figure 5: Left: Data from 2014 to 2018; Right: Holt-Winters forecast

Following the current trend, the projection suggests that daily gun violence injuries would steadily increase into the near future, with upwards of 100 incidents per day becoming a new normal during summer months, following profound increases after stark drop-offs that characterize winter months.

While research has shown that the COVID-19 pandemic has increased firearm injury to an extent that would have been difficult to anticipate based on previous data, the gradual relaxation of lockdown measures and the return to normalcy may allow for gun violence related-data to revert to levels consistent with previous trend.[10] Nevertheless, in 2019, the most recent year before the COVID-19 pandemic had transformed daily life in the United States, the National Gun Violence Archive reports 29,501 gun-related injuries in the United States, up from 28,333 in 2018. These translate into 80.824 and 77.624 daily events respectively (lines drawn on graphic above), which appear relatively consistent with our forecasting. In 2020, a total of 39,446 gun violence-related injuries (108.071 per day) stood out as a clear outlier; no yearly firearm-related injury count has *ever* exceeded 31,300.[11] As of 2020, media reporting on gun violence continues to suggest a seasonal trend, with summer serving as a time of noteworthy violence.[12]

[9] https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2594804
[10] https://www.journalacs.org/article/S1072-7515(20)32413-3/fulltext
[11] https://www.gunviolencearchive.org
[12] https://www.opb.org/news/article/gun-violence-statistics-covid-19-pandemic/

Interpretation

While gun violence is often perceived as unpredictable or even "random", broad trends on both a micro level (social contagion) and macro level (additive increase with seasonality) can help policymakers more effectively characterize and address the issue. Though "black swan" events, such as the COVID-19 pandemic, may make it challenging to perfectly foresee future changes, even the broadest of indicators may aid policymakers, law enforcement agencies, and community members in preparing for (and preventing) an onslaught of violence.

Although some general trends, including broad notions of summer flare-ups and a slight but often-interrupted yearly increase, are well-known, more comprehensive estimates of specific injury counts and long-term likelihoods can provide relevant authorities with a more complete idea of circumstances, which can inform mitigation strategy.

Perhaps the most important take-away from this result is that, in the long run, gun violence injury is poised to increase steadily, and summer flare-ups continue to persist perennially. While the quantitative forecasting results can be useful in broad-based policy formation, the most substantial result of this model may be that it underscores the necessity of an equilibrium-disturbing shift in gun violence-related policy.

Question Three: How do Seasonality and Climate Relate to Patterns of Gun Violence?

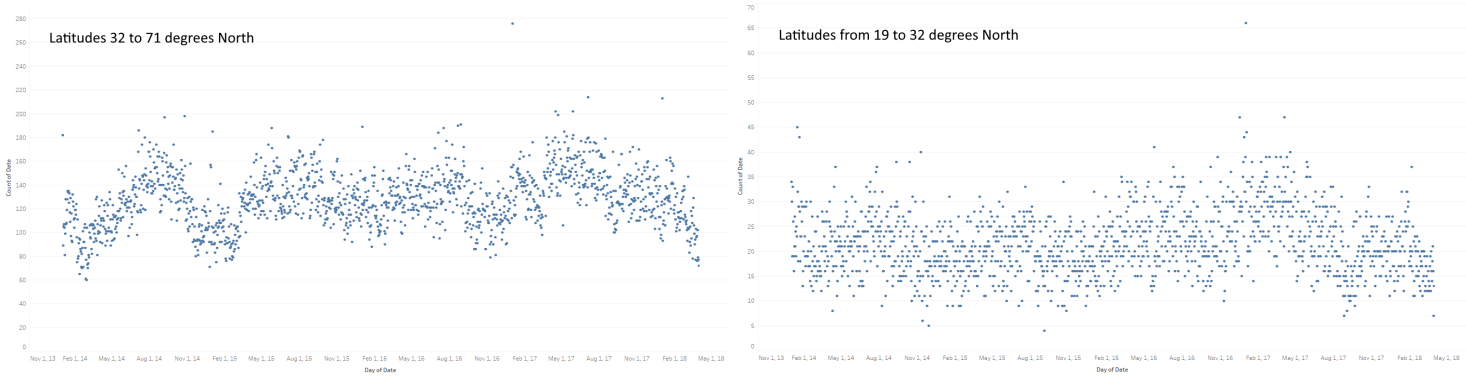


Figure 6: Left: Latitudes between 19 and 32 degrees North; Right: Latitudes between 32 and 71 degrees North

Climate and Seasonality: The Big Picture

While the Holt-Winters forecast may have clarified that gun violence is seasonal, current understanding of the topic remains cursory and broad. A *New York Times* exploration in 2017 found a connection between heat and gun violence burden – specifically, that gun violence and temperature in Philadelphia, Detroit, and Baltimore are positively correlated, though this correlation does not exist in the somewhat seasonless climate of San Francisco.[13]

Nevertheless, it may be surprising that temperature alone would have such an outsize effect on gun violence. If this were the case, it would evade intuition as to why Chicago, Saint Louis, Baltimore, and Philadelphia, cities with cold winters, would continue to post the highest numbers of gun-related injuries in the United States. Other plausible explanations for seasonal spike may include school recesses, changes in work and leisure schedules, migration patterns, or shifts in organized crime. Intuitively, temperature should have some effect, but can seasonality in gun-related injury *all* be due to climate?

This information may prove more pertinent than it appears at first glance. While it is clearly not possible for policymakers to transform the climate of New York into that of Miami, if seasonality is at least partially caused by items other than climate, "summer spike" mitigation may prove to be a more actionable goal than if climate is the sole culprit for the pattern. In keeping consistent with the *New York Times* analysis, we decided to examine overall casualties in this analysis, rather than nonfatal injuries alone.

Linear and Sinusoidal Regression

In order to best evaluate the relationship between climate and seasonality, communities within the dataset were divided based on whether they were north or south of 32°latitude, which runs through Arizona, New Mexico, Texas, Louisiana, Mississippi, Alabama, and Georgia. This parallel forms a natural dividing line between US cities considered to be tropical and relatively seasonless, including New Orleans, Houston, Miami, and Jacksonville, and cities with greater temperature fluctuation, such as Atlanta, New York, and Washington D.C. A caveat to this method is that cities on the Pacific coast often display muted seasonality despite being much further north. By coincidence, however, few Pacific Northwest cities exhibit especially high levels of gun violence (Oakland, California being the main exception).

[13] https://www.nytimes.com/2018/09/21/upshot/a-rise-in-murder-lets-talk-about-the-weather.html

We ran a linear regression on firearm casualties and month for US cities north (left) and south (right) of the 32nd parallel, and found no meaningful relationship whatsoever between date and casualty count:

OLS Regression Results

Dep. Variable:	Count of Date	**R-squared:**	0.049
Model:	OLS	**Adj. R-squared:**	0.048
Method:	Least Squares	**F-statistic:**	79.78
Date:	Wed, 12 May 2021	**Prob (F-statistic):**	1.16e-18
Time:	00:52:49	**Log-Likelihood:**	-7030.5
No. Observations:	1548	**AIC:**	1.406e+04
Df Residuals:	1546	**BIC:**	1.408e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	118.9897	1.154	103.068	0.000	116.725	121.254
num	0.0115	0.001	8.932	0.000	0.009	0.014

OLS Regression Results

Dep. Variable:	Count of Date	**R-squared:**	0.005
Model:	OLS	**Adj. R-squared:**	0.005
Method:	Least Squares	**F-statistic:**	8.483
Date:	Wed, 12 May 2021	**Prob (F-statistic):**	0.00364
Time:	00:52:50	**Log-Likelihood:**	-5064.6
No. Observations:	1548	**AIC:**	1.013e+04
Df Residuals:	1546	**BIC:**	1.014e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	20.7973	0.324	64.144	0.000	20.161	21.433
num	0.0011	0.000	2.912	0.004	0.000	0.002

Figure 7: Left: Linear regression on northern cities; Right: Linear regression on southern cities

It happens that slight increase over time is more observable in the northern latitudes (perhaps by chance), though the fit is clearly nonlinear (below, left), while there is no clear trend whatsoever in the southern latitude data (below, right). Interestingly, the overall combined data does show signs of a moderate linear increase (below, center). Hypothesized fit lines are superimposed for clarity, on these regressions of daily casualties on days after January 1, 2014:

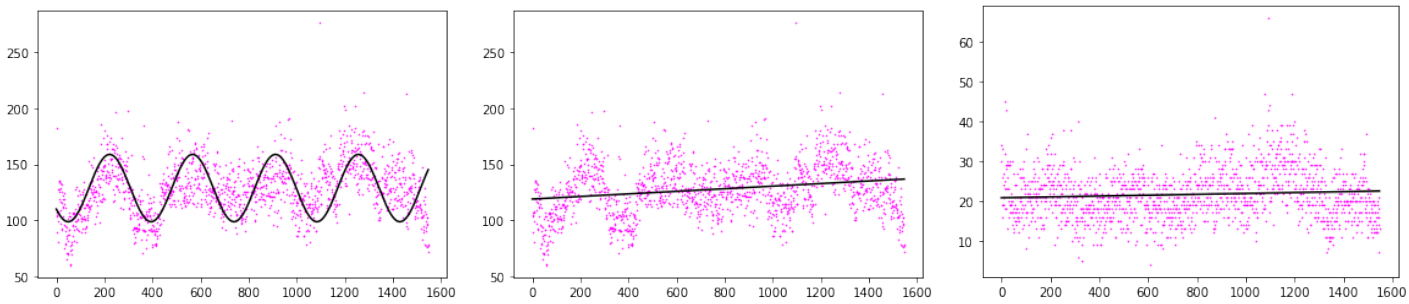


Figure 8: Hypothesized fit lines for cold and warm areas

A quick examination of the northern latitude data suggests a potential sinusoidal pattern. Using the following SciPy test function, which optimizes parameters for a sinusoidal fit given starting parameter estimates, we obtained a sin curve approximate for the northern latitude data:

```
def test_func(x, a, b, c, d):
    return -a * np.sin(b+d*x) + c
params, params_covariance = scipy.optimize.curve_fit(test_func, cold['num'],
                            cold['Count of Date'], p0=[100, 0, 127, 0.021])
```

When optimized, we received the function

$$y = -1.488 \sin(0.674 + 0.01821x) + 0.012877,$$

where y represents daily casualties, and x represents the number of days after January 1, 2014.

As a means of evaluating the sinusoidal model, we performed an average sum of squares comparison between the fit and the data points. We found that for the linear model, the result was 515.6397, whereas for the sinusoidal model, the result was a comparatively smaller 432.1545. Given that the linear model was a poor fit, however, we hesitate to conclude that the northern latitude data is particularly sinusoidal.

A further look at the northern latitude data, shown below, demonstrates a rather inconsistent trend that cannot be characterized well by one particular sinusoidal function. Amplitude and period fluctuate each cycle, without a clear trend. It is worthwhile to consider, however, that these amplitude and period fluctuations may not be completely random. As the "real world" meaning of amplitude is the size of the "summer flare up" in gun violence, and the "real world" meaning of the period is the length of each year's "summer season", there may be related, non-random variables that dictate the behavior of each yearly cycle on this piecewise, semi-sinusoidal curve. An example is illustrated below:

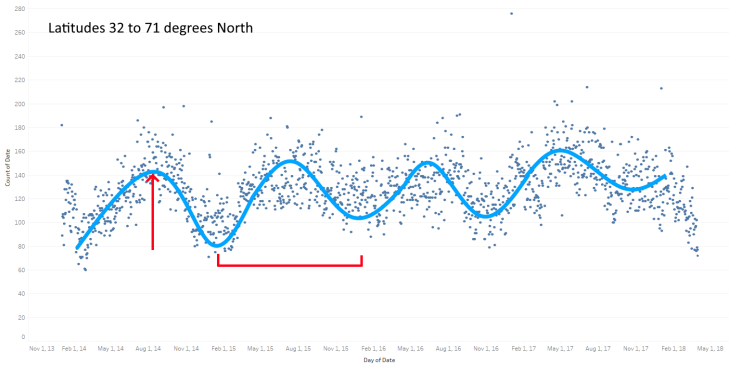


Figure 9: Example curve; amplitude and period shown in red

Interpretation

While it would be simplest for policymakers to interpret gun violence in light of a well-defined sinusoidal (or linear) function, no such function exists. One striking finding, however, is the stark contrast between the gun casualty landscape in tropical and temperate climates; seasonality appears hardly present at all south of the 32nd parallel, while north of the 32nd parallel, seasonality is a significant determinant of gun casualty count. Nevertheless, the length of each year's "summer", and the extent to which seasonality factors into gun violence, varies from year to year. A nonspecific explanation for this would be to attribute these differences to "broader national trends", with more specific rationales related to yearly changes in drug markets or organized crime, specific policy initiatives, or even weather-related events worth considering.

Though it may be tempting to chalk the lack of an obvious model-able fit up to "randomness", this is far from likely to be the case. The main prescriptive conclusion from this analysis is a call to action in support of further research into why seasonality, as it relates to gun violence casualty, behaves as it does. Understanding these factors may provide insight into both generic and year-specific causes of the "summer spike". Ideally, this can point policymakers and community members in a productive direction.

Conclusion

Gun violence in the United States remains an astoundingly complex issue, though not one that is impossible to quantify, model, and characterize. Through a detailed analysis of clustering (the "where"), a long-term forecast with seasonality (the "what"), and an investigation of determinants of seasonal trend (the "why"), upticks in American gun violence can be better understood. Our core results, as well as suggestions for future investigation, are summarized below.

Topic	Research Question	Key Finding	Future Investigation
Clustering	Is gun violence clustered or dispersed?	Gun violence is *highly* clustered, and nationwide flare-ups are linked with a higher proportion of firearm injuries in "high-casualty" areas	An examination of community-specific solutions for gun violence mitigation in particularly-affected areas
Forecasting	Can gun violence be forecasted?	Broad trends in gun violence-related injury can be roughly forecasted through a model that incorporates seasonality, though this model may be derailed by unexpected events (e.g. COVID-19)	A model that links micro-level forecasting (social contagion networks) to macro-level forecasting (Holt-Winters Method) to comprehensively examine gun violence nationwide
Seasonality	How do seasonality and climate relate to patterns of gun violence?	Seasonality in gun violence casualties exists at temperate latitudes, though cannot be easily characterized through a linear or sinusoidal model	An investigation into predictors of "summer spike" length and magnitude, which appear inconsistent between years

■

Appendix: Python code

Note: This DOES NOT count towards the page limit (see Piazza post @1309)

Logistic Regression

```
# casualties data
casualties_df = pd.read_csv("casualties.csv")
casualties_df['high casualty'] = [1 if i > 900 else 0 for i in casualties_df['Casualties']]

# high-casualty data
hc = pd.read_csv('hc_clusters.csv')

# comparing plots of high-incident areas vs. total by date
# see Figure 3 above
plt.plot(hc['month'], hc['HC_prop'])
plt.plot(hc['month'],hc['SUM(n_inj)'])

# looking at proportions of total injuries from high-incident vs. monthly injuries
# see Figure 2: left
plt.plot(hc['SUM(n_inj)'], hc['HC_prop'], '.')

# logit regression -- see Figure 2: right
y = hc['HC_prop']
x = hc['SUM(n_inj)']
x = sm.add_constant(x)
model = sm.OLS(y, x).fit()
model.summary()

# getting general trendline for logit regression -- see Figure 2: center
a_sorted = [1 if i > 0.2 else 0 for i in hc_sorted['HC_prop']]
b_sorted = sm.add_constant(hc_sorted['SUM(n_inj)'])
model = sm.Logit(a.astype('float64'), b.astype('float64')).fit()
q_sorted = model.predict()
q.sort()
plt.plot(q,'.')
```

Holt-Winters

Rolling average of incidents

```
# read in data and only take relevant columns
gunviolence = pd.read_csv("gun-violence-data_01-2013_03-2018.csv")
gunviolence = gunviolence[['incident_id', 'date', 'state', 'city_or_county', 'n_killed', 'n_injured']]

# daily incident totals
incidents_by_date = gunviolence.groupby(['date']).sum().drop(columns=['incident_id'])
incidents_by_date.index = pd.DatetimeIndex(incidents_by_date.index).to_period('D').drop(columns=['n_killed'])

# remove NULL outlier that was causing issues
incidents_by_date = incidents_by_date.drop(index = incidents_by_date.index[list(range(177))])
incidents_by_date.reset_index(level=0, inplace=True)

# rolling averages of daily incident totals -- see Figure 4: right
x = incidents_by_date['date'].astype(int)
y = incidents_by_date['n_injured']
plt.plot(x,y, '.', markersize=3, color='magenta')
rolling_df = incidents_by_date.rolling(30, on='date').mean()
plt.plot(rolling_df['date'].astype(int),rolling_df['n_injured'], markersize=3, color='black')
```

Holt-Winters seasonal forecasting

```
# forecasting -- see Figure 5: right
a = ExponentialSmoothing(rolling_df, trend = 'add', seasonal = 'add', seasonal_periods = 365).fit()
fore1 = a.forecast(1500)
```

```
fore1.plot(xlabel = 'date', ylabel = 'daily incidents', ylim=(40,110))

# plotting rolling average for use alongside forecast -- see Figure 5: left
plt.plot(rolling_df['date'].astype(int),rolling_df['n_injured'], markersize=3, ylim=(40,110))
```

Sinusoidal Regression

Linear regression by latitude areas

```
# reading in data
cold = pd.read_csv('cold_latitudes.csv')
warm = pd.read_csv('warm_latitudes.csv')

# see Figure 7: left
model = sm.OLS(cold['Count of Date'], sm.add_constant(cold['num'])).fit()
model.summary()

# see Figure 7: right
model = sm.OLS(warm['Count of Date'], sm.add_constant(warm['num'])).fit()
model.summary()

# plotting the actual linear prediction from the OLS results
# see Figure 8: center
plt.plot(cold['num'], cold['Count of Date'], '.', markersize=1, color = 'magenta')
plt.plot(cold['num'], [0.0115*i + 118.9897 for i in range(1548)], color = 'black')

# see Figure 8: right
plt.plot(warm['num'], warm['Count of Date'], '.', markersize=1, color = 'magenta')
plt.plot(warm['num'], [0.0011*i + 20.7973 for i in range(1548)], color = 'black')
```

Sinusoidal regression

```
# sine fit for cold data
# p0 is our initial guess for parameters
def test_func(x, a, b, c, d):
    return -a * np.sin(b+d*x) + c
params, params_covariance = scipy.optimize.curve_fit(test_func, cold['num'],
                                                     cold['Count of Date'],
                                                     p0=[100, 0, 127, 0.021])

# used to generate potential sine plot
# see Figure 8: left
plt.plot(cold['num'], cold['Count of Date'], '.', markersize=1, color = 'magenta')
plt.plot(cold['num'], test_func(cold['num'], params[0], params[1], params[2], params[3]), color = 'black')

# arbitrary metric for determining "goodness of fit" b/w linear and sinusoidal:

# difference -- average the squares of the differences between the # points in the data and in the graph

# linear
print('linear:', (sum([(cold['Count of Date'][i] - (0.0115*i + 118.9897))**2 for i in range(1548)]))/1548)

# sinusoidal
print('sinusoidal:', (sum([(cold['Count of Date'][i] - (-params[0] * np.cos(params[1]+params[3]*i) +
      params[2]))**2 for i in range(1548)]))/1548)

# results:
#   linear:         515.6397343421314
#   sinusoidal:     432.1545271462703
```