How Can Societal Phenomena During Specific Eras Affect Prediction in Classification Models?

Carissa Chen, Iliana Vasslides, Kayla Kim, Ryan Ermovick, Yuthi Madireddy

Abstract

This study focuses on analyzing how large-scale national or global phenomena can affect the accuracy and method of representing statistical data and predicting results using machine learning models. While there have been many global societal phenomena over the decades, the most recent widely influential event, COVID-19, was chosen to compare and analyze the effects on calculations and predictions using machine learning models. To explore this, crime data was compiled through different sources, including the FBI, containing murders throughout the United States from 1976 to 2023. For this analysis question, the year range was narrowed to start from 2016. It contained data about what state and city the crime took place in, what the data source was, when it happened, the nature of the homicide, including the situation, weapon & victim count, victim's age, race, sex, and ethnicity, as well as the offender's age, sex, and race.

Using this data set, two decision trees were created using factors of year, offender age, month, state, offender race, and situation to predict the victim's race for murder and non-negligent manslaughter. One model was trained on pre-COVID data, and the other was trained on post-COVID data. After the models were trained, they would then predict the victim's race. Several comparisons were made, comparing White vs. Black victims, Asian vs. White victims, Asian vs. Black victims, and Asian vs. Native American & Alaskan Native victims. Both decision trees made the most incorrect predictions when compared to White victims, no matter which combination, likely because a large proportion of victims are White among all data

observations. The dataset for both pre- and post-COVID was split into training and testing datasets. For the prediction, both decision tree models were tested for accuracy on the post-COVID testing data. It was found that the post-COVID model performed slightly better in regards to predicting Asian victims.

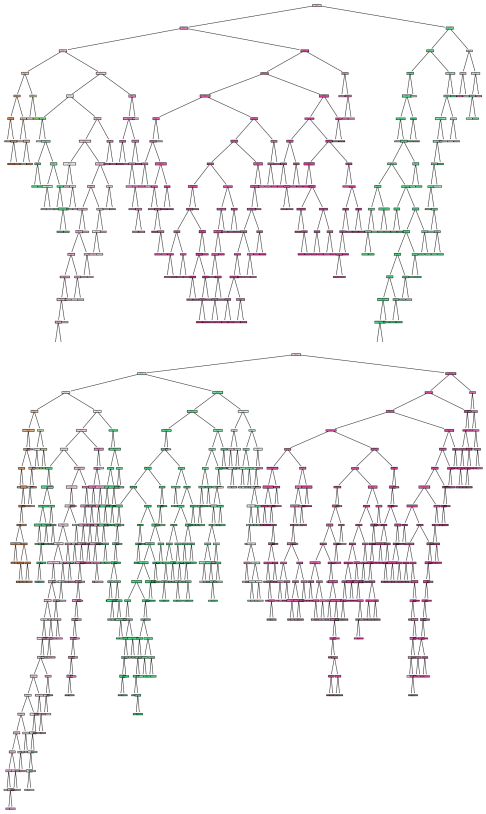
Another set of graphs was created comparing COVID-19 cases with the crime data set, starting from 2020-2023, filtering for crimes against Asian individuals. The results only had data in California, Arizona, Alaska, Alabama, and Arkansas. Throughout these five states, California seemed to have some clear trends in cases and crimes against Asians. Specifically, when there was a spike in COVID-19 cases, there would usually be a spike in crimes against Asians 1-2 months following. These factors may be correlated with each other, but without more testing, it is not possible to state as such.

Introduction

With the recent rise of global events, it was decided to focus on the COVID timeframe and the societal aftereffects it brought about, specifically levels of possible xenophobia, which was measured through crime incidents. During the rise of COVID, there had been a massive change in global society as people navigated the rapid changes to safety, health, and personal well-being. In America, a troubling shift occurred where there was an observed increase in racially motivated hate crimes throughout the country, particularly towards Asians. In fact, BBC's 2021 article on hate crimes rising against Asian Americans states, “Advocates and activists say these are hate crimes, and often linked to rhetoric that blames Asian people for the spread of Covid-19” (“Covid 'hate Crimes' against Asian Americans on Rise”). With this context, it’s suggested that a major cause behind this shift was the idea that some Asians caused the virus or brought it to the US from foreign lands, causing an urge for violence and discrimination. This study aims to examine patterns in these violent crimes, specifically murder and non-negligent

manslaughter, to see if the pandemic influenced these changes in crime frequency and to what degree.

Using crime data compiled from various sources like the FBI, this study investigates homicide incidents from 2016- 2023 using variables of the victim's age, race, location, offender's situation & race, and time of incident to create decision trees and predict the victim's race.



Two machine learning models were developed, splitting the range into pre-COVID, which is 2016-2019, and post-COVID, which is 2020-2023. The models used in this research were Decision Trees. Decision Trees are useful models when working with categorical data because the tree flows top-down through different nodes, or "decision", for each data value and checks if certain criteria are satisfied or not at each split. Then, the tree uses the resulting pathway destination as a prediction for what variable category a certain data point belongs to. This type of model was ideal for predicting the race of a victim as the dataset had many binary variables after one-hot encoding. To train these models, the minimum samples per leaf parameter was tested. Above are two images of the pre-COVID and post-COVID model decision trees respectively.

For both models testing on the post-COVID test data, the victim's predicted race was compared to what the victim's actual race was, utilizing these values in a cross table to display the false positive, false negative, true negative, and true positive values. The race comparisons made were Black vs. White, Asian vs. White, Asian vs. American Indian or Alaskan Native, and Asian vs. Native Hawaiian or Pacific Islander. In addition to the cross-table, the percentage ratio

of each race over all victims will be calculated as a frame of comparison for pre-COVID and post-COVID. Pre-COVID, the ratio of White victims to all victims was the highest compared to all other races. After COVID, the ratio of Asian victims to all victims increased considerably. As a result, the decision tree most often predicted these two categories for each model, resulting in numerous inaccuracies. The predicted number of White victims decreased post-COVID, just like the actual number of White victims, while the number of predicted Black victims also decreased, following the same trend as the actual recorded number of Black victims. The accuracy of predicting an Asian victim also rose when using the post-COVID model.

To compare this crime data with COVID cases from 2020-2023, four time series plots will be generated, one for each year. Across both datasets, the only states that had available data to compare were Alaska, Alabama, Arkansas, Arizona, and California. For California, there were clear spikes in both crimes against Asians and COVID-19 cases. The interesting matter with these is that they lined up with each other throughout the three-year time frame. As this observation has not been thoroughly studied, it cannot be concluded that these graphs have definite relationships.

Data

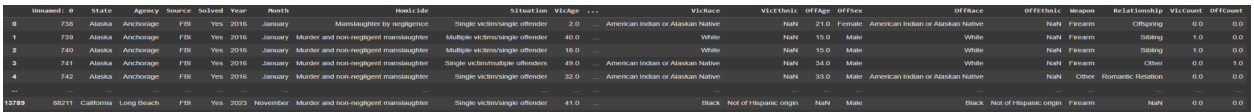
Crime Dataset

The dataset is about murders throughout the U.S. from the Murder Accountability Project, which is a free, publicly accessible website. The dataset spans from 1976 to 2023, but for this project, the dataset will be narrowed down to start from 2016. In the original dataset, some columns were dropped due to them not being relevant, such as the `file data`. The columns in the current data set are:

- `State`: The state of the Crime

- `Agency`: The city of the Crime
- `Source`: Where this data came from (i.e., the FBI)
- `Solved`: Was the case solved?
- `Year`: Year of Crime
- `Month`: Month of Crime
- `Homicide`: Type of Crime
- `Situation`: The Situation in which the crime occurred
- `VicAge`: Age of Victim
- `VicSex`: Sex of Victim
- `VicRace`: Race of Victim
- `VicEthnic`: Ethnicity of Victim
- `OffAge`: Age of Offender
- `OffSex`: Sex of Offender
- `OffRace`: Race of Offender
- `OffEthnic`: Ethnicity of Offender
- `Weapon`: Weapon Used
- `Relationship`: The Relationship between the Victim and the Offender
- `VicCount`: Number of Victims
- `OffCount`: Number of Offenders

Some analysis questions that this dataset could be used to analyze include how COVID years have affected crime rates (including who's doing the crime and where it's happening), if certain places are more inclined to use a certain type of weapon, states with the highest crime frequencies, season with highest crime rates and which months, whether or not crime rates have decreased since the start of data collection, most common types of relationships that incite violence, and more. Many questions can be explored using this dataset, but for this research, it was narrowed down to the current headline topic.



	Unnamed: 0	State	Agency	Source	Solved	Year	Month	Homicide	Situation	VicAge	...	VicRace	VicEthnic	OffAge	OffSex	OffRace	OffEthnic	Weapon	Relationship	VicCount	OffCount
0	738	Alaska	Anchorage	FBI	Yes	2016	January	Manslaughter by negligence	Single victim/single offender	2.0	...	American Indian or Alaskan Native	NaN	21.0	Female	American Indian or Alaskan Native	NaN	Firearm	Offspring	0.0	0.0
1	739	Alaska	Anchorage	FBI	Yes	2016	January	Murder and non-negligent manslaughter	Multiple victims/single offender	40.0	...	White	NaN	15.0	Male	White	NaN	Firearm	Sibling	1.0	0.0
2	740	Alaska	Anchorage	FBI	Yes	2016	January	Murder and non-negligent manslaughter	Multiple victims/single offender	18.0	...	White	NaN	15.0	Male	White	NaN	Firearm	Sibling	1.0	0.0
3	741	Alaska	Anchorage	FBI	Yes	2016	January	Murder and non-negligent manslaughter	Single victim/multiple offenders	49.0	...	American Indian or Alaskan Native	NaN	34.0	Male	White	NaN	Firearm	Other	0.0	1.0
4	742	Alaska	Anchorage	FBI	Yes	2016	January	Murder and non-negligent manslaughter	Single victim/single offender	32.0	...	American Indian or Alaskan Native	NaN	33.0	Male	American Indian or Alaskan Native	NaN	Other	Romantic Relation	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13789	88211	California	Long Beach	FBI	Yes	2023	November	Murder and non-negligent manslaughter	Single victim/single offender	41.0	...	Black	Not of Hispanic origin	NaN	Male	Black	Not of Hispanic origin	Firearm	NaN	0.0	0.0

Weekly COVID Cases Dataset

This dataset was taken from the CDC spanning from 2020 to 2023. The columns in this dataset that were used are:

- `date_updated`: When this observation was added
- `State`: The state where the crimes took place
- `start_date`: The start of the date range recorded in the observation
- `end_date`: The end of the date range recorded in the observation
- `new_cases`: New cases that came up during this time range
- `new_deaths`: Number of new deaths during this date range

	date_updated	state	start_date	end_date	tot_cases	new_cases	tot_deaths	new_deaths	new_historic_cases	new_historic_deaths
0	01/23/2020	AK	01/16/2020	01/22/2020	0	0	0	0	0	0
1	01/30/2020	AK	01/23/2020	01/29/2020	0	0	0	0	0	0
2	02/06/2020	AK	01/30/2020	02/05/2020	0	0	0	0	0	0
3	02/13/2020	AK	02/06/2020	02/12/2020	0	0	0	0	0	0
4	02/20/2020	AK	02/13/2020	02/19/2020	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
10375	04/13/2023	WY	04/06/2023	04/12/2023	186284	158	2023	3	0	0

	Unnamed: 0	state	month_number	new_cases
0	0	AK	1	0
1	1	AK	2	0
2	2	AK	3	63
3	3	AK	4	292
4	4	AK	5	72
...	...	...	...	...
2455	2455	WY	37	919
2456	2456	WY	38	1068
2457	2457	WY	39	1246
2458	2458	WY	40	544
2459	2459	WY	41	495

Monthly COVID Dataset

This dataset was created from the previously shown weekly COVID-19 cases dataset. The dates were compiled into monthly totals, and totals were dropped while new cases were added together. This new dataset included the total cases for each month in the timeframe, per state. The months are labeled starting from early 2020 and ending in mid-2023, with 1 representing January 2020 and 41 representing May 2023.

Methods

The examined dataset consisted of one specific instance of a crime as an observation that includes six different features relevant to the crime: year, region, offender race, offender age, and victim race, set as the predictor variable. The data was further separated into two categories, with pre-COVID representing crimes that occurred within the years 2016-2019, and

post-COVID being crimes within 2020-2023. Using this dataset, it was observed whether a specific race would see a proportional increase or decrease as victims depending on if the crime occurred pre- or post-COVID (beginning year 2020).

The data was prepared by evaluating and imputing appropriate missing values that may have been missed from the earlier EDA. The main categorical variables, aside from the predictor variable,were one-hot encoded. Variables such as month, situation, offender race, and state were one-hot encoded, while the relationship between numeric variables, such as age, offender sex, offender age, and year, were evaluated using correlation matrices.

The two models implemented, the pre- and post-COVID, were both decision trees, which was a type of supervised learning because the outcomes were labeled in training. The decision trees utilized the same features across the data, with the only difference being the time range or years of crime incidence. The trees were trained through supervised learning on the set features from pre-2020 data (pre-COVID) and post 2020 data (post-COVID). Both models were evaluated using the same post-COVID test set to compare how well pre- and post-COVID training data predict post-COVID outcomes. To measure how successful the model was, there was a comparison made between the predicted victim race in both models against the actual race of the victim in the post-COVID data. This was done with the creation of confusion matrices. Resulting patterns in both matrices were analyzed for any type of correlation.

In the confusion matrix, model performance was assessed through the distribution of true positives, false positives, and false negatives per class, with a focus on underrepresented racial groups. If there was a significant amount of false positives and negatives compared to the true values, this could bring about future and significant discussions in relation to societal phenomena and data.

Prediction and accuracy graphs, along with the confusion matrices, were further analyzed in the results section to determine the type of correlation between the Classification tree model's accuracies and the pre-COVID vs post-COVID condition.

Improvements may include tuning decision tree hyperparameters (e.g., depth, min samples per leaf), engineering additional features, or trying models like Random Forests. To note, the methods implemented here differ slightly from the original plan of testing the accuracy of each model on separate training data.

Results

The tables below display the confusion matrices on predictions vs actual for the pre-COVID and post-COVID models on the post-COVID data for Black vs Asian victims. On the rows labeled by the leftmost columns, one can see the actual values of Asian or Black that the victims represented. The columns labeled by the topmost rows display whether the observation was predicted as Asian or Black. All were based on the predicted/actual values in the post-COVID testing dataset.

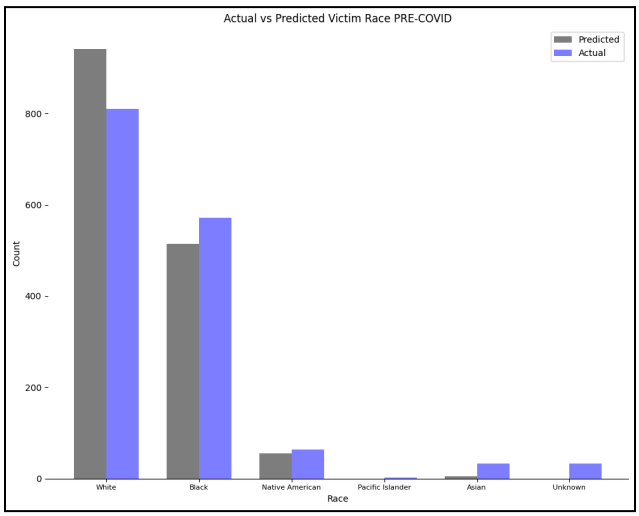
Although small, the difference in the Asian-Asian prediction-accurate score still exists between the 2 models. In addition, the same “true-positive” score increases from pre- to post-COVID for the confusion matrices of Asians to the other Victim Races, like White.

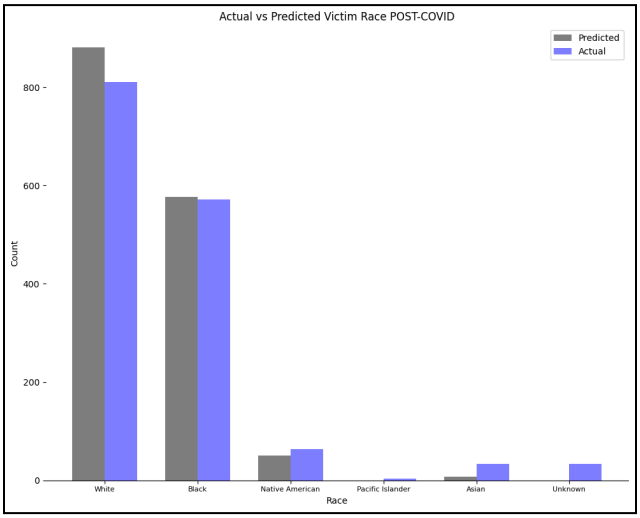
This finding could indicate that the post-COVID model more accurately predicted the post-COVID testing data compared to the pre-COVID model. With more experiments, there could exist a significance in temporal data from major events on how accurately models can predict. In other words, models can predict with different accuracies based on the timeframe they’re trained and tested on. This finding might carry importance in societally significant data, like this murder/manslaughter one from the pre- and post-COVID era.

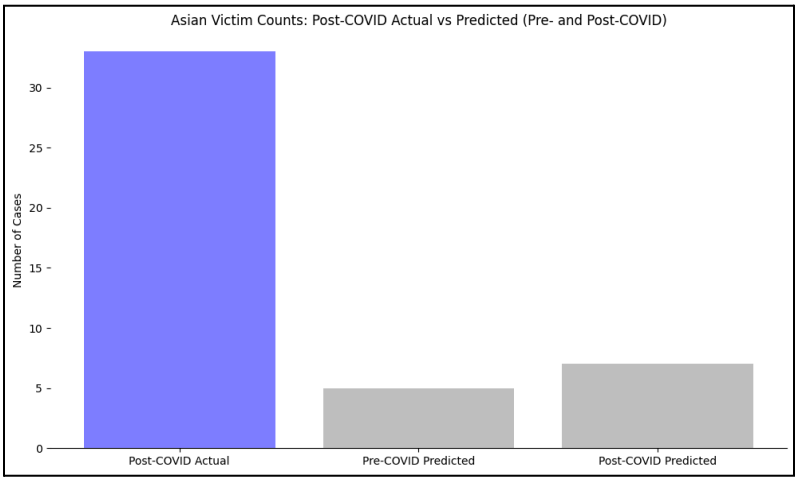
Pre-COVID Model Confusion Matrix: Black to Asian in Victim Race:		Predictions	
		Asian	Black
Actual	Asian	0.375000	0.625000
	Black	0.083612	0.916388

Post-COVID Model Confusion Matrix: Black to Asian in Victim Race:		Predictions	
		Asian	Black
Actual	Asian	0.400000	0.600000
	Black	0.002469	0.997531

Below are some graphs that display the predicted vs actual values of Victim Race done by the models with the post-COVID testing dataset. The first graph is with the pre-COVID model, while the second is with the post-COVID one. Focusing on the Asian statistics, it might be difficult to tell that there's a change in predicted and actual values, but the third graph on the Asian pre-COVID and post-COVID predictions displays the increase discussed in the above findings, with the tables.







Conclusion

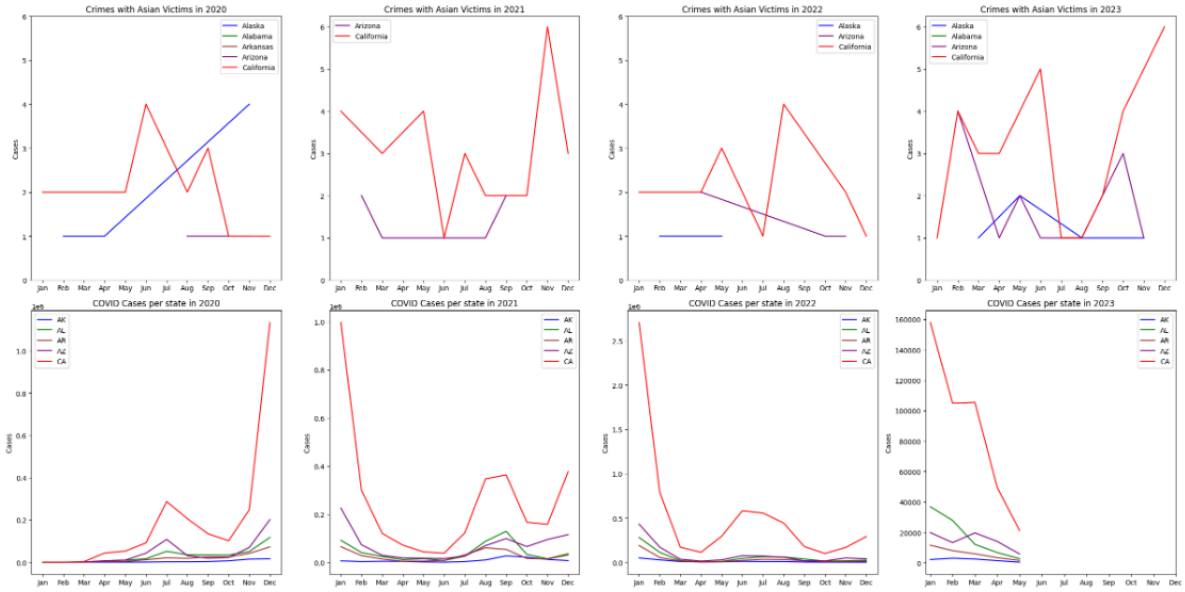
Two separate models were analyzed to see if the training data affected the model's accuracy in predicting the race of victims. The same dataset was used for both models, but the data was split into two different periods: pre- and post-COVID-19. The first model was trained on data from 2016 to 2019, while the second model was trained on data from 2020 to 2023. This allowed us to see if there was a significant difference in the accuracy of pre versus post-COVID model prediction on the post-COVID dataset.

The pre-COVID model performed slighty worse than the post-COVID model on accurately predicting Asian victims. This was most likely due to the fact that the model was trained on data after the large scale social phenomenon of COVID. In the future, models should be updated with the most accurate data as possible, while testing to see if their old data is still relevant. Models such as these would be useful to governments or law enforcement so as to help certain communities that may be targeted after a large social phenomenon.

Something to note would be that the post-COVID dataset might have contained a slightly larger number of Asian victims, which does align with the beliefs that Asians were being targeted more, though this finding isn't definite. However, the model demonstrated mismatched predictions that revealed that certain parameters regarding the offenders, such as races, ages, and months, could have impacted the models to become less accurate in prediction. The model seemed to misclassify many crimes as Black or White victims, which may be because there were a majority Black and White victims in the pre-COVID-19 dataset. This may suggest that the model was biased towards predicting these groups, leading to a higher number of false positives and false negatives. In the context of anti-Asian hate crimes, many crime reports have suggested that certain offender demographics were over- or underrepresented in datasets overall, which could have led to the model's misclassification of the manslaughter/murder crimes.

Even though the model was able to achieve a high accuracy, the difficulties reported in incorrectly predicting victim race when the offender was either Black or White suggested that hate crimes were not the result of a victims race. This meant that any race was capable of being assaulted/hatecrimed. Further training and exploration would be necessary to strengthen possible beliefs. Some of this further training could include testing different hyperparameters or using having a different minimum samples per leaf.

The graphs below have displayed the incorporation of the monthly COVID-19 cases dataset. Represented by the top layer of graphs were the monthly totals of attacks on Asian Americans in a few select states. Specifically, California was represented by the red line. On the bottom level of the graphs were new COVID-19 cases by state, with California again represented by the red line. There was a clear trend throughout these graphs. One of the clearest examples of this trend was at the end of 2020 and into the beginning of 2021. There was a significant spike in the number of COVID cases in December 2020 that continued into February 2021 in California. Looking at January 2021, there was also a spike in the number of crime cases with Asian victims. Additionally, after there was a big drop in the number of COVID cases in March 2021 in California, there was also a drop in the number of crimes against Asians shortly following in May 2021. A third example, where the correlating drop was a bit delayed, was at the beginning of 2023. Starting in January 2023 to May 2023, there was a significant drop in the number of COVID cases in California. Even though it took a month for this drop to be reflected in the number of Crime cases in California, there was a steep drop in the number of cases in June 2023 as well.



It was also important to take into account any other factors that might have influenced the trends of Crime cases against Asian victims in these states, including California, where the biggest increases and decreases were seen. These factors could be a small contribution or a significant contribution to the trends that were displayed in these graphs, and the trends of COVID cases could possibly have no effect. This was an example of correlation but not causation. Unless another dataset or social issues were incorporated, there could be no way to know for sure.

References

1. *Murder Accountability Project*. (2023). [Dataset]. https://www.murderdata.org/p/data-docs.html

2. Cabral, S. (2021, May). COVID-19 ‘hate Crimes’ against Asian Americans on Rise. *BBC News*. www.bbc.com/news/world-us-canada-56218684.

3. *Weekly United States COVID-19 Cases and Deaths by State*. (2025). [Dataset]. https://data.cdc.gov/Case-Surveillance/Weekly-United-States-COVID-19-Cases-and-Deaths-by-/pwn4-m3yp/about_data