HATE CRIME IN AMERICA

Miles A. Peña
DSC 540 – Data Preparation
Catherine Williams
Project Milestone 5
06/01/2024

Project Milestone 1

Project Subject Area:

The project will aim to answer questions related to the amount of hate crimes happening in the United States by state as well as what type of crime these victims are experiencing. It will focus primarily on the year 2019 and aim to find relationships in order to best determine which states require a bit more attention in order to reduce these crime rates.

Data Sources:

- **Flat File:**
  - The CSV file holds the data for Hate Crime in the United States between the years of 1991 and 2020. It includes variables like state, date, and offense name among others.
  - https://www.kaggle.com/datasets/lyonabido/hate-crime-statistics-annual-report-for-2020
- **API:**
  - It includes information about each offense, such as the time of day an incident occurred, the demographics of the offenders/victims, the known relationships between the offenders and victims, and many other details around how and where crime occurs.
  - FBI Crime Data API [Base URL: api.usa.gov/crime/fbi/cde/]
- **Website:**
  - The website provides data in the tabular form of crimes that occurred in the United States in the year 2019. It separates everything by state as well as by crime type among other variables.
  - https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/tables/table-5

Relationships:

- **CSV File:** Contains a column titled state name that stores the state that the crime took place in as well as offense name.
- **Website:** Contains a column title state as well as individual columns denoting the type of crime that is being recorded.
- **API:** I am unsure if I have done this part correctly. I am not quite sure if I have the correct URL and how to access the information correctly. I know we have worked with APIs before, but I feel that I am still struggling in this area. From what I understand however, this divides crimes by state as well and also shares how the crime occurred.

All 3 of these data sources are related by state and very possibly by type of crime as well but for the sake of this project I will be primarily focused on states. The website has a one-to-many relationship with the CSV file as there is only one row per state, but the CSV file has multiple instances of each state.

Project approach/plan:
The approach that I am intending with this project is blending the data based on the state that reported each specific crime. The intent is to find if there is a state where hate crimes happen more often or which states are more prone to these types of crime as well as looking to see if there is a "crime of choice" or so to speak, when it comes to hate crimes. Ultimately, the goal would be to take the data and implement better policies to assist with the rise in hate crime in America as well as decide where and how to better police the neighborhood and communities.

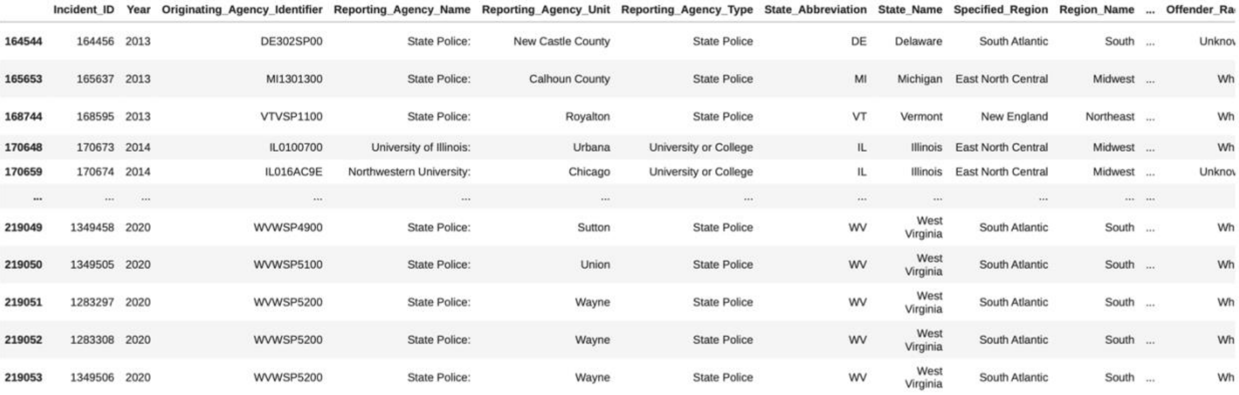
What concerns/challenges you think you will face with the data/project topic:
The challenges that I think I might face with the data/project is that the data might not line up the way that I want it to. I was unsure of how to work with the API in order to see that data that might be found in it. I am hoping that what I am looking for will be provided there (from the research I did, it should) however, if push comes to shove, I might have to find a different API to work from. I am hoping that there will be enough data to measure what I am looking for and actually find patterns in. I am unsure if am being too ambitious in trying to answer two questions with the information provided with the dataset and that may be my downfall in this project.

Ethical Implications of your project topic:
The biggest concern with ethical implications is the quality of the data. Ensuring that the data is accurate and reliable as well as complete in order to prevent erroneous or misleading conclusions that could impact (in the long run) law enforcement strategies or public policies.

Project Milestone 2

In order to clean up the hate crime dataset, the column names were changed to reflect names that were both easier to read and more descriptive of the content within them. The whitespace between column names was replaced with an underscore to align with Python conventions and best practices and increase readability and compatibility. The next steps look for both null values and duplicate values in the dataset. Finally, rows with null values were dropped from the dataset. It is important to ensure that the data we are using is obtained legally and provides accurate information about the crimes. The risk that could be created based on the transformations done is the removal of a row containing a null value that may include additional important values elsewhere. No assumptions were made in cleaning/transforming the data. The dataset was sourced from the Federal Bureau of Investigation - Crime Data Explorer, indicating its credibility and reliability.

Milestone 2 Dataset After Transformations:

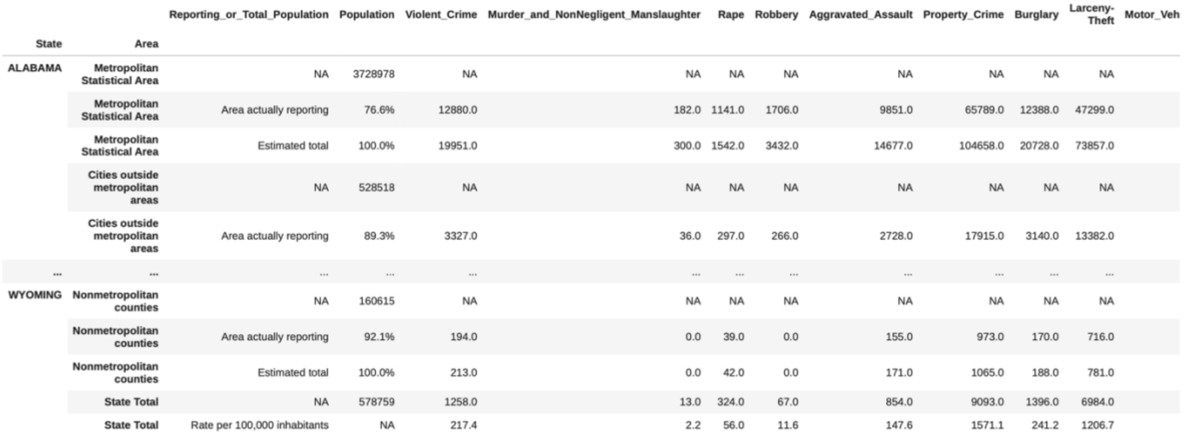


	Incident_ID	Year	Originating_Agency_Identifier	Reporting_Agency_Name	Reporting_Agency_Unit	Reporting_Agency_Type	State_Abbreviation	State_Name	Specified_Region	Region_Name	...	Offender_Ra
164544	164456	2013	DE302SP00	State Police:	New Castle County	State Police	DE	Delaware	South Atlantic	South	...	Unknov
165653	165637	2013	MI1301300	State Police:	Calhoun County	State Police	MI	Michigan	East North Central	Midwest	...	Wh
168744	168595	2013	VTVSP1100	State Police:	Royalton	State Police	VT	Vermont	New England	Northeast	...	Wh
170648	170673	2014	IL0100700	University of Illinois:	Urbana	University or College	IL	Illinois	East North Central	Midwest	...	Wh
170659	170674	2014	IL016AC9E	Northwestern University:	Chicago	University or College	IL	Illinois	East North Central	Midwest	...	Unknov
...	...	...	...	...	...	...	...	...	...	...	...	
219049	1349458	2020	WVWSP4900	State Police:	Sutton	State Police	WV	West Virginia	South Atlantic	South	...	Wh
219050	1349505	2020	WVWSP5100	State Police:	Union	State Police	WV	West Virginia	South Atlantic	South	...	Wh
219051	1283297	2020	WVWSP5200	State Police:	Wayne	State Police	WV	West Virginia	South Atlantic	South	...	Wh
219052	1283308	2020	WVWSP5200	State Police:	Wayne	State Police	WV	West Virginia	South Atlantic	South	...	Wh
219053	1349506	2020	WVWSP5200	State Police:	Wayne	State Police	WV	West Virginia	South Atlantic	South	...	Wh

Project Milestone 3

In order to clean up the FBI's Crime in the United States by State (2019) dataset, the column names were changed to remove numbers that were superscripted on some of the column titles as well as to name the unnamed column that existed. The whitespace between column names was replaced with an underscore to align with Python conventions and best practices and increase readability and compatibility. The next steps look for both null values and duplicate values in the dataset. Rows with null values were changed from NaN to NA since the cells were technically not empty but were cells that did not require input and were more so meant to be data-less. The final step was to create a hierarchical index that would provide a structured way to organize and represent the dataset, increase readability, and make it cleaner. It is important to ensure that the data we are using is obtained legally and provides accurate information about the crimes. I do not believe that there were any risks that may have been created based on the transformations done as the data was not changed other than the change from NaN to NA. No assumptions were made in cleaning/transforming the data. The dataset was sourced from the Federal Bureau of Investigation website through their Uniform Crime Reporting Program, indicating its credibility and reliability.

Milestone 3 Dataset After Transformations:

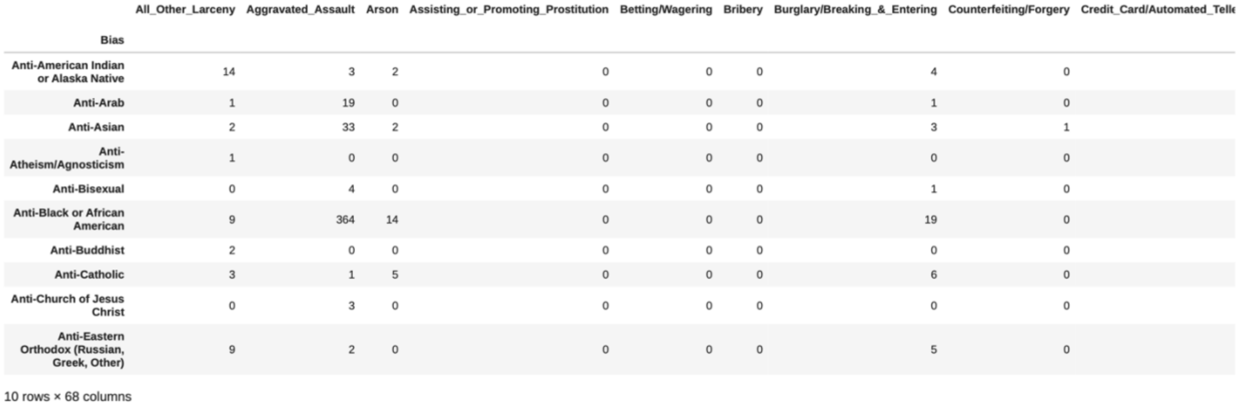


		Reporting_or_Total_Population	Population	Violent_Crime	Murder_and_NonNegligent_Manslaughter	Rape	Robbery	Aggravated_Assault	Property_Crime	Burglary	Larceny-Theft	Motor_Veh
State	Area											
ALABAMA	Metropolitan Statistical Area	NA	3728978	NA	NA	NA	NA	NA	NA	NA	NA	
	Metropolitan Statistical Area	Area actually reporting	76.6%	12880.0	182.0	1141.0	1706.0	9851.0	65789.0	12388.0	47299.0	
	Metropolitan Statistical Area	Estimated total	100.0%	19951.0	300.0	1542.0	3432.0	14677.0	104658.0	20728.0	73857.0	
	Cities outside metropolitan areas	NA	528518	NA	NA	NA	NA	NA	NA	NA	NA	
	Cities outside metropolitan areas	Area actually reporting	89.3%	3327.0	36.0	297.0	266.0	2728.0	17915.0	3140.0	13382.0	
...	...	...	...	...	...	...	...	...	...	...	...	
WYOMING	Nonmetropolitan counties	NA	160615	NA	NA	NA	NA	NA	NA	NA	NA	
	Nonmetropolitan counties	Area actually reporting	92.1%	194.0	0.0	39.0	0.0	155.0	973.0	170.0	716.0	
	Nonmetropolitan counties	Estimated total	100.0%	213.0	0.0	42.0	0.0	171.0	1065.0	188.0	781.0	
	State Total	NA	578759	1258.0	13.0	324.0	67.0	854.0	9093.0	1396.0	6984.0	
	State Total	Rate per 100,000 inhabitants	NA	217.4	2.2	56.0	11.6	147.6	1571.1	241.2	1206.7	

Project Milestone 4

In order to clean up the information pulled from the FBI Crime Data API, the first two column names were changed to follow the same capitalization rules that the rest of the column names had. The whitespace between column names was replaced with an underscore to align with Python conventions and best practices and increase readability and compatibility. The next steps look for both null values and duplicate values in the dataset. The 'Year' column was dropped from the dataset as it didn't provide any pertinent information since all of the data was from the same year (2019). The final step was to create a hierarchical index that would provide a structured way to organize and represent the dataset, increase readability, and make it cleaner. It is important to ensure that the data we are using is obtained legally and provides accurate information about the crimes. I do not believe that there were any risks that may have been created based on the transformations done to the data. No assumptions were made in cleaning/transforming the data. The dataset was sourced from the Federal Bureau of Investigation - Crime Data Explorer – FBI Crime Data API, indicating its credibility and reliability.

Milestone 4 Dataset After Transformations:



	All_Other_Larceny	Aggravated_Assault	Arson	Assisting_or_Promoting_Prostitution	Betting/Wagering	Bribery	Burglary/Breaking_&_Entering	Counterfeiting/Forgery	Credit_Card/Automated_Telle
Bias									
Anti-American Indian or Alaska Native	14	3	2	0	0	0	4	0	
Anti-Arab	1	19	0	0	0	0	1	0	
Anti-Asian	2	33	2	0	0	0	3	1	
Anti-Atheism/Agnosticism	1	0	0	0	0	0	0	0	
Anti-Bisexual	0	4	0	0	0	0	1	0	
Anti-Black or African American	9	364	14	0	0	0	19	0	
Anti-Buddhist	2	0	0	0	0	0	0	0	
Anti-Catholic	3	1	5	0	0	0	6	0	
Anti-Church of Jesus Christ	0	3	0	0	0	0	0	0	
Anti-Eastern Orthodox (Russian, Greek, Other)	9	2	0	0	0	0	5	0	

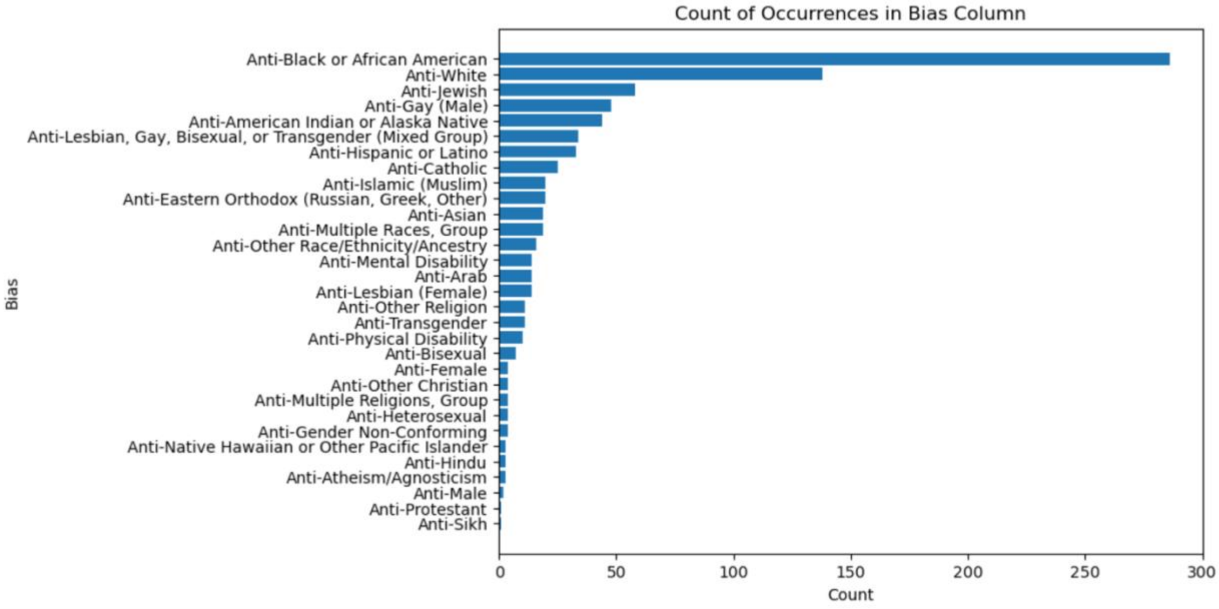
10 rows × 68 columns

Project Milestone 5

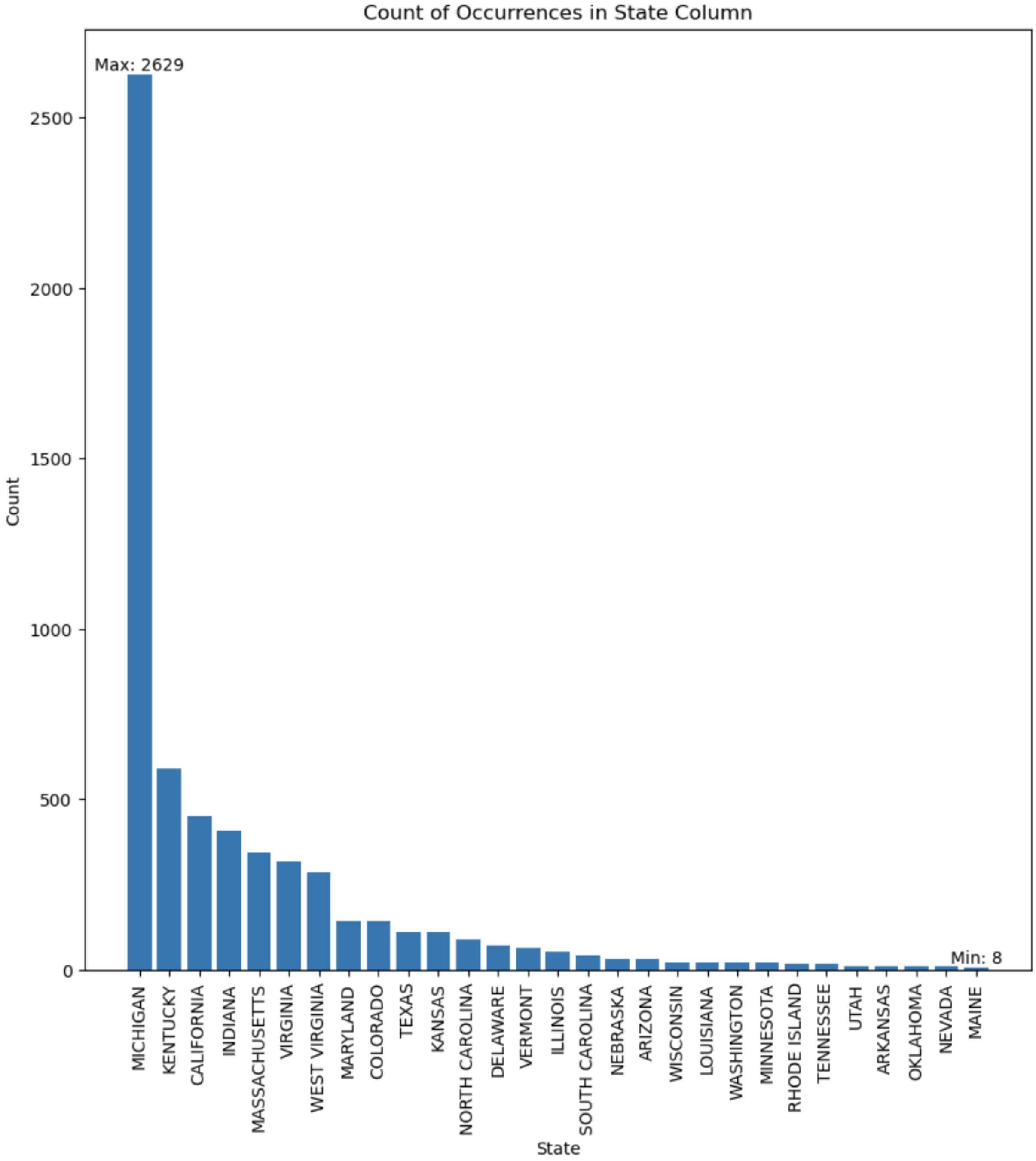
The final milestone aimed to join all three of the previously cleaned datasets into one dataset. Regrettably, my initial plan to join the datasets based on the State columns fell through due to limitations in accessing the API data in the desired format. Additionally, during Milestone 4, I erroneously assumed I could join the datasets based on offense types, which turned out to be unfeasible. To complete the assignment to the best of my abilities given the constraints, I decided to join only two tables on the bias associated with the crime. To fulfill the assignment requirements more comprehensively, I also opted to conduct an additional join operation in addition to the one previously mentioned. The second join merges the datasets "hate_crime_stats" and "fbi_table" based on the "State" column. It is important to ensure that the data we are using is obtained legally and provides accurate information about the crimes. I do not believe that there were any risks that may have been created based on the transformations done to the data. No assumptions were made in cleaning/transforming the data. The datasets were sourced from the Federal Bureau of Investigation - Crime Data Explorer , the Federal Bureau of Investigation website through their Uniform Crime Reporting Program, and the Federal Bureau of Investigation - Crime Data Explorer – FBI Crime Data API, indicating their credibility and reliability. Given that they all originated from the FBI, we can confidently assert that the acquisition of this data was conducted ethically, adhering to legal and regulatory guidelines.

Graphical Analyses:

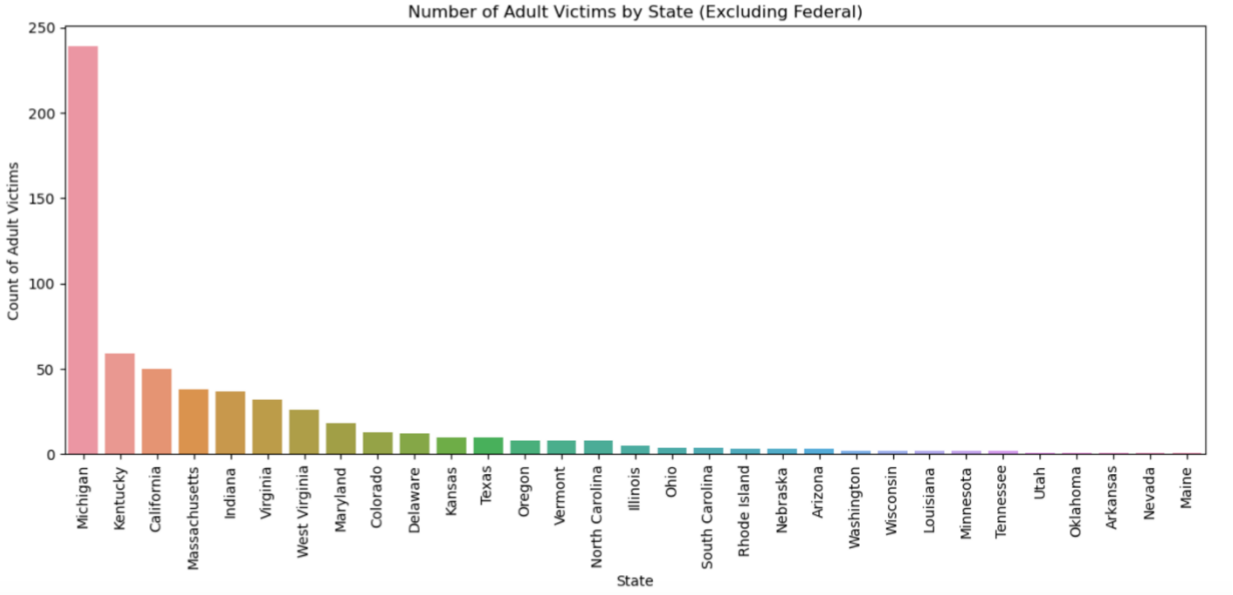
Count of Occurrences in Bias Column:



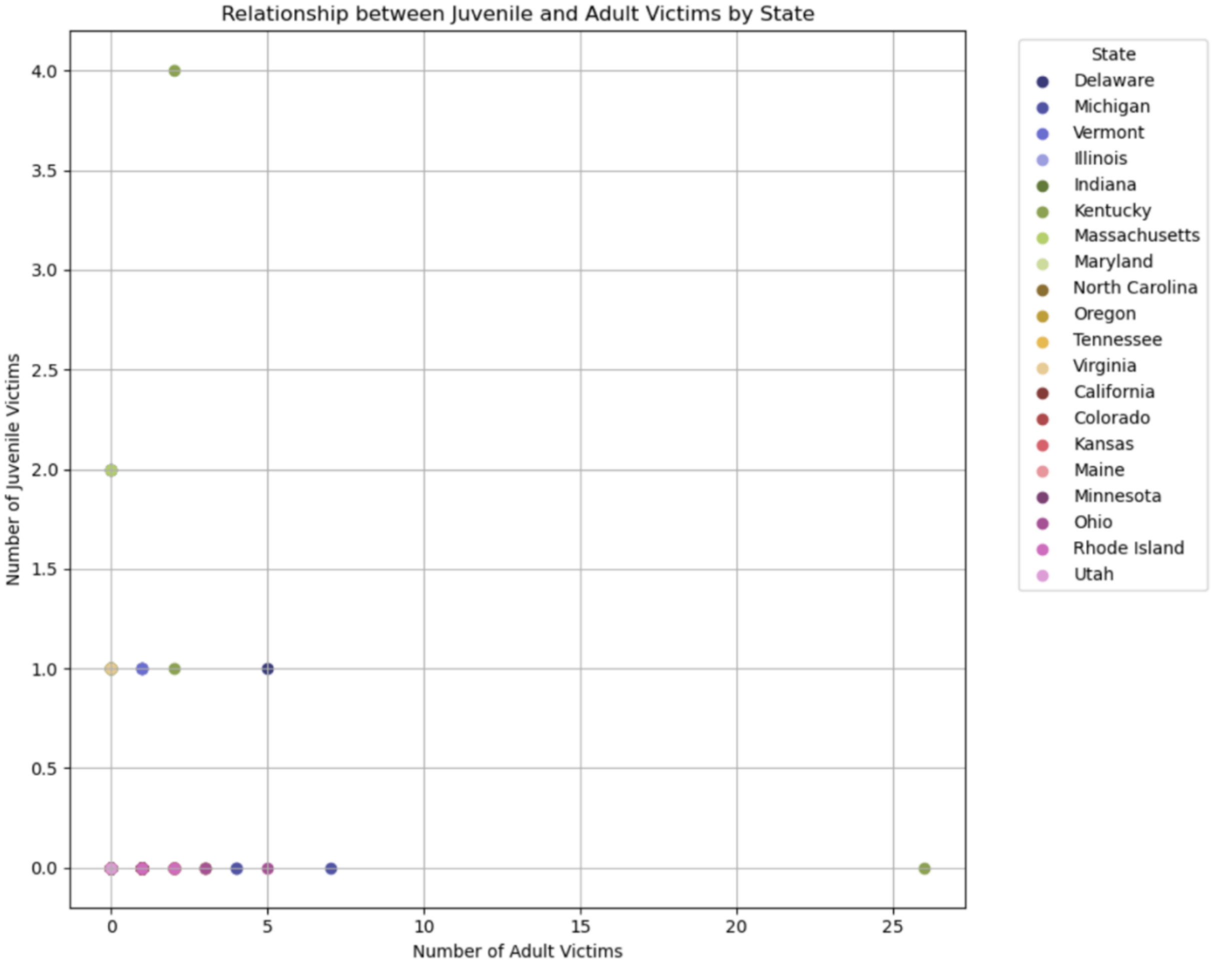
Count of Occurrences in State Column:



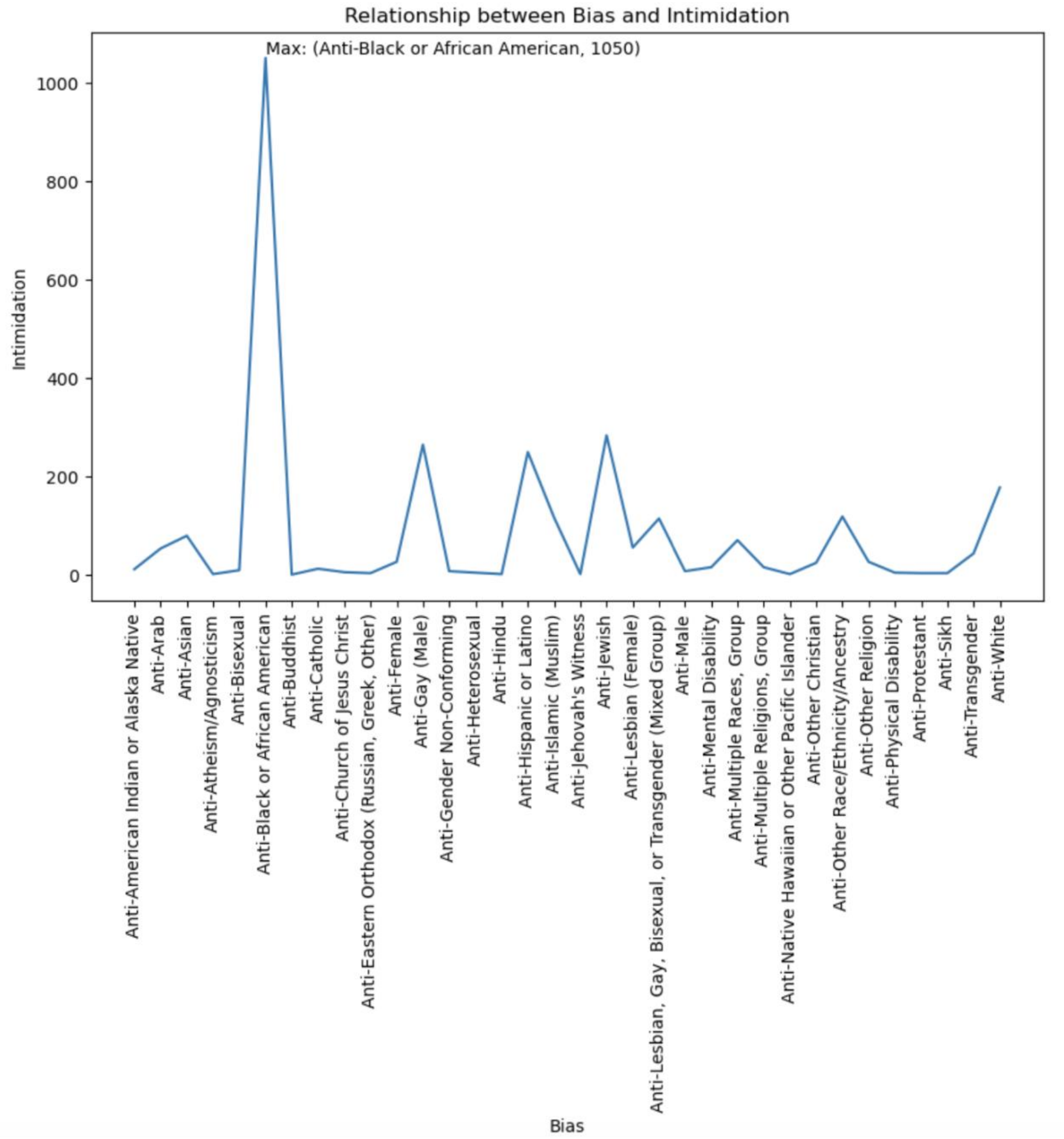
Number of Adult Victims by State (Excluding Federal):



Relationship between Juvenile and Adult Victims by State:



Relationship between Bias and Intimidation:


Relationship between Bias and Intimidation
Max: (Anti-Black or African American, 1050)
Intimidation
Anti-American Indian or Alaska Native
Anti-Arab
Anti-Asian
Anti-Atheism/Agnosticism
Anti-Bisexual
Anti-Black or African American
Anti-Buddhist
Anti-Catholic
Anti-Church of Jesus Christ
Anti-Eastern Orthodox (Russian, Greek, Other)
Anti-Female
Anti-Gay (Male)
Anti-Gender Non-Conforming
Anti-Heterosexual
Anti-Hindu
Anti-Hispanic or Latino
Anti-Islamic (Muslim)
Anti-Jehovah's Witness
Anti-Jewish
Anti-Lesbian (Female)
Anti-Lesbian, Gay, Bisexual, or Transgender (Mixed Group)
Anti-Male
Anti-Mental Disability
Anti-Multiple Races, Group
Anti-Multiple Religions, Group
Anti-Native Hawaiian or Other Pacific Islander
Anti-Other Christian
Anti-Other Race/Ethnicity/Ancestry
Anti-Other Religion
Anti-Physical Disability
Anti-Protestant
Anti-Sikh
Anti-Transgender
Anti-White
Bias