1. Motivation

For my project, I decided to take a look at some data to try and determine if the common attribution of video games to violence was really true. I feel that this is an often repeated phrase, but I have never seen any data to back it up. Because of this apparent lack of information on this topic, I decided that I could prove once and for all, at least to myself, whether this adage carries any weight at all whatsoever.

2. Data Sources

All of my data came from two sources. For video game information, I used a kaggle dataset titled simply "Video Game Dataset" The link to this data is provided below this paragraph. According to the documentation on the kaggle site, the data was obtained using a RAWG API, however when I used it, I simply downloaded the dataset as a CSV file from the website. The data contained mayn different variables, however I was mostly interested in the ESRB rating and the release date. The full list of variables is listed below as well. I processed all 474,417 games listed in the dataset for my project, but not all were needed. The time frame was also limited from 1995 to 2020 in order to match up better with the crime data I had collected.

- id: An unique ID identifying this Game in RAWG Database
- slug: An unique slug identifying this Game in RAWG Database
- name: Name of the game
- metacritic: Rating of the game on Metacritic
- released: The date the game was released
- tba: To be announced state
- updated: The date the game was last updated
- website: Game Website
- rating: Rating rated by RAWG user
- rating_top: Maximum rating
- playtime: Hours needed to complete the game
- achievements_count: Number of achievements in game
- ratings_count: Number of RAWG users who rated the game
- suggestions_count: Number of RAWG users who suggested the game
- game_series_count: Number of games in the series
- reviews_count: Number of RAWG users who reviewed the game
- platforms: Platforms game was released on. **Separated** by ||
- developers: Game developers. **Separated** by ||
- genres: Game genres. **Separated** by ||
- publishers: Game publishers. **Separated** by ||
- esrb_rating: ESRB ratings
- added_status_yet: Number of RAWG users had the game as "Not played"
- added_status_owned: Number of RAWG users had the game as "Owned"
- added_status_beaten: Number of RAWG users had the game as "Completed"

- added_status_toplay: Number of RAWG users had the game as "To play"
- added_status_dropped: Number of RAWG users had the game as "Played but not beaten"
- added_status_playing: Number of RAWG users had the game as "Playing"

https://www.kaggle.com/datasets/jummyegg/rawg-game-dataset?resource=download

For my second source, I looked at the crime data as provided by a usa.gov API. I have provided the link below as well as a secondary way to obtain the data from the site as it requires an API key. The API returns a JSON file containing the data for the category and range that you choose. I was interested in the US national estimates for crime from 1995 to 2020, a date range that included both datasets. This contains the year, the population, and a number of crime data estimates. I was interested particularly in violent crimes, homicide, aggravated assault, and arson, but other crimes such as rape, robbery, property crime, burglary, larceny, and motor vehicle theft were included as well. Though there was quite a bit of data to look at, I only retrieved 25 records containing all of this data for each year, as the time period was the limiter here. Again, the links I used are displayed below, first for the API site I used and second for the actual request with API key included.
https://crime-data-explorer.fr.cloud.gov/pages/docApi
https://api.usa.gov/crime/fbi/sapi/api/estimates/national/1995/2020?API_KEY=iiHnOKfno2Mgkt5AynpvPpUQTEyxE77jo1RU8PIv

3. Data Manipulation Methods

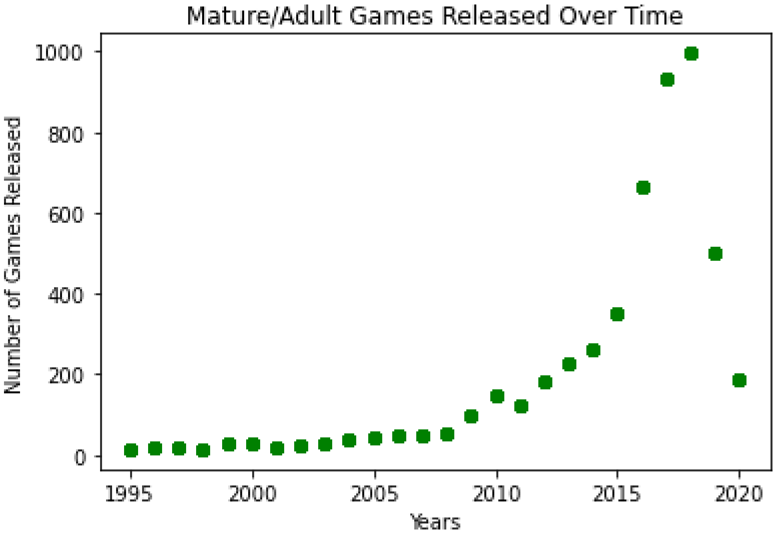
For my first data source, the video game data, there was a good amount of data manipulation I had to do. First off, I needed to reduce the hundreds of thousands of games included to only the ones I wanted to look at. This was done by selecting only the games with an ESRB rating of Mature or Adults Only. From there, I needed to then look at only the games that fit the time frame of 1995 to 2020. The format of the released column was too much information for my use, so I also added a column of the release year in order to quickly get a count for each year. From there, I dropped any N/A values in the year category, just in case. Then, I finally filtered the date range down to the 1995 to 2020 range once the data was cleaned up. The last thing I did was sort the game data by the release year in descending order. As a final step, I also added another column for the count for the number of games with Mature or Adults Only ESRB ratings that were in the timeframe. I did some text processing to separate the game genres into a more easy form to work with, however, I didn't end up using the new column of genres in my final analysis. I ran into a fair bit of problems getting the columns to show the data in the proper form. As an example, the release year's type was a float, and I went through a bit of time trying to change it to an int before realizing it wouldn't matter since there are no fractional years anyway. I had another struggle trying to figure out how to take the value counts numbers and store them somehow before finally solving the problem by adding the counts column for every row of the table. The bulk of my time was spent trying to resolve this issue and the problems that stemmed from it.

For my second data source, the US crime estimates, most of my work was trying to figure out how to best display the data. The first hurdle was changing the JSON format into a DataFrame. This was quickly resolved, but I was unable to read the data at first before looking closer at how it was formatted. Next, I needed to figure out how I would compare the numbers when there is a change in the population every year. I fixed this by adding new columns for the parts of the data that I wanted to look at with the proportion of the population for each. Since I was interested in violence, I chose to add columns for violent crime, homicide, aggravated assault, and arson, as those seemed to be the most direct comparisons to common actions in violent video games, at least to me. From there, I sorted the data by the year to ensure it would be similarly ordered to the video game data. One major problem I had with this data came with plotting the results. The proportion of the population that actually commits these violent crimes is fairly low, so the data could not be easily visualized on one plot. To resolve this, I added a second visualization for this data with four smaller plots to show how each value changed individually. I also multiplied the proportions by 100 to get a percentage which showed up better since the numbers were larger. Even with the highest value, there was no data points even close to one percent of the population, so the various violent crimes were not commonly occurring, even when the numbers were higher.

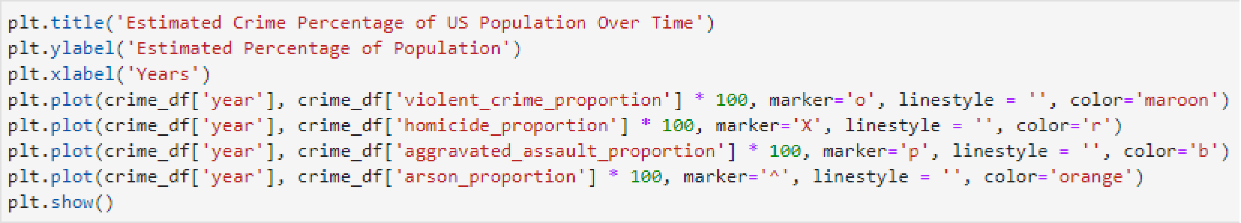
Working from top to bottom, my code essentially does exactly as described above. First, I import everything I need. From there, I read in the data, first from the video game CSV file, then from the crime JSON file. Once both are available as DataFrames, I move on to sorting and filtering through the data for each, starting with the video games. I filter out everything except the Mature and Adults Only ESRB rating, then add the release year column. Once the release year is separate, I drop any N/A values and then filter the data down to the data range I wish to work with, finally sorting the data by the year in descending order and adding the game count column. For the crime data, I first add the proportions for the crimes previously discussed, and then sort by the year in descending order to match the video game format. Lastly, I have a visualization for each data source, with two for the crime data to show both all on one graph and to show the trend of each on its own. There were a couple issues with the visualizations, but simply reviewing the previous in-class notebooks and the matplotlib documentation solved all of the problems fairly easily.

4. Analysis and Visualization

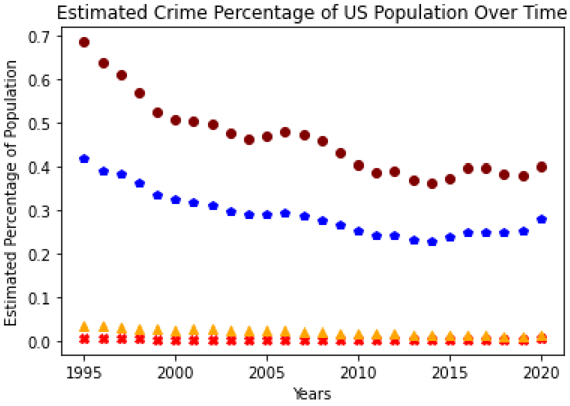
While I didn't run any actual analytical calculations, I think that the results and the visualizations show enough to draw conclusions from. Though this data may not translate to the full population, I feel comfortable saying that we can assume it is accurate for the data obtained at the very least. I set out to try and determine if there was any truth to the commonly heard belief that violent video games cause violence in the real world. In order to find out if this belief was true or not, I took all of the violent video games, or those with Mature or Adults Only ratings, and looked at how many were released per year from 1995 to 2020. I also looked at the estimated crime data generated by the FBI in the same range of 1995 to 2020, looking particularly at violent crime, homicide, aggravated assault, and arson. I then took the data compiled and created visualizations for each to see if there was any correlation.



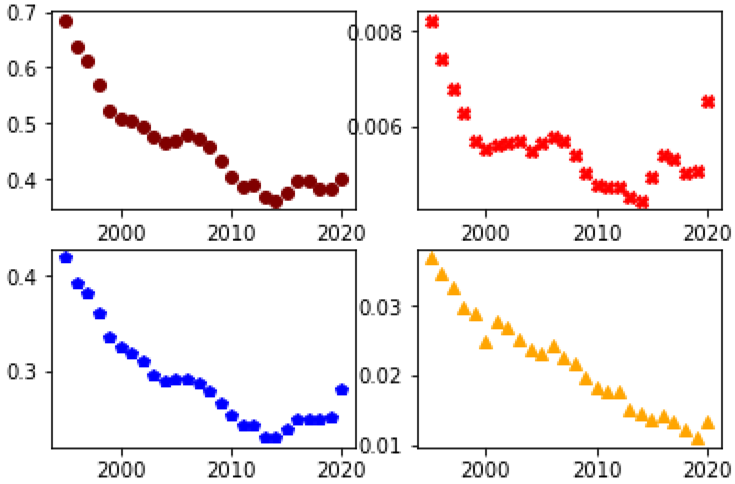
This plot shows the video game release year of these Mature and Adult games over time. On its own, this does not tell us much about any possible correlation between violence in games and violence in society. However,if we compare this to the plot for the percentage of the population estimated to be committing these violent crimes, we can begin to draw conclusions.



```
plt.title('Estimated Crime Percentage of US Population Over Time')
plt.ylabel('Estimated Percentage of Population')
plt.xlabel('Years')
plt.plot(crime_df['year'], crime_df['violent_crime_proportion'] * 100, marker='o', linestyle = '', color='maroon')
plt.plot(crime_df['year'], crime_df['homicide_proportion'] * 100, marker='X', linestyle = '', color='r')
plt.plot(crime_df['year'], crime_df['aggravated_assault_proportion'] * 100, marker='p', linestyle = '', color='b')
plt.plot(crime_df['year'], crime_df['arson_proportion'] * 100, marker='^', linestyle = '', color='orange')
plt.show()
```



This second visualization shows the estimated crime statistics over time for the crimes chosen for analysis: violent crime, homicide, aggravated assault, and arson. Above the plot is the cell that generates the plot, shown in order to give explanation for which colors and shapes are associated with which crime. Below is a breakdown of each crime shown with a better scale for each in order to display trends in the data more effectively.



As you can see, the percentage of the population committing these crimes actually seems to decrease as time goes on with some uptick in 2020 in particular. Overall, however, the data seems to trend downwards, suggesting that less crime is occurring relative to the size of the population, not more. If video games truly did cause violence, we would expect the opposite to occur, as a large number of Mature or Adult Only games were released between 2015 and 2020. In this same timeframe on the crime plots, we actually see the data stay relatively unchanged or level off, with the exception of homicide which did see a bit of a spike before dropping off. Another point of note is 2020 itself. While less games were released in 2020 than in previous years, all crime seems to see an uptick, although a small one. Again, this is the opposite of what we would expect from the data if video games truly were causing violence and murder. Unfortunately, I was unable to combine this data into one dataset as the scale for each didn't seem to be comparable. Despite this lack of combination, I think that we can see from the shapes of the graphs that there is likely no link between video games and violence in the US.

Though these insights and observations are interesting, I don't believe that we can make much out of them as many more factors contribute to crime than just video game releases. Other factors should be considered, such as economic conditions and political or social changes, as well as many others that weigh in. Since this is purely an observational study, it is worth noting as well that there is no control group or really any control at all with this data, it is simply pulled from publicly available sources. I would caution against assigning any meaning to this data or conclusively stating any conclusion since there are so many factors that go into violent crimes and other offenses being committed. Despite this, it does appear that video games don't seem to have much of an effect on crime rates, and if they do have any, it appears to be inversely proportional, meaning as violent video game releases increase, violent crimes in the US go down.