ENPM808W Homework #1

Daniel M. Sahu

October 2, 2022

1. Analysis of simulated New York Times webpage activity.

   Problem #1 involves loading and analyzing data which simulates page activity on the New York Times website over the course of a month. The simulated data includes user metadata like age, $gender$, etc. (if the user is logged in), as well as $impressions$ (page views) and $clicks$.

   All code for Problem #1 can be found on GitHub. See this script in particular.

   Data Cleaning:

   The data is contained in 31 CSV files, one for each date. Data cleaning was fairly straightforward - the only handling required was to avoid lumping metadata for logged in users with metadata for anonymous users, as certain field defaults cause misleading results. For example, the $gender$ field for anonymous users defaults to 0, which is the exact same value as female logged in users. To avoid this (and similar) issues we explicitly cast anonymous user data as $UNKNOWN$.

   Parts (a,b):

   After data cleaning our first goal was to plot the click-through-rate (CTR) for a variety of different age categories. Additionally, we wanted to split the data across other interesting categories to see what sort of trends presented themselves. Figures 1 and 2 summarize our results.

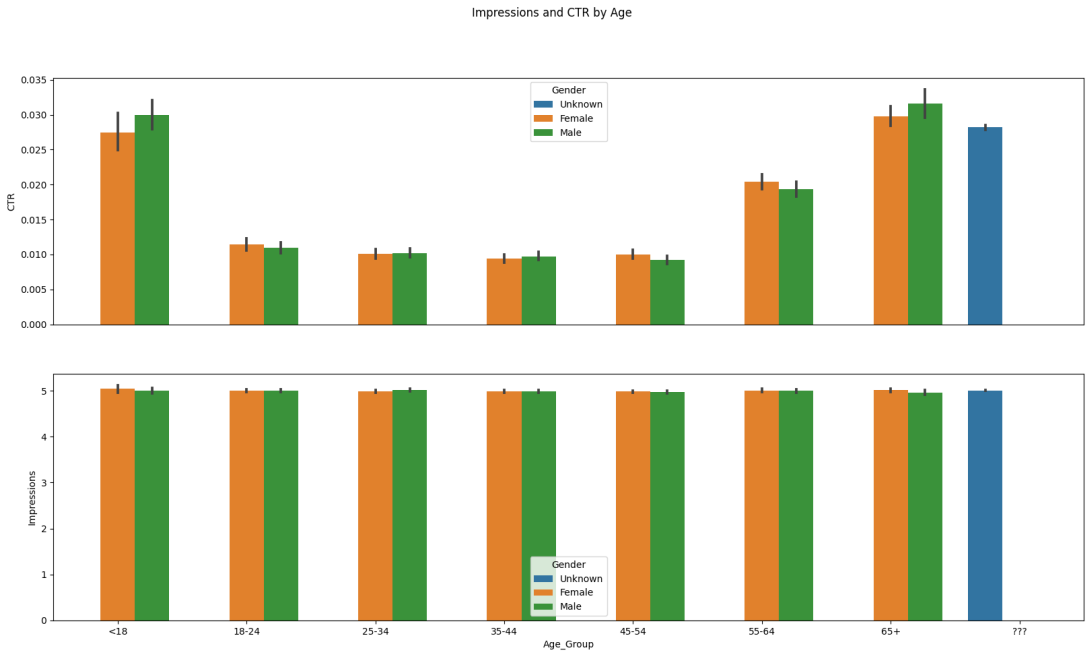

Figure 1: Impressions and CTR by Age

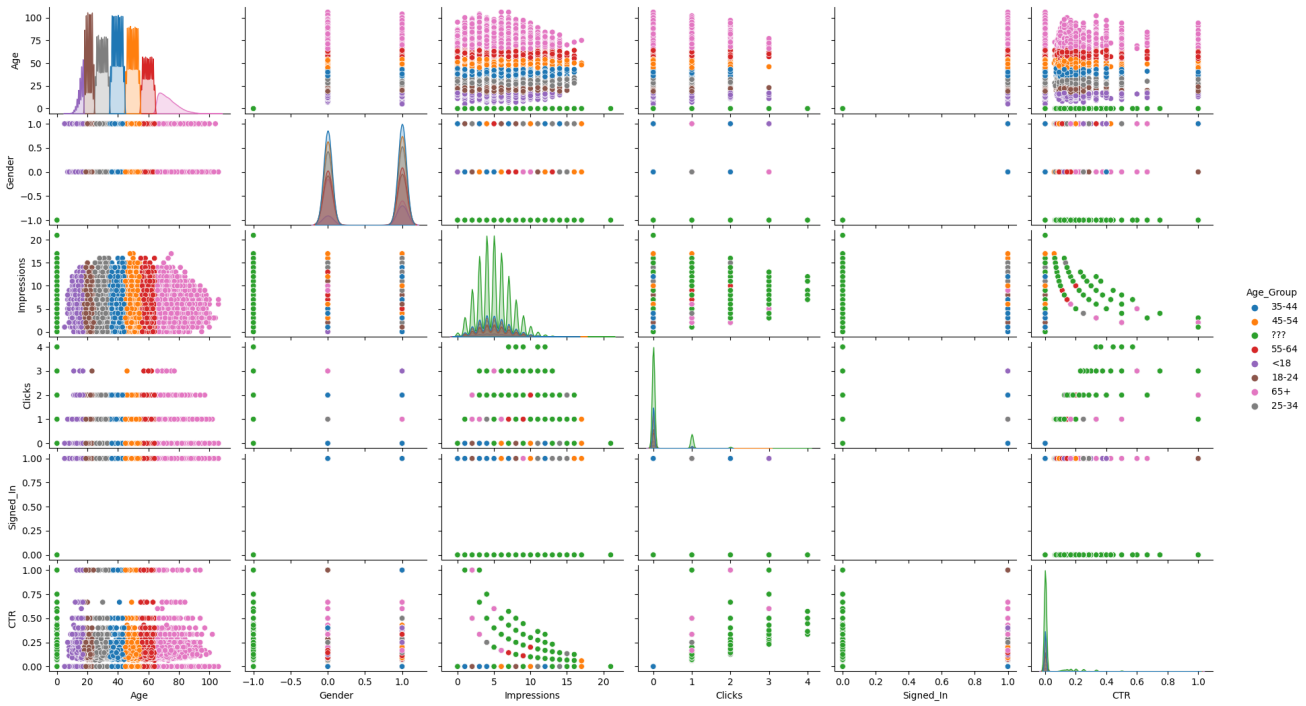

Figure 2: Categorical Pairplot

Figure 1 shows the *Impressions* and *CTR* of our various age groups. To handle anonymous users we deliberately split them out into a separate category. Each *agegroup* is also split along *gender* lines (except anonymous users, of course). Although the average number of *Impressions* is fairly stable for all users, we see a marked difference in *CTR* based on *agegroup*. There's a clear trend where younger (< 18) and older (65+) users have relatively high *CTR*s. Users in the middle tend to click less for the same number of *Impressions*.

The investigating of Figure 1 led to the conclusion that a more holistic view of the data would be useful. To that end we used the Python seaborn package's pairplot function to see if any other relationships stand out. Figure 2 shows the result. Although most of the pairwise relationships aren't interesting, there are a couple of notworthy plots:

- The Histogram plot of *Impressions* vs. *Impressions* demonstrates that the majority of our users are *not* logged in.
- The plots of *Impressions* vs. *CTR* show trend lines that clearly indicate that this is fake data. They look like this because the number of *click*s are bounded by [0,4], which is much too clean for real-world data.

Parts (c,d):

In Parts (a,b) the only interesting trend is a marked relationship between *AgeGroup* and *CTR*. All that analysis was done on a single day's data though, so the next step is to analyze trends over time. To that end we developed a number of metrics which summarize the overall information for a given date and compare those metrics over time.

The metrics we settled upon are:

- *Count*: The number of users collected in a given day.
- *CTRMean*: The mean of the CTR.
- *CTRStddev*: The standard deviation of CTR.
- *ClicksMean*: The mean of the clicks.
- *ClicksStddev*: The standard deviation of Clicks.
- *ImpressionsMean*: The mean of the impressions.
- *ImpressionsStddev*: The standard deviation of impressions.

In addition to the above we also pass forward the following metadata for ease of analysis:

- *Day*: The day of the month this data comes from.
- *AgeGroup*: Metrics are collected per-age group for a given day, not collectively.

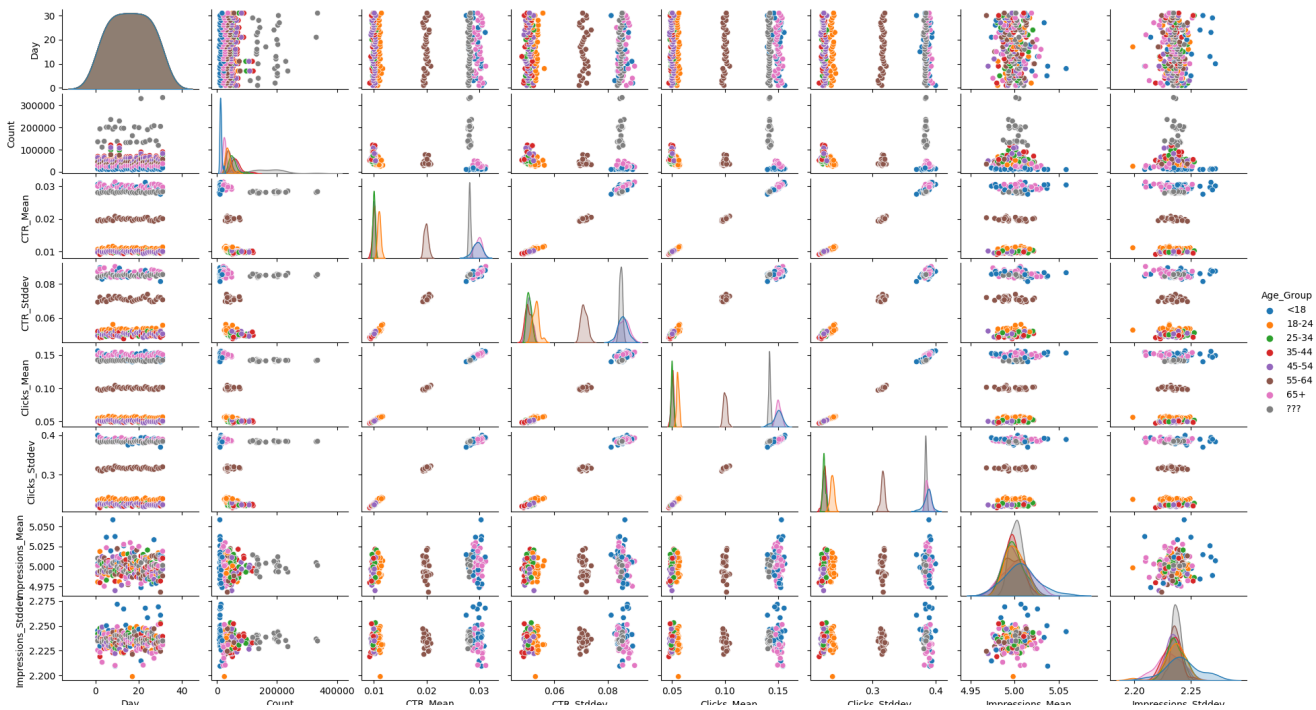

Figure 3: Metric Pairplot

Our metric information across time is summarized in another pairplot in Figure 3. Although this plot can be somewhat overwhelming, it gives a high enough snapshot to allow us to spot a few interesting trends:

- Although there's no clear temporal trend like "CTR increases over time" it is clear that all the days are not homogeneous. There are a number of days which stand out as "fast news days" where the user count is significantly higher than average. See the "Day vs. Count" plot.
- Although there's no clear trend for number of Impressions, the same marked relationship between Age Group and CTR / Clicks is present over time.

The trend between Age Group and Clicks / CTR is so strong that we can visually infer that the data was simulated with (very roughly) the following Gaussian distributions:

Metric	<18	18-24	25-34	35-44	45-54	55-64	65+	???
Clicks (mean)	0.15	0.05	0.05	0.05	0.05	0.10	0.15	0.15
Clicks (stddev)	0.40	0.25	0.25	0.25	0.25	0.30	0.40	0.40
CTR (mean)	0.03	0.01	0.01	0.01	0.01	0.02	0.03	0.03
CTR (stddev)	0.90	0.05	0.05	0.05	0.05	0.07	0.90	0.90

2. Analysis of a self-selected dataset.

   The mandate of this problem was to find and analyze a non-trivial dataset. We chose to analyze data on police shootings in the U.S. from Kaggle. This dataset includes select information about police shootings that resulted in fatalities between 2015 and 2022. The goal here was mainly to get a sense for *what* information is actually available on such a controversial topic for my edification.

   As with Problem #1 all the code used to generate plots is found on Github, specifically this script.

   Raw Data Overview:

   The raw dataset contains the following columns. Items marked "unused" were not used in further analysis, either due to resource constraints or because a use couldn't be determined.

   - *id* (unused): Unique identifier.
   - *name* (unused): The name of the deceased.
   - *date*: The date of the encounter.
   - *manner of death* (unused): How the deceased was killed.
   - *armed*: Weapon the deceased had (if any).
   - *age*: The age of the deceased.
   - *gender*: The gender of the deceased.
   - *race*: The race of the deceased.
   - *city* (unused): The city in which the encounter occurred.
   - *state*: The state in which the encounter occurred.
   - *signs* of mental illness (unused): Whether or not the deceased exhibited signs of mental illness.
   - *threat level* (unused): Perceived threat level of the deceased.
   - *flee*: The means of attempted escape, if applicable.
   - *body camera* (unused): Whether or not the officer(s) were wearing a body camera.
   - *longitude* (unused): Longitude of the incident.
   - *latitude* (unused): Latitude of the incident.
   - *is geocoding exact* (unused): Whether or not the coordinates are exact.

   Data Munging:

   Much of the raw information presented above was desired, but some of it wasn't immediately useful in its present form. To that end the following operations were performed to extract salient information:

(a) *date*, *age*, *gender* were used directly.

(b) *race* was used directly, with NaNs replaced by *O* (Other).

(c) *year* was extracted from *date*.

(d) *region* was extracted by assigning *state* information to the corresponding geographic region (e.g. SOUTHWEST).

(e) *weapon* was extracted by binning *armed* into *GUN*, *OTHER*, *UNARMED* as a rough proxy of dangerousness.

(f) *armed* was converted to a boolean.

(g) *escaping* was extracted by converting *flee* to a boolean.

(h) All NaNs found after the above were dropped.

Cleaned Data Overview:

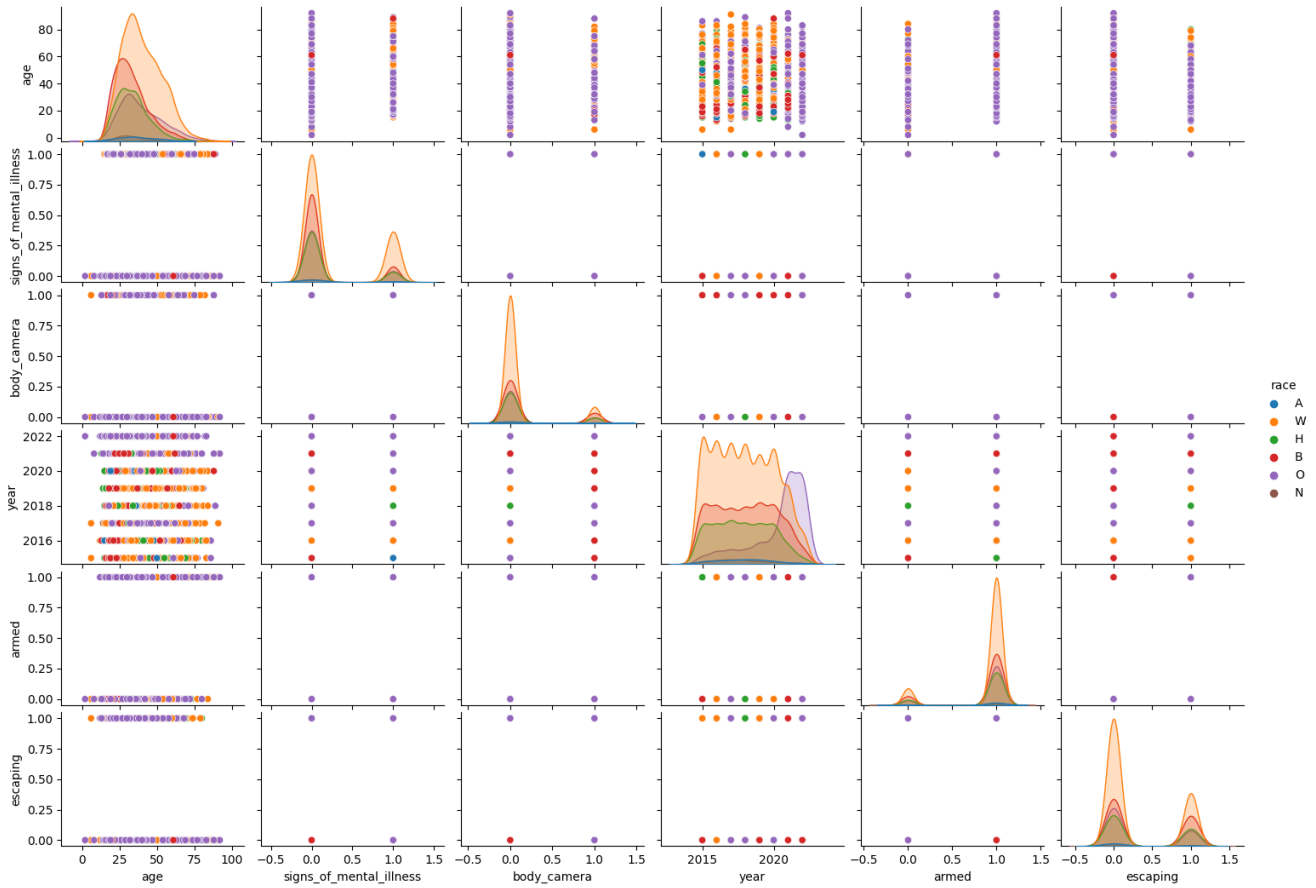

Figure 4: Cleaned Data Pairplot

We began the analysis of salient information with the overview plot given in Figure 4. Unlike in Problem #1 we don't see any obvious standout trends in this figure. We see that the racial breakdown is significant, but without information about the percentage of each race in the overall population this information isn't very helpful.

Temporal Trend Analysis:

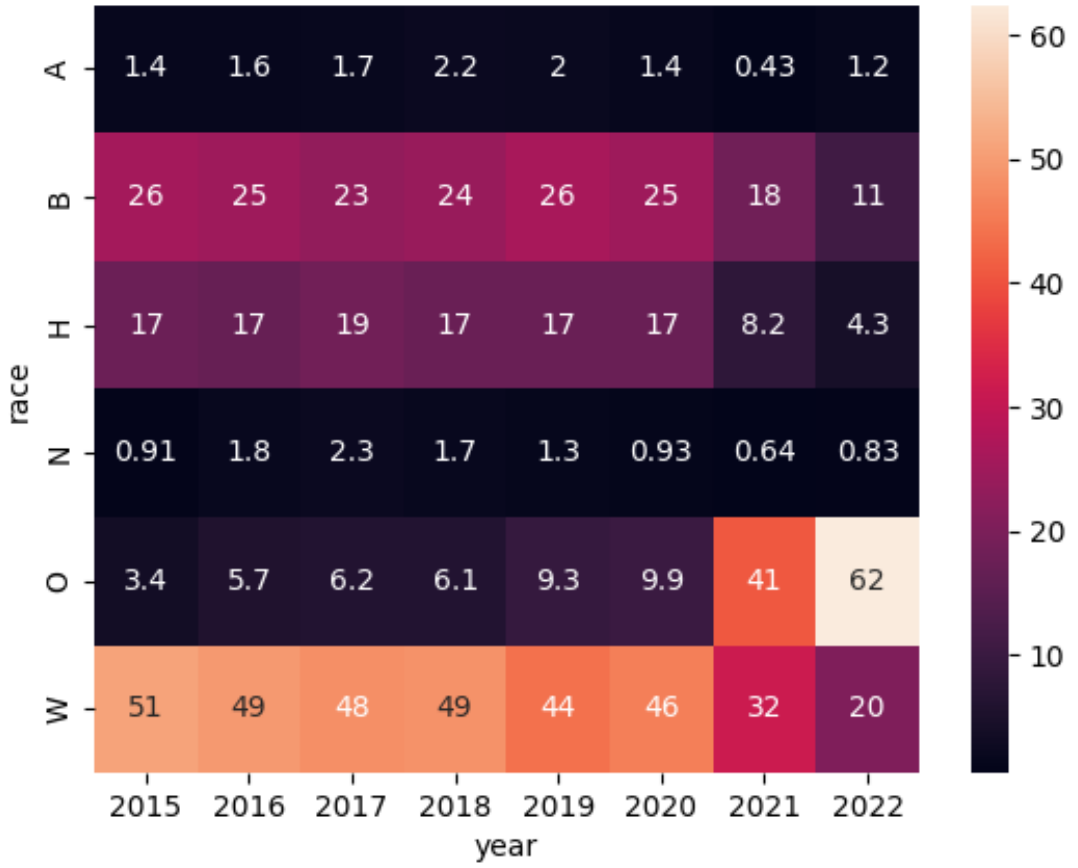

Figure 5: Count by Race over Time

Figures 5, 6, and 7 demonstrate the general trend for the percentage of a given variable (*race*, *region*, and *weapon*, respectively) over time. The goal of these sorts of visualizations is twofold. First, we get an overview of the percentage of the proportions of the population broken along, e.g., racial bounds. Second, we can see how that proportion changes over time, if at all. Note that these plots do not capture changes in magnitude over time. For example, if the overall number of shooting victims doubled in 2017 we would not detect that. In our case the magnitudes did not change significantly.

There are a number of smaller trends which seem detectable, although they may not be statistically significant, like the *SOUTHEAST* increasing in proportion of shooting victims, or a drop in the percentage of deceased who were *UNARMED* in 2020. But one trend stands out above all others: the proportion of deceased racially identified as "Other" went from a minority to a significant majority. We discuss this more in our conclusion, but it seems likely that this is a data artifact, not a true societal trend.

Other Analyses:

Figure 8 is a Violin plot which shows the relationship between *age* and *weapon*. These were good to look at for due diligence, but there's no obvious trend to note.

Interpretation / Interesting Notes:

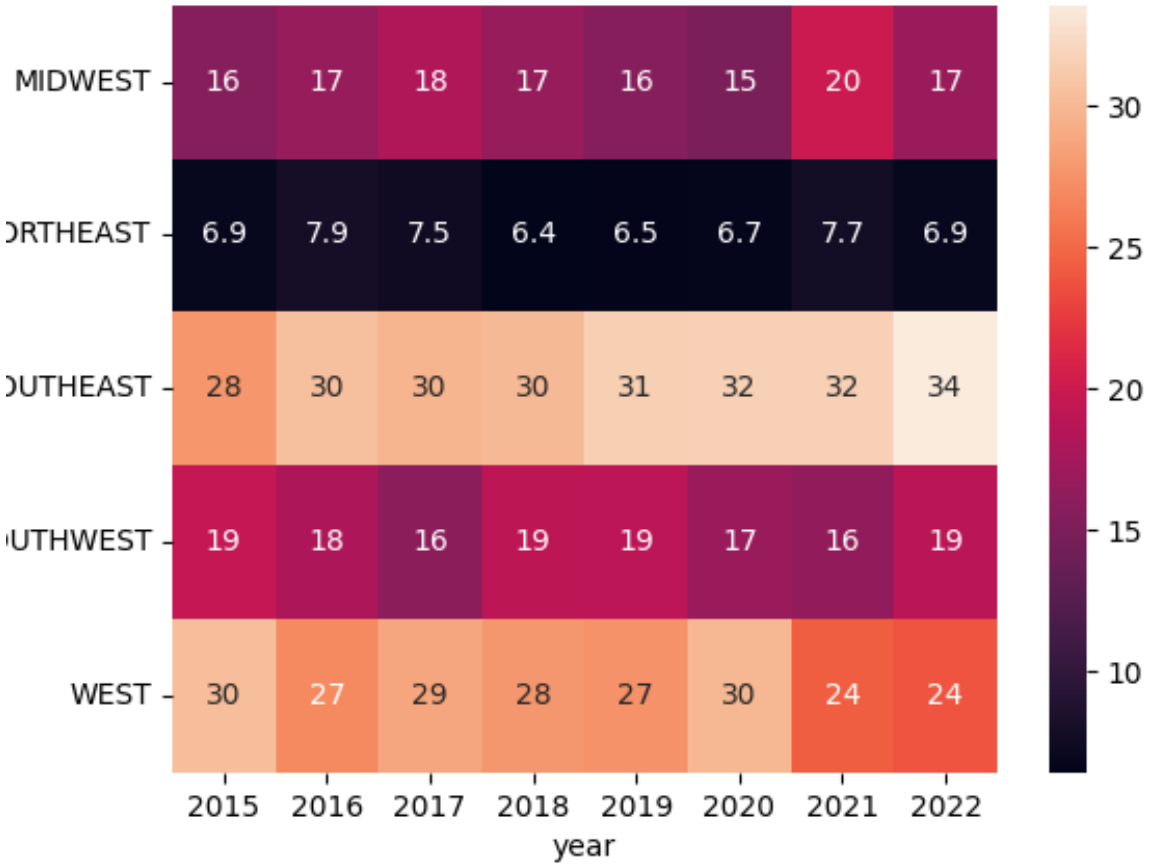

Figure 6: Count by Region over Time

The following trends were noted from observing the data, but do not have the statistical backing to draw any real conclusions.

- The age of the shooting victim appears to vary by race. Further analysis could be done to determine how much age affects the likelihood of being killed by race.
- Fewer of the deceased exhibited signs of mental illness than those that didn't, but one would need more information about the classification methodology and rates of mental illness in the general population to draw conclusions.
- The vast majority of encounters did not occur with a body camera filming. Futher analysis could determine how this trend changes as adoption rates increase.
- Significantly more people killed by police were armed (in some fashion) than unarmed.
- The proportion of deceased racially identified as "Other" increased from about 3% in 2015 to over 60% in 2022. This change seems too large to reflect social changes and instead points to a change in data collection methodology, or perhaps a concatenation of discrete datasets.

Broadly speaking it has been very difficult to draw societal conclusions from this dataset. Proper interpretation relies on deeply understanding the collection methodology and coupling this data with

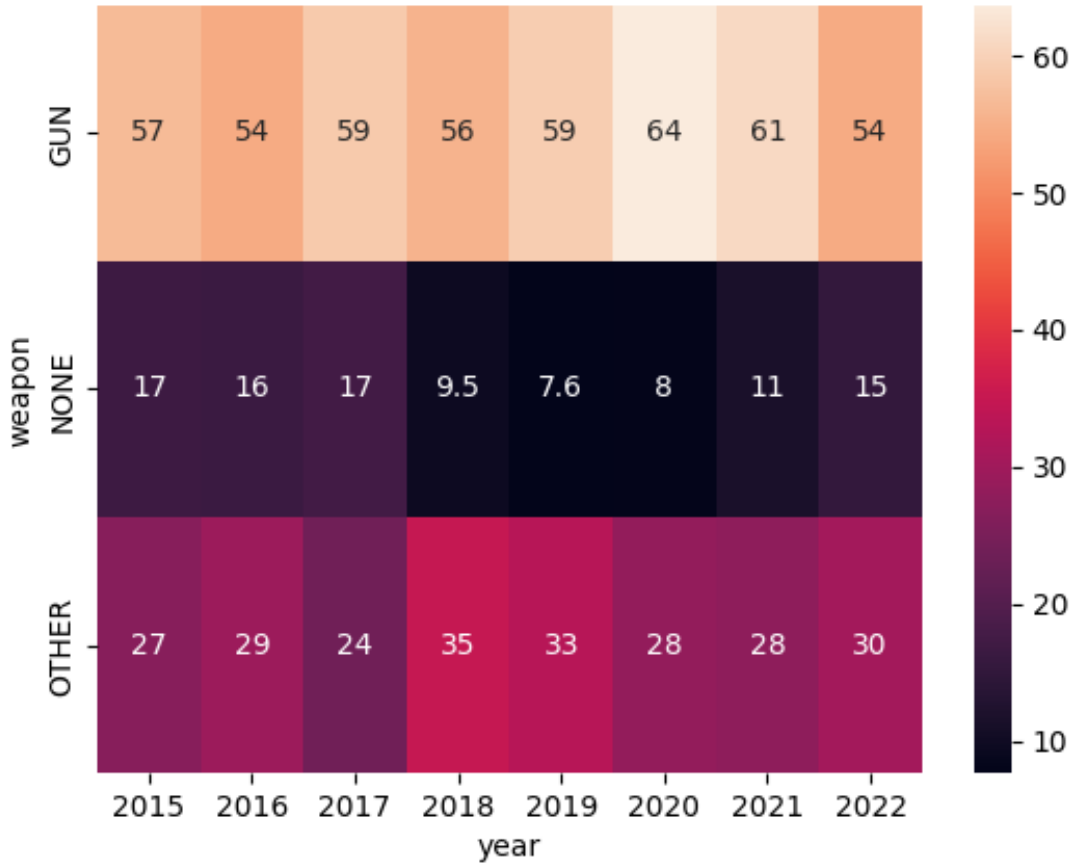

Figure 7: Count by Weapon over Time

other information from, e.g., local census data. In a sense that fulfills my original goal in analyzing this dataset: self-education about the complexity of analysis of this sort. All this is not to say that there isn't a clear conclusion to be drawn; rather, it is important to note how carefully and critically one must proceed to arrive at that conclusion.

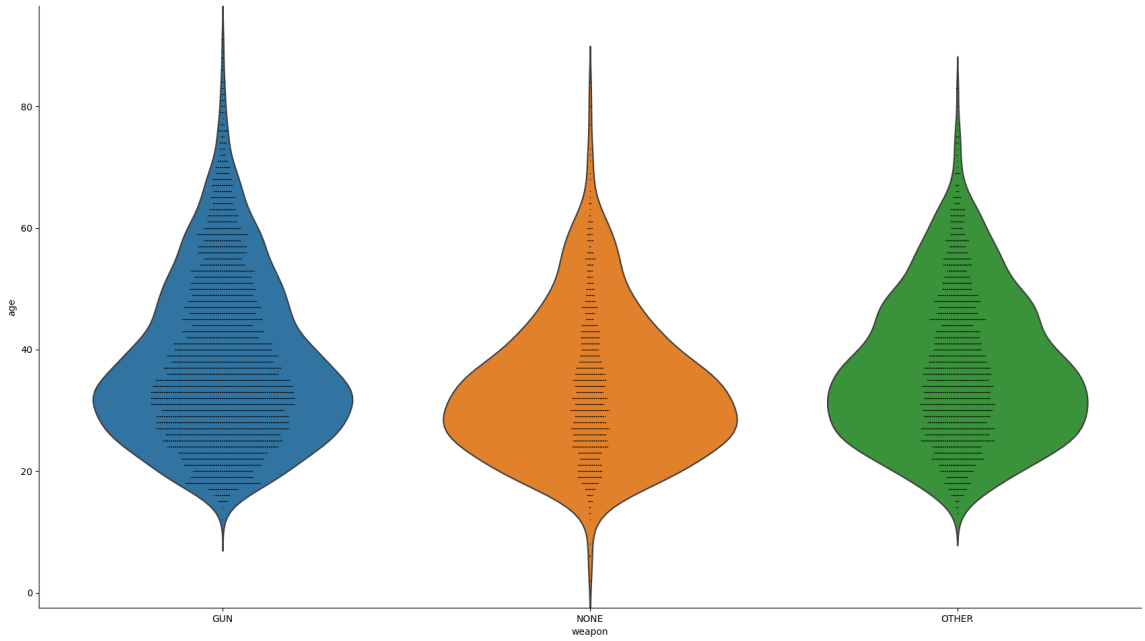

Figure 8: Age vs. Weapon