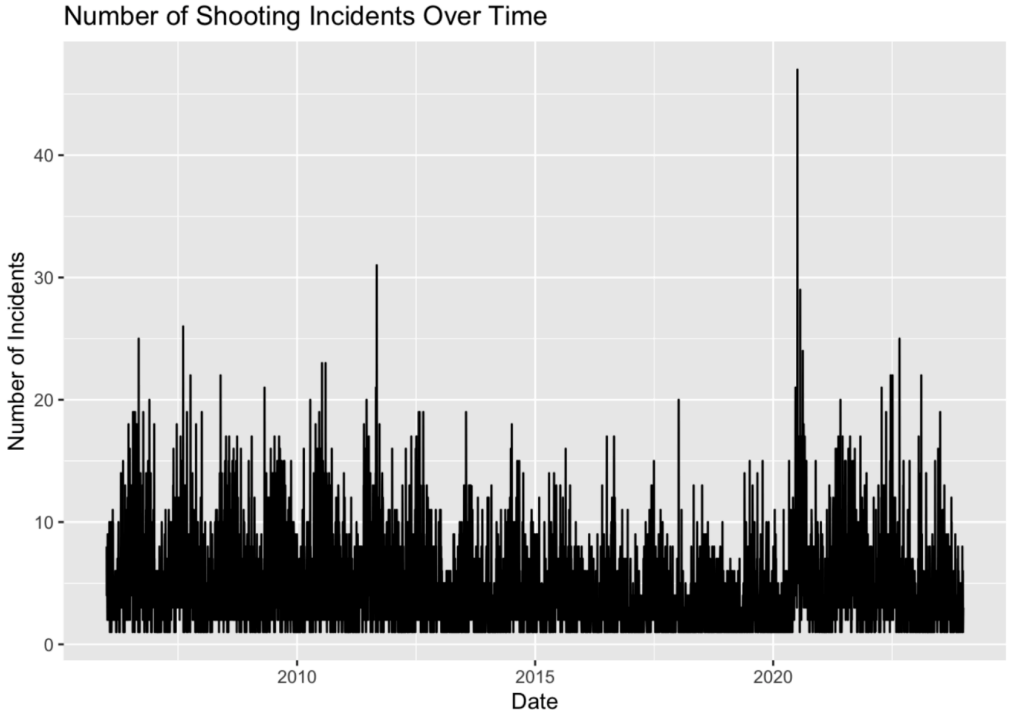
Code link: https://github.com/okaheel/DTSA-5304-final

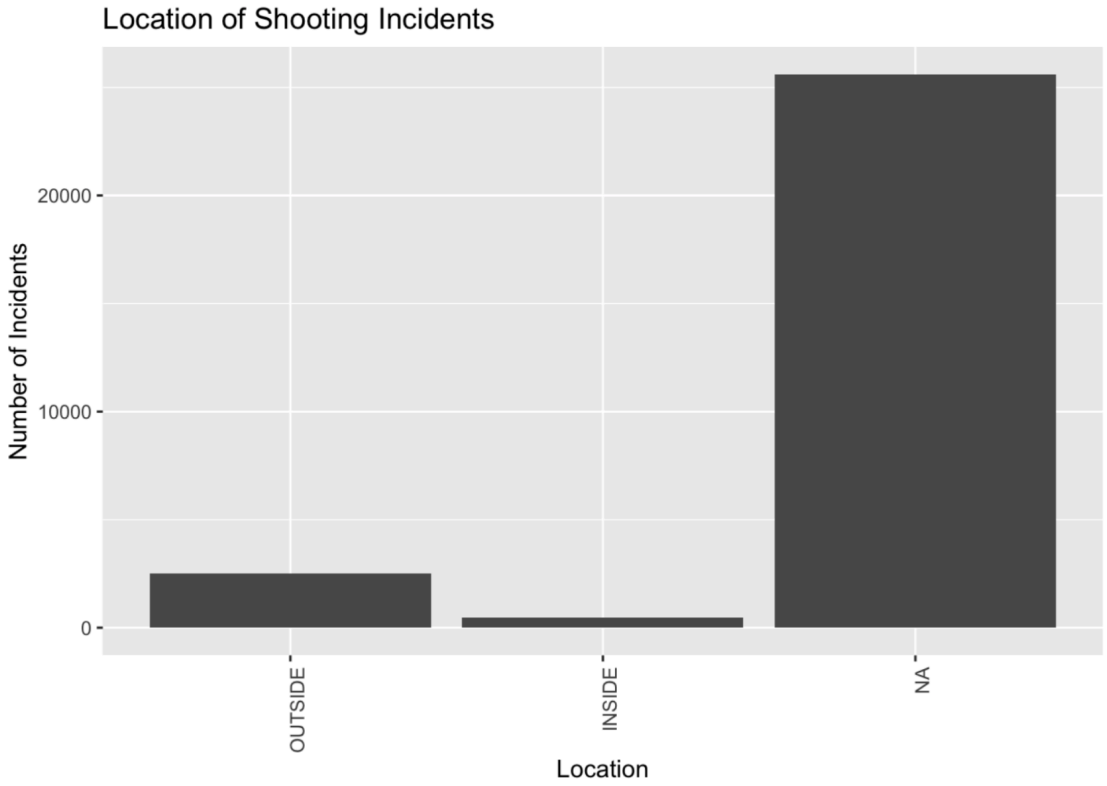
This dataset includes a breakdown of all recorded shooting incidents that occurred in NYC's boroughs. This data was published by the City of New York on the GSA's data.gov. The dataset includes time information about the incident, location descriptions along with boroughs, information about the perps if known, and the victims With this dataset I am primarily interested in a couple of items.

The 4 visualizations I created in the first pass:

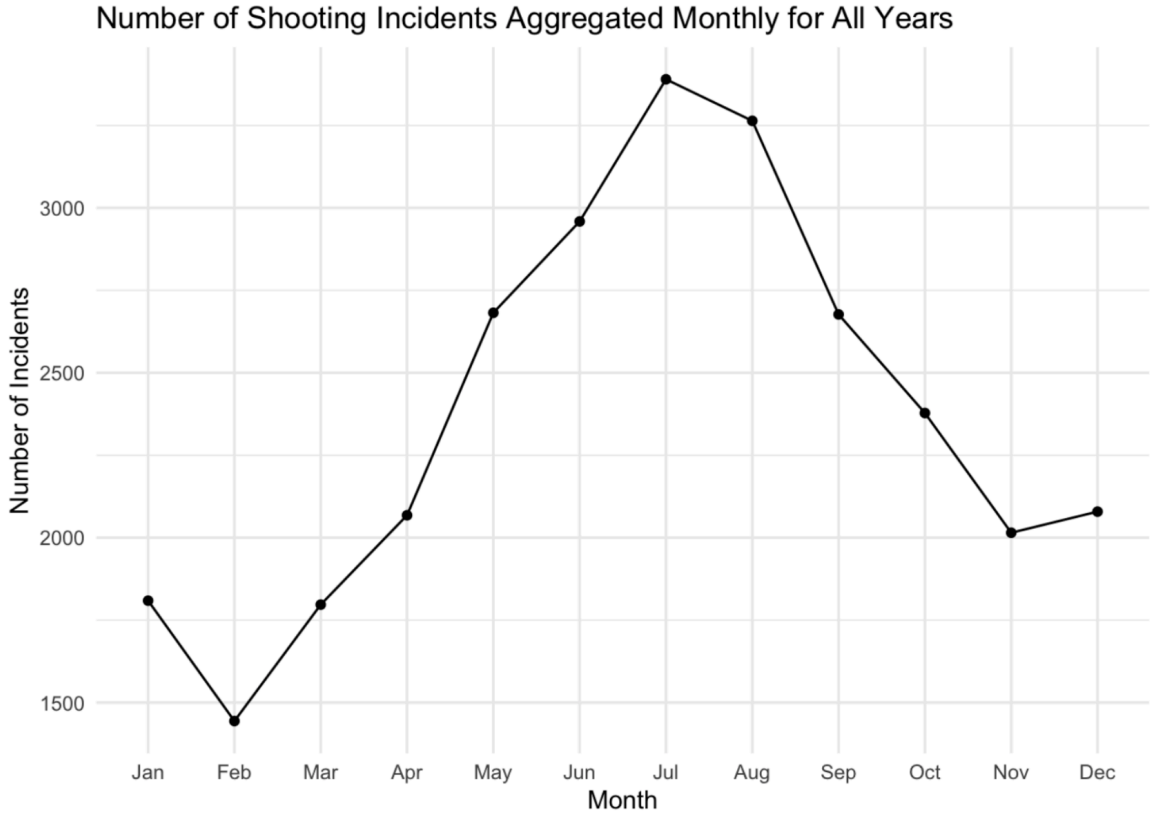
- A visualization to look at the time series counts of incidents:



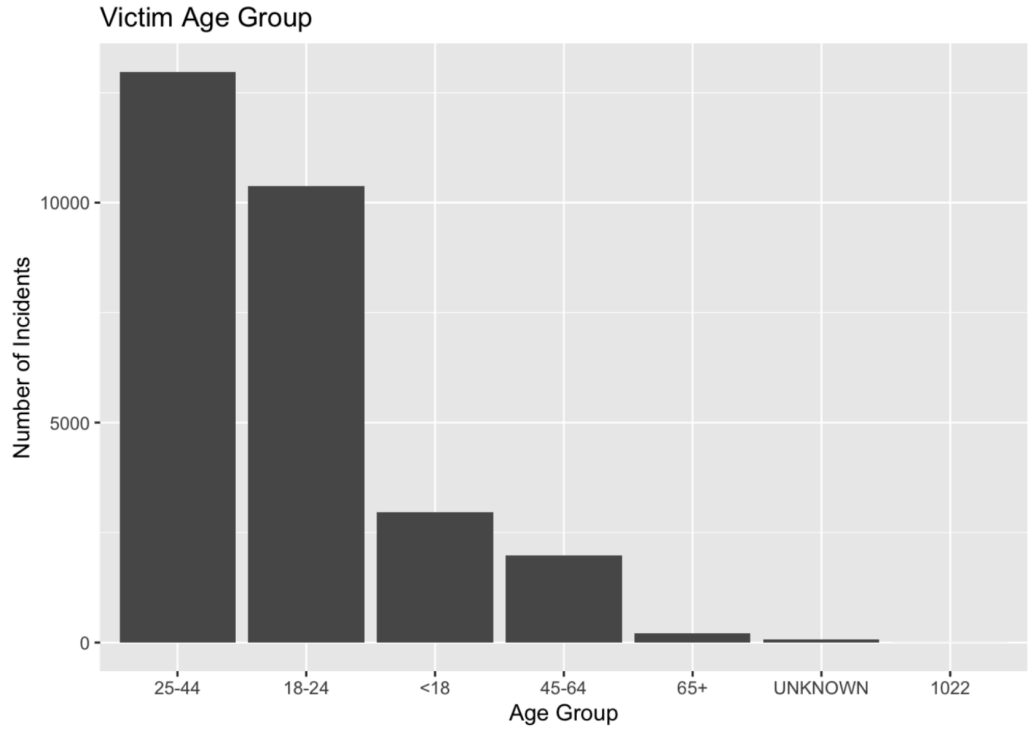
- A visualization to look at the location of the incidents (inside/outside)



- A visualization that looked at the overall number of incidents by month for all the years



- A bar chart visualizing the distribution of victim age groups



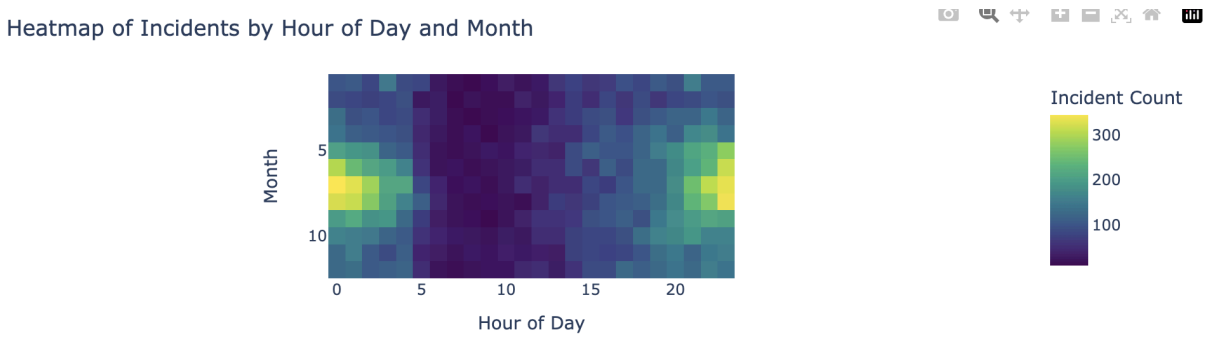
My first pass was in my mind an opportunity to look at the overall summaries to show relevant demographics. I used them as a way to structure the interview as an interactive working session. The people I talked to and got feedback from are data scientists/software engineers in the finance field so they had some very interesting insights. We spent some time talking more about the dataset itself rather than what they wanted to learn in the initial pass. These initial charts sort of answered some of those types of questions that they had like “Oh what’s the most common age group to be a perp or victim”, “Is there a seasonal trend”, “What’s the month with the most amount of shootings”, “do most of these happen inside or outside”.

They were great initial discussing but actually led to a very interesting point where we each reached the thought of, okay what data do we need to be able to make “actionable” insights not just insights? These things that we as data scientists can use to come up with recommendations, patterns, or things to look for that relate to these shootings and can possibly be used to prevent or even just understand these incidents.

And we used the discussion to work on building a new design. It was a great working group format. The design that we created involves fixes for some of the critiques that we had of the initial pass of charts.

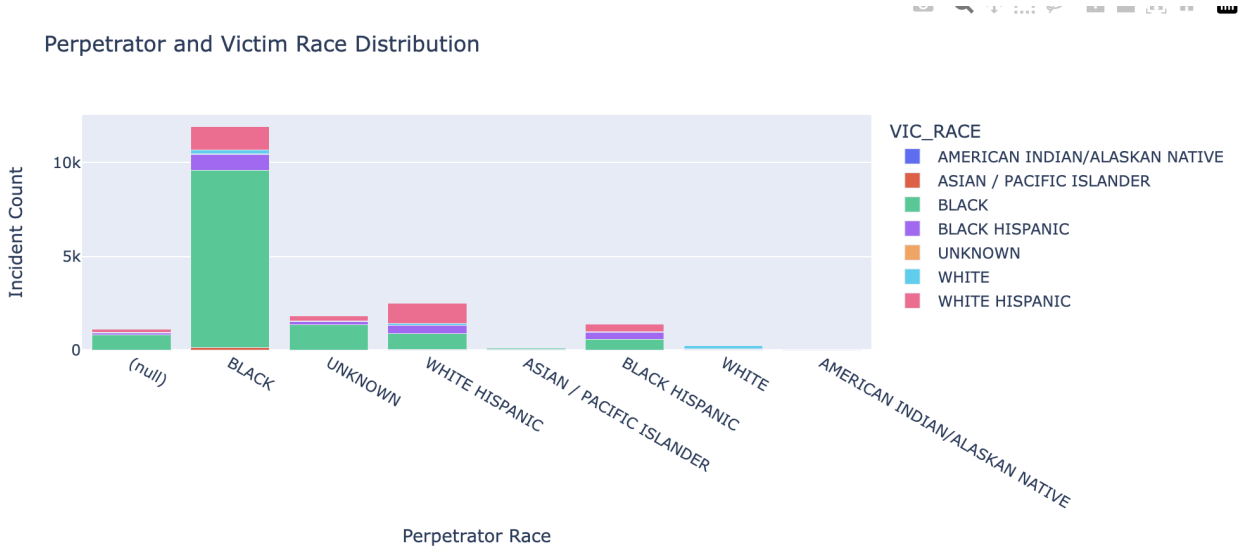
The first big critique was that the breakdown but month is helpful but it is not something easy to try to prevent crime in a certain set of months (or at least we hypothesize), we thought it might be more valuable to see if certain hours of given months are more problematic. For example maybe 9PM -12 AM in winter months so in this second pass I came up with this month-hour

chart:



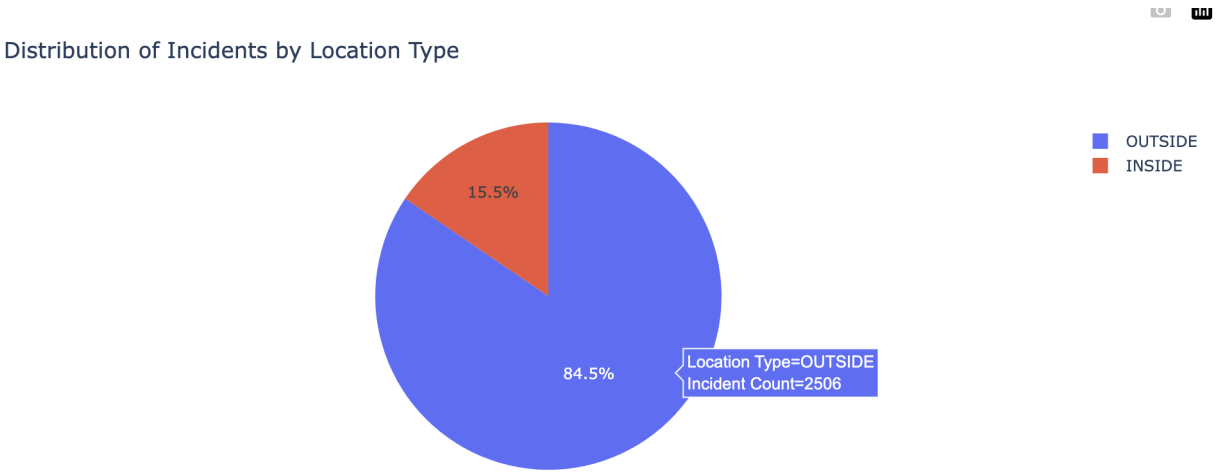
And actually, this chart shows us something super cool. Which is that a very large amount of the incidents that happen do happen around the hours of 9 PM - 2 or 3 AM of Summer months. This is not something that would be easy to discern from just a bar or a line chart.

The next interesting thing was one of the subjects suggested that we look into how race plays into it. I had made an initial low-fidelity chart with just Victim and perp race but they didn't overlay so each one was just a distribution chart on its own and that led to a not-so-unexpected but intriguing conversation relating to population demographics and race on race crime. To resolve that conversation I made this chart to look at the perp and victim distribution and sort of unexpectedly it turned out that the largest number of incidents was Black Perp and Black Vic incidents.



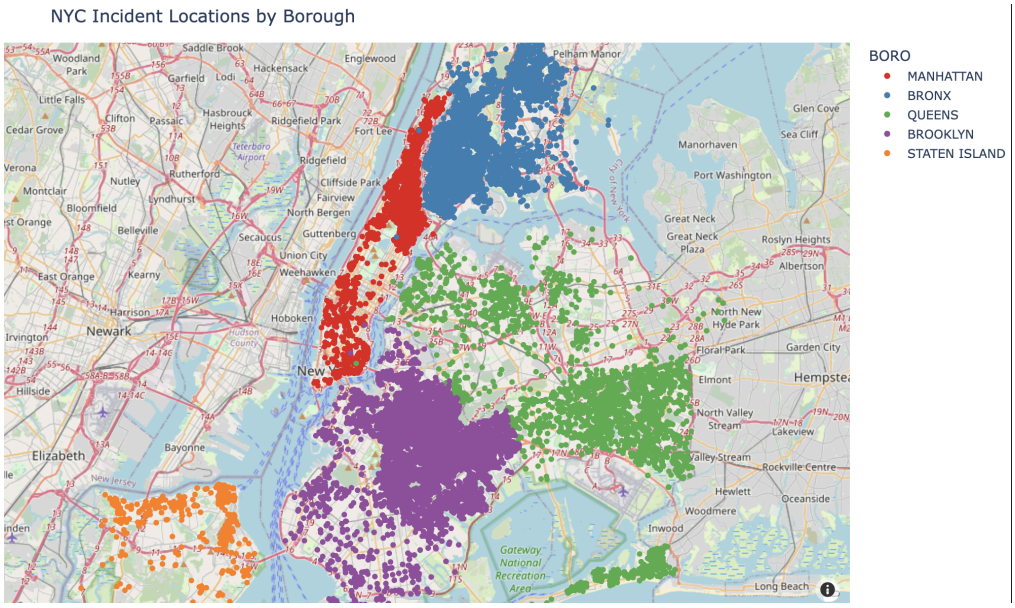
This looked very interesting and led me to want to look at two more things. I had wanted to look at the location and inside vs outside distribution of incidents.

So I started with a bar chart comparing the number of inside vs outside incidents and outside outweighed the inside by a lot to the point that a bar chart was really meaningless so I pivoted to this pie chart which helps illustrate it a lot more

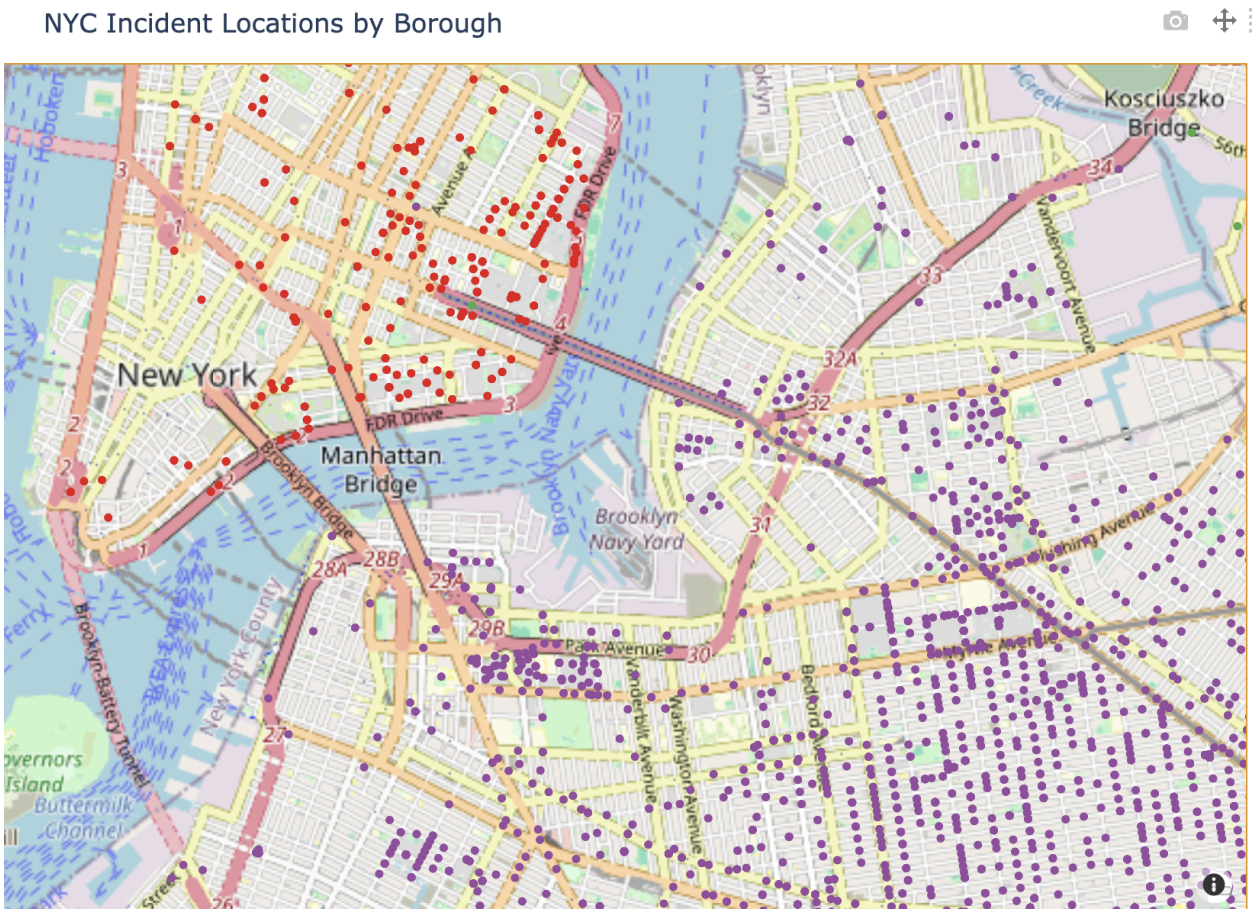


This helps understand how much the different is significantly. Bar charts are nice for categorical variables sometimes and they make sub-categories sometimes easy to understand but for something like this once of the interviewees recommended a pie instead and that was a great recommendation.

Finally, one last thing I wanted to look at was the breakdown of incidents by location (Boro), so I had initially used a bar and pie and that had worked fine but was still not very telling of anything. So I tested something a little different:



I tested a couple of these, and made a few different iterations. This is great for looking at in detail where exactly (by street or such) these incidents happen. I really some part of this chart and hate some others.



On a zoomed-in scale, it illustrates really well where there are clusters around certain blocks or locations or neighborhoods. But on the other hand, a person intuitively on the zoomed-out scale can't really tell the density/real relative frequency in one area compared to the other because the dots turn into blobs. Two approaches we discussed that seemed viable were - First, Creating a more aggregate view as opposed to showing all the individual incidents since people are not very good at distinguishing very large numbers like this (so show things in the form of a colored plot that is grouped on a Boro level) but that would require some code to handle Boro polygons which I didn't feel the need to implement. The other one would be to create bigger markets that represent clusters of small markers to act as a sort of heatmap and that could be helpful for this.

Overall this experience was awesome and I learned a lot about how different people think about various datasets and representations. It was also cool to get feedback and build it into some charts in real-time since they were relatively easy to implement.