Introduction

This dataset goes over dog bites in New York City (NYC), New York from the end of 2014 to 2021. It has multiple columns:

Unique ID: Each bite incident has a unique ID associated with it.
DateOfBite: Each bite incident has a date for when the incident happened.
Species: Each bite incident specifies that the biter was a dog. This is a redundant column since this dataset is already specific only to dogs.
Breed: Each bite incident specifies the breed of the dog that bit, the majority of breeds are categorized as “unknown”.
Age: Each bite incident specifies the age of the dog.
Gender: Each bite incident specifies the gender of the dog, including an “unknown” parameter.
SpayNeuter: Each bite incident specifies a boolean value on whether or not the dog was spayed or neutered when the bite took place.
Borough: Each bite incident specifies the NY Borough that the bite took place in.
ZipCode: Each bite incident specifies the zip code of the location the bite took place in.

Methodology

I decided to create geo maps based mostly off of the zip code of each bite incident, since I felt it would be more accurate and helpful than the Boroughs, which are much larger. I was concerned data would be lost by defining bites by Boroughs, as well as it being much more difficult to find a way to visualize data based on Boroughs than by zip code.

Pretty quickly, I found that creating maps by zip code was going to be difficult, and that I’d need a ‘geojson’ file for NYC zip codes to feed through when making maps as the ‘geojson’ parameter to help “zoom” into NYC and map the zip codes. This proved to be really effective after just a few techniques to clean up my zip codes to fit the geojson.

The dataset has a large amount of missing/null data, as well as lots of ‘UNKNOWN’ data. For the missing breeds, I filled the missing data with ‘UNKNOWN’ to match what was already there. I filled missing age values with the Age columns’ median. I also filled missing zip codes with ‘UNKNOWN’ to match the dataset’s trend. The breed column also had known mixed breeds, like an American Pit Bull Terrier/Pit Bull mix, but also many ‘mixed breed’ or ‘mixed’ as rows in the column. I corrected these by grouping any undefined mixed breed as ‘UNKNOWN’ as a way to only retain known breeds, either single-breed dogs, or known mixed breed dogs.

I made two maps; a choropleth map and a bubble map. The choropleth map shows bite count density by zip code and uses the geojson to map the zip codes. I played around with several color scales and ended up using the Yellow to Red color scale because I liked color scales with two colors only, and yellow to red seemed to match the dataset, where more incidents was

‘worse’. Other color scales that went into a cool color with frequent incidents didn’t feel accurate to the data I was presenting; they gave off a feeling that more was good.

I decided to do a bubble map because I found it really difficult to do any other kind of geomap with the data I had. I tried to make density maps of each breed per bite around NYC, but it never ended up looking right, especially since the most common breed for each bite incident was ‘UNKNOWN’, which makes visualizing the breeds difficult.

The bubble map mostly reflects the choropleth, but does add a new element where bite count is shown as bubble size instead of just as a gradient. I did end up adding a gradient to add more information, but I really wanted to “group” the bubbles to make it easier to read through the data, but found that that would’ve been very difficult and time-consuming without being all that worth it.

Had this been a full EDA project, I would’ve love to explore into more parts of the dataset including spayed & neutered dog bites vs. non-spayed and non-neutered, to see if those affect aggressiveness in dogs, or see if age or gender affects a dog’s aggressiveness, etc. I crumbled and did end up including at least a small bar graph showing the top ten most common dog bite breeds, making sure to exclude unknown as a breed.

Results

The bar plot shows the most common breed in these incidents is by far the Pitbull, and even so, the 5th and 6th most common breeds were pit bull mixes. Together, these three breeds would make up 4,995 of the 16,798 incidents with known dog breeds, or roughly 30% of dog bites. This seemed to be a significant portion of dog bites by pit bulls.

The choropleth and bubble maps both show the zip codes with the highest bite occurring in zip code 10029 right next to City Park. The other zip codes around City Park seem to have an increased amount of bites compared to the rest of NYC, with the next highest being in zip codes 11368 and 11208.

Conclusion

From the data set, I mostly found that the highest number of bites come from around City Park, and the next highest are in Brooklyn and some other seemingly random areas. Additionally, the dataset revealed pit bulls bite people in NYC more than any other breed.