Cremieux 6/29/2024

View in browser

The Dogs of New York

Analysis of the Department of Health and Mental Hygiene Dog Datasets Shows Pit Bulls Like To Bite

CREMIEUX
JAN 3 • PAID

I prepared the plots for this before writing it and then wrote the piece in less than one hour. Because of how much preparation I did for this specifically rather than incidentally, I don't consider it fair to add to my timed rapid writing series.

Dog may be man's best friend, but breeds differ and experience tells us they're not all equally his pal. Who's the best? Who's the worst? To start

answering this question, I've opened up the New York City Department of Health and Mental Hygiene's (DOHMH) dog licensing and dog bite datasets.

The Data

The Dog Licensing Dataset. This post uses the latest version of the DOHMH Dog Licensing Dataset, updated as of February 6, 2024. Dog owners in New York City are required by law to register their dogs, whereafter the information they registered is placed into this dataset. Dogs are recorded with a unique ID, the name of the dog, its birth year, sex, the zip code of the owner, the breed, and so on. The dataset presently covers licenses issued between September 2014 and November 2023.

The Dog Bite Dataset. This post uses the latest version of the DOHMH Dog Bite Dataset, updated as of February 21, 2024. New York City Health Code Section 11.03 requires reporting animal bites within 24 hours of the event occurring. The DOHMH collects information on breed, sex, age, spayed or neutered status, and so on. The dataset presently covers dog bites that occurred between January 2015 and December 2022.

Dog (Data) Handling

One of the most annoying aspects of working with registrations where dog owners can pick the breed name is that some dog owners are persnickety. They want to tell you their dog's very unique breed. It's not a Cavalier King Charles Spaniel, it's an award-winning Cavalier King Charles Spaniel MX (unclear what this means) with a great-grandparent that was a Bichon Frise. Did you write that down? Great, now it's a one-off entry in the dataset, and they put it down as "CCKS" instead.

There are over 600,000 dogs registered in the Dog Licensing Dataset and over 26,000 in the Dog Bites Dataset. I will not be going through by hand and reclassifying everyone's more or less obviously registered dog breeds, since the vast majority of people used sensible names. Instead, I will dump any dog

breed listing that is entirely unique and not identified with a simple grep into the "Other" category. This has little effect on the results, but feel free to download the data and play around with different permutations of it to see if you can find one where it matters.

First, I did what I'm dubbing the 'luxury' classification. This is where I went through the Dog Licensing Dataset and grep'd all the breeds with decent numbers of dogs. After I did a first run of this, I found lots of sensibly-defined, common breeds, threw the rest in "Other", and then I inspected "Other" again to see how many of those had decent numbers of dogs in them. I ended up with hundreds of breeds, but I noticed there were lots that I didn't think needed to be kept separate, such as "Miniature Schnauzer", "Standard Schnauzer", "Giant Schnauzer", and "Schnauzer". So, I grouped those into "Schnauzer" and did similarly with groupings like

{Malti-Poo, Maltipoo, MALTIPOO (with an extra space)}

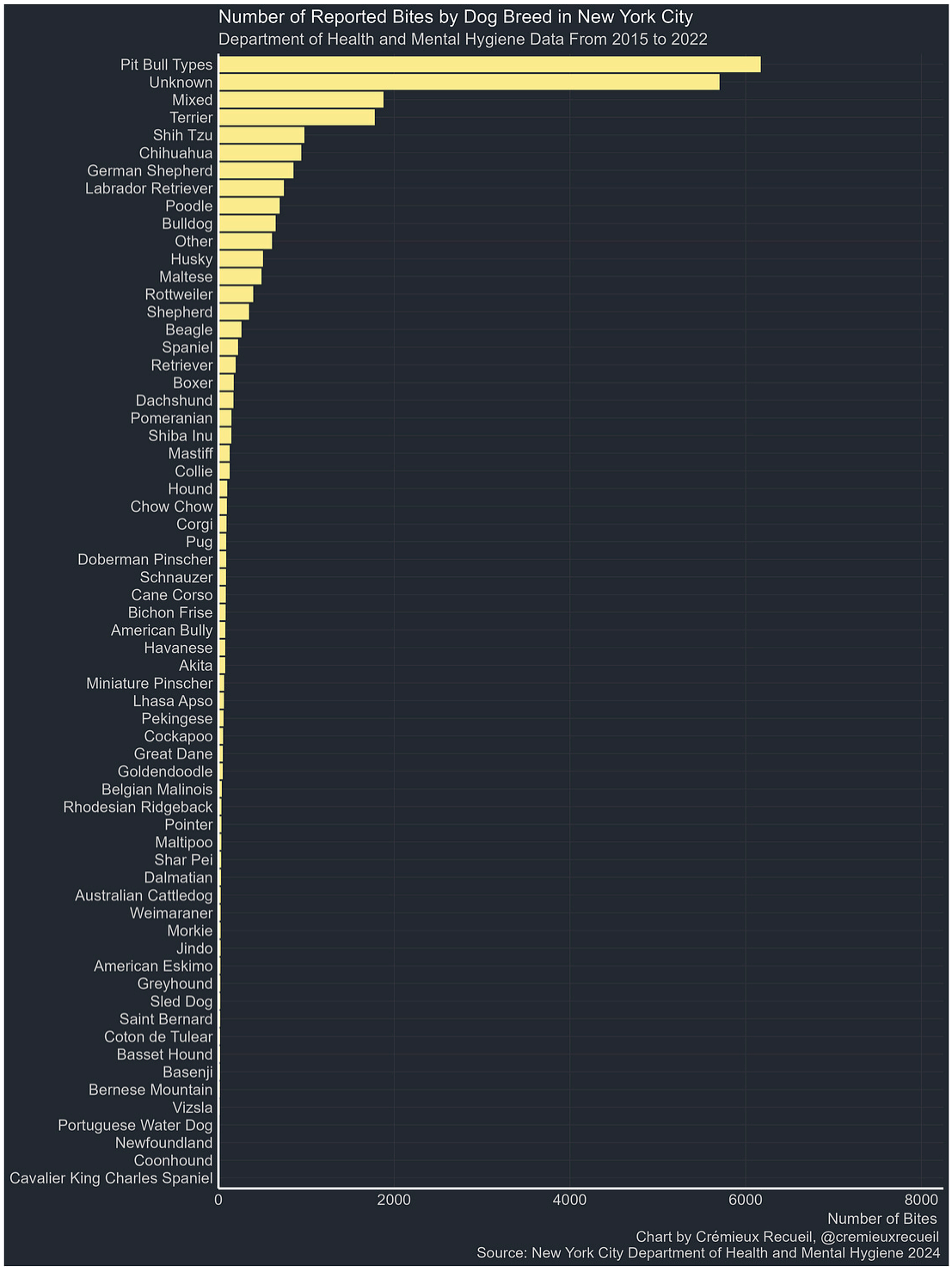
{(Mastiff, French), (Mastiff, Neapolitan), (Mastiff, English), (Mastiff, Tibetan), (Mastiff, Bull), Mastiff, Neapolitan Mastiff, Tibetan Mastiff, Bullmastiff}

{Vizsla, Wirehaired Vizsla}, etc.

Breeds with fewer than ten entries after consolidation were classed as "Other". Those classified as "Pit Bull Mix", "Afghan Hound Crossbreed", etc., were classed with Pit Bulls, Afghan Hounds, etc., so the remaining "Mixed" category was just for nonspecific mixes. I moved some mixed categories like "Hound Mix" and "Hound Crossbreed" into a more general "Hound" category. Lots of obvious misspellings were afoot as well, and I simple classed those to the correct spellings. For example, "Schipperke" and "Schipperkee" became "Schipperke", while the "Pharaoh Hound" and the "Pharoh hound" became the "Pharaoh Hound".

There's ultimately a lot of judgment calls here, so if you have criticisms, go ahead and tell me. The complicated part of this was decidedly *not* specifying the queries, but how to do the classifications.[1] You can see the results of using this information in the ***Luxury Dog Information*** section below.

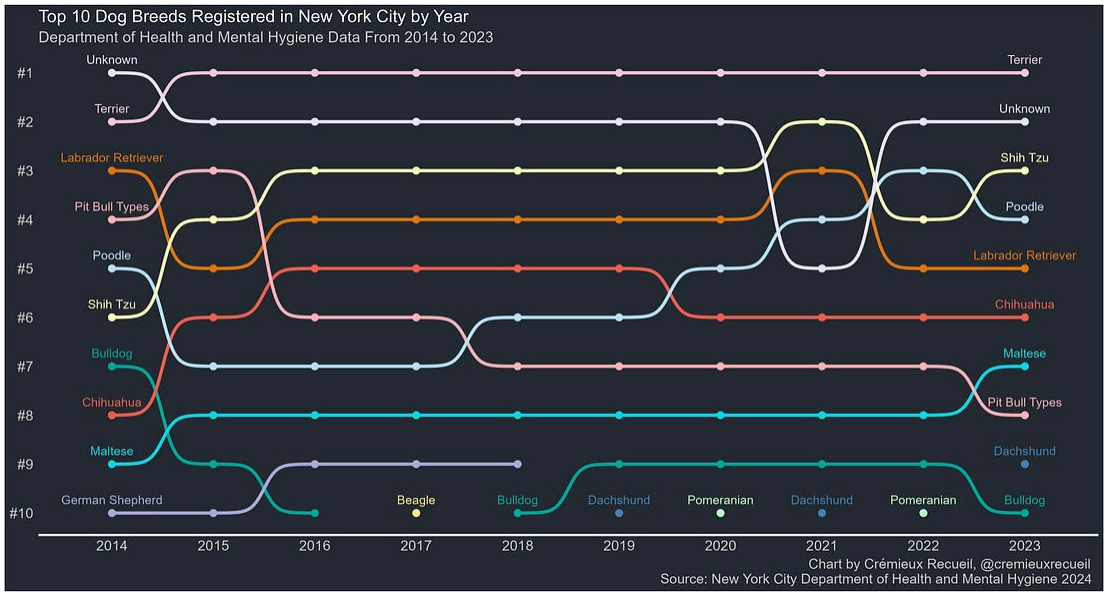
Next, I classified the breeds in the Dog Bite Dataset, and then I back-classified the breeds in the Dog Licensing Dataset accordingly, because the dogs listed in the former are needed for the Dog Bite analyses, and they are not labeled with as much granularity as the dogs in the Dog Licensing Dataset. The Dog Bite frequencies by breed are listed here, with zeros removed.



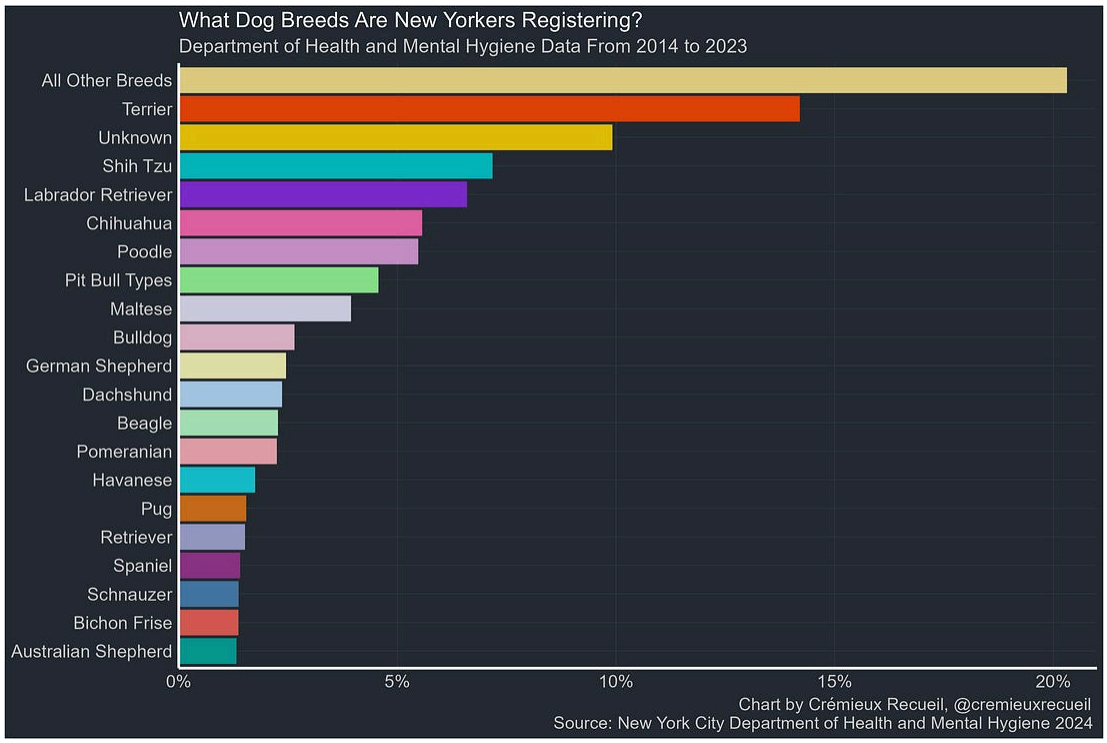
The problem for this data is that many breeds haven't been reporting attacking anyone, and there are a lot of unknowns.

Luxury Dog Information

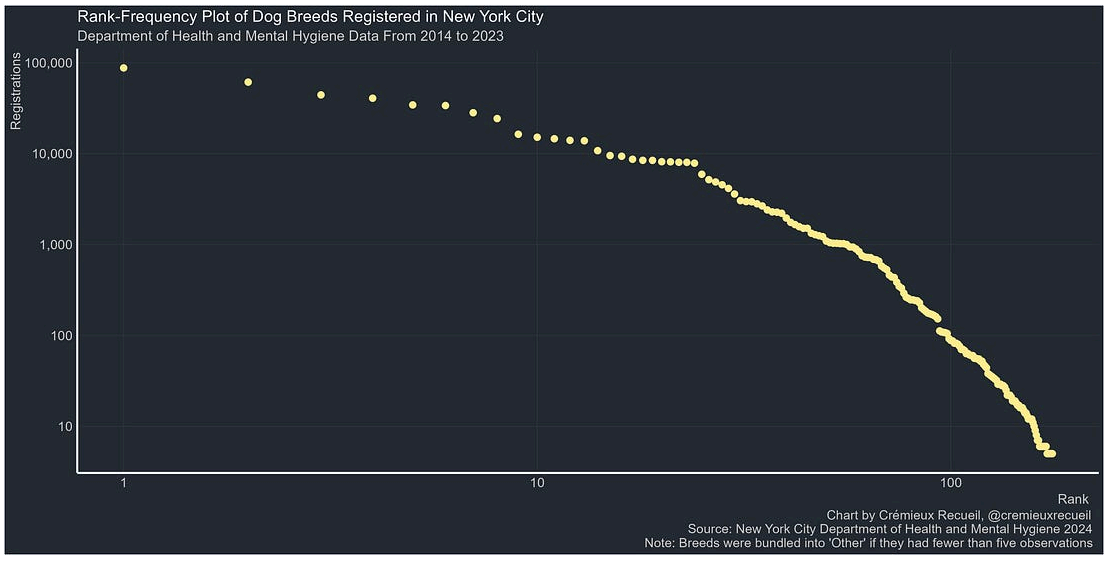
Visualizing New York City's top ten registrations by year, we see that Unknowns are pretty common, but Terriers are the top dog. Pit Bull Types also started off in fourth place and dropped to eighth by 2023. Aside from tenth place, there was a good amount of consistency over the years.



A few breeds decisively dominate the registrations, with about 70% being known breeds in the top-20, Unknowns constituting about 10%, and All Other Breeds making up the final 20% or so.



One question is: Does this fit a power law-like distribution? Maybe this looks like Zipf's Law. Cutting off the breed counts at five, the rank-frequency plot looks like so:

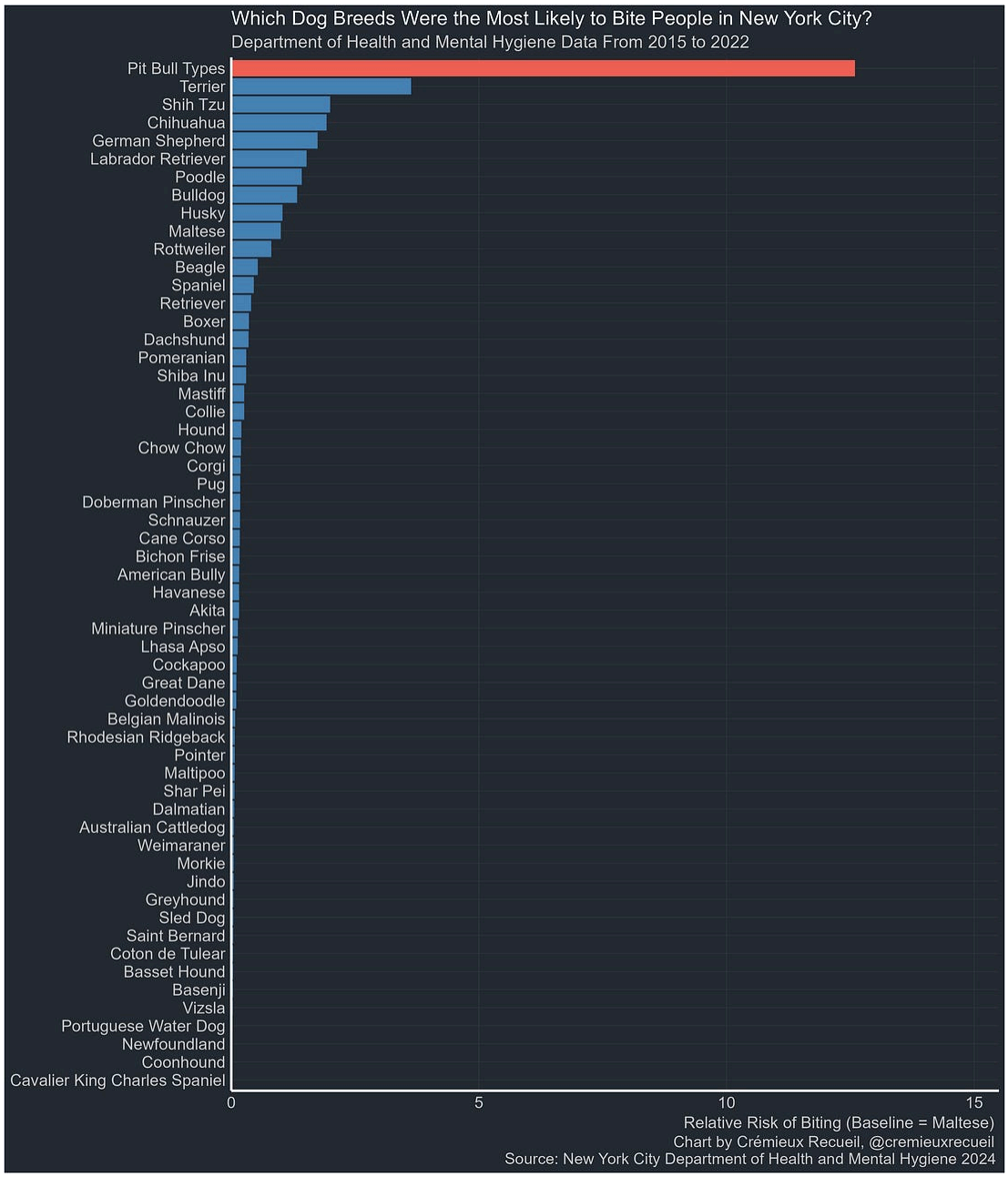


Not quite Zipf.[2]

Which Dogs Bite?

To determine which dogs are the most likely to bite people, we need a baseline. For that, I could use the Maltese Dog; you know, these guys:

Risk is \frac{number of bites}{number of dogs}, and the relative risk is \frac{Risk for Breed X}{Risk for the Maltese Dog}. There are strategies for dealing with zeros in relative risk calculations; to avoid any dubiousness that comes with smoothing decisions, I'll just go ahead with the available breeds in the Dog Bite Dataset, noting that this does lead to their risks being underestimated absolutely and that some breeds are sorted into "Other" in the Dog Bite and not the Dog Licensing data, so we're dropping it and the unclear categories of "Mixed" and "Unknown". These exclusions are good for improving the connection between the datasets and curbing noisy estimates. The baseline is the Maltese:[3]



Identification Questions

Breed Identification. Some people worry that the breed identification in the Dog Bite Dataset is bad because the people doing the reporting are not all dog identification experts. Identifying common types of dog by breed is straightforward, and identifying pit bull types—including pit mixes—is especially so. It does not take a genius to go through randomly-selected American shelter websites to see the problem with pit bulls and pit mixes, and the fact that people are averse to the breed. People discriminate against

pit bulls with good reason and good capability: they are recognizable, and their behavior is understood by many people to be as bad as they were bred to be, which is to say *very bad*.

But some people don't like the different means we have for identifying pit bulls, they argue about admixture and edge cases, some argue (wrongly) that pit bulls aren't a legitimate classification, and they don't treat the identification of pit bulls versus other breeds of dogs in a serious manner, seemingly because they're not looking for an answer.

To address identification problems, in a future post, I will be discussing Swiss data. I've been meaning to get around to that for a while, but the post has instead been languishing half-written in my drafts. The identification by breed is far better in Switzerland. Switzerland requires registration and has surveys where they count up dogs in different cantons and experts classify them, sometimes with genetic testing. The situation with respect to identification in Switzerland is relatively excellent compared to anywhere else I've seen, but—and here's the spoiler—*pit bulls are still extremely overrepresented among the perpetrators of bites and other dog attacks in Switzerland, where they're identified comparatively well.*

Until that time, enjoy the post, leave suggestions, do your own breed classifications and re-run the analyses, and if you're looking for a good dog, I suggest buying one that's from a good breed, not a pit bull.

Compliance. Some people do not register their dogs. This is more likely in impoverished and non-Asian minority communities.[4]

Misnaming. When people seek out a license for their dog, there are undoubtedly many who register a breed their dog actually isn't, or as a pure breed when their dog is mixed, unknowingly or knowingly. There are also individuals who misregister so as to be able to have their dog in public housing, with a restrictive landlord, and so on.

Robustness. Theoretically, the mislabeling of dogs in the Dog Licensing Dataset could be a big issue for computing their relative risk of attack. For that reason, I'll outline a scenario that shows pit bull bite disproportionality

is still extreme, even if we fudge the numbers based on these sources of error.

Scenario: All breeds are under-registered by 5% due to noncompliance, over-registered by 10% due to misidentification when they *should* be registered as pit bull types, and fully a tenth of the pit bulls reported biting are actually the other dogs. To make this even more favorable to pit bulls, we'll say that a tenth of all the dog bites are undocumented and undistributed, but at least pit bulls are responsible for zero of them. **The relative risk of an attacking dog being a pit bull remains highly elevated.** The "ground-truth" relative risk was 12.59 times the Maltese' risk; with these adjustments, it's 11.11 times.

To make pit bulls as apparently friendly as the Maltese, we would need there to be around 350,000 of them in New York City. In the ground truth case, we would need the pit bull population to number a bit over 4% of New York's human population.

Pit bulls are at an extremely elevated risk of causing harm to humans.[5]

---

1 Luxury Classifications:

DogLicense$LuxuryClassification	n	percent
Affenpinscher	176	2.853021e-04
Akita	753	1.220639e-03
American Bully	1656	2.684433e-03
American Eskimo	1330	2.155976e-03
American Water Spaniel	19	3.079966e-05
Anatolian Shepherd Dog	385	6.240983e-04
Australian Cattledog	3039	4.926324e-03
Australian Kelpie	161	2.609866e-04
Australian Shepherd	8145	1.320333e-02
Baladi	18	2.917862e-05
Barbet	22	3.566276e-05
Basenji	717	1.162282e-03
Basset Hound	1507	2.442899e-03
Beagle	14000	2.269448e-02
Beauceron	34	5.511517e-05
Belgian Griffon	25	4.052586e-05

Belgian Malinois	551	8.931900e-04
Belgian Sheepdog	80	1.296828e-04
Belgian Tervuren	33	5.349414e-05
Berger Picard	37	5.997828e-05
Bernese Mountain Dog	1753	2.841674e-03
Bichon Frise	8424	1.365560e-02
Boerboel	63	1.021252e-04
Bolognese	56	9.077793e-05
Borzoi	82	1.329248e-04
Bouvier Des Flandres	92	1.491352e-04
Boxer	5167	8.375885e-03
Boykin Spaniel	63	1.021252e-04
Bracco Italiano	12	1.945241e-05
Briard	60	9.726207e-05
Brittany	108	1.750717e-04
Brussels Griffon	1282	2.078166e-03
Bulldog	16327	2.646663e-02
Canaan Dog	192	3.112386e-04
Cane Corso	1031	1.671287e-03
Carolina Dog	19	3.079966e-05
Catahoula Leopard Dog	834	1.351943e-03
Cavalier King Charles Spaniel	8019	1.299908e-02
Chesapeake Bay Retriever	109	1.766928e-04
Chihuahua	34345	5.567443e-02
Chinese Crested	246	3.987745e-04
Chinook	19	3.079966e-05
Chow Chow	1000	1.621035e-03
Cirneco dell	16	2.593655e-05
Clumber Spaniel	60	9.726207e-05
Cockapoo	2401	3.892104e-03
Collie	5925	9.604630e-03
Coonhound	2206	3.576002e-03
Corgi	4540	7.359497e-03
Coton de Tulear	1507	2.442899e-03
Curly-Coated Retriever	22	3.566276e-05
Dachshund	14578	2.363144e-02
Dalmatian	461	7.472969e-04
Dalmatian Mix	55	8.915690e-05
Doberman Pinscher	1307	2.118692e-03
Dogo Argentino	152	2.463973e-04
Dogue de Bordeaux	68	1.102303e-04
Dutch Shepherd	228	3.695959e-04
English Springer Spaniel	439	7.116342e-04
English Toy Spaniel	38	6.159931e-05

Entlebucher Mountain Dog	35	5.673621e-05
Eurasier	17	2.755759e-05
Field Spaniel	52	8.429380e-05
Fila Brasileiro	11	1.783138e-05
Finnish Spitz	61	9.888311e-05
Flat-Coated Retriever	184	2.982704e-04
French Spaniel	22	3.566276e-05
German Shepherd	15137	2.453760e-02
German Spitz	70	1.134724e-04
Goldendoodle	7848	1.272188e-02
Great Dane	727	1.178492e-03
Great Pyrenees	936	1.517288e-03
Greater Swiss Mountain Dog	46	7.456759e-05
Greek Shephard	10	1.621035e-05
Greyhound	2266	3.673264e-03
Harrier	48	7.780966e-05
Havanese	10777	1.746989e-02
Hound	1048	1.698844e-03
Hovawart	15	2.431552e-05
Husky	8033	1.302177e-02
Icelandic Sheepdog	16	2.593655e-05
Japanese Chin	242	3.922904e-04
Japanese Chin/Spaniel	255	4.133638e-04
Japanese Spitz	201	3.258279e-04
Jindo	2951	4.783673e-03
Kai Ken	16	2.593655e-05
Keeshond	166	2.690917e-04
Kooikerhondje	32	5.187311e-05
Kuvasz	28	4.538897e-05
Labrador Retriever	40674	6.593396e-02
Lagotto Romagnolo	346	5.608780e-04
Lancashire Heeler	14	2.269448e-05
Leonberger	29	4.701000e-05
Lhasa Apso	2655	4.303847e-03
Lowchen	36	5.835724e-05
Maltese	24306	3.940087e-02
Maltipoo	4143	6.715946e-03
Mastiff	942	1.527015e-03
Miniature American Shepherd	529	8.575273e-04
Miniature Pinscher	3594	5.825998e-03
Mixed	2965	4.806367e-03
Morkie	4878	7.907407e-03
Mountain Feist	52	8.429380e-05
Newfoundland	331	5.365624e-04

Norwegian Buhund	12	1.945241e-05
Nova Scotia Duck Tolling Retriever	170	2.755759e-04
Old English Sheepdog	433	7.019080e-04
Other	681	1.103925e-03
Papillon	1570	2.545024e-03
Pekingese	2284	3.702443e-03
Perro de Presa Canario	29	4.701000e-05
Peruvian Inca Orchid	12	1.945241e-05
Petit Basset Griffon Vendeen	107	1.734507e-04
Pharaoh Hound	173	2.804390e-04
Pit Bull Types	28171	4.566616e-02
Plott	687	1.113651e-03
Pointer	1956	3.170744e-03
Polish Lowland Sheepdog	28	4.538897e-05
Pomeranian	13840	2.243512e-02
Pomsky	893	1.447584e-03
Poodle	33790	5.477476e-02
Portuguese Podengo Pequeno	70	1.134724e-04
Portuguese Water Dog	678	1.099061e-03
Pug	9523	1.543711e-02
Puli	76	1.231986e-04
Pumi	14	2.269448e-05
Pyrenean Shepherd	17	2.755759e-05
Retriever	9352	1.515992e-02
Rhodesian Ridgeback	1091	1.768549e-03
Rottweiler	2803	4.543760e-03
Russian Toy	29	4.701000e-05
Saint Bernard	246	3.987745e-04
Saluki	55	8.915690e-05
Schipperke	291	4.717211e-04
Schnauzer	8439	1.367991e-02
Setter	240	3.890483e-04
Shar Pei	1032	1.672908e-03
Shetland Sheepdog	1249	2.024672e-03
Shiba Inu	8121	1.316442e-02
Shih Tzu	44278	7.177617e-02
Shih-Poo	12	1.945241e-05
Sled Dog	1022	1.656697e-03
Spaniel	8667	1.404951e-02
Spanish Water Dog	44	7.132552e-05
Spinone Italiano	88	1.426510e-04
Sussex Spaniel	27	4.376793e-05
Swedish Vallhund	21	3.404173e-05
Terrier	87619	1.420334e-01

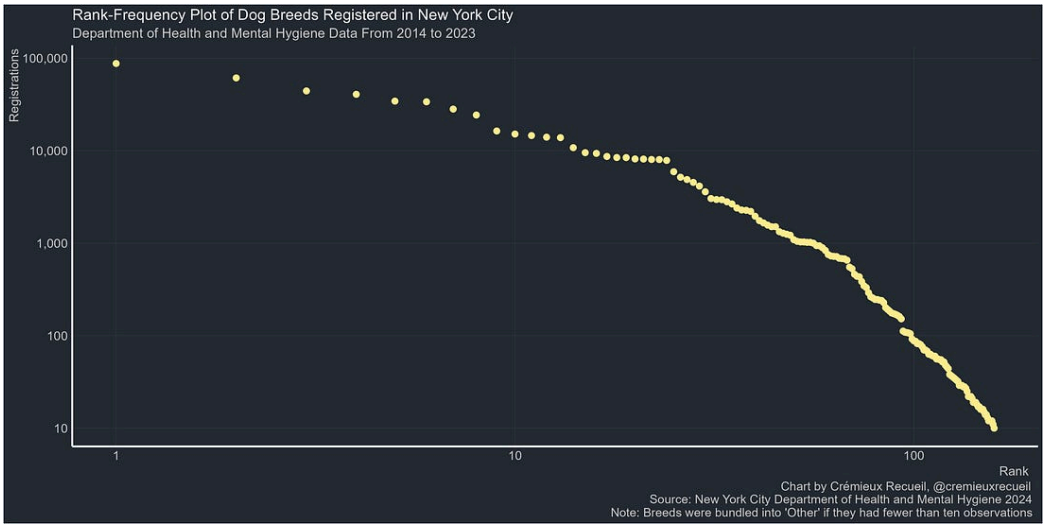
```
                    Thai Ridgeback    56 9.077793e-05
                  Tibetan Spaniel   263 4.263321e-04
                          Unknown 61185 9.918300e-02
                           Vizsla  1021 1.655076e-03
                       Weimaraner   655 1.061778e-03
           Welsh Springer Spaniel   105 1.702086e-04
                          Whippet   720 1.167145e-03
  Wirehaired Pointing Griffon   112 1.815559e-04
                       Xoloitzcuintli    82 1.329248e-04
                           Yorkie Poo    13 2.107345e-05
```

For identifying a “Pit Bull Type”, I used the definition:

```
grepl("Staffordshire Bull Terrier|American Pit Bull
Terrier|APBT|Pit Bull|Pit Bull Mix|American Pit Bull Mix
/ Pit Bull Mix|American Pit Bull Terrier/Pit
Bull|PITBULL|PITBULL MIX|STAFFORDSHIRE MIX", Breed,
ignore.case = TRUE) ~ "Pit Bull Types"
```

Everything else is more obvious.

2 Cutting the “Other” limit at ten, we get a similar picture:



3 Population numbers were assumed to be constant the same throughout the period dog bites were not monitored in for the purposes of

computing relative risk. Dropping the observations outside the dataset increases the pit bull relative risk somewhat, but not meaningfully.

4 As is pit bull ownership. I plan to update this analysis on the zip code level.

5 This denotes them attacking more often. Their attacks are also usually much more severe and they affect a broader age range of people.