Analysis Of Dog Biting Trends in New York

Team:

Alex Moore

Thripura Pakala

Joy Weishan

Introduction:

This project is to demonstrate data cleaning, data analysis and data presentation skills to show the team grasped the concepts provided the past 6 weeks from the Data analysis and Visualization Certification Course.We chose the dog biting dataset because we found a common interest of dogs and all own dogs. We as well hoped this presentation educated others on the importance of proper data collection as well as data visualization techniques.

Project Scope:

The Scope was limited to analysis on a CSV dataset; no more than 10MB of data. A minimum of three research questions were answered including inspiration as we built upon previous topics. The dataset chosen came from Kaggle.com as Kaggle has a wealth of data and previous analysis to build upon. The dataset could come from other sources if needed.. Five visualizations were included, extra visualizations out of scope, yet will be considered. The visualizations included Bar charts, line graphs, pie charts (donut) as well as other useful visualizations to accommodate proper presentation of the distribution, spread, min, max, mean, median and mode of the data. A color pallet was included for consistency across the graphs. Roles and responsibilities were initially defined and presented in the project proposal with modifications as needed throughout the project.

Dataset Cleaning Steps:

Breed Column:

A large problem we had with our dataset was that the data was manually entered. This caused our dog breed column to be over 1600 different distinct breeds. According to the World Canine Organization, they only recognize 360 so obviously there was a problem. When we did a value count on the column we soon figured out why this was the case. A large majority of our columns had multiple misspellings, they had added letters such as x's or -'s, and a large amount of mixed breeds. What we had to do due to the time constraints on this project was drop a lot of rows. We first created a new column so that we would still have the original column but also have a new cleaner column. In this new column we made all letters lowercase because that was another issue in the original column that we had. Multiple different breeds had separated value counts due to there being "Pit Bull", "PIT BULL", "pit bull", etc.. We started with the assumption that we could clean it all but found out it wasn't going to happen. We had started with cleaning any data entries for pit bulls because they were the largest breed count to begin with. So, we found a row containing a version of pit bull and defined them all as pit bulls. We did the same for jack Russ's, poodles, retrievers, and dachshunds. By the time we had gotten here it was already a class and a half of data cleaning, so we decided to turn mixed breeds, and any remaining breed under 50 distinct values to an unknown breed. We then dropped all rows where there were unknown breeds to get us down to the top 16 breeds with the most bites in NYC between 2015 and 2021.

Age:

Upon initial analysis of the age date, it was immediately determined through viewing the age data that there were 11,442 rows out of 22663 rows available for age analysis. The other 11,221 rows had null (no age data). Once reviewing the actual data is was quickly noted the data must have been manually entered. This immediately was a concern for analysis and manually entered data is not always reliable in terms of format as well as increase error probability. Some of the formats included 4Y, 3 MTHS, 2WK, 5M & 3WK, timestamps, etc. There were many age formats, some of which were invalid in the data. The invalid data was due to manual entry. For cleaner data collection in the future, we recommend drop down selections split by year, month, weeks to provide a better opportunity of data reporting. During the data cleaning it was chosen to use regular expressions via google searches and XPERT consultation. The formatting was split into 4 categories; week, month, year and in some cases year and month. Four formats via re.search were setup as a variables and used to convert the string age data. These formats were as follows:

```
week_match = re.search(r'(\d+)\s*(WKS|W|WEEK|WEEKS)', age_str)

month_match = re.search(r'(\d+)\s*(MTHS|M|MTH|MONTH)', age_str)

year_match = re.search(r'(\d+)\s*(YRS|Y|YR|YEAR)', age_str)

combined_match = re.search(r'(\d+)\s*(&)', age_str)
```

After the analysis and formatting was completed, the rest of the data cleaning was fairly simple. An If/Else statement was used to convert the data to decimal row by row checking in week, month, year and finally year/month combined data order. Of course, with an If/Else statement there's always that error situation that must be accounted for. In this case, it was our invalid data with one example was the timestamp used for age. The catch all end of the IF/Else converted the invalid data to None (null) so that it could be dropped with the other null values. If the data was valid it was converted from string to decimal so that better analysis could be

performed on the data. Initially stated, there were 11,221 null age rows. After the conversion of the invalid data replacing the invalid data with null, the age data was analyzed again. It was noted there were 11,387 age rows with proper decimal data place in a new column age_decimal so that the original age_data could be maintained. Pre-conversion of the age data we started with 22663 rows of data, 11442 had age data and 11,221 rows were null. Out of the 11,442 converted 11,387 rows had valid data. This means the conversion change 55 rows to null. As stated above these were rows with data such as timestamp. Between nulls and bad data, this left us with only 50% fo the data in the dataset to analyze. Hence why manually entered data without required dropdown field entry is not ideal for data collection. Further analysis below will also show how the collection of Breed information as well was compromised. We hope this age analysis shows the importance of proper data collection and required data collection.

Datetime/season:

DateOfBite column data type was in text so converted into Datetime and created Year, Moth, Day columns. Based on Month column created Seasons column by creating a dictionary of seasons mapped to month column. Based on Day column created DayOfWeek column by creating a dictionary of days and mapped to Day column. Dropped Species column which has “Dog” word in all the columns thought it is not useful for analysis to reduce the data size. We had a ZipCode column which had ‘?’ in one of the rows so eliminated that row in the data cleaning process.

Dog bites dataset Visualization

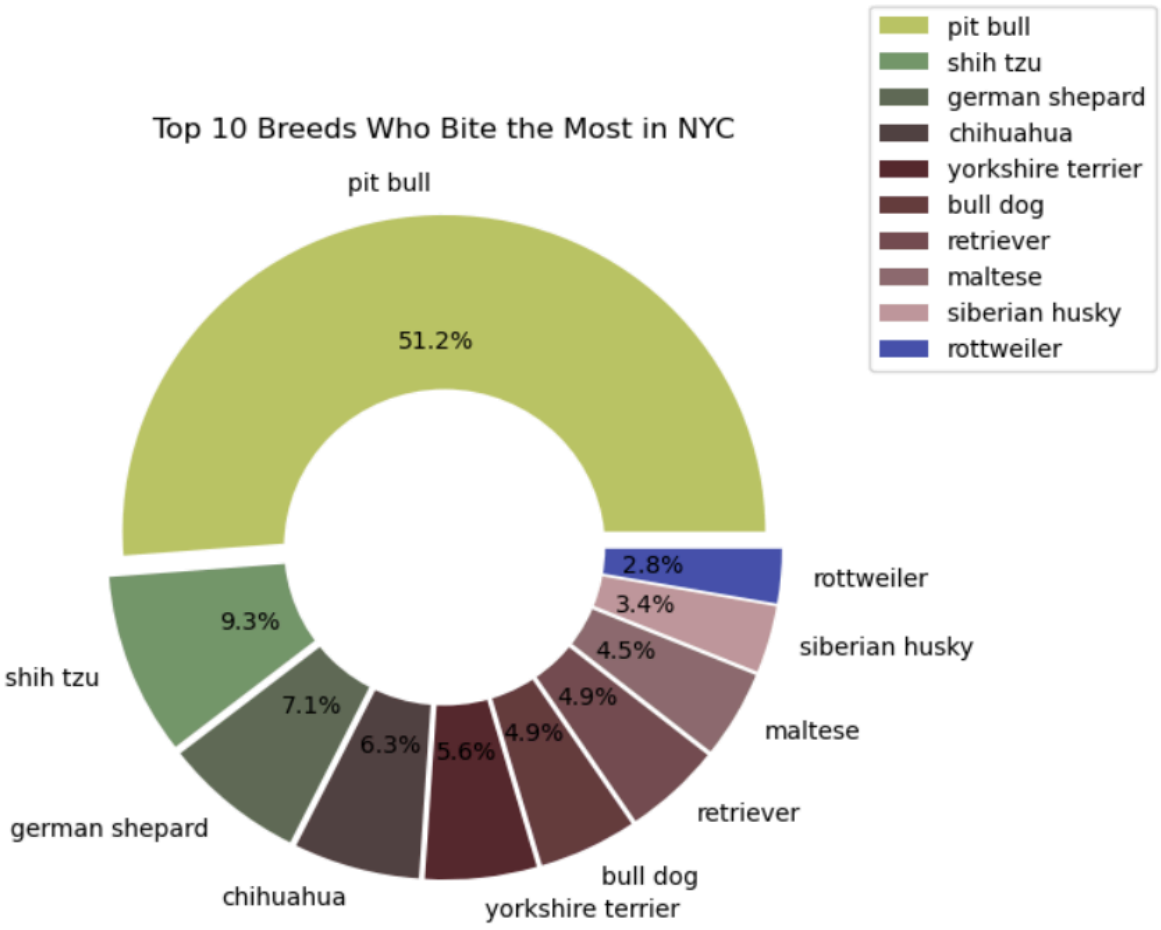
Visualizations:

Through data visualizations we show what potentially increases a dogs chances of biting and answers data questions through Data Visualization. We hope this educates our larger audience as well. Despite the unfortunate error prone data collection techniques, we were still able to provide visualizations which coincides with other prior data analysis.

Below are questions which will be answered through Visualization:

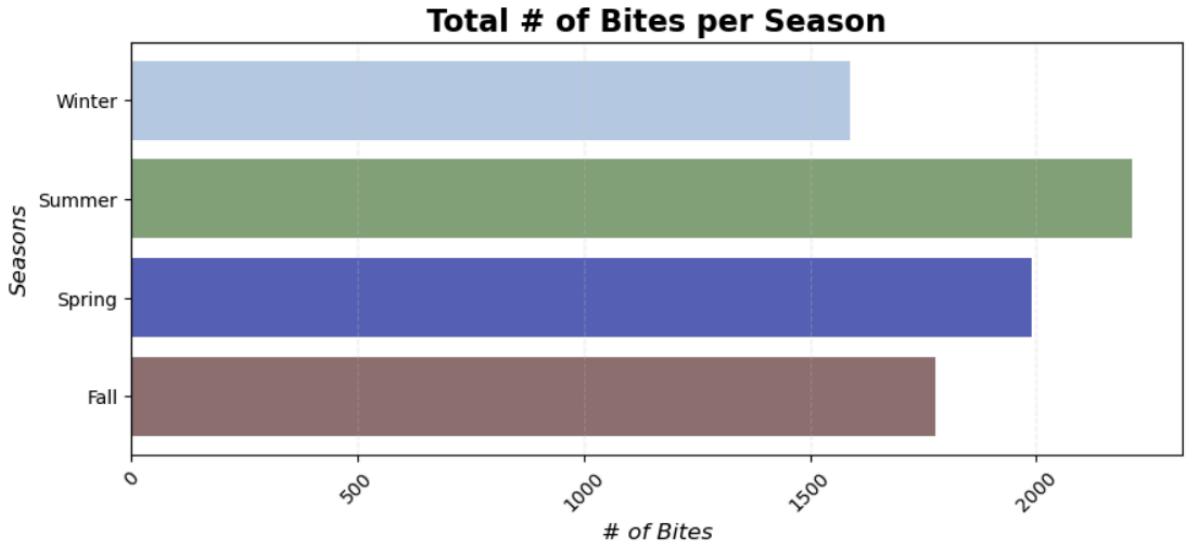
1. What type of Breeds are more likely to bite?
2. How do Seasons and Days of Week impact biting?
3. Does the age of a dog show biting trends?
4. Are dog bites increasing or decreasing?

Dog bites vs # of bites:



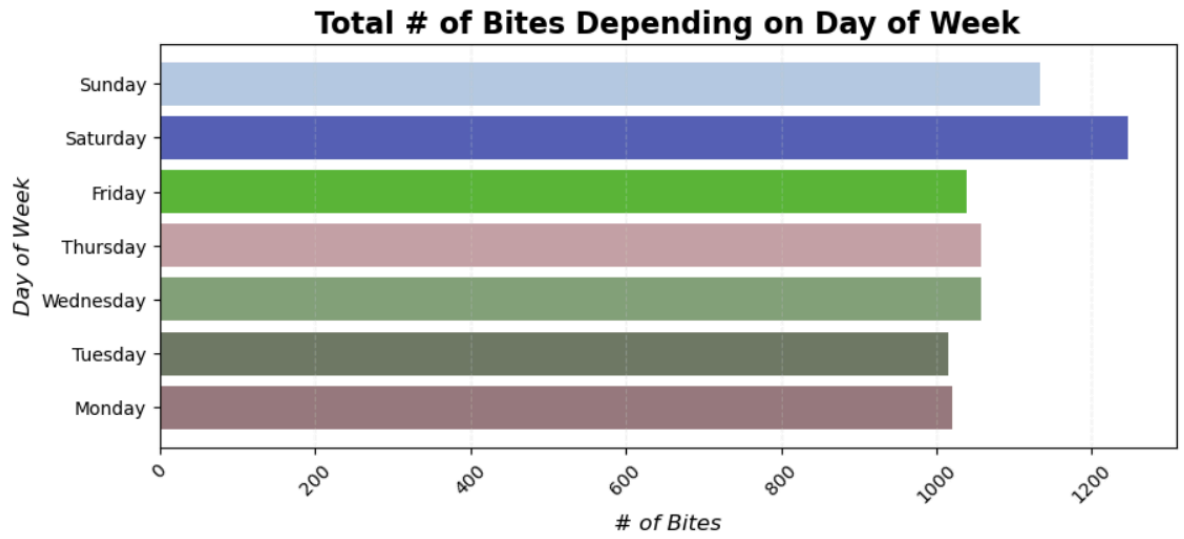
To answer our first question of what dog breed bites the most we figured using a donut chart would best display this info because it's a great way to represent which breeds bite the most as a percentage out of 100. With this donut chart we can clearly see the breed that bites the most are pit bulls by quite a large margin. With over 50% of the documented bites it's pretty obvious they are leading the pack. Followed not very closely by shih tzu's, german shepherd's, and chihuahuas which are all much closer together in number of bites compared to the disparity between pit bulls and the next leading breed in bites, shih tzu's. Of the top 10 breeds shown with this donut chart, four of them are smaller dogs. Everyone knows that small dogs have the biggest bark so maybe the saying all bark and no bite isn't as true as we thought it was. According to the American kennel Club the most common breed types in NYC are Bull dogs, poodles, retrievers, and pit bulls. So because there are a larger number of these breeds in NYC could this explain why these breeds show up more than the rest, or it is truly because they bite more than the others. Unfortunately our data doesn't include what we need to find out this information; however, if we had everything we needed to figure this out, we would have to find the total number of dogs in NYC and compare the total number of a certain breed, such as a pit bull and the total number of bites they have recorded to find the percentage of pit bulls that bite and then compare the percentage to that of other breeds. This would have given us a better understanding of the data because maybe pit bulls represent 50% of the dogs in NYC and really they don't bite more than other dogs, there are just more of them.

Bites per season/day of week:



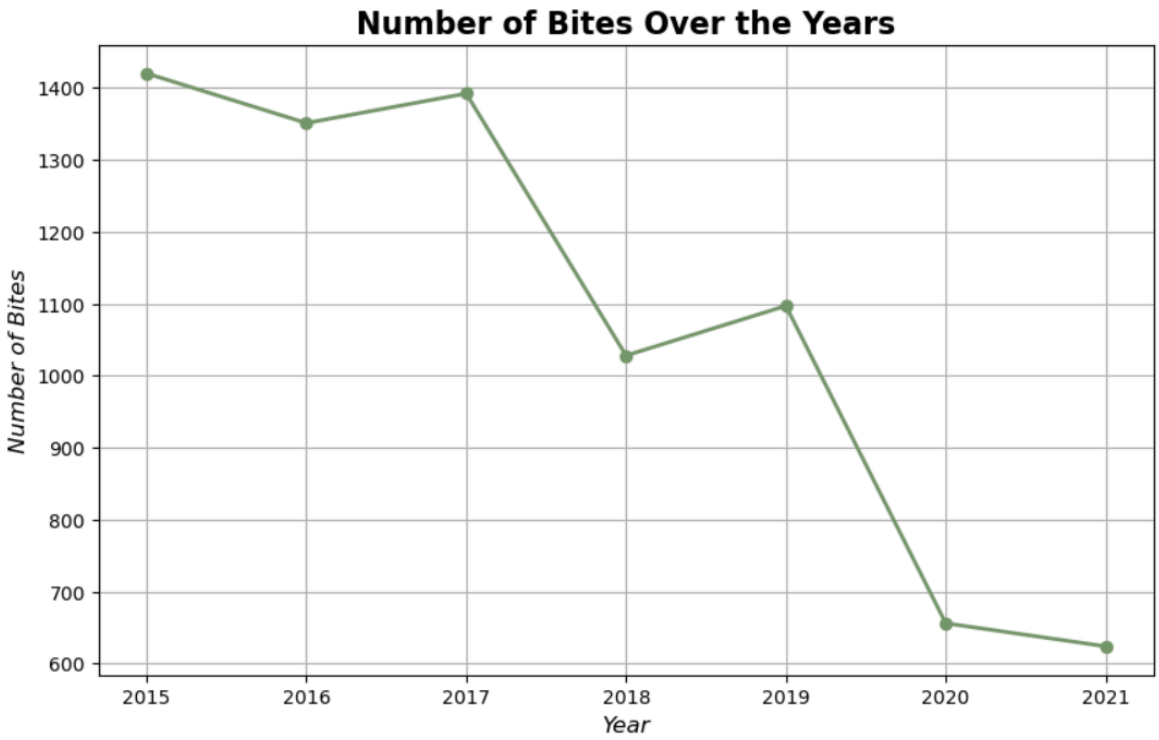
To answer our second question of how Seasons impact biting, using a horizontal bar chart works best to compare the seasons, each bar represents the number of dog bites reported in a specific season. The length of the bar corresponds to the number of bites. There are more dog bites reported during summer, followed by spring, the lowest number of bites recorded in winter. During Summer, people have more visitors, spend time outdoors, visit dog parks which makes dogs exposed to more people causing an increased no of bites in summer than any other season. To reduce dog biting it is recommended to socialize your dog when young, provide chew toys and use positive reinforcement.

[Bites per Day of Week]



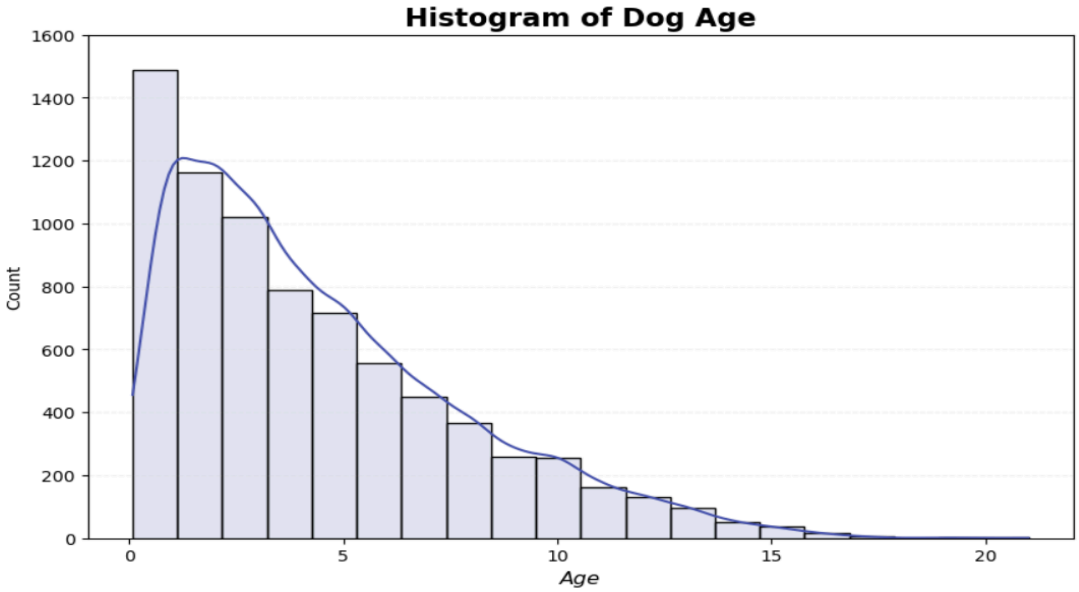
To answer our second part of the second question of how the day of the week impacts biting, using a bar chart best displays the comparison of the number of bites versus day of week. Each bar represents the number of dog bites reported in a specific day of a week. The length of the bar corresponds to the number of bites. There are more dog bites reported during Saturdays, followed by Sunday, the lowest number of bites recorded on Tuesday. During weekends people have more visitors, spend time outdoors, visit dog parks which makes dogs exposed to more people, causing an increased number of bites on weekends than weekdays. Socializing your dog when it is a puppy, providing chew toys, using positive reinforcement, will help to reduce bites.

Bites over the years:



To answer our fourth question, are dog bites increasing or decreasing over the years? We have a line chart which shows the number of bites over the years.There is a trend that dog bites are dropping from the years of 2017 through 2021.The year 2021 has the lowest dog bites recorded with the largest drop between 2019 and 2020. During Covid, people started spending more time with their dogs at home & social distancing, the results of social distancing are that dogs spend less time with other people reducing the number of bites. Hopefully, the trend continues to decrease. It would be great to see the data post 2021; post COVID restrictions.

Bites by dog age:



The next 3 Visualizations will answer question three above: Does the age of a dog show biting trends? A Histogram and Violin plot easily show the beautiful shape and distribution of the data. Along the Vertical axis we see the number of bites based on age, the horizontal axis. Despite more than 50% of age data deleted due to poor data collection, we still had enough data to show the age distribution and right skew of bites per dog age. In fact when first reviewing the histogram, we found an outlier; a dog at an age of 41 which we as well chose to delete. Upon further valuation we left the 21 year old Dalmation as well as a few other older dogs. We see the skew of those outliers as well as the trend for dogs under 5 years of age with more dog bites. The dataset does not tell us what constitutes a dog bite, nor does it tell us if the dog bite was provoked or the age of the person bit for further analysis. Obviously, most dogs are old by

the age of 10 with the fewest bites for older dogs. Showing that nice distribution curve as age increases.

Besides the Violin plot showing the data distribution and spread of the data as well as the widest distribution for younger dogs, it shows our 5 M's in data analysis; min, max, mean (average), median and mode. You can see those outliers, the 21 year old Dalmation at the top. As well as the week old puppies at the bottom. Some of those very you puppy bites we are attributing to normal puppy bites as owners are expected to tame their dogs. It's natural for puppies to nip.

The data calculations below are clearly shown in the data below, except you will see we deleted the 41 year old outlier, shown by this data post data cleaning. Keep in mind age was converted to decimal.

```
Age Max: 41.0

Age Min: 0.08

Age Mean: 4.52

Age Media: 4.0

Age Mode:  2.0

Name: age_decimal, dtype: float64
```

The min dog bite was .08 which is 4 weeks old. Therefore, we question the data collection again. Was there any data validation during data collection? As we stated, all puppies are prone to biting. What constitutes a dog bite and what was the true situation of a 4 week old dog bite.

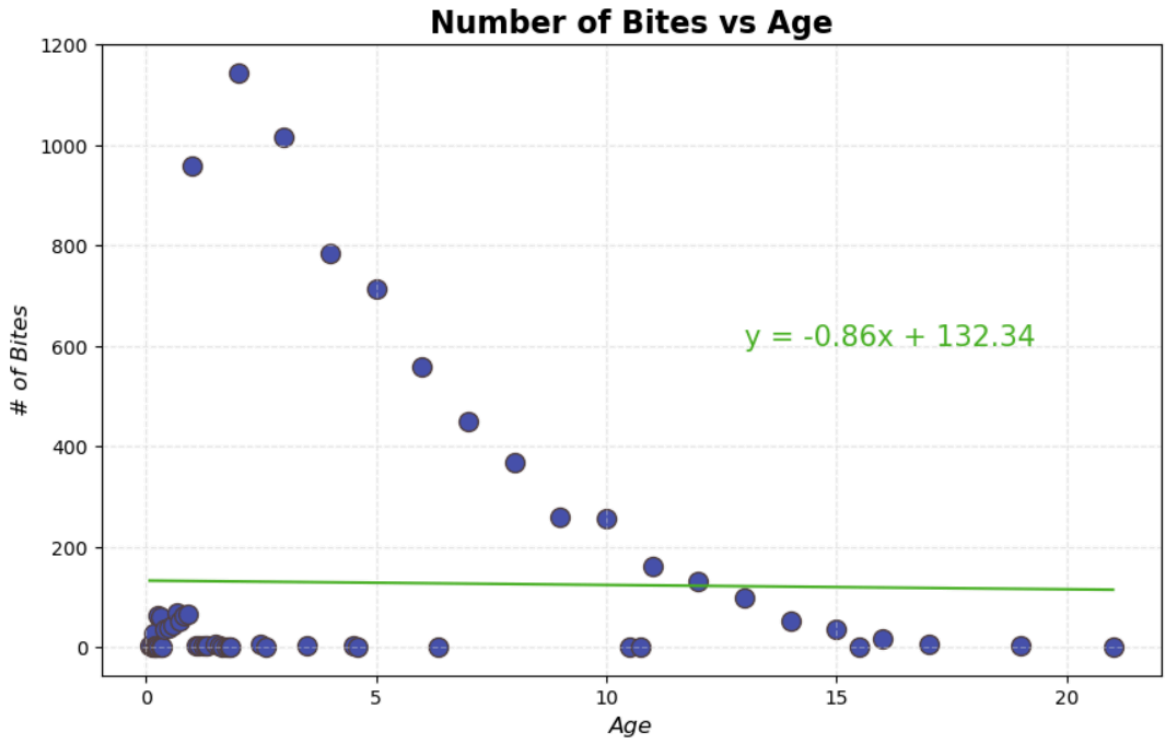
The Mean (average) dog bite was 4 ½ years old. By the age of 4, we are anticipating the owner should have socialized their dog, provided proper chew toys and trained their dogs via proper positive reinforcement. We see the Mean at 4 ½ years old while the Median is 4 years old, hence above in the histogram we see the right skew. And, in the violin chart we see that central dot at the Mean.

The Mode is the value represented most often; the wide distribution in the violin chart of 2 years old shows the mode. The violin chart clearly shows the 1st and 3rd quartile of the data.

Regression:

The Regression Scatter plot displays the concentration of younger dog bites. We barely see the negative correlation and the r-squared=.0003. The regression explain very little about the variation. Our data had such a large concentration of young age dog entries that we can't use the regression. In this case, we would choose to ignore the regression. When removing outliers regression went up slightly, not enough to make any measurable difference.

We are able to see the scatter plot mode and the trend of younger dog bites with some older dog outliers with fewer bites with a steep decrease in bites as a dog ages.



Call to Action:

With all of the results we found from our data visualizations we have a few different suggestions to make that we think could help decrease the number of bites in NYC in the future. First one being, that all pit bulls be required to wear leashes when outside with or without their owner. Since pit bulls are clearly the most common breed to bite, this is the breed we should target the most. This would help with pit bulls not being able to roam away from their owners and bite unpromptedly and hopefully the owners would keep them on a shorter leash so even when walking by they wouldn't really have the opportunity to bite any on goers. Second suggestion would be to have any dog under the age of five wear a leash as well when outside or not in a dog park. Looking at what we found with our histograms it is quite clear that the most common ages to bite are under the age of five and clearly the age of 2 being the most common. With these dogs on leashes when outside or not in a dog park, just like with the pit bulls, it won't let them roam away from their owners and bite anyone passing by. Our third suggestion is to enlist your dog in training within the first two weeks of adopting your dog. According to "Dog See Chew", an online site, they recommend that you start training your dog around eight weeks old; however, if your adopting your dog as a puppy you typically can't take them home until around eight weeks. This is why we recommend enlisting them into training within the first two weeks of taking them home, so that you're starting to get them trained as close to that eight week old age. If you're adopting and the dog is over the 10 weeks, we recommend you enlist them into training within the first two weeks of bringing them home. This will help instill great behaviors within your dog, so that when you start taking them out for walks they won't bite anyone passing by. Our final recommendation is to go to the vet before the summer every year to have your dog checked for any allergies or medical issues. We recommend this because biting is most common during the summer and we don't want any of these medical issues transferred over to

a human if the dog was to bite someone. Vets as well can also review biting precaution measures with owners prior to summer.

Limitations/Biases:

We had a good amount of limitations/biases with this project. This was a good example of what reality will be like and how some datasets can be very messy and questions not able to be fully answered with only what was provided as well as how poor data collection methods can minimize the amount of useful data. Our first limitation was with our breed column. As mentioned earlier, there were over 1600 breeds according to this dataset, obviously that's not accurate. Yet the data entry showed this to be true. The data appeared to be manually entered which caused the obnoxious amount of distinct breeds which quickly became a rather large limitation for analysis causing many judgment calls during data cleaning. If they were to have been entered in through drop down selections and choosing which breed it was as well as required entry, it could have removed the misspelling, added letters, and would have had consistent capitalization across all breeds. Age as well as breed faced similar issues as discussed above. Less than 50% of the rows were used because of invalid age data. Another limitation we had was the dataset only included NYC and its five boroughs. To make our results more significant, we would need more data across the United States to determine if the same 10 breeds that bite the most across the U.S. are the same as the ones just in NYC or would there be a difference? A big bias in this project was what this dataset defined as a dog bite. We had a dog that was 4 weeks old recorded as a dog bite. It could be a dog excited to see someone and run towards them and sort of nibble on them, or it could be only when they bite and cause injury. Without the information to which way they defined a bite to be we have to assume it whenever a dog nibbled/bit someone, rather than only when they caused injury. We weren't able to use the spay/neuter column due to the lack of verification of the data. If we had attempted to use it we may have misrepresented the data. Not so much a limitation but a good

addition to the dataset could have been the age of the person bit. This could have told a very interesting story, for example, are there more younger kids getting bitten by dogs and could this be because they may run up to dogs when they see one and the dog gets defensive and bites. The lack of time to complete this project was another limitation we encountered in terms of breed as every type of breed had to be reviewed and categorized. Was it a terrier? Was it a bull? Do we just lump them all under mixed breed? Was bias being created by our own interpretation of breed during data cleaning? If we had as much time as we needed we could have completed the breed column cleaning, found out more about the total number of dogs in NYC and compared % of bites per breed to one another, and added in more detailed visualizations.

Future work:

Breed Column:

More data cleaning is needed under the breed column. Addressing data bias such as more years, better data collection for breed and age, gender verification. In the Future, we would review a larger dataset with better data collection techniques and verified data that expands to the U.S., covers all breeds and ages as well as required entry for breed, age and gender. We would require the age of the person bit as well as a drop down of bite situations to further understand the situation. We would require a dataset with specific definition of bite to avoid a 4 week old dog entry without understanding the situation.

Borough Column:

This data set is limited to these places Brooklyn, Bronx, Staten Island, Queens, Manhattan from New York, others indicate the bite took place outside New York City, but we

don't have data for in any other columns so In the future we can collect data from more places to use this data to compare and make a good decision.

Gender:

In the description of the date, it states gender is not verified information during collection. We would expect any dataset collecting data on dog bites to have all data entry verified if used to improve understanding of dog bites. If gender data was reliable, more descriptive and very visual violin charts would have been created. We specifically left gender out because it was stated this was unverified data.

DateOfBite Column:

Dog Bite Incidents dataset has data from 2015-2021, having only 7 years of data, future work would be collecting data for more years and ensure that we can make fair and ethical decisions on this dataset. We can't say for sure dog biting is continually decreasing as the data ended during COVID restrictions. COVID had a great impact on many data visualizations including pet behavior.

Work Responsibilities:

Responsibilities were realigned after the proposal as the 4th person in the group never showed up. Joy took the lead on the project proposal creation, Alex created the group GitHub project, Thripura took the lead on the daily standup communication. Each person did some data cleaning due to the data collection complexities. Data cleaning required individual effort with this dataset. Alex took the lead on the breed data cleaning with the assistance of a mentor and the team - Breed was the most complex. Joy worked on age with Alex merging the age function into the final code and the team perfecting that merge. Thripura added columns for day of week, year and season analysis which started off the data cleaning. The team overall worked together to finalize the data cleaning. The team worked together on the visualizations including regression with each person working further to answer the questions during the presentation and the write-up. Alex took the lead on the breed, Joy took the lead on Age, Regression and the limited Gender data, Thripura took the lead on the Date columns(day of week, years and seasons). The creation of a simple Project Presentation Template started with Joy with all contributing their sections. The creation of the Project Writeup started with Alex with all contributing their sections. We all did quality reviews and met outside of class to coordinate roadblocks; mostly with Github due to the initial use of GitKraken. GitHub issues were definitely a roadblock as well as the poor data. We all worked very closely on the project coordination, each taking the lead when needed. Assistance from TAs as needed occurred during the project as well as suggestions from the professor. The team was prepared to lump several breeds into one future breed cleaning bucket so that we could use that data for age, day of week, season and year analysis. Professor okayed dropping some of the breeds due to the poor data collection of breed.

Sources:

"Most Popular Dog Breeds by City in 2021." *American Kennel Club,* 01 March 2024, https://www.akc.org/expert-advice/news/most-popular-dog-breeds-city-2021/

"Dog Bite Incidents Dataset." *Kaggle,* n.d., https://www.kaggle.com/datasets/michaelbryantds/dog-bite-incidents/data

"Coolors." *Coolors.co,* n.d., https://coolors.co/visualizer/bdc667-4bd61d-626d58-544343-4f282c-653c40-734e51-737171-8c6d6f-7885e7

"What's the Right Age to Start Puppy Training?" *Dogsee Chew Blog,* 22 March 2022, https://www.dogseechew.in/blog/whats-the-right-age-to-start-puppy-training#:~:text=Having%20your%20pup%20socialise%20and,%2C%20stand%2C%20stay%20and%20come

"Microsoft Template Designs" *Microsoft* https://create.microsoft.com/en-us/template/blob-design-62339b2c-5cf2-47ff-90b5-2207f1f6a677