Epidemiological Study: NYC Dog Bites

Ian Young

Table of contents

Background

Pet ownership has continued to rise in the past few decades.[1] Dogs are the most widely owned pet across the United States with an estimated 68 million households in the United states owning at least one dog.[2] Using data on dog bite occurrences in New York City this report seeks to analyze trends over time in dog bites as well as examine any spatial and demographic trends related to dog bite occurrences.

The data was collected and provided by the New York City Department of Health and Mental Hygiene. NYC requires all animal bites to be reported within 24 hours, the data is compiled from reports submitted online, by mail, fax, or phone. Each record represents a single dog bite incident occurring between the start of 2015 and the end of 2023.

Purpose & Limitations

A major focus of this report is to demonstrate a range of analytics skills, including API integration, data tidying and transformation, data visualization, statistical modeling, and professional reporting. The primary packages used in this report come from the tidyverse and tidyverse/Posit developers (httr2, tidymodels, gt, etc.).

This project did not involve extensive research into animal bite epidemiology. Additional data sources could be included for more complex or broadly applicable analysis.

Initial Data Collection & Cleaning

Using an API call the data was pulled from NYC OpenData. Following the extraction of the data the date column was used to generate individual columns for year, month, day, and day of the week. The existing uniqueid column had duplicate values. A new uniqueid column was made by sorting the dataset by the date that the dog bite occurred and using the row number as a unique identification number. Next, the species column was dropped as it provides no information (all values in it are "DOG").

Below is a sample of 10 observations following the initial cleaning. There are a total of 29,992 observations and 12 variables (4 date columns not displayed in the table below).

[1] https://www.statista.com/statistics/198086/us-household-penetration-rates-for-pet-owning-since-2007/

[2] https://www.statista.com/statistics/198095/pets-in-the-united-states-by-type-in-2008/

NYC Dog Bite Instances 2015-2023

Sample of 10 Observations

uniqueid	dateofbite	zipcode	borough	breed	gender	age	spayneuter[1]
6474	2016-11-21	11694	Queens	Pit Bull	M	3	FALSE
23858	2022-05-15	10303	Staten Island	Chihuahua	F	1	FALSE
27566	2023-05-25	NA	Other	Pit Bull	M	1	FALSE
19725	2020-12-28	NA	Manhattan	Shih Tzu	U	NA	FALSE
9676	2017-10-20	NA	Bronx	Pit Bull	U	NA	FALSE
26122	2022-12-31	10463	Bronx	UNKNOWN	U	NA	FALSE
14651	2019-04-30	10463	Bronx	PIT BULL BLUE NOSE	M	5	FALSE
9245	2017-09-05	10306	Staten Island	American Staffordshire Terrier	F	5	TRUE
4284	2016-04-06	11362	Queens	MALTESE POODLE MIX	U	NA	FALSE
26035	2022-12-18	11229	Brooklyn	UNKNOWN	U	NA	FALSE

[1] **FALSE** represents dogs that have not been neutered **OR** it was unknown.
Source: https://data.cityofnewyork.us/Health/DOHMH-Dog-Bite-Data/rsgh-akpg/about_data

Given the way the data was generated, certain columns like zipcode, breed, and age all contain a wide range of values and missing data. The value 'FALSE' in the spayneuter column represents dogs that have not had their reproductive organs surgically removed, as well as cases where it is unknown if the dog involved in the incident has or has not been neutered.

Additional Data Cleaning

Using the zipcodeR package data on census and geographical data are joined with the dog bite data (e.g., borough, city, population, housing units, housing value, etc.). This data reflects 2010 census and zip code data. For simplicity sake records outside of New York City and the records missing location data were removed. The breed column contained open text responses with a wide variety of inputs. Some simple steps were taken to format including making text lower case, removing white spaces, and removing trailing non letter characters.

Lastly, the age column contained open text responses reporting the age of the dog involved in the incident. The data in this column was very messy with no clear and consistent parameters used in the data collection. To clean this data up values that were simply numbers were treated as years, while other values had some text indicating months or weeks old. Other columns listed multiple weeks, months, or years in these instances the largest value was taken. A couple values had nonsensical values such as 65 years old and those were assumed to be typos and changed to months. It is also possible some respondents when considering the age of the dog thought in dog years. For all these reasons even the cleaned column should be used with caution or disregarded. In the future the form used to generate the data should outline clear parameters for responding to the question or force specific answers. In the repository

is an excel sheet named "cleaned_age.xlsx" that can be used as a reference for how certain values in the age column were treated and transformed into the age_years and age_months columns.

Final Data Dictionary

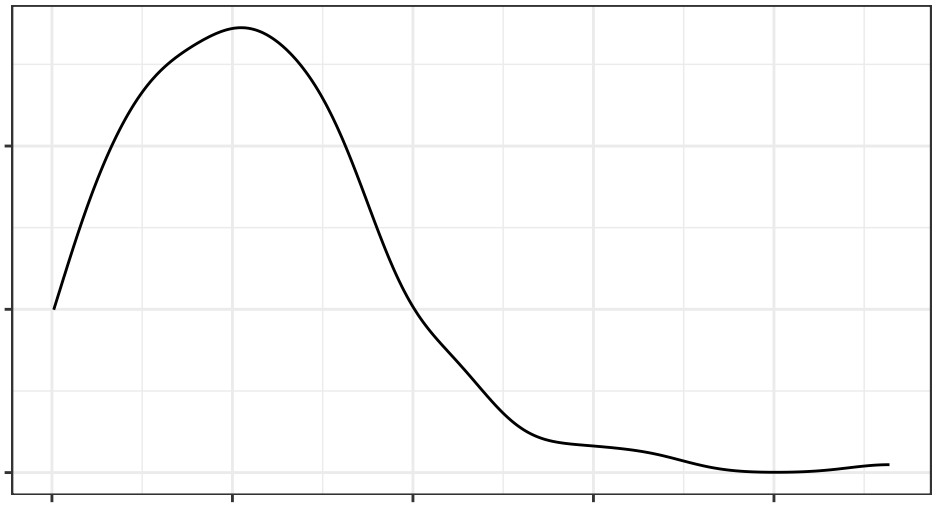
Analysis

Exploratory Data Analysis

The mean dog bites in the different ZIP Codes is 101.2 while the median is 93. There is a noticeable right skew in the distribution with a few ZIP Codes having upwards of 300 dog bites.

```
# Distribution of Count of Dog Bites in Each Zip Code
dog_bites_full |>
  group_by(zipcode) |>
  summarise(n = n()) |>
  ggplot(aes(n)) +
  geom_density() +
  labs(title = "Dog Bite in NYC Zip Codes Distribution",
       caption = "Source: NYCOpenData | Design: Ian Young") +
  theme_bw() +
  theme(text = element_text(family = "Century Gothic"),
        plot.title = element_text(face = "bold"),
        axis.text.x = element_text(color = "black"),
        axis.text.y = element_text(color = "black"))
```

mean	median
112.8	107.5



```
# Summary statistics
dog_bites_full |>
  group_by(zipcode) |>
  summarise(count = n()) |>
  summarise(mean = round(mean(count),1),
            median = median(count)) |>
  gt()
```

```
# month, day of week, gender, age, spay nueter, borough,
# land area, water area, populaton, population density
# housing units, occupied housing units, home value, household income
```

Mapping Dog Bites In NYC

Below is a heat map of dog bites in NYC with lines drawn around the boroughs. The ZIP code with the most dog bites is right off of Central Park

```
# Load mapping package
library(sf)
```

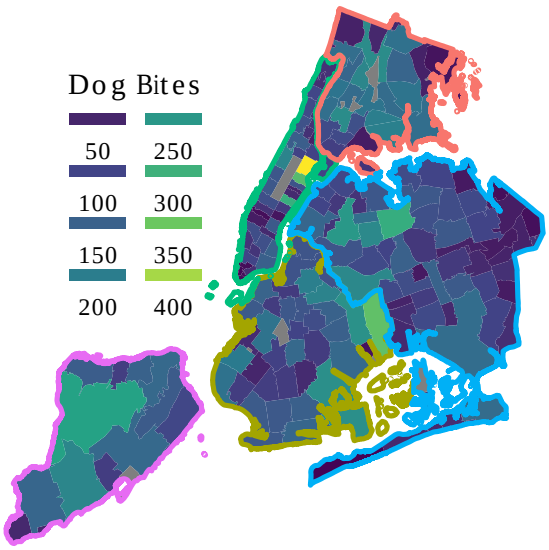
```r
# Read in shape files
shape_nyc_boro <- st_read("Borough Boundaries.geojson", quiet = TRUE)
shape_nyc_zip <- st_read("NYC ZIP Code Boundaries.geojson", quiet = TRUE)

# Merge spatial data and count data
dog_zip_counts <- dog_bites_full |> count(zipcode)

shape_nyc_zip <-
  shape_nyc_zip |>
  left_join(dog_zip_counts, by = join_by(modzcta == zipcode))

# Create Map
(p <-
  ggplot() +
  geom_sf(data = shape_nyc_zip,
          aes(fill = n),
              color = NA) +
  geom_sf(data = shape_nyc_boro,
          aes(color = boro_name),
          show.legend = FALSE,
          linewidth = 0.9,
          fill = NA) +
   scale_fill_viridis_c(
    breaks = c(50, 100, 150, 200, 250, 300, 350, 400),
    name = "Dog Bites",
    guide = guide_legend(
      keyheight = unit(2, units = "mm"),
      keywidth = unit(8, units = "mm"),
      label.position = "bottom",
      title.position = "top",
      nrow = 4
    )
  ) +
  labs(title = "NYC Dog Bites (2015-2023)",
       caption = "Source: NYCOpenData | Design: Ian Young") +
  theme_void() +
  theme(
    text = element_text(family = "Century Gothic"),
    panel.background = element_rect(fill = "white", color = NA),
    plot.background = element_rect(fill = "white", color = NA),
    plot.title = element_text(hjust = 0.5),
    plot.caption = element_text(hjust = 0.5),
```

```
    legend.position = c(0.15, 0.85),
    legend.justification = c(0, 1),
    legend.key = element_rect(fill = "transparent", color = NA),
    legend.background = element_rect(fill = "transparent", color = NA, size = 0)
  )
)
```



Predicting The Number Of Dog Bites In A Given ZIP Code

```
# Split the data
```

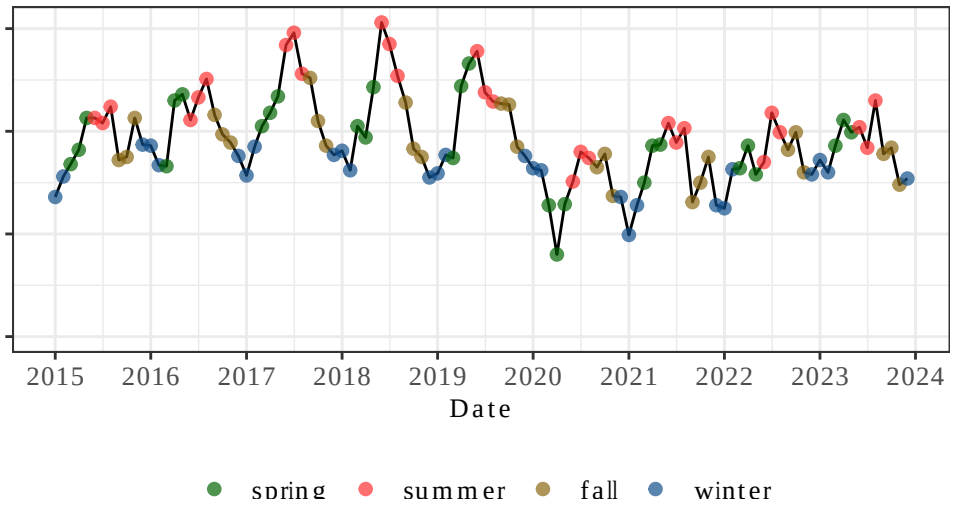
Future Projects

Forecasting Dog Bites In NYC

```
# Create data frame for time series analysis
dog_ts <-
  dog_bites_full |>
  group_by(biteyear, bitemonth) |>
  count() |>
```

```
  mutate(
    season = case_when(
    bitemonth %in% c(3, 4, 5) ~ "spring",
    bitemonth %in% c(6, 7, 8) ~ "summer",
    bitemonth %in% c(9, 10, 11) ~ "fall",
    bitemonth %in% c(12, 1, 2) ~ "winter"),
    date = make_date(biteyear, bitemonth)
    ) |>
  ungroup() |>
  select(date, season, n)

# Initial plot of time series data
dog_ts |>
mutate(season = fct_relevel(season, "spring", "summer", "fall", "winter")) |>
ggplot(aes(x = date, y = n)) +
  geom_line() +
  geom_point(aes(color = season), size = 3, alpha = 0.7, shape = 20) +
  ylim(0, NA) +
  scale_x_date(date_labels = "%Y", date_breaks = "1 year") +
  labs(x = "Date",
       y = "Dog Bites",
       title = "NYC Dog Bites 2015-2023",
       subtitle = "Dog bites regularly spike in the summer months.",
       caption = "Source: NYCOpenData | Design: Ian Young") +
  scale_color_manual(values = c("spring" = "darkgreen",
                                "summer" = "firebrick1",
                                "fall" = "goldenrod4",
                                "winter" = "dodgerblue4")) +
  theme_bw() +
  theme(
    text = element_text(family = "Century Gothic"),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 10),
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 10),
    plot.caption = element_text(size = 8, face = "bold"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_blank(),
    legend.text = element_text(size = 10)
    )
```


2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Date
spring
summer
fall
winter