New York City Dog Bite Analysis

Ian Young

Contents

Load Packages

```
library(tidyverse); library(extrafont); library(scales)
loadfonts(quiet = TRUE)
library(httr2)
library(gt)
```

Connect to API & Compile Dataset

The httr2 and purrr packages are used to connect to an API from New York City and convert dog bite data for the years 2015 to 2022 from an unstructured format into a data frame.

```
# Collect data on dog bites in NYC
dog_bites <-
  request("https://data.cityofnewyork.us/resource/rsgh-akpg.json") |>
  req_url_path_append("?$limit=30000") |> # change call limit to collect all records
  req_perform() |>
  resp_body_json() |>
  map_dfr(~ as_tibble(.)) # convert lists into one tibble
```

Initial Data Cleaning

An initial organizing and cleaning of the data takes place. This involves converting the date column that is stored as character data type to a date type, and adding a individual column for year, month, day, and day of the week. This is done using the lubridate package. The uniqueid column was not recorded properly in the initial data set, and a more helpful uniqueid column is created. Lastly, the species column is dropped as it provides no information (all values in it are "DOG") and the columns are rearranged.

```
# Correct unique ID, adjust/create date columns, select relevant variables
dog_bites_clean <-
  dog_bites |>
  arrange(dateofbite) |>
  mutate(uniqueid = row_number(),
         dateofbite = as_date(dateofbite),
         biteyear = year(dateofbite),
         bitemonth = month(dateofbite),
         biteday = day(dateofbite),
         dayofweek = wday(dateofbite, label = TRUE)) |>
  select(uniqueid, dateofbite, biteyear,
         bitemonth, biteday, dayofweek,
         zipcode, borough, breed,
         gender, age, spayneuter)
```

Background & Initial Exploration

Below is a sample of 10 observations from the initial cleaned dog bite data. There are a total of 26,127 observations and 12 variables (4 data columns not displayed in the table below). Each observation represents a single unique dog bite incident collected by NYC between the start of 2015 and the end of 2022. The data was collected online, by mail/fax, and by phone by the health departments animal bite unit. As a result columns like zipcode, breed, and age all contain a range of values including missing data. For the spayneuter column FALSE 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. Note: The skimr package and function skim are not rendered below, they were used in gathering a quick general understanding of the data.

```
# Sample of 10 observations
set.seed(30)
dog_bites_clean |>
  select(-biteyear, -bitemonth, -biteday, -dayofweek) |>
  mutate(dateofbite = format(dateofbite, "%Y-%m-%d")) |>
  sample_n(10) |>
  gt() |>
  tab_header(
    title = md("**NYC Dog Bite Reports 2015-2022**"),
    subtitle = md("Sample of 10 Observations")
  ) |>
  tab_source_note(
    source_note = md("Source:
    https://data.cityofnewyork.us/Health/DOHMH-Dog-Bite-Data/rsgh-akpg/about_data")
  ) |>
  tab_footnote(
    footnote = md("**FALSE** represents dogs that have not been neutered **OR**
                  it was unknown."),
    locations = cells_column_labels(
      columns = spayneuter
    )
  )
```

NYC Dog Bite Reports 2015-2022

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
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
2884	2015-10-20	11224	Brooklyn	NA	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

Further Cleaning

After the initial data cleaning a few more steps occur to prepare the data for analysis. Using the zipcodeR package data on census and geographical data are joined with the dog bite data. This data reflects 2010 census and zip code data.

The data contains values input from outside of New York City and has missing location data values. For simplicity sake records outside of New York City and the records missing location data are filtered out.

The breed column contains open text responses with a wide variety of inputs. Some simple steps are taken to format including making text lower case, removing white spaces, and removing trailing non letter characters.

Lastly, the age column contains a variety of open text responses reporting age of the dog involved in the incident. The data in this column is very messy and no clear and consistent parameters were used in the generation of this data. To clean this data up values that were simply numbers were treated as years, while other values had something 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. In the future the form used to generate the data should outline clear parameters for responding to the question or force specific answers.

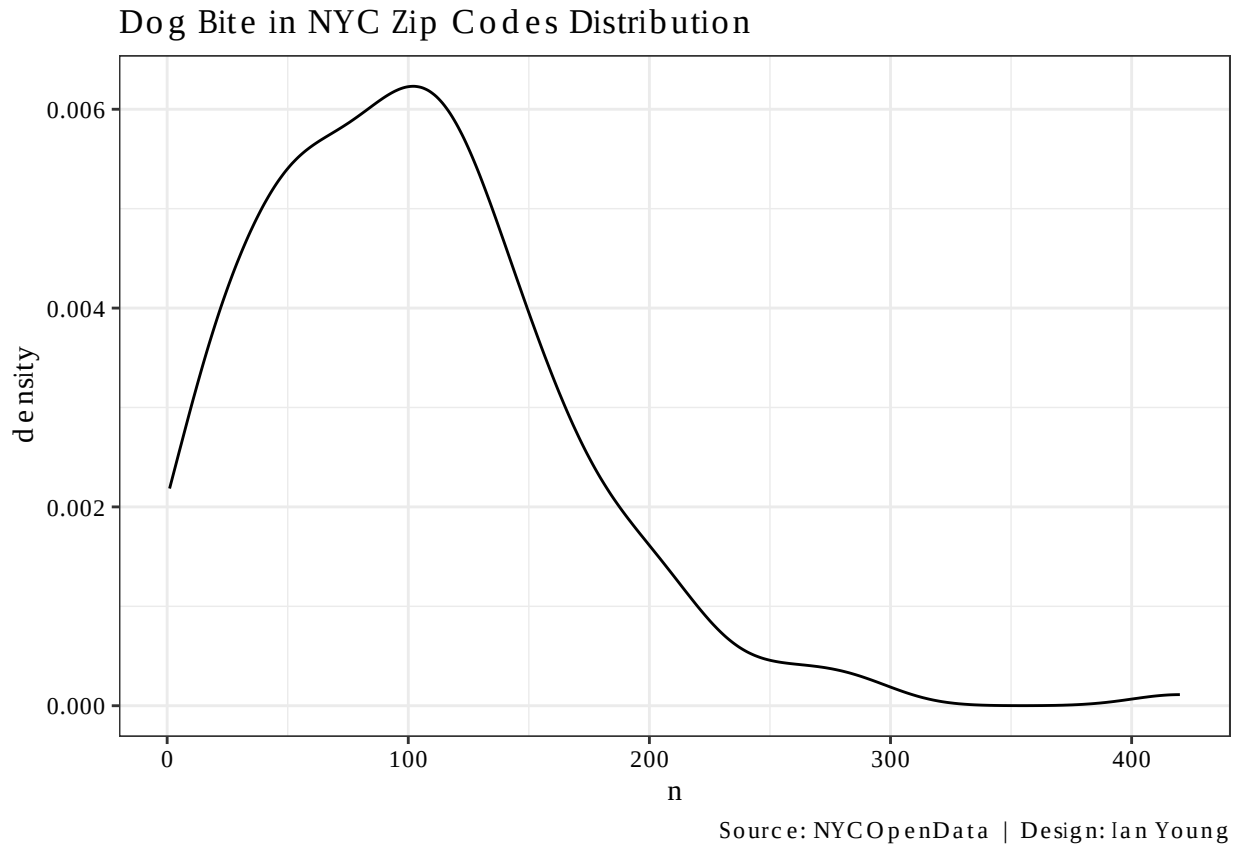
```
# Load in U.S. ZIP code data from zipcodeR to join with dog bite data
library(zipcodeR)

zip_db <-
  zip_code_db |>
  select(zipcode, lat, lng, major_city, state,
         county, radius_in_miles, population, population_density,
         land_area_in_sqmi, water_area_in_sqmi, housing_units,
         occupied_housing_units, median_home_value, median_household_income)

# Join dog bite and zip code data
dog_bites_full <-
  dog_bites_clean |>
  left_join(zip_db, by = "zipcode")

# Rearrange columns
dog_bites_full <-
  dog_bites_full |>
  relocate(zipcode, .after = spayneuter) |>
  relocate(borough, .after = lng) |>
  relocate(land_area_in_sqmi, water_area_in_sqmi, .after = radius_in_miles)

# Filter missing location data and down to NYC records
nyc_counties <- c("Bronx County", "Kings County", "New York County",
                  "Queens County", "Richmond County")

dog_bites_full <-
  dog_bites_full |>
  drop_na(lat) |>
  filter(state == "NY") |>
  filter(county %in% nyc_counties)

# Clean up breed column text
dog_bites_full <- dog_bites_full %>%
```

```
  mutate(breed = str_to_lower(breed), # change text to lower case
         breed = str_trim(breed), # remove start and end white space
         breed = str_squish(breed), # internal white space replaced with single space
         breed = str_remove(breed, "[^a-z]+$") # remove non letter trailing characters
         )

# Clean up age column
clean_age <- readxl::read_excel(
  "cleaned_age.xlsx",
  col_types = c("text", "numeric", "numeric")
  )

dog_bites_full <-
  dog_bites_full |>
  left_join(clean_age, by = "age") |>
  relocate(age_months, age_year, .after = age)
```

Analysis

Exploratory Data Analysis

```
# 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"))
```

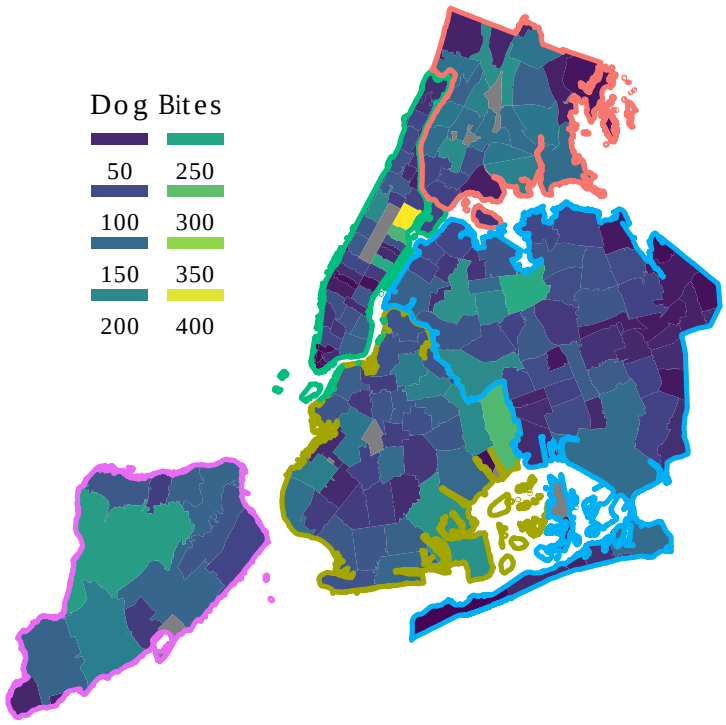



```
# month, day of week, gender, age, spay nueter, borough,
# land area, water area, populaton, population density
# housing units, occupied housing units, home value, household income
```

Mapping Dog Bites In NYC

```
# Load mapping package
library(sf)

# 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 (2015-2022)",
       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)
  )
)
```



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 2015-2022",
       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)
    )
```