FIT5147 Data exploration and visualisation

DEP Part 2: Project Report

Dog Ownership, Infrastructure Access, and Dog Bite Incidents Across NYC Neighbourhoods (2016–2023)

Name: Wanting Zhao

Table of Contents

1. Introduction

Dog ownership is common in large cities, yet supporting infrastructure—such as off-leash parks and dog runs—is not always evenly distributed. In New York City, open datasets show clear spatial variation in both registered dog ownership and dog bite incidents. This raises a practical urban question: whether neighbourhoods with high dog populations have adequate access to dog-friendly spaces, and how this relates to public safety outcomes.

This project is partly motivated by personal observation. In China, dogs are often restricted in public parks, while cities like New York actively provide dedicated dog infrastructure. This contrast suggests that policy and infrastructure may shape how dog ownership interacts with urban environments.

Based on this, the project explores three related questions. First, how dog ownership density varies across NYC boroughs between 2016 and 2023, particularly during the COVID-19 period. Second, whether off-leash spaces are distributed equitably relative to dog ownership density, and which areas show the largest gaps. Third, whether neighbourhoods with higher dog density and limited infrastructure experience higher dog bite rates, and how this pattern relates to neighbourhood income.

2. Data Wrangling and Checking (no more than 3 pages)

2.1 Data Sources

This project draws on five datasets. All data is openly accessible and was accessed between March and April 2026.

Dataset A— NYC Dog Licensing Dataset (NYC Department of Health and Mental Hygiene). Tabular CSV dataset (~617k records × 17 attributes) including breed, gender, zipcode, and licence dates. Records represent individual dog licence transactions rather than unique dogs. The dataset is structured as annual extracts identified by an extract_year field, covering 2016, 2017, 2018, 2022, and 2023. Data accessed from: https://data.cityofnewyork.us/Health/NYC-Dog-Licensing-Dataset/nu7n-tubp

Dataset B — DOHMH Dog Bite Data (NYC Department of Health and Mental Hygiene). Tabular CSV data containing 29,992 incident-level records across 9 attributes, including species, breed, age, gender, spay/neuter status, borough, zipcode, and date of bite. Records span January 2015 to December 2023. Data accessed from: https://data.cityofnewyork.us/Health/DOHMH-Dog-Bite-Data/rsgh-akpg

Dataset C — NYC Dog Runs and Off-Leash Areas (NYC Parks Department). Spatial data containing 91 records across 16 attributes, including park name, borough, facility type, surface type, and polygon boundary geometry, enabling precise spatial joins. The dataset was originally proposed to be accessed as a JSON file via the NYC Open Data API; however, the endpoint returned a format crash with R's sf package. The equivalent CSV export from the same source was used instead, with geometry parsed using st_as_sf(wkt = "the_geom", crs = 4326). Data accessed from: https://catalog.data.gov/dataset/dogruns-20190417

Dataset D — ACS Median Household Income by ZCTA (US Census Bureau). Tabular data accessed programmatically via the R tidycensus package using the 2018–2022 American Community Survey (ACS) five-year estimates (variable B19013_001). The 2022 ACS vintage was selected as the most temporally centred estimate relative to the 2016–2023 dog licensing window. The national extract contains 33,774 ZCTAs; after filtering to NYC zipcodes this reduced to 320 records, which further reduced to 182 after the minimum dog threshold was applied (see Section 2.3), at which point all remaining ZCTAs had valid income estimates. Data accessed via: US Census Bureau ACS, retrieved using tidycensus::get_acs() with a registered API key.

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

Dataset E — NYC Zipcode Boundaries (ZCTA) (US Census Bureau TIGER/Line). Spatial polygon data accessed programmatically via the R tigris package. The national ZCTA dataset was filtered to NYC using a bounding box (longitude −74.26 to −73.70, latitude 40.49 to 40.92), initially yielding 352 polygon geometries. After removing out-of-state and non-residential ZCTAs (see Section 2.3), 182 ZCTAs were retained as the spatial framework for all choropleth mapping and aggregation. Data accessed via: tigris::zctas(cb = TRUE, year = 2020)

Dataset F — NYC Borough Population Estimates (US Census Bureau). Annual population estimates accessed programmatically via the R tidycensus package using the American Community Survey 1-year estimates (variable B01003_001) for five NYC boroughs (Bronx, Kings, New York, Queens, Richmond) across seven years: 2016 to 2023. The 2020 ACS 1-year estimates were not released by the Census Bureau. Perhaps impacted by the COVID-19 pandemic. This dataset was used to normalise licensed dog counts as a per-capita ownership rate for borough-level temporal analysis. Data accessed via: tidycensus::get_acs(geography = "county", survey = "acs1").

2.2 Data Wrangling and Transformation

All data preparation was conducted in R (version 4.5.2) using packages including tidyverse, sf, tigris, tidycensus, lubridate, and janitor. The overall goal was to standardise datasets to a consistent geographic unit (ZCTA) and ensure compatibility for spatial analysis.

Dataset A (dog licensing) was cleaned by removing nine columns with complete missingness, as they provided no usable information. Relevant attributes such as breed, gender, zipcode, and licensing dates were retained. Zipcodes were standardised to a five-digit character format to support consistent joins across datasets.

Dataset B (dog bite incidents) required more extensive cleaning. Date fields were parsed into a standard format using lubridate, and records were restricted to 2016–2023 to align with the licensing data. Zipcodes were padded to five digits and validated against NYC ZCTA boundaries. Invalid or out-of-region zipcodes were set to missing rather than removed, allowing these records to still contribute to borough-level summaries. Records labelled as “Other” boroughs were excluded from spatial analysis, as they could not be reliably located.

Dataset C (dog runs) was converted into a spatial object by parsing WKT geometry from the CSV export. Borough abbreviations were recoded for consistency. A spatial join (st_within) was then used to assign each dog run to its corresponding ZCTA, enabling aggregation of infrastructure counts at the zipcode level.

Datasets D (income) and E (ZCTA boundaries) were accessed programmatically. Income data was filtered to NYC and aligned using the “GEOID” field, while ZCTA polygons were subset using a bounding box covering the NYC region. This initial spatial filter required further refinement, as it also captured adjacent New Jersey areas.

A unified spatial dataset was then constructed at the ZCTA level by joining all sources onto the boundary polygons. This ensured consistent geographic alignment for all subsequent analysis. From this dataset, several derived variables were computed, including dog density (licensed dogs per km²), bite rate (bites per 1,000 licensed dogs), number of runs per ZCTA, and dogs per run (for ZCTAs with at least one facility).

During validation, 88 ZCTAs with zero licensed dogs were identified and traced to New Jersey postal areas unintentionally included by the bounding box. These were removed to ensure the analysis remained within NYC. After filtering, 182 ZCTAs were retained as the final analytical dataset.

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

2.3 Data Checking and Error Correction

Dataset A required several corrections. Duplicate records (6.4%) were removed, consistent with the dataset's structure as annual extracts. Missing years (2019–2021) were identified, limiting temporal continuity. Zipcode validation was performed using ZCTA boundaries rather than prefix filtering to avoid misclassification. A minimum threshold of 50 dogs per ZCTA was applied to remove unreliable denominators.

Dataset B — Data quality issues were identified in several fields. The zipcode variable was missing for 29.1% of records (n = 8,737); these were retained for borough-level analysis but excluded from ZCTA-level spatial analysis, reducing the precision of bite rate estimates. The age variable showed substantial missingness (51.6%) and inconsistent formats (e.g., "4", "4Y"), and was therefore excluded. All records were confirmed as species "DOG", with no non-canine entries.

Dataset C — Completeness. All 91 dog run records contained valid WKT geometry. No records were removed from this dataset.

Dataset D — Missing income estimates. Thirteen ZCTAs returned NA for median household income from the ACS, corresponding to areas with insufficient survey sample sizes for reliable estimation. These were excluded from income-related analyses (Question 3). After the minimum dog threshold was applied, all 182 retained ZCTAs had valid income estimates, so no further exclusions were required for Q3.

Dataset F — 2020 population gap. Consistent with the licensing data gap, no ACS 1-year population estimate exists for 2020. Borough-level ownership rates are therefore computed for 2016, 2017, 2018, 2022, and 2023 only — the five years where both numerator (licensed dogs) and denominator (population) are available.

3. Data Exploration

The data exploration was conducted in R using ggplot2 for all visualisations, with sf for spatial operations and patchwork for multi-panel layouts. The exploration followed the structure of the three research questions, progressing from temporal and spatial ownership patterns (Q1), to infrastructure distribution and gap analysis (Q2), to the relationship between dog density, infrastructure access, and bite rates alongside neighbourhood income (Q3). Visualisation types were selected to match the nature of each variable: choropleths for spatial distributions, bar charts for discrete categorical comparisons, line charts for temporal trends, and scatter plots for relationships. Statistical methods are described where applied.

3.1 Question 1: Dog Ownership Density Across NYC Boroughs (2016–2023)

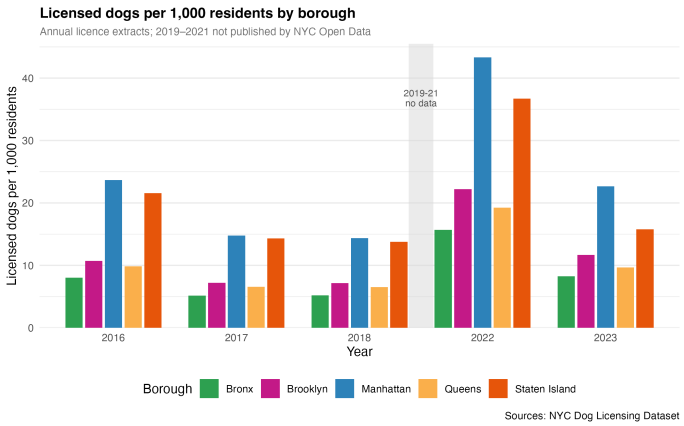
To examine how dog ownership varies across boroughs over time, two temporal visualisations and one spatial map were used. Due to missing licensing data for 2019–2021 in Dataset A, ownership trends and bite trends are analysed separately.

[Figure 0] Borough reference map of New York City

Figure 0 provides a borough reference map for spatial context. A distinct categorical colour palette is assigned to the five boroughs, ensuring each can

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

be easily distinguished.

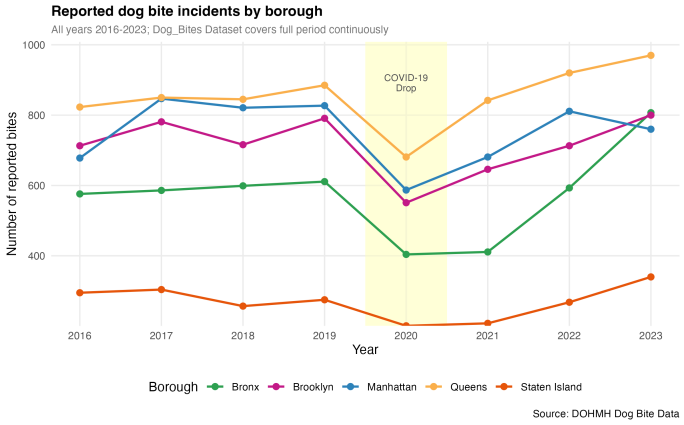


[Figure 1] Licensed dogs per 1,000 residents by borough for five available years (2016–2023)

Figure 1 shows licensed dogs per 1,000 residents by borough for the available years (2016, 2017, 2018, 2022, 2023), calculated using Dataset A and population estimates from Dataset F. A grouped bar chart is used because the data consists of discrete annual extracts rather than a continuous time series.

Manhattan consistently records the highest ownership rate, peaking at 43.3 per 1,000 residents in 2022. From 2016 to 2018, all boroughs show a clear decline (around 32–35%), suggesting a broader reduction in licensing uptake rather than isolated fluctuations. In contrast, 2022 shows a sharp increase across all boroughs, which aligns with widely reported increases in pet ownership during the COVID-19 period.

Rates fall again in 2023, indicating that part of this increase was temporary. Overall, the pattern suggests that dog ownership is influenced by external factors such as COVID-related lifestyle changes.



[Figure 2] Reported dog bite incidents by borough, 2016–2023.

Figure 2 shows total dog bite incidents by borough from 2016 to 2023 using Dataset B. A line chart is used because the data forms a continuous yearly time series, allowing trends and disruptions to be clearly identified.

Queens consistently records the highest number of incidents, despite not having the highest dog ownership. This indicates that bite counts are not directly proportional to dog population size.

All boroughs show a noticeable drop in 2020, likely reflecting reduced human–dog interaction and reporting during COVID-19 restrictions. By 2022, most boroughs had returned to close to or slightly above pre-pandemic levels, although Manhattan and Brooklyn remained marginally lower.

Notably, these are absolute counts rather than rates, due to missing licensing data for 2019–2021.

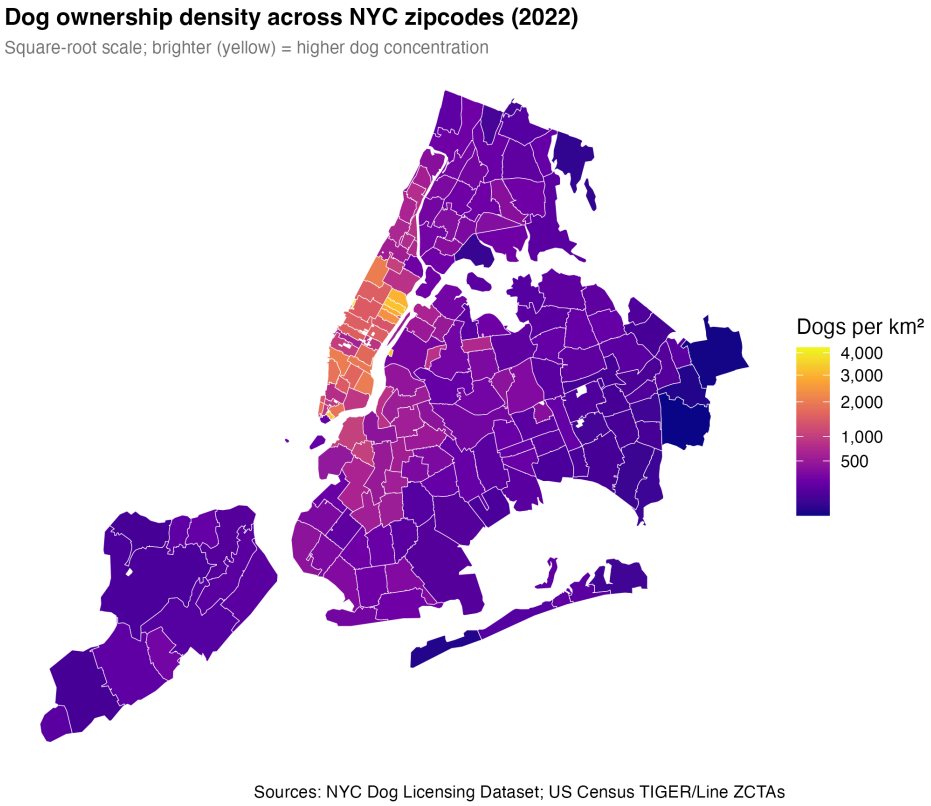
This further suggests that bite incidents are shaped by behavioural and environmental factors, not just dog population size.

Figure 3 presents dog ownership density (licensed dogs per km²) across ZCTAs for 2022. A choropleth map is used because the variable is spatially continuous and aggregated by geographic units, making it suitable for visualising regional variation.

A square-root colour scale is applied to reduce the dominance of extreme values in Manhattan and improve interpretability across lower-density areas. The “plasma” colour palette is used as it provides a perceptually uniform continuous scale, suitable for representing quantitative variation. Its high contrast between dark purple (low values) and bright yellow (high values) helps emphasise areas of high density, drawing attention to potential hotspots. A sequential palette with a neutral midpoint was avoided, as it may imply a “balanced” or desirable state, which is not meaningful for this variable. This choice ensures that higher intensity is consistently interpreted as higher density without introducing misleading semantic cues.

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

High-density clusters are concentrated in Manhattan, particularly in residential areas such as the Upper West and Upper East Side. Parts of Brooklyn and Queens show moderate density, while the Bronx and Staten Island remain relatively low. This spatial pattern broadly reflects differences in population density and housing structure across boroughs.

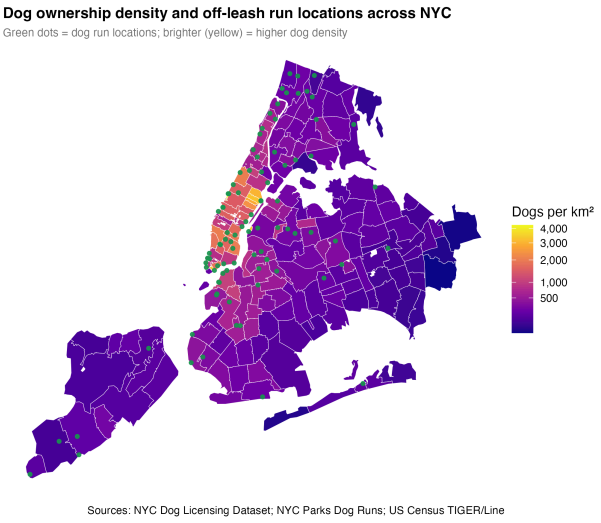


[Figure 3] Dog ownership density across NYC zip code tabulation areas (2022)

Overall, dog ownership shows consistent spatial patterns across boroughs, with temporary disruption during the COVID-19 period.

3.2 Question 2: Equity of Off-Leash Infrastructure Relative to Dog Ownership Density

All spatial analyses in this section use 2022 licensing data (Dataset A, n = 250,259) as the reference year. This year was selected because it provides the most complete and largest single extract, and aligns reasonably with the current snapshot of dog run infrastructure (Dataset C). Therefore, this section is treated as a cross-sectional analysis rather than a temporal one.



[Figure 4] Dog ownership density and off-leash run locations across NYC (2022)

Figure 4 overlays dog run locations onto the dog density choropleth. An overlay map is used to combine

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

two spatial variables—continuous density (area fill) and discrete infrastructure points(green dots)—allowing direct visual comparison between demand and provision within the same geographic context.

At a broad level, infrastructure appears to follow demand — Manhattan, which has the highest density, also contains the largest number of runs. However, this relationship breaks down at finer spatial scales. Many high-density ZCTAs in northern Manhattan, Brooklyn, and Queens have no runs within their boundaries. In total, 118 out of 182 ZCTAs (64.8%) contain zero off-leash facilities, indicating that lack of access is widespread rather than isolated.

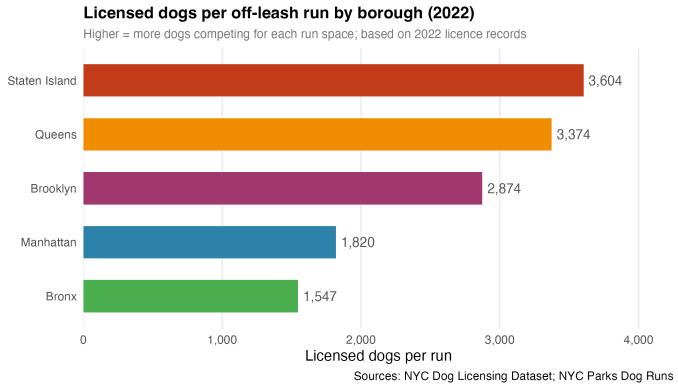


Figure 5] Licensed dogs per off-leash run by borough (2022)

Figure 5 compares the number of licensed dogs per run at the borough level. A bar chart is used here as boroughs form a small set of discrete categories, making differences in ratios easy to compare.

Staten Island records the highest ratio (3,604 dogs per run), followed by Queens (3,374) and Brooklyn (2,874). This reflects limited infrastructure relative to demand rather than high density alone. Manhattan, despite having the most runs (n = 38), shows a moderate ratio (1,820), suggesting that its infrastructure partly offsets high dog density. The Bronx records the lowest ratio (1,547), though this is largely driven by a smaller dog population rather than better provision.

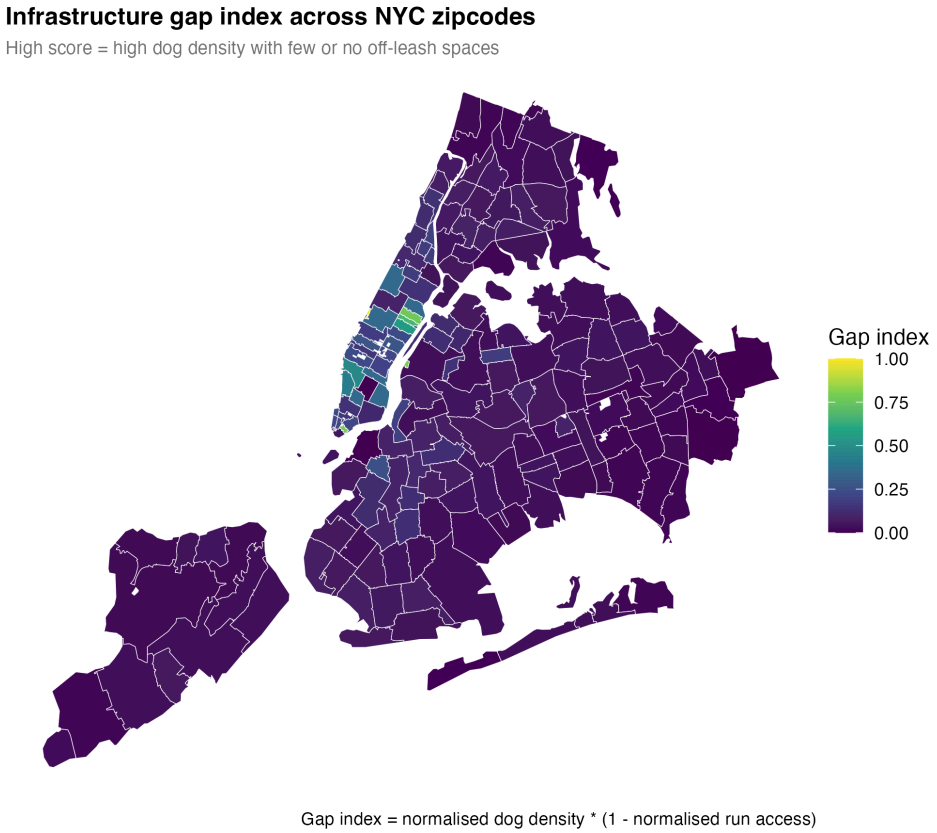
Figure 6 maps the resulting gap index using a choropleth map. A choropleth is appropriate here because the index is a spatially aggregated continuous variable, allowing differences in relative access (demand versus provision) to be compared across ZCTAs.

High scores are concentrated in parts of Manhattan, particularly dense residential areas along the west side. In contrast, Staten Island shows consistently low scores despite having few runs, as demand is also low. This reinforces that infrastructure counts alone are insufficient to assess equity and must be interpreted relative to underlying demand.

A different colour palette (“viridis”) is applied compared to Q1 to distinguish this metric from ownership density. As the gap index represents relative imbalance rather than absolute intensity, using a separate palette reduces the risk of misinterpretation and helps audiences clearly differentiate between the two measures.

Overall, the results show that off-leash infrastructure is not evenly distributed relative to dog ownership. While there is some alignment at a broad scale, many high-density neighbourhoods remain underserved.

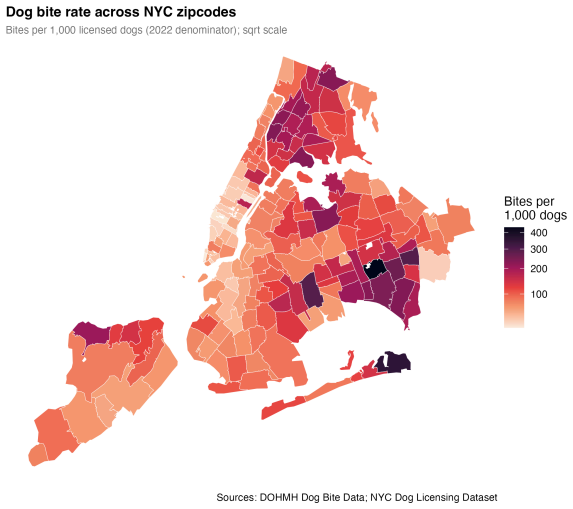
Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.



[Figure 6] Infrastructure gap index across NYC zip code tabulation areas (2022)

Overall, the results show that off-leash infrastructure is not evenly distributed relative to dog ownership. While there is some alignment at a broad scale, many high-density neighbourhoods remain underserved.

3.3 Question 3: Dog Bite Rates, Infrastructure Access, and Neighbourhood Income



[Figure 7] Dog bite rate across NYC zip code tabulation areas.

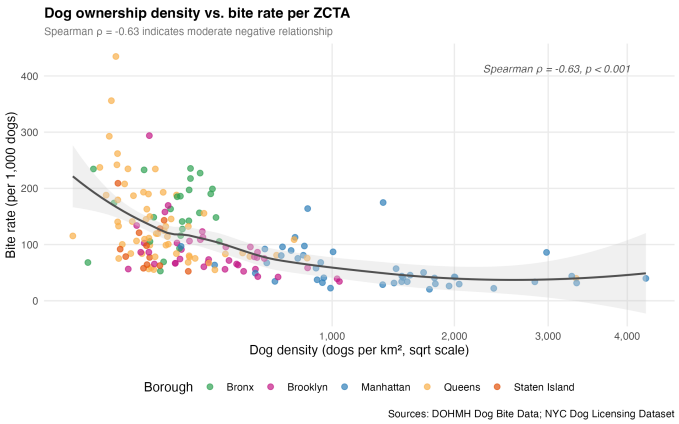
Figure 7 maps bite rates across ZCTAs using 2022 as the reference year. A choropleth map is used because bite rate is a spatially aggregated continuous variable, allowing regional variation to be compared across geographic units.

A sequential “rocket” colour palette is applied, where darker red indicates higher bite rates. This palette is chosen to align with intuitive risk perception, as red is commonly associated with danger or alert. The use of a continuous sequential scale ensures that higher intensity consistently represents higher risk without implying artificial thresholds.

Higher rates are concentrated in Queens and the Bronx, although the spatial patterns differ. Queens contains several extremely high-rate ZCTAs, while the Bronx shows more consistently elevated rates across a wider area. This distinction highlights a limitation of choropleth maps: visually dominant extremes may draw attention, even when another region has a higher overall central tendency. Borough-level summaries confirm that the Bronx has

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

the highest median bite rate despite fewer extreme outliers.

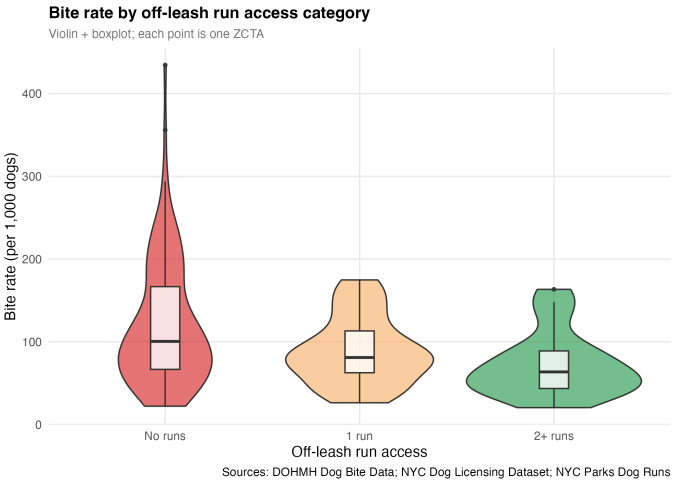


[Figure 8] Dog ownership density versus bite rate per ZCTA

Figure 8 examines the relationship between dog density and bite rate. A scatter plot is used because both variables are continuous, making it appropriate for assessing correlation and non-linear patterns. A LOESS smoothing curve is added to capture the overall trend without assuming linearity.

There is a moderate negative relationship (Spearman ρ = −0.63), indicating that higher dog density is associated with lower bite rates. The curve is steep at low densities but flattens at higher densities, suggesting a threshold effect where bite rates stabilise beyond a certain density level.

Manhattan clusters in the high-density, low-bite-rate region, while Queens frequently lies above the trend line, indicating higher-than-expected bite rates given its density. This suggests that additional local factors, such as housing conditions or dog management practices, may influence outcomes.



[Figure 9] Bite rate by off-leash run access category

Figure 9 compares bite rate distributions across three categories of run access (0, 1, and 2+ runs). A combined violin and boxplot is used to display both distribution shape and summary statistics, which is appropriate given the right-skewed nature of the data.

Median bite rates decrease as run access increases (100.0 → 81.0 → 63.6 per 1,000 dogs). The group with no runs shows the widest spread, indicating greater variability and potential instability in outcomes.

A traffic-light colour scheme (red → yellow → green) is used to represent increasing levels of infrastructure access, aligning with intuitive interpretations of risk and improvement.

A Kruskal–Wallis test is applied due to non-normality, showing a statistically significant difference between groups ($\chi^2 = 11.52$, df = 2, $p = 0.003$). Pairwise Wilcoxon tests (Bonferroni-adjusted) indicate a significant difference between areas with no runs and those with two or more runs ($p = 0.010$), suggesting a potential threshold effect rather than a linear relationship.

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

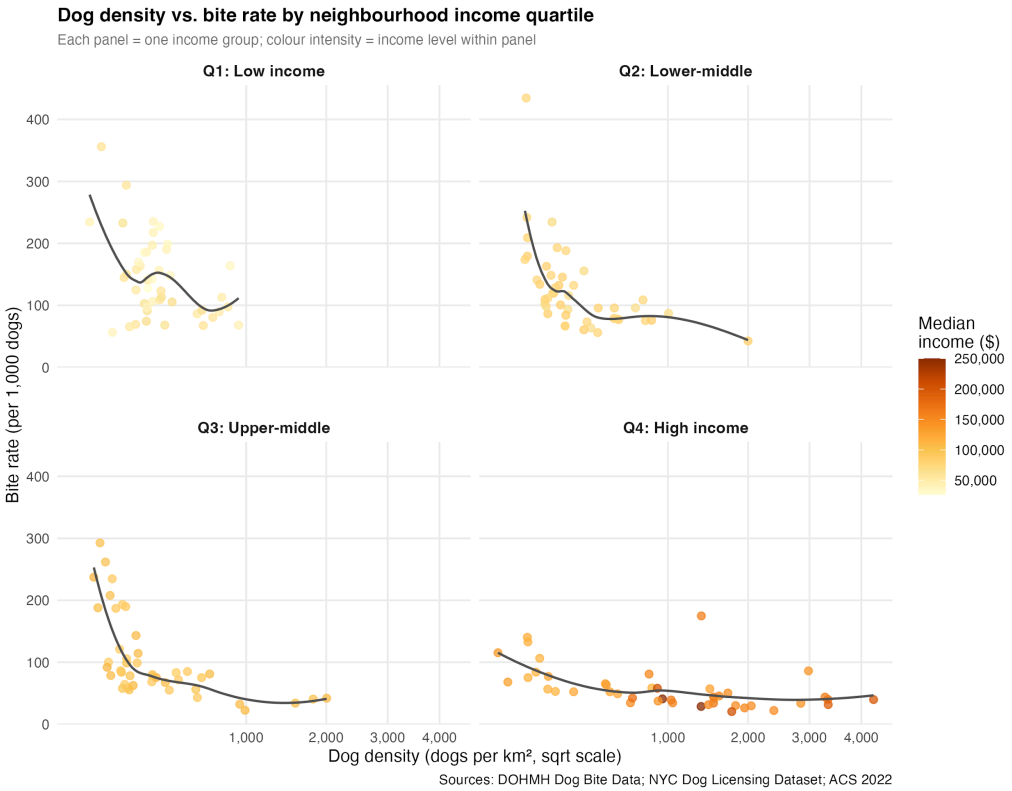


Figure 10. Dog density versus bite rate by neighbourhood income quartile

Figure 10 examines how income modifies these relationships using a faceted scatter plot by income quartile. Faceting is used to compare the same relationship across subgroups while maintaining a consistent scale, enabling clearer cross-group comparison.

The negative relationship between density and bite rate is present across all income groups but varies in strength. It is weakest in the lowest-income quartile ($\rho = -0.303$) and strongest in upper-middle income areas ($\rho = -0.751$). Lower-income areas also show consistently higher bite rates across density levels, with a median of 128.0 per 1,000 dogs, compared to 47.4 in the highest-income quartile.

A sequential colour scale is used to represent income quartiles, ensuring that higher income levels are consistently encoded with higher visual intensity. This avoids introducing conflicting semantic cues while maintaining comparability across facets.

The results suggest that income is a stronger and more consistent predictor of bite rates than density alone, with infrastructure access playing a secondary and potentially confounded role.

Overall, the findings do not support the initial assumption that higher dog density leads to higher bite rates. Instead, socioeconomic context appears to play a more dominant role in shaping public safety outcomes.

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

4. Conclusion

This project examined dog ownership patterns, access to off-leash infrastructure, and dog bite outcomes across 182 NYC ZCTAs using multiple open datasets from 2016–2023.

For Question 1, dog ownership patterns were relatively stable across boroughs, with Manhattan consistently showing the highest ownership rate. A clear spike in 2022 aligns with broader COVID-related changes in pet ownership, followed by a decline in 2023. Bite incident trends show a temporary drop in 2020 and recovery afterwards. Due to missing licensing data for 2019–2021, these trends should be interpreted as indicative rather than continuous.

For Question 2, off-leash infrastructure is not evenly distributed relative to dog ownership density. While provision broadly follows demand at a city level, most ZCTAs (64.8%) have no off-leash facilities. The gap index highlights that dense areas in Manhattan still experience limited access when demand is taken into account. In contrast, Staten Island shows low gap scores largely because of lower demand rather than better provision.

For Question 3, higher dog density does not correspond to higher bite rates. The relationship is moderately negative (Spearman $\rho = -0.63$), suggesting that denser areas may have more controlled or experienced dog ownership environments. Infrastructure access shows some association with bite rates, particularly where there are at least two runs, but this effect is not consistent across comparisons. Instead, neighbourhood income appears to be a stronger and more consistent factor, with lower-income areas showing substantially higher bite rates.

Overall, the analysis suggests that public safety outcomes related to dog ownership are shaped more by socioeconomic context than by dog population size alone. Infrastructure still plays a role, but its effect is uneven and likely interacts with other unobserved factors.

5. Reflection

A key lesson from this project was the importance of checking data completeness early. The missing licensing data for 2019–2021 meant I had to adjust my approach partway through, particularly for the COVID period. In hindsight, I would spend more time validating data coverage before finalising the research questions.

Choropleth maps were the most effective visualisation in this project, especially for showing spatial inequality. The gap index map and bite rate map together made it clear that areas with limited infrastructure do not necessarily have the highest bite risk. At the same time, choropleths are limited in their ability to show change over time, which requires separate temporal charts in Q1. Choosing 2022 as the reference year was consistent, but it also meant that most of the licensing data were not used in spatial analysis.

I also found that colour choices had a bigger impact than expected. Different palettes needed to match the meaning of the data (e.g., density vs risk vs income), and small adjustments like using a square-root scale made a noticeable difference in readability. These design decisions took more iterations than I initially planned.

One thing I could not complete was a clearer integration of income trends over time. Including annual income data alongside ownership and bite trends would likely provide a more complete picture of how socioeconomic factors influence outcomes. This would be a useful extension if the project were developed further.

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

6. Bibliography

1. American Pet Products Association. (2023). Post-COVID rise in pet ownership: Evident and persistent. https://americanpetproducts.org/blog/post-covid-rise-in-pet-ownership-evident-persistent

2. Atlassian. (n.d.). Violin plot: Complete guide.

https://www.atlassian.com/data/charts/violin-plot-complete-guide

3. Brewer, C. A. (1994). Color use guidelines for mapping and visualization. In A. M. MacEachren & D. R. F. Taylor (Eds.), Visualization in modern cartography (pp. 123–147). Pergamon.

4. DogsBite.org. (2025). Dog bite injury trends, top-biting dog breeds and the geography of bite incidents in New York City pre- and post-COVID (2015–2023).

https://blog.dogsbite.org/2025/12/dog-bite-incidents-new-york-city-pre-post-covid-2015-2023.html

5. Field, A. (2018). Discovering statistics using IBM SPSS Statistics (5th ed.). SAGE Publications.

6. Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621. https://doi.org/10.2307/2280779

7. NYC Department of Health and Mental Hygiene. (2025). NYC dog licensing dataset [Data set A]. NYC Open Data. https://data.cityofnewyork.us/Health/NYC-Dog-Licensing-Dataset/nu7n-tubp

8. NYC Department of Health and Mental Hygiene. (2025). DOHMH dog bite data [Data set B]. NYC Open Data. https://data.cityofnewyork.us/Health/DOHMH-Dog-Bite-Data/rsgh-akpg

9. NYC Parks Department. (2026). NYC dog runs and off-leash areas [Data set C]. Data.gov. https://catalog.data.gov/dataset/dogruns-20190417

10. Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72–101. https://doi.org/10.2307/1412159

11. US Census Bureau. (2022). American Community Survey 5-year estimates, 2018–2022 [Data set D & F]. Retrieved via R tidycensus package. https://www.census.gov/programs-surveys/acs

12. US Census Bureau. (2020). TIGER/Line shapefiles: ZIP code tabulation areas [Data set E]. Retrieved via R tigris package. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

13. Walker, K. (2023). tidycensus: Load US census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames (R package version 1.6). https://CRAN.R-project.org/package=tidycensus

14. Walker, K., & Herman, M. (2023). tigris: Load census TIGER/Line shapefiles (R package version 2.0). https://CRAN.R-project.org/package=tigris

15. Wickham, H., & others. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

Appendix

1. **Table of the violin chart**

Group	n	Median	bite rate	IQR
No runs	114	100.0	66.7	167.0
1 run	45	81.0	62.6	113.0
2+ runs	19	63.6	43.6	89.0

2. **Table of Gap_index**

borough	n	median_bite_rate	mean_bite_rate	max_bite_rate
Bronx	25	**156.0**	157.0	236.0
Queens	59	115.0	138.0	**435.0**
Staten Island	12	85.3	98.7	209.0
Brooklyn	38	74.8	86.8	294.0
Manhattan	44	43.9	58.6	175.0

3. **Table of Income**

Quartile	n	rho	Median income	Median bite rate
Q1 Low income	45	−0.303	$51,194	128.0
Q2 Lower-middle	45	−0.685	$75,947	109.0
Q3 Upper-middle	44	−0.751	$95,854	80.6
Q4 High income	44	−0.667	$145,899	47.4

4. **Guihub repository**
   https://github.com/EchoZhao1998/NYC_dog_neighbour/tree/main

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.

5. Generative AI Declaration

This assessment was completed with the assistance of generative AI tools, primarily Claude, which was used as a personalised tutoring aid throughout the project. Its use included guidance on R package selection and debugging, explanation of statistical methods (e.g., Spearman correlation and Kruskal–Wallis test), and discussion of visualisation design choices.

ChatGPT and Grammarly were used to support language refinement, including correcting grammatical errors, improving clarity, and reducing redundancy to meet academic writing standards. All such outputs were reviewed and revised by the author.

All analytical decisions, data processing steps, and interpretations were independently conducted and verified against actual R outputs. The final report reflects the author's own understanding and judgement. Prompts used with AI tools were conversational and focused on learning, clarification, and refinement, rather than direct content generation.

Tools used: R 4.5.2, tidyverse 2.0.0, sf 1.1.0, tigris 2.2.1, tidycensus 1.7.5, lubridate 1.9.5, janitor 2.2.1. No programming code is submitted with this report but is available via Github repo.