A1: Sightings

Weikang Yang

Introduction

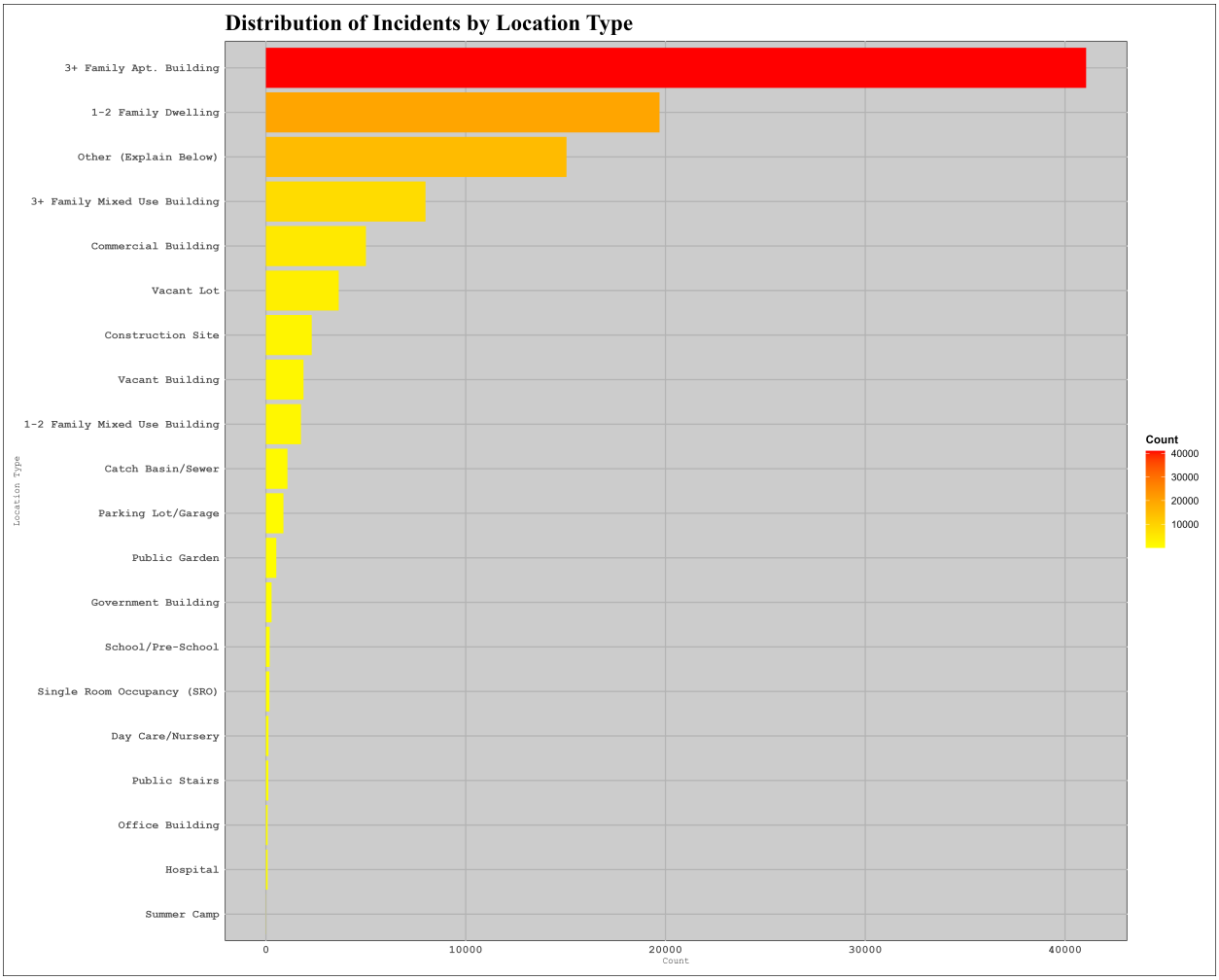
Major cities face continuous rat infestations that have significant impacts on public health and sanitation standards. The combination of high population density and complex infrastructure in New York City creates continuous difficulties in controlling rodent populations. Researchers analyzed rat sighting reports from 2010 to 2017 to determine patterns, geographical distribution, and seasonal changes in rat infestations.

Kaggle provides the dataset analyzed in this study which consists of public records about rat sightings submitted to New York City's 311 service. The data set records the specific dates and types of locations as well as the boroughs and specific position where rat sightings were reported.

Through report visualization and analysis the study aims to determine which areas experience the worst infestations and what time periods see the highest activity along with potential causal factors. The study's results present important information about rat population distribution patterns in urban settings which will benefit the planning of future urban development and pest management strategies.

Analysis by Visualization

Rat Sighting Distribution for Address Type



The bar chart displays how rat sightings in New York City vary according to location types. The x-axis on the chart displays the number of reported sightings and the y-axis enumerates the various location types where these sightings took place. The color gradient from yellow to red indicates incident frequency levels where red signifies high counts and yellow indicates lower counts. The highest occurrence of rat sightings is found in buildings with three or more family apartments which surpasses both 1-2 family dwellings and mixed-use buildings. Hospitals, offices, and summer camps experience significantly lower numbers of reported incidents compared to other locations. The data visualization ranks locations by rat infestation reports while showing residential and commercial areas as primary infestation zones.

Application of the CRAP Principles

- Contrast: The yellow-to-red color gradient successfully emphasizes the differences in sighting counts and makes heavily impacted locations more visible. Using dark text on a light background improves the ease with which readers can understand written information.

- Repetition: Maintaining a professional look requires consistent font styles together with grid lines and bar alignment throughout the presentation.

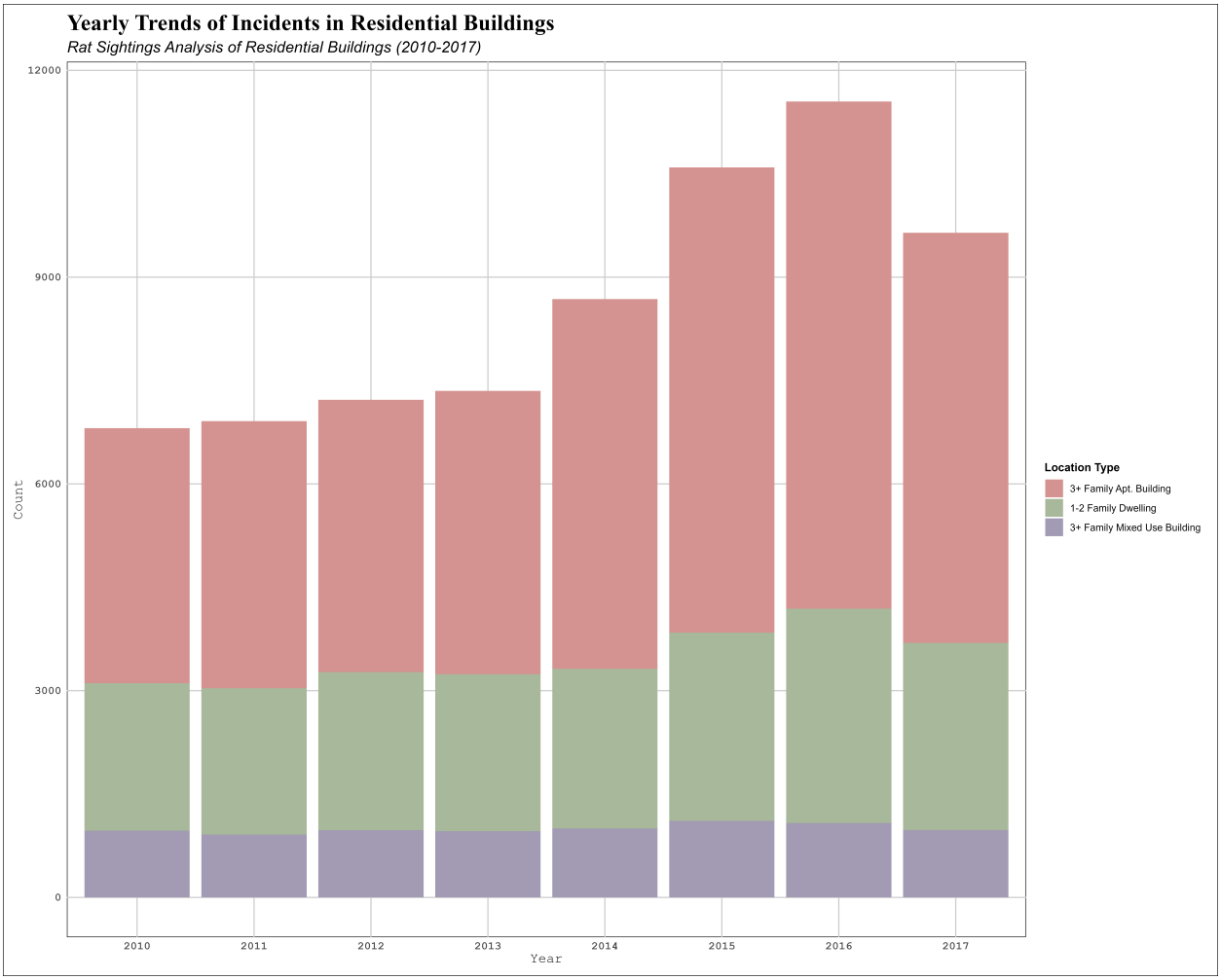
- Alignment: The correct positioning of the title alongside axis labels and the legend maintains chart organization for straightforward interpretation.

- Proximity: Each bar matches its label closely which reduces the possibility of misinterpretation. The legend is located on the right side to maintain accessibility while preventing visual clutter.

Application of Great Visualization Principles (Kieran Healy & Alberto Cairo)
- Kieran Healy's Clarity: The visualization presents rat sightings for different location types in a straightforward manner that excludes extraneous elements.
- Alberto Cairo's Truthfulness: The visual representation maintains true scale and avoids any distortion of data. The horizontal bar format presents an accurate ranking of locations without any distortion.
- Comparability: The bar chart format enables straightforward comparison between different location types to highlight areas with highest rat infestation frequencies.
- Functionality: The gradient color scheme helps viewers to identify problematic regions at first glance without having to examine precise data values.
- Aesthetic Appeal: The visualization combines a smooth gradient with clean typography and minimalistic design to provide an informative and visually pleasing display.

Rat Sighting Trends for Residential Building

Rat sightings show uneven distribution across location types while multi-family apartment buildings experience the highest levels of infestation. The results prompt an investigation into the temporal evolution of rat infestations within residential spaces. The upcoming visualization analyzes yearly residential trends to show how various housing types influence New York City's rat problem.



The stacked bar chart displays the annual patterns of rat sightings across various residential building types throughout the years 2010 to 2017. The chart categorizes incidents into three building types: The data separates rat sightings into three categories which include 3+ Family Apartment Buildings depicted in red and 1-2 Family Dwellings shown in green along with 3+ Family Mixed-Use Buildings highlighted in purple. The 3+ Family Apartment Buildings show the highest number of rat sightings throughout the years according to the data with a significant rise beginning in 2014 reaching its highest point in 2016 followed by a minor drop in 2017. 1-2 Family Dwellings follow a consistent trend while 3+ Family Mixed-Use Buildings show steady numbers that remain low. The visualization accurately demonstrates the increasing rat problem in multi-family residential buildings throughout the observed period.

Application of the CRAP Principles

- Contrast: Distinct color representation of the three categories ensures that they can be easily distinguished from each other. The white background improves readability while grid lines give necessary structure to data without dominating it.

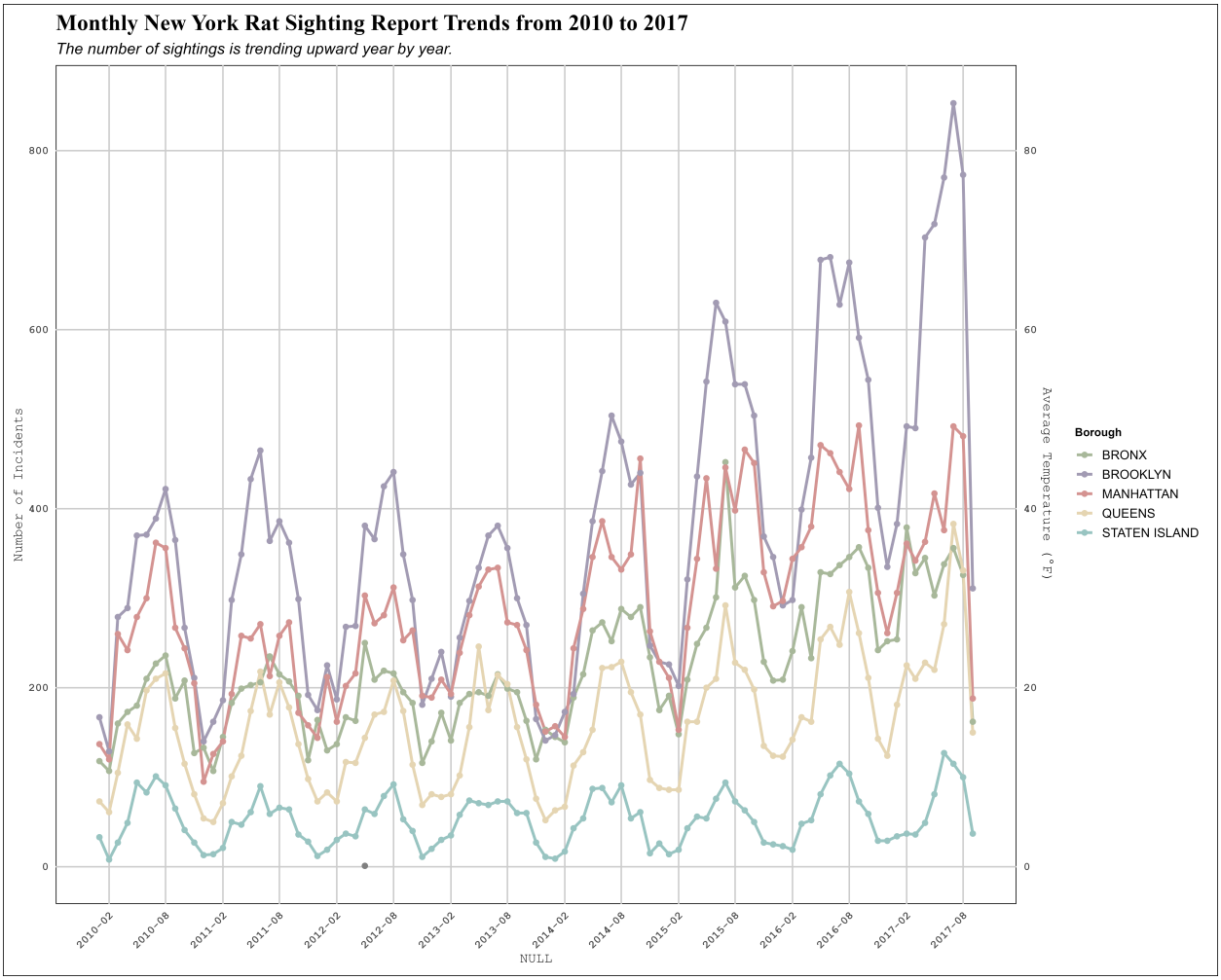
- Repetition: The uniform widths of bars together with regular grid spacing and consistent font styles

produce a structured and professional visual presentation.
- Alignment: All elements including the title, axis labels and legend are correctly positioned to improve readability and interpretation.
- Proximity: The legend's position next to the bars allows users to easily match colors to different building types. The close alignment of the bars enables straightforward comparisons between different years.

Application of Great Visualization Principles (Kieran Healy & Alberto Cairo)
- Kieran Healy's Clarity: The stacked bar chart format presents temporal changes efficiently and demonstrates the individual contribution of each building type towards rat sighting trends.
- Alberto Cairo's Truthfulness: The visual representation of data maintains accuracy throughout by avoiding any scaling manipulations or distortion techniques to depict genuine infestation patterns.
- Comparability: Combining these three categories into one stacked visualization makes it possible to analyze both the aggregate data volume and year-to-year proportions.
- Functionality: Yearly trends become clear to observe through the bar chart format which presents changes over time effectively.
- Aesthetic Appeal: The data representation achieves visual appeal and readability through its combination of a well-selected color palette and organized layout.

Rat Sighting Trends for Each Borough

After known the trends for Rat Sightings in the building/address type, I also want to check how this trend differs among New York City's boroughs. This visualization demonstrates borough-specific patterns by showing which areas have seen the largest growth in rat activity along with the changes in rat sightings over time.



The visualization shows the monthly patterns of rat sightings in all five boroughs of New York City including Bronx, Brooklyn, Manhattan, Queens and Staten Island throughout the period from 2010 to 2017. Rat sightings show a steady increase over time with peaks during warm months followed by declines in cold months. Rat incidents reach their peak numbers in Brooklyn and Manhattan but remain minimal in Staten Island. A cyclical pattern emerges in each borough which reflects the seasonal changes in rat activity. The peak in reported rat sightings occurred during 2016 and 2017 which demonstrates a growing issue as time progresses. A multi-line format combined with color coding facilitates straightforward comparisons between boroughs which enables easy identification of city-wide trends.

Application of the CRAP Principles

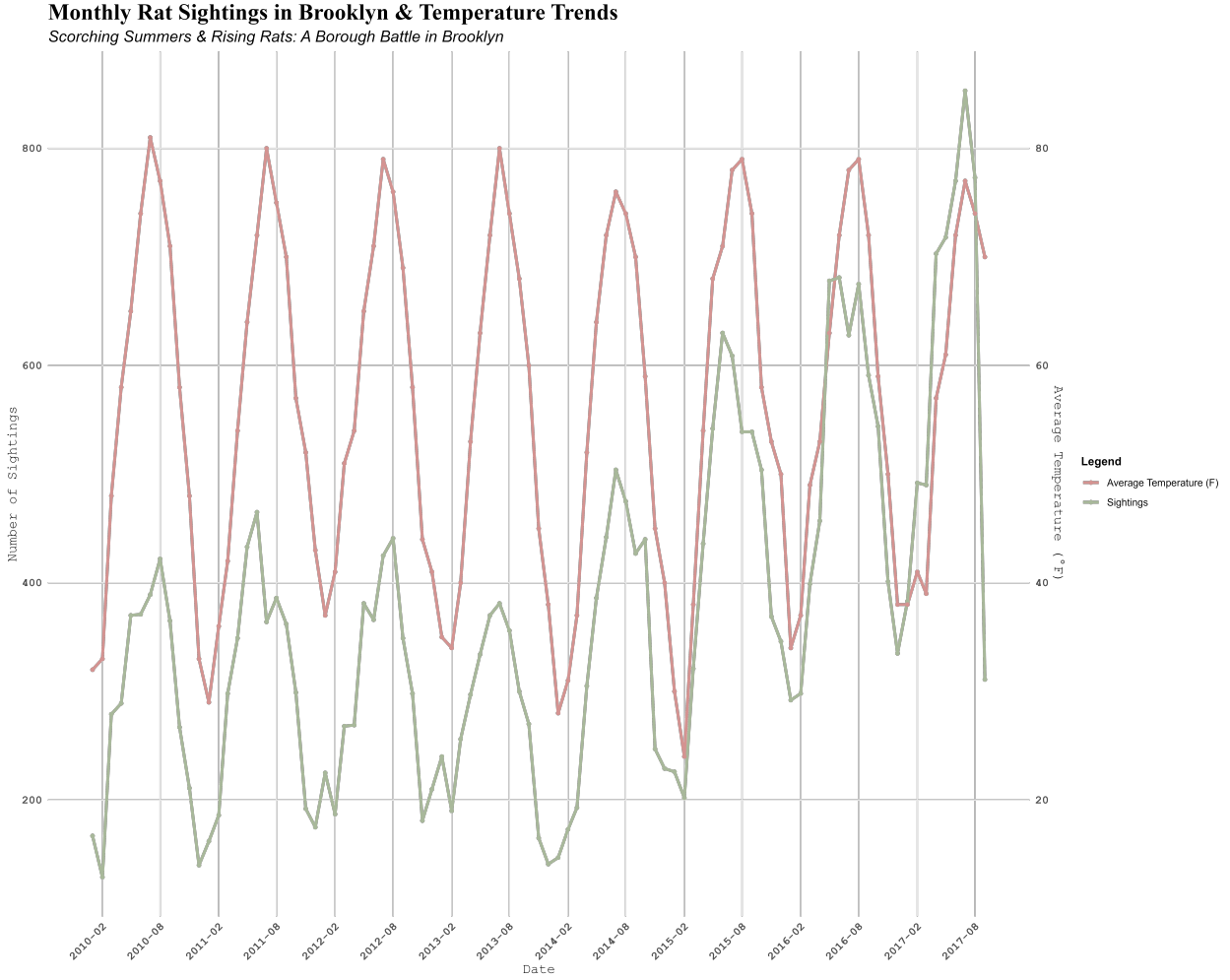
- **Contrast:** Each borough features its own unique color to enable simple recognition. White background and grid lines maintain visual clarity by providing an organized and clear presentation.
- **Repetition:** Uniform application of grid lines and font styles alongside axis formatting creates a structured appearance across the chart.
- **Alignment:** The chart maintains readability through well-aligned title and axis labels along with a subtitle. The legend appears on the right side of the chart to maintain data visibility while remaining easily

reachable.
- Proximity: The legend's location near the visualization helps viewers quickly connect colors to boroughs. The time series data points maintain equal spacing to prevent visual congestion.

Application of Great Visualization Principles (Kieran Healy & Alberto Cairo)
- Kieran Healy's Clarity: The chart delivers rat sighting trends through streamlined elements that emphasize data analysis.
- Alberto Cairo's Truthfulness: This visualization faithfully displays rat sighting numbers over time without manipulating axes or data for visual bias.
- Comparability: Displaying all five boroughs on a single graph allows observers to easily compare their relative trends in rat infestation.
- Functionality: The line format allows users to track temporal changes while clear labeling maintains accessibility across various audiences.
- Aesthetic Appeal: The professional and visually pleasing presentation results from selecting appropriate colors combined with a clean background and organized gridlines.

Rat Sighting with Tempreture in Brooklyn

Brooklyn and Manhattan show the most rat sightings according to borough-level data yet it remains to investigate which external elements are driving these trends. The relationship between temperature fluctuations and rat activity patterns emerges from seasonal variations which we examine by matching temperature data with rat sightings in Brooklyn through subsequent visualization.



This line chart displays data on Brooklyn's monthly rat sightings from 2010 to 2017 together with trends in average temperature. The purple line tracks rat sightings and the red line shows average temperature using a secondary axis for comparison. The chart demonstrates that rat sightings reach their highest levels during warmer periods but drop during colder periods which indicates temperature changes may influence rat activity levels. The data demonstrates a rising trend in rat sightings throughout the years especially after 2014 though significant fluctuations exist. The visualization uses a dual-axis format to merge two interrelated variables which demonstrates potential environmental impacts on rodent population patterns.

Application of the CRAP Principles

- Contrast: The dual-color line representation makes rat sightings and temperature trends easily distinguishable for simple comparison. The dark background makes the white grid lines stand out which enhances readability.

- Repetition: The chart maintains visual coherence through uniformity in font style, grid pattern and line width.

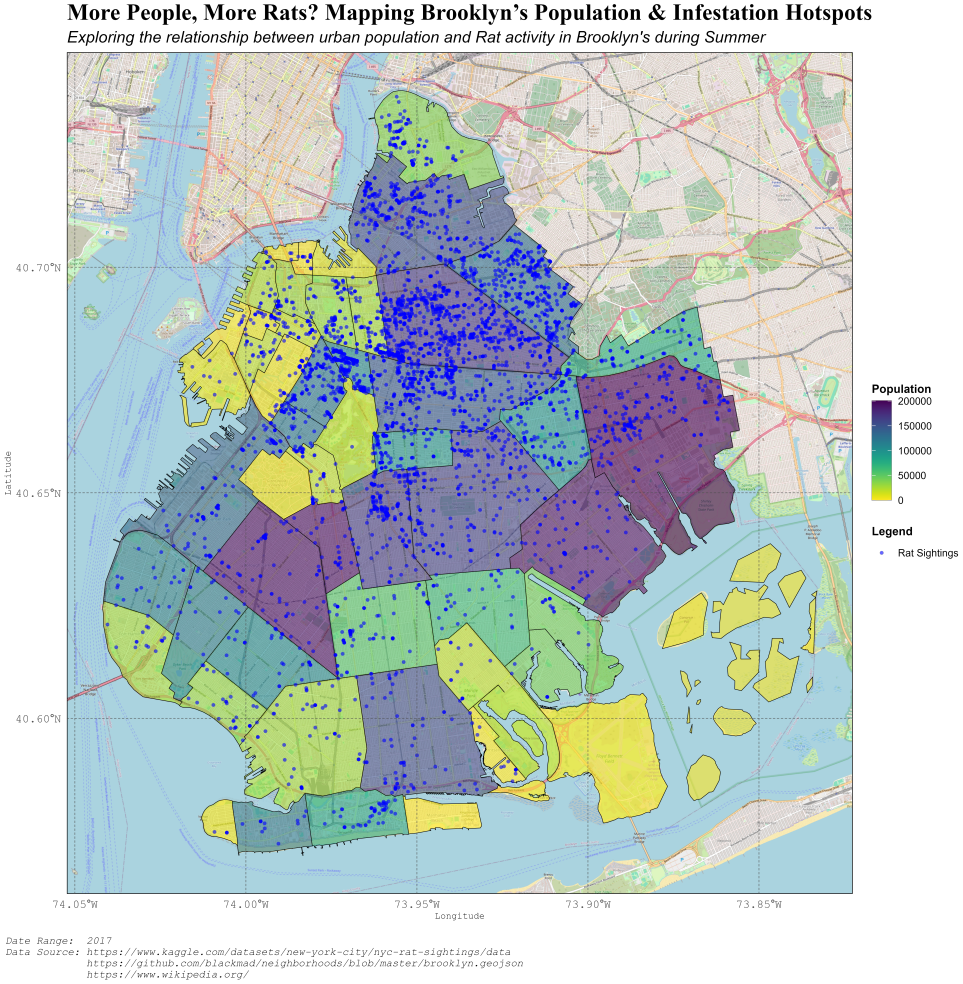
- Alignment: Both variables remain easy to interpret because of the well-balanced dual-axis structure. The legend placement ensures it does not obstruct the data presentation.

- Proximity: The visualization of rat sightings alongside temperature patterns enables immediate visual analysis due to their close plotting. The legend is positioned close to the data lines to facilitate quick reference.

Application of Great Visualization Principles (Kieran Healy & Alberto Cairo)
- Kieran Healy's Clarity: The visualization maintains clarity by eliminating superfluous elements to highlight trends in rat sightings and temperature across seasons and long-term periods.
- Alberto Cairo's Truthfulness: The visualization ensures accurate data representation by using a scaled secondary axis which helps prevent temperature exaggeration while maintaining proper context.
- Comparability: The dual-line format delivers an immediate comparison of temperature fluctuations alongside rat sightings throughout different periods.
- Functionality: The dual-axis layout successfully displays two connected datasets and the time-series format highlights seasonal trends with ease.
- Aesthetic Appeal: The visualization achieves informative clarity while being visually appealing through its soft color palette and organized grid structure.

Rat Sighting and Population in Brooklyn Borough

Seasonal rat activity depends heavily on temperature readings while population density and human activities remain critical factors in controlling infestation rates. This visualization enables spatial analysis by mapping Brooklyn's population together with rat sighting locations to show how the different human activity level will effect the rat sightings.



The interactive map demonstrates how Brooklyn's population density intersects with reported rat sightings from 2017 to analyze potential relationships between popultaion and rat sighting report. The map uses colored regions to show population density where darker colors represent areas with greater population density. The blue dots on this map represent locations where rat sightings have been reported to show how they are distributed throughout various neighborhoods. Northern and central Brooklyn display more rat sightings in their densely populated regions, and with this area as the center, it affects the surrounding area.. The provided visualization enables users to discover areas with intense rodent activity which might be shaped by waste disposal practices and human density alongside infrastructure conditions. The map demonstrates where rat infestations peak by superimposing rat sighting data onto Brooklyn's detailed geographic representation.

Application of the CRAP Principles

- Contrast: The population gradient color scale allows users to easily differentiate between varying levels. The blue dots indicating rat sightings are clearly visible because they stand out from the background.

- Repetition: Consistent gridlines and fonts along with uniform data layering methods ensure the map remains visually coherent.

- Alignment: The title, legend, and axis labels maintain proper positioning which facilitates easy understanding while supporting a balanced and organized layout.
- Proximity: By placing rat sighting points directly on the population density map their relationship becomes more apparent. The legend placement at the right side ensures visibility while avoiding disruption to the map visualization.

Application of Great Visualization Principles (Kieran Healy & Alberto Cairo)
- Kieran Healy's Clarity: Viewers get a straightforward depiction of Brooklyn's population densities alongside rat infestations using an uncluttered map layout.
- Alberto Cairo's Truthfulness: The map provides precise representation of data while maintaining geographical correctness and avoiding any distortions in the locations of rat sightings or population density scales.
- Comparability: The combination of population density information with rat sighting data allows viewers to make straightforward comparisons between different regions while recognizing infestation trends.
- Functionality: Through a map format viewers can effectively perform spatial analysis to observe infestation concentration areas in relation to population density levels.
- Aesthetic Appeal: The map's smooth color gradients together with clear labeling and structured elements deliver an informative and visually appealing depiction of urban rodent activity.

Counclution

New York City rat sightings between 2010 and 2017 demonstrate the expansion of the rodent problem in densely packed residential zones. The most frequent rat sightings occur in multi-family apartment buildings which suggests that both housing density and waste management methods are likely contributing factors to these infestations. After 2014 there was a significant rise in rat reports which points to possible alterations in urban infrastructure development or reporting processes. The impact of climate on rodent behavior becomes evident through seasonal patterns that show warmer temperatures lead to increased rat activity. Data comparing boroughs shows Brooklyn and Manhattan as the most heavily impacted by rat infestations in their densely populated neighborhoods. The spatial analysis of rat sightings in Brooklyn reveals that densely populated areas suffer the most infestations because human activities like improper waste disposal and poor sanitation facilitate rodent growth. The findings demonstrate that pest control methods must be specifically developed for densely populated urban areas with high infestation rates.

Working Process

Data Cleaning and Processing

```
# Read the original data set and check the column
# to check the information that contained in the dataset
library(readr)
raw_df <- read_csv("data/A1_sightings.csv")
colnames(raw_df)
```

```
##  [1] "Unique Key"                    "Created Date"
##  [3] "Closed Date"                   "Agency"
##  [5] "Agency Name"                   "Complaint Type"
##  [7] "Descriptor"                    "Location Type"
##  [9] "Incident Zip"                  "Incident Address"
## [11] "Street Name"                   "Cross Street 1"
## [13] "Cross Street 2"                "Intersection Street 1"
## [15] "Intersection Street 2"         "Address Type"
## [17] "City"                          "Landmark"
## [19] "Facility Type"                 "Status"
## [21] "Due Date"                      "Resolution Action Updated Date"
## [23] "Community Board"               "Borough"
## [25] "X Coordinate (State Plane)"    "Y Coordinate (State Plane)"
## [27] "Park Facility Name"            "Park Borough"
## [29] "School Name"                   "School Number"
## [31] "School Region"                 "School Code"
## [33] "School Phone Number"           "School Address"
## [35] "School City"                   "School State"
## [37] "School Zip"                    "School Not Found"
## [39] "School or Citywide Complaint"  "Vehicle Type"
## [41] "Taxi Company Borough"          "Taxi Pick Up Location"
## [43] "Bridge Highway Name"           "Bridge Highway Direction"
## [45] "Road Ramp"                     "Bridge Highway Segment"
## [47] "Garage Lot Name"               "Ferry Direction"
## [49] "Ferry Terminal Name"           "Latitude"
## [51] "Longitude"                     "Location"
```

```
# select the column that we need to do the visulization and analysis
library(dplyr)
ini_df <- raw_df %>%
  select(-c(
    "Resolution Action Updated Date","Status","Due Date","Closed Date",
    "Agency", "Agency Name", "Complaint Type", "Descriptor", "Landmark",
    "Facility Type", "Park Facility Name", "Park Borough",
    "School Name", "School Number", "School Region", "School Code",
    "School Phone Number", "School Address", "School City",
    "School State", "School Zip", "School Not Found",
    "School or Citywide Complaint", "Vehicle Type",
    "Taxi Company Borough", "Taxi Pick Up Location",
    "Bridge Highway Name", "Bridge Highway Direction",
    "Road Ramp", "Bridge Highway Segment", "Garage Lot Name",
    "Ferry Direction", "Ferry Terminal Name",
```

```
    "Intersection Street 1","Intersection Street 2"
  ))


colnames(ini_df)
```

```
##  [1] "Unique Key"                 "Created Date"
##  [3] "Location Type"              "Incident Zip"
##  [5] "Incident Address"           "Street Name"
##  [7] "Cross Street 1"             "Cross Street 2"
##  [9] "Address Type"               "City"
## [11] "Community Board"            "Borough"
## [13] "X Coordinate (State Plane)" "Y Coordinate (State Plane)"
## [15] "Latitude"                   "Longitude"
## [17] "Location"
```

```
table(ini_df$`Location Type`)
```

```
##
##           1-2 Family Dwelling 1-2 Family Mixed Use Building
##                         19702                          1754
##       3+ Family Apt. Building  3+ Family Mixed Use Building
##                         41061                          7991
##             Catch Basin/Sewer           Commercial Building
##                          1079                          5007
##             Construction Site              Day Care/Nursery
##                          2293                           130
##           Government Building                      Hospital
##                           282                            83
##               Office Building         Other (Explain Below)
##                            87                         15044
##            Parking Lot/Garage                 Public Garden
##                           875                           515
##                 Public Stairs             School/Pre-School
##                           120                           185
##   Single Room Occupancy (SRO)                   Summer Camp
##                           175                             7
##               Vacant Building                    Vacant Lot
##                          1879                          3639
```

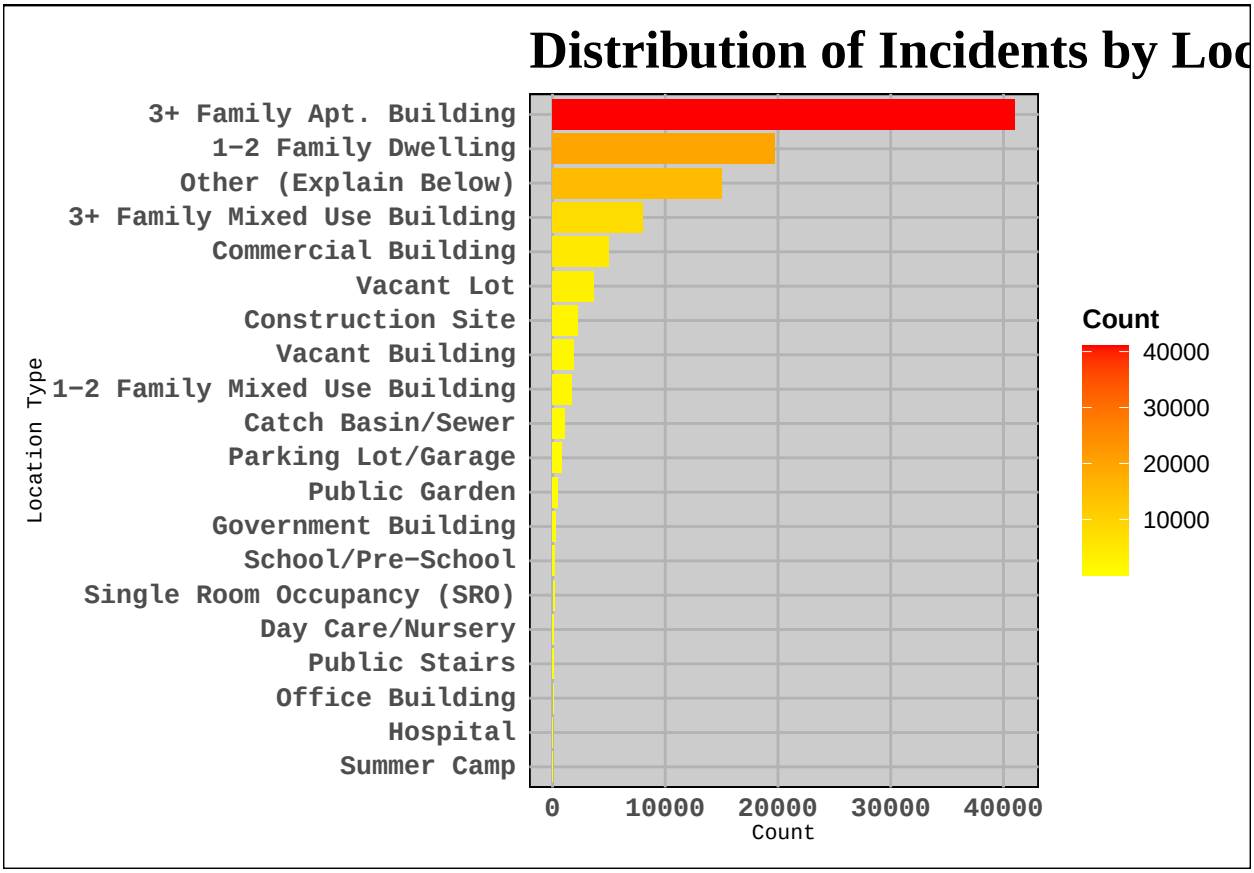
```
library(dplyr)
library(lubridate)
library(tidyr)
library(readr)
library(stringr)
ini_df <- ini_df %>%
  mutate(
    `Incident Address` = replace_na(`Incident Address`, "Unknown"),
    `Street Name` = replace_na(`Street Name`, "Unknown"),
    `Cross Street 1` = replace_na(`Cross Street 1`, "Unknown"),
    `Cross Street 2` = replace_na(`Cross Street 2`, "Unknown"),
    # `Latitude` = replace_na(`Latitude`, 0),
```

```
    # `Longitude` = replace_na(`Longitude`, 0),
    # `X Coordinate (State Plane)` = replace_na(`X Coordinate (State Plane)`, 0),
    # `Y Coordinate (State Plane)` = replace_na(`Y Coordinate (State Plane)`, 0)
  )

# Transform the Date Format
ini_df <- ini_df %>%
  mutate(
    `Created Date` = mdy_hms(`Created Date`),
  )

# 3. Clean & Normalize the address column
ini_df <- ini_df %>%
  mutate(
    `Incident Zip` = as.character(`Incident Zip`),
    `City` = str_to_title(trimws(`City`)),
    `Borough` = str_to_title(trimws(`Borough`))
  )

# Clean the position data, we will make a map later
ini_df <- ini_df %>%
  mutate(
    `Latitude` = as.numeric(`Latitude`),
    `Longitude` = as.numeric(`Longitude`)
  )

# remove the duplicated information
ini_df <- ini_df %>% distinct()

write_csv(ini_df[1:100, ], "ini_df.csv")

# check the dataframe that we already cleaned
glimpse(ini_df)
```

```
## Rows: 101,914
## Columns: 17
## $ ‘Unique Key‘                <dbl> 31464015, 31464024, 31464025, 31464026, 3~
## $ ‘Created Date‘              <dttm> 2015-09-04, 2015-09-04, 2015-09-04, 2015~
## $ ‘Location Type‘             <chr> "3+ Family Mixed Use Building", "Commerci~
## $ ‘Incident Zip‘              <chr> "10006", "10306", "10310", "11206", "1046~
## $ ‘Incident Address‘          <chr> "Unknown", "2270 HYLAN BOULEVARD", "758 P~
## $ ‘Street Name‘               <chr> "Unknown", "HYLAN BOULEVARD", "POST AVENU~
## $ ‘Cross Street 1‘            <chr> "Unknown", "Unknown", "CARY AVENUE", "HUM~
## $ ‘Cross Street 2‘            <chr> "Unknown", "Unknown", "GREENLEAF AVENUE",~
## $ ‘Address Type‘              <chr> "INTERSECTION", "LATLONG", "ADDRESS", "AD~
## $ City                        <chr> "New York", "Staten Island", "Staten Isla~
## $ ‘Community Board‘           <chr> "01 MANHATTAN", "Unspecified STATEN ISLAN~
## $ Borough                     <chr> "Manhattan", "Staten Island", "Staten Isl~
## $ ‘X Coordinate (State Plane)‘ <dbl> 980656, 955207, 949033, 1000550, 1021648,~
## $ ‘Y Coordinate (State Plane)‘ <dbl> 197137, 148858, 169278, 197585, 250489, 1~
## $ Latitude                    <dbl> 40.70777, 40.57521, 40.63124, 40.70899, 4~
## $ Longitude                   <dbl> -74.01296, -74.10455, -74.12688, -73.9412~
## $ Location                    <chr> "(40.70777155363643, -74.01296309970473)"~
```

Plot the Visualization

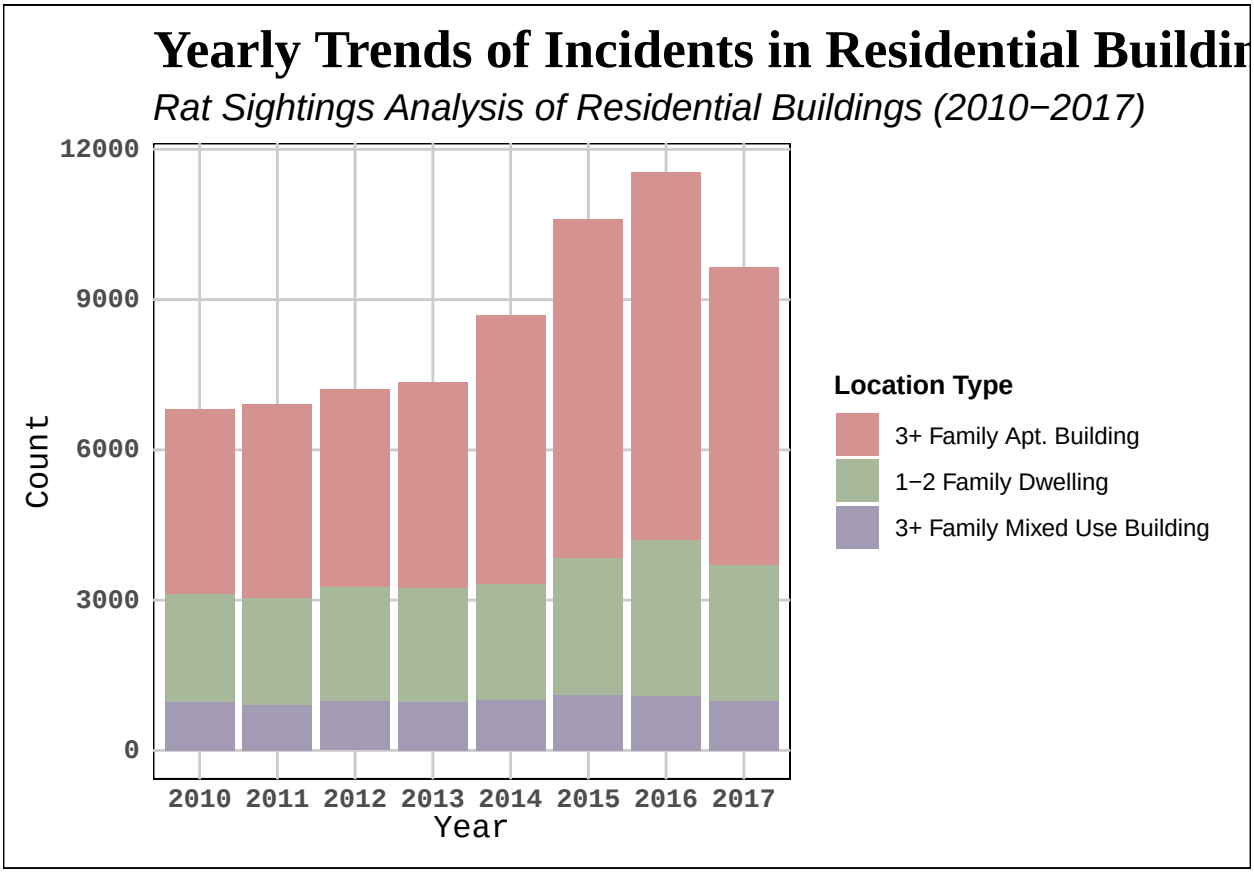
```
library(ggplot2)
library(dplyr)
library(readr)

# count the Location Type, and sort in Decreasing
location_counts <- ini_df %>%
  filter(!is.na(`Location Type`)) %>%
  count(`Location Type`, name = "Count") %>%
  arrange(desc(Count))

# plot the bar chart, use the color to show the level of count
# plot the bar chart horizontal, make the label of building type more easier to read
ggplot(location_counts, aes(x = reorder(`Location Type`, Count), y = Count, fill = Count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_gradient(low = "yellow", high = "red") +
  labs(title = "Distribution of Incidents by Location Type",
       x = "Location Type", y = "Count") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0, size = 20, face = "bold", family = "serif"),
    plot.subtitle = element_text(size = 14, face = "italic", family = "sans"),
    plot.caption = element_text(size = 9, face = "italic", family = "mono", hjust = 0),
    axis.title = element_text(size = 14, family = "sans"),
    axis.title.x = element_text(size = 8, family = "mono"),
    axis.title.y = element_text(size = 8, family = "mono"),
    axis.text = element_text(size = 10, face = "bold", family = "mono"),
    plot.caption.position = "plot",
    legend.title = element_text(size = 10, face = "bold"),
    panel.spacing = unit(1.3, "lines"),
    plot.background = element_rect(fill = "white"),
    panel.background = element_rect(fill = "gray80"),
    panel.grid.major = element_line(color = "gray70"),
    panel.grid.minor = element_blank(),
    legend.position = "right",
    plot.margin = margin(10, 10, 10, 10)
  )
```




```
# save the plot
ggsave("output_vis/Distribution.pdf", width = 15, height = 12, dpi = 600)
ggsave("output_vis/Distribution.png", width = 15, height = 12, dpi = 600)
```

```
library(ggplot2)
library(dplyr)
library(readr)

#select the Residential building
selected_types <- c("3+ Family Apt. Building",
                    "1-2 Family Dwelling",
                    "3+ Family Mixed Use Building")

# filter the data from original data set, plot the stack with decreasing order
filtered_df <- ini_df %>%
  filter(`Location Type` %in% selected_types) %>%
  mutate(Year = format(`Created Date`, "%Y"),
         `Location Type` = factor(`Location Type`, levels = c("3+ Family Apt. Building",
                                                              "1-2 Family Dwelling",
                                                              "3+ Family Mixed Use Building"))) %>%
  count(Year, `Location Type`)

# plot the stack bar chart
ggplot(filtered_df, aes(x = Year, y = n, fill = `Location Type`)) +
  geom_bar(stat = "identity", position = "stack") +
  theme_minimal() +
```

```
  theme(
    plot.title = element_text(hjust = 0, size = 20, face = "bold", family = "serif"),
    plot.subtitle = element_text(size = 14, face = "italic", family = "sans"),
    plot.caption = element_text(size = 9, face = "italic", family = "mono", hjust = 0),
    axis.title = element_text(size = 14, family = "sans"),
    axis.title.x = element_text(size = 12, family = "mono"),
    axis.title.y = element_text(size = 12, family = "mono"),
    axis.text = element_text(size = 10, face = "bold", family = "mono"),
    plot.caption.position = "plot",
    legend.title = element_text(size = 10, face = "bold"),
    panel.spacing = unit(1.3, "lines"),
    plot.background = element_rect(fill = "white"),
    panel.background = element_rect(fill = "white"),
    panel.grid.major = element_line(color = "gray80"),
    panel.grid.minor = element_blank(),
    legend.position = "right",
    plot.margin = margin(10, 10, 10, 10)
  ) +
  scale_fill_manual(values = c("#d49391", "#a8b89a", "#a29bb3")) +
  labs(title = "Yearly Trends of Incidents in Residential Buildings",
       subtitle = "Rat Sightings Analysis of Residential Buildings (2010-2017)",
       x = "Year", y = "Count", fill = "Location Type")
```




```
# save the plot
ggsave("output_vis/Residential_trends_stack.pdf", width = 15, height = 12, dpi = 600)
```

```r
ggsave("output_vis/Residential_trends_stack.png", width = 15, height = 12, dpi = 600)
```

```r
library(readr)
library(sf)

suppressMessages(suppressWarnings({
  boroughs <- st_read("data/borough.geojson")
}))
```

```
## Reading layer ‘borough’ from data source
##   ‘D:\R_individual\A1_Sightings\data\borough.geojson’ using driver ‘GeoJSON’
## Simple feature collection with 5 features and 4 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -74.25545 ymin: 40.49613 xmax: -73.70002 ymax: 40.9154
## Geodetic CRS:  WGS 84
```

```r
library(dplyr)
library(lubridate)

# Read the weather data
weather_data <- read.csv("./data/weather_data.csv", stringsAsFactors = FALSE)

# Transform the DATE column to date format
weather_data$DATE <- as.Date(weather_data$DATE)

# filrter the weather data to the range 2010-01 to 2017-09
filtered_data <- weather_data %>%
  filter(DATE >= as.Date("2010-01-01") & DATE <= as.Date("2017-09-30"))

# calculate the daily average temperature(integer)
filtered_data <- filtered_data %>%
  mutate(AVG_TEMP = floor((TMAX + TMIN) / 2))

# calculate the monthly average temperature base on the daily average
weather_df <- filtered_data %>%
  mutate(Year_Month = format(DATE, "%Y-%m")) %>%
  group_by(Year_Month) %>%
  summarise(AVG_TEMP = round(mean(AVG_TEMP, na.rm = TRUE)))

# check the result
print(weather_df)
```

```
## # A tibble: 93 x 2
##    Year_Month AVG_TEMP
##    <chr>         <dbl>
##  1 2010-01          32
##  2 2010-02          33
##  3 2010-03          48
##  4 2010-04          58
##  5 2010-05          65
##  6 2010-06          74
```

```
##  7 2010-07          81
##  8 2010-08          77
##  9 2010-09          71
## 10 2010-10          58
## # i 83 more rows
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

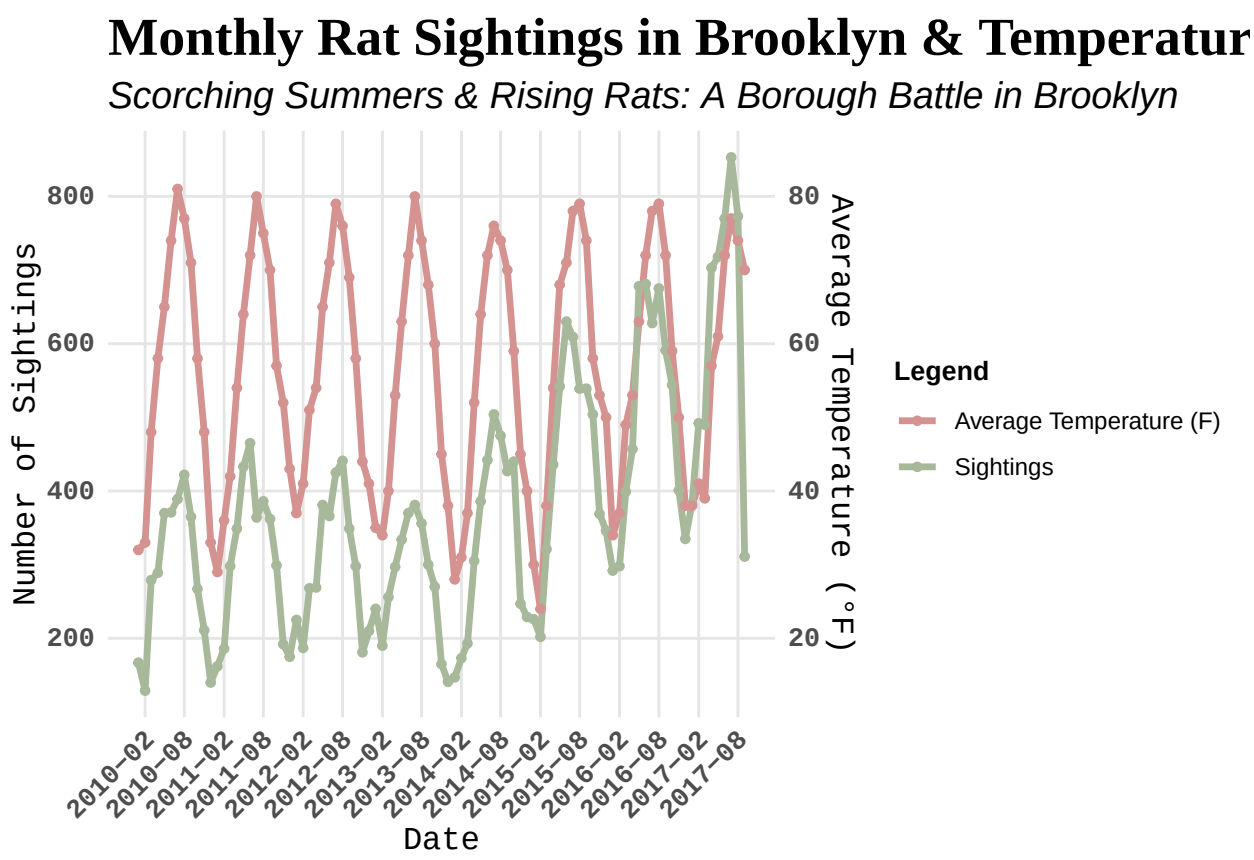
```
library(dplyr)

# filter the Brooklyn borough with a subset
df_brooklyn <- temp_df %>% filter(Borough == "Brooklyn")

# check the sub-dataset
head(df_brooklyn)
```

```
## # A tibble: 6 x 19
##   ‘Unique Key‘ ‘Created Date‘      ‘Location Type‘              ‘Incident Zip‘
##          <dbl> <dttm>              <chr>                        <chr>
## 1     31464026 2015-09-04 00:00:00 3+ Family Apt. Building      11206
## 2     31464188 2015-09-04 00:00:00 3+ Family Apt. Building      11231
## 3     31464199 2015-09-04 00:00:00 3+ Family Mixed Use Building 11205
## 4     31464803 2015-09-04 00:00:00 Commercial Building          11226
## 5     31464858 2015-09-04 00:00:00 3+ Family Apt. Building      11238
## 6     31464871 2015-09-05 00:00:00 Other (Explain Below)        11201
## # i 15 more variables: ‘Incident Address‘ <chr>, ‘Street Name‘ <chr>,
## #   ‘Cross Street 1‘ <chr>, ‘Cross Street 2‘ <chr>, ‘Address Type‘ <chr>,
## #   City <chr>, ‘Community Board‘ <chr>, Borough <chr>,
## #   ‘X Coordinate (State Plane)‘ <dbl>, ‘Y Coordinate (State Plane)‘ <dbl>,
## #   Latitude <dbl>, Longitude <dbl>, Location <chr>, Year <dbl>, Month <chr>
```

```
library(ggplot2)
library(dplyr)
library(lubridate)

# make sure the Created Date =column is DATE format
df_brooklyn$Created_Date <- as.Date(df_brooklyn$`Created Date`, format="%Y-%m-%d")

# abstract the year and month
df_brooklyn$YearMonth <- format(df_brooklyn$Created_Date, "%Y-%m")

# count the event for each month
brooklyn_monthly <- df_brooklyn %>%
  group_by(YearMonth) %>%
  summarise(Sightings = n()) %>%
  arrange(YearMonth)

# make sure the data in the YearMonth column is match the fomat in sub-Brookly set
weather_df <- weather_df %>%
```

```r
  mutate(YearMonth = format(as.Date(paste0(Year_Month, "-01")), "%Y-%m"))

# merge the data
brooklyn_weather <- brooklyn_monthly %>%
  left_join(weather_df, by = "YearMonth")

# check the Date format again after the merge operation
brooklyn_weather$YearMonth <- as.Date(paste0(brooklyn_weather$YearMonth, "-01"))

# adjust the format of data, transform the "Sightings" and "AVG_TEMP" to different variables
brooklyn_weather_long <- brooklyn_weather %>%
  mutate(AVG_TEMP_Scaled = AVG_TEMP * 10) %>%  # make temperature and sightings in the same data range
  select(YearMonth, Sightings, AVG_TEMP_Scaled) %>%
  pivot_longer(cols = c(Sightings, AVG_TEMP_Scaled),
               names_to = "Variable",
               values_to = "Value")


color_map <- c("AVG_TEMP_Scaled" = "#d49391","Sightings" = "#a8b89a" )

# plot the data
ggplot(brooklyn_weather_long, aes(x = YearMonth, y = Value, color = Variable, group = Variable)) +
  geom_line(size = 1.2) +
  geom_point(size = 1.2) +

  # theme
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0, size = 20, face = "bold", family = "serif"),
    plot.subtitle = element_text(size = 14, face = "italic", family = "sans"),
    plot.caption = element_text(size = 9, face = "italic", family = "mono", hjust = 0),
    axis.title = element_text(size = 14, family = "sans"),
    axis.title.x = element_text(size = 12,family = "mono"),
    axis.title.y = element_text(size = 12,family = "mono"),
    axis.text.x = element_text(angle = 45,  hjust = 1,size = 10, face = "bold", family = "mono"),
    axis.text.y = element_text(size = 10, face = "bold", family = "mono"),
    plot.caption.position = "plot",
    legend.position = "right",
    legend.title = element_text(size = 10, face = "bold"),
    panel.spacing = unit(1.3, "lines"),
    panel.grid.major = element_line(color = "gray90"),
    panel.grid.minor = element_blank()
  ) +

  # label and title
  labs(
    title = "Monthly Rat Sightings in Brooklyn & Temperature Trends",
    subtitle = "Scorching Summers & Rising Rats: A Borough Battle in Brooklyn",
    x = "Date",
    y = "Number of Sightings",
    color = "Legend"
  ) +
```

```
  scale_x_date(date_labels = "%Y-%m", date_breaks = "6 months") +

   scale_color_manual(values = color_map, labels = c( "Average Temperature (F)","Sightings"))+

  scale_y_continuous(
    sec.axis = sec_axis(~ . / 10, name = "Average Temperature (°F)")
  )
```




```
ggsave("output_vis/brooklyn_weather_sightings.pdf", width = 15, height = 12, dpi = 600)
ggsave("output_vis/brooklyn_weather_sightings.png", width = 15, height = 12, dpi = 600)
```

```
library(extrafont)
```

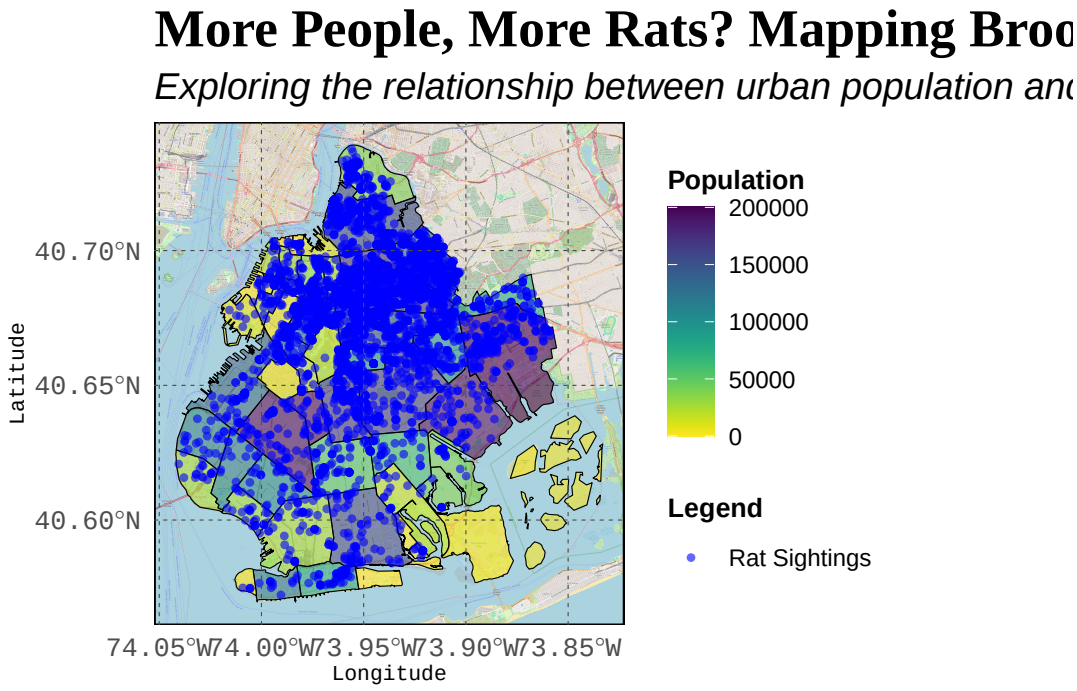
```
library(ggplot2)
library(sf)
library(dplyr)
library(lubridate)
library(ggspatial)  # OpenStreetMap for map layer
library(viridis)  # for color assign

# read the Brooklyn community data with the population from wikipidia
brooklyn_geo <- st_read("./data/brooklyn_updated.geojson")
```

```
## Reading layer 'brooklyn_updated' from data source
```

```
##   'D:\R_individual\A1_Sightings\data\brooklyn_updated.geojson'
##   using driver 'GeoJSON'
## Simple feature collection with 54 features and 5 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -74.0419 ymin: 40.56953 xmax: -73.83304 ymax: 40.73913
## Geodetic CRS:  WGS 84
```

```r
# read the rat sighting events from the data frame we set before
rats_df <- df_brooklyn
rats_df$Created_Date <- as.Date(rats_df$Created_Date, format="%Y-%m-%d")

# filter the data for 2017
rats_july_2017 <- rats_df %>%
  filter(year(Created_Date) == 2017 ) %>%
  select(Latitude, Longitude) %>%
  na.omit()

# transform the rat data into sf object
rats_sf <- st_as_sf(rats_july_2017, coords = c("Longitude", "Latitude"), crs = 4326)

# filter the data again with the coordinator make sure the data point is in the Brooklyn range
brooklyn_rats <- rats_sf[apply(st_within(rats_sf, brooklyn_geo, sparse = FALSE), 1, any), ]


ggplot() +
  # add OpenStreetMap as backgroud layer
  annotation_map_tile(type = "osm", zoom = 14) +

  # population layer
  geom_sf(data = brooklyn_geo, aes(fill = population), color = "black", size = 0.5, alpha = 0.6) +

  # set the color for the level of population
  scale_fill_viridis(option = "viridis", direction = -1, na.value = "grey80") +

  # rat sighting layer
  geom_sf(data = brooklyn_rats,aes(color = "Rat Sightings", shape = "Rat Sightings"), size = 1.2, alpha

  scale_color_manual(name = "Legend", values = c("Restaurants" = "blue", "Rat Sightings" = "blue")) +
  scale_shape_manual(name = "Legend", values = c("Restaurants" = 17, "Rat Sightings" = 16)) +

  # add the title and label and reference
  labs(title = "More People, More Rats? Mapping Brooklyn's Population & Infestation Hotspots",
       subtitle = "Exploring the relationship between urban population and Rat activity in Brooklyn's du
       fill = "Population",
       caption =
       "
Date Range:  2017
Data Source: https://www.kaggle.com/datasets/new-york-city/nyc-rat-sightings/data
             https://github.com/blackmad/neighborhoods/blob/master/brooklyn.geojson
             https://www.wikipedia.org/
       ",
       x = "Longitude", y = "Latitude") +
  theme_minimal() +
```

```
  theme(
    # make sure the grid is on the top layer
    panel.ontop = TRUE,

    # transplate the color make sure the background map is visible
    panel.background = element_rect(fill = NA),

    # theme
    panel.grid.major = element_line(color = "gray30",linetype = 'dashed', size = 0.2),
    panel.grid.minor = element_line(color = "gray30",linetype = 'dashed', size = 0.2) ,
    plot.title = element_text(hjust = 0, size = 20, face = "bold", family = "serif"),
    plot.subtitle = element_text(size = 14, face = "italic", family = "sans"),
    plot.caption = element_text(size = 9, face = "italic", family = "mono", hjust = 0),
    axis.title = element_text(size = 14, family = "sans"),
    axis.title.x = element_text(size = 8,family = "mono"),
    axis.title.y = element_text(size = 8,family = "mono"),
    axis.text = element_text(size = 10, face = "bold", family = "mono"),
    plot.caption.position = "plot",
    legend.position = "right",
    legend.title = element_text(size = 10, face = "bold"),
    panel.spacing = unit(1.3, "lines")
  )
```



Date Range: *2017*
Data Source: *https://www.kaggle.com/datasets/new-york-city/nyc-rat-sightir*
https://github.com/blackmad/neighborhoods/blob/master/brookly
https://www.wikipedia.org/

```
ggsave("output_vis/brooklyn_population_density_map.pdf", width = 15, height = 12, dpi = 600)
ggsave("output_vis/brooklyn_population_density_map.png", width = 15, height = 12, dpi = 600)
```