Assignment A1: Rat Sightings

Abi Joshua George (46656697)

2025-03-01

Embedded Code Chunks for Summarizing the Data and Creating the Visualization

```
# Loading the required libraries
library(tidyverse)
library(lubridate)
library(ggrepel)
```

```
# Reading the dataset and handling missing values
data <- read_csv("data/A1_Sightings.csv", na = c("", "NA", "N/A"))

# Converting 'Created Date' column to DateTime format
data$created_date <- mdy_hms(data$`Created Date`)

# Displaying the first few rows
head(data)
```

```
## # A tibble: 6 x 53
##   `Unique Key` `Created Date`         `Closed Date`         Agency `Agency Name`
##          <dbl> <chr>                  <chr>                 <chr>  <chr>
## 1     31464015 09/04/2015 12:00:00 AM 09/18/2015 12:00:00 ~ DOHMH  Department o~
## 2     31464024 09/04/2015 12:00:00 AM 10/28/2015 12:00:00 ~ DOHMH  Department o~
## 3     31464025 09/04/2015 12:00:00 AM <NA>                  DOHMH  Department o~
## 4     31464026 09/04/2015 12:00:00 AM 09/14/2015 12:00:00 ~ DOHMH  Department o~
## 5     31464027 09/04/2015 12:00:00 AM 09/22/2015 12:00:00 ~ DOHMH  Department o~
## 6     31464188 09/04/2015 12:00:00 AM 09/22/2015 12:00:00 ~ DOHMH  Department o~
## # i 48 more variables: `Complaint Type` <chr>, Descriptor <chr>,
## #   `Location Type` <chr>, `Incident Zip` <dbl>, `Incident Address` <chr>,
## #   `Street Name` <chr>, `Cross Street 1` <chr>, `Cross Street 2` <chr>,
## #   `Intersection Street 1` <chr>, `Intersection Street 2` <chr>,
## #   `Address Type` <chr>, City <chr>, Landmark <lgl>, `Facility Type` <lgl>,
## #   Status <chr>, `Due Date` <chr>, `Resolution Action Updated Date` <chr>,
## #   `Community Board` <chr>, Borough <chr>, ...
```

```
# Creating new date-related variables
data <- data %>%
  mutate(
    sighting_year = year(created_date),
```

```
    sighting_month = month(created_date, label = TRUE),
    sighting_weekday = wday(created_date, label = TRUE)
  )

# Displaying the first few rows with new columns
head(data)
```

```
## # A tibble: 6 x 56
##   `Unique Key` `Created Date`         `Closed Date`         Agency `Agency Name`
##          <dbl> <chr>                  <chr>                 <chr>  <chr>
## 1     31464015 09/04/2015 12:00:00 AM 09/18/2015 12:00:00 ~ DOHMH  Department o~
## 2     31464024 09/04/2015 12:00:00 AM 10/28/2015 12:00:00 ~ DOHMH  Department o~
## 3     31464025 09/04/2015 12:00:00 AM <NA>                  DOHMH  Department o~
## 4     31464026 09/04/2015 12:00:00 AM 09/14/2015 12:00:00 ~ DOHMH  Department o~
## 5     31464027 09/04/2015 12:00:00 AM 09/22/2015 12:00:00 ~ DOHMH  Department o~
## 6     31464188 09/04/2015 12:00:00 AM 09/22/2015 12:00:00 ~ DOHMH  Department o~
## # i 51 more variables: `Complaint Type` <chr>, Descriptor <chr>,
## #   `Location Type` <chr>, `Incident Zip` <dbl>, `Incident Address` <chr>,
## #   `Street Name` <chr>, `Cross Street 1` <chr>, `Cross Street 2` <chr>,
## #   `Intersection Street 1` <chr>, `Intersection Street 2` <chr>,
## #   `Address Type` <chr>, City <chr>, Landmark <lgl>, `Facility Type` <lgl>,
## #   Status <chr>, `Due Date` <chr>, `Resolution Action Updated Date` <chr>,
## #   `Community Board` <chr>, Borough <chr>, ...
```

```
# Getting unique values of Location Type
unique_location_types <- unique(data$`Location Type`)

# Displaying the unique location types
print(unique_location_types)
```

```
##  [1] "3+ Family Mixed Use Building"  "Commercial Building"
##  [3] "1-2 Family Dwelling"           "3+ Family Apt. Building"
##  [5] "Public Stairs"                 "Other (Explain Below)"
##  [7] "Vacant Lot"                    "Construction Site"
##  [9] "Hospital"                      "Parking Lot/Garage"
## [11] "Catch Basin/Sewer"             "Vacant Building"
## [13] "1-2 Family Mixed Use Building" "Public Garden"
## [15] "Government Building"           "Office Building"
## [17] "School/Pre-School"             "Day Care/Nursery"
## [19] "Single Room Occupancy (SRO)"   "Summer Camp"
## [21] NA
```

```
data <- data %>%
  mutate(Dwelling_Type = case_when(
    `Location Type` %in% c("3+ Family Mixed Use Building", "3+ Family Apt. Building",
                           "1-2 Family Dwelling", "1-2 Family Mixed Use Building") ~ "Residential",
    `Location Type` %in% c("Commercial Building", "Office Building",
                           "Single Room Occupancy (SRO)") ~ "Commercial",
    `Location Type` %in% c("Public Stairs", "Public Garden", "School/Pre-School",
                           "Day Care/Nursery", "Hospital", "Government Building") ~ "Public Spaces",
    `Location Type` %in% c("Vacant Lot", "Vacant Building", "Construction Site",
                           "Parking Lot/Garage", "Catch Basin/Sewer", "Other (Explain Below)") ~ "Indus
```

```
    TRUE ~ NA_character_  # Assigns NA to unclassified values
  )) %>%
  filter(!is.na(Dwelling_Type))  # Remove NA dwelling types
```

```
# Summarizing the sightings by Borough and Year
rat_summary <- data %>%
  filter(!is.na(sighting_year)) %>%
  group_by(Borough, sighting_year, Dwelling_Type) %>%
  summarize(total_sightings = n(), .groups = 'drop')

# Displaying the summary
head(rat_summary)
```

```
## # A tibble: 6 x 4
##   Borough sighting_year Dwelling_Type    total_sightings
##   <chr>           <dbl> <chr>                      <int>
## 1 BRONX            2010 Commercial                    70
## 2 BRONX            2010 Industrial/Other             381
## 3 BRONX            2010 Public Spaces                 15
## 4 BRONX            2010 Residential                 1601
## 5 BRONX            2011 Commercial                    69
## 6 BRONX            2011 Industrial/Other             368
```

```
# Summarizing the sightings by Borough and Year
rat_summary <- data %>%
  filter(!is.na(sighting_year)) %>%
  group_by(Borough, sighting_year) %>%
  summarize(total_sightings = n(), .groups = 'drop')

# Displaying the summary
head(rat_summary)
```

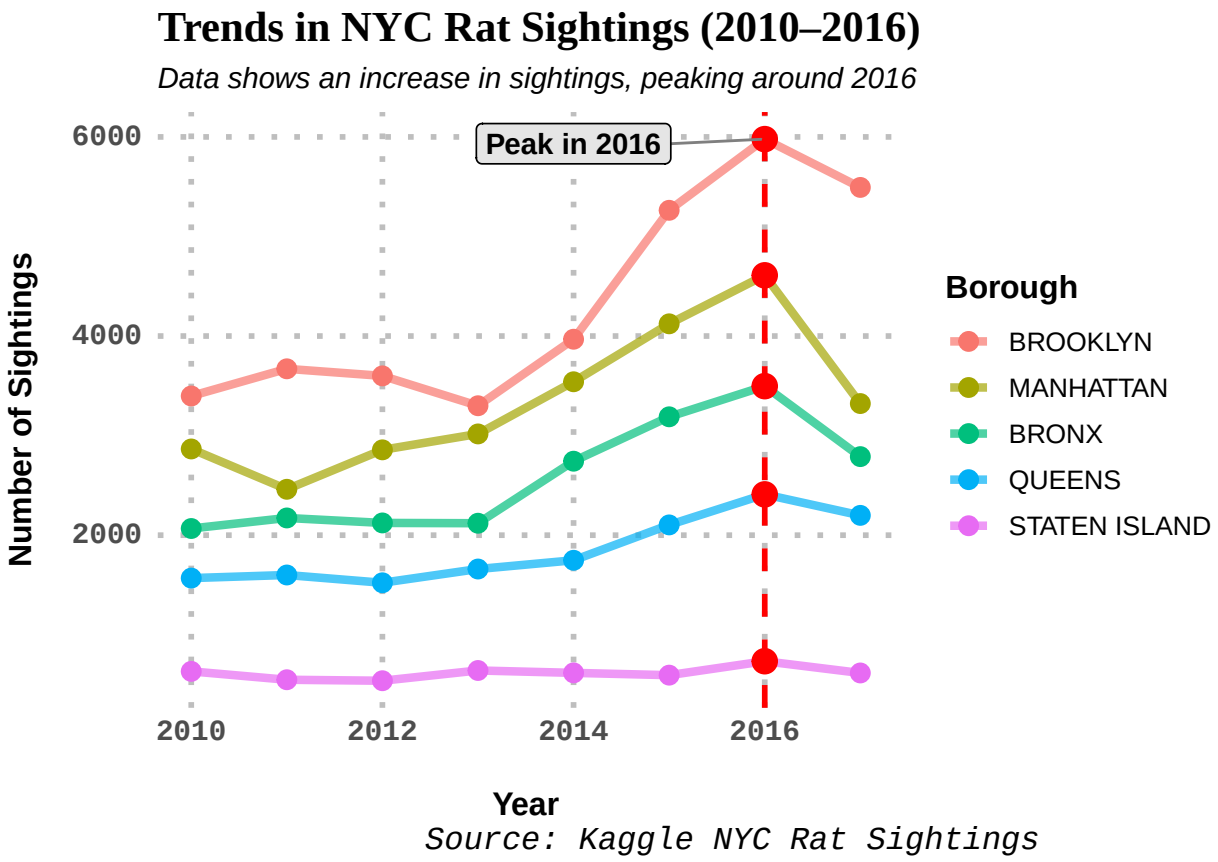
```
## # A tibble: 6 x 3
##   Borough sighting_year total_sightings
##   <chr>           <dbl>           <int>
## 1 BRONX            2010            2067
## 2 BRONX            2011            2173
## 3 BRONX            2012            2123
## 4 BRONX            2013            2120
## 5 BRONX            2014            2743
## 6 BRONX            2015            3189
```

```
# Reordering the factor levels based on the peak total_sightings for each Borough
rat_summary$Borough <- factor(rat_summary$Borough, levels = rat_summary %>%
                                group_by(Borough) %>%
                                summarize(max_sightings = max(total_sightings)) %>%
                                arrange(desc(max_sightings)) %>%
                                pull(Borough))
```

Visualization (Final Figure)

```
# Filtering the data for 2016 peaks
peak_data <- rat_summary %>% filter(sighting_year == 2016)

# Identifying the highest peak borough for labeling
highest_peak <- peak_data %>% filter(total_sightings == max(total_sightings))

ggplot(subset(rat_summary, Borough != "Unspecified"),
       aes(x = sighting_year, y = total_sightings, color = Borough, group = Borough)) +
  geom_line(size = 1.5, alpha = 0.7) +  # Thicker lines with transparency
  geom_point(size = 3) +  # Larger points for emphasis

  # Highlight 2016 peak points in red
  geom_point(data = peak_data, aes(x = sighting_year, y = total_sightings),
             color = "red", size = 4, shape = 21, fill = "red") +

  # Add a single repelled label inside a text box for the highest peak borough (Brooklyn)
  geom_label_repel(data = highest_peak, aes(x = sighting_year, y = total_sightings, label = "Peak in 20
                   size = 4, fontface = "bold",
                   nudge_x = -2, nudge_y = 0, color = "black",
                   fill = "gray90",
                   segment.angle = 0,
                   box.padding = 0.3,  # Adjusting padding for better fit
                   label.size = 0.3,   # Slightly increasing label border
                   segment.color = "gray50",
                   segment.size = 0.5) +

  # Add a highlighted vertical dashed line for the year 2016
  geom_vline(xintercept = 2016, color = "red", linetype = "dashed", size = 1) +

  labs(
    title = "Trends in NYC Rat Sightings (2010-2016)",
    subtitle = "Data shows an increase in sightings, peaking around 2016",
    x = "Year",
    y = "Number of Sightings",
    color = "Borough",
    caption = "Source: Kaggle NYC Rat Sightings"
  ) +
  theme_minimal(base_size = 13) +
theme(
    legend.position = "right",
    legend.title = element_text(size = 12, face = "bold"),
    plot.title = element_text(face = "bold", size = 16, family = "serif"),
    plot.subtitle = element_text(size = 11, family = "sans", face = "italic"),
    plot.caption = element_text(size = 12, face = "italic", family = "mono", hjust = 2.2, vjust = -1.5)
    panel.grid.major.x = element_line(size = 1, color = "gray", linetype = "dotted"),
    panel.grid.major.y = element_line(size = 1, color = "gray", linetype = "dotted"),
    panel.grid.minor = element_blank(), #
    panel.spacing = unit(1.5, "lines"),
    axis.title = element_text(size = 12, family = "sans"),
    axis.title.x = element_text(size = 12,face = "bold",margin = margin(t = 10, r = 5),vjust = -2 ),
    axis.title.y = element_text(size = 12,face = "bold",margin = margin(r = 10),  vjust = 2),
```

```
    axis.text = element_text(size = 11, face = "bold", family = "mono"),
    strip.text = element_text(size = 12,face = "bold", family = "mono")
  )
```



Saving the Visualization as PNG and PDF

```
ggsave("output/A1_final.png", width = 12, height = 8, dpi = 300)
ggsave("output/A1_final.pdf", width = 12, height = 8, dpi = 300)
```

Memo

Introduction:

New York City monitors rat sightings with care, and records them in a wide dataset that offers informative data on urban rodent wildlife patterns. The objective of this analysis was to find meaningful trends in rat sightings over time, categorize sightings by borough and dwelling type, and present the findings in an informative and engaging visualization. The original dataset consisted of over 100,000 rows and required a lot of preprocessing and summarization in order to be suitable before inferring useful patterns from it. This memo outlines the cleaning and data structuring process, and how the final figure effectively tells a superior and more informative story.

Data Summarization Process:

The raw data contained a number of attributes like date of sighting, borough, and location type. The data summarization process involved the following key steps:

Data Loading and Cleaning,

- The data was loaded into R, and missing values were handled by treating empty cells and "N/A" as NA.
- The 'Created Date' column was converted to a proper DateTime format using the lubridate package.

Generating Temporal Features,

- Parsed year, month, and day of week out of the 'Created Date' column to allow time-based analysis.

Location Type Categorization,

- The 'Location Type' column contained very specific descriptions (e.g., "3+ Family Apt. Building", "Public Stairs").
- These were categorized into four high-level categories: Residential, Commercial, Public Spaces, and Industrial/Other.

Summarizing data by Borough, Year, and Dwelling Type,

- Binned the number of rat sightings per year per borough using group_by() and summarize().
- Removed any missing or misplaced types of dwellings.
- This structured data provided a neater and more informative basis for visualization.

New Visualization and Its Story:

A new, better visualization was created, showing the trends in a more clearer and structured manner. The most significant improvements are:

- **Trend Highlighting:** A dashed vertical line was included to mark 2016, the year with the most rat sightings, with a label annotation.
- **Increased Color Discrimination:** Every borough had a distinct color, making comparison easier.
- **Decluttering and Legibility:** The visualization provided enough spacing so that both axes and legend could be easily read.
- **Storytelling through Data:** The graph easily illustrates how rat sightings increased over time and peaked in 2016 and then reduced.

Application of CRAP Principles:

The CRAP principles (Contrast, Repetition, Alignment, and Proximity) were applied such that the final visualization was aesthetically pleasing as well as functional:

- **Contrast:** Strong contrast between boroughs through choice of color.
- **Repetition:** Consistent font faces, gridlines, and axis styling ensured uniformity.
- **Alignment:** Legend, title, and labels aligned in a consistent manner.
- **Proximity:** Annotation of the 2016 peak was placed near the respective data point without obscuring other data.

Application of Kieran Healy's Principles:

Kieran Healy's principles for good data visualization have been employed as follows:

- **Clarity:** The visualization was intuitive so that the audience could easily read trends at a glance.
- **Comparability:** The clear borough-wise color coding made direct comparisons across various areas possible.
- **Emphasis on Key Trends:** The 2016 peak was highlighted so that the audience could easily see the most important insight.
- **Use of Context:** Annotations and labels provided context, allowing the audience to read the significance of changes in rat sightings over time.

Application of Alberto Cairo's Principles:

Alberto Cairo's five elements of good visualizations were included in the design:

- **Truthful:** The information is being depicted in an unmanipulated and undistorted way.
- **Functional:** The visualization effectively conveys borough-by-borough trends across time with low cognitive load meaning that it is easy on the eyes of the viewer.
- **Beautiful:** Balanced design choices, thoughtful spacing, and the use of attractive colors support readability.
- **Insightful:** The identification of the peak year provides immediate understanding of a significant data trend.
- **Educational:** It is easy for viewers to identify differences between boroughs and long-term trends in rat sightings, increasing their knowledge of urban rodent activity.

Conclusion: By structured data summarization and visualization design, this final figure presents a clearer and more effective story of NYC rat sightings. The structured data, improved layout, and thoughtful annotations provide significant information on the increasing trend of rat sightings, peaking in 2016. Finally, this figure makes the key findings both usable and interpretable to the audience and raises their awareness of NYC urban rodent trends.