Assignment A1

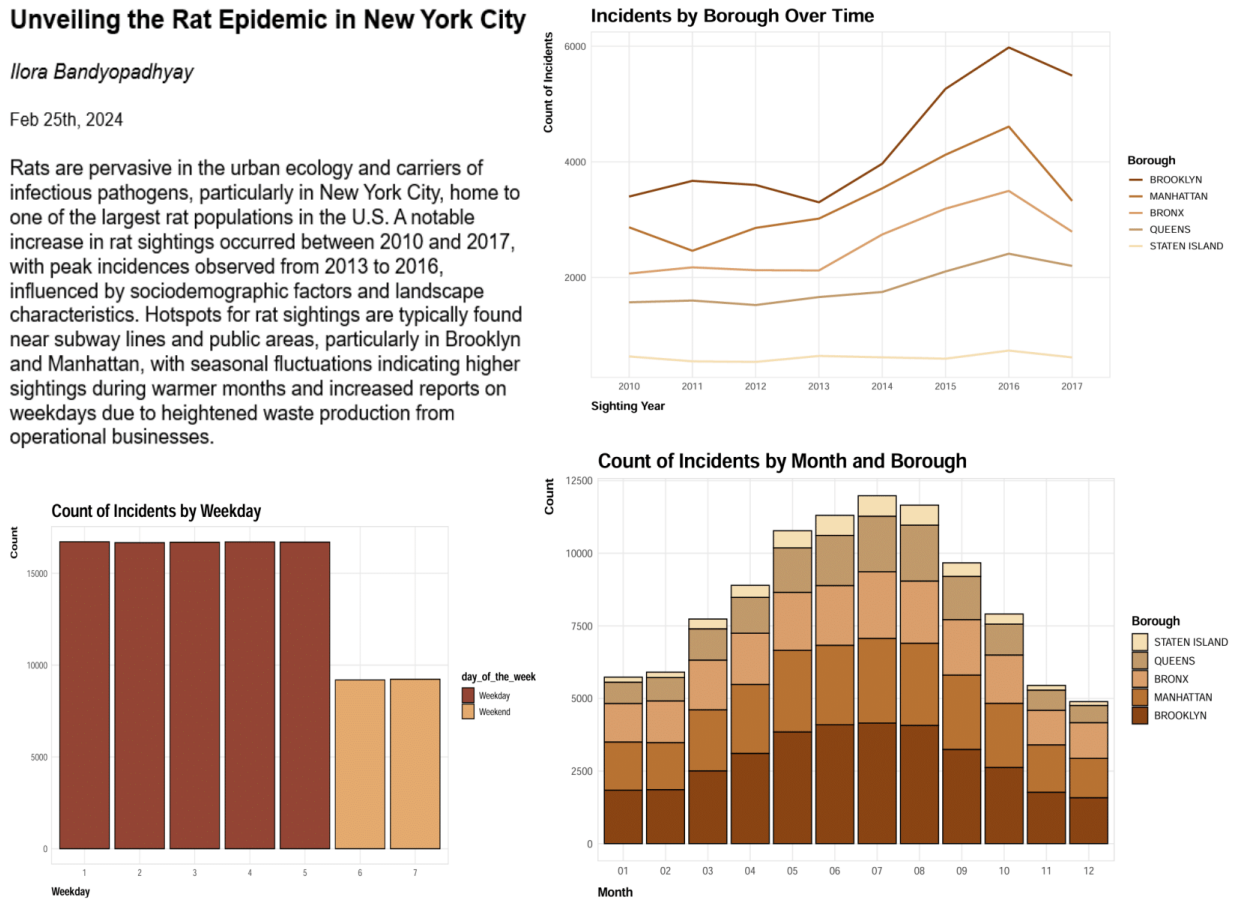
Ilora Bandyopadhyay

02/24/2024

Executive Summary

Unveiling the Rat Epidemic in New York City

```
knitr::include_graphics("output/A1_graphs_Rat sightings.png")
```



Changes made to one of the original graphic (pie chart for Rat sightings across NYC)

The rat sightings in every borough faceted by year presentation in pie charts was misleading as this does not have clarity in conveying the yearly trend. Pie charts are not effective in representing trend on time series. Also, it is difficult to visualize the difference in values in a pie chart briefly. For showing trend over year,

line chart is the most appropriate and that has been used. Also, for each borough a separate line chart has been plotted that also shows how the incidents varied over year in each borough.

The story with new graphic: Unveiling the Rat Epidemic in New York City

Rats are pervasive in the urban ecology and are carriers of infectious pathogens. Among cities in the United States, New York city hosts one of the most extensive rat populations. This is of great concern for public health. Between the years 2010 and 2017, New York city has experienced a significant surge in rat sightings. The years 2013 to 2016 has witnessed a significant spike in rat sightings due to sociodemographic and physical landscape. Rats are highly prevalent near subway lines and public spaces [1]. Brooklyn and Manhattan have the highest number of subways [2]. This reflects in our analysis where we see Brooklyn is the borough with the highest incidents of rat sightings followed by Manhattan, Bronx, Queens and Staten Island. Moreover, the months May, June, July and August have the highest incidences and November and December have the lowest incidences of rat sightings as observed in our analysis. The possible reason is the breeding season during warm season (like late summers and early fall) and also rats aggressively find food to stockpile for the winters [3]. We see in our analysis that weekends have lower number of reported rat sightings than weekends, as weekends have lower garbage and food disposal than weekdays, when businesses are operational [4]. Hence, the analysis performed gives an insight on the rat epidemic in New York city.

The principles of CRAP

The principles of CRAP (Contrast, Repetition, Alignment, Proximity) have been meticulously applied while designing the visualizations for rat sightings in NYC. This has been done to ensure effective presentation of data, by using distinctive colours and contrast and different font to identify data and the underlying insights from each borough in New York City. Same colour scheme for every borough has been used to ensure repetition. Like, in the line chart incidents over time, all the boroughs have been represented in different shades of brown. In fact, in the other charts wherever boroughs have been represented, the same colour combination have been used. The charts have been clearly aligned and organized and elements have been positioned for a clear logical flow of the eyes. Proximity has been employed for grouping of related elements together and strategic placement and visual cues were used to indicate the relatable data points and annotations and facilitating a clear, intuitive, and easy understanding of complex data relationships. The CRAP design principles have been appropriately used to enhance the report and make it easily and effectively comprehensible for the reader.

Kieran Healy's principles of great visualizations:

Kieran Healy's principles of effective visualization were applied to analyse rat sightings in New York city. It was ensured that the visualizations were clear, comprehensible, and relevant. In the first two graphs, line graph has been chosen to represent trend of incidents of rat sightings over time. This is aligned with the selection of an appropriate graph for time series. This graph helped to understand trend over the boroughs. Distinct shades of brown have been used to represent each borough. All visualizations have been kept in shades of brown as brown has a natural association with dirt, grime and decay which is strongly associated with rat infestation. Also, dark shades of brown show a good contrast with light shades, like shown in the graphs, with higher incidents of rat sightings darker shades of brown have been used and lighter shades for lower values of the same. This helped to prevent confusion and enhance intuitive comparison. According to Healy's principles, the visualization process aimed to have a balance between complexity and simplicity. This helped to convey deep insights in a simple way for audiences, even without specialized knowledge to interpret the data. The equilibrium was maintained to inform and engage the audiences on the rat sightings in the city.

```
# Load necessary libraries
library(dplyr)
library(tidyverse)
library(lubridate)
library(scales) #for nice labels in charts
library(countrycode)
library(plotly)
```

```r
library(rnaturalearth)
library(sf)
library(ggplot2)



rat_data <- read_csv("data/A1_sightings.csv")

rat_data
```

```
## # A tibble: 101,914 x 52
##    `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~
##  7     31464195 09/04/2015 12:00:00 AM 09/22/2015 04:26:36~ DOHMH  Department o~
##  8     31464199 09/04/2015 12:00:00 AM <NA>                 DOHMH  Department o~
##  9     31464220 09/04/2015 12:00:00 AM 08/19/2015 12:00:00~ DOHMH  Department o~
## 10     31464802 09/04/2015 12:00:00 AM 09/25/2015 12:00:00~ DOHMH  Department o~
## # i 101,904 more rows
## # i 47 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` <chr>,
## #   Status <chr>, `Due Date` <chr>, `Resolution Action Updated Date` <chr>, ...
```

```r
na_values <- c("", "NA", "N/A")
rat_data <- replace(rat_data, rat_data %in% na_values, NA)
```

```r
# Rename columns
colnames(rat_data)[colnames(rat_data) == "Location Type"] <- "location_type"
colnames(rat_data)[colnames(rat_data) == "Incident Zip"] <- "incident_zip"
colnames(rat_data)[colnames(rat_data) == "Created Date"] <- "created_date"
```

```r
# Convert 'Created Date' to Date format
rat_data$created_date <- as.Date(rat_data$created_date, format = "%m/%d/%Y")

# Extract year, month, and day
rat_data$sighting_year <- format(rat_data$created_date, "%Y")
rat_data$sighting_month <- format(rat_data$created_date, "%m")
rat_data$sighting_date <- format(rat_data$created_date, "%d")

# Display the modified dataframe
print(rat_data)
```

```
## # A tibble: 101,914 x 55
##    `Unique Key` created_date `Closed Date` Agency `Agency Name` `Complaint Type`
```

```
##           <dbl> <date>       <chr>          <chr>  <chr>          <chr>
##  1     31464015 2015-09-04   09/18/2015 1~ DOHMH  Department o~ Rodent
##  2     31464024 2015-09-04   10/28/2015 1~ DOHMH  Department o~ Rodent
##  3     31464025 2015-09-04   <NA>          DOHMH  Department o~ Rodent
##  4     31464026 2015-09-04   09/14/2015 1~ DOHMH  Department o~ Rodent
##  5     31464027 2015-09-04   09/22/2015 1~ DOHMH  Department o~ Rodent
##  6     31464188 2015-09-04   09/22/2015 1~ DOHMH  Department o~ Rodent
##  7     31464195 2015-09-04   09/22/2015 0~ DOHMH  Department o~ Rodent
##  8     31464199 2015-09-04   <NA>          DOHMH  Department o~ Rodent
##  9     31464220 2015-09-04   08/19/2015 1~ DOHMH  Department o~ Rodent
## 10     31464802 2015-09-04   09/25/2015 1~ DOHMH  Department o~ Rodent
## # i 101,904 more rows
## # i 49 more variables: 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` <chr>,
## #   Status <chr>, `Due Date` <chr>, `Resolution Action Updated Date` <chr>, ...
```

```r
# Subset dataframe to keep only specified columns
rat_data_subset <- rat_data[, c("created_date", "location_type","sighting_year","sighting_month","sighti

# Display the modified dataframe
print(rat_data_subset)
```

```
## # A tibble: 101,914 x 7
##    created_date location_type         sighting_year sighting_month sighting_date
##    <date>       <chr>                 <chr>         <chr>          <chr>
##  1 2015-09-04   3+ Family Mixed Use ~ 2015          09             04
##  2 2015-09-04   Commercial Building   2015          09             04
##  3 2015-09-04   1-2 Family Dwelling   2015          09             04
##  4 2015-09-04   3+ Family Apt. Build~ 2015          09             04
##  5 2015-09-04   3+ Family Mixed Use ~ 2015          09             04
##  6 2015-09-04   3+ Family Apt. Build~ 2015          09             04
##  7 2015-09-04   3+ Family Apt. Build~ 2015          09             04
##  8 2015-09-04   3+ Family Mixed Use ~ 2015          09             04
##  9 2015-09-04   Public Stairs         2015          09             04
## 10 2015-09-04   1-2 Family Dwelling   2015          09             04
## # i 101,904 more rows
## # i 2 more variables: incident_zip <dbl>, Borough <chr>
```

```r
# Convert 'Created_Date' to Date format
rat_data_subset$created_date <- as.Date(rat_data$created_date, format = "%m/%d/%Y")

# Extract weekday
rat_data_subset$sighting_weekday <- weekdays(rat_data$created_date)

# Map weekdays to numerical values (1-5 for Monday to Friday, 6-7 for weekend)
rat_data_subset$sighting_weekday <- ifelse(rat_data_subset$sighting_weekday %in% c("Monday", "Tuesday",
                                    1:5,
                                    ifelse(rat_data_subset$sighting_weekday %in% c("Saturday", "Sunday"

# Display the modified dataframe
print(rat_data_subset)
```

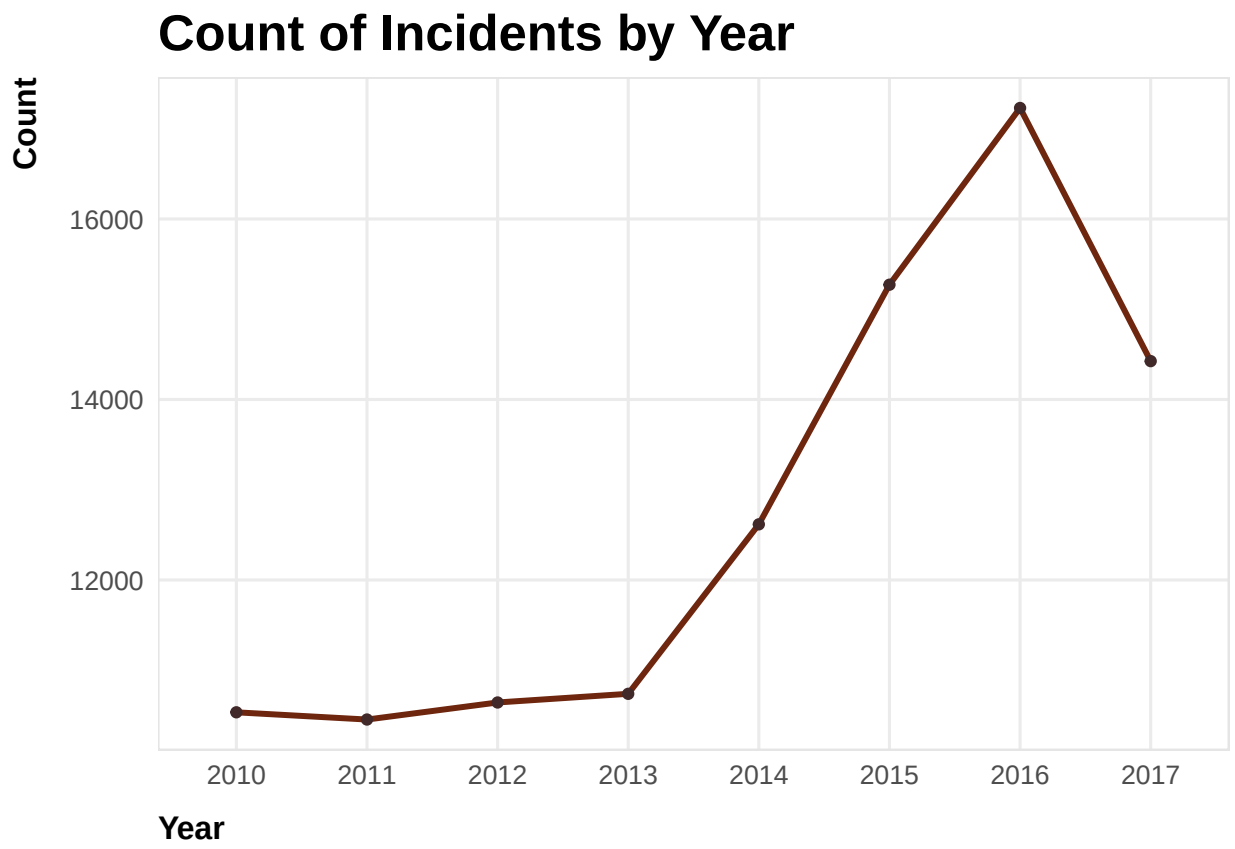
```
## # A tibble: 101,914 x 8
##    created_date location_type         sighting_year sighting_month sighting_date
##    <date>       <chr>                 <chr>         <chr>          <chr>
##  1 2015-09-04   3+ Family Mixed Use ~ 2015          09             04
##  2 2015-09-04   Commercial Building   2015          09             04
##  3 2015-09-04   1-2 Family Dwelling   2015          09             04
##  4 2015-09-04   3+ Family Apt. Build~ 2015          09             04
##  5 2015-09-04   3+ Family Mixed Use ~ 2015          09             04
##  6 2015-09-04   3+ Family Apt. Build~ 2015          09             04
##  7 2015-09-04   3+ Family Apt. Build~ 2015          09             04
##  8 2015-09-04   3+ Family Mixed Use ~ 2015          09             04
##  9 2015-09-04   Public Stairs         2015          09             04
## 10 2015-09-04   1-2 Family Dwelling   2015          09             04
## # i 101,904 more rows
## # i 3 more variables: incident_zip <dbl>, Borough <chr>, sighting_weekday <int>
```

```
my_pretty_theme <- theme_minimal(base_family = "Helvetica", base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        # Bold, bigger title
        plot.title = element_text(face = "bold", size = rel(1.6)),
        # Plain, slightly bigger subtitle that is grey
        plot.subtitle = element_text(face = "plain", size = rel(1.1), color = "grey70"),
        # Italic, smaller, grey caption that is left-aligned
        plot.caption = element_text(face = "italic", size = rel(0.7),
                                    color = "grey70", hjust = 0),
        # Bold legend titles
        legend.title = element_text(face = "bold"),
        # Bold, slightly larger facet titles that are left-aligned for the sake of repetition
        strip.text = element_text(face = "bold", size = rel(1.2), hjust = 0),
        # Bold axis titles
        axis.title = element_text(face = "bold"),
        # Add some space above the x-axis title and make it left-aligned
        axis.title.x = element_text(margin = margin(t = 10), hjust = 0),
        # Add some space to the right of the y-axis title and make it top-aligned
        axis.title.y = element_text(margin = margin(r = 10), hjust = 1),
        # Add a light grey background to the facet titles, with no borders
        strip.background = element_rect(fill = "grey90", color = NA),
        # Add a thin grey border around all the plots to tie in the facet titles
        panel.border = element_rect(color = "grey90", fill = NA))
```

```
rat_data_subset$sighting_year <- as.factor(rat_data_subset$sighting_year)

# Plot line chart
plot <- ggplot(rat_data_subset, aes(x = sighting_year)) +
  geom_line(stat = "count", aes(group = 1), color = "#6E260E", size=1.05) +
  geom_point(stat = "count", color = "#422929") +
  labs(title = "Count of Incidents by Year",
       x = "Year",
       y = "Count") +
  theme_minimal()+
  my_pretty_theme

plot
```




```
# Group the data by Borough and count the number of incidents in each Borough
incident_counts <- rat_data_subset %>%
  group_by(Borough) %>%
  summarise(Incident_Count = n())

# Print the Borough and the number of incidents in each Borough
print(incident_counts)
```

```
## # A tibble: 6 x 2
##   Borough       Incident_Count
##   <chr>                  <int>
## 1 BRONX                  20706
## 2 BROOKLYN               34673
## 3 MANHATTAN              26803
## 4 QUEENS                 14811
## 5 STATEN ISLAND           4920
## 6 Unspecified                1
```

Since there is only one incidence of Unspecified Boprough, hence filtering it out before further analysis.

```
#Plotting trend over years for boroughs in New York City
# Filter out 'Unspecified' borough
filtered_data <- rat_data_subset %>%
  filter(Borough != "Unspecified")
```

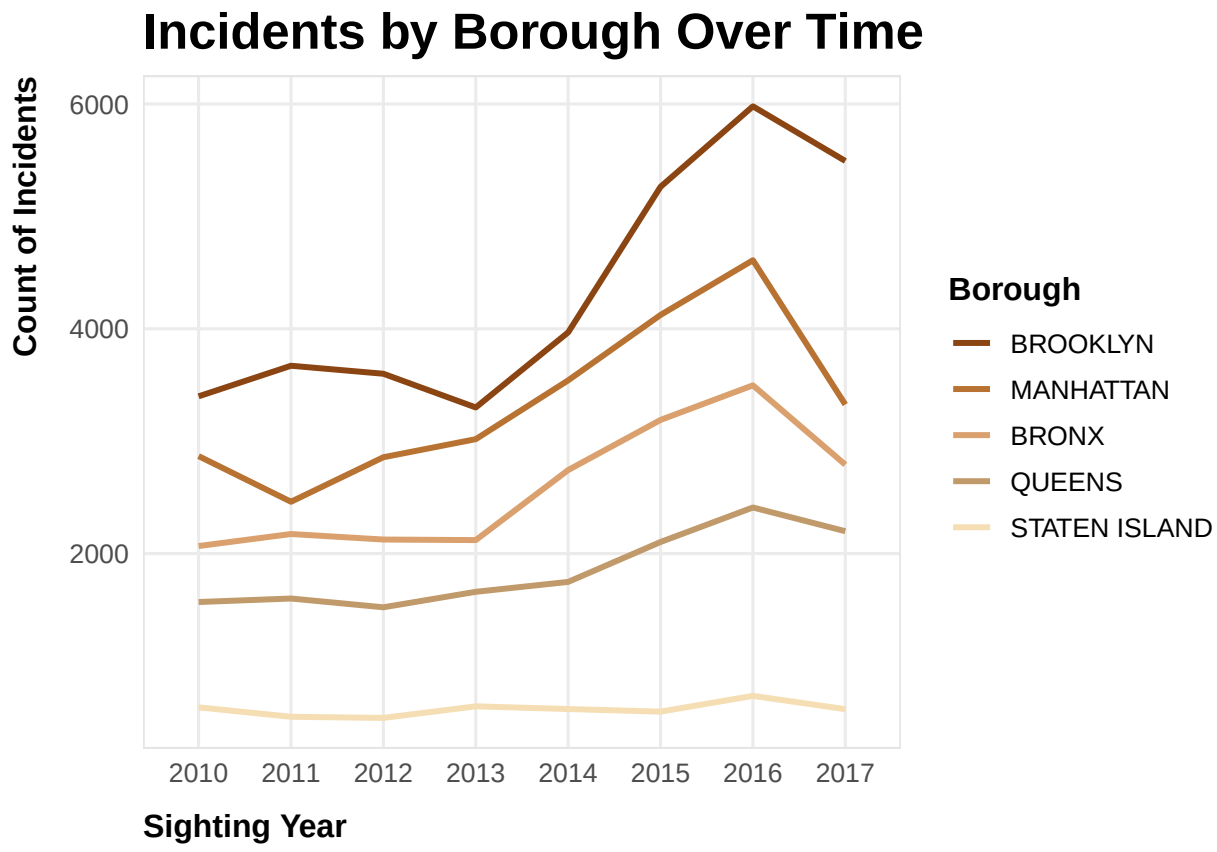
```r
# Define custom colors for each Borough
borough_colors <- c("BROOKLYN" = "#8B4513",
                    "MANHATTAN" = "#B87333",
                    "BRONX" = "#DAA06D",
                    "QUEENS" = "#C19A6B",
                    "STATEN ISLAND" = "#F5DEB3")

# Filter out 'Unspecified' borough
filtered_data <- rat_data_subset %>%
  filter(Borough != "Unspecified")

# Reorder the levels of Borough in the dataframe
filtered_data$Borough <- factor(filtered_data$Borough, levels = c("BROOKLYN", "MANHATTAN", "BRONX", "QU

# Create the line graph with custom colors and reordered legend
plot_1<-ggplot(filtered_data, aes(x = sighting_year, group = Borough, color = Borough)) +
  geom_line(stat = "count", size=1.05) +
  scale_color_manual(values = borough_colors) +
  labs(title = "Incidents by Borough Over Time",
       x = "Sighting Year",
       y = "Count of Incidents") +
  theme_minimal()+
  my_pretty_theme

plot_1
```

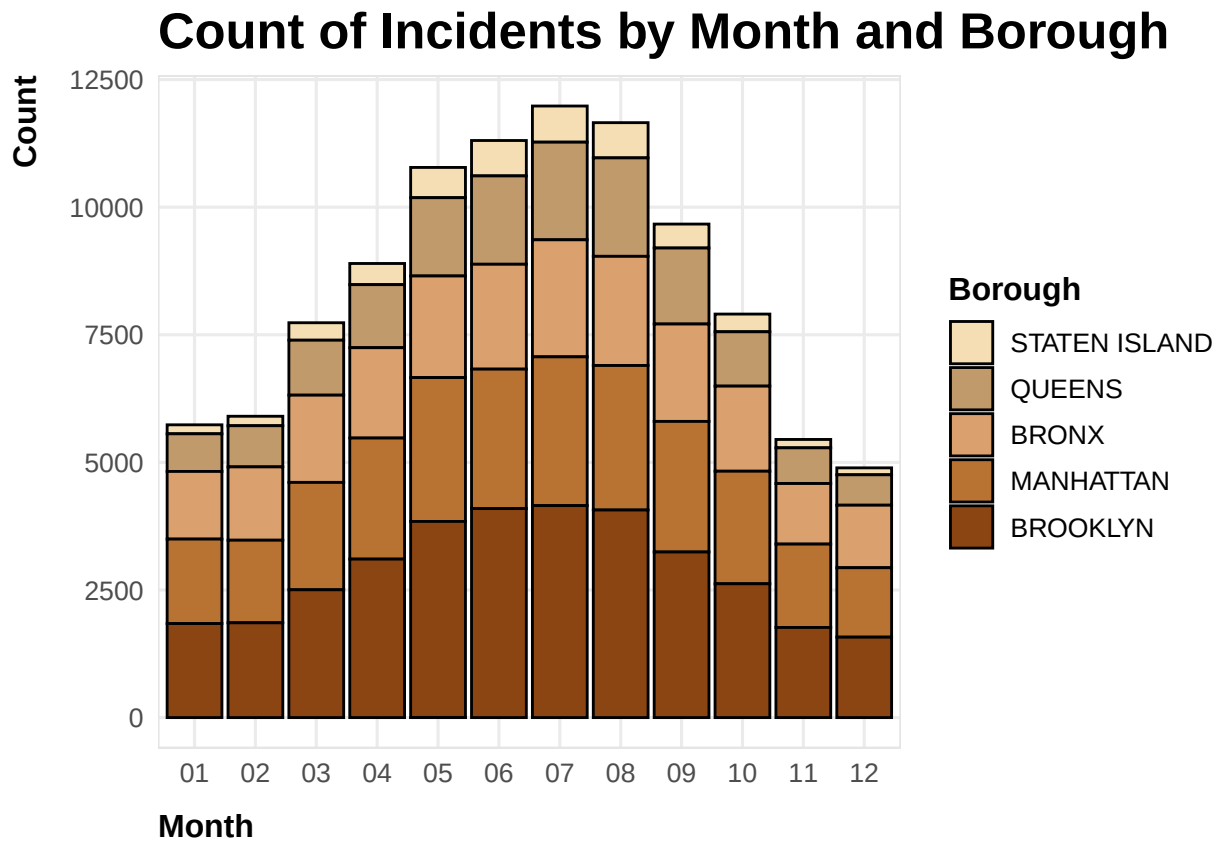


```
ggsave(plot_1, filename = "Incidents by Borough Over Time.pdf", width = 10, height = 6)
```

```
#Plotting count of incidents by Month and Borough
# Define the order of Boroughs for stacking
borough_order <- c("STATEN ISLAND", "QUEENS", "BRONX", "MANHATTAN", "BROOKLYN")

# Plot count of incidents in months for all years with stacked columns in specified order and legend ti
plot_2<-ggplot(rat_data_subset, aes(x = sighting_month, fill = factor(Borough, levels = borough_order))
  geom_bar(color = "black") +
  scale_fill_manual(values = borough_colors, name = "Borough") +
  labs(title = "Count of Incidents by Month and Borough",
       x = "Month",
       y = "Count") +
  theme_minimal()+
  my_pretty_theme

plot_2
```

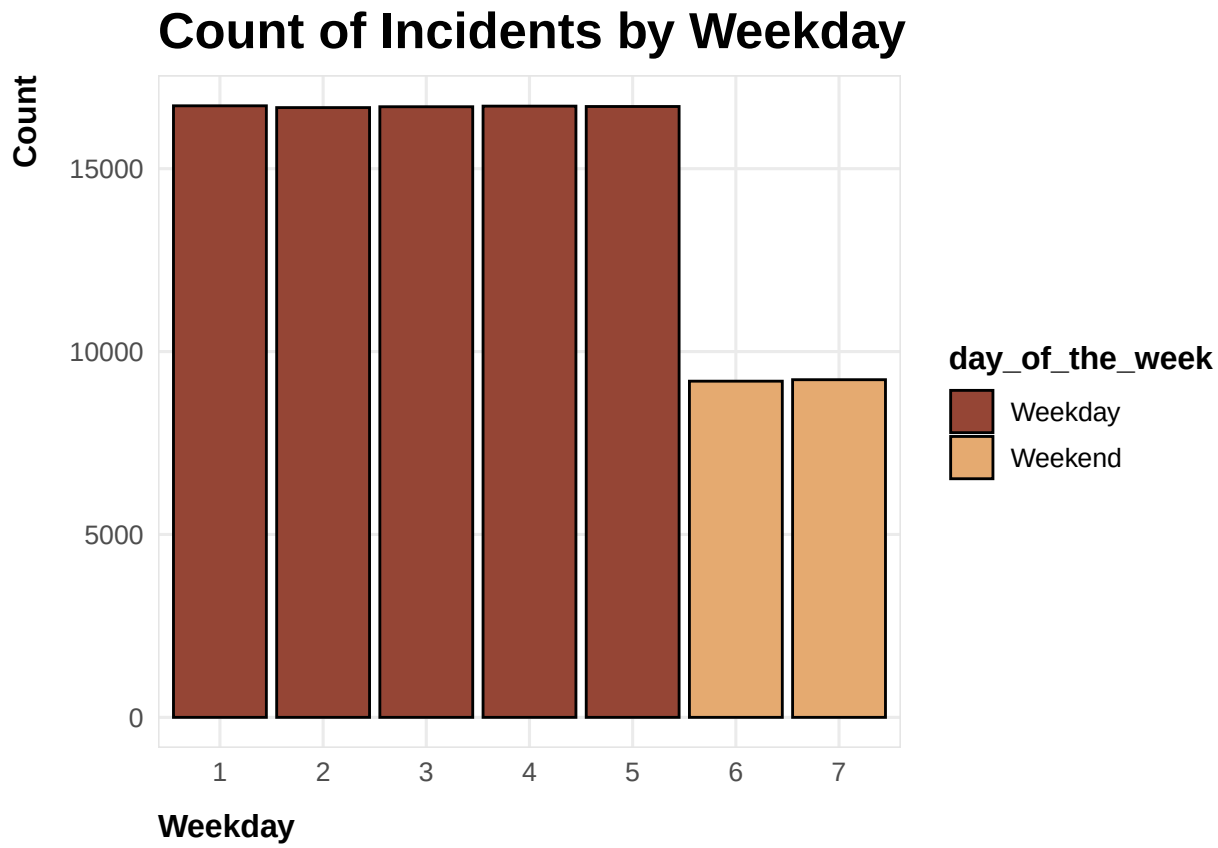



```
ggsave(plot_2, filename = "Count of Incidents by Month and Borough.pdf", width = 10, height = 6)
```

```
# Plot count of incidents based on sighting_weekday
plot_3<-ggplot(rat_data_subset, aes(x = factor(sighting_weekday), fill = ifelse(sighting_weekday %in% c
  geom_bar(color = "black") +
  scale_fill_manual(values = c("#954535","#E5AA70"), labels = c("Weekday", "Weekend")) +
```

```
  labs(title = "Count of Incidents by Weekday",
       x = "Weekday",
       y = "Count",
       fill = "day_of_the_week") +  # Set legend title
  theme_minimal()+
  my_pretty_theme


plot_3
```




```
ggsave(plot_3, filename = "Count of Incidents by Weekday.pdf", width = 10, height = 6)
```

References:

1. Walsh MG. Rat sightings in New York City are associated with neighborhood sociodemographics, housing characteristics, and proximity to open public space. PeerJ. 2014 Aug 26;2:e533. doi: 10.7717/peerj.533. PMID: 25237595; PMCID: PMC4157232. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4157232/
2. Wikipedia contributors. (2024, February 25). New York City Subway stations. Wikipedia. https://en.wikipedia.org/wiki/New_York_City_Subway_stations
3. RentHop. (2018, June 12). Rats continue to thrive in New York City, especially during summer. RentHop. https://www.renthop.com/research/rats-continue-to-thrive-in-new-york-city-especially-during-summer/

4. New York Post. (2024, January 5). NYC tourists seeking 'authentic' experience take 'rat tours'. New York Post. https://nypost.com/2023/08/26/nyc-tourists-seeking-authentic-experience-take-rat-tours/