# A1 Assignment

SHIH HSUAN, YAN

2024-02-23

```
# setup
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
```

```
# import library
library(tidyverse)
library(lubridate)
library(sf)
library(tigris)
library(gganimate)
options(tigris_use_cache = TRUE)
```

```
# read csv
df <- read_csv("data/Rat_Sightings.csv", na = c("", "NA", "N/A"), show_col_types = FALSE)
```

**Remove all the null's and single unique value's column**

```
# drop null columns
df <- df %>%
  select(where(~ !all(is.na(.))))
```

```
# drop columns with single unique value
df <- df %>%
  select(where(~ n_distinct(.) > 1))
```

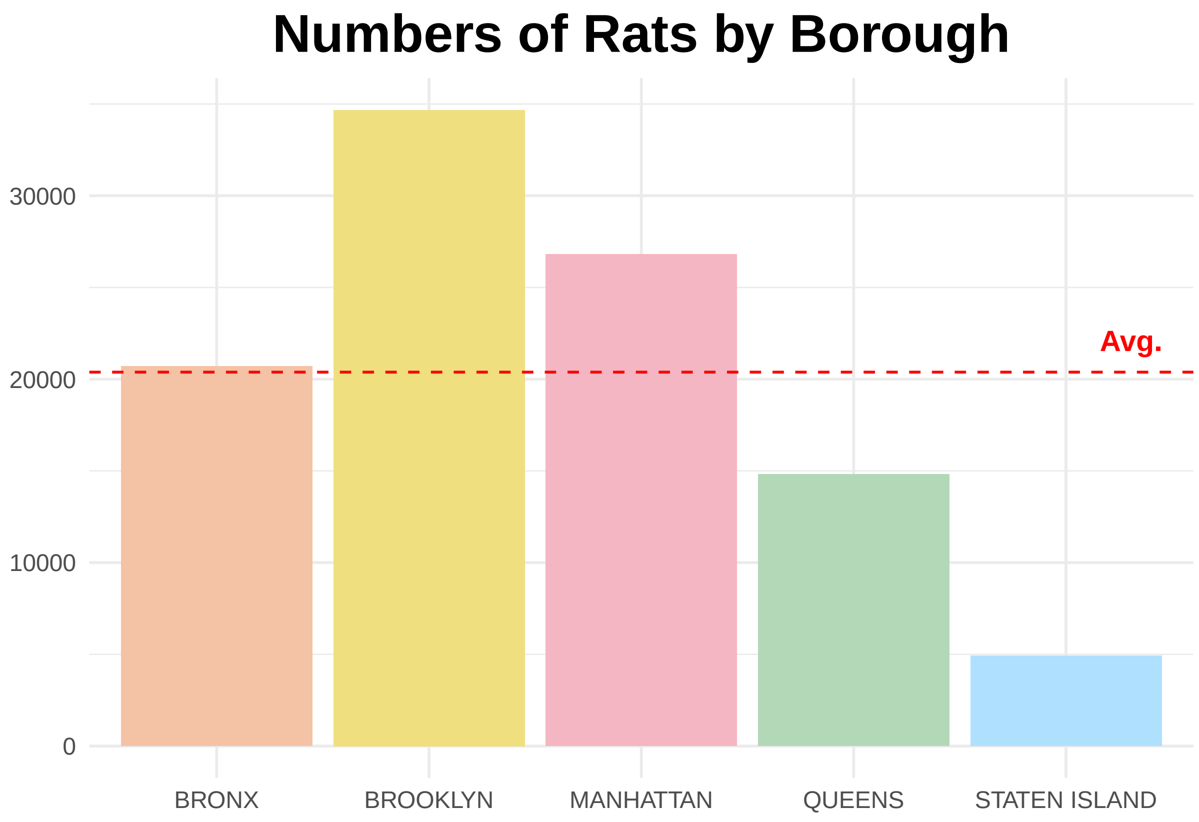
**Mutate all the date columns to the format that I want to work with**

```
# reformat all the datetime columns
df <- df %>%
  mutate (across(where(is.character),
                  ~ if (any(!is.na(.x) &
                        grepl("^\\d{2}/\\d{2}/\\d{4} \\d{2}:\\d{2}:\\d{2} [AP]M$", .x)))
                    mdy_hms(.x) else .x))
```

**Group the data by borough to calculate the total count of each borough**

```
# rate count by borough
borough_counts <- df %>%
  filter(Borough != "Unspecified") %>%
  group_by(Borough) %>%
  summarise(Count = n()) %>%
  arrange(Count)
```

```
# rat count by borough
plot1<-ggplot(borough_counts, aes(x = Borough, y = Count, fill = Borough)) +
geom_bar(stat = "identity") +
labs(title = "Numbers of Rats by Borough", x = "", y = " ") +
theme_minimal() +
theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 20),
  legend.position = "none") +
scale_fill_manual(values = c("#f4c2a5","#efdf7f","#f4b6c2","#b2d8b8","#afe0ff"))+
geom_hline(yintercept = mean(borough_counts$Count), linetype = "dashed", color = "red")+
annotate("text", x = Inf, y = mean(borough_counts$Count),
         label = "Avg.", vjust = -1, hjust = 1.5, color = "red", fontface = "bold")
plot1
```



This graph shows that, over the years, Brooklyn has the highest number of rats, followed by Manhattan as the second highest. These two boroughs not only have a large rat population but also have numbers much higher than the average in New York City. Since Brooklyn has the highest population of rats, I am going to dive into Brooklyn and analyze where the rats gather.

I used different colors to distinguish the different boroughs and added an average line to the chart, making it very easy for the audience to understand that the rat populations of these two boroughs are higher than the average at first glance. I removed the x-axis and y-axis labels since the graph can be easily comprehended and the labels inferred by looking at the title.

```r
# save graph
ggsave("/Users/jamie/Desktop/Hult/R/A1/output/Numbers_of_Rats_by_Borough.png",
       plot1, width = 8, height = 6, units = "in", dpi = 300,bg = "white")
ggsave("/Users/jamie/Desktop/Hult/R/A1/output/Numbers_of_Rats_by_Borough.pdf",
       plot1, width = 8, height = 6, units = "in", dpi = 300)
```

## Create a new column to save year from `Created Date` column

```r
# create `Year` column
df$Year <- year(df$`Created Date`)
```

## Download NYC map from library `sf`

```r
# download nyc map from library 'sf'
ny_counties <- counties(state = "NY", class = "sf")
nyc_counties <- ny_counties %>%
  filter(NAME %in% c("Bronx", "Kings", "New York", "Queens", "Richmond"))
```

## Create a dataset that focus on Brooklyn

```r
# remove data that are missing longitude and latitude
df_to_map <- df[!is.na(df$Longitude) & !is.na(df$Latitude), ]
```

```r
# create data that focus on brooklyn
brooklyn <- ny_counties %>%
  filter(NAME %in% "Kings")

brooklyn_map<-df_to_map%>%
  filter(Borough == "BROOKLYN")
```

```r
# create animate to show how rats located at over time in brooklyn
animated_brooklyn_map <- ggplot(data = brooklyn) +
  geom_sf(fill = "#efdf7f", color = "black") +
  geom_point(data = brooklyn_map, aes(x = Longitude, y = Latitude),
             color = "#c01c07", size = 2, alpha = 0.7) +
  ggtitle("BROOKLYN MAP WITH RAT") +
  labs(subtitle = 'Year: {as.integer(current_frame)}') +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5,vjust = -5, face = "bold",size = 20),
        plot.subtitle = element_text(hjust = 0.1,vjust = -15,size = 15, face = "bold")) +
  transition_manual(Year) +
  ease_aes('linear')
```

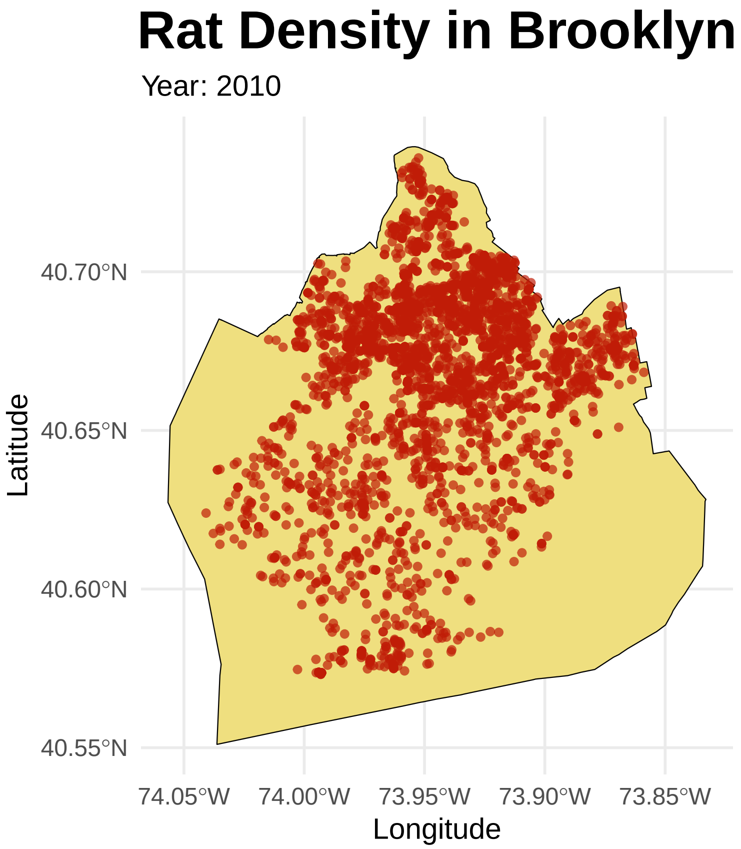
```
# show the animate
animated_brooklyn_map <- animate(animated_brooklyn_map, nframes = 20,
                                 end_pause = 3, width = 500, height = 600)
```

```
# save animation
anim_save("/Users/jamie/Desktop/Hult/R/A1/output/brooklyn_rat.gif", animated_brooklyn_map)
```

```
# create first frame png of animation
first_frame_data <- brooklyn_map %>%
  filter(Year == 2010)

# Create the plot for the first frame
first_frame_plot <- ggplot(data = brooklyn) +
  geom_sf(fill = "#efdf7f", color = "black") +
  geom_point(data = first_frame_data, aes(x = Longitude, y = Latitude),
             color = "#c01c07", size = 1, alpha = 0.7) +
  ggtitle("Rat Density in Brooklyn") +
  labs(subtitle = 'Year: 2010') +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 20))

first_frame_plot
```



This graph and the attached GIF show that rats are likely to gather in the northern part of Brooklyn over the years, and this pattern has not changed. According to the article referenced, there is a higher density of people in the northern part of Brooklyn, which is closer to Manhattan. This may indicate that areas with more people tend to have more rats. Building on this observation, my analysis will further explore the rat population in Brooklyn by distinguishing between residential and non-residential areas. This distinction is crucial as it may shed light on whether rats are more attracted to areas based on human habitation patterns or other factors associated with urban development.

In adherence to the CRAP design principles, this map utilizes a consistent background color for visual continuity and red dots for high contrast, enhancing clarity and focus on key data points. Following Alberto Cairo's and Kieran Healy's guidelines, the map's simplicity aids in directly comparing rat density, offering insightful and aesthetically pleasing visuals without extraneous elements. The design effectively highlights rat concentration trends in Northern Brooklyn, encouraging further exploration of the underlying causes.

```r
# save graph
ggsave("/Users/jamie/Desktop/Hult/R/A1/output/brooklyn_gif_first_frame.png",
       first_frame_plot, width = 500, height = 600, units = "px", dpi = 100,bg = "white")
ggsave("/Users/jamie/Desktop/Hult/R/A1/output/brooklyn_gif_first_frame.pdf",
       first_frame_plot, width = 500, height = 600, units = "px", dpi = 100)
```

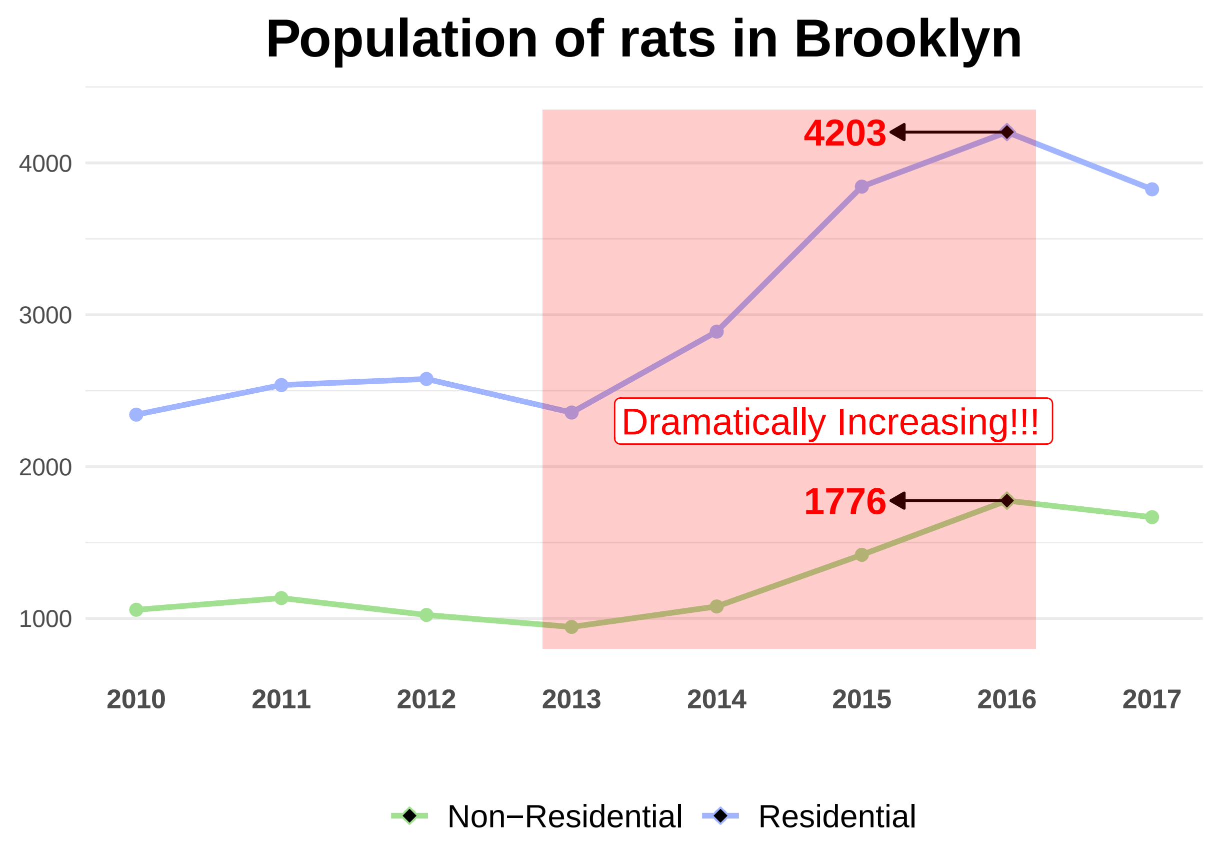
## Create a new dataframe focusing on Brooklyn and aggregate the data by residential and non-residential

```r
# aggregate whether location type is residential or not
aggr_loc <- df %>%
  group_by(Year, `Location Type`) %>%
  filter(Borough=="BROOKLYN")%>%
  summarise(Count = n(), .groups = 'drop') %>%
  mutate(AggregatedType = case_when(
    grepl("Family|SRO|Vacant Building", `Location Type`) ~ "Residential",
    TRUE ~ "Non-Residential"
  )) %>%
  group_by(Year, AggregatedType) %>%
  summarise(TotalCount = sum(Count, na.rm = TRUE), .groups = "drop")
```

```r
# plot numbers of rat in brooklyn over years
geoms <- list(
  geom_line(size = 1),
  geom_point(size = 1.8),
  geom_point(data = aggr_loc %>%
               group_by(AggregatedType) %>%
               mutate(MaxCount = max(TotalCount)) %>%
               filter(TotalCount == MaxCount),
             aes(x = Year, y = TotalCount, color = AggregatedType)
             , size = 2, shape = 23, fill = "black"),
  geom_text(data = aggr_loc %>%
              group_by(AggregatedType) %>%
              mutate(MaxCount = max(TotalCount)) %>%
              filter(TotalCount == MaxCount) %>%
              ungroup() %>%
              mutate(YearLabel = Year - 1.4),
            aes(label = TotalCount, x = YearLabel, y = TotalCount),
            color = "red", hjust = 0, fontface = "bold", size = 5),
  geom_segment(data = aggr_loc %>%
                 group_by(AggregatedType) %>%
                 mutate(MaxCount = max(TotalCount)) %>%
                 filter(TotalCount == MaxCount) %>%
                 ungroup() %>%
                 mutate(YearLabel = Year - 0.8),
               aes(x = Year, y = TotalCount, xend = YearLabel, yend = TotalCount),
               arrow = arrow(length = unit(0.08, "inches"),
                             type = "closed"), color = "black")
)

plot2 <-ggplot(aggr_loc, aes(x = Year, y = TotalCount,
                     group = AggregatedType, color = AggregatedType)) +
  geoms +
  theme_minimal() +
  scale_x_continuous(breaks = aggr_loc$Year) +
  scale_color_manual(values = c("#a1df91", "#a1b5ff")) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 20),
        axis.text.x = element_text(face = "bold", size = 10),
        legend.position = "bottom",
        legend.text = element_text(size = 12),
        legend.title = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())+
  labs(title = "Population of rats in Brooklyn", x = "", y = " ")+
  annotate("rect", xmin = 2012.8, xmax = 2016.2,
           ymin = 800, ymax = 4350, alpha = 0.2, fill = "red")+
  annotate(geom = "label", x = 2013.9, y = 2300, label = "Dramatically Increasing!!!",
           hjust = 0.2 ,vjust = 0.5, lineheight = 0.5,  size = 5, color = "red")

plot2
```



This chart suggests that the population of rats may be correlated with human gatherings. Starting from 2013, the rat population has increased dramatically, and both residential and non-residential places recorded their highest rat populations in 2016. However, the rat population in residential areas in 2016 is almost 2.5 times higher than that in non-residential areas.

This graph, while akin to one of the original graphs, adopts a distinct presentation method. The original graph utilized a bar chart to compare rat populations by various location types over the years. However, I found that approach ineffective due to the excessive number of location types. Consequently, I opted to consolidate the location data into residential and non-residential categories. This method clearly shows that sightings of rats in residential areas are significantly higher than in non-residential areas. I have used an annotated rectangle to highlight the trend of increasing rat populations and annotated arrows to indicate the peak values in both lines.

In crafting this visualization, I adhered to the CRAP principles by employing alignment and proximity to organize the data, and contrast to differentiate between residential and non-residential areas. Kieran Healy's emphasis on clarity is reflected in the decision to simplify location types, and Alberto Cairo's quality of functionality is honored by using annotations that guide the viewer towards key findings. The graph achieves insightfulness by clearly demonstrating the disparity between rat sightings in different types of locations, and the annotations serve to further enlighten the audience about specific trends and values.

```
# save graph
ggsave("/Users/jamie/Desktop/Hult/R/A1/output/Rat_Count.png",
       plot2, width = 8, height = 6, units = "in", dpi = 300,bg = "white")
ggsave("/Users/jamie/Desktop/Hult/R/A1/output/Rat_Count.pdf",
       plot2, width = 8, height = 6, units = "in", dpi = 300)
```

## References

- Matt Coneybeare. “This Density Map Shows How We Crowd 8.5 Million People in New York City.” Viewing NYC, April 29, 2017. URL