Car Crashes in New York City

Sawyer Cremer, Sophie Chikhladze, Adarsh Gadepalli, Rahul Prakash

Table of contents

Introduction

1.3 million people die each year as a result of road traffic crashes, and between 20 and 50 million people every year suffer non-fatal injuries resulting from car accidents. Other than the millions of lives lost, the U.S. Department of Transportation's most recent estimate of the annual economic cost of crashes is $340 billion. This is an issue affecting virtually every country in the world, including the United States. In New York City, New York, the New York Police Department is required to be fill out a report for every collision where someone is injured or killed, or where there is at least $1000 worth of damage.

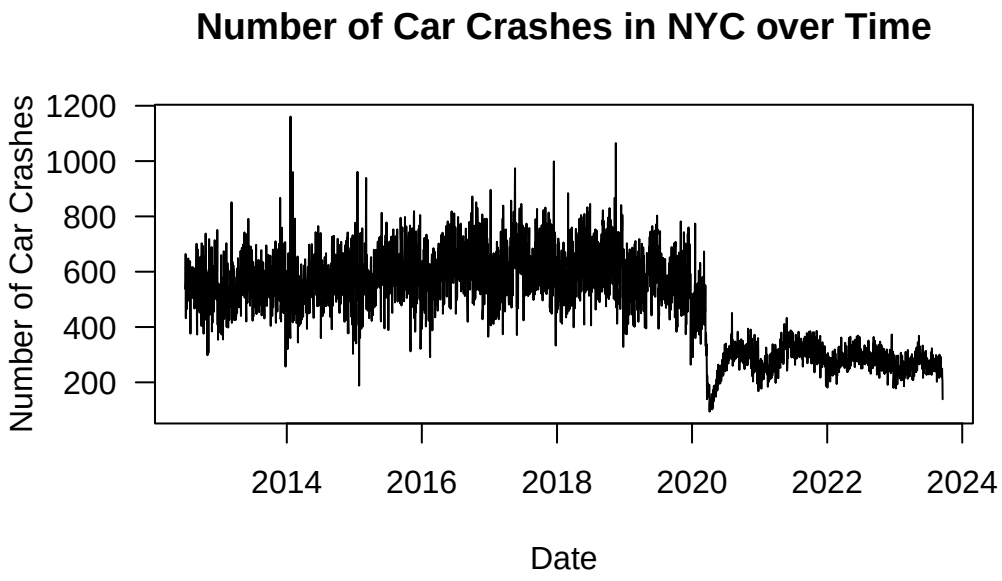
Primary Dataset

The Motor Vehicle Collisions dataset contains information about car crashes in NYC from the years 2016-2023. It contains information from all police reported motor vehicle collisions in NYC. We accessed the dataset from data.gov, provided by the City of New York. Each entry represents a car crash that occured in New York City. The dataset contains 28 columns that describe a variety of data, including time of accident, accident severity(deaths/injuries), pinpointed locations (neighborhoods/streets), causes for the accident, and vehicle type. Using this data, we hope to analyze and identify chronological and geographical patterns in New York City to pinpoint high-risk environments and circumstances. We are assuming that the data is continuous in terms of time, and that the specific timestamps and event details are accurate.

Plots

Number of Car Crashes over time

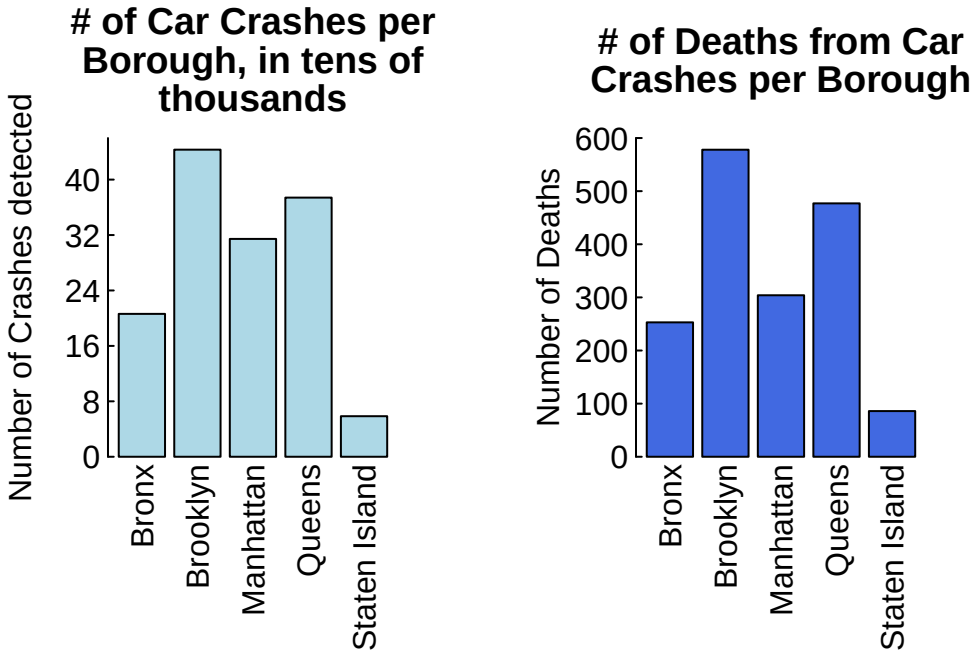
First, we decided to plot a time series of the number of car crashes each day starting from 2012 up until 2023. Each point represents the total number of crashes that occurred on each day, plotted chronologically.



Looking at the number of crashes over time, from 2012 to 2023, we can see that the number of crashes remained fairly consistent from 2012 to the beginning of 2020, with certain exceptions like big spikes or drops. Until 2020, there is no year with an exceptionally low or high number of crashes. However, we see a clear change in the number of crashes in the beginning of 2020, which can be interpreted as the effects of COVID-19 as that is roughly the date it started. Less people had to go to work, school became virtual, and therefore less people were driving, leading to a lowered number of car crashes per day. We could also further study the exact dates of this drop versus when the lockdown started.

Number of Car Crashes and Deaths per Borough Bar Plot

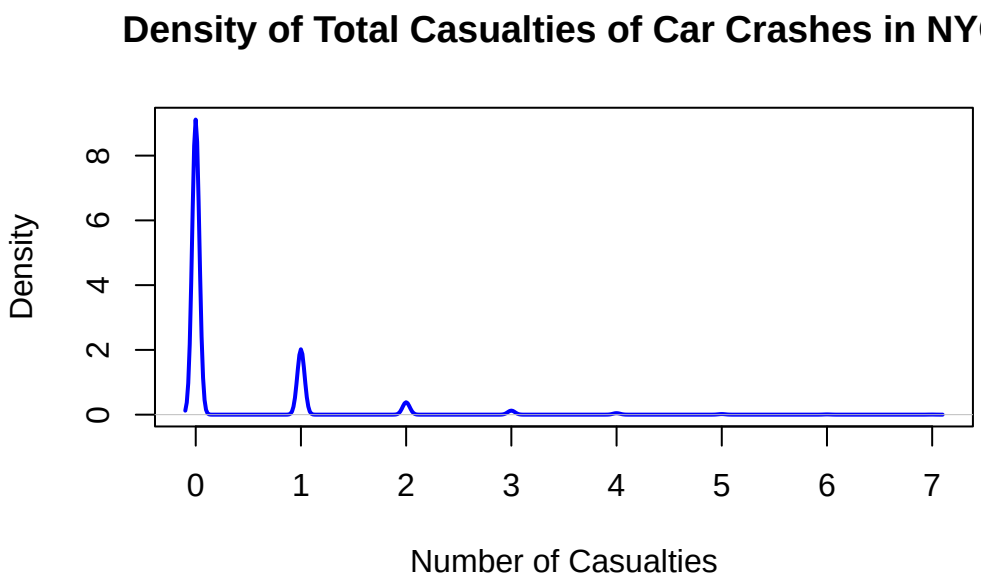
For this plot, we grouped the data by 5 boroughs of New York City: Bronx, Brooklyn, Manhattan, Queens, and Staten Island. First, we calculated the total number of car crashes in each borough, and plotted it in tens of thousands since there was a large amount. Then, we wanted to see if the number of casualties in each borough was proportional to the number of car crashes. Therefore, we calculated the total number of deaths from car accidents per borough and plotted that beside the number of car crashes.



From the left plot, we can see that Brooklyn has the highest amount of crashes, followed by Queens and Manhattan. Staten Island has the lowest number of crashes recorded, by a significant amount. Despite the fact that Staten Island has the highest number of cars per household, it is sensible that the rest of the boroughs have more car accidents, since there is more general traffic there. Furthermore, if we compare the number of crashes per borough to the number of casualties per borough, We can see that the proportions of deaths are similar to the number of accidents, except for Manhattan, which is slightly lower.

The Density of Total Number of Casualties

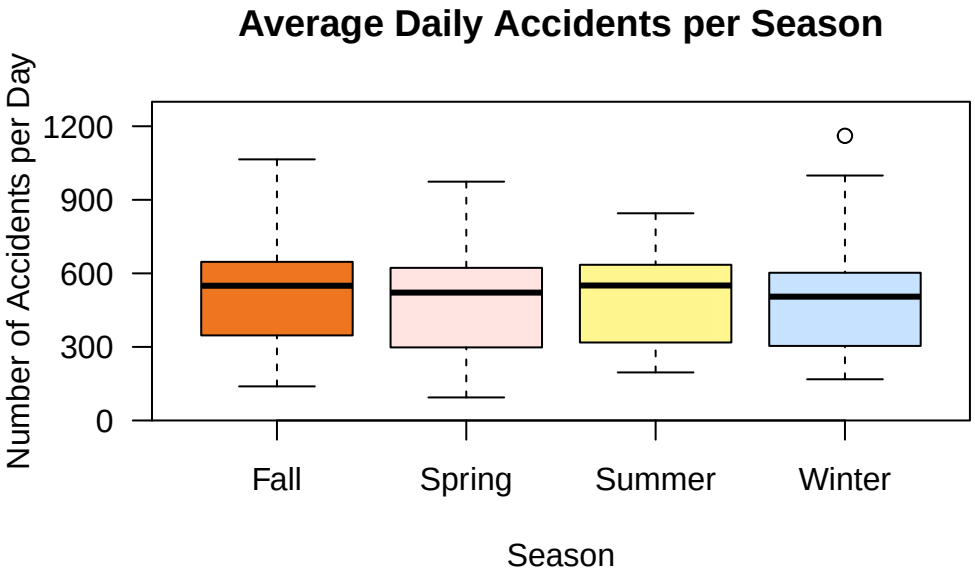
This is a density plot visualizing total casualties for all accidents in our dataset, where a casualty is either a person injured or a person killed. The plot visualizes the probability density function (pdf) of total casualties of car crashes in NYC. It shows the distribution of car crash casualties in NYC and how likely each amount is.



This plot suggests that based on our data, car crashes in NYC are most likely to result in no casualties, evidenced by the density for 0 casualties being by far the highest amount. As the number of casualties increases, a large drop is seen. However, there is likelihood for a car accident to result in 1 or 2 casualties, but past that the likelihood drops to virtually 0. This shows that more severe crashes are much less likely in NYC.

Average Accidents per Day per Season Boxplot

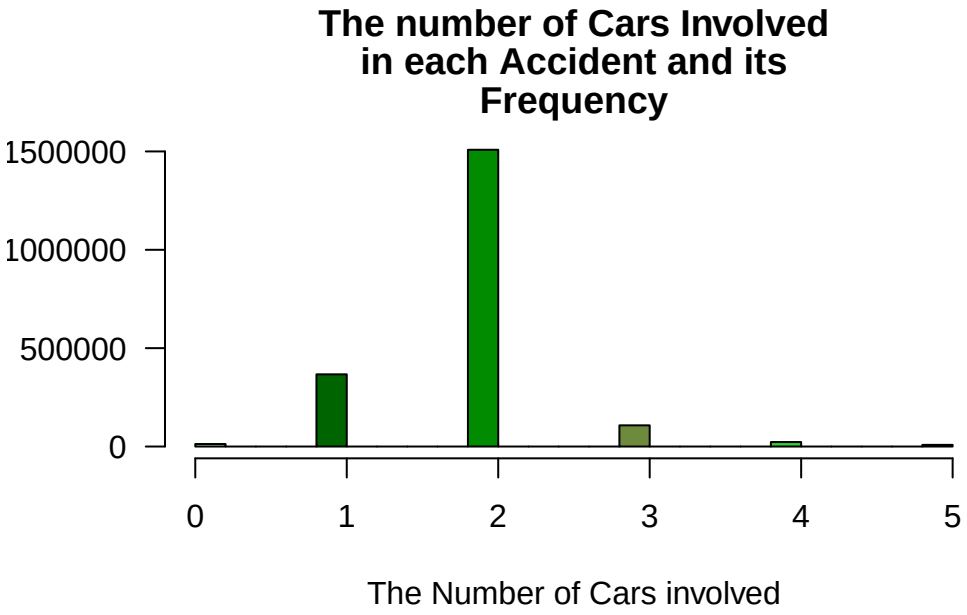
Next, we asked ourselves - is an accident more or less likely to occur during a specific season? For this, we turned each date into a season and calculated the number of car crashes per day for each season. The, we plotted 4 boxplots to represent each season portraying the number of car crashes per day.



As we can see, there is a relatively similar amount of car crashes across all the seasons, suggesting that the season does not significantly affect the likelihood of crashes occurring. One interesting point is that there is a much larger variance and spread of car crashes per day in the fall while there is a significantly smaller variance of car crashes per day in the summer. There is also an outlier in the Winter, where over 1200 crashes happened on one day. However, this is just an outlier since the median number of crashes in Winter is slightly lower than the median number of crashes in the Summer and Fall.

Number of Cars Involved in a Car Crash Histogram

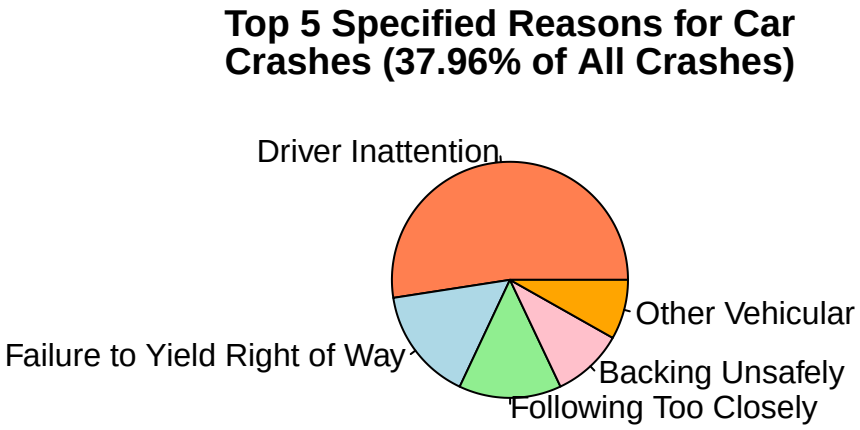
We decided to find out how many cars were involved in each accident. While there wasn't a column for the number of cars involved, there were 5 columns for vehicle type, so we counted each nonempty entry as a car involved.



As the histogram shows, the highest frequency of cars involved in a given car crash peak at 2 cars, with lower frequencies otherwise. The number of crashes at 3+ cars are relatively low, implying the rarity of largescale accidents and the commonality of lower scale, regular accidents involving 2 cars. There is a little under 500 000 accidents involving just one car, which can mean the driver was inattentive and crashed into something or crashed into a pedestrian/cyclist.

5 Most Common Reasons for Car Crashes

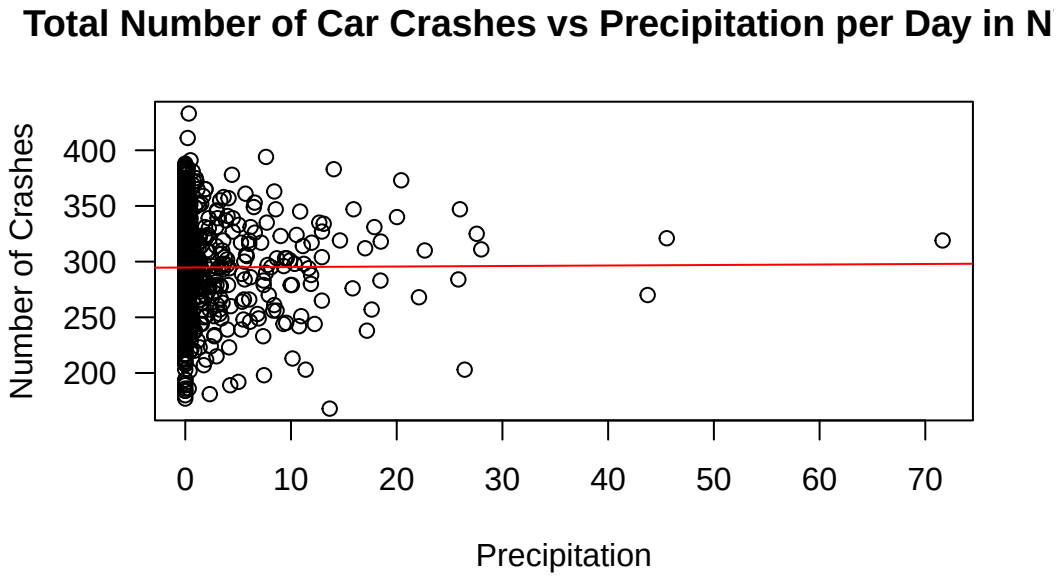
We decided to find the 5 most common specified reasons for car crashes in New York. While a majority of the causes was unspecified or other, we decided not to include those in our chart.



This plot displays the distribution of the most common reasons for car crashes in New York. As we can see, a high portion of crashes are due to driver inattention (possibly due to using their phone while driving, or generally not paying attention), while the other reasons are similarly represented at a much smaller amount. These reasons include Failure to Yield Right of Way, Following Too Closely, and Backing Unsafely.

Regression

```
Call:
lm(formula = Count ~ precip, data = merged_data3)

Residuals:
     Min       1Q   Median       3Q      Max
-127.314  -28.693    0.307   28.307  138.292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 294.69303    1.53948 191.424   <2e-16 ***
precip        0.04545    0.29216   0.156    0.876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.4 on 850 degrees of freedom
Multiple R-squared:  2.847e-05, Adjusted R-squared:  -0.001148
F-statistic: 0.0242 on 1 and 850 DF,  p-value: 0.8764
```



This plot shows a linear regression over a plot of car crashes vs precipitation in NYC. The summary of the regression indicates that as there is a one unit increase in precipitation, the average number of car crashes increases by 0.04545. The intercept shows that the estimated

average number of car crashes on a day with no precipitation in NYC is 294.69. The precipitation coefficient is not statistically significant because of its high p-value of 0.876, and the t-value of 0.156 further reinforces the lack of statistical significance in the coefficient.

ggplot2 Plots

Number of Car Accidents Over/Under average per Day of Week

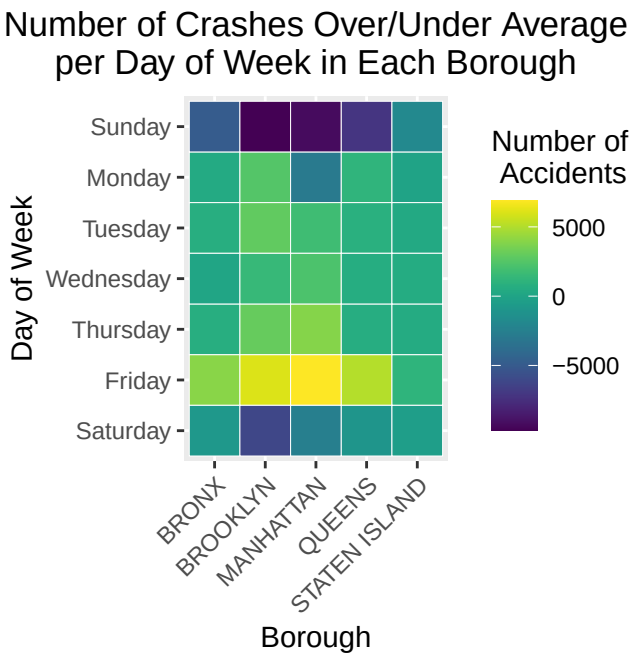
```
# Create the faceted heatmap with sorted weekdays
library(ggplot2)
library(viridis)
```

```
Loading required package: viridisLite
```

```
nyc_crashes$WEEKDAY <- weekdays(nyc_crashes$CRASH.DATE)
borough_weekday_subset <- nyc_crashes[(nyc_crashes$BOROUGH!=""),]

borough_crash_counts <- as.data.frame(table(borough_weekday_subset$BOROUGH, borough_weekda

colnames(borough_crash_counts) <- c("Borough", "Weekday", "Count")
weekday_order <- c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Satu

avg_borough <- data.frame(Borough = unique(borough_crash_counts$Borough))
avg_borough$Average <- sapply(avg_borough$Borough, function(borough){
  mean(borough_crash_counts$Count[borough_crash_counts$Borough == borough])
})

borough_crash_counts$Count_avg <- apply(borough_crash_counts, MARGIN = 1, FUN = function(r
  return(as.numeric(row[3]) - as.numeric(avg_borough$Average[avg_borough$Borough == row[1]
})

ggplot(borough_crash_counts, aes(x = Borough, y = factor(Weekday, levels = rev(weekday_ord
  geom_tile(color = "white", size = 0.1) +
  coord_equal() +
  labs(
    x = "Borough",
    y = "Day of Week",
    title = "Number of Crashes Over/Under Average\nper Day of Week in Each Borough",
    fill = "Number of\n Accidents"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +
  scale_fill_viridis() +
```

```
    theme(plot.title = element_text(hjust = 0.5))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Number of Car Crashes per Zipcode Heatmap

```
  library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v lubridate 1.9.2     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

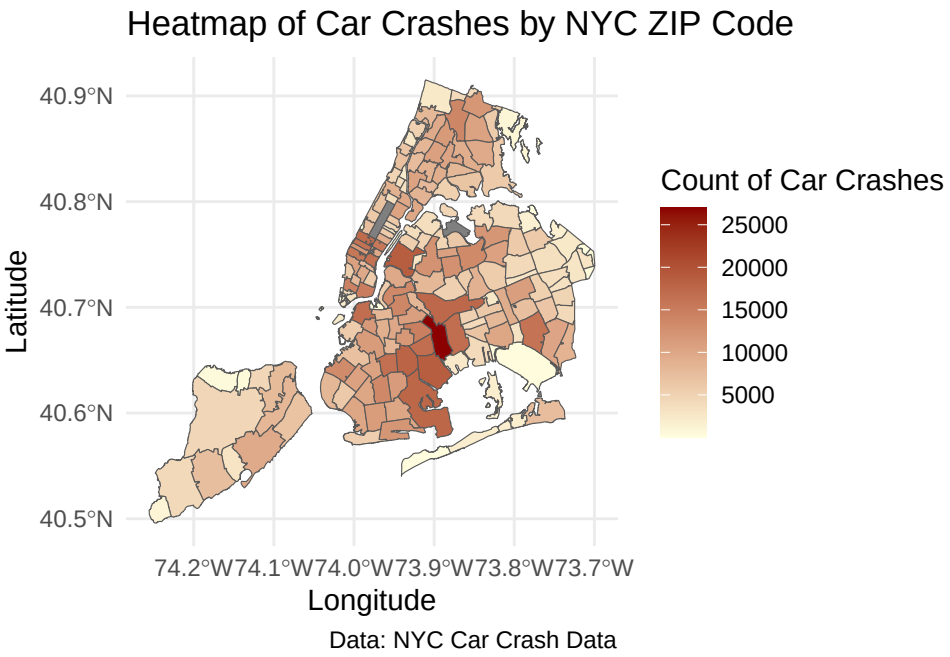
```r
library(sf)
```

```
Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE
```

```r
nyc_crashes$ZIP.CODE <- as.character(nyc_crashes$ZIP.CODE)
nyc_crashes$ZIP.CODE <- sub("\\.0$", "", nyc_crashes$ZIP.CODE)
zip_counts <- nyc_crashes %>%
group_by(ZIP.CODE) %>%
summarize(count = n())
nyc_zip_boundaries <- st_read("zipcodes.geojson")
```

```
Reading layer `ZIP_CODE_040114' from data source
  `/Users/sophiochikhladze/Desktop/Stat_405/Stat405_Data/zipcodes.geojson'
  using driver `GeoJSON'
Simple feature collection with 263 features and 12 fields
Geometry type: POLYGON
Dimension:     XY
Bounding box:  xmin: -74.25576 ymin: 40.49584 xmax: -73.6996 ymax: 40.91517
Geodetic CRS:  WGS 84
```

```r
merged_zip_data <- left_join(nyc_zip_boundaries, zip_counts, by = c("ZIPCODE" = "ZIP.CODE"
ggplot(merged_zip_data) +
  geom_sf(aes(fill = count)) +
  scale_fill_gradient(low = "lightyellow", high = "darkred",
                      name = "Count of Car Crashes") +
  ggtitle("Heatmap of Car Crashes by NYC ZIP Code") +
  labs(x = "Longitude", y = "Latitude",
       caption = "Data: NYC Car Crash Data") +
  theme_minimal()
```



Trash code:

Resources

- Primary Dataset - Motor Vehicle Collisions in New York City, New York
- Weather data from Visual Crossing - Weather in New York City, New York from 2020-09-01 to 2022-12-31
- Road Traffic Injuries - World Health Organization