NYPD_Data

T. Grey

2024-10-15

Introduction

In this analysis, we are exploring the dataset provided by NYC Open Data from the following source:

[NYC Open Data - Link to the dataset]

https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Setup

```
library(tidyverse)
library(lubridate)
library(janitor)
library(corrplot)

url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nyc_data <- read_csv(url)

nyc_data <- janitor::clean_names(nyc_data)

head(nyc_data)
```

```
## # A tibble: 6 x 21
##   incident_key occur_date occur_time boro      loc_of_occur_desc precinct
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # i 15 more variables: jurisdiction_code <dbl>, loc_classfctn_desc <chr>,
## #   location_desc <chr>, statistical_murder_flag <lgl>, perp_age_group <chr>,
## #   perp_sex <chr>, perp_race <chr>, vic_age_group <chr>, vic_sex <chr>,
## #   vic_race <chr>, x_coord_cd <dbl>, y_coord_cd <dbl>, latitude <dbl>,
## #   longitude <dbl>, lon_lat <chr>
```

```
summary(nyc_data)
```

```
##   incident_key        occur_date         occur_time           boro
##  Min.   :  9953245   Length:28562       Length:28562      Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms        Class :character
##  Median : 92711254   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  loc_of_occur_desc     precinct     jurisdiction_code loc_classfctn_desc
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  location_desc      statistical_murder_flag perp_age_group
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    perp_sex           perp_race         vic_age_group        vic_sex
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    vic_race            x_coord_cd        y_coord_cd        latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    longitude         lon_lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

```
str(nyc_data)
```

```
## spc_tbl_ [28,562 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ incident_key           : num [1:28562] 2.45e+08 2.48e+08 8.50e+07 2.03e+08 2.71e+07 ...
##  $ occur_date             : chr [1:28562] "05/05/2022" "07/04/2022" "05/27/2012" "09/24/2019" ...
```

```
##  $ occur_time             : 'hms' num [1:28562] 00:10:00 22:20:00 19:35:00 21:00:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ boro                   : chr [1:28562] "MANHATTAN" "BRONX" "QUEENS" "BRONX" ...
##  $ loc_of_occur_desc      : chr [1:28562] "INSIDE" "OUTSIDE" NA NA ...
##  $ precinct               : num [1:28562] 14 48 103 42 83 23 113 77 48 49 ...
##  $ jurisdiction_code      : num [1:28562] 0 0 0 0 0 2 0 0 0 0 ...
##  $ loc_classfctn_desc     : chr [1:28562] "COMMERCIAL" "STREET" NA NA ...
##  $ location_desc          : chr [1:28562] "VIDEO STORE" "(null)" NA NA ...
##  $ statistical_murder_flag: logi [1:28562] TRUE TRUE FALSE FALSE FALSE FALSE ...
##  $ perp_age_group         : chr [1:28562] "25-44" "(null)" NA "25-44" ...
##  $ perp_sex               : chr [1:28562] "M" "(null)" NA "M" ...
##  $ perp_race              : chr [1:28562] "BLACK" "(null)" NA "UNKNOWN" ...
##  $ vic_age_group          : chr [1:28562] "25-44" "18-24" "18-24" "25-44" ...
##  $ vic_sex                : chr [1:28562] "M" "M" "M" "M" ...
##  $ vic_race               : chr [1:28562] "BLACK" "BLACK" "BLACK" "BLACK" ...
##  $ x_coord_cd             : num [1:28562] 986050 1016802 1048632 1014493 1009149 ...
##  $ y_coord_cd             : num [1:28562] 214231 250581 198262 242565 190105 ...
##  $ latitude               : num [1:28562] 40.8 40.9 40.7 40.8 40.7 ...
##  $ longitude              : num [1:28562] -74 -73.9 -73.8 -73.9 -73.9 ...
##  $ lon_lat                : chr [1:28562] "POINT (-73.9935 40.754692)" "POINT (-73.88233 40.854402)"
##  - attr(*, "spec")=
##   .. cols(
##   ..   INCIDENT_KEY = col_double(),
##   ..   OCCUR_DATE = col_character(),
##   ..   OCCUR_TIME = col_time(format = ""),
##   ..   BORO = col_character(),
##   ..   LOC_OF_OCCUR_DESC = col_character(),
##   ..   PRECINCT = col_double(),
##   ..   JURISDICTION_CODE = col_double(),
##   ..   LOC_CLASSFCTN_DESC = col_character(),
##   ..   LOCATION_DESC = col_character(),
##   ..   STATISTICAL_MURDER_FLAG = col_logical(),
##   ..   PERP_AGE_GROUP = col_character(),
##   ..   PERP_SEX = col_character(),
##   ..   PERP_RACE = col_character(),
##   ..   VIC_AGE_GROUP = col_character(),
##   ..   VIC_SEX = col_character(),
##   ..   VIC_RACE = col_character(),
##   ..   X_COORD_CD = col_double(),
##   ..   Y_COORD_CD = col_double(),
##   ..   Latitude = col_double(),
##   ..   Longitude = col_double(),
##   ..   Lon_Lat = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
colnames(nyc_data)
```

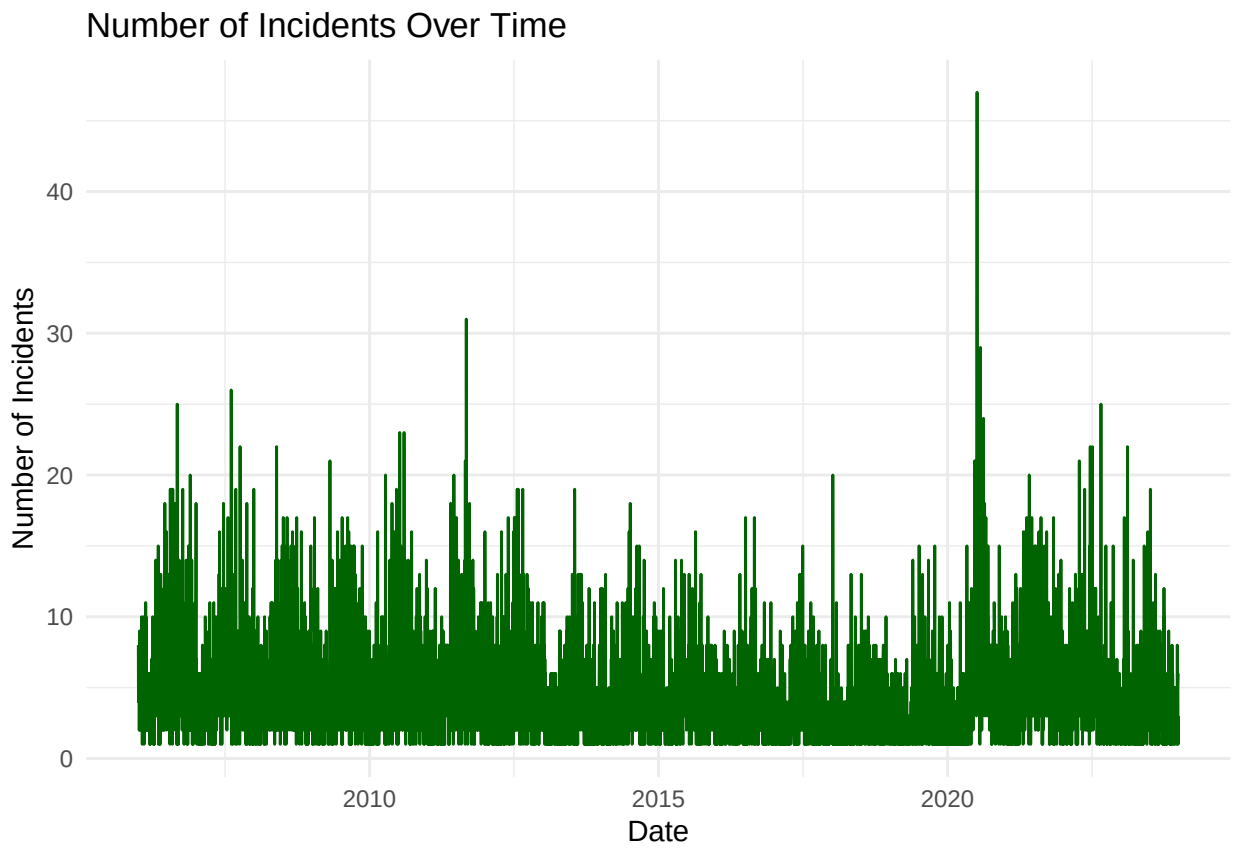
```
##  [1] "incident_key"           "occur_date"
##  [3] "occur_time"             "boro"
##  [5] "loc_of_occur_desc"      "precinct"
##  [7] "jurisdiction_code"      "loc_classfctn_desc"
##  [9] "location_desc"          "statistical_murder_flag"
## [11] "perp_age_group"         "perp_sex"
```

```
## [13] "perp_race"               "vic_age_group"
## [15] "vic_sex"                 "vic_race"
## [17] "x_coord_cd"              "y_coord_cd"
## [19] "latitude"                "longitude"
## [21] "lon_lat"
```

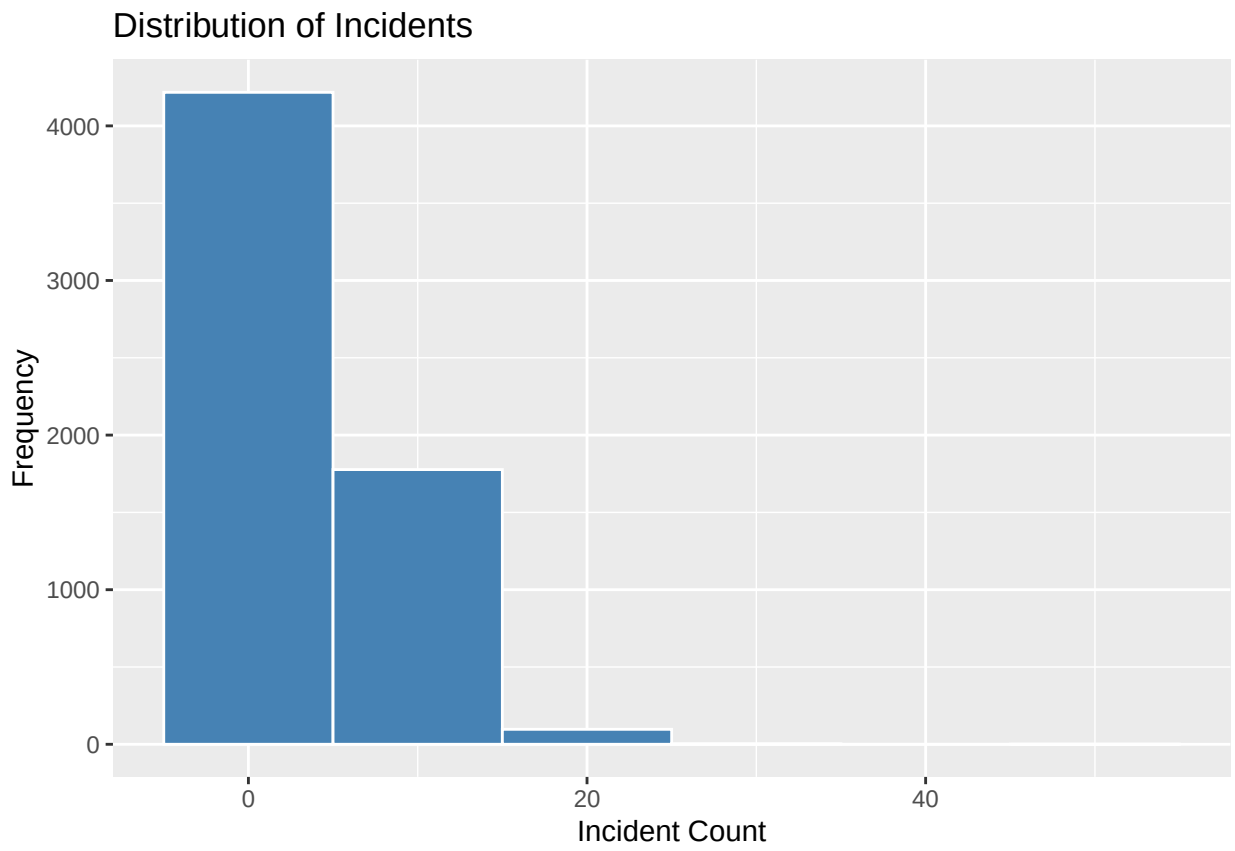
```
colSums(is.na(nyc_data))
```

```
##            incident_key               occur_date               occur_time
##                       0                        0                        0
##                   boro       loc_of_occur_desc                precinct
##                       0                    25596                        0
##       jurisdiction_code      loc_classfctn_desc           location_desc
##                       2                    25596                    14977
## statistical_murder_flag          perp_age_group                perp_sex
##                       0                     9344                     9310
##               perp_race           vic_age_group                 vic_sex
##                    9310                        0                        0
##                vic_race              x_coord_cd              y_coord_cd
##                       0                        0                        0
##                latitude               longitude                 lon_lat
##                      59                       59                       59
```

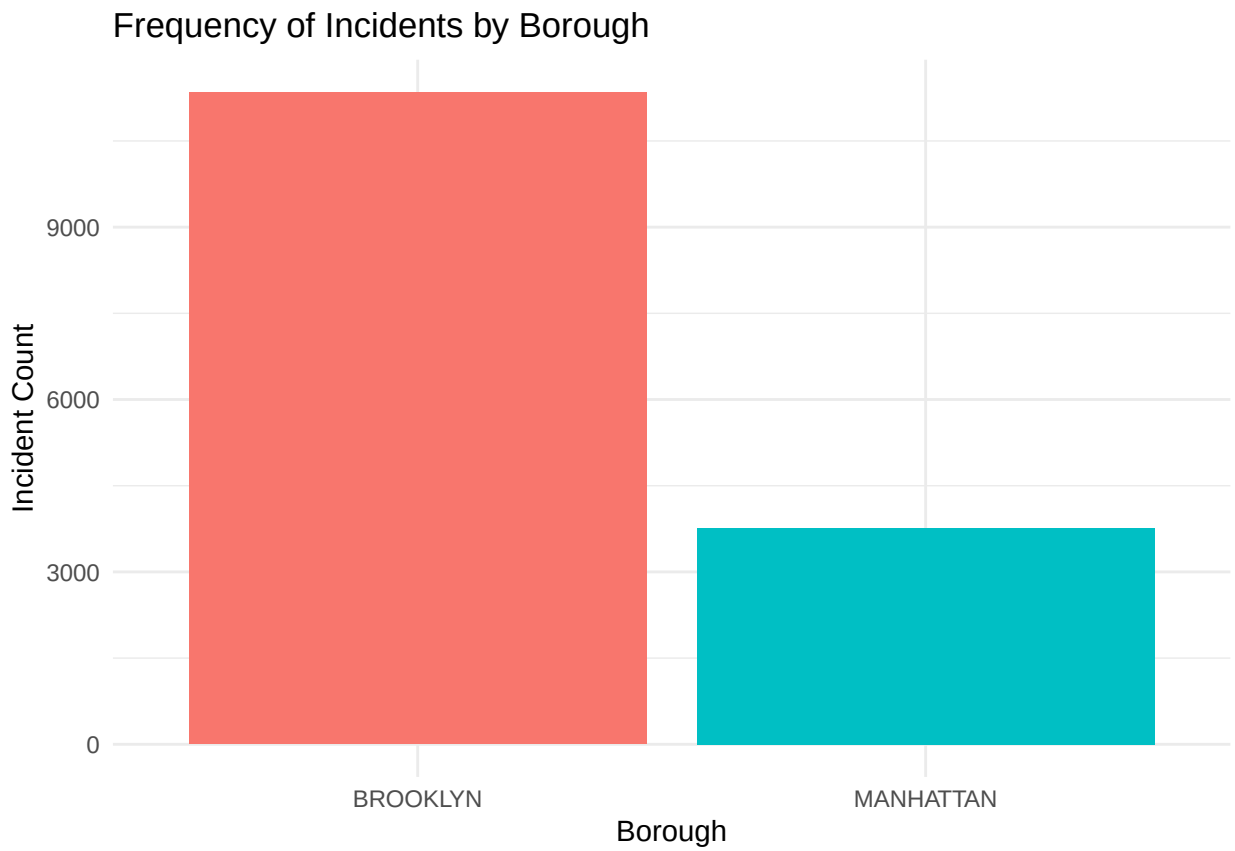
```
nyc_data <- nyc_data %>%
  mutate(occur_date = mdy(occur_date))

nyc_data <- nyc_data %>%
  mutate(
    hour_of_day = hour(occur_date),
    day_of_week = wday(occur_date, label = TRUE),
    boro_num = as.numeric(factor(boro))
  )

# Plot number of incidents over time
nyc_data_summary <- nyc_data %>%
  group_by(occur_date) %>%
  summarise(incident_count = n(), .groups = 'drop')

ggplot(nyc_data_summary, aes(x = occur_date, y = incident_count)) +
  geom_line(color = "darkgreen") +
  labs(title = "Number of Incidents Over Time",
       x = "Date",
       y = "Number of Incidents") +
  theme_minimal()
```



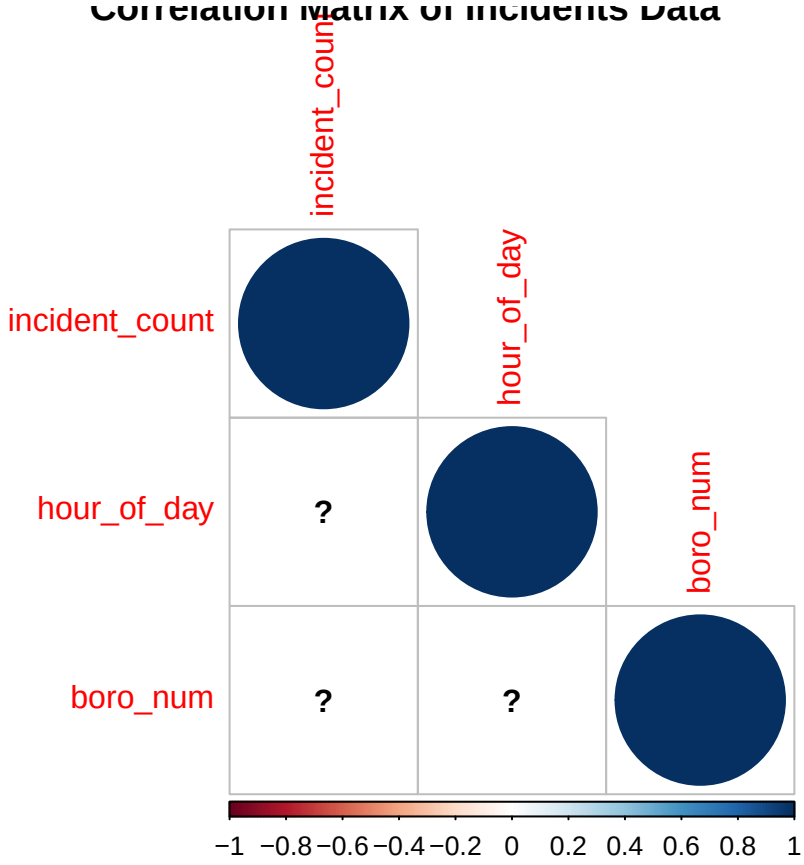

```
# Histogram of incident counts
numeric_cols <- nyc_data_summary %>% select_if(is.numeric)
if (nrow(numeric_cols) > 0) {
  ggplot(nyc_data_summary, aes(x = incident_count)) +
    geom_histogram(binwidth = 10, fill = "steelblue", color = "white") +
    labs(title = "Distribution of Incidents", x = "Incident Count", y = "Frequency")
} else {
  print("No numeric columns found for histogram.")
}
```




```
boroughs_to_compare <- c("MANHATTAN", "BROOKLYN")

filtered_data <- nyc_data %>%
  filter(boro %in% boroughs_to_compare)

incident_counts <- filtered_data %>%
  group_by(boro) %>%
  summarise(incident_count = n(), .groups = 'drop')

# Visualize the frequency of incidents by borough
ggplot(incident_counts, aes(x = reorder(boro, -incident_count), y = incident_count, fill = boro)) +
  geom_bar(stat = "identity") +
  labs(title = "Frequency of Incidents by Borough",
       x = "Borough",
       y = "Incident Count") +
  theme_minimal() +
  theme(legend.position = "none")
```




```r
# Correlation analysis
filtered_incident_counts <- filtered_data %>%
  group_by(boro) %>%
  summarise(incident_count = n(),
            hour_of_day = hour(occur_date),
            boro_num = as.numeric(factor(boro)),
            .groups = 'drop')

correlation_data <- filtered_incident_counts %>%
  select(incident_count, hour_of_day, boro_num)

if (nrow(correlation_data) > 0) {
  cor_matrix <- cor(correlation_data, use = "complete.obs")

  # Visualize the correlation matrix
  corrplot(cor_matrix, method = "circle", type = "lower",
           title = "Correlation Matrix of Incidents Data")
} else {
  print("No numeric columns found for correlation analysis.")
}
```


Correlation Matrix of Incidents Data
incident_count
hour_of_day
boro_num
incident_count
hour_of_day
boro_num
?
?
?
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1