Road accidents in New York

Luca Renz & Alvaro Cervan

30/08/2024

Contents

0.1 Road accidents in New York

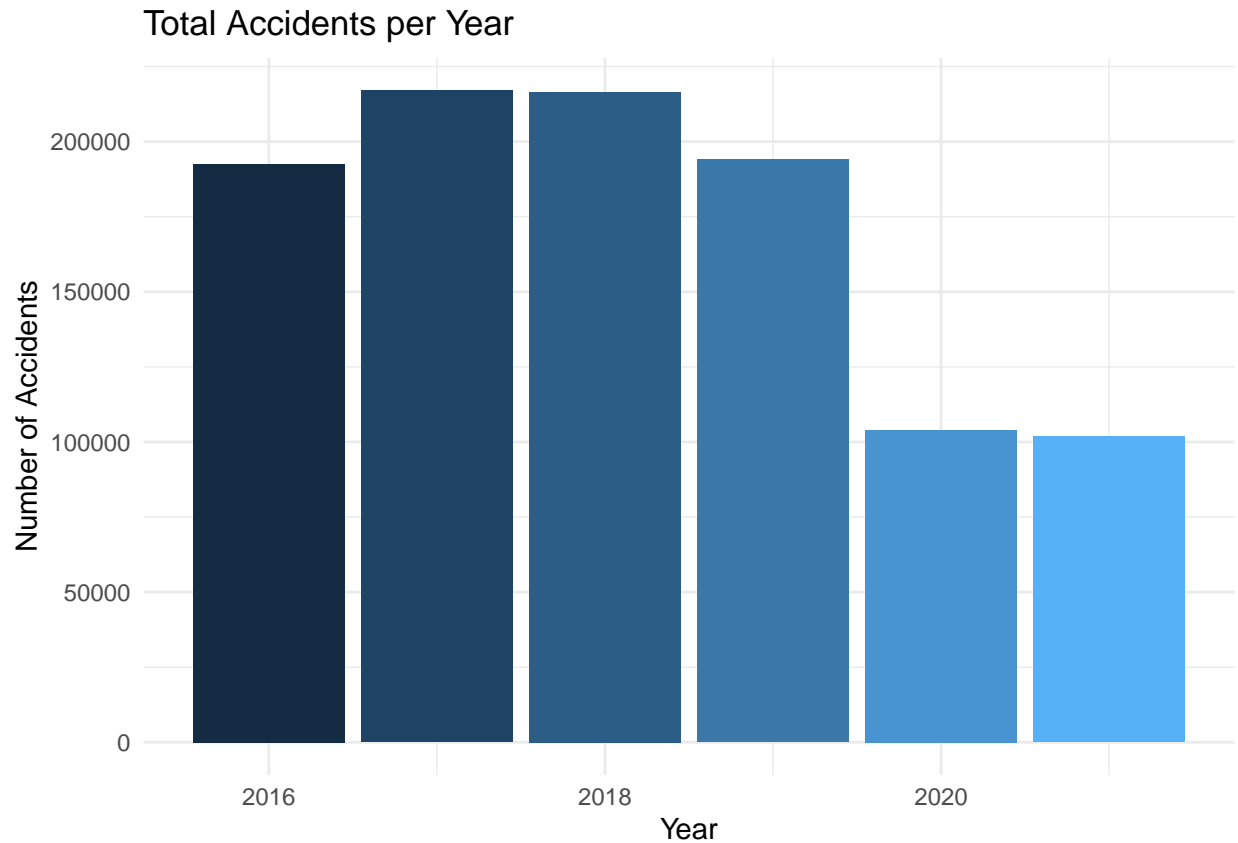### 0.1.1 Total Accidents in New York

```r
# Loading necessary libraries
library(ggplot2)
library(dplyr)

# Load the data from a CSV file
data <- read.csv("data/merged_all_years.csv", header = TRUE)

# Convert the CRASH.DATE into a Date object
data$CRASH.DATE <- as.Date(data$CRASH.DATE, "%d/%m/%Y")

# Aggregating data by Year
yearly_accidents <- data %>%
  group_by(Year) %>%
  summarise(Total_Accidents = n(), .groups = 'drop')

# Plotting the data
ggplot(yearly_accidents, aes(x = Year, y = Total_Accidents, fill = Year)) +
  geom_bar(stat = "identity") +  # Set a single color for the bars
  labs(title = "Total Accidents per Year",
       x = "Year",
       y = "Number of Accidents") + theme_minimal() + theme(legend.position = "none")
```

Total Accidents per Year


```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(lubridate)
library(readxl)

# Load the data
accident_data <- read.csv("data/merged_data_2021.csv", header = TRUE)
weather_data <- read_excel("data/weather_2021.xlsx")

# Step 1: Extract Date and Hour from the 'time' column in the weather data
weather_data <- weather_data %>%
  mutate(time = as.POSIXct(time, format = "%Y-%m-%dT%H:%M"))

# Step 2: Convert CRASH.DATE in accident_data to Date type
accident_data <- accident_data %>%
  mutate(CRASH.DATE = as.Date(CRASH.DATE, format = "%Y-%m-%d"))

# Step 3: Match accidents with corresponding hourly rainfall data
accident_data <- accident_data %>%
  mutate(CRASH.DATETIME = as.POSIXct(paste(CRASH.DATE, CRASH.TIME), format = "%Y-%m-%d %H:%M")) %>%
  mutate(CRASH.DATETIME = floor_date(CRASH.DATETIME, "hour"))

accidents_with_rainfall <- left_join(accident_data, weather_data, by = c("CRASH.DATETIME" = "time"))

# Step 4: Categorize rainfall into meaningful bins
```

```r
accidents_with_rainfall <- accidents_with_rainfall %>%
  mutate(Rainfall_Category = case_when(
    rain..mm. == 0 ~ "0 mm (No rain)",
    rain..mm. > 0 & rain..mm. <= 4 ~ "1 mm - 4 mm (Light rain)",
    rain..mm. > 4 & rain..mm. <= 7 ~ ">4 mm - 7 mm (Moderate rain)",
    rain..mm. > 7 ~ ">7 mm (Heavy rain)"
  ))

# Step 5: Group data by Month and Rainfall Category
accidents_by_month_rainfall <- accidents_with_rainfall %>%
  mutate(Month = month(CRASH.DATETIME, label = TRUE)) %>%
  group_by(Month) %>%
  summarise(Total_Accidents = n(), .groups = 'drop')

# Step 6: Calculate total rainfall per month
total_rainfall_per_month <- weather_data %>%
  mutate(Month = month(time, label = TRUE)) %>%
  group_by(Month) %>%
  summarise(Total_Rainfall = sum(`rain (mm)`))

# Step 7: Merge the accident data with the rainfall data
correlation_data <- left_join(accidents_by_month_rainfall, total_rainfall_per_month, by = "Month")

# Step 8: Calculate the correlation coefficient
correlation_coefficient <- cor(correlation_data$Total_Accidents, correlation_data$Total_Rainfall)
print(paste("Correlation coefficient:", round(correlation_coefficient, 2)))
```
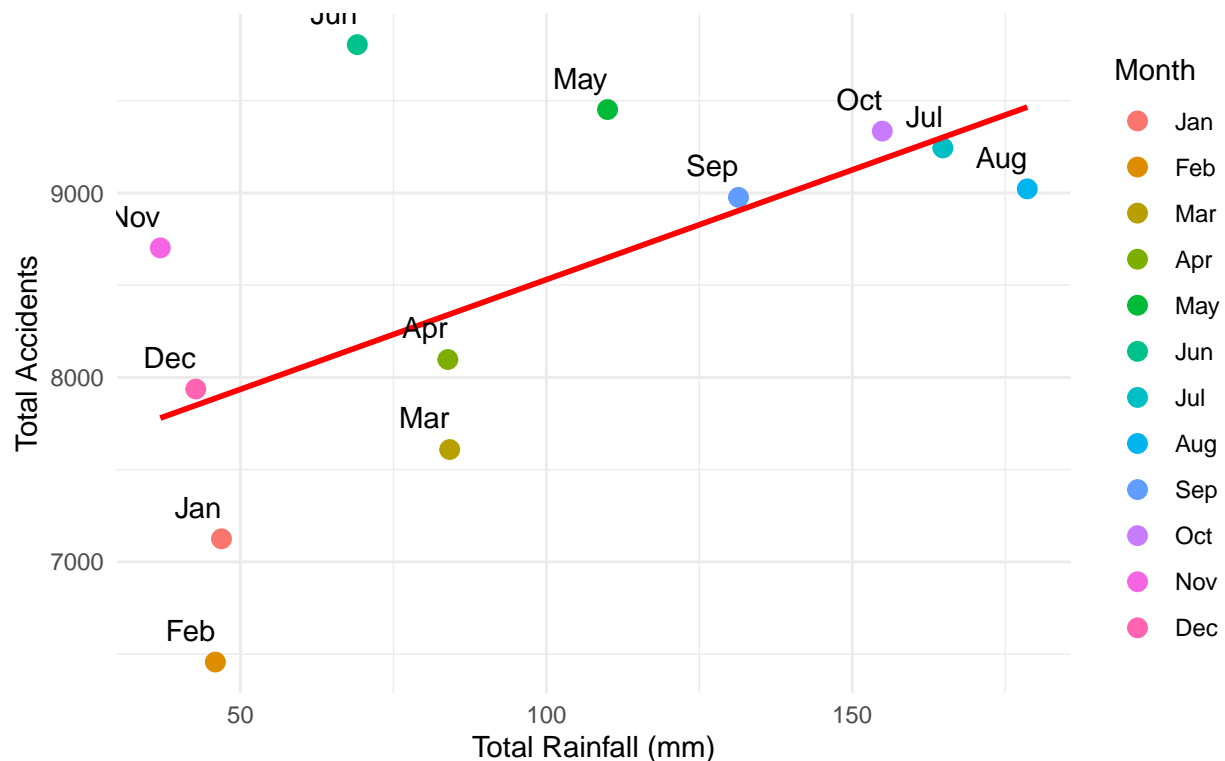
```
## [1] "Correlation coefficient: 0.59"
```

```r
# Step 9: Plot the correlation with labels
ggplot(correlation_data, aes(x = Total_Rainfall, y = Total_Accidents)) +
  geom_point(aes(color = Month), size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  geom_text(aes(label = Month), vjust = -1, hjust = 1, size = 4) +
  labs(title = paste("Correlation between Total Accidents and Total Rainfall per Month\nCorrelation coe
       x = "Total Rainfall (mm)",
       y = "Total Accidents") +
  theme_minimal() +
  theme(legend.position = "right") +
  scale_color_discrete(name = "Month")
```

Correlation between Total Accidents and Total Rainfall per Month
Correlation coefficient: 0.59


```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(lubridate)
library(readxl)

# Load the data
accident_data <- read.csv("data/merged_data_2021.csv", header = TRUE)
weather_data <- read_excel("data/weather_2021.xlsx")

# Step 1: Extract Date and Hour from the 'time' column in the weather data
weather_data <- weather_data %>%
  mutate(time = as.POSIXct(time, format = "%Y-%m-%dT%H:%M"))

# Step 2: Convert CRASH.DATE in accident_data to Date type
accident_data <- accident_data %>%
  mutate(CRASH.DATE = as.Date(CRASH.DATE, format = "%Y-%m-%d"))

# Step 3: Match accidents with corresponding hourly rainfall data
accident_data <- accident_data %>%
  mutate(CRASH.DATETIME = as.POSIXct(paste(CRASH.DATE, CRASH.TIME), format = "%Y-%m-%d %H:%M")) %>%
  mutate(CRASH.DATETIME = floor_date(CRASH.DATETIME, "hour"))

accidents_with_rainfall <- left_join(accident_data, weather_data, by = c("CRASH.DATETIME" = "time"))

# Step 4: Categorize rainfall into meaningful bins
```

```r
accidents_with_rainfall <- accidents_with_rainfall %>%
  mutate(Rainfall_Category = case_when(
    rain..mm. == 0 ~ "0 mm (No rain)",
    rain..mm. > 0 & rain..mm. <= 4 ~ "1 mm - 4 mm (Light rain)",
    rain..mm. > 4 & rain..mm. <= 7 ~ ">4 mm - 7 mm (Moderate rain)",
    rain..mm. > 7 ~ ">7 mm (Heavy rain)"
  ))

# Step 5: Group data by Month and Rainfall Category
accidents_by_month_rainfall <- accidents_with_rainfall %>%
  mutate(Month = month(CRASH.DATETIME, label = TRUE)) %>%
  group_by(Month, Rainfall_Category) %>%
  summarise(Total_Accidents = n(), .groups = 'drop') %>%
  filter(!is.na(Rainfall_Category))  # Drop NA values

# Step 6: Calculate the total number of accidents per month across all categories
total_accidents_per_month <- accidents_by_month_rainfall %>%
  group_by(Month) %>%
  summarise(Total_Accidents_Month = sum(Total_Accidents))

# Step 7: Calculate total rainfall per month
total_rainfall_per_month <- weather_data %>%
  mutate(Month = month(time, label = TRUE)) %>%
  group_by(Month) %>%
  summarise(Total_Rainfall = sum(`rain (mm)`))

# Step 8: Plot the data with dots for each category and a line for the total accidents
ggplot() +
  # Add dots for each category with transparency
  geom_point(data = accidents_by_month_rainfall, aes(x = Month, y = Total_Accidents, color = Rainfall_Ca
  # Add a line for the total number of accidents per month
  geom_line(data = total_accidents_per_month, aes(x = Month, y = Total_Accidents_Month, group = 1), col
  # Add a secondary bar plot for total rainfall per month with a secondary y-axis
  geom_bar(data = total_rainfall_per_month, aes(x = Month, y = Total_Rainfall * 50), stat = "identity",
  scale_y_continuous(
    name = "Number of Accidents",
    sec.axis = sec_axis(~./50, name = "Total Rainfall (mm)")
  ) +
  labs(title = "Monthly Number of Accidents by Rainfall Category with Monthly Rainfall",
       x = "Month") +
  theme_minimal() +
  theme(legend.position = "right")
```
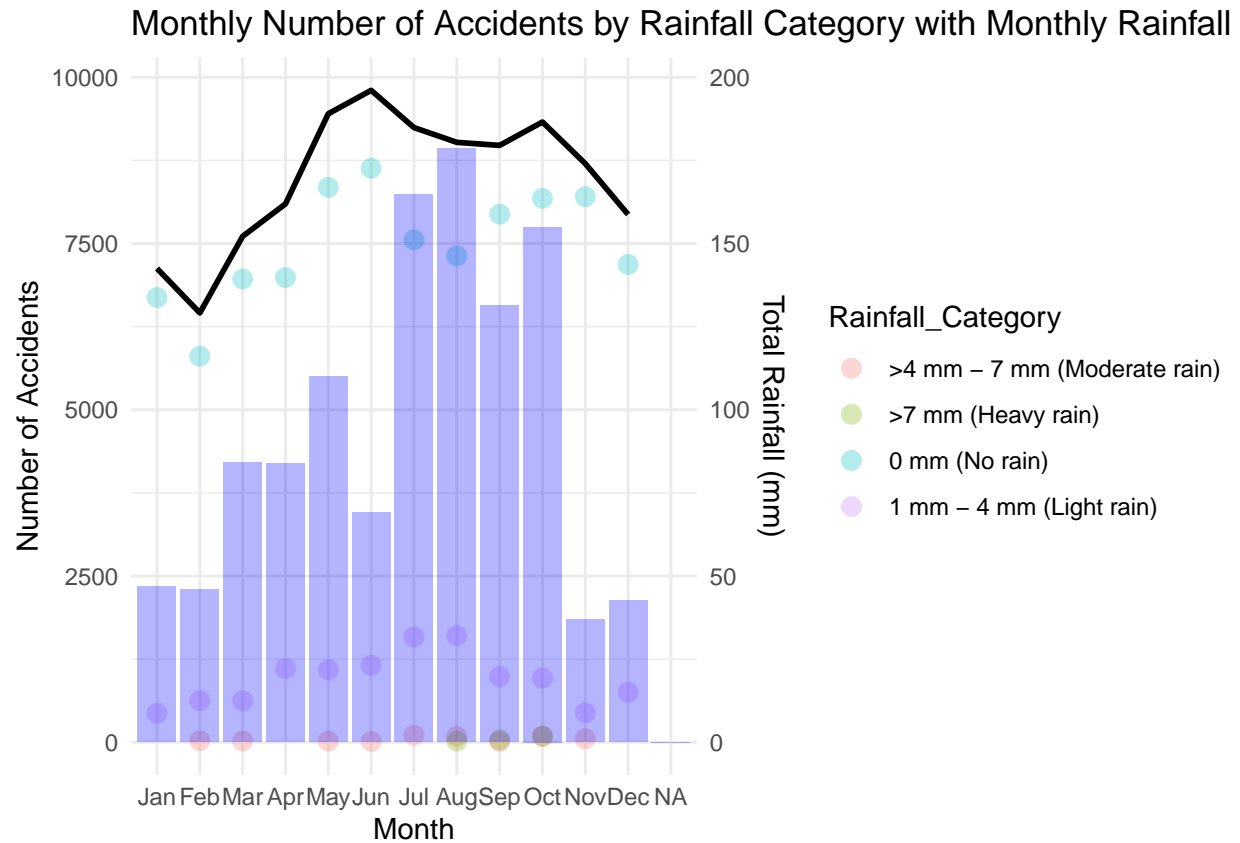
Monthly Number of Accidents by Rainfall Category with Monthly Rainfall

0.2 Chapter of Choice (Esquisse (Web Server App))

```r
#install.packages("esquisse")
#install.packages("plotly")
library(esquisse)
library(plotly)

esquisser(data) # Opens the web app in the Viewer pane or in the browser with the data loaded
```