IS 445 Final Project

1. Introduction

Vehicle usage is increasing in daily life, with an increase in traffic violations and car accidents. Regulations and laws have explicit penalties for different types of traffic violations as global reminders and warnings for drivers. However, not every traffic violation is deliberate, and some types of traffic violations are greatly affected by environmental factors like geographic information or road conditions. Investigation of such traffic violations with clustering effect and certain geographic relationships is essential and meaningful for drivers to avoid unintentional traffic violations. Since different areas have unique geographic features, this work focuses on the impact of geographic factors on traffic violations in Montgomery County in the United States.

I focus on traffic violations in Montgomery County in the United States. This dataset is publicly released with the link. There are 1015271 rows and 35 columns in this dataset, with 3 numeric variables, 29 categorical variables, and 3 date/time variables. Since the variables are mostly categorical, considering there are 1015271 data entries to be chosen, I plan to use a sample size of around 50000-100000 (5%-10%) of the dataset size. If possible, I would also like to increase the sample size from small to large gradually from approximately 1%-50%, thus drawing a reliable and representative conclusion based on sampled data visualization results. All the information will be stored in a single .csv file called Traffic_Violations.csv.

R is the main media to obtain data visualization results between the number/types of traffic violations and geographical locations. As an auxiliary, Python is used to derive quantitative approximations of such relationships via interpolations or regression.

2. Literature Review

Existing works have already reached some success in figuring out the geographic impact of traffic violations. Sukhaia et al. (2013) correlate road conditions to traffic violations using road traffic fatalities in South Africa. Li et al. (2020) estimate the space and time patterns of traffic violation behaviors to investigate the relationship between traffic violations and urban surroundings. Elfahim et al. (2023) perform a comparison between different clustering optimizations to detect common types of traffic violations within certain districts. However, these existing methods are mostly generalized to simple impacts of common geographic features or concentrating on fatal traffic violations, indicating their weaknesses in analyzing area-specific and geographically sensitive data acquired from places like Montgomery County. Those disadvantages of existing work motivate us to develop new and targeted research on traffic violations in Montgomery County.

Back to our focus, investigating geographic effects on traffic violations in Montgomery County is an unprecedented research problem. I would like to figure out the potential impacts of geographical locations or surroundings on the number and types of traffic violations. Specifically, it makes sense to verify if some categories of traffic violations are

dominant in areas with certain geographical locations. There might also be some hidden patterns of geographical locations influencing the distribution of traffic violations. Finally, I hope with the relationship between geographical locations and certain types of traffic violations, we can further think of how certain traffic violations can be avoided, from the drivers' side and society's side.

3. Methods

When it comes to resolving the above research question, several separate steps can be considered in sequence. First of all, it is necessary to obtain a visualization plot showing the positions of all traffic violations. Moreover, it makes sense to investigate the individual effect of geographic locations on the number and types of traffic violations via model fitting and visualization. Furthermore, it is advised to figure out potential patterns hidden between different types of traffic violations with multivariate visualization tools. While visualizing the relationships between abundant variables, it is worth considering if those relationships make sense according to our life experience. If the relationships violate our intuition, investigation into additional geographic information might be necessary. Last but not least, time-series analysis can be applied to traffic violations in Montgomery County to visualize the trends of traffic violations. For the specific model, I performed Principal Component Analysis (PCA) when visualizing the effect of geographic locations on the number and types of traffic violations on a sampled subset of traffic violation data (0.5% out of 1015271 samples).

4. Discussion and Results

Most traffic violations are citations and warnings with similar occurrence frequencies. Drivers in Montgomery County are advised to be cautious of possible citations and warnings when driving and should be aware of traffic laws in Montgomery County beforehand.

Most traffic violations occur around -77±0.05 longitude degrees and 39 to 39.2 latitudes. The shape of traffic violation distribution is like a tilted rectangle heading towards the southeast (where the latitude decreases and longitude increases). If drivers in Montgomery County are equipped with navigation systems, it is highly recommended to set up a location reminder when they drive approaching (39.10, -77.00) or heading southeast.

Only a small part of traffic violations led to accidents eventually, with most accidents happening after citations, indicating well-conditioned roads and a safe traffic atmosphere. For drivers living in Montgomery County, it is wise to keep calm when driving to avoid citations and warnings in case of internal accidents.

There are no obvious patterns between violation types and geological locations since citations and warnings seem to have covered areas around (39.10, -77.00), where the most traffic violations happen. However, citations might be more common as the longitude increases (to about -76.90 degrees), and warnings are still observed as the longitude decreases. Combined with Figure 1 and Figure 3, it is recommended to set up reminders in the area around (39.10, -77.00) and be aware of potential citations and accidents when heading east.

To obtain a better view of common geological locations of traffic violations. I located the area of frequent traffic violations. The results in real maps demonstrate that traffic violations occurred mostly along 270 road and Maryland 200 road from Summer Place to 2501 Green Valley Rd as well as from 8804 34th Avenue to 3111 Woodbine Road.

5. Conclusion

For traffic violations triggered by misleading surroundings, new settings along roads can be tested via similar data collection and visualization methods. For traffic violations related to certain times, further attributes like drivers' ages, jobs, or work hours can be included in visualization and modeling. Simple but explicit instructive signs can be pulled up along the road to indicate potential traffic violations as well. Finally, I hope that diverse relationships between geographic factors and traffic violations are useful reminders and warnings for all drivers in Montgomery County.

6. Reference

[1] Elfahim, O., El Midaoui, M., Youssfi, M., & Bouattane, O. (2023). Traffic violations analysis: Identifying risky areas and common violations. *Heliyon*, *9*(9).
[2] Sukhaia, A., & Jones, A. P. (2013). Understanding geographical variations in road traffic fatalities in South Africa. *South African Geographical Journal= Suid-Afrikaanse Geografiese Tydskrif*, *95*(2), 187-204.
[3] Li, Y., Abdel-Aty, M., Yuan, J., Cheng, Z., & Lu, J. (2020). Analyzing traffic violation behavior at urban intersections: A spatio-temporal kernel density estimation approach using automated enforcement system data. Accident Analysis & Prevention, 141

7. Tables and Figures

Table 1. Basic Information of the dataset

Row No.	Col No.	Numeric Var.	Categorical Var.
1015271	35	3	29

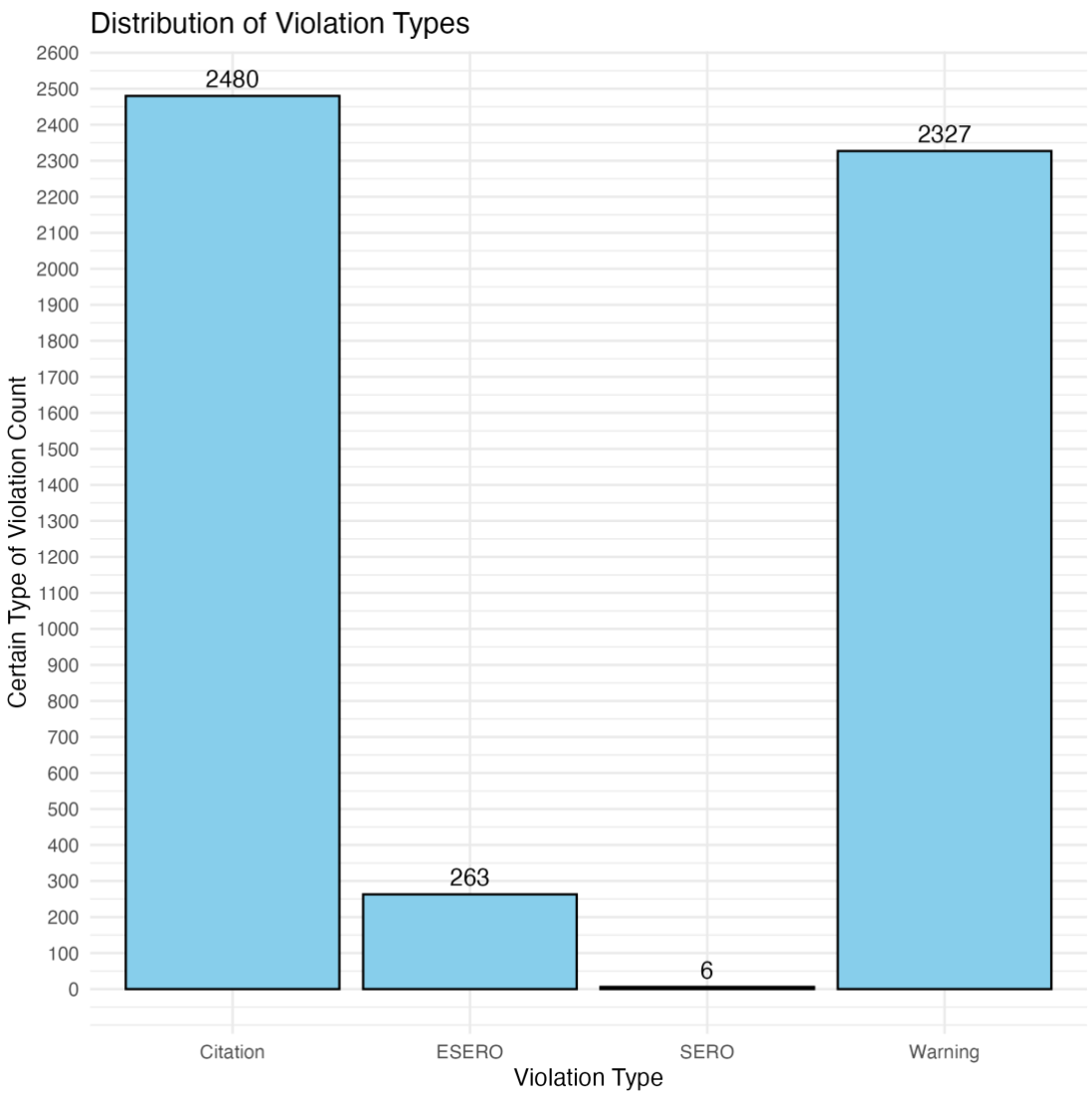

Figure 1. Distribution of Traffic Violation Types via Bar Chart. An example of ESERO is wearing decorative items violating traffic regulations. For SERO violations, most are due to destroyed exterior car parts or headlights.

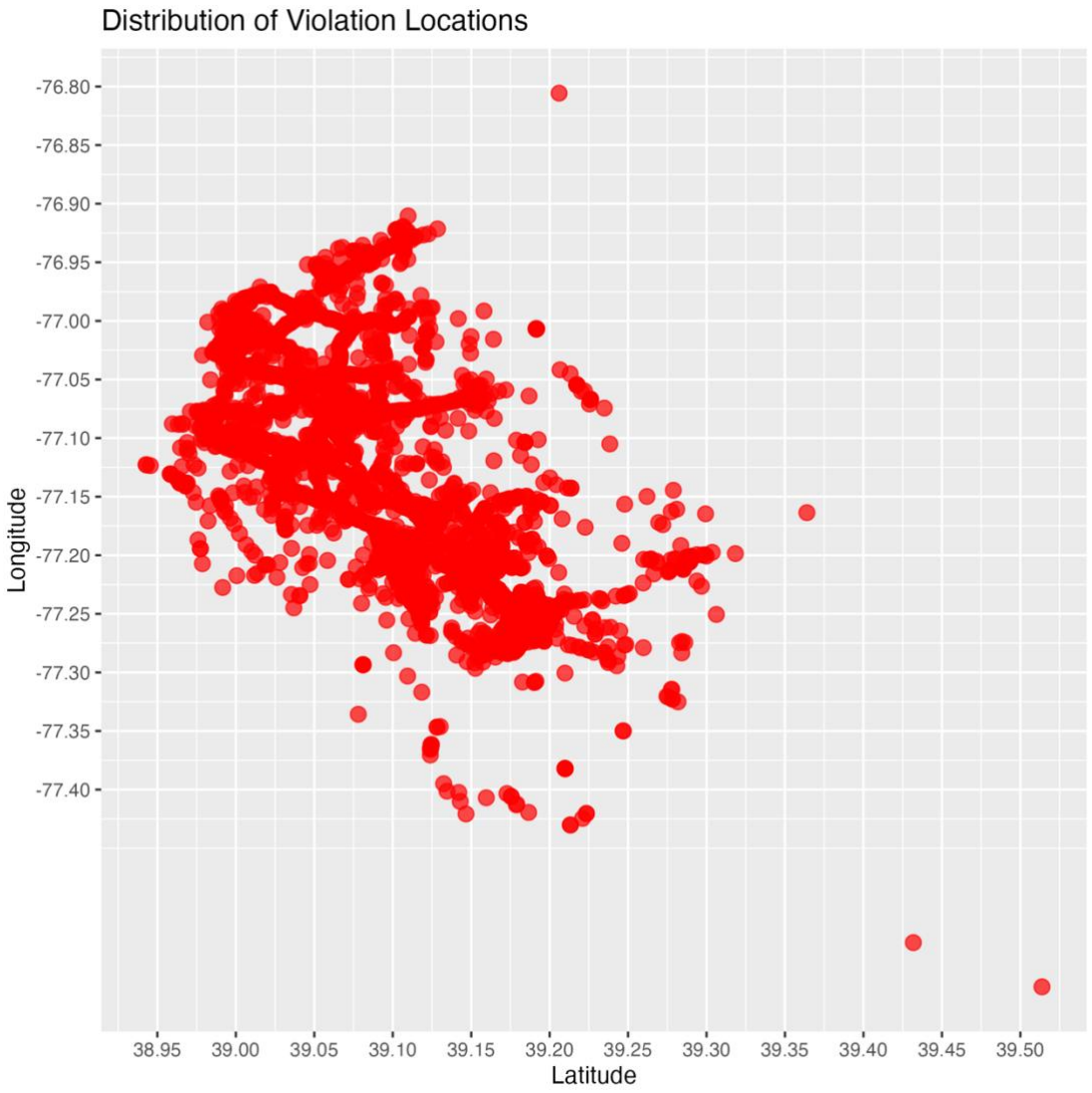

Figure 2. Distribution of Traffic Violation Locations

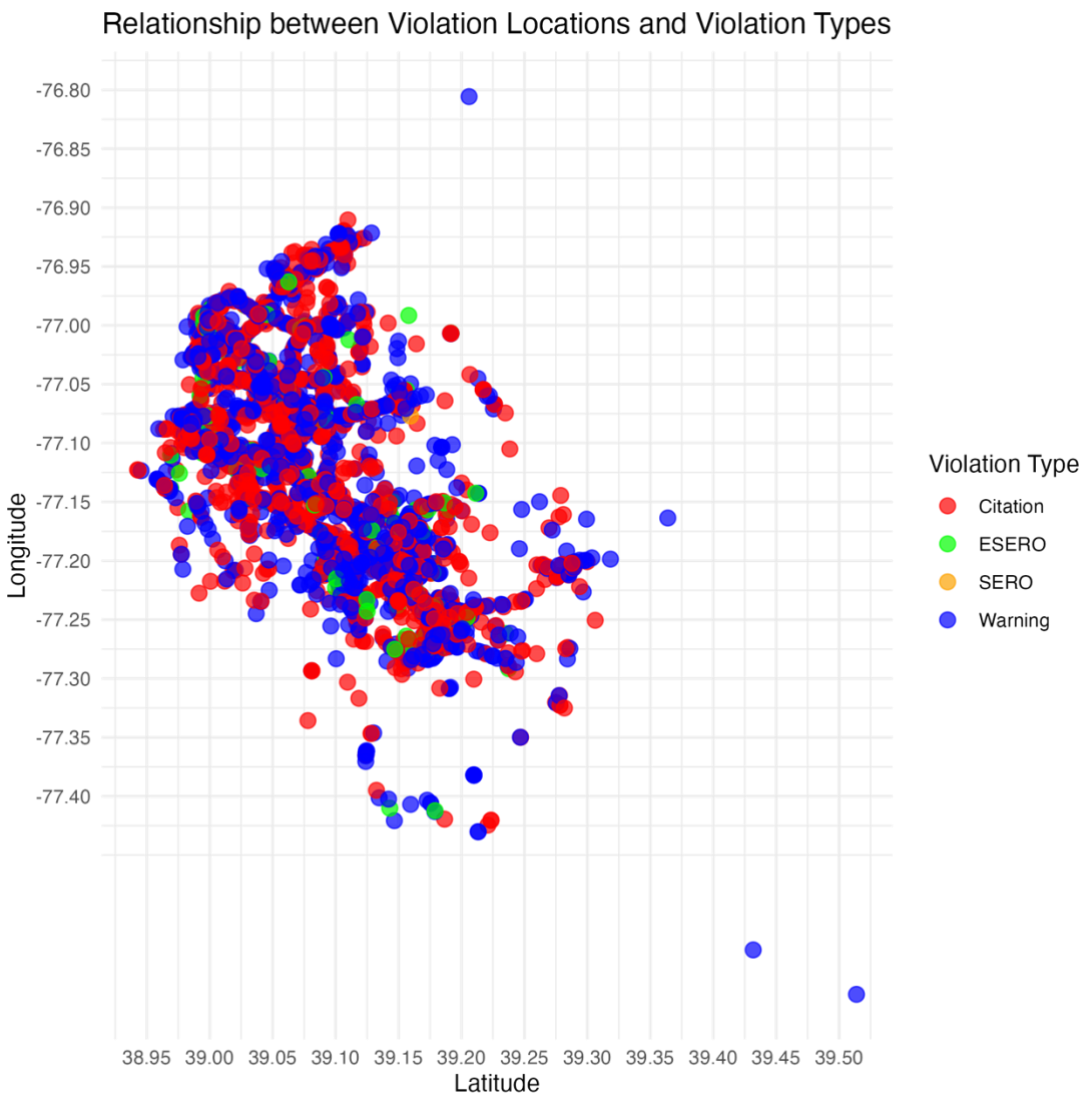

Figure 3. Relationship between Traffic Violation Locations and Violation Types

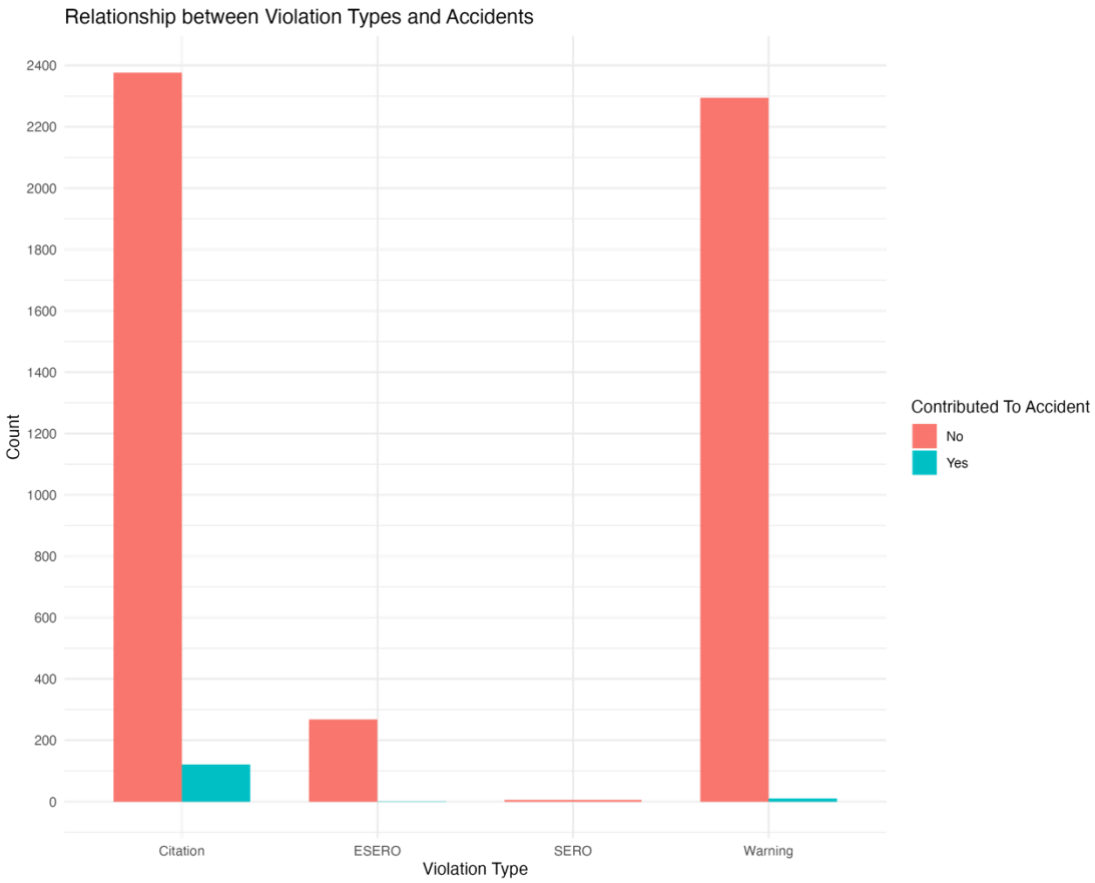

Figure 4. Relationship between Traffic Violation Types and Accidents

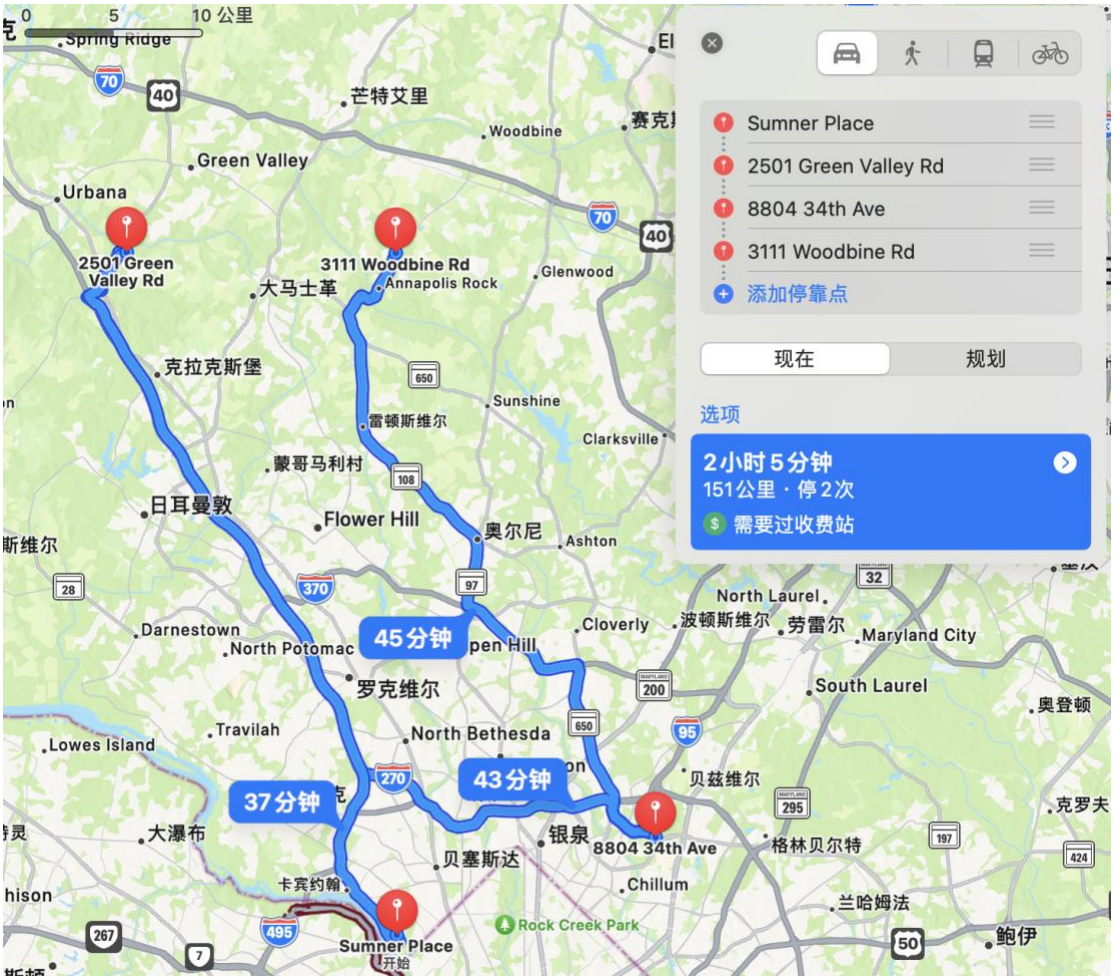

Figure 5. Real-world locations of common traffic violations.

8. Appendix (R Script)

```
library(readr)
library(readxl)
library(ggplot2)
library(ggpubr)
library(VIM)
library(GGally)
library(dplyr)
library(tidyverse)
library(superheat)
library(scales)
library(ggradar)
library(ggrepel)
library(waterfalls)

Attrition <- read_csv("/Users/wendiwang/Downloads/Traffic_Violations-2.csv")
# Attrition <- sample_frac(Attrition, 0.05, replace = FALSE)
df <- sample_frac(Attrition, 0.005, replace = FALSE)
```

```
print(df)
print(colnames(Attrition))

title <- Attrition %>%
  filter(`Contributed To Accident` %in% c("Yes")) %>%
  select(`Contributed To Accident`, Latitude, Longitude)
plot <- ggplot(title, aes(x = Latitude, y = Longitude)) +
  geom_point(color = "orange", size = 3, alpha = 0.7) +
  scale_x_continuous(
    breaks = seq(38.7, 39.9, by = 0.05)
  ) +
  scale_y_continuous(
    breaks = seq(-77.4, -68.8, by = 0.05)
  ) +
  labs(
    title = "Distribution of Accident Locations",
    x = "Latitude",
    y = "Longitude"
  ) +
  theme(
    plot.title = element_text(size = 24, hjust = 0.5),
    axis.title.x = element_text(size = 20),
    axis.title.y = element_text(size = 20),
    axis.text.x = element_text(size = 16),
    axis.text.y = element_text(size = 16),
    legend.title = element_text(size = 18),
    legend.text = element_text(size = 16)
  ) +
  theme_minimal()
ggsave("dv-0113-5.png", plot=plot)


ift <- ggplot(Attrition, aes(x = `Violation Type`)) +    # Ensure column name is correct
(use backticks if there are spaces)
  geom_bar(fill = "skyblue", color = "black") +    # Create bar plot
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  labs(title = "Distribution of Violation Types", x = "Violation Type", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(
    breaks = seq(0, 5000, by = 100)
  ) +
  theme_minimal()
# ggsave("dv-0113-1.png", plot=ift)
```

```
df <- Attrition %>%
  select(Latitude, Longitude)

plot <- scatter_plot <- ggplot(df, aes(x = Latitude, y = Longitude)) +
  geom_point(color = "orange", size = 3, alpha = 0.7) +
  scale_x_continuous(
    breaks = seq(38.7, 39.9, by = 0.05)
  ) +
  scale_y_continuous(
    breaks = seq(-77.4, -68.8, by = 0.05)
  ) +
  labs(title = "Distribution of Violation Locations", x = "Latitude", y = "Longitude")
# ggsave("dv-0113-2.png", plot=plot)

plot <- scatter_plot <- ggplot(Attrition, aes(x = Latitude, y = Longitude, color =
`Violation Type`)) +
  geom_point(size = 3, alpha = 0.7) +
  scale_color_manual(values = c("Citation" = "red", "Warning" = "blue", "ESERO" =
"green", "SERO" = "orange")) +
  scale_x_continuous(
    breaks = seq(38.7, 39.9, by = 0.05)
  ) +
  scale_y_continuous(
    breaks = seq(-77.4, -68.8, by = 0.05)
  ) +
  labs(title = "Relationship between Violation Locations and Violation Types", x =
"Latitude", y = "Longitude", color = "Violation Type") +
  theme_minimal()
# ggsave("dv-0113-3.png", plot=plot)

cat_cat <- ggplot(Attrition, aes(x = `Violation Type`, fill = `Contributed To Accident`)) +
  geom_bar(position = "stack", width = 0.3) +
  theme(
    plot.title = element_text(size = 24),             # Title font size
    axis.title.x = element_text(size = 20),       # X-axis title font size
    axis.title.y = element_text(size = 20),       # Y-axis title font size
    axis.text.x = element_text(size = 16, angle = 45, hjust = 1),    # X-axis text size and
angle
    axis.text.y = element_text(size = 16),         # Y-axis text size
    legend.title = element_text(size = 18),       # Legend title font size
    legend.text = element_text(size = 16)           # Legend text font size
  ) +
  scale_y_continuous(
    breaks = seq(0, 2500, by = 200)    # Set Y-axis breaks
```

```
  ) +
  labs(
    title = "Relationship between Violation Types and Accidents",
    x = "Violation Type",
    y = "Count"
  ) +
  theme_minimal()

cat_cat <- ggplot(Attrition, aes(x = `Violation Type`, fill = `Contributed To Accident`)) +
  geom_bar(position = "dodge", width = 0.7) +
  theme(
    plot.title = element_text(size = 24),
    axis.title.x = element_text(size = 20),
    axis.title.y = element_text(size = 20),
    axis.text.x = element_text(size = 16, angle = 45, hjust = 1),
    axis.text.y = element_text(size = 16),
    legend.title = element_text(size = 18),
    legend.text = element_text(size = 16)
  ) +
  scale_y_continuous(
    breaks = seq(0, 2500, by = 200)
  ) +
  labs(
    title = "Relationship between Violation Types and Accidents",
    x = "Violation Type",
    y = "Count"
  ) +
  theme_minimal()
ggsave("2-0120.png", plot = cat_cat, width = 10, height = 8)
# Save the plot
# ggsave("dv-0113-4.png", plot = cat_cat, width = 10, height = 8)
```