TriMet Analysis

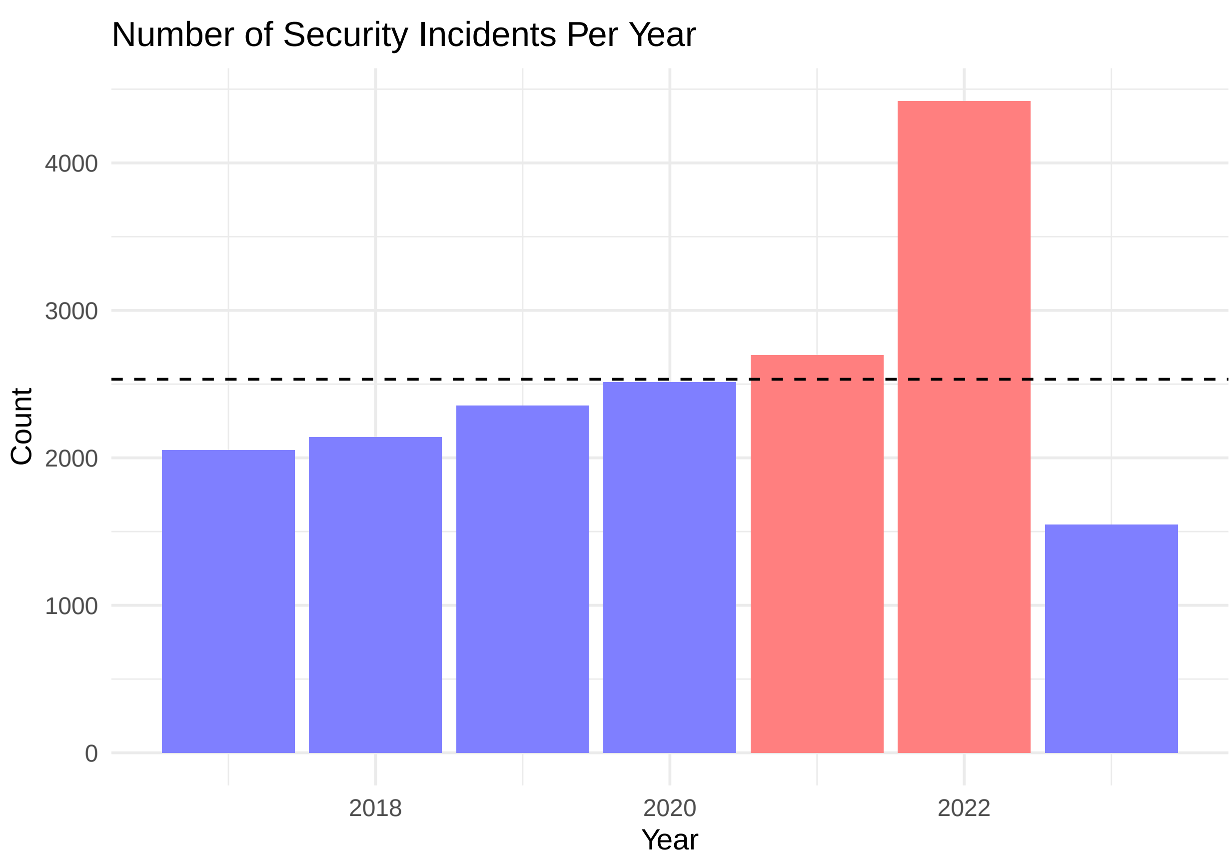
Karol Orozco, Corey Cassell, Justus Eaglesmith, Charles Hanks, & CorDarryl Hall

2023-06-05

Exploratory Data Analysis of Trimet Security Data, 2017 - 2023

Yearly Counts

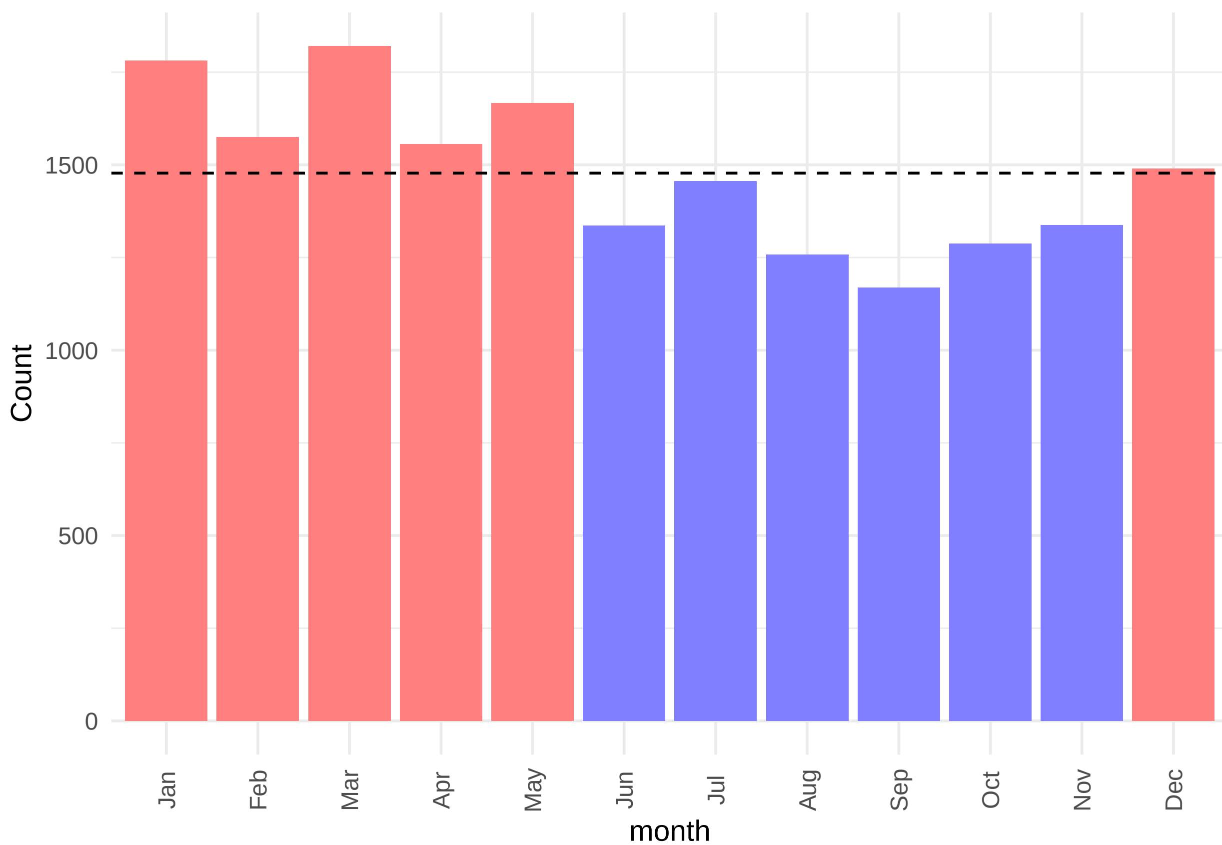
```
# Plot yearly counts
ggplot(df_year_counts, aes(x = year, y = n, fill = color)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_hline(aes(yintercept = avg), linetype = "dashed", color = "black") +
  scale_fill_manual(values = c("Above Average" = "#FF7F7F", "Below Average" = "#7F7FFF")) +
  theme_minimal() +
  labs(x = "Year", y = "Count", fill = "", title = "Number of Security Incidents Per Year")
```



The incidents are increasing annually so it is important to get a handle on this. Especially considering 2022 had tremendously more results than 2021. A better plot for this would be using a timeseries plot.

Monthly Counts

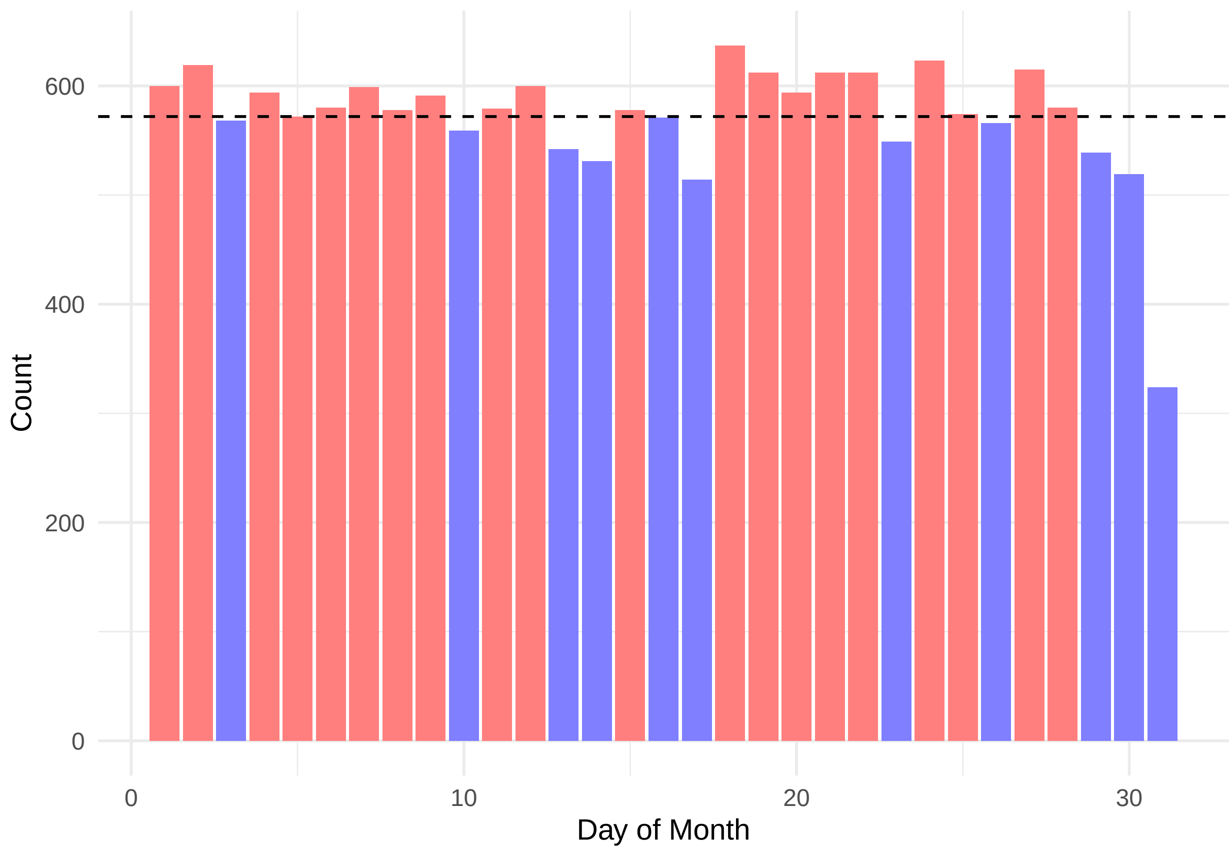
```
# Calculate counts and average
df_month_counts <- df %>%
  count(month) %>%
  mutate(avg = mean(n),
         color = ifelse(n > avg, "Above Average", "Below Average"))

ggplot(df_month_counts, aes(x = month, y = n, fill = color)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_hline(aes(yintercept = avg), linetype = "dashed", color = "black") +
  scale_fill_manual(values = c("Above Average" = "#FF7F7F", "Below Average" = "#7F7FFF")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
  labs(y = "Count", fill = "")
```



We should focus on the months Dec - May as those have the most incidents and are above the average. It would be valuable to understand what could be causing the increase in these specific months as well.

Daily Counts

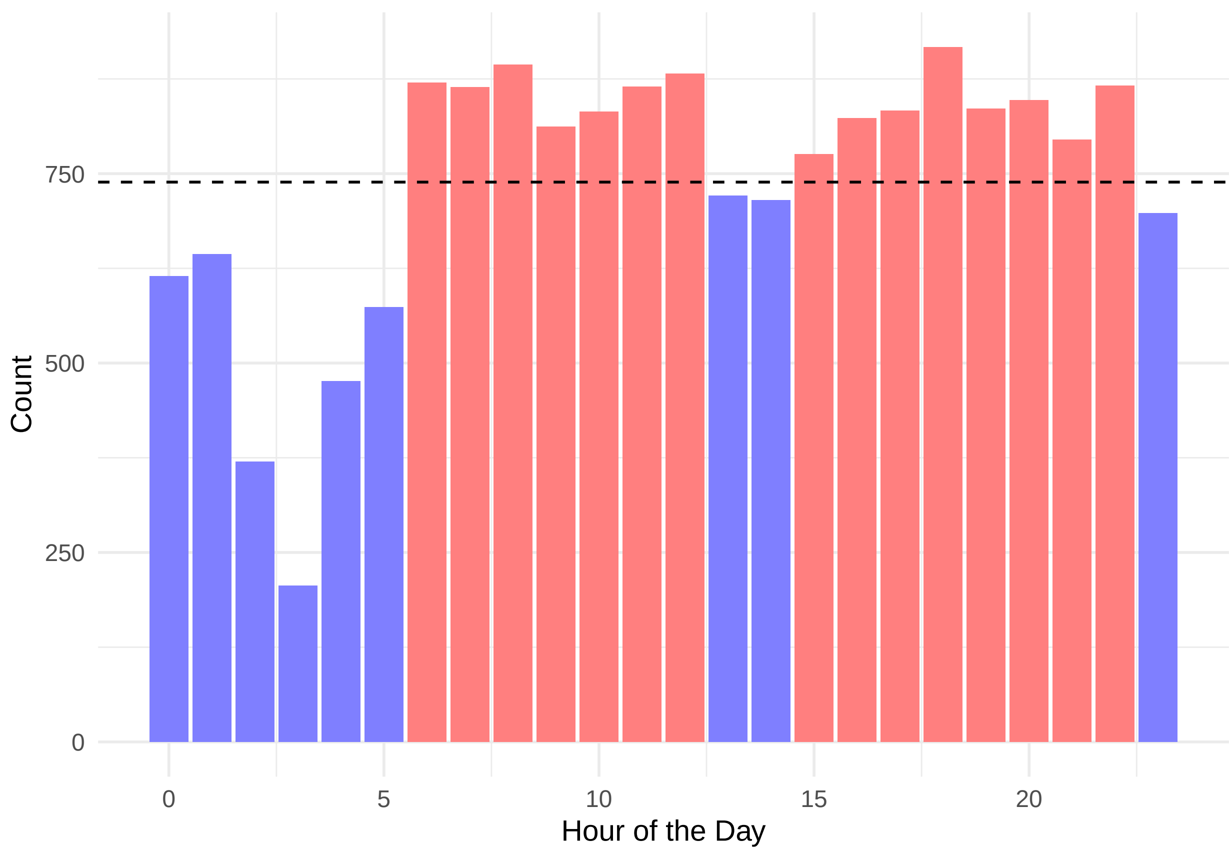
```r
# Calculate counts and average for each day
df_day_counts <- df %>%
  count(day) %>%
  mutate(avg = mean(n),
         color = ifelse(n > avg, "Above Average", "Below Average"))

# daily counts
ggplot(df_day_counts, aes(x = day, y = n, fill = color)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_hline(aes(yintercept = avg), linetype = "dashed", color = "black") +
  scale_fill_manual(values = c("Above Average" = "#FF7F7F", "Below Average" = "#7F7FFF")) +
  theme_minimal() +
  labs(x = "Day of Month", y = "Count", fill = "")
```



Days are a little sporatic but it looks like earlier in the month and later in the months there are more incidents, first and last week of the month specifically.

Hourly Incidents

```r
#Hourly incidents
library(lubridate)

# Create a new hour field
df$hour <- hour(as.POSIXct(df$incident_date, format="%m/%d/%Y %H:%M"))

# Calculate counts and average for each hour
df_hour_counts <- df %>%
  count(hour) %>%
  mutate(avg = mean(n),
         color = ifelse(n > avg, "Above Average", "Below Average"))

# Plot hourly counts
ggplot(df_hour_counts, aes(x = hour, y = n, fill = color)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_hline(aes(yintercept = avg), linetype = "dashed", color = "black") +
  scale_fill_manual(values = c("Above Average" = "#FF7F7F", "Below Average" = "#7F7FFF")) +
  theme_minimal() +
  labs(x = "Hour of the Day", y = "Count", fill = "")
```



Clearly it shows that many incidents occur at hour 6-12 and 3-12a. Makes sense as most people are commuting or using the transportation services in the morning before work/school and after work/school.

```r
#Parsing text to extract common themes amoungst reported incidents
comments_corpus <- Corpus(VectorSource(df$comments))
```

```r
comments_corpus <- tm_map(comments_corpus, content_transformer(tolower))
```

```
## Warning in tm_map.SimpleCorpus(comments_corpus, content_transformer(tolower)):
## transformation drops documents
```

```r
comments_corpus <- tm_map(comments_corpus, removePunctuation)
```

```
## Warning in tm_map.SimpleCorpus(comments_corpus, removePunctuation):
## transformation drops documents
```

```r
comments_corpus <- tm_map(comments_corpus, removeNumbers)
```

```
## Warning in tm_map.SimpleCorpus(comments_corpus, removeNumbers): transformation
## drops documents
```

```r
comments_corpus <- tm_map(comments_corpus, removeWords, stopwords("english"))
```

```
## Warning in tm_map.SimpleCorpus(comments_corpus, removeWords,
## stopwords("english")): transformation drops documents
```

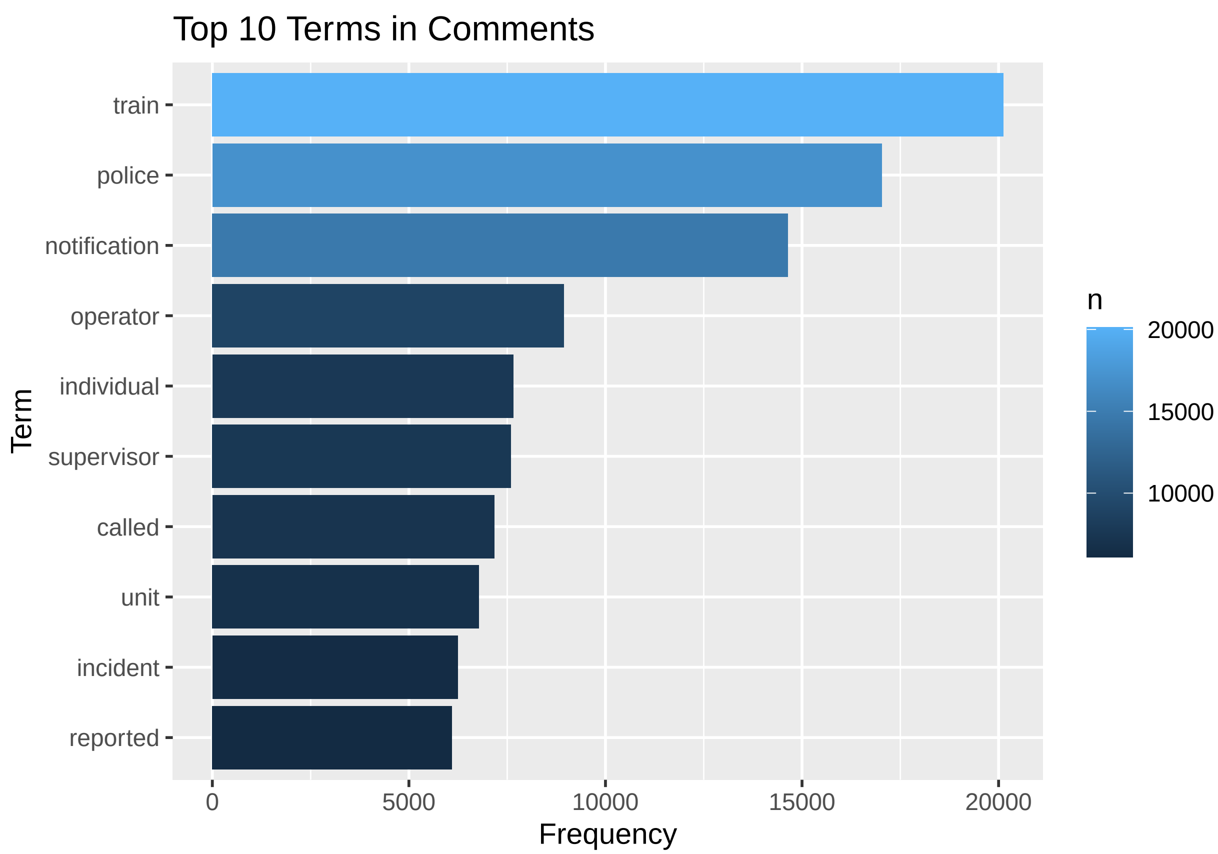
```r
comments_corpus <- tm_map(comments_corpus, stemDocument)
```

```
## Warning in tm_map.SimpleCorpus(comments_corpus, stemDocument): transformation
## drops documents
```

```r
library(tidytext)
# Converting the text to lower case
df$comments <- tolower(df$comments)

# Removing punctuation, numbers, stop words and white spaces
df$comments <- removePunctuation(df$comments)
df$comments <- removeNumbers(df$comments)
df$comments <- removeWords(df$comments, stopwords("english"))
df$comments <- stripWhitespace(df$comments)

# Tokenizing the words
df_tokens <- df %>%
  unnest_tokens(word, comments)

# Counting the frequency of each word
df_word_counts <- df_tokens %>%
  count(word, sort = TRUE)

# Filtering out words with less than 3 characters
df_word_counts <- df_word_counts[nchar(df_word_counts$word) > 2, ]

# Displaying the top 10 words
top_10_words <- df_word_counts %>%
  top_n(10) %>%
  mutate(word = reorder(word, n))
```

```
## Selecting by n
```

```
ggplot(top_10_words) +
  geom_col(aes(x = word, y = n, fill = n)) +
  labs(x = "Term", y = "Frequency", title = "Top 10 Terms in Comments") +
  coord_flip()
```



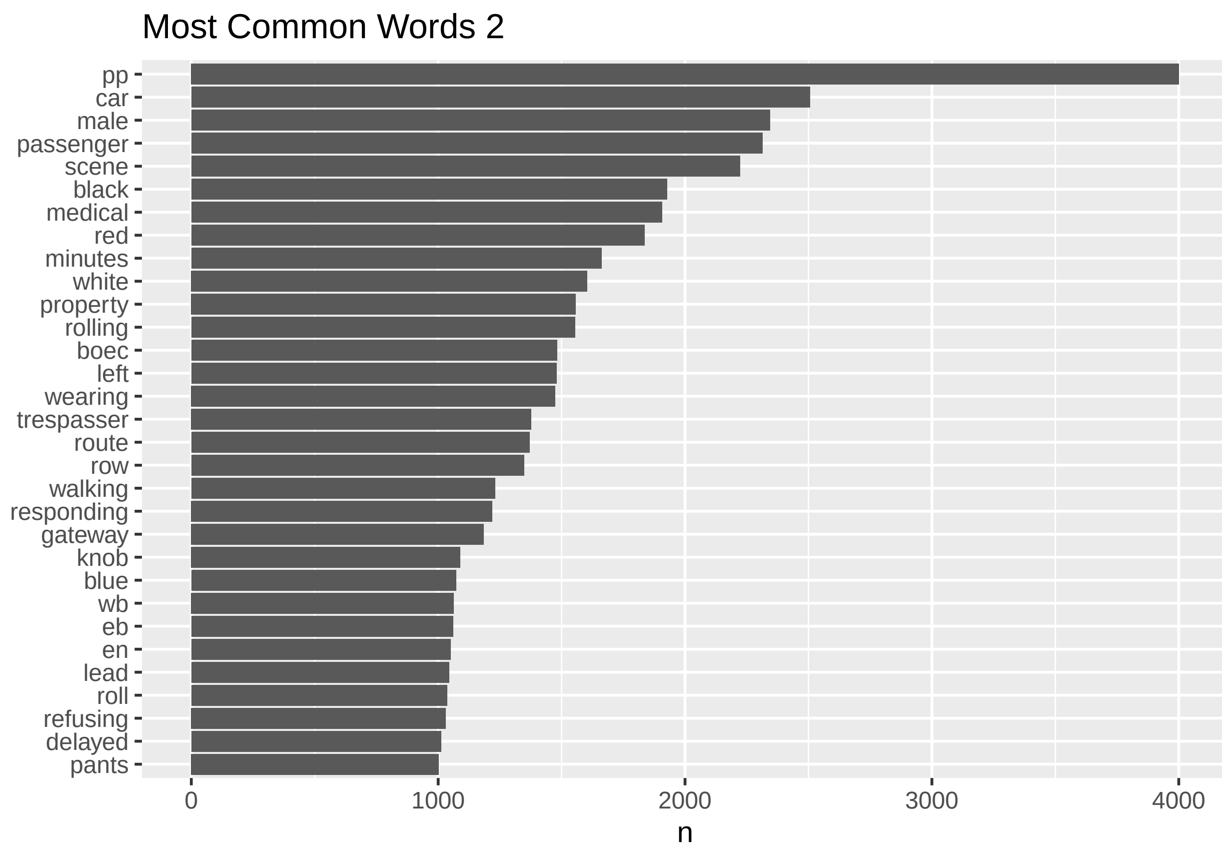
Removing common words:

```
dtm <- df %>%
unnest_tokens(word, comments) %>%
anti_join(stop_words) %>% # get rid of stop words
filter(!(word %in% c("train","notification", "police", "reports","cleared","trains","time","ave","due",
count(incident_id, word) %>%
group_by(incident_id) %>%
mutate(freq = n/sum(n)) %>%
mutate(exists = (n>0)) %>%
ungroup %>%
group_by(word) %>%
mutate(total = sum(n))
```

```
## Joining with `by = join_by(word)`
```

```
dtm %>%
count(word, sort = TRUE) %>%
```

```r
  filter(n > 1000) %>%
    ggplot(aes(x = n , y= reorder(word,n))) + geom_col() + labs(y = NULL) + labs(title = "Most Common W
```



Incidents occur mostly on the train it appears, however we should look at the next most common phrases or nouns/adjectives to get a better understanding.

Analyzing Security Incidents at Night

What type of security incidents occurs most frequently as night ? This will require subcategory per each incident. At present in this dataset, 82% of incidents have subcategory of 'other'. We could generate subcategories through text analysis of `comments` column.

```r
df %>% group_by(subtype_desc) %>% count()
```

```
## # A tibble: 20 x 2
## # Groups:   subtype_desc [20]
##    subtype_desc                      n
##    <chr>                         <int>
##  1 Assault-Employee                254
##  2 Bomb                              6
##  3 Facility                         31
##  4 Fight                           260
##  5 Hijack                            1
```

```
##   6 Homicide                          5
##   7 Hostage                           1
##   8 Park & Ride                     199
##   9 ROW Trespasser                 1159
## 10 ROW Trespasser -Non-reportable   185
## 11 Robbery                          47
## 12 Robbery w/weapon                  2
## 13 Suspicious Package               17
## 14 TVM Break-in                     29
## 15 Theft to Gain Access             12
## 16 Tow (Non-TriMet Vehicle)        312
## 17 Vandalism                       391
## 18 WES                               1
## 19 Weapon                          225
## 20 [Other]                       14594
```

```r
14591/nrow(df)
```

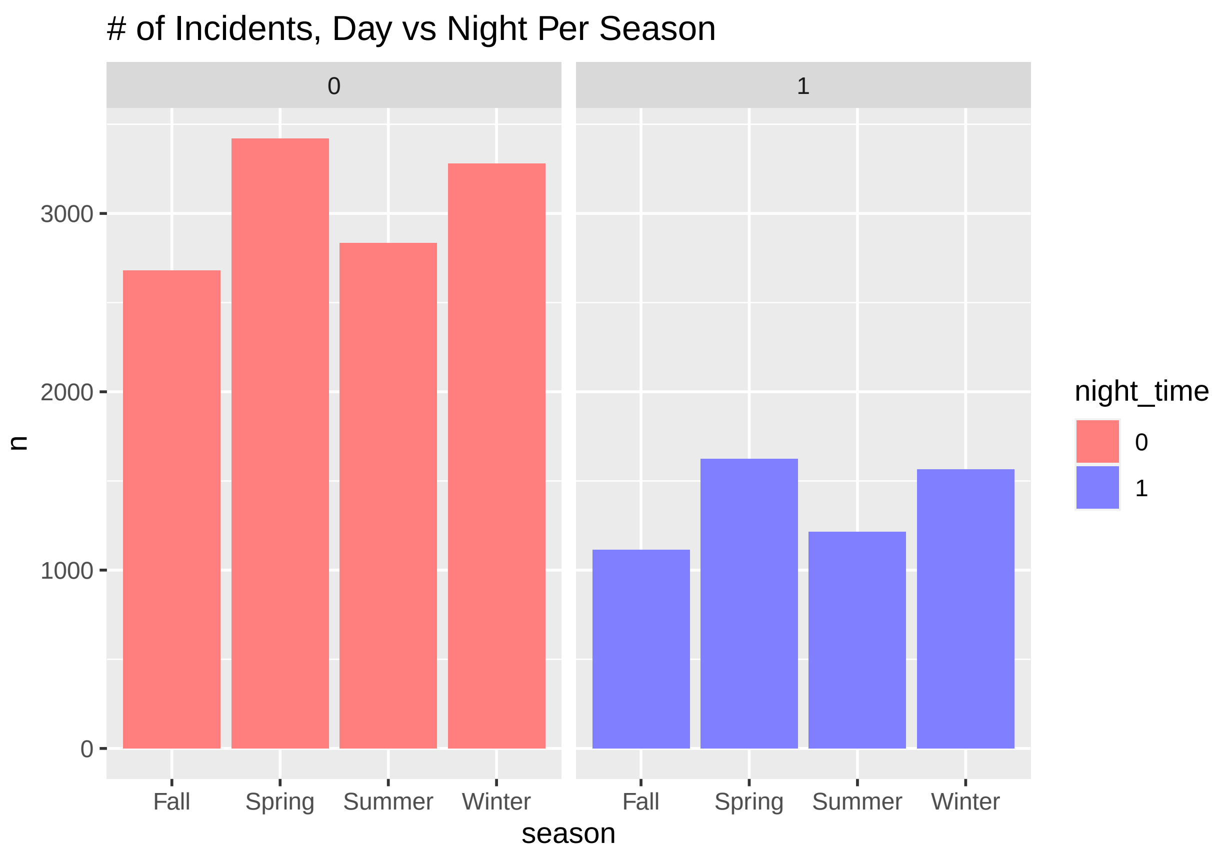
```
## [1] 0.822909
```

Winter : December - February Spring : April - June Summer : July - August Fall : September - November

```r
df = df %>% mutate(season = factor(case_when(
                          month %in% c('Dec','Jan','Feb') ~ 'Winter',
                          month %in% c('Mar','Apr', 'May') ~ 'Spring',
                          month %in% c('Jun', 'Jul', 'Aug') ~ 'Summer',
                          month %in% c('Sep', 'Oct','Nov') ~ 'Fall')),
              night_time = factor(ifelse(hour >= 20 | hour <= 4, 1, 0)))
```

Approximating nighttime as between the hours of 20:00 and 05:00, how many incidents occur at night vs during day?

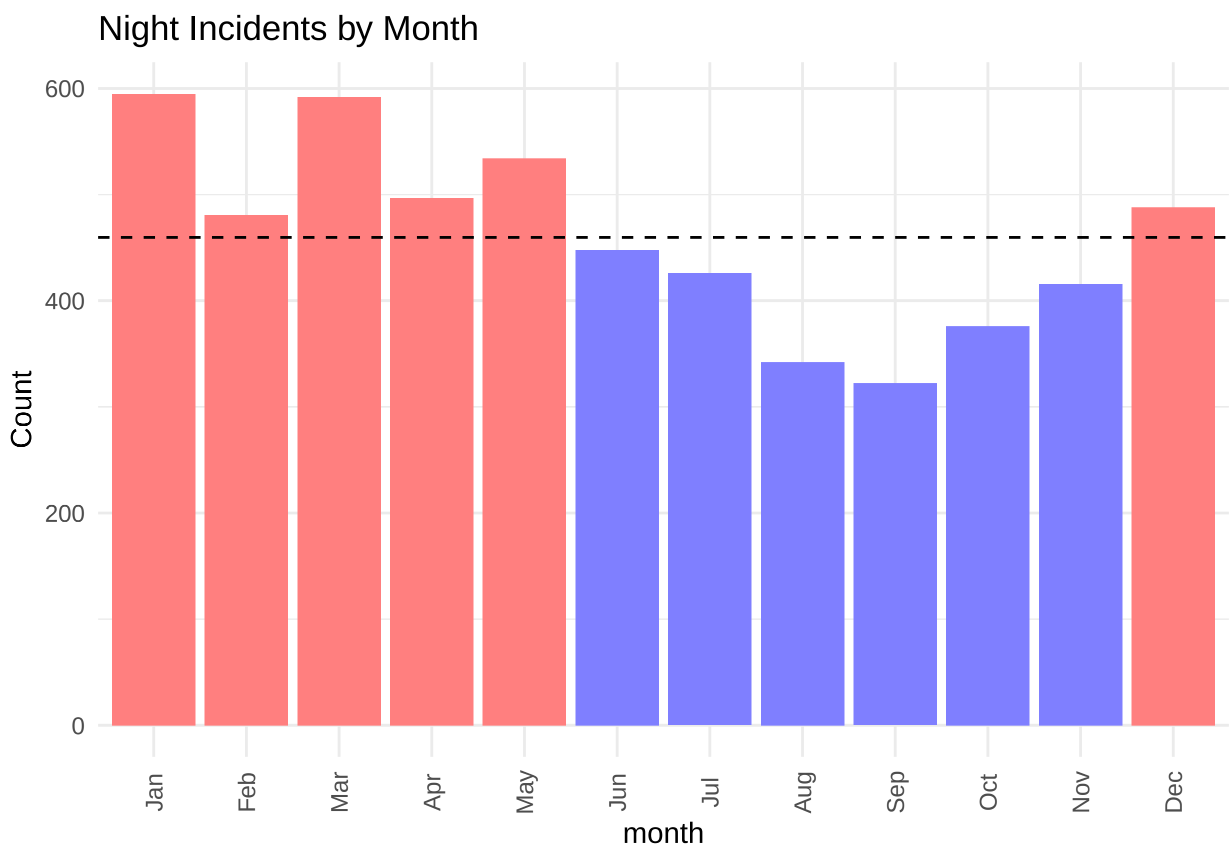
```r
df %>% group_by(night_time, season) %>% count()  %>%
            ggplot(aes(x = season, y = n, fill = night_time)) + geom_col() + facet_grid(~ night_time) +
            scale_discrete_manual(aesthetics = c("fill"), values = c("#FF7F7F", "#7F7FFF")) + labs(titl
```



We see a decrease in night time incidents during summer time. perhaps this suggests that more incidents occur when more people are using TriMet services due to weather. For example, people seeking shelter in max trains due to cold and wet conditions.

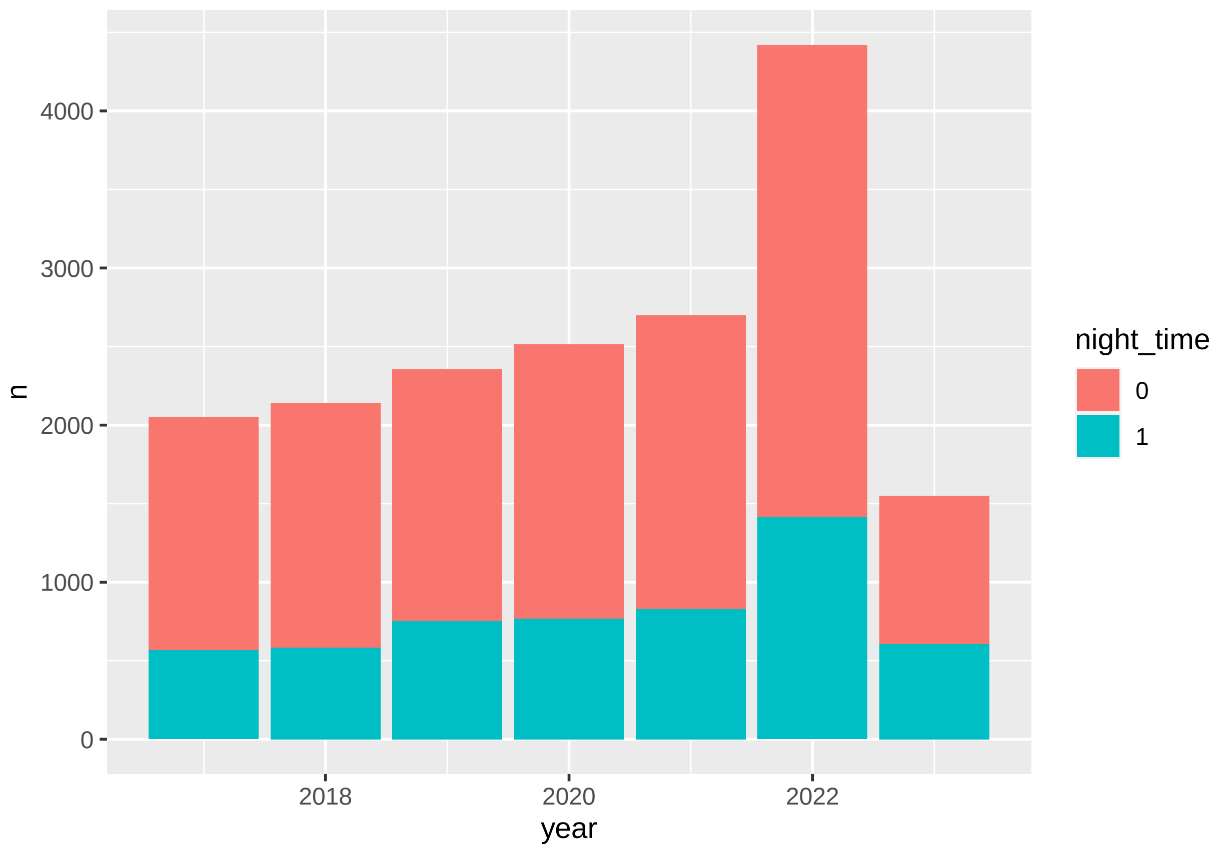
What is spread of night incidents in one year?

```
ggplot(df %>% filter(night_time == 1) %>%
  count(month) %>%
  mutate(avg = mean(n),
         color = ifelse(n > avg, "Above Average", "Below Average")), aes(x = month, y = n, fill = color)
        geom_bar(stat = "identity", show.legend = FALSE) +
        geom_hline(aes(yintercept = avg), linetype = "dashed", color = "black") +
        scale_fill_manual(values = c("Above Average" = "#FF7F7F", "Below Average" = "#7F7FFF")) +
        theme_minimal() +
        theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
        labs(y = "Count", fill = "") + labs(title = "Night Incidents by Month")
```



In summer and fall we see a decrease in incidents at night.

```
df %>% group_by(year, night_time) %>% count() %>% ggplot(aes(x = year, y = n, fill = night_time)) + geom
```



The number of incidents at night is proportional to total number of incidents per year.

Where are the most incidents occurring at night?

```
df %>% filter(night_time == 1) %>%
  group_by(location, type_desc ) %>%
    count() %>%
      arrange(desc(n))
```

```
## # A tibble: 164 x 3
## # Groups:   location, type_desc [164]
##    location             type_desc     n
##    <chr>                <chr>     <int>
##  1 Gateway Tc           Security    380
##  2 Elmonica/Sw 170th    Security    351
##  3 Cleveland Avenue     Security    294
##  4 Ruby Jct/197th Ave   Security    282
##  5 Rose Quarter Tc      Security    211
##  6 Willow Crk/185th Tc  Security    154
##  7 Hollywood/42nd Ave   Security    139
##  8 <NA>                 Security    136
##  9 82nd Avenue          Security    116
## 10 Beaverton Tc         Security    105
## # i 154 more rows
```

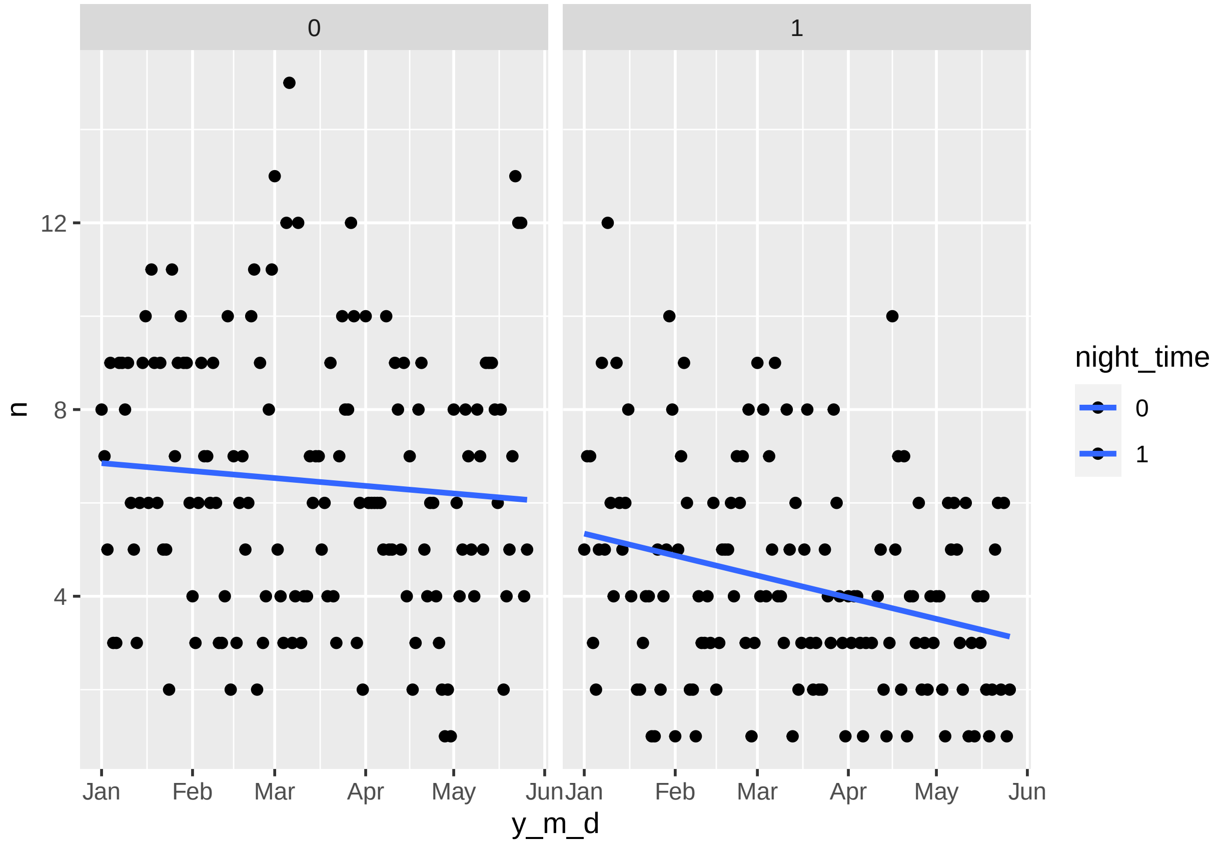
Elmonica is where the MAX trains are stored and serviced - most trains in the morning start here. It would make sense that security personnel are reporting from here. Gateway is a hot spot for all types of activity.

`Cleveland Avenue` we assume is in reference to MAX station in Gresham (final stop for the Blue Line) Cleveland is mentioned as a place where 'sleepers' are found.

In 2023, have we seen a downward trend in incidents at night given the increased presence of security personnel starting in March 2023?

```
df = df %>% mutate(y_m_d = date(as.POSIXct(incident_date, format="%m/%d/%Y %H:%M"))) # adding date only

df %>% filter(year == 2023) %>% group_by(y_m_d, night_time) %>% count() %>%  ggplot(aes(x = y_m_d, y = n
  geom_smooth(method = 'lm', se = FALSE) +
  facet_grid(~night_time)
```

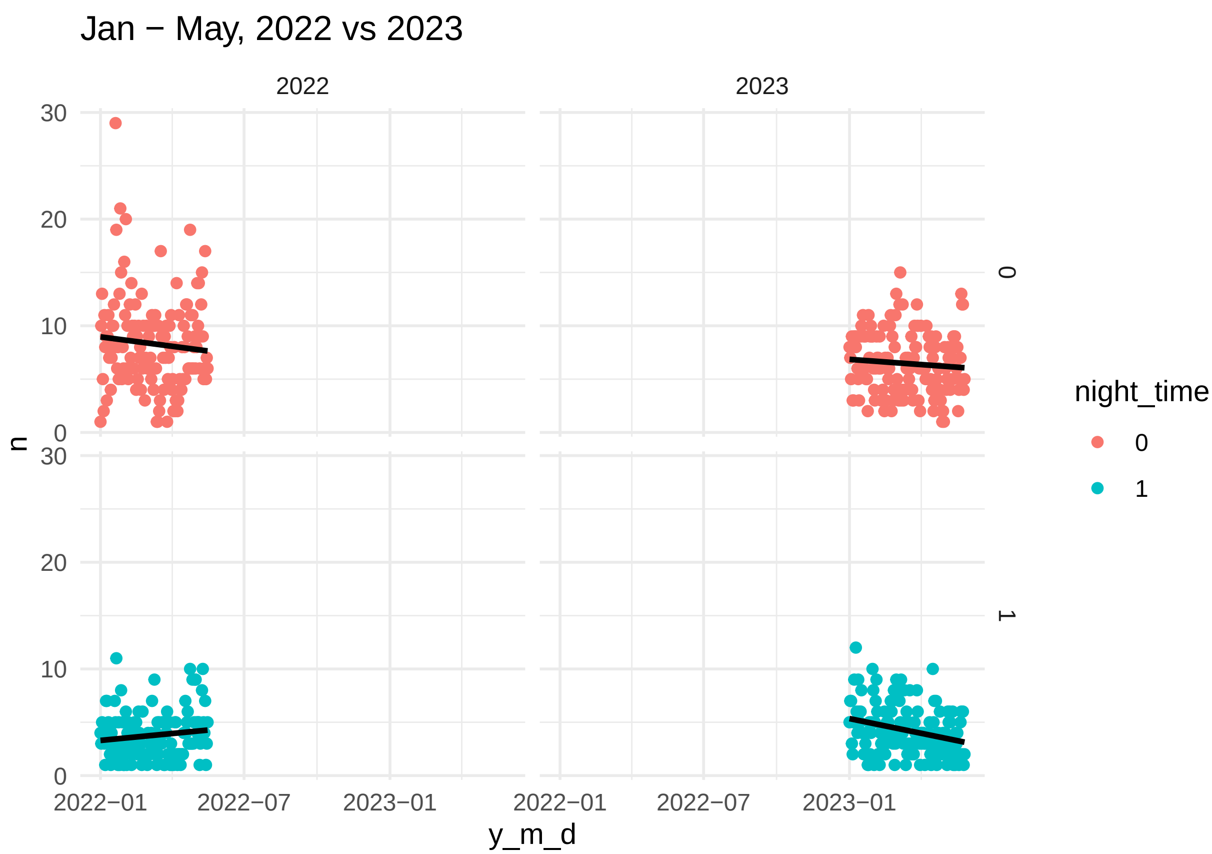
```
## `geom_smooth()` using formula = 'y ~ x'
```



What if we compare the same time interval, one year ago?

```
df %>% filter(year %in% c(2022,2023))%>%
  filter((y_m_d >= "2022-01-01" & y_m_d <= "2022-05-16") | y_m_d >= "2023-01-01")%>%
  group_by(y_m_d,year, night_time) %>% count() %>%  ggplot(aes(x = y_m_d, y = n, color = night_time))  +
  geom_smooth(method = 'lm', se = FALSE, color = "black") +
  facet_grid(night_time~year) +
  theme_minimal() + labs(title = "Jan - May, 2022 vs 2023")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Nighttime incidents were trending upward in Jan - May 2022, and now they are trending downward Jan - May 2023.