VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

RLAB PROJECT

MATHS LAB

NAME Mainak Chattopadhyay
SLOT L21+L22
COURSE NAME BMAT202P
COURSE TITLE PROBABILITY AND STATISTICS LAB
FACULTY Dr. Dhivya P

TOPIC -

TAKING A CRIME REPORT DATASET WITH AREA PARAMETRE , DOING PREPROCESSING ON THE DATA ,AND PLOTTING THE CRIME RATE ON THE MAP

21bai1217-rproject

July 8, 2023

21BAI1217 MAINAK CHATTOPADHYAY

[ ]:
```
# Load required packages
library(tidyverse)
library(lubridate)

# Read in incidents dataset
incidents <- read_csv("datasets/downsample_police-department-incidents.csv")

# Read in calls dataset
calls <- read_csv("datasets/downsample_police-department-calls-for-service.csv")

print('Done!')
```

```
-- Attaching packages --------------------------------------- tidyverse 1.2.1 --
v ggplot2 3.2.1     v purrr   0.2.5
v tibble  2.1.3     v dplyr   0.7.6
v tidyr   0.8.1     v stringr 1.3.1
v readr   1.3.1     v forcats 0.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

Attaching package: 'lubridate'

The following object is masked from 'package:base':

    date

Parsed with column specification:
cols(
  IncidntNum = col_double(),
  Category = col_character(),
  Descript = col_character(),
  DayOfWeek = col_character(),
  Date = col_datetime(format = ""),
  Time = col_time(format = ""),
  PdDistrict = col_character(),
```

```
  Resolution = col_character(),
  Address = col_character(),
  X = col_double(),
  Y = col_double(),
  Location = col_character(),
  PdId = col_double()
)
Parsed with column specification:
cols(
  `Crime Id` = col_double(),
  Descript = col_character(),
  `Report Date` = col_datetime(format = ""),
  Date = col_datetime(format = ""),
  `Offense Date` = col_datetime(format = ""),
  `Call Time` = col_time(format = ""),
  `Call Date Time` = col_datetime(format = ""),
  Disposition = col_character(),
  Address = col_character(),
  City = col_character(),
  State = col_character(),
  `Agency Id` = col_double(),
  `Address Type` = col_character(),
  `Common Location` = col_character()
)

[1] "Done!"
```

[ ]:
```
# Glimpse the structure of both datasets
glimpse(incidents)
glimpse(calls)

# Aggregate the number of reported incidents by Date
daily_incidents <- incidents %>%
    count(Date, sort = TRUE) %>%
    rename(n_incidents = n)

# Aggregate the number of calls for police service by Date
daily_calls <- calls %>%
    count(Date, sort = TRUE) %>%
    rename(n_calls = n)
```

```
Observations: 84,000
Variables: 13
$ IncidntNum <dbl> 176122807, 160569314, 160362475, 160435298, 90543656, 18…
$ Category   <chr> "LARCENY/THEFT", "ASSAULT", "ROBBERY", "KIDNAPPING", "MI…
$ Descript   <chr> "GRAND THEFT FROM UNLOCKED AUTO", "BATTERY", "ROBBERY, B…
$ DayOfWeek  <chr> "Saturday", "Thursday", "Tuesday", "Friday", "Tuesday", …
$ Date       <dttm> 2017-05-13, 2016-07-14, 2016-05-03, 2016-05-27, 2009-05…
```

```
$ Time       <drtn> 10:20:00, 16:00:00, 14:19:00, 23:57:00, 07:40:00, 18:00…
$ PdDistrict <chr> "SOUTHERN", "MISSION", "NORTHERN", "SOUTHERN", "TARAVAL"…
$ Resolution <chr> "NONE", "NONE", "ARREST, BOOKED", "ARREST, BOOKED", "LOC…
$ Address    <chr> "800 Block of BRYANT ST", "MISSION ST / CESAR CHAVEZ ST"…
$ X          <dbl> -122.4034, -122.4182, -122.4299, -122.4050, -122.4612, -…
$ Y          <dbl> 37.77542, 37.74817, 37.77744, 37.78512, 37.71912, 37.806…
$ Location   <chr> "{'latitude': '37.775420706711', 'human_address': '{\"ad…
$ PdId       <dbl> 1.761228e+13, 1.605693e+13, 1.603625e+13, 1.604353e+13, …
Observations: 100,000
Variables: 14
$ `Crime Id`        <dbl> 163003307, 180870423, 173510362, 163272811, 17281…
$ Descript          <chr> "Bicyclist", "586", "Suspicious Person", "911 Dro…
$ `Report Date`     <dttm> 2016-10-26, 2018-03-28, 2017-12-17, 2016-11-22, …
$ Date              <dttm> 2016-10-26, 2018-03-28, 2017-12-17, 2016-11-22, …
$ `Offense Date`    <dttm> 2016-10-26, 2018-03-28, 2017-12-17, 2016-11-22, …
$ `Call Time`       <drtn> 17:47:00, 05:49:00, 03:00:00, 17:39:00, 08:54:00…
$ `Call Date Time`  <dttm> 2016-10-26 17:47:00, 2018-03-28 05:49:00, 2017-1…
$ Disposition       <chr> "GOA", "HAN", "ADV", "NOM", "GOA", "ADV", "REP", …
$ Address           <chr> "The Embarcadero Nor/kearny St", "Ingalls St/van …
$ City              <chr> "San Francisco", "San Francisco", "San Francisco"…
$ State             <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "…
$ `Agency Id`       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ `Address Type`    <chr> "Intersection", "Intersection", "Intersection", "…
$ `Common Location` <chr> NA, NA, NA, NA, NA, NA, "Midori Hotel Sro #612, S…
```

```
Warning message:
"The `printer` argument is deprecated as of rlang 0.3.0.
This warning is displayed once per session."Warning message:
"`lang()` is deprecated as of rlang 0.2.0.
Please use `call2()` instead.
This warning is displayed once per session."
```

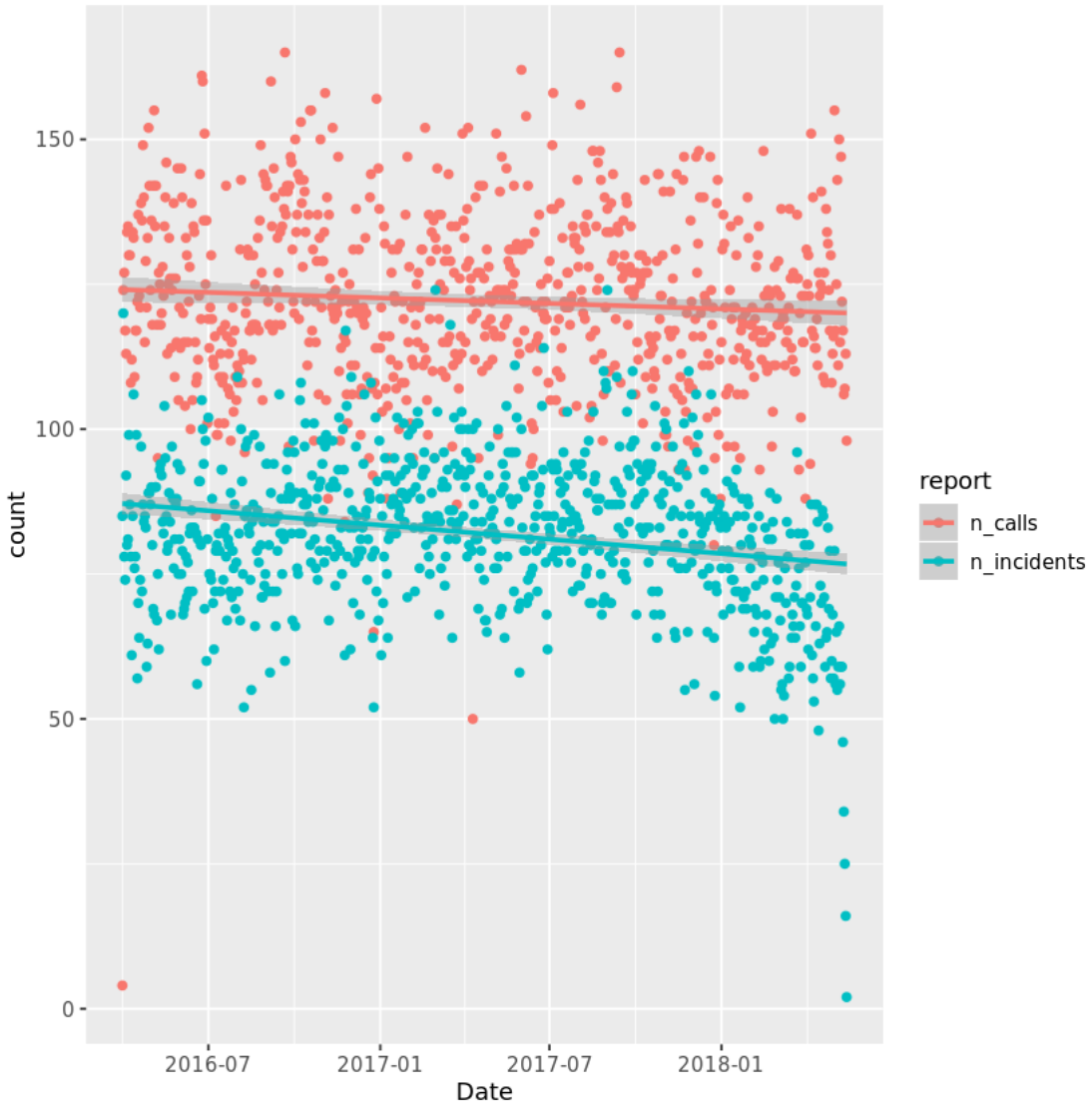
[ ]:
```
# Join data frames to create a new "mutated" set of information
shared_dates <- daily_calls%>%
                inner_join(daily_incidents, by="Date")

# Take a glimpse of this new data frame
glimpse(shared_dates)
```

```
Warning message:
"`chr_along()` is deprecated as of rlang 0.2.0.
This warning is displayed once per session."
```

```
Observations: 776
Variables: 3
$ Date        <dttm> 2016-09-21, 2017-09-14, 2017-06-01, 2016-06-24, 2016-0…
$ n_calls     <int> 165, 165, 162, 161, 160, 160, 159, 158, 158, 157, 156, …
$ n_incidents <int> 60, 97, 100, 105, 100, 89, 109, 97, 93, 72, 73, 68, 80,…
```

[ ]:
```r
# Gather into long format using the "Date" column to define observations
plot_shared_dates <- shared_dates %>%
  gather(key = report, value = count, -Date)

# Plot points and regression trend lines
ggplot(plot_shared_dates, aes(x = Date, y = count, color = report)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x)
```

```
Warning message:
"`new_overscope()` is deprecated as of rlang 0.2.0.
Please use `new_data_mask()` instead.
This warning is displayed once per session."Warning message:
"`overscope_eval_next()` is deprecated as of rlang 0.2.0.
Please use `eval_tidy()` with a data mask instead.
This warning is displayed once per session."
```



[ ]:
```r
# Calculate correlation coefficient between daily frequencies
daily_cor <- cor(shared_dates$n_calls, shared_dates$n_incidents)
daily_cor

# Summarize frequencies by month
correlation_df <-shared_dates %>%
  mutate(month = month(Date)) %>%
  group_by(month) %>%
  summarise(n_incidents = sum(n_incidents),
            n_calls = sum(n_calls))

# Calculate correlation coefficient between monthly frequencies
monthly_cor <- cor(correlation_df$n_incidents, correlation_df$n_calls)
```

```
monthly_cor
```

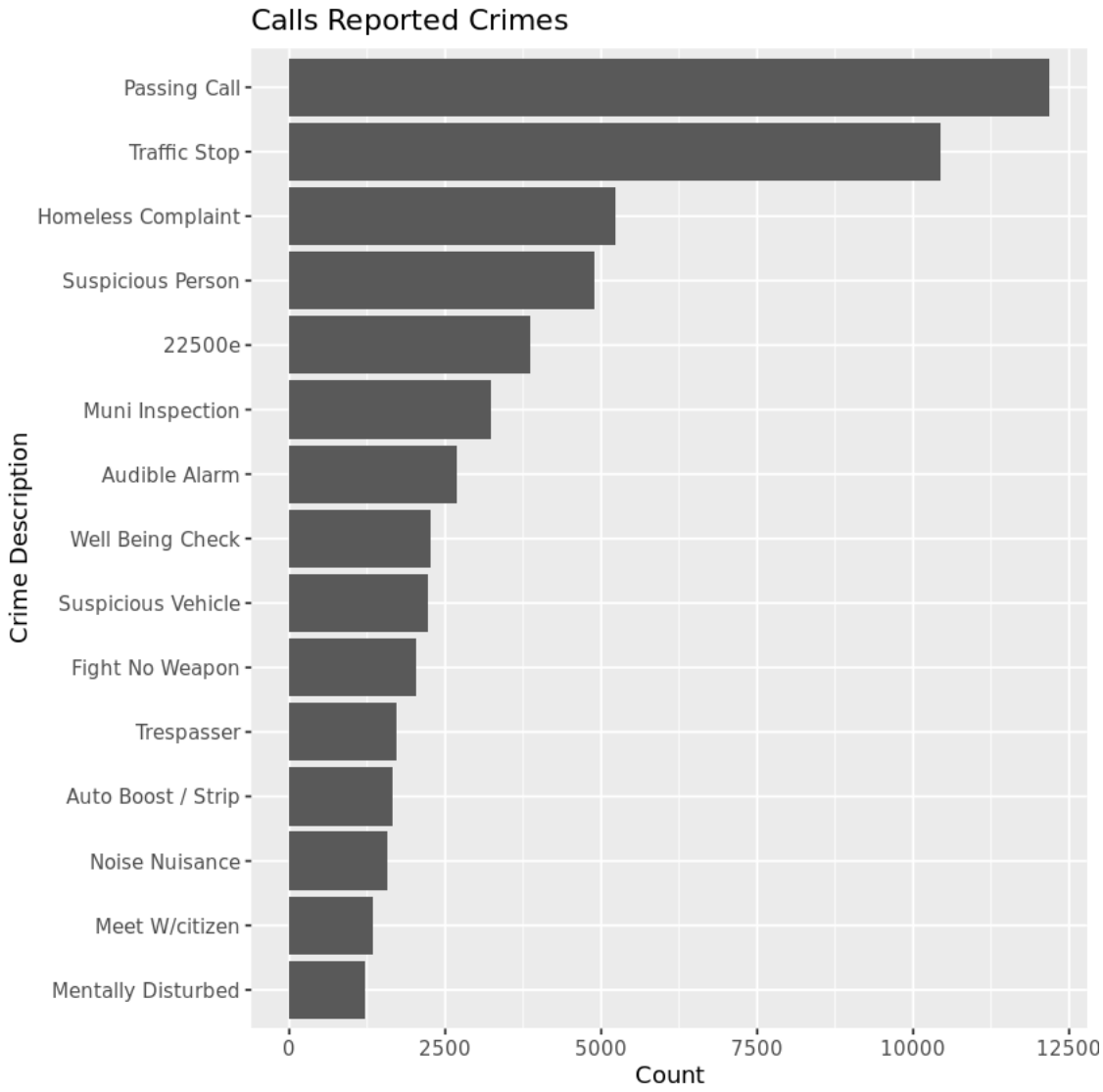
0.146968821239074

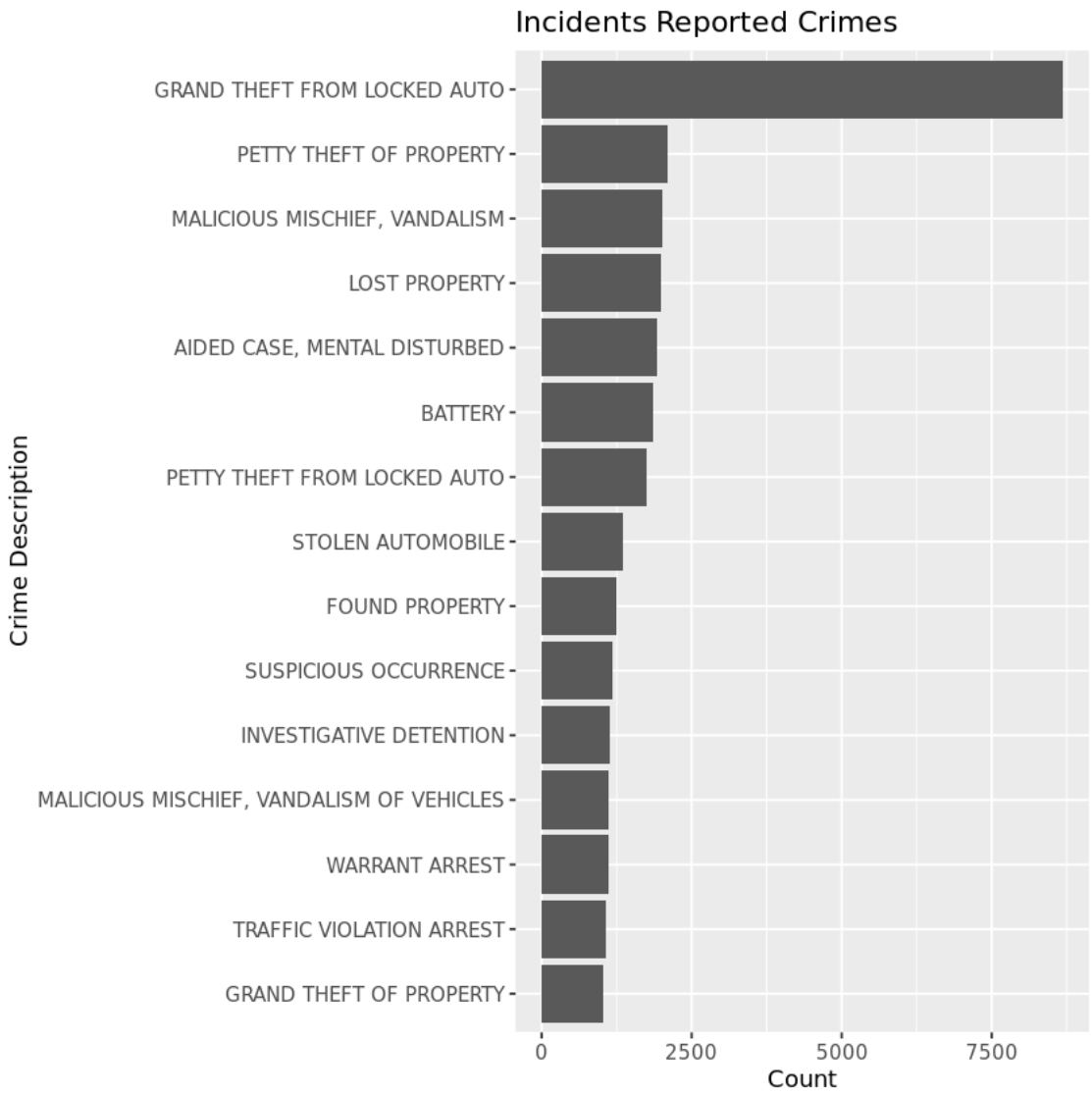
0.970682977463344

[ ]:
```
# Subset calls to police by shared_dates
calls_shared_dates <- semi_join(calls, shared_dates, by = c("Date" = "Date"))

# Perform a sanity check that we are using this filtering join function␣
 ↪appropriately
identical(sort(unique(shared_dates$Date)),␣
 ↪sort(unique(calls_shared_dates$Date)))

# Filter recorded incidents by shared_dates
incidents_shared_dates <- filter(incidents, Date %in% shared_dates$Date)
```

TRUE

[ ]:
```
# Create a bar chart of the number of calls for each crime
plot_calls_freq <- calls_shared_dates %>%
  count(Descript) %>%
  top_n(15, n) %>%
  ggplot(aes(x = reorder(Descript, n), y = n)) +
  geom_bar(stat = "identity") +
  ylab("Count") +
  xlab("Crime Description") +
  ggtitle("Calls Reported Crimes") +
  coord_flip()


# Create a bar chart of the number of reported incidents for each crime
plot_incidents_freq <- incidents_shared_dates %>%
  count(Descript) %>%
  top_n(15, n)  %>%
  ggplot(aes(x = reorder(Descript, n), y = n)) +
  geom_bar(stat = 'identity') +
  ylab("Count") +
  xlab("Crime Description") +
  ggtitle("Incidents Reported Crimes") +
  coord_flip()

# Output the plots
plot_calls_freq
plot_incidents_freq
```


Calls Reported Crimes
Passing Call
Traffic Stop
Homeless Complaint
Suspicious Person
22500e
Muni Inspection
Audible Alarm
Well Being Check
Suspicious Vehicle
Fight No Weapon
Trespasser
Auto Boost / Strip
Noise Nuisance
Meet W/citizen
Mentally Disturbed
Crime Description
0
2500
5000
7500
10000
12500
Count



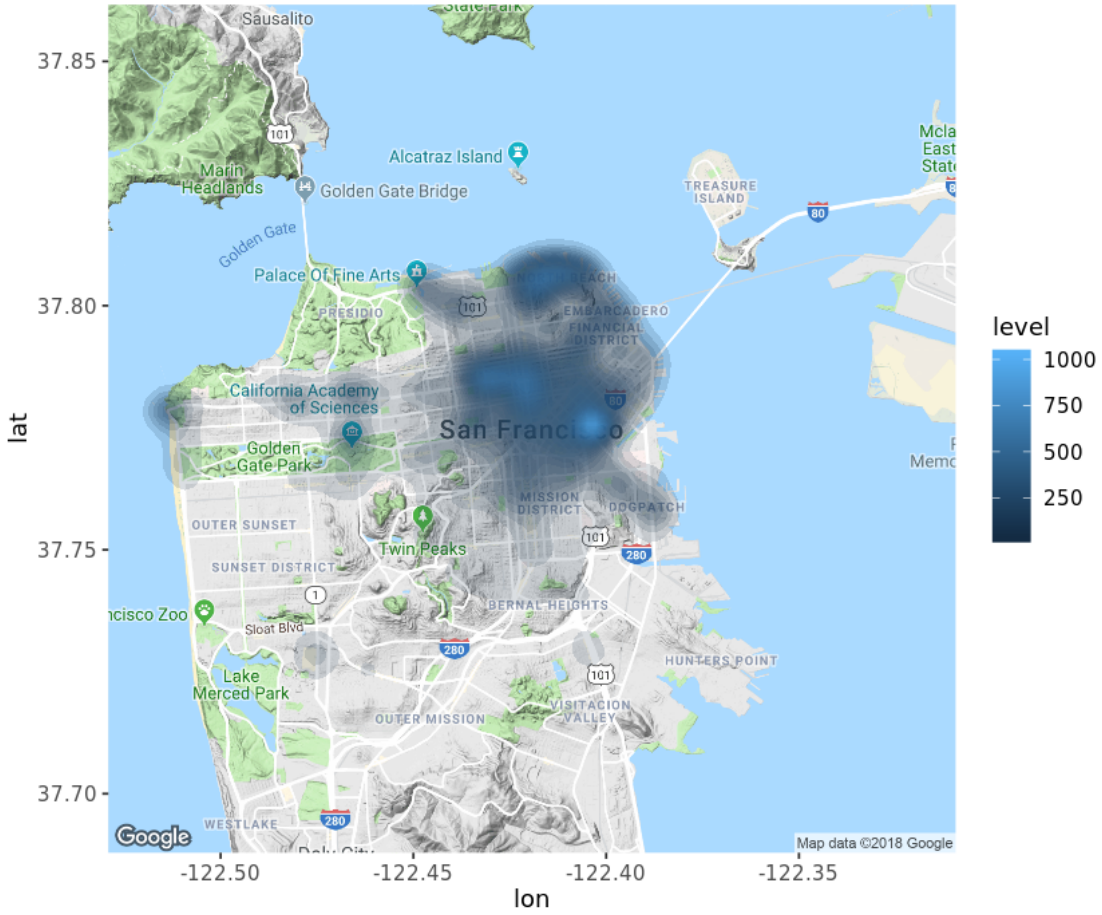
[ ]:
```r
# Arrange the top 10 locations of called in crimes in a new variable
location_calls <- calls_shared_dates %>%
  filter(Descript == "Auto Boost / Strip")%>%
  count(Address) %>%
  arrange(desc(n))%>%
  top_n(10, n)

# Arrange the top 10 locations of reported incidents in a new variable
location_incidents <- incidents_shared_dates %>%
  filter(Descript == "GRAND THEFT FROM LOCKED AUTO") %>%
  count(Address) %>%
  arrange(desc(n))%>%
  top_n(10, n)
```

```
# Print the top locations of each dataset for comparison
location_calls
location_incidents
```

Address	n
1100 Block Of Point Lobos Av	21
3600 Block Of Lyon St	20
100 Block Of Christmas Tree Point Rd	18
1300 Block Of Webster St	12
500 Block Of 6th Av	12
800 Block Of Vallejo St	10
1000 Block Of Great Hy	9
100 Block Of Hagiwara Tea Garden Dr	7
1100 Block Of Fillmore St	7
3300 Block Of 20th Av	7
800 Block Of Mission St	7

Address	n
800 Block of BRYANT ST	441
500 Block of JOHNFKENNEDY DR	89
1000 Block of POINTLOBOS AV	84
800 Block of MISSION ST	61
2600 Block of GEARY BL	38
3600 Block of LYON ST	36
1300 Block of WEBSTER ST	35
1100 Block of FILLMORE ST	34
22ND ST / ILLINOIS ST	33
400 Block of 6TH AV	30

[ ]:
```
# Load ggmap
library(ggmap)

# Read in a static map of San Francisco
sf_map <- readRDS("datasets/sf_map.RDS")

# Filter grand theft auto incidents
auto_incidents <- incidents_shared_dates %>%
    filter(Descript == "GRAND THEFT FROM LOCKED AUTO")

# Overlay a density plot of auto incidents on the map
ggmap(sf_map) +
  stat_density_2d(
    aes(x = X, y = Y, fill = ..level..), alpha = 0.15,
    size = 0.01, bins = 30, data = auto_incidents,
    geom = "polygon")
```


Sausalito
Alcatraz Island
Marin Headlands
Golden Gate Bridge
Golden Gate
Palace Of Fine Arts
PRESIDIO
TREASURE ISLAND
EMBARCADERO FINANCIAL DISTRICT
California Academy of Sciences
Golden Gate Park
San Francisco
MISSION DISTRICT
DOGPATCH
OUTER SUNSET
Twin Peaks
SUNSET DISTRICT
BERNAL HEIGHTS
Sloat Blvd
Lake Merced Park
HUNTERS POINT
VISITACION VALLEY
OUTER MISSION
WESTLAKE
Google
Map data ©2018 Google
37.85
37.80
37.75
37.70
lat
-122.50
-122.45
-122.40
-122.35
lon
level
1000
750
500
250