post02-haoyu-chen

Is Berkeley Really a Dangerous Place?

Have you ever felt that Berkeley, CA is a very dangerous place to live in? You may find some new scratches on your car or a window smashed when you return within only 5 minutes from buying a Boba. You may feel threatened all the way walking back and forth to class because the homeless, the beggars, and the robbers are everywhere on the streets. Sometimes, a bunch of emails came in in one day from University of California, Police Department as alerts to students about crimes which happened around the campus. As such, many people ask: is Berkeley really a dangerous city? How many crimes happen over there every day and what kinds of crimes are they? When and where are crimes most likely to happen in a day? Knowing these, we will take the first step to better understand our surroundings and to be more able to protect ourselves. But how do we obtain those information? Luckily, we have a powerful data analysis tool, RStudio, to help us figure everything out. In this post, I will do some basic data manipulation (mainly with function "ggplot") first to analyze the crime types and the times that crimes happened. I will also do some geospatial analysis, with function "ggmap", to visualize where the reported crimes occurred.

What are new here beyond the materials from class include:

a. A pie chart is used to analyze the percentage of each type of crime in the database.

b. The idea of time series is introduced when analyzing the distribution of each type of crime over a period of 24 hours.

c. Satellite map, heat map and contour heat map are used for the geospatial analysis.

1. About the Data:

The data I used for this post is from the website "City of Berkeley Open Data", which can be obtained here on https://data.cityofberkeley.info/browse?limitTo=datasets&utf8. The data is a detailed record of calls for Berkeley Police Department Service within the last 180 days, updated lastly on November 27, 2017. There are 11 columns in the data which include: CASENO: case number

```
    OFFENSE: types of crimes reported

    EVENTDT: dates that events occurred

    EVENTTM: event times

    CVLEGEND: descriptions for events

    CVDOW: day of week events occurred

    InDbDate: date that the dataset was updated on the website

    Block_Location: addresses for events on block level

    BLKADDR: plain text of street names

    City: the city where events occurred

    State: the state where events occurred
```

Also, the data has 5621 rows, which means 5621 events were reported to Berkeley Police Department within the last 180 days until November 27, 2017.

2. Data Preparation

I. The first step in data preparation includes loading the needed packages and reading in the data from Berkeley_PD.csv file.

```
library(readr)       #To read in a csv file
library(dplyr)       #To rearrange sections in pie chart
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(forcats)     #To rearrange sections in pie chart
library(ggplot2)     #To plot pie charts and bar plots
library(stringr)     #To separate the hours and minutes, as well as longtitudes and latitudes
library(ggmap)       #To generate maps
```

```
## Google Maps API Terms of Service: http://developers.google.com/maps/terms.
```

```
## Please cite ggmap if you use it: see citation("ggmap") for details.
```

```
rawdata <- read.csv("Berkeley_PD.csv", stringsAsFactors = FALSE)  #Read in data
```

II. The second step is to extract data and add new columns which will be used for the data manipulation part.

a. Separate the hours and minutes, and save the data of hours under the column event_hour.

```
event_hour <- str_sub(rawdata$EVENTTM, start = 1, end = 2)  #extract only the hour part of the times
event_hour <- as.numeric(event_hour)                #convert the characters into numeric values
rawdata$event_hour <- event_hour                #save values under a new column event_hour
```

b. Obtain longitudes and latitudes

Since the coordinates were given as a part of the entries under the column Block_Location, we just need to extract them, separate them, and save them into two new columns, namely “lon” and “lat” for longitude and latitude, respectively.

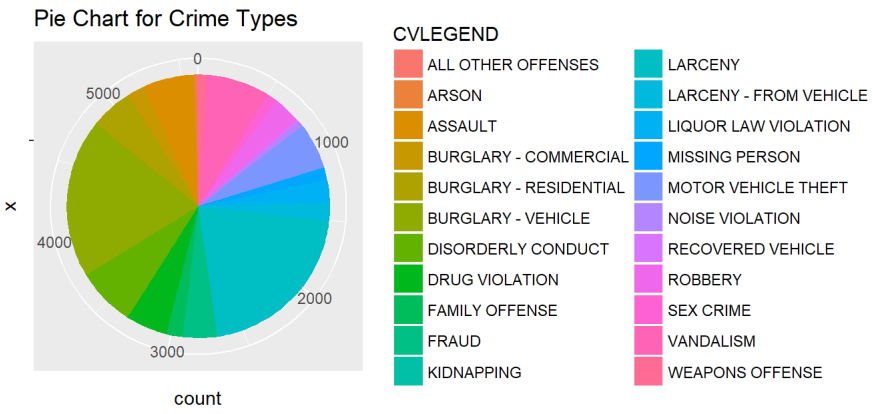
```
for(i in 1:nrow(rawdata)){                                  #for loop to reach all rows
  loc <- rawdata$Block_Location[i]                          #assign values from column Block_Location to vari
able loc
  loc <- gsub(".*\\(","",loc)                               #extract the part only after the parenthesis "("
  loc <- str_replace(loc, pattern = '\\)', replacement = '')      #remove the other parenthesis ")"
  loc <- unlist(str_split(loc, pattern = ','))              #unlist the coordinates and remove the comma
  rawdata$lon[i] <-as.numeric(loc[2])                       #assign the first part of coordinates to column l
on
  rawdata$lat[i] <- as.numeric(loc[1])                      #assign the second part of coordiates to column l
at
}
```

3. Data Manipulation and Graphics

I. Pie Chart for Crime Types

First, I want to generate a pie chart to show the percentage of each type of crime with function “ggplot”.

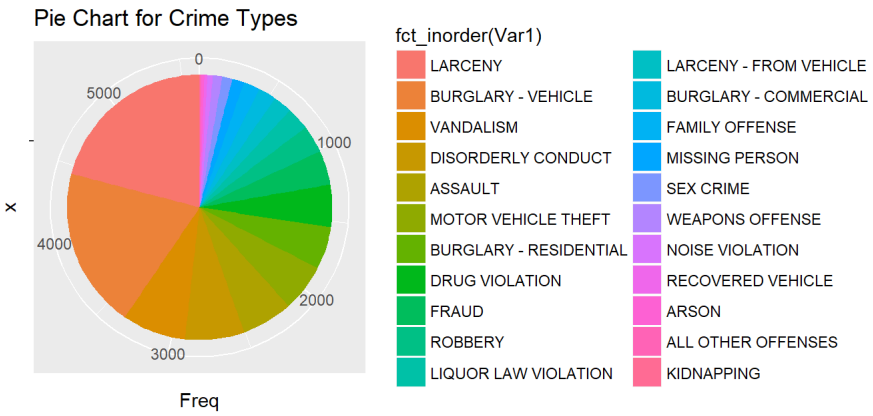
```
ggplot(rawdata, aes(x="", fill = CVLEGEND)) +   #apply function ggplot
  geom_bar(width = 1) +                         #pie char
  coord_polar(theta = "y") +                    #set y-axis to be the coordinate polar
  ggtitle("Pie Chart for Crime Types")          #set title
```



The colors are mixing up with each other! But don’t worry. I can reorder the pieces in ascending order, which means the pie chart will start with the piece with the smallest percentage clockwise.

```
crime_distr <- table(rawdata$CVLEGEND)
crime_distr <- as.data.frame(crime_distr)
crime_distr %>%
  arrange(desc(Freq)) -> crime_distr         #reorder the pieces in ascending order

ggplot(crime_distr, aes(x="", y= Freq, fill = fct_inorder(Var1))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y", start = 0) +
  ggtitle("Pie Chart for Crime Types")
```



From the pie chart, we get information as followings:

a. Larceny and burglary from vehicles rank as the top two, having occurred much more frequently than other types of crimes. These two almost take up 40% of all crimes reported within the last 180 days;

b. Vandalism, disorderly conduct, and assault rank from the 3rd to the 5th position. They are potential threats to people's life.

c. Kidnapping, arsons, and recovered vehicles rarely happened.

Some conclusions and self-protection tips:

a. Always remember to lock buildings or vehicles whenever people leave them.

b. Don't leave bags, purses, laptops, or other luxury stuff in cars when no one is sitting in the car.

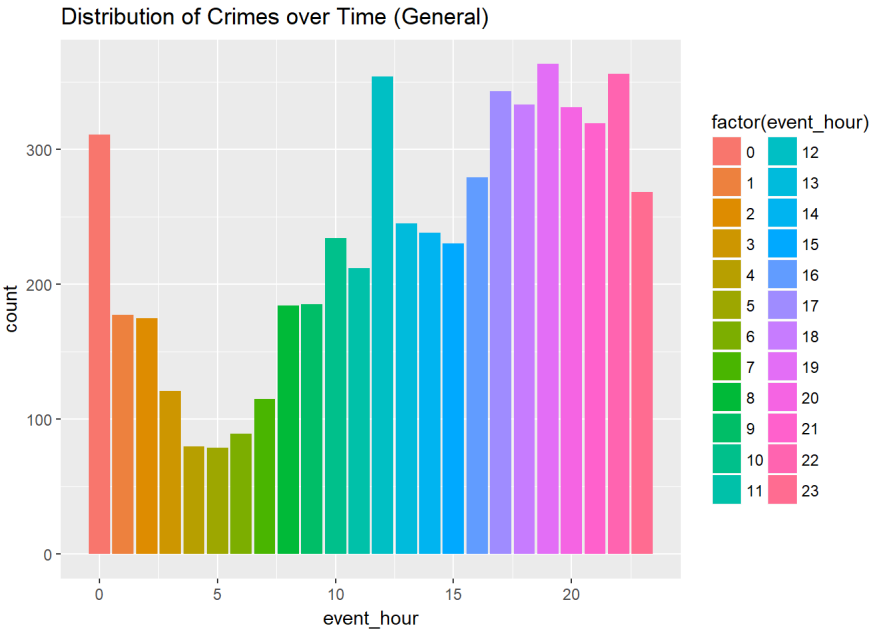
c. Park cars in public parking place, or at least in some places where people or cars will frequently pass by.

d. Avoid talking to or hanging out with people you do not know well.

II. Bar Chart for Distribution of Crimes Over Time

It is also good to know when crimes are most likely to happen so that we should keep alerted to avoid being alone by ourselves during these times. I will analyze the distribution of reported events over a time period of 24 hours. By dividing 24 hours in a day into 24 intervals, I get 1 hour for each time interval. For example, [0, 1) will include events occurred between 0:00am-1:00am and [19, 20) will include events occurred between 7:00pm-8:00pm.

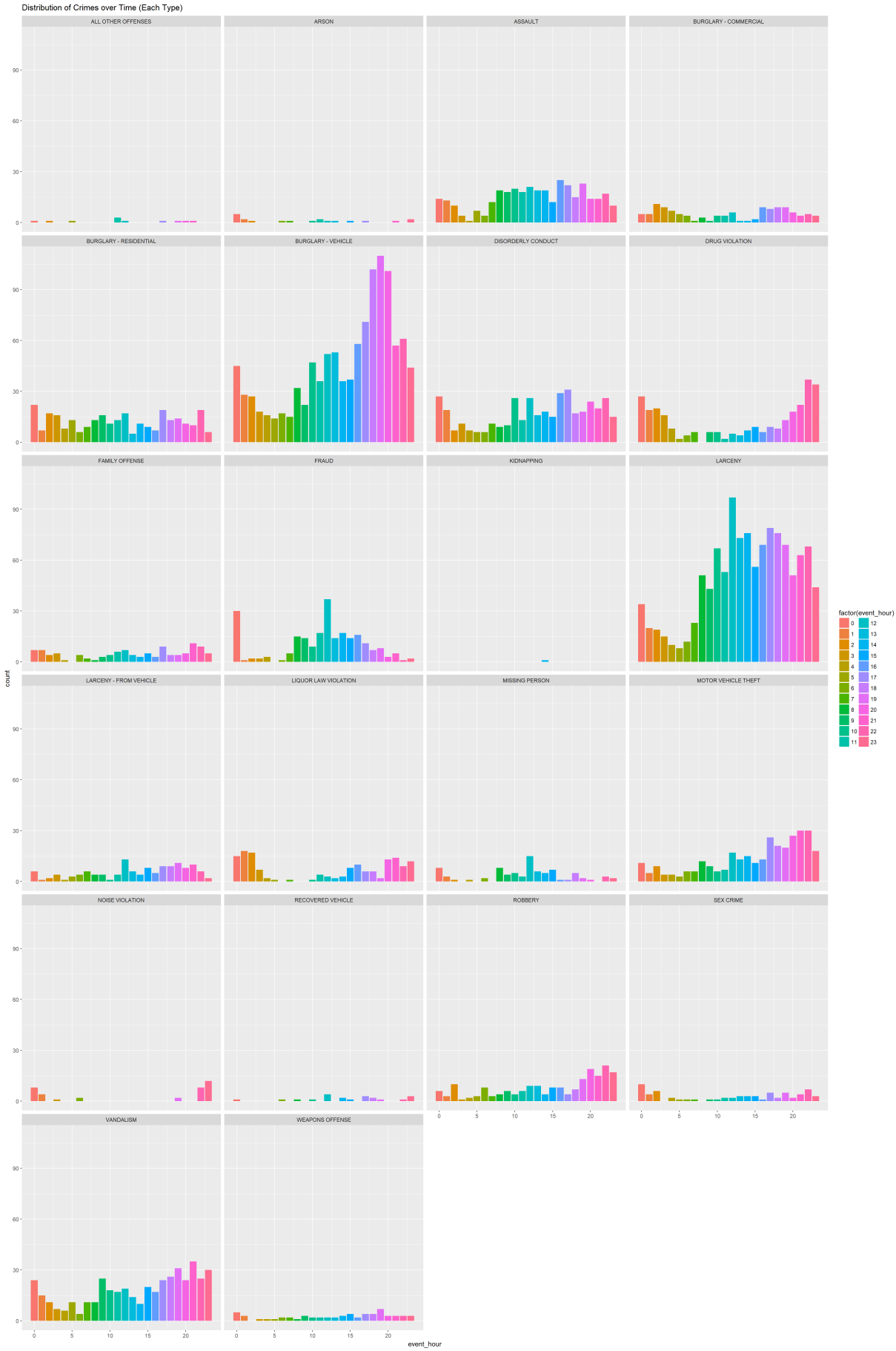
```
ggplot(rawdata, aes(x=event_hour)) +                          #bar plot ranged by hours
  ggtitle("Distribution of Crimes over Time (General)") +
  geom_bar(aes(fill = factor(event_hour)))
```



We can clearly tell a trend that after 5am the bars keep growing and reach a peak in the time interval from 17:00-22:00 and then gradually fall to the bottom at 5am. This is a general illustration that all kinds of crimes, together, are most likely to occur from evening, about 5pm (usually the time when it is getting dark) to midnight, about 12am. This makes sense because the group of crimes is dominated by larceny, burglaries, and assaults, and those kinds of events usually happen at night. Therefore, a general tip here to avoid such troubles is: go home early!

However, people may wonder: does every single type of crime maintain the same trend? Does fraud also happen mostly at night? Here I used function “facet” to display the time distribution of each kind of crime.

```
ggplot(rawdata, aes(x=event_hour)) +
  ggtitle("Distribution of Crimes over Time (Each Type)") +
  geom_bar(aes(fill = factor(event_hour))) +
  facet_wrap(~CVLEGEND, nrow = 6)          #Use facet to display each type of crime
```



The facet graphic well displayed the trend for each type of crime. The distributions of different crimes vary largely. For example, burglary from vehicle follows the general trend very nicely, with the peak at evening after 5pm. Larceny also shares a similar trend except that it reaches the

peak slightly earlier. However, fraud contrasts those two tremendously by reaching the peak at the middle of days, which makes sense because fraud can only occur when a company is running or an individual (the victim) is awake during daytime. Crimes such as burglary from commerce, family offence, or vandalism widely distributes throughout the day. Moreover, sex crime and drug violation, as expected, happen more often at late night.

III. Positions of Crimes that Occurred

After figuring out the percentage and the time distribution of each type of crimes, it is then wise to know where the crimes happened so that we can intentionally avoid those areas, especially the areas that robberies and assaults occurred most often. I will use “ggmap” to generate a map, including dots with different colors for each type of crime representing the positions where the crimes actually took place.

```
coordi <- make_bbox(lon = na.omit(rawdata$lon), lat = na.omit(rawdata$lat), f = .1)     #get cooridates data
berkeley_map <- get_map(location = coordi, maptype = "terrain", source = "google")     #terrain map
```

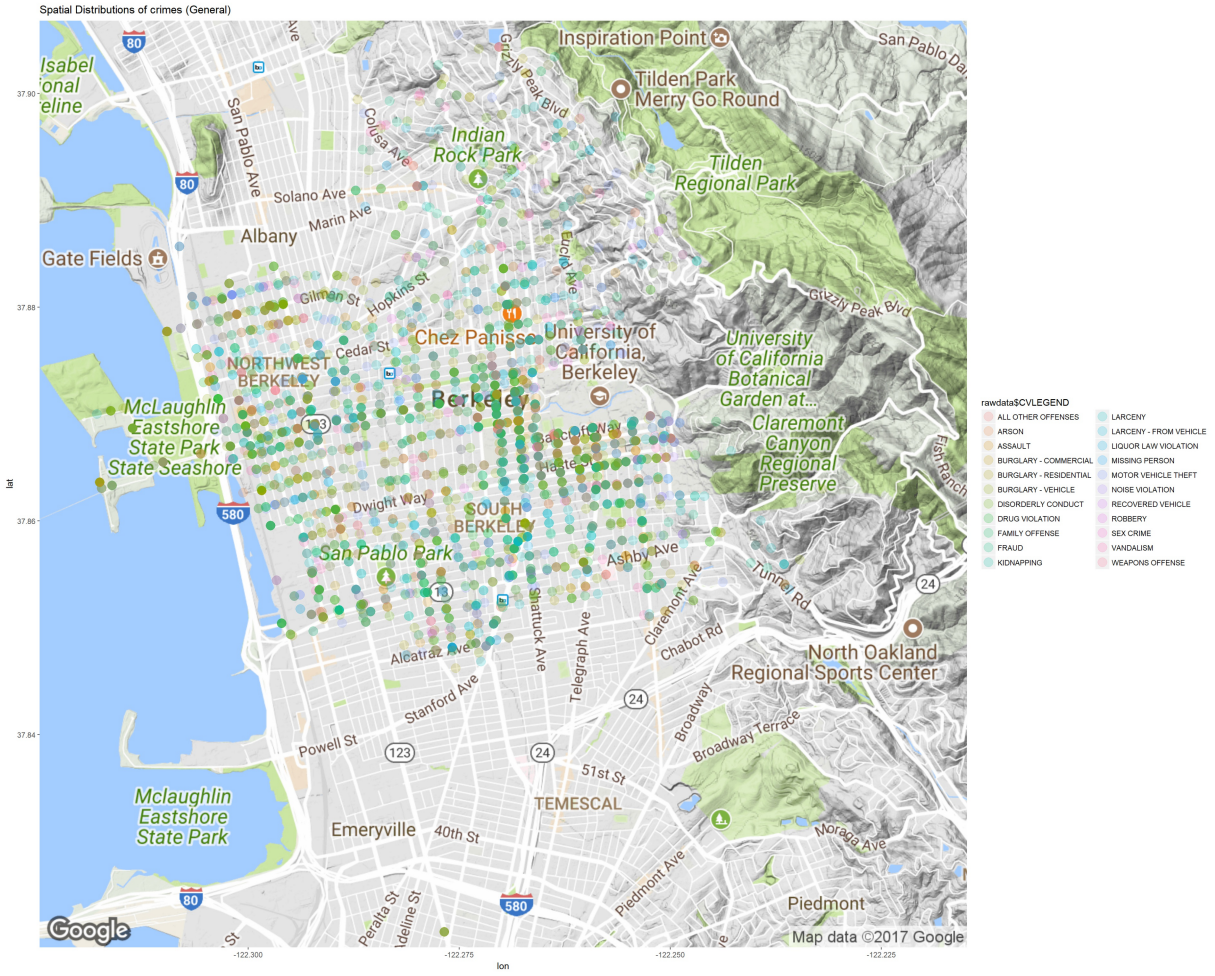
```
## Warning: bounding box given to google - spatial extent only approximate.
```

```
## converting bounding box to center/zoom specification. (experimental)
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=37.863519,-122.269847&zoom=13&size=640x640&scale
=2&maptype=terrain&language=en-EN
```

```
ggmap(berkeley_map) +
  geom_point(rawdata,
             mapping = aes(x = lon, y = lat,
             color = rawdata$CVLEGEND), alpha = 0.2,
             size = 5) +
  ggtitle("Spatial Distributions of crimes (General)")
```

```
## Warning: Removed 269 rows containing missing values (geom_point).
```



As expected, more points could be seen around the campus area than other areas, which means the surrounding areas of campus of UC Berkeley are indeed very dangerous. If you still doubt about this, a stronger evidence could be obtained by using the heat map, as shown below.

Also, this time for a better visualization, I used “satellite” for map type instead of “terrain”.

```
berkeley_map <- get_map(location = coordi, maptype = "satellite", source = "google")  #satellite map
```

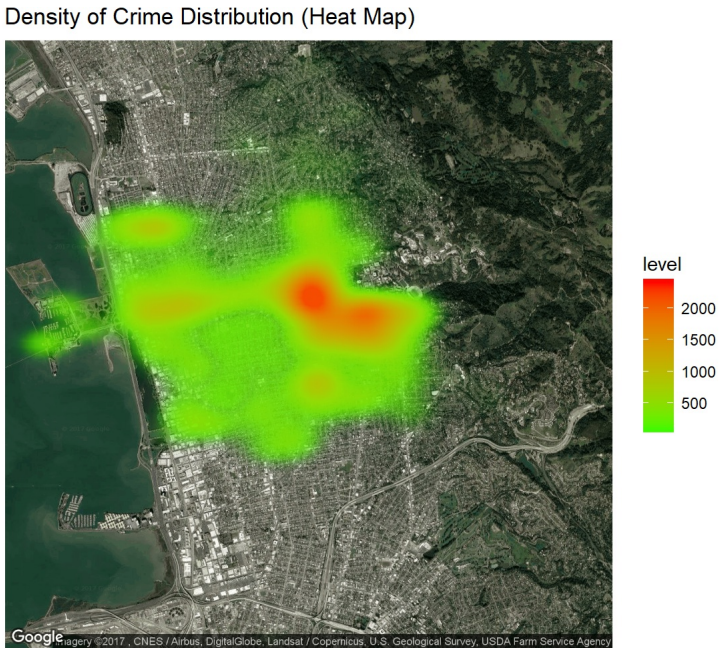
```
## Warning: bounding box given to google - spatial extent only approximate.
```

```
## converting bounding box to center/zoom specification. (experimental)
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=37.863519,-122.269847&zoom=13&size=640x640&scale
=2&maptype=satellite&language=en-EN
```

```
ggmap(berkeley_map, extent = "device") +
      stat_density2d(data = rawdata,                                          #density2d for density plot
                     aes(x = lon, y = lat, fill = ..level.., alpha = ..level..), size = 0.01,
                     bins = 500, geom = "polygon") +
      scale_fill_gradient(low = "green", high = "red") +        #color filled from green(low density) to red (high
density)
      scale_alpha(range = c(0, 0.3), guide = FALSE) +
  ggtitle("Density of Crime Distribution (Heat Map)")
```

```
## Warning: Removed 269 rows containing non-finite values (stat_density2d).
```



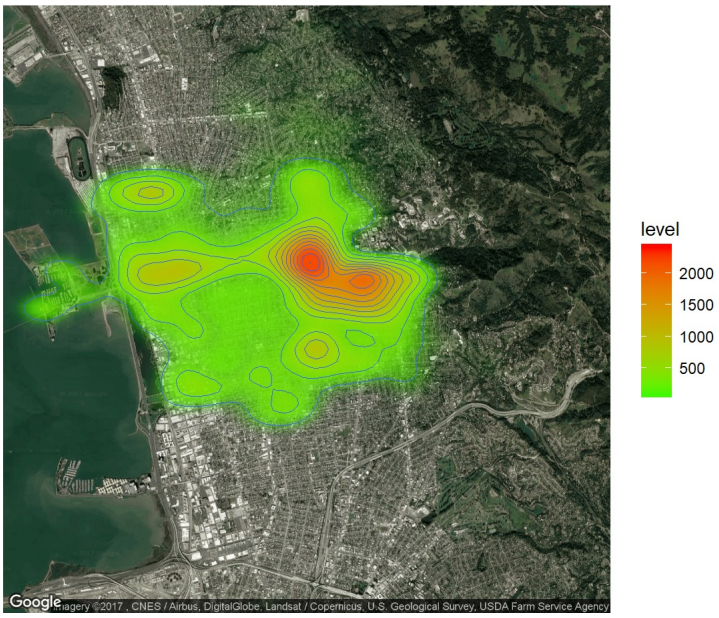
To keep up with our efforts, we can create contour heat map by including function “geom_density2d”.

```
ggmap(berkeley_map, extent = "device") +
      stat_density2d(data = rawdata,
                     aes(x = lon, y = lat, fill = ..level.., alpha = ..level..), size = 0.01,
                     bins = 500, geom = "polygon") +
      geom_density2d(data = rawdata, aes(x=lon, y = lat), size =0.3) +
      scale_fill_gradient(low = "green", high = "red") +
      scale_alpha(range = c(0, 0.3), guide = FALSE) +
  ggtitle("Density of Crime Distribution (Contour Heat Map")
```

```
## Warning: Removed 269 rows containing non-finite values (stat_density2d).

## Warning: Removed 269 rows containing non-finite values (stat_density2d).
```

Density of Crime Distribution (Contour Heat Map



However, as before, I am wondering the distribution of each type of crimes. I will make a facet graph to show that.

```
coordi <- make_bbox(lon = na.omit(rawdata$lon), lat = na.omit(rawdata$lat), f = .1)
berkeley_map <- get_map(location = coordi, maptype = "terrain", source = "google")    #terrain for map
```

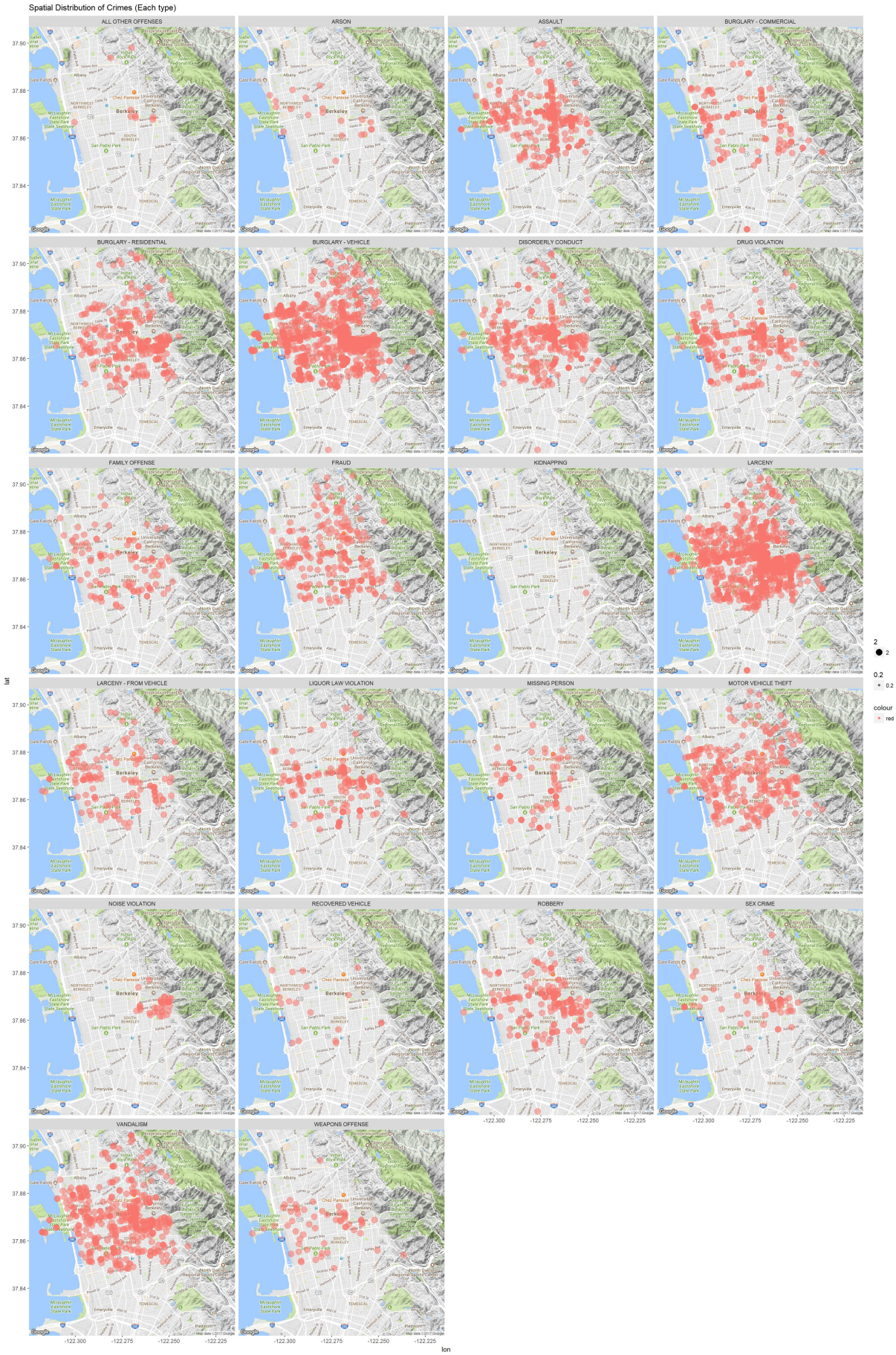
```
## Warning: bounding box given to google - spatial extent only approximate.
```

```
## converting bounding box to center/zoom specification. (experimental)
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=37.863519,-122.269847&zoom=13&size=640x640&scale
=2&maptype=terrain&language=en-EN
```

```
ggmap(berkeley_map) +
  geom_point(rawdata,
             mapping = aes(x = lon, y = lat,
             color = "red", alpha = 0.2,
             size = 2)) +
  facet_wrap(~CVLEGEND, nrow = 6) +          #facet
  ggtitle("Spatial Distribution of Crimes (Each type)")
```

```
## Warning: Removed 269 rows containing missing values (geom_point).
```



Some insights and tips:

a. Most types of crimes. including assaults, burglaries from residentials, disorderly conducts, drug violations, liquor law violations, robberies, noise violations, and sex crimes are most likely to happen on the south and west side of campus. Those places, however, are where most student apartments and houses locate. So, students should be very careful if they live in those areas, or if possible, choose apartments or houses on the north or east side of campus.

b. Crimes such as burglaries from vehicles, frauds, larcenies, motor vehicle thefts, or vandalisms are almost everywhere.

c. Arsons, recovered vehicles, weapon offenses barely happen around the school.

Conclusion: Berkeley, CA is, as the data showed us, a dangerous place to live in because 5621 events reported to the police department within the last 180 days until November, 2017. There could possibly be even more events that occurred but have not been reported to the officials. However, as the pie chart showed, most of crimes are larcenies, thefts, burglaries, frauds, or noise and liquor violations that are not actually threats to people's life. By contrast, no murders, shootings, or suicides were reported, which is good for us students. Moreover, self-protection could be achieved by avoiding the peak times of some kinds of crimes, which was shown by the bar charts. For example, park the car at public parking places or somewhere people will pass by most often when it is getting dark. Robberies can also be greatly avoided by not being alone outside after 8pm. Furthermore, try to choose apartments or houses on the north and east side of campus, or to live far away from campus is a good idea to avoid most of crimes occur at Berkeley, as shown in the maps.

In this post, I first showed how a pie chart could be plotted by using function "ggplot", which was not covered in homework or lectures so far. Also, I introduced the idea of time series where the distribution of data are shown in the order of time. When I reordered the pieces in the pie chart, a new package "forcats" and a new function "fct_inorder" were introduced. It is also worth to note that except for the basic terrain map covered in the previous homework, satellite plot, heat plot, and contour heat plot are also available by function "ggmap". Lastly, the posts show that when multiple categories are included in one data set, it is always a good way to do facet in order to show the feature of each category by showing each of them individually.

References:

1. http://www.sthda.com/english/wiki/ggplot2-pie-chart-quick-start-guide-r-software-and-data-visualization
2. https://stackoverflow.com/questions/38592691/ggmap-heatmap-with-value
3. https://yihui.name/knitr/options/
4. http://data-analytics.net/cep/Schedule_files/geospatial.html
5. http://www.geo.ut.ee/aasa/LOOM02331/heatmap_in_R.html
6. https://data.cityofberkeley.info/browse?limitTo=datasets&utf8
7. https://stackoverflow.com/questions/41338757/adding-percentage-labels-on-pie-chart-in-r
8. https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html
9. https://stackoverflow.com/questions/31481885/r-map-with-color-points-depending-on-the-category