Statistics 133 Final Project
Inquiry of Criminal Activity in Berkeley

Group: True Detectives
Members: Malayandi Palaniappan (Andy Palan), Yakub Akhmerov, Rushil Sheth, Avi Sen, Try Khov

Introduction

With a recent myriad occurrence of crime reported in the city of Berkeley, CA, it appears that criminal activity has become prevalent throughout the city. Although various attempts have been made to reduce criminal activities, various reports have illustrated the failure of these methods. As a result, we sought to investigate criminal activity throughout Berkeley, CA and identify methods through which residents of the city could keep themselves safe.

The primary objective of this investigation is to analyze the pattern of criminal activity in the city of Berkeley and suggests potential applications in terms of how this information can be utilised in order to improve the safety of residents in the city.

The primary motivation behind this project are the following two questions:

1. Which areas in Berkeley are the safest to live in?

2. At what particular time and day should we expect a significant number of criminal occurrences throughout Berkeley?

By answering these questions, we hope to identify particularly risky areas and times, hopefully helping residents of the city keep themselves safe.

However, we did not want to just stop there. We entertained the following through experiment: Suppose someone or a group of people possessed this data and sought to commit a crime. When would be an ideal time to commit a crime with respect to other crimes occurring at that particular time? This led to the formulation of the following third question:

3. If one were to commit a crime, what time and place would be ideal with respect to the severity of other criminal activities occurring at that particular time?

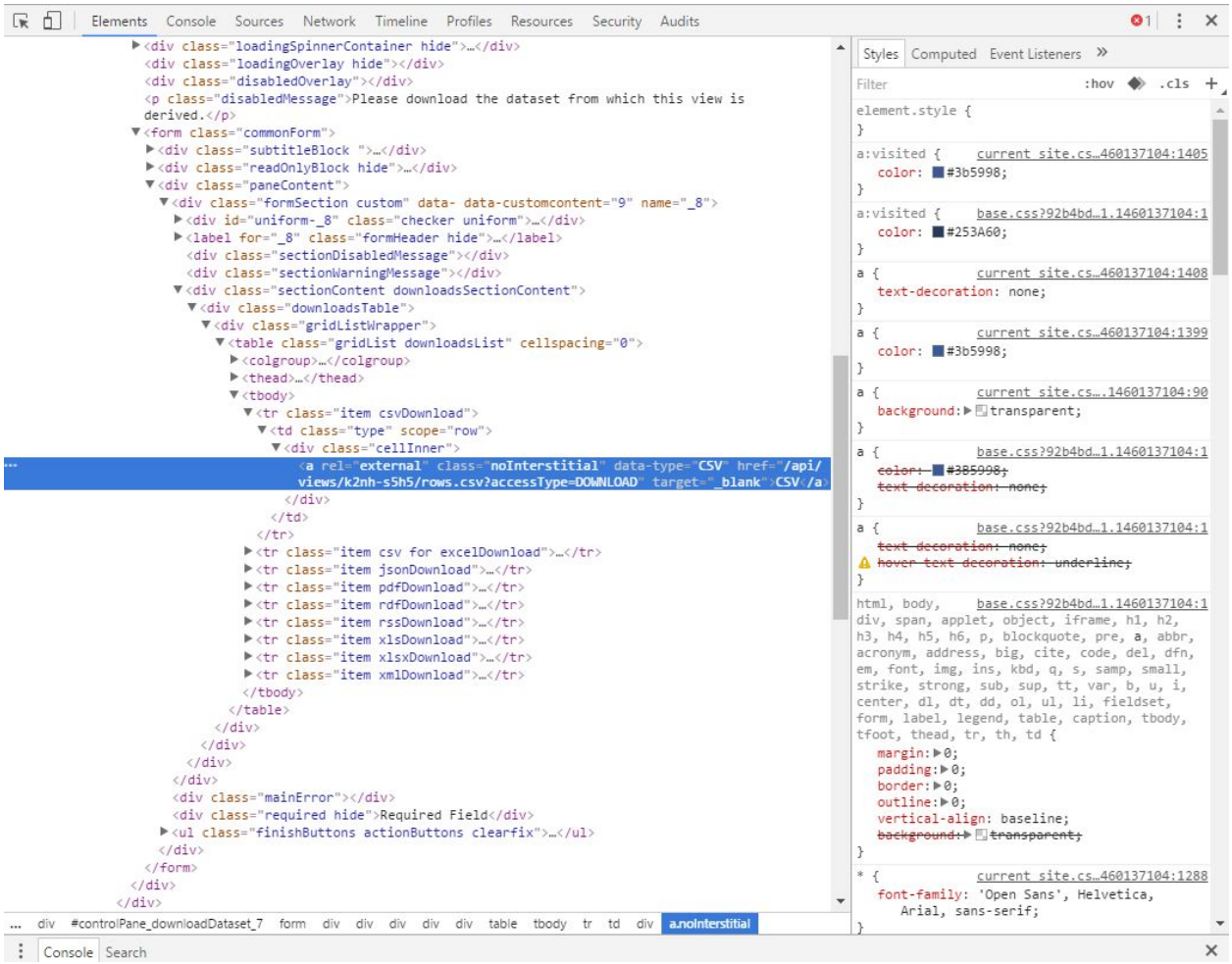
It seems ideal to commit a crime when a large number of more severe crimes are statistically likely to occur simultaneously; law enforcement would be forced to prioritize

their responses to those more severe crimes, increasing the chances that the perpetrator of these lower-severity can go free.

While this question initially began as a fun thought, it led to us some interesting observations, helping us improve our answers to the first two questions.

Data Collection

Our data source came from [City of Berkeley Open Data](#) website.



We imported the data set by reading in the .csv file from a URL within the HTML code of the City of Berkeley Crime Dataset. By reading in the data from this URL, each time our code is run, the data used in the following wrangling the plotting is the most up to date. The analysis found in this report was done using data read in on May 5, 2016 at approximately 6 PM.

Data Cleaning & Wrangling

Wrangling the Data

Upon reading in the data, the table looked as follows:

```
> head(data)
    CASENO                  OFFENSE               EVENTDT EVENTTM               CVLEGEND CVDOW              InDbDate
1 15092300 THEFT FELONY (OVER $950) 12/12/2015 12:00:00 AM   12:30                LARCENY     6 05/07/2016 03:30:10 AM
2 16090558 THEFT MISD. (UNDER $950) 03/27/2016 12:00:00 AM   23:00                LARCENY     0 05/07/2016 03:30:08 AM
3 16024307 THEFT MISD. (UNDER $950) 04/25/2016 12:00:00 AM   12:44                LARCENY     1 05/07/2016 03:30:09 AM
4 16022266     BURGLARY RESIDENTIAL 04/16/2016 12:00:00 AM   16:52 BURGLARY - RESIDENTIAL     6 05/07/2016 03:30:09 AM
5 16017867     ASSAULT/BATTERY FEL. 03/27/2016 12:00:00 AM   20:46                ASSAULT     0 05/07/2016 03:30:08 AM
6 15075141        DOMESTIC VIOLENCE 12/27/2015 12:00:00 AM   09:39         FAMILY OFFENSE     0 05/07/2016 03:30:10 AM
                                      Block_Location           BLKADDR     City State
1 2000 KITTREDGE ST\nBerkeley, CA\n(37.868204, -122.270054) 2000 KITTREDGE ST Berkeley    CA
2    2400 DURANT AVE\nBerkeley, CA\n(37.867513, -122.26127)   2400 DURANT AVE Berkeley    CA
3    2000 OREGON ST\nBerkeley, CA\n(37.857409, -122.268473)    2000 OREGON ST Berkeley    CA
4     1600 BLAKE ST\nBerkeley, CA\n(37.861653, -122.278529)     1600 BLAKE ST Berkeley    CA
5 1500 BERKELEY WAY\nBerkeley, CA\n(37.871311, -122.281978) 1500 BERKELEY WAY Berkeley    CA
6     1400 NINTH ST\nBerkeley, CA\n(37.878274, -122.297556)     1400 NINTH ST Berkeley    CA
```

We recognized that there were a lot of variables that were not going to be of use to us. The only variables we were interested in were:

- “Offense" (the specifics of the crime committed)
- “EVENTDT” (the date on which the crime was committed)
- “EVENTTM” (the approximate time at which the crime was committed)
- “CVLEGEND” (the broad category under which the crime is categorized)
- “CVDOW” (the day of week when the crime was committed)
- “Block_Location” (the location and coordinates at which the crime was committed).

We removed the unnecessary columns by using the “select” function to select columns that we wanted (and to give them more reader-friendly names).

A negligible number of cases in the table did not have coordinates for the location of the crime and considering our goal, we filtered out such cases using “grepl” and a ReGEx pattern.

Every entry in the date variable was listed as the date of the crime and the time “12:00:00 AM” in the following form as a string: “mm/dd/yyyy 12:00:00 AM”. To make this variable usable, we trimmed the string to only contain the first 10 characters (i.e. the date characters) and then used the “mdy” function contained within the “lubridate”

package to convert this variable into a usable format. Finally, we used the function “as.Date” to store this variable as a date variable.

To make the day of week variable usable, we had to convert the day of week into a factor variable. The levels of the factor were then 0, 1, 2, … , 6 where 0 corresponded to Sunday, 1 to Monday, 2 to Tuesday, etc. To make this variable user-friendly, we renamed the levels of the variable as such, hence resulting in the levels of the Vector being Sunday, Monday, Tuesday, … , Saturday.

The time variable was similarly listed as a string in the form “hh:mm”. To make this variable useful for our needs, we needed to reformat the variable to be a decimal representation (so that time would then be a continuous variable). To do this, we first converted the time variable into a “POSIXct” variable. We then extracted the minute portion of the time variable using the “lubridate” package, divided this value by 60, and added it to the hour portion of the time variable (which was similarly extracted).

In order to extract the coordinates for the location of each individual crime, we had to wrangle the “Block_Variable” variable in the initial, raw data that was read in. The “Block_Location” variable stored data in the form “ADDRESS CITY, STATE (LATITUDE, LONGITUDE)”. Some instances of the variable were missing the latitude and longitude entirely but were filtered out in a previous data step using the “grepl” function within the “filter” function; the variable was also renamed to “location” for future ease of use.

Loading in the “stringr” package for R helped immensely for the following required steps to extract the coordinates. The “str_extract_all” function included in the “stringr” package for R takes in a string (or a vector of strings) and a regular expression. The “str_extract_all” function then returns a list of the parts of the strings which matched the regular expression. For example, if we ran “str_extract_all(c(“then”, “them”, “there”), “the”)”, R would return a list of length three, each with one element with the value “the”. In our case, a vector of strings (specifically, the “location” variable from the original extracted dataset) and a regular expression matching the format of the “LATITUDE, LONGITUDE” is found in the “location” variable. The function then returned a list with each element of the list containing one element with the string of coordinates for each crime.

Next, this list was turned into a vector using the “unlist” function so then the “strsplit” function could be used to separate the latitudes and longitudes. The “strsplit” function from base R takes in a string (or a vector of strings) and a string with which the strings should be split. The function then returns a list of vectors with two elements each per element of the list with the first element being the left side of the split string and the second element being the right side of the split string. For example, in our case, “strsplit(“LATITUDE, LONGITUDE”, “, ”)” would return a list with one element containing

a vector with two elements, the first being “LATITUDE” and the second being “LONGITUDE”.

Using the “vector” function, two vectors with the length of the number of the coordinates (found using “length” function) was created, “lat” and “long”. Initially, we used a for-loop to populate the two vectors, but decided that using a vectorized loop would increase our code’s efficiency. First, the list of coordinates were made into a vector using the “unlist” function. This was a vector such that each odd element is a latitude and each even element is a longitude. Using the fact that logical vectors used for indexing have their values recycled should the length of the index vector be less than that of the vector possessing those values, we created a simple vectorized loop to fill the “lat” and “long” vectors. “c(TRUE, FALSE)” was used as an index for the “lat” loop since the latitudes were stored in each odd index and “c(FALSE, TRUE)” was used as an index for the “long” loop since the longitudes were stored in each even index. Finally, using the “as.numeric” function, the character vectors, “lat” and “long”, were turned into numeric vectors and appended onto our data frame.

Dividing Berkeley into Four Areas

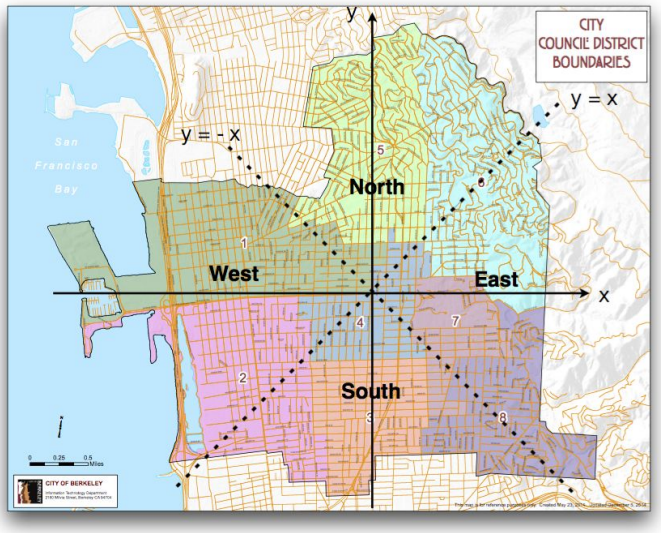
In order to determine the safest area to live in Berkeley, we first had to decide on how we wanted to group areas in Berkeley. We decided to group Berkeley into four regions – North, South, East and West – since that would provide us with a clearer picture on specific regions of the city, while still being easy to interpret for readers. Our next challenge was therefore to divide Berkeley into these four regions. However, these specific regions were not well defined, so our first challenge was to define these regions.

By using the website citylatitudelongitude.com, which contains a database of coordinates for the city center of all cities in the world, we found the official coordinates of the city center in Berkeley to be (37.8716°, -122.2727°). Using this fact, we then proceeded to normalize all the coordinates in our data set according to this new center. To execute this normalization, we subtracted every entry in the latitude column of our data set by 37.8716, calling this the new x-coordinate, and every entry in the longitude column by -122.2727, calling this the new y-coordinate. We made two new columns in the data set, titled “x” and “y”, which contained the new normalized x-coordinate and y-coordinates.

Once we did this, any crime that occurred at (37.8716, -122.2727) would have the normalized coordinates (0, 0) and any crime that occurred, for example, at (37.86820, -122.2701) – as in the first entry of our data set – would have the normalized coordinates (0.002646, -0.003396).

This made our task of dividing Berkeley into four regions much easier. We decided to split Berkeley into four diagonal quadrants at the center of the city. The top quadrant would be classified as North, the right quadrant as East, the bottom quadrant as South and the left quadrant as West, as in the diagram below.

Map of Berkeley split into N, S, E, W



We now needed to categorize each crime into one of the four quadrants (which are formed by the intersections of the Cartesian lines y = x and y = -x). We did this by deriving a system of equations (involving our normalized coordinates x and y) that would automatically determine which quadrant a specific coordinate pair fell into. The system was as follows:

North: y >= 0, -y < x <= y
South: y < 0, y <= x < -y
East: x >= 0, -x <= y < x
West: x < 0, x < y <= -x

The job of categorizing each crime in our data set into the four regions was then trivially completed by creating a new variable in our data set “area”, which would contain the area that each crime was committed in. We filled the column by using a vectorized calculation that would determine the area by comparing the x and y coordinates with the system of equations above.

Categorizing Severity of Crime

Prior to obtaining our wrangled dataset for our analysis, we were presented with numerous categories of crimes such as disturbance, homicide, arson, etc. As a result, we sought to condense these categorization by rating the labeled crimes on a scale of 1 to 5, with 5 being the most severe of those criminal acts (such as murder). This categorization allows us to analyze criminal activities in a quantitative manner. The severity ratings we decided on ar as follows:

Category 1	Category 2	Category 3	Category 4	Category 5
* Disturbance * Fraud * Identity theft * Municipal code violation * Alcohol offence * Second response	* Domestic violence * Commercial burglary * Noise disturbance * Brandishing * Missing adult * Missing juvenile * Narcotics * Misdemeanor theft	* Misdemeanor sexual assault * Vehicle burglary * Auto theft * Gun/Weapon * Person theft * Vice * Vandalism	* Misdemeanor battery assault * Arson * Felony theft * Robbery * SUSCIR - Robbery * Residential burglary * Vehicle Stolen	* Homicide * Felony sexual assault * Felony battery assault * Kidnapping

Once we added those severity ratings into our table, we were done with the cleaning and wrangling of the data. Our clean data set looked as follows:

```
> head(dc)
        date      day     time                 type                  details
1 2015-12-12 Saturday 12.50000              LARCENY THEFT FELONY (OVER $950)
2 2016-03-27   Sunday 23.00000              LARCENY THEFT MISD. (UNDER $950)
3 2016-04-25   Monday 12.73333              LARCENY THEFT MISD. (UNDER $950)
4 2016-04-16 Saturday 16.86667 BURGLARY - RESIDENTIAL     BURGLARY RESIDENTIAL
5 2016-03-27   Sunday 20.76667                ASSAULT     ASSAULT/BATTERY FEL.
6 2015-12-27   Sunday  9.65000         FAMILY OFFENSE        DOMESTIC VIOLENCE
       lat      long    wday severity         x         y area
1 37.86820 -122.2701 weekend        5  0.002646 -0.003396    S
2 37.86751 -122.2613 weekend        2  0.011430 -0.004087    E
3 37.85741 -122.2685 weekday        2  0.004227 -0.014191    S
4 37.86165 -122.2785 weekend        4 -0.005829 -0.009947    S
5 37.87131 -122.2820 weekend        5 -0.009278 -0.000289    W
6 37.87827 -122.2976 weekend        2 -0.024856  0.006674    W
```

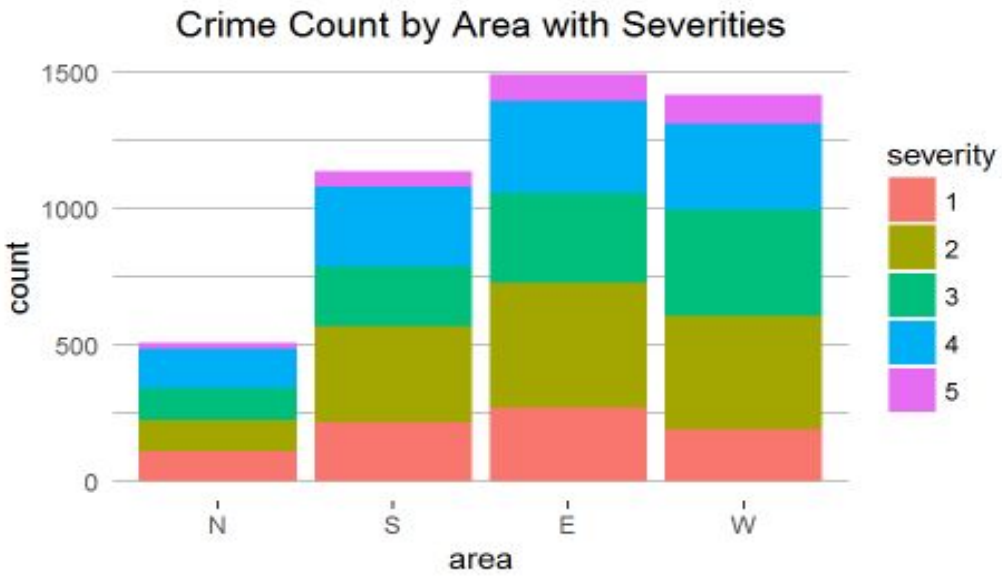
Data Analysis

Where is the safest place to live in Berkeley?

In order to investigate which areas are the safest to live in Berkeley, we created various visualisations to provide visual insight into the data. With our division of Berkeley and our categorization of the severity of crimes, we were able to analyze the data in a much more coherent manner.

We started with the question: “Where is the safest place to live in Berkeley?” We approached this question by first looking at the count of crime in each area. The following graph shows us that North Berkeley is far and away the safest area in all of Berkeley, with the count of crimes being only one half that of South Berkeley and one third that of both East and West Berkeley; the difference in count is unlikely to be caused by chance due to the extreme variation.

South Berkeley is the next safest area on the account of the count, while East and West Berkeley are the two most dangerous areas. While the count of crime is higher in East Berkeley, we should keep in mind that the difference in the count of crime between East and West Berkeley is marginal (less than 10%) and since the data we have only represents six months of data, we cannot rule that East Berkeley is significantly more dangerous; it is completely possible that this difference is caused by chance. However, what is evident is that in terms of the sheer count of crimes, East and West Berkeley are much more dangerous than both North and South Berkeley.



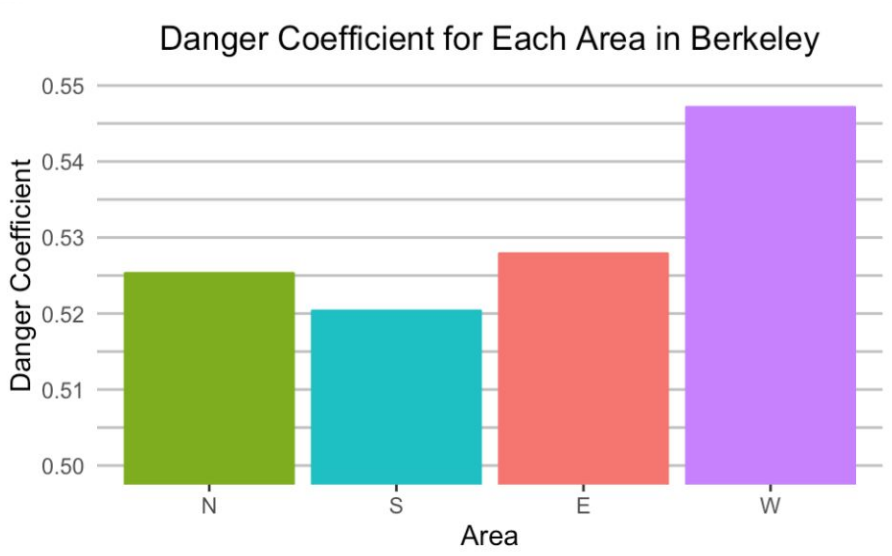
However, looking at the count of crime alone as a representative of how dangerous a region is, is a very naive approach. We need to keep in mind that different areas may be prone to different types of crime (which may have varying severity). Hence, it is completely possible that certain areas may be more prone to more severe crime – a fact we need to take into account before making a judgement as to which area is the most dangerous.

To do this, we created a measure that we call the “Danger Coefficient”. The aim of the danger coefficient is to provide further insight into which area is prone to the most severe crimes. Our Danger Coefficient is the sum of all the severities divided by the total number of crimes in a particular area and then divided by 5 and multiplied by the number of crimes in that area, as in the formula below.

$$Danger\ Coefficient\ = \frac{sum\ of\ severities\ per\ area}{5 \times number\ of\ crimes\ in\ an\ area}$$

The Danger Coefficient is a measure of the severity of crimes that gives more weight to high severity crimes. Hypothetically, if five crimes of severity 1 happened in North Berkeley while two crimes of severity 5 occurred in South Berkeley and we were to compare them, the latter area would be more “dangerous” according to the Danger Coefficient.

After computing the Danger Coefficient for each area, we arrived at the following graph.



Here, we note that North, South and East Berkeley all have roughly the same danger coefficient, while the danger coefficient in West Berkeley is significantly higher than those in all other areas.

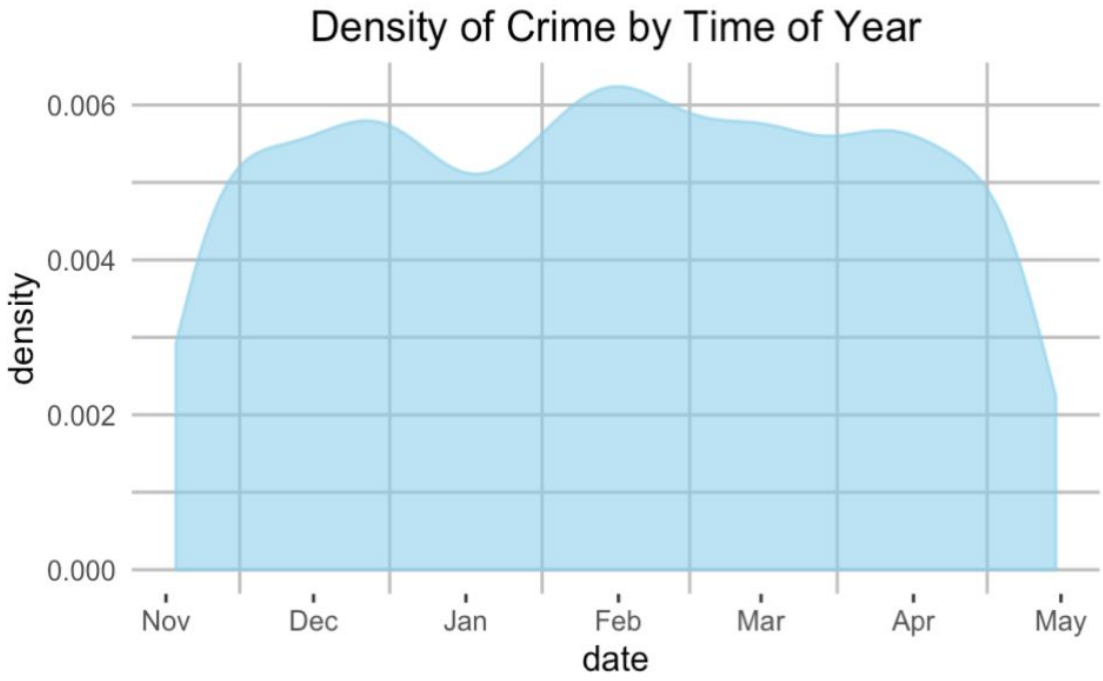
Hence, while the count of crimes in East Berkeley is marginally higher than the count of crimes in West Berkeley, it seems like West Berkeley is the more dangerous of the two areas since the difference in the count of crimes is negligible but the severity of crimes in West Berkeley is much higher than that in East Berkeley – i.e. the average crime in West Berkeley inflicts much more damage than the average crime in East Berkeley.

Hence, we make the claim that East and West Berkeley are likely to be the two most dangerous areas in Berkeley, with West Berkeley being marginally more dangerous. Both South and North Berkeley are relatively safe in comparison, while it is seemingly evident that North Berkeley is the safest of all four areas. Hence, if avoiding crime is the number one priority for residents, they should choose to live in North Berkeley.

When should we expect a significant number of criminal occurrences?

Although the Danger Coefficient and count of crimes present us with information regarding the level of severity to expect in certain areas, it doesn't provide us with information as to when one can expect a criminal activity to occur. Hence, we sought to determine the month of the year most prone to crime. Our initial hypothesis was that this month would be February since students are just getting back from break and there are a higher volume of people in Berkeley. To test our hypothesis, we analyzed the data of

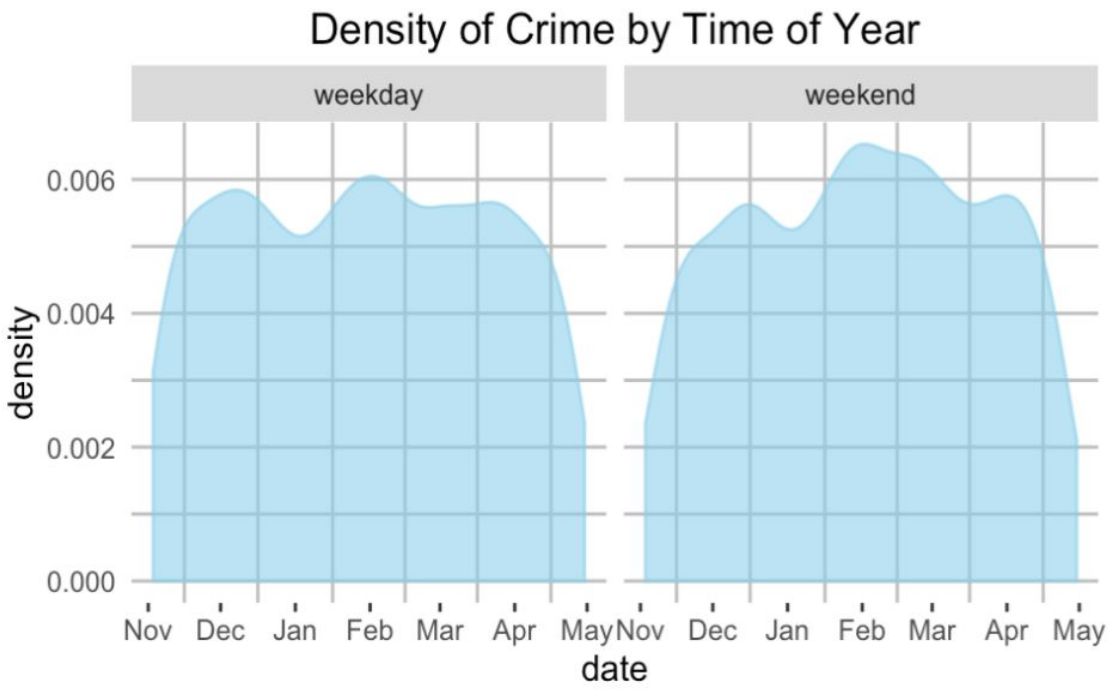
the past six months via a density graph.



As per the graph, our hypothesis appears to be right: early February does happen to have the highest number of criminal occurrence. Interestingly, there appears to be a drastic drop of criminal activities from December to February. While we were initially confused by this drop, we recognised that this drop-off in criminal activity was likely caused by the fact UC Berkeley has its winter break during that period – most of the students, who form a large part of the thriving city, would have returned home for the break, drastically reducing the number of people in and around the city. It seems safe to suggest that since there are fewer people in town, there are both fewer targets and fewer perpetrators of crime, causing the amount of criminal activity to fall.

This further suggests a plausible explanation as to why February is the month most rife with crime – this is month when students return from winter break. During the early parts of the month of February, when the schooling semester is yet to be in full throttle, students are more likely to be out and about in the town, leaving themselves vulnerable to crime. This analysis agrees with the graph since we note that as we go deeper into the semester when students become busier with school and are less likely to roam around the town, the amount of criminal activity decreases consistently.

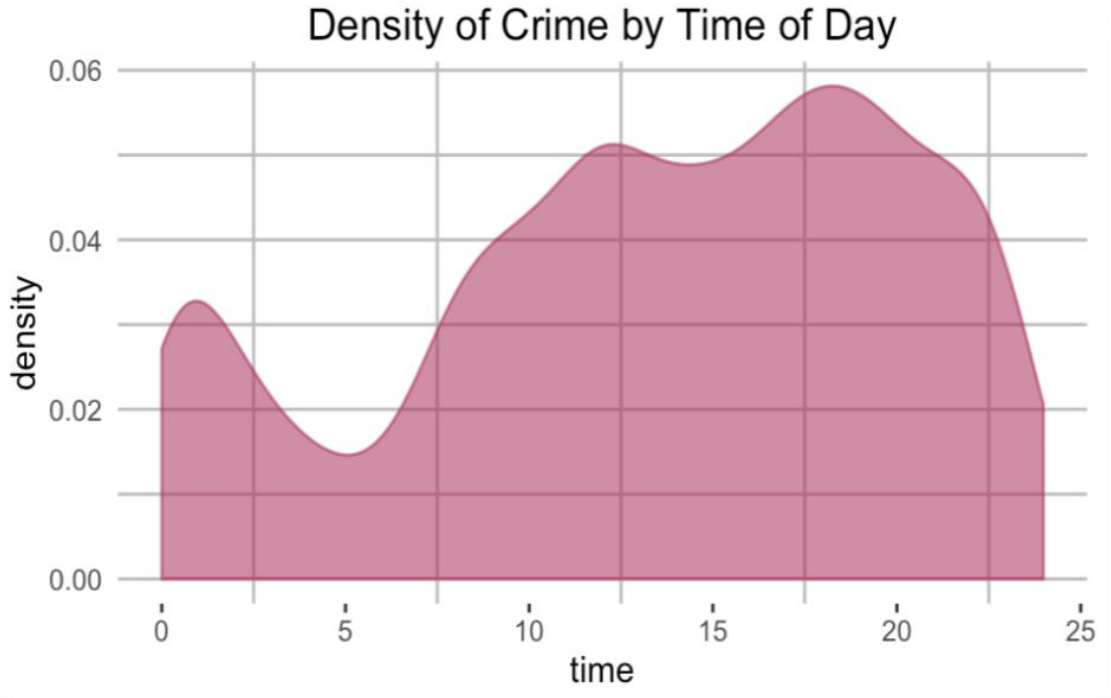
Continuing our analysis, we decided to look into the daily pattern of crime across the year.



These graphs are largely consistent with our previous analysis, showing us that throughout the year, on both weekdays and weekends, the pattern of crime is generally as expected, with a collapse in the number of crimes during the month of December and a peak in February. However, it was especially fascinating to note that the peak in February was much higher on the weekends than it was on the weekdays while the density of crimes in November (when the fall semester is at its most intense) is much lower on weekends than on weekdays.

The fact that there is a large dip in criminal activity on weekends in November, when students are most likely to be working hard and not going out (as their semester workload is at a peak), suggests that much of the criminal activity on weekends is related to students roaming about. This leads us to believe that our prior suggestion, that the peak in February is caused by an increase in the number of students roaming about, was indeed accurate since we see here that this peak is largely caused by an increase in the number of crimes during weekends when students are most likely to be out and about. This suggests to us that much of the variation in the pattern of criminal activity is driven by changes in the behaviour of student and that crime is most likely to occur when there are a large number of people out and about.

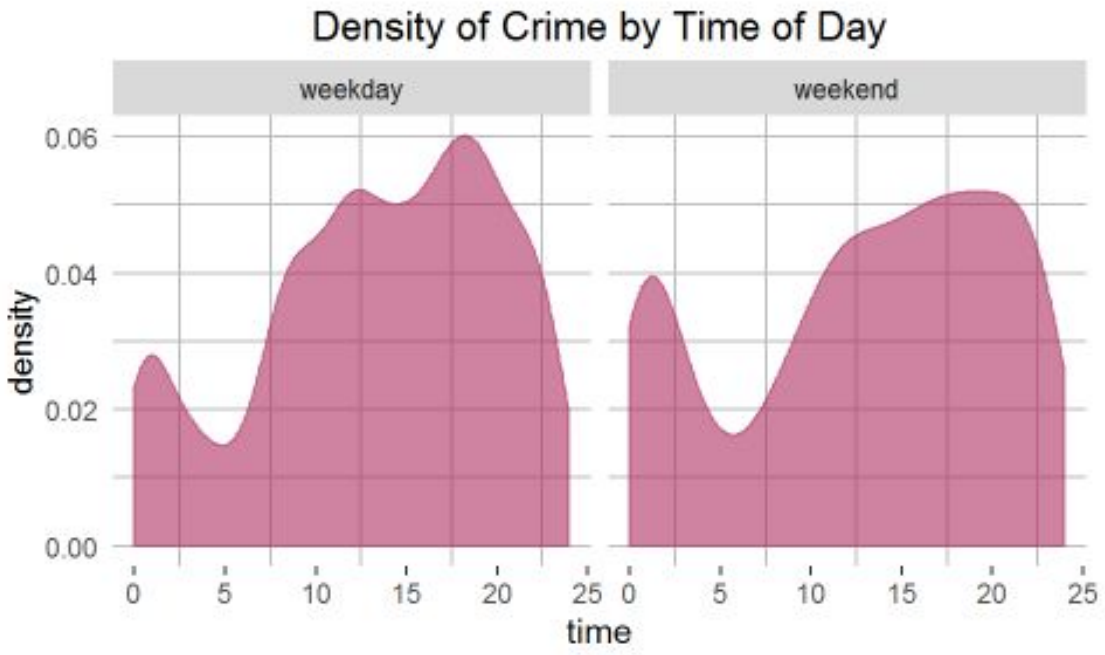
We thought to take another step deeper and see which time of the day had a higher volume of crimes. Our initial hypothesis was that 11 PM - 12 AM would have the highest density of crimes since this is when the streets would be the most desolate and the most dangerous.



First off, we noticed there were 2 large spikes with a decline in the early day. It is not surprising that 5:00 A.M has the lowest occurrence of crime, since people are either finishing up a very late night or starting a very early day. We see the first spike at around 1 - 2 PM and we see the second, larger spike at roughly 6 PM. This is fairly unintuitive and clashed with our hypothesis.

We hypothesised that one potential reason why this is the case is that that at both 1 PM – lunchtime for most people – and 6 PM – the time when most people get off work or class and get dinner – there are a large number of people out and about in Berkeley. The fact that spikes in criminal activity occur at these times of the day add further credibility to out suggestion that criminal activity is most likely when there are a large number of people out and about on the streets.

We were curious to investigate if the time of the occurrence of crime varied across weekends and weekdays. Our hypothesis was that the weekend would have a higher peak with a larger density of crimes being at its peak.



It was interesting to note that on weekdays, the two spikes at 1 PM and 6 PM were more clear. This is of no surprise since on weekdays, when people are in work and class, their lunchtime is more likely to be at 1 PM and they are more likely to get off work or class and get dinner at 6 PM, hence it is more likely that there are a large conglomeration of people at these times.

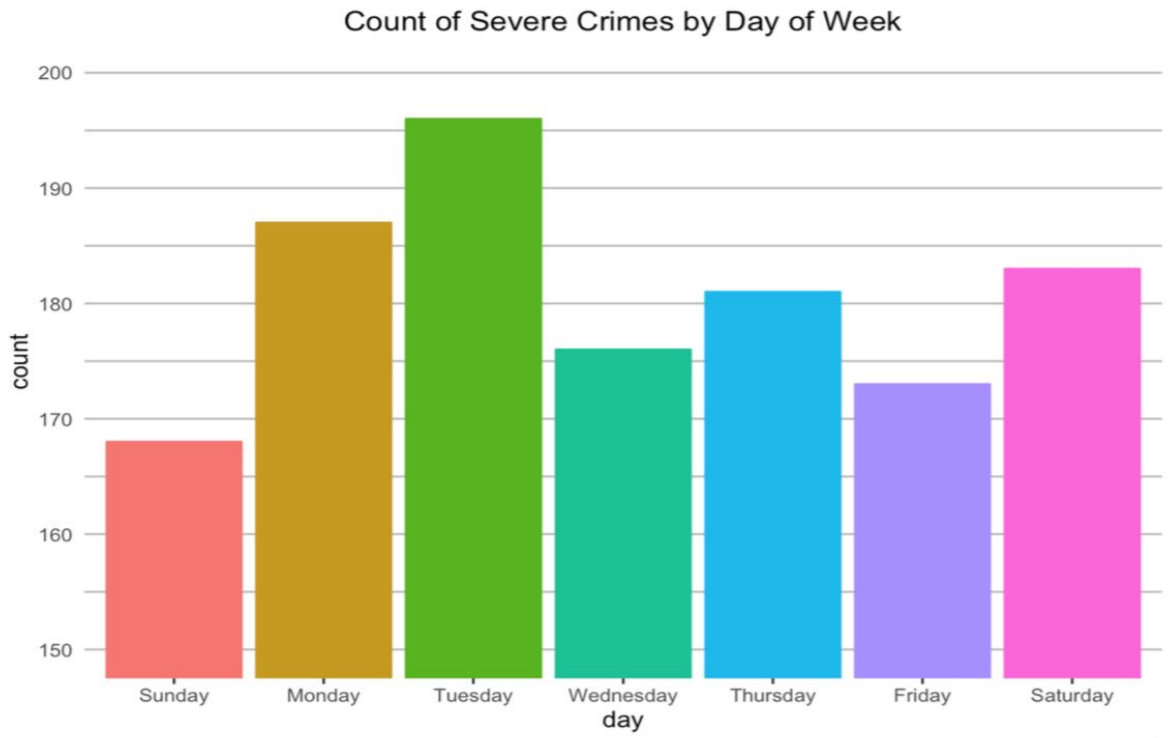
During the weekends, especially in a college town, lunch and dinner times are less defined, explaining why the peak isn't as clear as it is on weekdays, however, there still exists a peak at around 6 - 7 PM, when most people would get dinner, although this peak is distinctly shorter that the peak on weekdays. What is interesting to note is that there is a second, shorter peak on weekends at 1 - 2 AM – the density of crime at this time on weekends is much higher than it is on weekdays. This is again perhaps of no surprise, since this is the time when most people may be returning home in the dark after a night out, leaving themselves especially vulnerable to crime.

Hence, we see that the most dangerous time to be about is most likely when there are a lot of people out and about. In the town of Berkeley, the most dangerous month is February when students return to school after winter break; the most dangerous time in a week, is either lunch (1 PM) or dinner time (6 PM) on weekdays, and dinner time again (6 PM) and the early morning (2 AM) on weekends. If you wanted to avoid crime, then these would be the times when you would stay away from the streets.

If one were to commit a low level crime, when would be the best time to do it?

For the purpose of our third question, we wanted to address the best day to commit a low level crime. Looking at the days of the week which had the highest count of severe crimes would seem best since the authorities are preoccupied with protecting the public from such heinous crimes. Our initial hypothesis was that some time during the middle of the week would be the best since much isn't going on, while the weekend would be most dangerous.

Our primary focus was on crimes of severity 4 and 5 since those crimes were rated for the reason that they have the highest risk to physical and psychological health (rape, assault, murder, etc.). Naturally, wandering around Berkeley when severe crimes (4 and 5) *weren't* happening would be ideal. This led us to construct a histogram with the counts of severe crimes faceted by day of the week.



What we found was unexpected, Tuesday had the highest count of severe crime with Sunday being the lowest. We originally thought Friday would be the day with the highest count of severe crimes, but it ended up having the 2nd lowest count. If one were looking to commit a lower level crime, Tuesday or Wednesday would be ideal since the authorities are occupied keeping an eye out for more severe crimes.

This did not agree with our hypothesis at all. The only potential reason we could suggest is that people are typically inside for work on weekdays and predators would think they have a better chance of getting away with a heinous crime.

Recall that our analysis indicates that February has the highest volume of criminal occurrences. Since February has the highest volume of crime, law enforcement would be stretched then and thus, this would be an ideal time to commit a lower-level crime. It's worth noting that overall crime density peaked at around 6PM. So the ideal time for committing a crime would be sometime during the peak return time for students at UC Berkeley (February), during the middle of the week (Tuesday)**,** roughly around dinner time (6PM).

Visualisation

Upon completing the analysis, we figured that it would be helpful for potential readers to have a visual aid to determine which areas in Berkeley were particularly rife with crime or to simply analyze what kind of crimes were particularly prevalent in the specific area of Berkeley in which they lived. To do this, we created a KML-based Google Earth visualization of crime in Berkeley, which we constructed in R.

Using the standard approach to creating a Google Earth visualization in R, we constructed a visualization for our data set, that put a time-sensitive pin in the location of each crime and which recorded the time at which the crime was committed. However, due to the sheer number of crimes committed in Berkeley in the last six months, the map was packed with pins that were all indistinguishable from each other, essentially rendering the visualization useless. We decided to add labels to the visualization and while this helped the reader read the nature of each crime, it did not assist the readability of the map.

We realized that one thing that could help the visualization was if the reader was able to distinguish between crimes based on varying severity - that way, the reader would be able to distinguish the vast number of crimes in his/her area and focus on those that they were particularly concerned about.

To do this, we had to add a new "Style" tag to each "Placemark" node. Contained within this node would be a further sequence of nodes that determined the exact style of the pin dropped on the location of each crime. By creating an if-case that determined the severity of each crime before creating the parent "Style" node, we were able to alter the color of a pin representing a crime based on the severity of the crime.

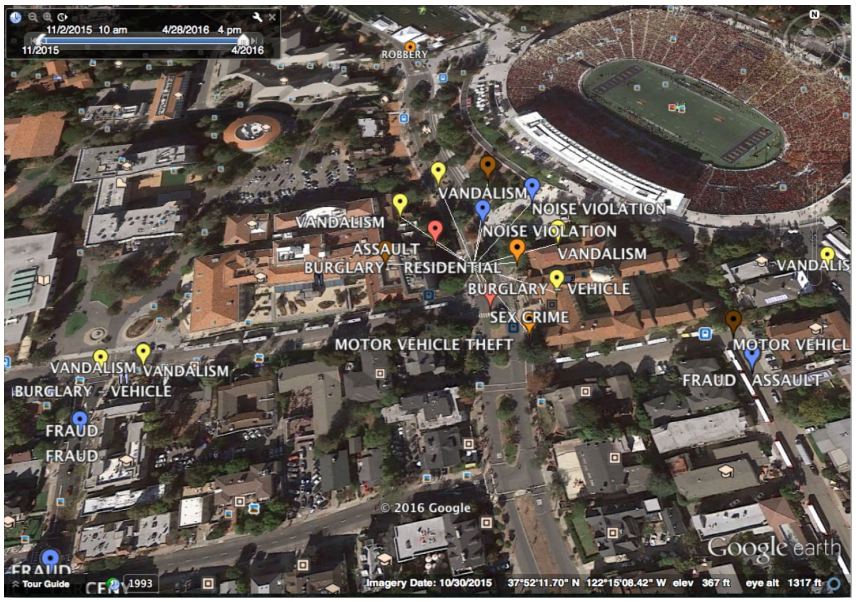
We decided to use the following colors to represent the different severities:

Severity 5 : Red
Severity 4 : Orange
Severity 3 : Yellow
Severity 2 : Green
Severity 1 : Blue

Note: Remember that 5 is the highest severity rating and 1 the lowest

Upon creating the visualization, we recognized that the size of the label was much too large and hence, we added a new node under the “Style” node, called “LabelStyle”, which reduced the size of the label. The full code for the KML can be found in the appendix.

Upon completing the visualization, we zoomed in onto the area in which several of the members of our group live (Piedmont Ave.) in order to look the severity crimes had occurred in our vicinity in the last six months. The following is a sample of our visualization:



Conclusion

The objective of our investigation was to analyze crime patterns in the city of Berkeley, CA. Despite the various degrees of criminal activities occurring throughout Berkeley, we found evidence that there are particular areas of Berkeley that are deemed to be more dangerous than others. Interestingly enough, the data indicates that February has the highest occurrence of criminal activity. Our consensus for this phenomenon is that the large influx of people, particularly students returning back to school, have settled down and with the semester still yet to be in full flow, are more likely to be out and about. Further analysis revealed that the day of the week with the

highest count of crimes was Tuesday, on average. This wasn't consistent with what we assumed, as we thought it would be a weekend since most people go out during those times and one would think that the large amount of people out would be more susceptible to becoming victims of crimes. One step further in the analysis was the time of day which most crimes occurred, which was 6 PM. This was against our assumption once again and a possible causation is that people typically go out during that time for dinner and things of the sort.

Overall, our data revealed interesting conclusions. Our analysis revealed that if we were to give suggestions for people exploring our content, we would say that the safest to live in Berkeley is North Berkeley. Our data revealed that while the danger coefficient is slightly higher than the South, the count is twice as much as the North. So statistically speaking, one should look to live in North Berkeley for a safe location in Berkeley. The safest time to walk home in the evening would be those times when there aren't large masses of people out and about – on weekdays, this would be after 8 pm or before 5 pm.

Finally, the best time to commit a crime is at roughly 6 - 7 PM in the middle of the week (Tuesday or Wednesday), somewhere around February or January because that is the highest volume of crime occurring. Since authorities are going to be preoccupied with dealing with more severe crimes that occur at a high rate at that time, you are more likely to get away with a crime if you commit one around that time.

However, we must admit that our investigation has its limitations with regards to our data collection and analysis. Our data consists of reports that have been reported in the past six months prior to this investigation. Although we requested the city of Berkeley's Online Service Center for more data that dates further past the six months, our request went ultimately unanswered. Our questions were aimed towards the safety of the individual and towards the opportunity to commit a seemingly harmless crime. That being said, our categorization system of severity of crime is based on these questions. It is prone to scrutiny due to its subjective categorization. For example, we categorized narcotics as a category 2 crime. Some might debate that narcotics ought to be placed higher. However, our justification for our rankings is based on how we perceived each crime's effect on the wellbeing of others. With regards to our analysis, we understand clearly that our inferences are prone to skepticism and we do not suggest that our inferences be taken as fact. There are various plausible explanations that are absent in this report and our suggestions are merely one of the many inferences that can be made with the data. Further investigation must be conducted before any claim can be made with certainty.

Despite the limitations present in this report, we attempted to present plausible analysis in an attempt for readers to utilize the information present in order to safely navigate around Berkeley, CA. Should one utilize the data present in this report in a

manner of their choosing, one ought to be aware of its limitations prior to execution of such actions.

Given the analysis conducted here, many other interesting questions can be raised. It would be especially interesting to analyse the relationship between specific crimes and specific areas, or between the demographic of particular areas in Berkeley with the prevalence of crime in the area. These suggestions could plausibly form the basis of a new investigation into continued exploration into the topic.

Appendix

```
library(DataComputing)
library(lubridate)
library(stringr)
library(XML)

# read in data from url
data <-
read.csv(url("https://data.cityofberkeley.info/api/views/k2nh-s5h5/rows.csv?accessType
=DOWNLOAD"))

### clean and wrangle data

# keep select variables
data_trimmed <- data %>%
  select(date = EVENTDT, day = CVDOW, time = EVENTTM, type = CVLEGEND,
details = OFFENSE, location = Block_Location)

# duplicate data frame to create a cleaned version
data_cleaned <- data_trimmed %>%
  filter(location = grepl("([0-9]+\\.[0-9]+)\\, -[0-9]+\\.[0-9]+", location))

# remove time from date
data_cleaned$date <- as.Date(mdy(strtrim(data_cleaned$date, 10)))

# name day factor
data_cleaned$day <- as.factor(data_cleaned$day)
```

```
levels(data_cleaned$day) <- c("Sunday", "Monday", "Tuesday", "Wednesday",
"Thursday", "Friday", "Saturday")

# convert time to number out of 24
data_cleaned$time <- hour(as.POSIXct(data_cleaned$time, format = "%H:%M")) +
minute(as.POSIXct(data_cleaned$time, format = "%H:%M")) / 60

## fix coordinates
# create vector containing the coordinate string for each crime
coord <- unlist(str_extract_all(data_cleaned$location, "([0-9]+\\.[0-9]+)\\,
-[0-9]+\\.[0-9]+"))
# create list with each element of list containing two more elements, latitude and
longitude
coord <- strsplit(coord, ", ")
# vectorized loop goes through coordinate vector and puts lats and longs into
appropriate vectors
l <- length(coord)
lat <- vector(length = l)
long <- vector(length = l)
coord <- unlist(coord)
# lat keeps odd indices and long keeps even indices using recycled logical vector
values
lat[1:l] <- coord[c(TRUE, FALSE)]
long[1:l] <- coord[c(FALSE, TRUE)]

# add lat and long column to and remove location column from cleaned data frame
data_cleaned$lat <- as.numeric(lat)
data_cleaned$long <- as.numeric(long)
data_cleaned$location <- NULL

# simplify data frame name for future ease of use
dc <- data_cleaned

# mutate wday variable to specify part of week for each day
dc <- dc %>%
  mutate(wday = ifelse(day %in% c("Sunday", "Saturday"), "weekend", "weekday"))
%>%
  # filter out
  filter(!as.character(details) %in% c("VEHICLE RECOVERED"))
```

```
## assign severity ratings and join rating to data frame
s <- dc %>%
  group_by(details) %>%
  summarise(count = n())
details <- s$details
# assign severity ratings
severity <- as.factor(c(1, 1, 4, 5, 4, 2, 3, 2, 4, 1, 2, 2, 1, 3, 1, 5, 2, 2, 1, 2, 4, 5, 3, 5, 4, 3,
2, 3, 4, 3))
# create severity rating data frame
detsev <- data.frame(details, severity)
# inner join to append severity variable to data frame
dc <- dc %>%
  inner_join(detsev, by = c("details" = "details"))

## create area label based on crime location coordinates
# create new variables for x and y coordinates recentered by Berkeley's city center
dc <- dc %>%
  mutate(x = long + 122.2727, y = lat - 37.8716)
# Create new vectors to increase readability of code
x <- dc$x
y <- dc$y
area <- vector(length = length(x))
# Categorizing location into one of four quadrants
area[y >= 0 & -y < x & x <= y] <- "N"
area[y < 0 & y <= x & x < -y] <- "S"
area[x < 0 & x < y & y <= -x] <- "W"
area[x >= 0 & -x <= y & y < x] <- "E"
# convert area quadrants as factors and append to data frame
dc$area <- as.factor(area)

## Making KML

# Create a new XML Document
doc <- newXMLDoc()
# Create the root node
root <- newXMLNode("kml",namespaceDefinitions = "http://www.opengis.net/kml/2.2",
doc = doc)
docmt <- newXMLNode("Document", parent = root)
```

```
# Name of the data set for Google Earth
newXMLNode("name", "Crimes in Berkeley", parent = docmt)
# Description of the data set for Google Earth
newXMLNode("description", "Crimes in Berkeley, colored by severity, in the last 6
months", parent = docmt)

# Looping through all cases in the table
for (i in 1:nrow(dc)) {
  # Create a new Placemark node
  pm <- newXMLNode("Placemark", parent = docmt)
  # Naming the pin as the name of crime committed
  name <- newXMLNode("name", dc[i, "type"], parent = pm)
  # Create a new Point node
  pt <- newXMLNode("Point", parent = pm)
  # Creating a string of the coordinate of the crime in KML format
  coordinates <- paste(dc[i, "long"], dc[i, "lat"], 0, sep = ", ")
  # Create the Coordinate node
  newXMLNode("coordinates", coordinates, parent = pt)

  # Create the TimeStamp node
  ts <- newXMLNode("TimeStamp", parent = pm)
  # Creating a string of the time of the crime in KML format
  when <- paste(dc[i, "date"], "T", dc[i, "time"], "-07:00", sep = "")
  # Create the When node
  newXMLNode("when", when, parent = ts)

  # Assigning a differently colored icon to the crime based on its severity
  color <- NULL
  if (dc[i, "severity"] == 1) {
    color <- "http://www.google.com/intl/en_us/mapfiles/ms/icons/blue-dot.png"
  } else if (dc[i, "severity"] == 2) {
    color <- "http://www.google.com/intl/en_us/mapfiles/ms/icons/green-dot.png"
  } else if (dc[i, "severity"] == 3) {
    color <- "http://www.google.com/intl/en_us/mapfiles/ms/icons/yellow-dot.png"
  } else if (dc[i, "severity"] == 4) {
    color <- "http://www.google.com/intl/en_us/mapfiles/ms/icons/orange-dot.png"
  } else if (dc[i, "severity"] == 5) {
    color <- "http://www.google.com/intl/en_us/mapfiles/ms/icons/red-dot.png"
  }
```

```
  # Create the Style node
  st <- newXMLNode("Style", parent = pm)

  # Create the Icon node
  icon_st <- newXMLNode("IconStyle", parent = st)
  icon <- newXMLNode("Icon", parent = icon_st)
  # Create the node that determines the color of the icon
  newXMLNode("href", color, parent = icon)

  # Create the LabelStyle node
  ls <- newXMLNode("LabelStyle", parent = pm)
  # Create the node that determines the scale of the label
  newXMLNode("scale", 0.5, parent = ls)
}

saveXML(doc, "/Users/AndyPalan/proj.kml")

# create simple project theme
theme_proj <- theme(axis.ticks.y = element_blank(),
          panel.background = element_rect(fill = "white"),
          plot.background = element_rect(fill = "white"),
          panel.grid.major = element_line(color = "grey", size=.5),
          panel.grid.minor = element_line(color = "grey", size=.5),
          panel.grid.major.x=element_blank())

dc %>%
  ggplot(aes(x = area, fill = severity)) +
    geom_bar() +
    scale_x_discrete(limits = c("N", "S", "E", "W")) +
    theme_proj +
    labs(title = "Crime Count by Area with Severities")

dc %>%
  ggplot(aes(x = date)) +
    geom_density(fill = "skyblue", alpha = 0.6, color = "skyblue") +
    theme_proj +
    labs(title = "Density of Crime by Time of Year")
dc %>%
  ggplot(aes(x = date)) +
```

```
    geom_density(fill = "skyblue", alpha = 0.6, color = "skyblue") +
    facet_grid(. ~ wday) +
    theme_proj +
    labs(title = "Density of Crime by Time of Year")

dc %>%
  ggplot(aes(x = time)) +
    geom_density(fill = "maroon", alpha = 0.6, color = "maroon") +
    theme_proj +
    labs(title = "Density of Crime by Time of Day")
dc %>%
  ggplot(aes(x = time)) +
    geom_density(fill = "maroon", alpha = 0.6, color = "maroon") +
    facet_grid(. ~ wday) +
    theme_proj +
    theme_proj +
    labs(title = "Density of Crime by Time of Day")

dc %>%
  group_by(area) %>%
  summarise(totsev = sum(as.numeric(severity)), count = n()) %>%
  mutate(dangcoeff = totsev / (count * 5)) %>%
  ggplot(aes(x = area, y = dangcoeff, color = area, fill = area)) +
    geom_bar(stat = "identity") +
    coord_cartesian(ylim=c(0.5, 0.55)) +
    scale_x_discrete(limits = c("N", "S", "E", "W")) +
    labs(title = "Danger Coefficient for Each Area in Berkeley", x = "Area", y = "Danger
Coefficient") +
    theme_proj +
    theme(legend.position = "none")

dc %>%
  filter(severity %in% c(4,5)) %>%
  ggplot(aes(x = date)) +
    geom_density(fill = "chartreuse1", alpha = 0.6, color = "chartreuse1") +
    theme_proj +
    labs(title = "Density of Severe Crimes by Time of Year")

dc %>%
```

```
filter(severity %in% c(4,5)) %>%
ggplot(aes(x = time)) +
  geom_density(fill = "chartreuse1", alpha = 0.6, color = "chartreuse1") +
  facet_grid(. ~ day) +
  theme_proj +
  labs(title = "Density of Severe Crimes by Time of Year")
```