Initial Plan

Group Members:

* Vanessa Chi- 24626664
* Matthew Cline - 24428540
* Yishan Han - 24374993
* Leon Liang - 24514488
* Siqi Zuo - 24445224

* Choose a name for your group.

Team ggplot

We chose ggplot because it is one of the core data analysis tools we have available for this project. It transforms the cold, meaningless mathematical data into heuristically powerful visual images that convey greats ideas. It is the basis of how our insights are derived from the data, thus our group name.

* What is your overall topic? What question(s) will you address?

Analyzing crime rates across all UC campuses based on location, time, and type of crime within the last year.

* Which side of Berkeley has the most crimes?
* What is the frequency of each crime?
* What time do the crimes tend to occur?
* What are the most common types of crime?
* Which UC campus out of the 3, in general, has the highest crime rate?
* Are certain UC campuses more prone to certain types of crime?
* Which academic period has the highest crime rate? (spring semester, fall semester, summer session)
* Where will you get the data? What is involved in obtaining it? What variables will you use?

We will get the data from a Crime Mapping website that maps out all the crimes by type, location and time around the UC campuses and a more detailed dataset for UC Berkeley split between west, north, and south side. The process is as follows: downloading the html file from the Crime Mapping website and importing into Excel to obtain the csv version of the data. Followed by cleaning the data in Rstudio to extract the appropriate data for our analysis. The variables that we will be using will include: crime type, date, location, and time.

The link to the website is here:

<http://www.crimemapping.com/map.aspx?aid=a673b3ba-84b4-4181-a3a4-acba5353ff65>

Scraped and partially cleaned data:

<https://drive.google.com/a/berkeley.edu/folderview?id=0B767Mps5_4FnbUJ4ZVljLWtnc3c&usp=sharing>

* Is there any initial data processing must be done to put the data in a form that is suitable for visualization or analysis?

The data found online is not a clean data set, so there are some missing data points on our locations of interest. We will need to take those into account when cleaning our data by excluding it from the data set and using a large sample size to not skew the results and analysis. The types of crime is given by a description of the crimes that may not always be consistent in terms of the language used, therefore, we will need to categorize the crimes by keywords to separate them. Since the date variable is given in the format of month/day/year and time, we will split this into date and time. From there, we will be able to analyze by the time of day, time of the week, and time of month.

* What plots will you make of the data? Will you perform any analyses, and if so, what?

We can perform multilinear regression models that look at trends of time, types of crime and location. We can aggregate the times of day, times of month, and time of year to analyze the changes in crime rate. From there, we can analyze the relationships between crime, time, and locations. For this, we can create scatter plots and lines of best fits. We would also like to generate histograms to show the frequency of different types of crimes happening around each campus and from this, we can find the most common crime happening around the UC campuses.

* List the responsibilities of each person in the group in completing the rest of the assignment.
* Vanessa: Write first part of report
* Matthew: Write second part of report
* Yishan: First part of data analysis
* Leon: Second part of data analysis
* Siqi: Data Scraping and Wrangling

Everyone will be contributing to data wrangling to create graphs and the interpretation of the trends from the graphs.

* So that I may help you resolve any issues, what do you think will be the most difficult part of doing this project?

We believe data scraping and wrangling will be the most difficult part of the project. From the initial look, some of the data have missing information, so it may be hard to categorize those. We’d also like to do an analysis based on location, but we do not know of a better way to categorize the location of the crimes in Berkeley other than adjusting the center of the circle on crime mapping to west, north, and south side respectively. We’d like to know if there is a better way to categorize the location provided in the dataset into different regions.