Draft Paper

Shangjun Jiang

8 August 2016

Introduction

For the STAT 133 final project, our group cleaned and explored the four police data sets available on the Berkeley Open Data website. We examined police stops, calls for service, arrests, and jail bookings made by the the police department over the past 3-6 month period. Rather than limiting the project to one set of data, these four sets were chosen in order to gain a more holistic and comprehensive understanding of the data. The additional resources of information each data set provided expanded our ability to understand and visualize the data.

One of our main goals was to find some ways of visualizing the relationship between the Berkeley Police Department and different groups of people. In order to compare the proportion of each group represented in the Berkeley police data relative to the population of those groups in Berkeley, we also considered three census data sets that are also available through Berkeley's open data site: first, we included the Berkeley 2000-2010 census data, which gave us the total population counts and percentages of each group of people in Berkeley for the 2010 census; next, we considered the Alameda County 2010 census tract population data, which allowed us to obtain the demographic information for 33 different neighborhoods (or "tracts") inside Berkeley; and finally, the Berkeley 2010 census tract polygons shapefile (.shp), which provided the information we needed to visually display the 33 neighborhoods.

After detailing our process of cleaning the police and census data, we will provided some visualizations of the data and some preliminary findings about the police data. Afterwards, we will discuss problems with the data and some ideas for future statistical analysis.

Berkeley Stop Data Cleaning Process

The original stop data set included 16,255 stops recorded by the Berkeley police over a 6 month period. For each stop, the police recorded the date and time, location (as a character string, usually in the form of a street block or a street intersection), incident type (referring to whether the stop was traffic, pedestrian, bike, or suspicious vehicle), and a column called disposition. The disposition column was a character string typically including 6 characters per individual as follows: the first character was race, the second gender, the third age range, the fourth reason (for the stop), the fifth enforcement (the result of the stop), and the sixth car search (yes or no). Each individual that was involved with the stop was recorded using comma separation in the same column followed by an additional 6 characters. Additional optional dispositions were sometimes included, again using comma-separation from the 6-character strings (though not necessarily in any order), ranging from one to three characters. These additional disposition abbreviations represented the following options: Primary Case Report, MDT Narrative Only, Arrest Reporty Only (No Case Report Submitted), Incident Report, Field Card, Collision Investigation Report, Emergency Psychiatric Evaluation, Impounded Vehicle, and Officer Made a Stop of More Than 5 Persons. We found it reasonable to assume that most of these additional dispositions were related to police paperwork procedures, and since they were used so sparingly, we decided to not consider most of them. However, in spite of only being used less than 20 times, we found the Emergency Psychiatric Evaluation disposition to be of interest, and so we chose to investigate this one in addition to the main 6 character set of dispostions, though we chose not to focus on it during the limited scope of this paper.

Cleaning the stop data was relatively time-consuming compared to the other data sets. First, the dispositions column in the original data set contained many columns and sometimes rows worth of information, though the comma-separated pieces were without order (for example, sometimes the optional disposition letters came before the 6-character string instead). We handled this problem using data tidying techniques, such

as separating disposition information into separate row entries for each individual assessed, in the case of multiple persons stopped, and splitting the column by isolating the six character dispositions into different columns. The optional disposition character strings were moved and separated into more columns.

Additional cleaning for the stop data included changing the stop date and time variable to the `lubridate` date format. We also created an hour column in order to examine the stop occurrences by hour over a 24-hour period.

Turning the stop data's character locations into location coordinates using a Google Maps geolocation library proved difficult because of common data entry mistakes. For example, there were common mispellings of street names and the use of several different abbreviations to mean the same thing. Additionally, we found that the geolocation provided by Google does not accept forward slash characters the data used in the colum to indicate street intersections, so these forward slash characters were changed to `and` for the location processing. Google also does not accept the term "block" to indicate a range of addresses, and so the word "block" was removed from the data set for location processing. Most of these types of errors were delt with on an automated basis before geolocation, using `str_detect` and `str_replace` from the `stringr` package. The number of problems with the location data that were left over afterwards were small enough to handle on a case-by-case basis. Further, because Google limits geolocation to 2500 queries per 24-hour period, it took about a week to process all the locations into coordinates from the stop data (this includes the time it took to fix spelling errors).

At the end of the cleaning process, the stop data was whittled down from 16,255 observations to 14,291 observations (due to missing disposition values, etc.) that we will use in the following analysis.

Berkeley Arrest and Jail Bookings Cleaning Process

The number of observations available for the arrest data set and the jail data set were significantly smaller than the number of observations for the stop data set. The arrest data had 205 observations and the jail bookings data had 223. These observations were obtained by the Berkeley Police department over a 3-month period, rather than a 6-month period. They both contained similar variables, including a case/arrest/booking number, date and time, type, and subject information (name, race, sex, D.O.B., age, height, weight, hair, eyes, and occupation) and statute information (type, description, agency, and disposition). Cleaning required the dates and times to be put into `lubridate` format, and we also created an hour column to consider the number of occurences during each hour of the day.

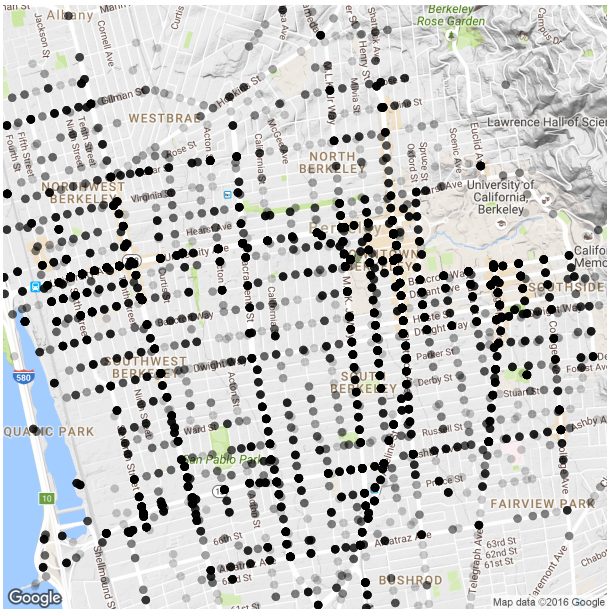
Berkeley Census 2010 Data Cleaning Process

The Berkeley census 2010 data was utilized from 3 different data sets. We first considered the summary information provided by the city website in an excel document, which was easy to clean and use in our analysis, since the information was already summarized. However, in order to understand the demographics of populations within Berkeley in a way that could provide a spatial visualization, we needed to use the census 2010 Alameda County population tract data, which provided the population and race data by tract number. For example, there are 33 census 2010 tracts within the city of Berkeley. Finally, to get the map information of these 33 tracts, we also needed the census 2010 Berkeley tract polygons shapefile. After converting the shapefile to a dataframe and changing the coordinate system to the one used by Google, we were able to `left_join` this new polygonal mapping data frame to the data frame containing the population information for the Berkeley tracts by the unique census 2010 tract number. Combining the tract map information with the tract population information gave us a data frame containing population information for 33 different areas that we could represent spatially. The goal was to visualize the population density of different groups of people within the city by neighborhood, which provided us with a context for the stop data.

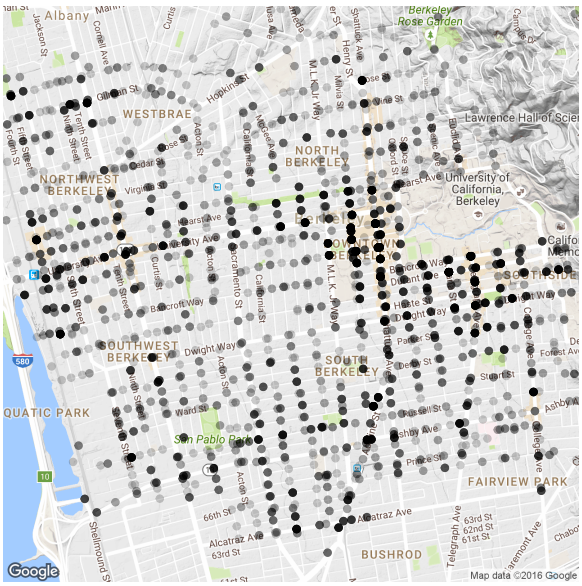
Berkeley Police Stop Data Analysis

Below is the stop data points mapped by location. We have excluded some observations on the outskirts of Berkeley in order to zoom in to the street level. Note that the `alpha` variable in `ggplot` is set to a small number in order to show the incidents that occur multiple times in the same place. In other words, the darker the point, the more police stops have occured at that location.

Berkeley Police Stops, 2015−2016

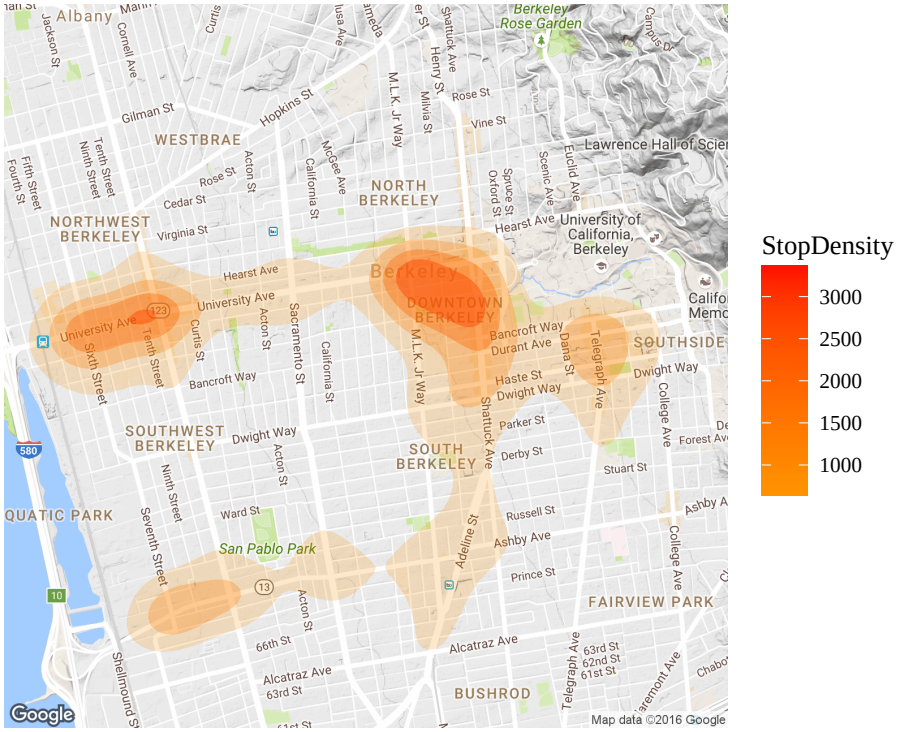


Calls for Service (not criminal reports) Within 180 days (Feb−July 2016)



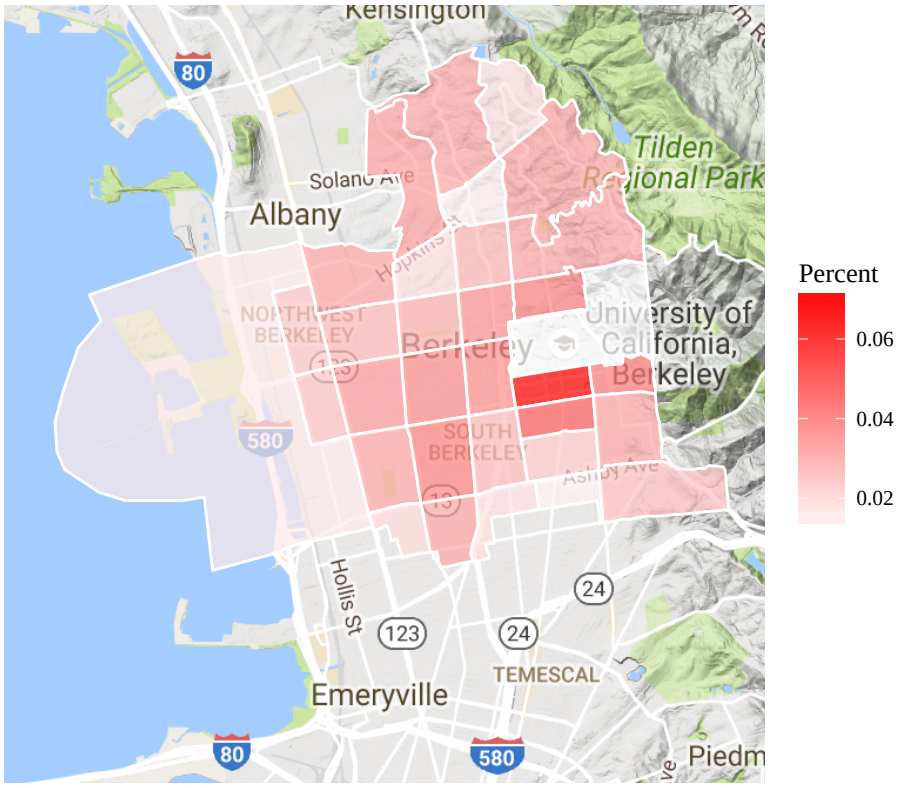
A more interesting visualization of this data is the following two-dimensional density map of the police stop data, displayed below. One of the most interesting aspects of the stop data in this image is that most police stops seem to be clustered west of UC Berkeley Campus in the Downtown area. This was contrary to what some of us might have thought before, which was that the area with the higher police stops would be corrolated to the area with the higher population. As you will see, however, this is not necessarily true.

Berkeley Police Stop Density Map, 2015−2016



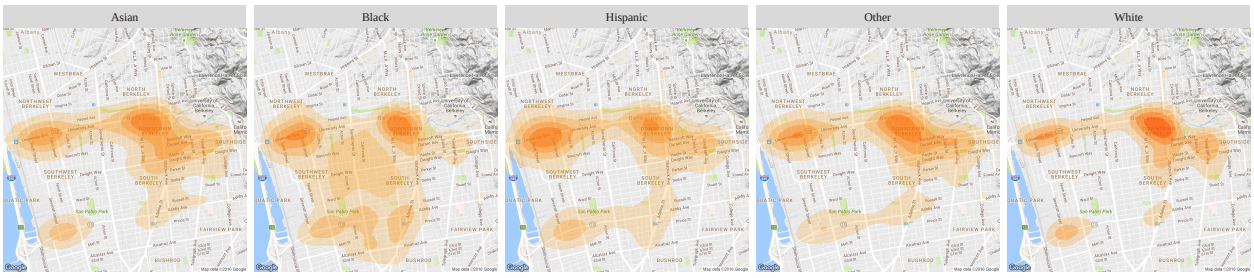
To visualize the relationship between police stops and population density, consider the following map, which visualizes the percent population of Berkeley for each of the 33 census 2010 tracts. As you can see, the least densely populated area is the census 2010 tract containing the University of California. This makes sense because the area of this tract includes a campus and some student housing areas, where permanent residents are less likeley to live. One of the most population-dense areas is Southside, the portion of Berkeley directly south of the UC campus. Note that even though the tract including the Southside area is the most densely populated of the 33 census 2010 tracts in Berkeley, it is not the area where most police stops occur.

2010 Population Density

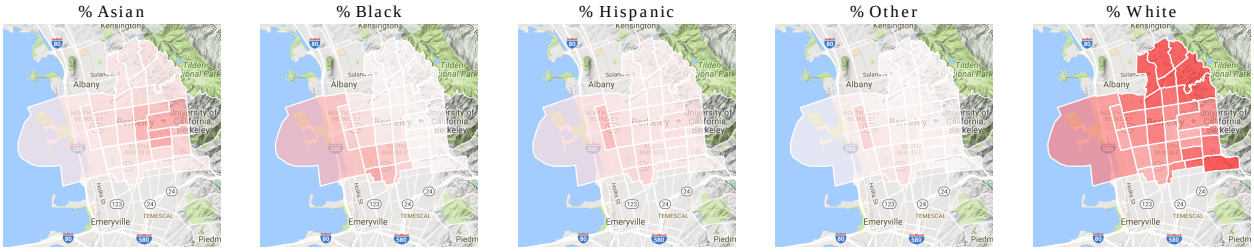


In order to break down the stop data further, we mapped the data by each of the disposition variables. We found race to be one of the most interesting variables. Consider the Berkeley stop data density map faceted by race:

Berkeley Police Stop Density Map by Race



The image above seems to imply that the group of people stopped the most are African Americans in Berkeley. For instance, in Centeral Berkeley, the stops for the races defined by the Berkeley Police Department as Asian, Hispanic, White, and Other seem comparatively lower than for Blacks in the same area. In order to consider the context of these stop incidents, we considered the population distribution for these areas, as shown below. Note that the population density visualized by the following map is conveyed by the range between white (0%) and red (100%). As you can see, White people make up a striking majority in all census 2010 tract areas, which seems to imply that the population of Black people in Berkeley is being disproportionally affected by police stops.

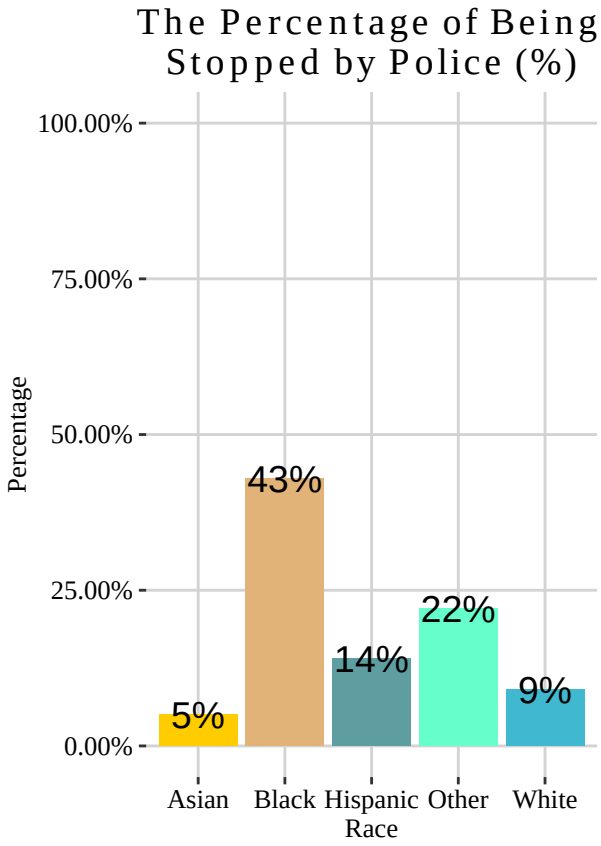


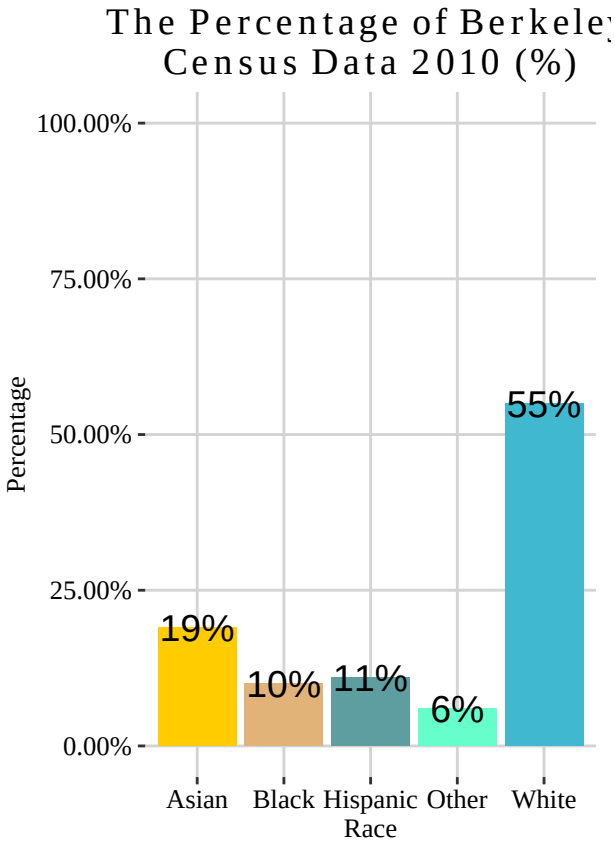
Now, with the population percentages of each group in hand, we were able to compare the results of our stop data analysis to the percent population of each group. The `Percent Stopped` column in the table below represents the percent of each group compared to the total number of people stopped, while the `Percent Population Stopped` column represents the proportion of people in each group stopped relative to the population of that group living in Berkeley. Even though the people stopped by police are not necessarily Berkeley residents, the `Percent Population Stopped` gives us some perspective of which groups of people in Berkeley might be more likely to be affected by police stops than others.

Race	Percent Stopped	Percent Population Stopped
Asian	8.0%	5.3%
Black	32.4%	41.2%
Hispanic	11.7%	13.7%
Other	9.7%	24.5%
White	38.2%	8.1%

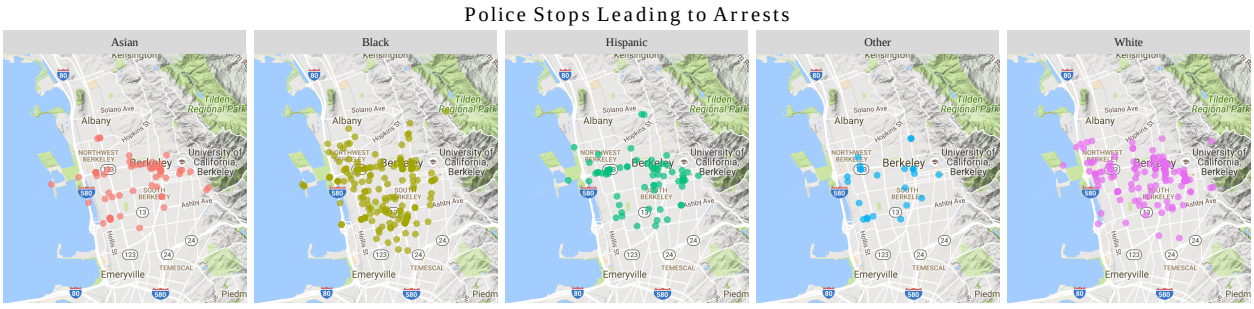
A visualization of the groups of people stopped in proportion to their population is provided by the bar chart below.

!!!!!!!!!!!!!!!!!!!!!!INCLUDE BAR GRAPH HERE of % race stopped vs. percent race population, as described above!!!!!!!!!!!!!!!!!!!!!!!!!

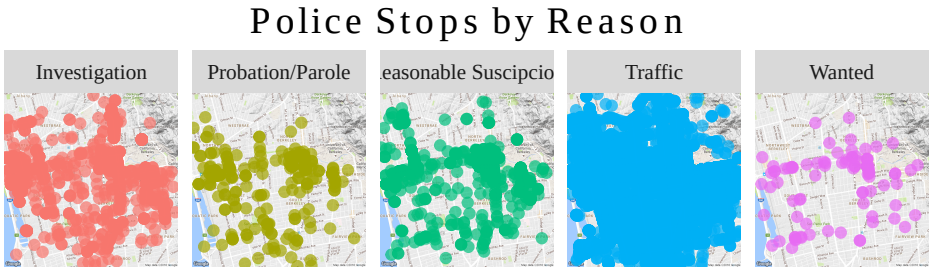




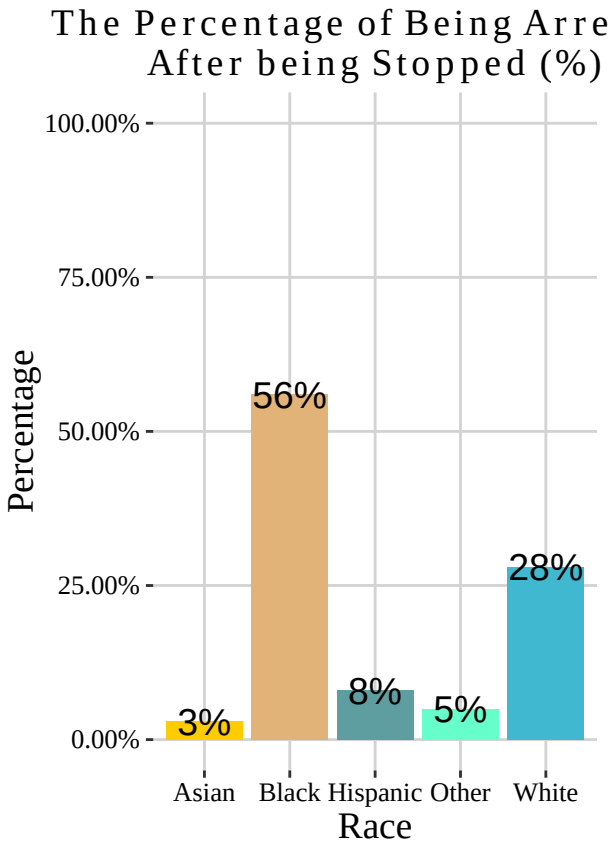
Another disposition variable we found to be of interest was the enforcement variable, which has the options Arrest, Citation, Warning, and Other. We were interested to see how these enforcements were applied to different groups of people. Consider the map below, which appears to suggest that the enforcement of arrest is applied differently to different groups of people.

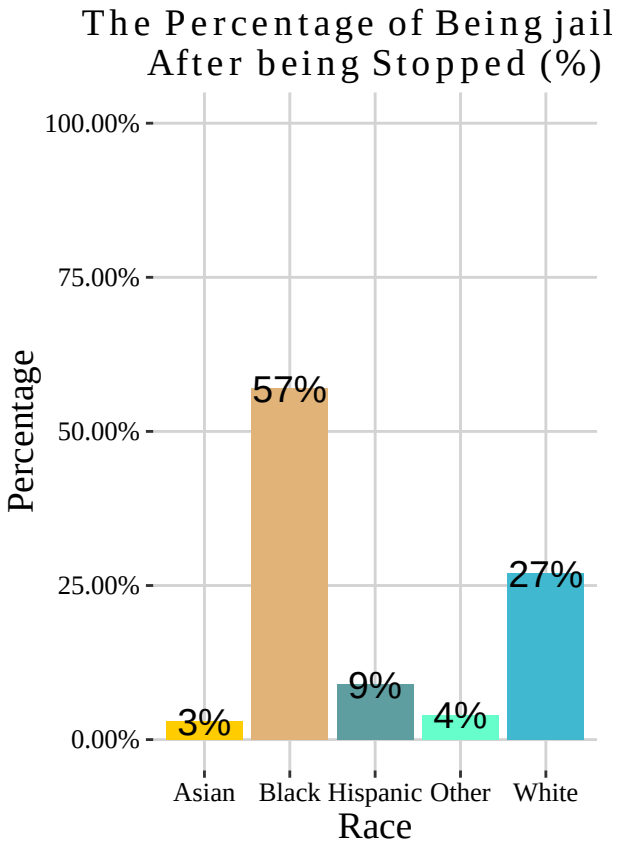


A third interesting disposition variable was the reason disposition, which indicates the reason for the police stop in each case. The reasons available for police stops are Investigation, Other, Probation/Parole, Reasonable Suspicion, Traffic, and Wanted. One way we explored the data initially was to see how these reasons affected different groups of people, and we would recommend exploring this idea in depth at a later time. Due to the length and scope of this paper however, we decided to end our discussion of the stop data by showing that most of the Berkeley police stops are traffic stops. Mapping the coordinates seems to suggest this fact, and it is confirmed when analyzing the data.



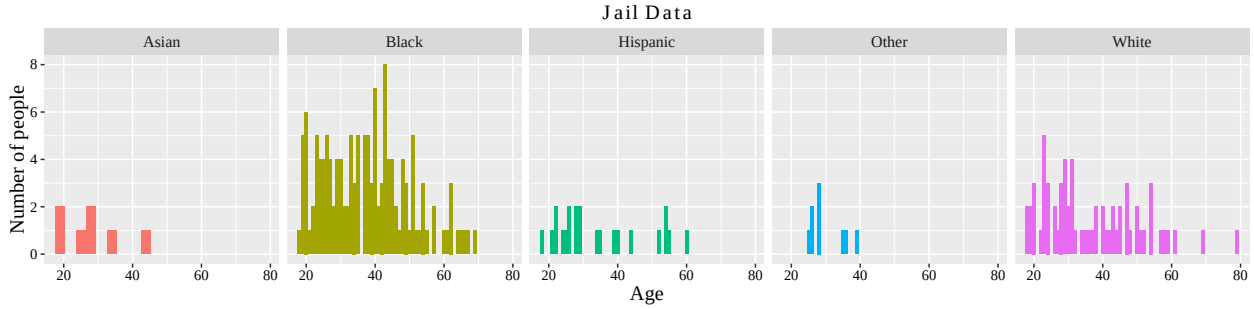
!!!!!!!!!!!!!!!!!!!!!!INCLUDE BAR GRAPH & TABLE OF POLICE STOPS BY REASON HERE (Use percentage, not count) !!!!!!!!!!!!!!!!!!!!!!!!

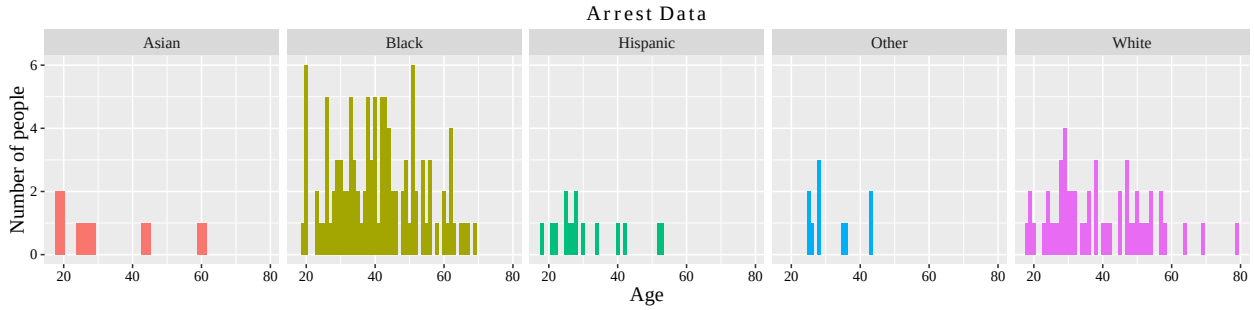




Berkeley Police Jail and Arrest Data Set Analysis

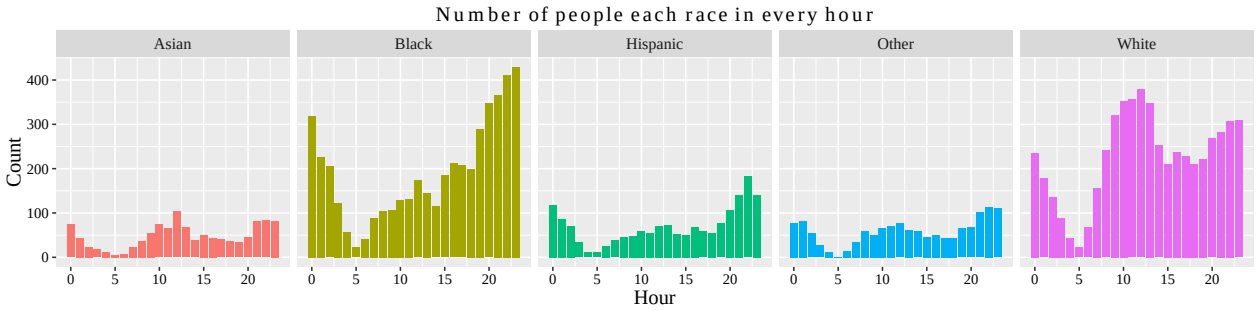
Moving on from the stop data, we considered the available jail data and arrest data for the city of Berkeley. Two of the most interesting variables to consider in these two data sets were race and age. We decided the examine the percent of people arrested and the percent of people jailed in Berkeley by both age and race. Below is a visualization. Note that the following percentages are related to the number of total people jailed or the number of people arrested, respectively, and we have not adjusted the data relative to the population of Berkeley.





One of the most interesting things about the above histograms is the number of black people arrested in Berkeley over the past 3 month period is higher than the numbers of people arrested in Berkeley over the past 3 months for other groups of people. Note that this statement is even stronger for young black people. Even though most of the population in Berkeley is white, most of the people recently arrested are black.

Another idea we had was to visualize the arrest data for each community by hour. !!!!!!!!!!!PLEASE FIX LABEL SO THAT IT IS PERCENT ARRESTED, NOT COUNT!!!! ALSO, PLEASE DISPLAY A SMALL TABLE OF THIS DATA FOLLOWING THE HISTOGRAM.!!!!!!!!!!!!!



!!!!!!!!!!!!!! 1-2 sentence: ANALYZE FINDINGS FROM chart & TABLE ABOVE HERE!!!!!!!!!!

Problems

The data sets provided by the police (stops, arrests, and jail bookings) have several problems. First, concerning the methods for recording race, especially for police stops that do not result in an arrest, it seems unreasonable to assume that racial information would have been provided by the individual stopped. It seems more likely that police officers are judging the race of the people they stop during stops for themselves in order to record it in the system that ultimately provided us with that data set. The problem with that method of data collection is that it opens up an opportunity for error; for example, if the data is recorded by the officers

through their judgment of the individuals physical appearance, they may unknowingly classify that individual as a race that the individual does not identify themselves as. Even though there doesn't seem to be a means to get around this problem with the actual method of collecting racial data for police stops, it is still important to address it as something that could induce error into a statistical analysis.

In further regards to the racial data available, it is important to note that the census data we used to compare the stop data to included many more categories for race than just the five used by police (Asian, Black, Hispanic, Other, and White). To deal with this problem, other races included by the census were moved to the category of other. Additionally, the census data had a category for Two or More Races, which we also included as Other in order to compare the data sets. While the census data allowed for overlap between racial categories, there was no overlap for the racial categories in the police data. This creates a problem when comparing the two data sets, and so it should be acknowledged.

Secondly, the police data used in this paper comes from a data set maintained by the City of Berkeley that is updated every 3 months (as is the case for arrests and jail bookings) or every 6 months (as is the case for stops). A more thorough analysis of this data may want to include an analysis of the data over a longer time frame.

Next, with regard to the stop data in particular, the many data-entry problems we faced in the character string location column (which sometimes included only 1 street), there is likely some stops that are not mapped correctly due to approximations made by the Google geolocation service (for example, in some cases, only having one street name instead of a corner caused Google to estimate the exact location based on few conditions). Though most of the location problems were fixed using string detection, there is likely some error in our visual representation of the stop data maps.

Finally, with regard to the calls for service, we were unfortunately unable to incorporate that particular data set due to constraints on time. However, a more full analysis of police interactions with the community might consider incorporating the calls for service data.

Conclusion

The goal of this paper was to investigate and visualize Berkeley Police interactions with the community through the available police data. Although our preliminary findings unfortunately seem to suggest that certain groups of people, especially black people, are disproportionately likely to be stopped, arrested, and jailed in the city of Berkeley, we recommend a more rigorous statistical analysis of the data be done in order to assess this claim. We also recommend an idea we had for this project to future researchers that could make the Berkeley police data more accessible. For example, one could make the data more accessible by creating a `shiny` app that allows the user to toggle between the different dispositions of the data (race, age range, enforcement, reason, etc.) and between modes of view (map or bar chart, for example) or order to play with the data in a visual way. Another `shiny` app idea we had was to filter the calls for service data by proximity to certain locations so that users could use the app to select a location and find a visual displaying the most common types of calls in that area. It would also be useful to find a way to create an app can easily be updated when the new data is posted (every 3-6 months), allowing users of the app to get the most up-to-date information.

Sources

- Stop Data (16,000)
- Arrest (200)
- Jail Bookings (250)
- Calls for Service (4,000)
- Berkeley Census Data *Census 2010 Population and race data by county tract polygons* Berkeley Census 2010 Tract Polygons