STAT 133 Final Project: Visualizing Berkeley Police Data

Amy Zhu, Mengyu Li, Rebecca Reus, and Shangjun Jiang

9 August 2016

Abstract

For the STAT 133 final project, our group was interested in finding some ways of visualizing the relationship between the Berkeley Police Department and different groups of people, especially within the context of different populations and neighborhoods within Berkeley. To obtain our results, we accessed, cleaned and explored the four Berkeley Police Department data sets available to the public on the Berkeley Open Data website using R. The police data included police stops, calls for service, arrests, and jail bookings made by the the police department over the past 3-6 month period. Further, we accessed 2010 census data in order to obtain demographic information about Berkeley as a whole and about neighborhoods inside of Berkeley. The three census data sets that we used are also available through Berkeley's open data site, including 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, the Alameda County 2010 census tract population data, which gave us the demographic information for 33 different neighborhoods (or "tracts") inside Berkeley, and the Berkeley 2010 census tract polygons shapefile (.shp), which provided the information we needed to visually display the 33 neighborhoods. Rather than limiting the project to one set of data, incorporating multiple data sets gave us a more holistic understanding of subject and expanded our ability to understand and visualize the data. Our findings suggest the hypothesis that (1) black people in Berkeley are more likely than others to be stopped by Berkeley police, and that (2) black people in Berkeley are arrested and jailed at disproportionate rates. We recommend future statistical analysis be done to test our hypothesis.

Introduction

After detailing our process of cleaning the police and census data, we will provide our results: visualizations of the data and some preliminary findings about the police data. Afterwards, we will discuss problems with the data and recommendation for future work on this subject.

Cleaning Data

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 row and 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 an 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.

We used the stop data's character string locations into location coordinates using a Google Maps geolocation service provided by the library `ggmap` with the function `geocode`. This task 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 column to indicate street intersections, so these forward slash characters were changed to `and` for the location processing, which was successful. Google also does not accept the term "block" to indicate a range of addresses, and so the word "block" was removed from the location column 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.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	Incident.Nun	Call.Date.Tim	Location	Incident.Typ	Other	Individual	Dispositions	Race	Gender	AgeRange	Reason	Enforcement	CarSearch	lat	long	Arrest	Emergency.P	Hour	Day
1	2015-00004	###########	2000 BLOCK	1194	NA	1	BM4ICN	Black	Male	40+	Investigation	Citation	No Search	37.8726801	-122.27074	Not Arrested	No	7	2
2	2015-00004	###########	1700 BLOCK	1194	NA	1	BM4ICN	Black	Male	40+	Investigation	Citation	No Search	37.8731719	-122.2938	Not Arrested	No	7	2
3	2015-00004	###########	M L KING JR	T	NA	1	OF4TCN	Other	Female	40+	Traffic	Citation	No Search	37.8716087	-122.27303	Not Arrested	No	9	2
4	2015-00004	###########	M L KING JR	T	NA	1	OM4TCN	Other	Male	40+	Traffic	Citation	No Search	37.8716087	-122.27303	Not Arrested	No	10	2
5	2015-00004	###########	UNIVERSITY	T	NA	1	OF2TCN	Other	Female	18-29	Traffic	Citation	No Search	37.8716087	-122.27303	Not Arrested	No	10	2

Figure 1:

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 (as was for the stop data). 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 this data set also 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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Arrest Numb	Date and Tim	Arrest Type	Subject	Race	Sex	Date of Birth	Age	Height	Weight	Hair	Eyes	Statute	Statute Type	Statute Desc	Case Number	
2	16347	06/20/2016	COURT FILED	Frank Moore	Black	Male	08/25/1982	32					Warr - Out (	PC;	Outside Warrant Misdemeanor;		
3	16431	06/30/2016	ON-VIEW BY	Michael Harc	Black	Male	05/07/1992	24					Warr - Out (	PC;	Outside Warrant Misdemeanor;		
4	16349	06/21/2016	COURT FILED	Julia Elizabet	White	Female	02/26/1998	18	5 Ft. 0 In.	90	BRO	HAZ	Warr - Out (	PC;	Outside Warrant Misdemeanor;		
5	16345	06/20/2016	ON-VIEW BY	Robert Lee C	Black	Male	06/02/1952	63	5 Ft. 9 In.	130	BLK	BRO	1203.2 - F; 4	PC;	Probation Vi	2016-00036449	
6	16346	06/20/2016	ON-VIEW BY	LAUREN LOU	White	Female	06/05/1985	30	5 Ft. 5 In.	138	BRO	GRN	243 (E)(1);	PC;	Battery: spou	2016-00036562	
7	16594	##########	SUSP. OF FEL	Jalisa Nicole	Black	Female	03/08/1990	26	5 Ft. 2 In.	140	BLK	BRO	459;	PC;	Burglary;	2016-00040113	
8	16582	07/17/2016	ON-VIEW BY	Ismael Valen	Hispanic	Male	05/26/1990	26	5 Ft. 7 In.	145	BLK	BRO	23152 (A) - N	VC;	DUI: Alcohol	2016-00042026	
9	16353	06/21/2016	ON-VIEW BY	Cheryl Denis	Black	Female	05/31/1983	33	5 Ft. 6 In.	130	BLK	BRO	1203.2 - M; 6	PC;	Probation Vi	2016-00036799	

Figure 2:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Booking Num	Booking Date	Subject	Race	Sex	Date Of Birth	Age	Height	Weight	Hair	Eyes	Occupatio	Statute	Statute T	Statute Desc	Arrest Date a	Case Numbe	Booking Age	Disposition	
2	2016-000017	06/20/2016	LAUREN LOU	White	Female	06/05/1985	31	5 Ft. 5 In.	138	BRO	GRN		243 (E)(1);	PC;	Battery: spou	06/20/2016	2016-000365	CA0010300	BAILED	
3	2016-000017	06/21/2016	ALONDRICK	Black	Male	01/14/1993	23	6 Ft. 2 In.	175	BLK	BRO		Warr - Out (	PC;	Outside War	06/21/2016	2016-000014	CA0019700	CITE-JAIL	
4	2016-000017	06/21/2016	Julia Elizabet	White	Female	02/26/1998	18	5 Ft. 0 In.	90	BRO	HAZ	UNEMPLO	11364 (A);	HS;	Possess narc	06/21/2016	2016-000366	CA0010300	SANTA RITA JAIL	
5	2016-000017	06/22/2016	JORDAN ELIJ	Black	Male	04/05/1997	19	6 Ft. 1 In.	145	BRO	BRO		647 (H);	PC;	Disorderly co	06/22/2016	2016-000368	CA0010300	CITE-JAIL	
6	2016-000017	06/22/2016	Anecia Johne	Black	Female	08/31/1996	19	5 Ft. 4 In.	115	BLK	BRO		Warr - Out (	PC;	Outside War	06/22/2016	2016-000368	CA0010300	SANTA RITA JAIL	
7	2016-000017	06/21/2016	Davin Williar	White	Male	04/01/1971	45	6 Ft. 3 In.	190	BRO	BLU		23152 (A) - N	VC;	DUI: Alcohol	06/21/2016	2016-000367	CA0010300	CITE-JAIL	

Figure 3:

Berkeley Census 2010 Data Cleaning Process

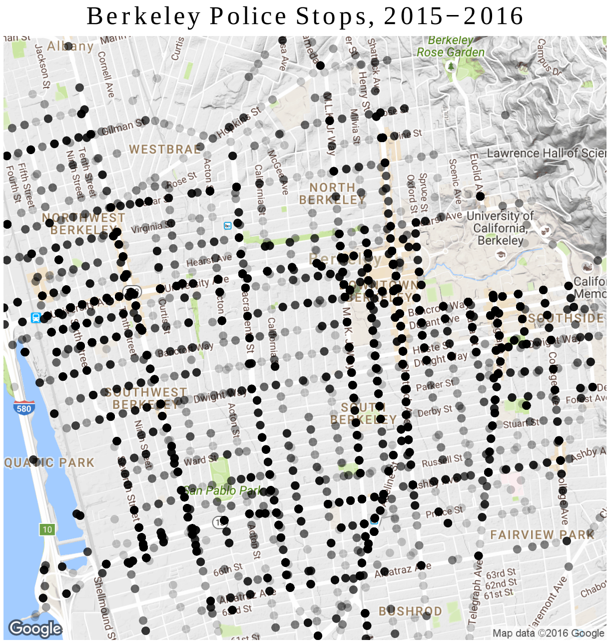
The Berkeley census 2010 data was utilized from 3 different data sets. We first considered the summary information provided by the city website in an excel document, which was easy to clean and use in our analysis, since the information was already summarized. However, in order to understand the demographics 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 (this was done using the `readOGR` function from the library `rgdal`), 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 sptatially visualized stop data.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	Population	Hispanic	White	Black	Asian	Other	Percent.Blac	Percent.Othe	Percent.Whit	Percent.Asia	Percent.Hisp	Percent.Berkeley
2	4238	2925	128	2549	49	182	29	0.01675214	0.00991453	0.87145299	0.06222222	0.04376068	0.02598152
3	4222	3144	358	1882	402	484	155	0.1278626	0.04930025	0.59860051	0.15394402	0.11386768	0.02792681
4	4224	4196	343	2324	191	1360	110	0.04551954	0.02621544	0.55386082	0.32411821	0.08174452	0.03727127
5	4225	4658	323	2889	104	1290	97	0.02232718	0.02082439	0.62022327	0.27694289	0.06934307	0.04137502
6	4223	3387	311	2224	196	651	106	0.05786832	0.03129613	0.65662828	0.19220549	0.09182167	0.03008527
7	4218	2007	100	1587	59	219	25	0.02939711	0.0124564	0.79073244	0.10911809	0.04982561	0.01782732
8	4236.01	2642	203	1934	92	385	90	0.0348221	0.0340651	0.7320212	0.14572294	0.07683573	0.02346776
9	4216	3558	188	2872	59	368	66	0.01658235	0.01854975	0.80719505	0.10342889	0.05283867	0.03160419

Figure 4:

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.



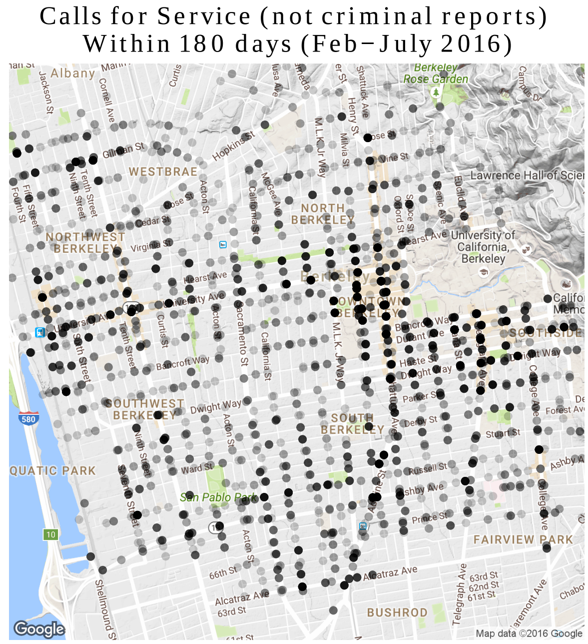


Figure 5: BP stop and Calls for service

PLEASE COMPARE THIS TO THE CALLS FOR SERVICE DATA

LOVE JENNY

Another visualization of the stop data mapped above 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 the area with the higher population. As you will see, however, this is not necessarily true.

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.

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:

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 suggest the hypothesis 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

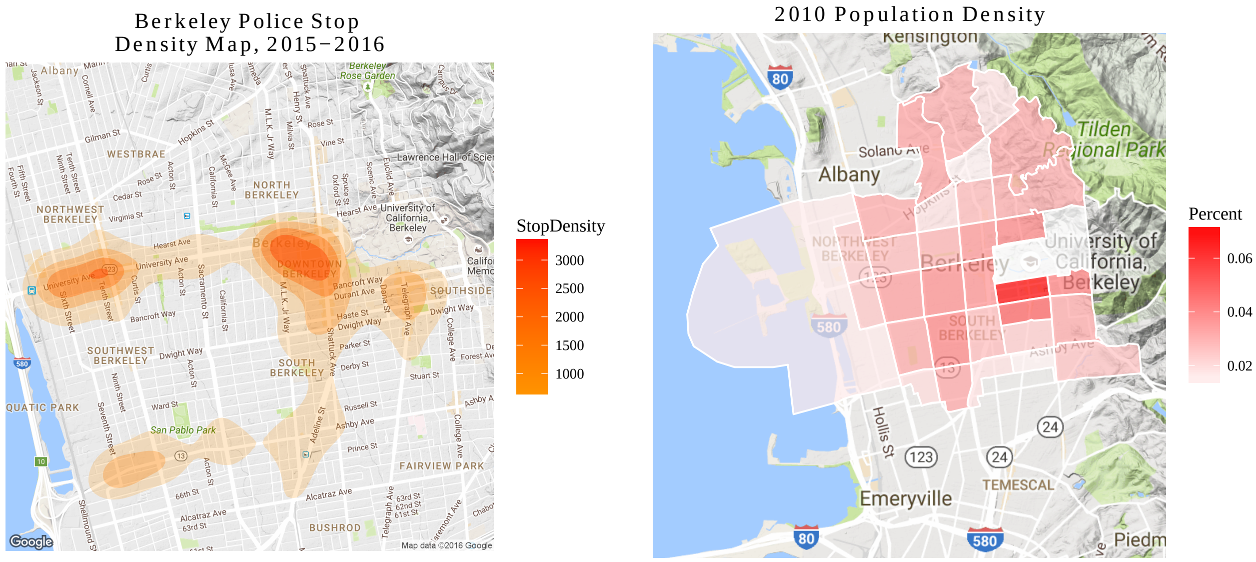


Figure 6: Berkeley Police Stop Density Map by Race

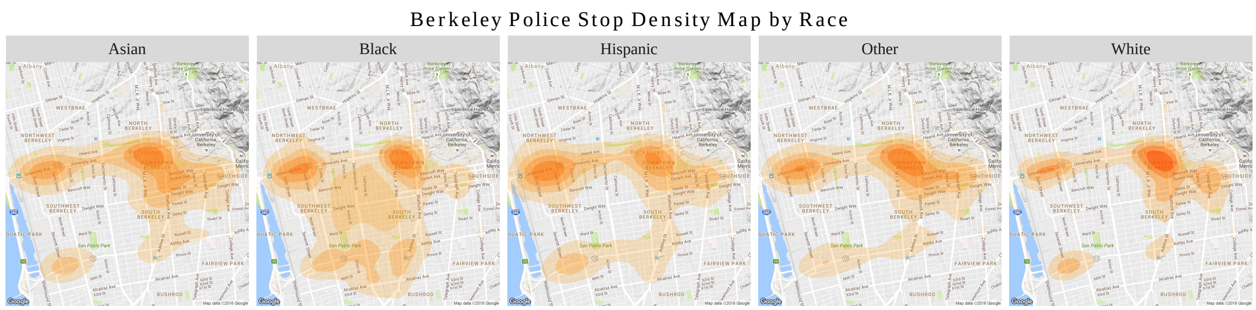


Figure 7: Police Stop density map VS. Population Density

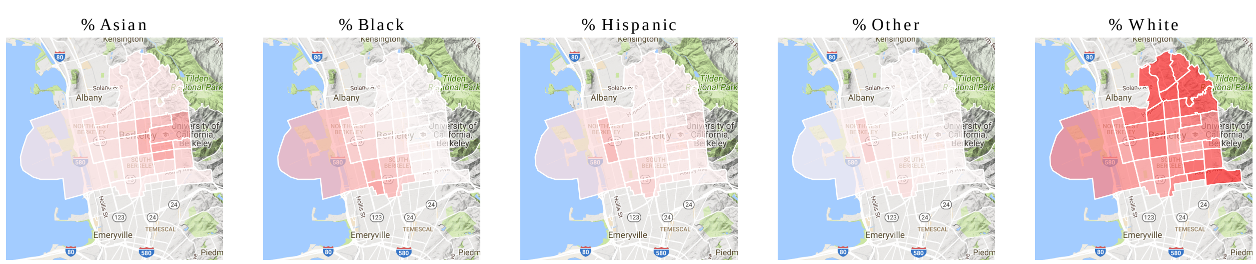


Figure 8: Berkeley Population Density Map by Race

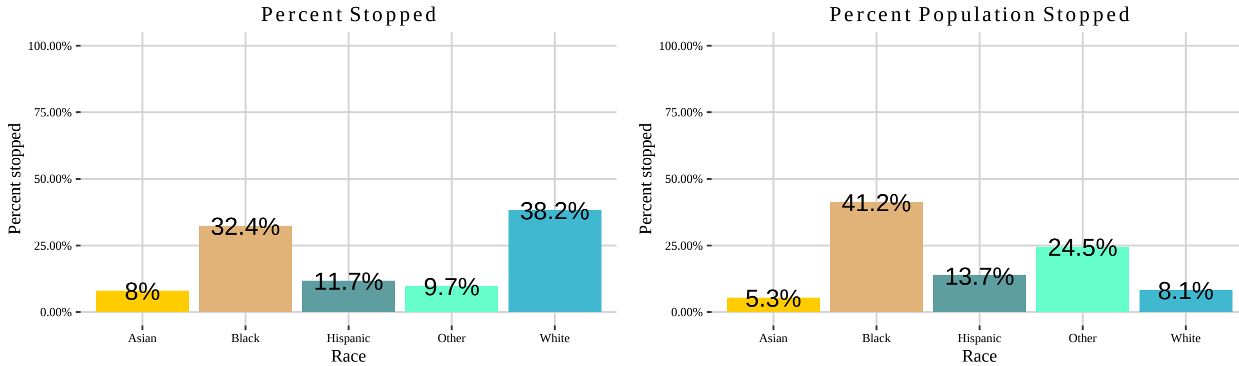


Figure 9: Percent stopped and Perecent Population Stopped

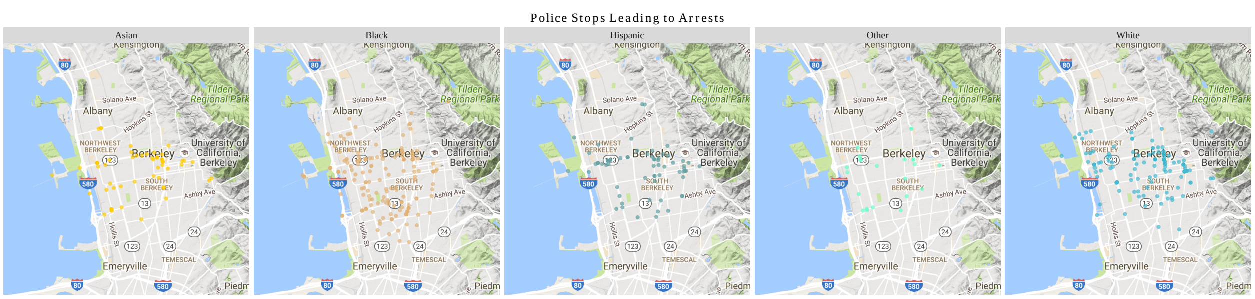
below represents the percent of each group compared to the total number of people stopped, while the `Percent Population Stopped` column represents the proportion of people in each group stopped relative to the population of that group living in Berkeley. Even though the people stopped by police are 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.

PLEASE mention that we delete tHE TABLE

LOVE JENNY

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

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.

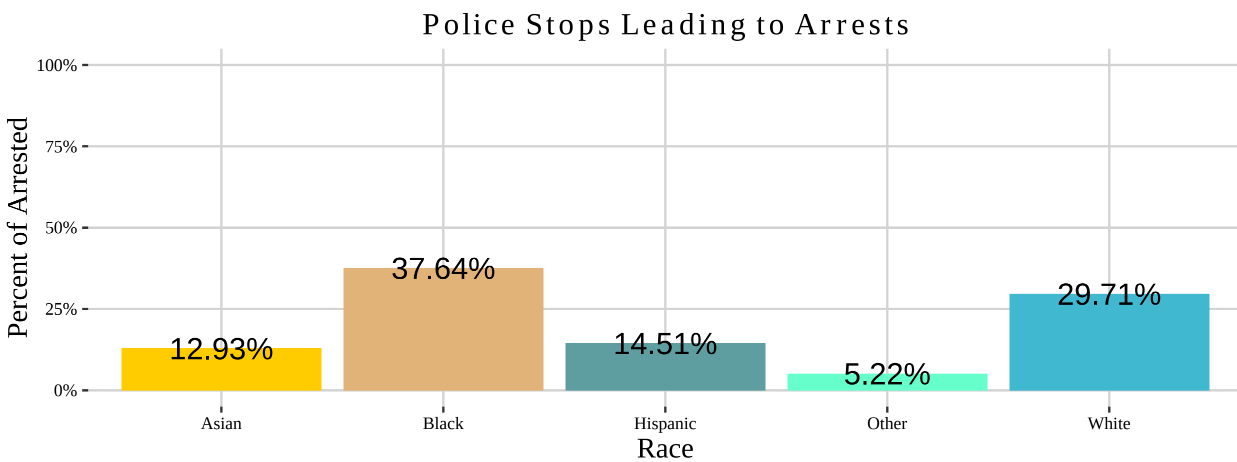


Figure 10: Police Stops Leading to Arrests

Figure 11: Police Stops by Reason

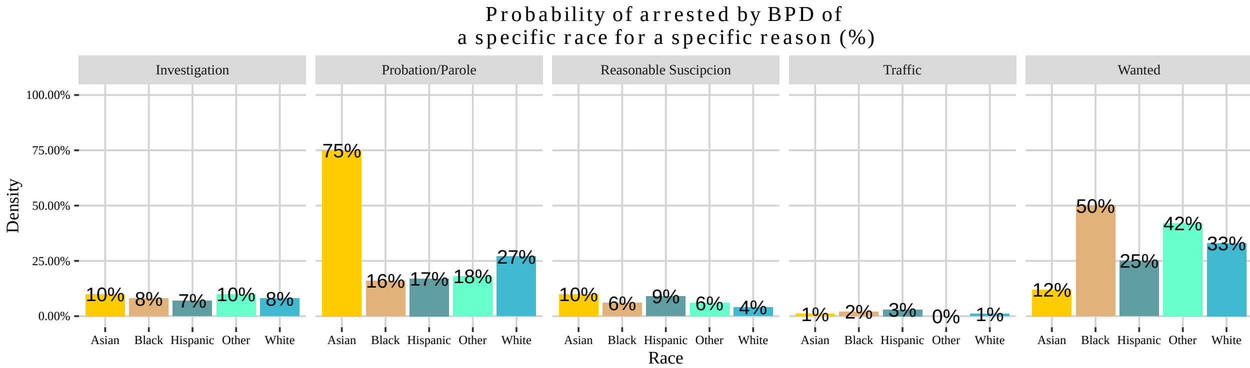


Figure 12: Probability of Arrested by BDP Race VS Reason

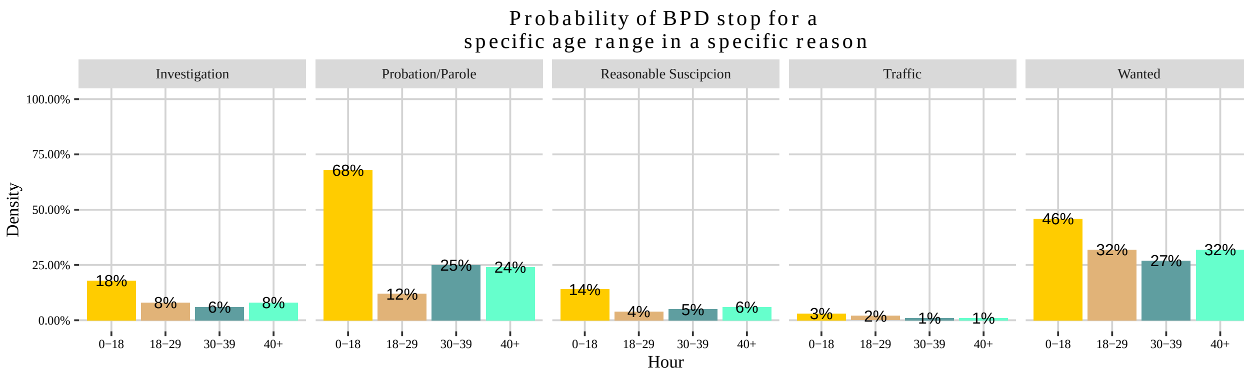


Figure 13: Probability of Arrested by BDP Age VS Reason

PLEASE EDIT THE ANALYSIS

LOVE JENNY

We also analyzed the preference of Berkeley PD when arresting people. We calculated the conditional probability of arrested given a specific reasion and given a specific race/ age range/ gender. That is to say, we calculated the probability of a person will be arrested by Berkeley PD if he was stopped by a specific reason and was of a specific race/ age range/ gender.

PLEASE EDIT THE ANALYSIS

LOVE JENNY

From the above plots, we can tell that the probability of arrested in a stop with the reason Traffic is much lower than other reasons. The average conditional probability of arrested given reason is traffic is 1.58%. Additionally, the probability of arrested in a stop with the reason Wanted is much higher than other reason. The average conditional probability of arrested given reason is wanted is 33.33%. Lastly, an interesting fact is that the conditional probability of arrested in a stop with the reason Probation or Parole and race Asian is 75%, which is much higher than that of any other race. Asian people is much more liable to be arrested by Berkeley Police Department during Probation or Parole if stopped by the BPD.

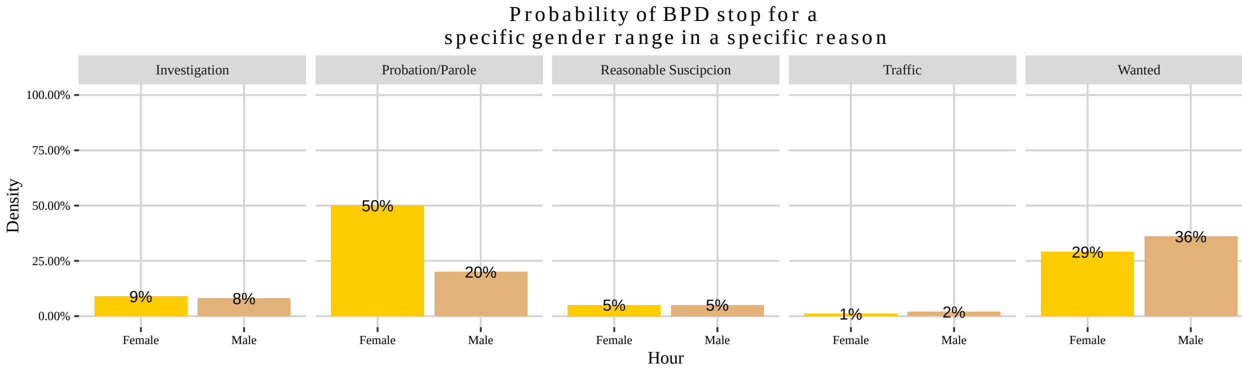


Figure 14: Probability of Arrested by BDP Gender VS Reason

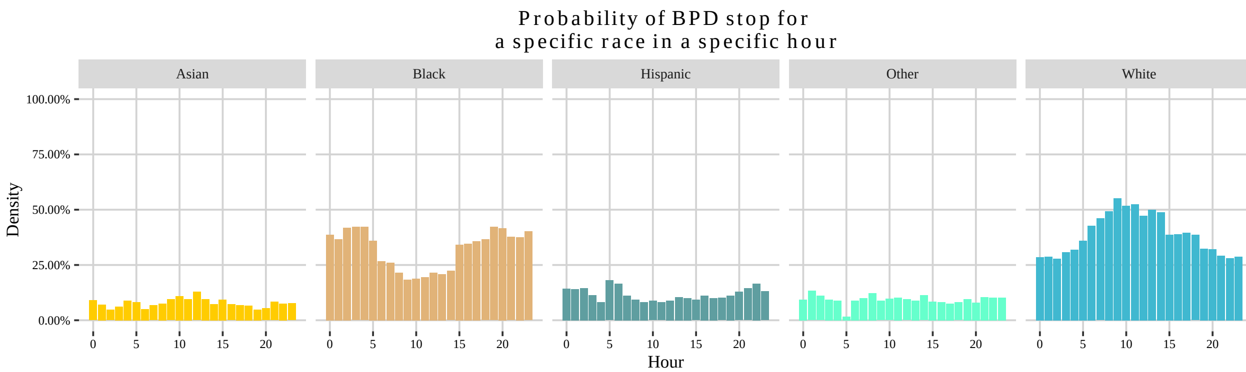


Figure 15: Probability of Arrested by BDP Race VS Hour

FINDINGS

LOVE JENNY

Similar to the conclusions above, the probability of arrested in a stop with the reason Traffic is the lowest and wanted is the highest. Again similar to the conclusion above, an interesting fact is that the conditional probability of arrested in a stop with the reason Probation or Parole and age range 0-18 is 67.56%, which is much higher than that of any other age range. Lastly, the conditional probability of arrested given person involved aged from 0 to 18 in a stop given any reason is more than that of given person with any other age range. Therefore, in a stop, teenagers is more liable to be arrested by Berkeley Police Department.

FINDINGS

LOVE JENNY

Similar to the conclusions above, the probability of arrested in a stop with the reason Traffic is the lowest and wanted is the highest. Again similar to the conclusion above, an interesting fact is that the conditional probability of arrested in a stop with the reason Probation or Parole of female is 50.00%, which is much higher than that of male, which is 20.11%.

Findings

LOVE JENNY

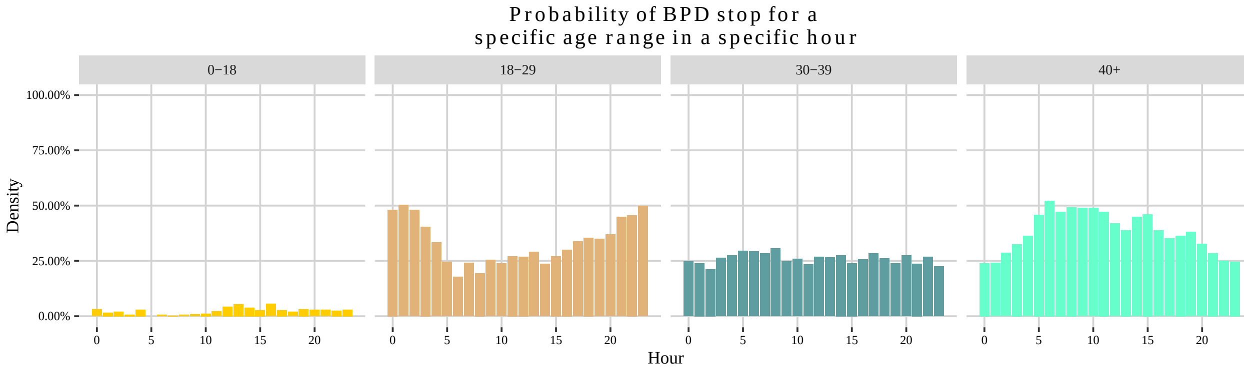


Figure 16: Probability of Arrested by BDP Age VS Hour

Black people are obviously liable to commit incidents at night. The average ratio of black people to all people stoped by the Berkeley Police Department at night is about 40%. On the other hand, white people are obviously liable to commit incidents in the daytime. The average ratio of white people to all people stoped by the Berkeley Police Deparment during the day is about 45%. While the Ratio of incidents commited by Asian people, Hispanic people and other people fluctuate during the daytime and the night, with an average ratio of 8%, 11% and 9% respectively.

Findings

LOVE JENNY

People aged from 18 to 29 are obviously liable to commit incidents at night. The average ratio of people aged from 18 to 29 to all people stoped by the Berkeley Police Department at night is greater than 40%. People aged greater than 40 are obviously liable to commit incidents in the daytime. The average ratio of people aged greater than 40 to all people stoped by the Berkeley Police Deparment during the day is greater than 40%. While Ratio of incidents commited by people aged between 0 and 18 and people aged from 30 to 39 fluctuates during the daytime and the night, with an average ratio of 2.5% and 25% respectively.

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.

Problems

PLEASE MENTION THE HOMELESS PEOPLE.

LOVE JENNY

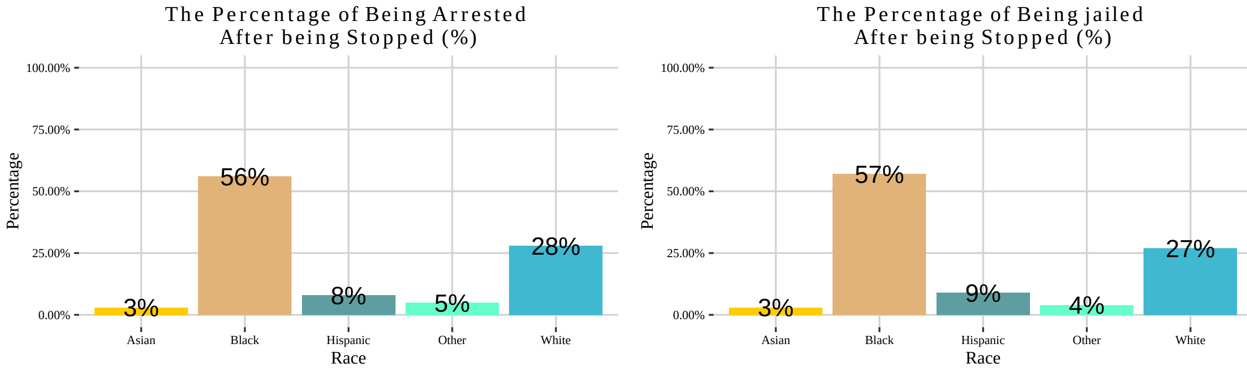


Figure 17: Percentage of being arrested and being jailed

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