Investigating for Bias in Houston, Texas Police Citations

Grace Wang, Virginia Baskin, Siddhi Narayan

2023-04-18

Introduction

Problem and Importance

For our project, we wanted to explore vehicular police stop data from Houston, Texas to investigate and identify any potential biases against particular populations. This problem is not trivial, as it could contribute towards revealing patterns of discrimination or disproportionate targeting of certain groups based on race or gender. Such biases can have serious consequences, such as perpetuating systemic inequality and eroding trust in law enforcement. By analyzing data on police stops and citations, we can illuminate any incongruencies we find to hopefully influence targeted interventions, such as training programs, policy changes, or even community policing initiatives. Overall, investigating data on police stops is an essential step towards creating a fair and just society that ensures equal treatment and protection under the law for all individuals, regardless of their demographic.

Primary Dataset: Police Citations in Houston, Texas

The primary dataset that we used was obtained from the Stanford Open Policing Project, which is a project based in Stanford University that works on collecting and standardizing data on vehicle and pedestrian stops from law enforcement departments across the country. For this project, we looked at the data based in Houston.

This dataset consists of a collection of 2,045,972 stops, providing us with an extensive amount of data to work with. The information collected in this dataset typically includes the date and time of the stop, the address of the stop, the latitude and longitude coordinates, and the geocode of the stop. We also have information on the police beat and police district where the stop was performed.

Additionally, if the stop was a vehicular stop, the dataset includes data on the speed of the vehicle and the speed limit of the area, which can be helpful in understanding the extent to which traffic violations are being targeted by law enforcement officers. We also have information on the color, make, and model of the vehicle. Finally, the dataset includes information on whether a citation was issued and what the outcome of the stop was. One thing to note is that most of these variables (all but three) were categorical, so we had to find unique and innovative ways to analyze and contextualize them.

Secondary Datasets: Census and GIS Data

Census Data

One of the supplementary datasets that our team used was the Census Data of the City of Houston from 2021, which we accessed through the United States Census Bureau website. This dataset gave population estimates of Houston, population percentage estimates of Age and Sex, population percentage estimates of Race and Hispanic Origin, estimates of Population Characteristics, and more. For the purpose of this report, we focused on the population percentage estimates of Race, which gave percentage estimates of White alone, Black or African American alone, American Indian and Alaska Native alone, Asian alone, Native Hawaiian and Other Pacific Islander alone and Two or More Races.

GIS/Beats Data

Another supplementary dataset our team utilized was GIS Data for the City of Houston Police Beats. This dataset contains 118 rows (one for each police beat in Houston) and has columns for police district and area (in square miles) among other geographical characteristics. We paired this dataset with our primary dataset in order to geographically contextualize our findings.

Exploration

The purpose of our exploration was to understand and contextualize the large primary dataset and investigate any trends that might exhibit potential for bias. We did this by examining a few different factors: distribution of citations over time, by demographic, and by location.

For distribution of citations over time, we aimed to see if there were any discrepancies between citations for each race over time, and generally see how citations differed from 2014-2019. We also looked at the frequency

of citations per hour to see if there were any particular patterns, as well as the miles per hour over speed limit per year.

For distribution by demographic, we examined the data in several ways: a waffle chart of citation by race, a table comparing citation proportions for each race to their population proportion in Houston, and a plot of MPH over Speed Limit for Race/Gender.

For distribution by location, we examined this through a population plot of citations by sex and district, as well as a table that displays the number of citations per square mile.

Finally, we combined our investigation of time, demographic, and location in our killer plot, which gave us insight into the distribution of citations by race, gender, and district over time.

Distribution Over Time

First, we began our exploration by looking at the distribution of citations over time. We first visualized citations issued to each race over the years 2014-2020 in order to evaluate the consistency of the distribution of citations over time.
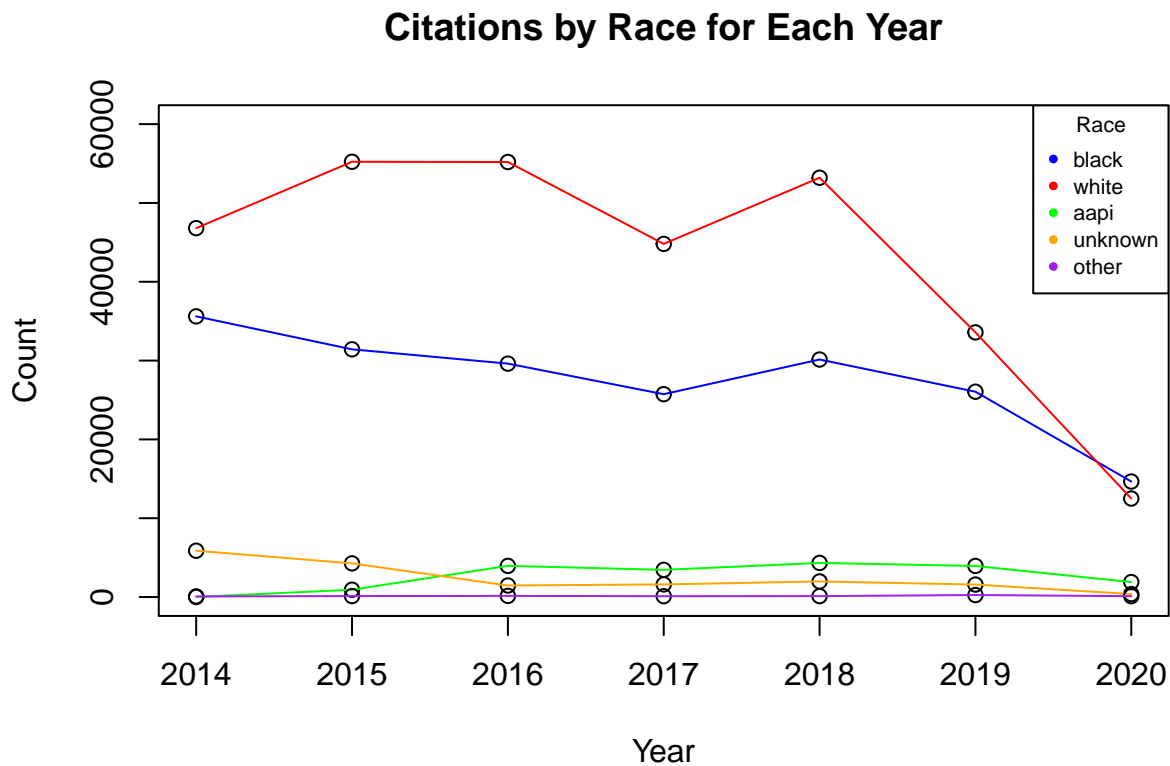
We found that the racial distribution of citations remained fairly consistent: from 2014-2019, the greatest number of individuals issued citations were white, but in 2020, black individuals narrowly surpassed white

individuals as the racial group with the most citations of any racial group. In each year, individuals of unknown race and asian/pacific islanders also received significantly fewer citations than white and black individuals. In 2019 and 2020, a decrease in citations issued to all races can be observed, likely due to the onset of the COVID-19 Pandemic. From this visualization, we did not conclude that there were any significant changes in the relative distribution of citations issued to each racial group over time.

Next in our investigation of citations throughout time, we decided to visualize the number of citations issued throughout the hours of the day.

Frequency of Citations vs Hour in Day
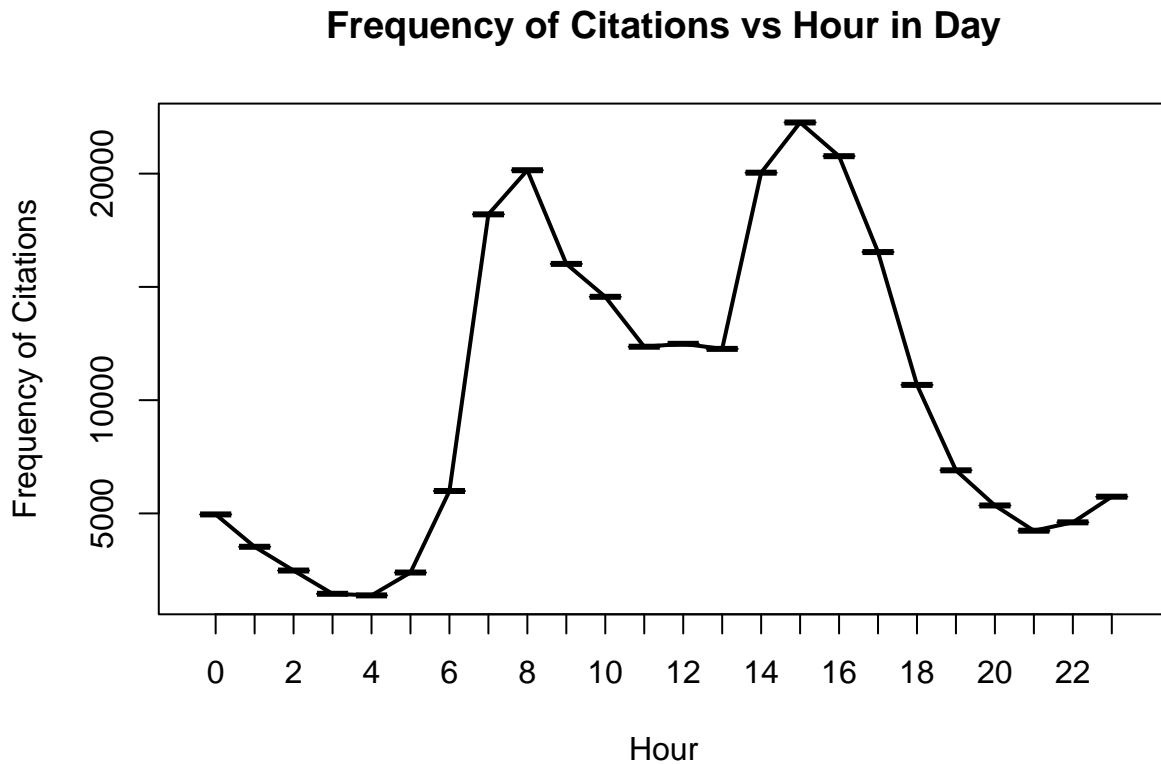
Based on the analysis of the frequency of citations over the hours of the day, it appears that there are two distinct peaks, one around 8 am and another around 3 pm. There are several possible explanations for these peaks. One possible explanation for the peak around 8 am is that it is due to high travel volume during rush hour. Many people are commuting to work at this time, which may increase the likelihood of traffic violations and, consequently, citations. This is further supported by the fact that there is a lull in citations during the super late and super early hours when there is typically less travel.

Another possible explanation for the peak around 3 pm is that it is related to school schedules. Children in K-12 typically get out of school around 3 pm, which may cause a spike in traffic volume as parents rush to pick up their children and return home before rush hour. This increased traffic volume could lead to a higher number of traffic violations and citations.

Distribution by Demographic

Waffle Chart of Citations per Race
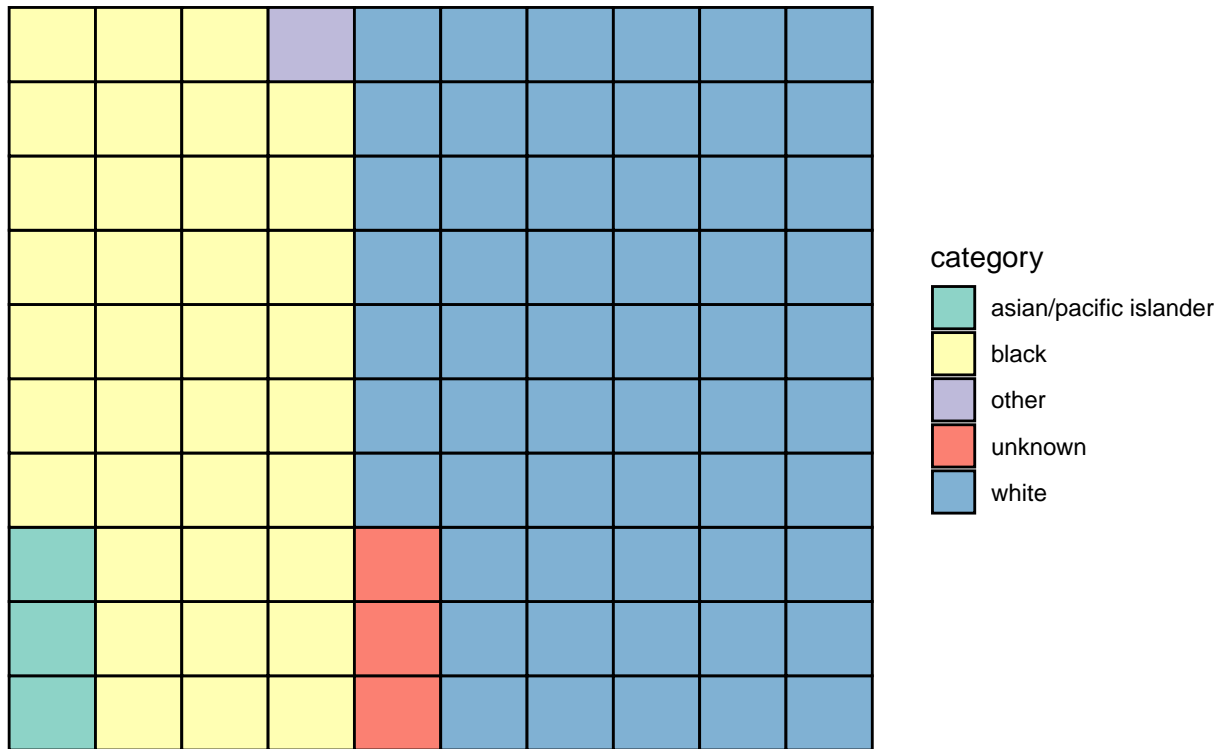
This representation of the percentage of citations by race across all years 2014-2020 provides a visual and geometric way to understand the distribution of citations across different racial groups. The graph shows that white individuals have the highest percentage of citations, followed by black individuals, and then Asian/Pacific Islanders and unknown races.

Following this visualization, our team determined that we would like to compare the distribution of citations by race to the demographics of Houston to see if any racial groups receive a disproportionate amount of citations as compared to their representation within the city of Houston.

Race	Population Proportion	Citation Proportion
white	0.468	0.5672991
black	0.226	0.3638231
other	0.108	0.0017566
asian/pacific islander	0.069	0.0348390

This table shows that the citation proportion for White people is higher than their population proportion in

Houston, indicating that they may be disproportionately cited. Similarly, the citation proportion for Black people is also higher than their population proportion, suggesting that they too may be disproportionately cited. On the other hand, the citation proportion for Asian/Pacific Islander people is lower than their population proportion, indicating that they may be underrepresented in the citations. However, it's essential to keep in mind that the exact reasons for these disparities are not clear, and there may be several factors contributing to them, such as socioeconomic factors or geographical location.

It's also worth noting that the "other" population proportion is significantly higher than the citation proportion. Still, this may be due to the fact that mixed-race individuals are classified as "other," while for citation purposes, people giving citations may mark them as a single race.

MPH over Speed Limit for Citation
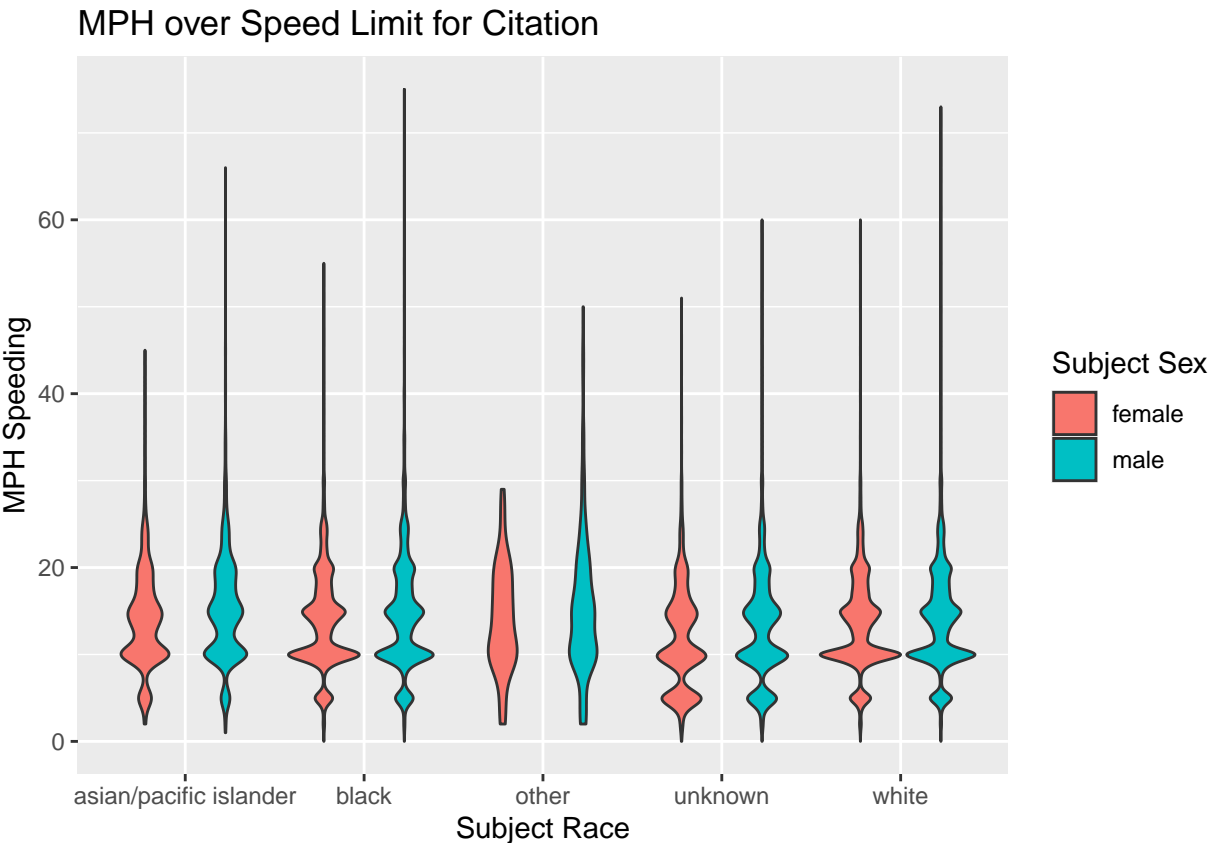
Moving forward, we chose to visualize MPH over Speed Limit for Citation by subject race and sex in order to identify any discrepancies between speeding amounts that warrant a ticket for different races and genders. We observed that the distribution of speeding amounts for individuals of white, black, and unknown race were similar, with the greatest peak around 10 mph, followed by another at 15 mph. Meanwhile, asian/pacific islanders demonstrated a more uniform distribution of speeding amounts from 10-20 mph that tapered off to a maximum of around 30 mph and a minimum of 5 mph. Lastly, individuals of "other" race received citations for speeding in amounts relatively uniformly distributed from 5-30 mph for men and 10-30 mph for

women. Density of speeding amounts for each sex within each race were also very similar, although males of every race besides "other" reached a greater maximum speeding amount than females. This is likely due to differences in driving behaviors between males and females (with males typically having risk tolerance than females), and not evidence of any biases.

Distribution by Location

Citation by Sex by District – Pyramid
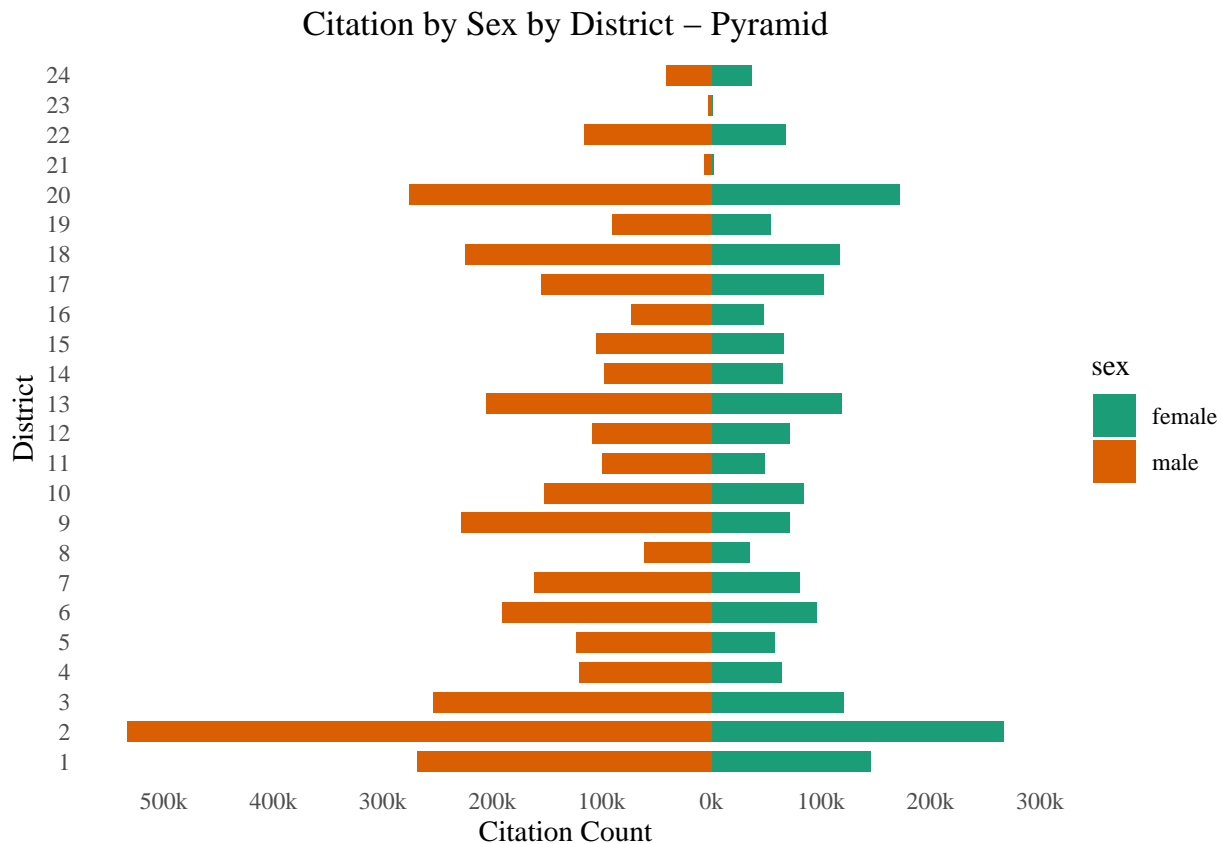
To begin our investigation by location, we made a pyramidal graph of citation count by sex for each district. The x-axis is the amount of citations in thousands, and the y-axis is the police district in Houston. This plot reveals interesting aspects of our dataset regarding the proportion of citations by gender. There are generally more citations for men than women for each district. This could be because men are more likely to commit infractions due to higher risk tolerance, or it could be that the rates of infraction are the same for the genders and women are less likely to be given a citation. We think it is most likely the former influencing the imbalance. While the core reason for the discrepancy is unknown, the unequal distribution is clear to see.

An interesting feature is that Districts 21 and 23 have significantly less citations overall, and that is because they are Hobby airport and IAH airport, respectively. Since it is not residential and technically doesn't have

a population, it is only logical that there are less citations for those districts.

District 2 has the most citations overall, which is one of the most densely populated districts, as it contains Montrose, Greater Heights, West University, and other densely populated neighborhoods within the inner loop of Houston. The inner loop (roughly Districts 1, 2, 9, and 10) have ½ million people in ~90 square miles, so it makes sense to see an increase of citations in those districts.

Next, our team aimed to understand the geographic distribution of citations by calculating the number of citations per square mile within each police beat. We hoped that this investigation would help us identify any beats with an abnormal quantity of citations so that we could narrow down the scope of our investigation for bias. As a part of our investigation, we first found the median, standard deviation, and interquartile range of the number of citations per square mile for all police beats.

Median	Standard Deviation	IQR
853.5304	1877.273	1094.37

We then found the top 5 police beats in terms of number of citations:

| Beats	Citations Per Sq. mi.
1A10	12162.841
2A40	10326.629
2A10	10174.400
10H40	6582.549
10H30	4685.895

Upon receiving the list of 5 police beats with the greatest number of citations, we cross-referenced a map of Houston's police beats to view the location and characteristics of these beats. We found that all five of these beats are adjacent, small, and located in the metropolitan heart of Houston. Additionally, we noted that both beats 1A10 and 2A40 are very small in terms of square mileage, but contain two and one police stations, respectively; proximity to police stations could therefore explain the extremely high citations/square mile values for these beats. Following this, we found the bottom 5 police beats in terms of number of citations:

Beats	Citations Per Sq. mi.
24C30	128.124500
8C40	91.977928
21I50	56.395863
24C40	8.352124
23J40	5.922349

Upon receiving the list of 5 police beats with the lowest number of citations, we once again cross-referenced a map of Houston's police beats to understand the characteristics of these beats. Our team found that these five beats were located in Houston's suburbs and are all relatively large in terms of square mileage, which explains their low citations/square mile values.

Killer Plot

Citation Breakdown by Sex, Race, & Year
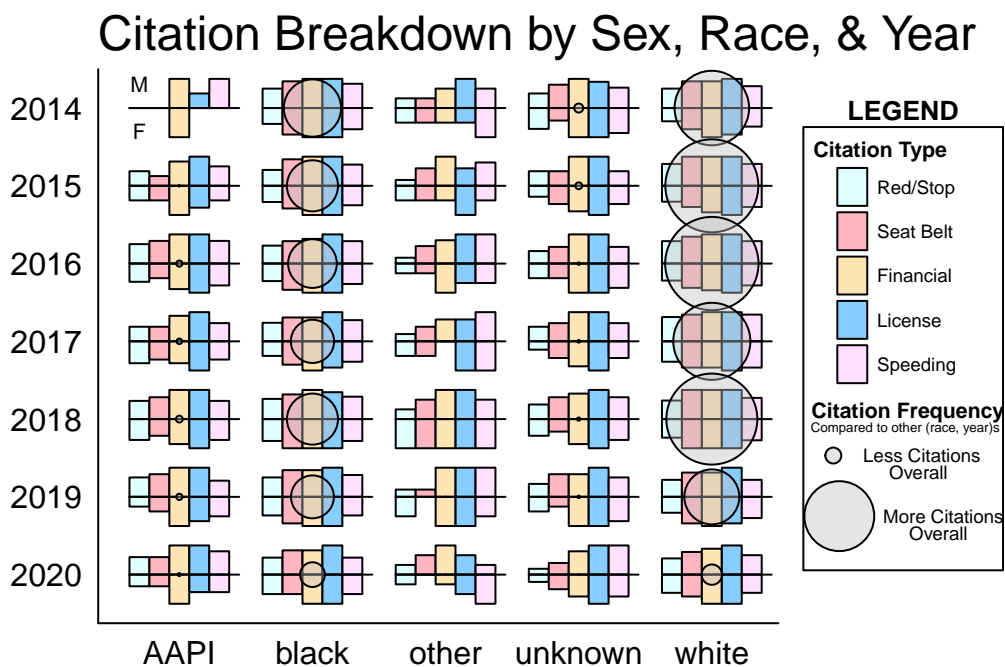
To truly analyze the key factors of our dataset–how citations change with regards to time, demographics, and location, we created a killer plot that could visually represent a combination of all these aspects of our data. We created a way to visualize citation breakdown by sex, race, year, and district all in one plot. NOTE: this plot is the overall killer plot, and is not subdivided by district. We created one in Shiny which cannot be inserted into a PDF in this manner. This plot here encompasses all districts.

On the x-axis, we have the different races in our dataset, while the y-axis are different years. For each (race,year) pair, we created a bar plot where bars above the line represent citations for men, while below the midline represent citations for women. We harvested the free text field of "Citation Type" from our dataset to split the citations into 5 main categories–running a red light or stop sign, failure to wear a seatbelt, failure to establish financial responsibility, operating a vehicle with an invalid driver's license, and driving over the speed limit. The height of the bars represent the frequency of that citation (compared to the other citation types) for that race that year. The gray circles on the killer plot have a radius that is relative to the greatest number of citations for any race, year combination. This plot is comparative by nature and does not have any objective representations of exact numbers of citations, just how they compare between different years and demographics. All of our important categorical information is represented here in one single killer plot.

Citation-Specific Analysis

Following exploration, our team had a more thorough understanding of the dataset and potential areas that we could investigate for bias. Specifically, our team took interest in the role of race in citations and sought to further analyze trends related to subject race.

First, looking at our data, our team noticed that there were 5 main types of violations: speeding, invalid license, failure to establish financial responsibility, failure to wear a seat belt, and running a stop sign/red light. We wanted to analyze the racial breakdown of each type of citation in order to see if any racial group disproportionately received any type of citation.

Subject Race	Speeding	Invalid License	Financial Responsibility	Seat Belt	Stop Sign/Red Light
white	120836	74202	36750	18368	33214
black	61548	58105	39280	11719	17689
NA	31988	51950	26702	5387	14812
asian/pacific islander	8311	2086	1155	549	2928
unknown	7620	3395	1768	532	1823
other	443	127	78	40	140

For nearly all types of citations, when considering subjects with defined races, white individuals received the most citations, followed by black and AAPI individuals; this aligns with the racial breakdown of the city of

Houston.

However, black individuals received the most citations for failure to establish financial responsibility, which refers to the inability of the subject to provide proof of insurance. Typically, this citation should only be issued given that the subject has committed some other infraction that necessitates police interaction and request for proof of insurance. So, suspicion may be raised around citations that record failure to establish financial responsibility as the sole infraction, as the officer has not recorded any indication of why the individual was pulled over in the first place. Citations that only reference failure to establish financial responsibility may be useful in identifying possible racial bias, as the officer may have pulled over the individual based on their appearance/race, since they did not indicate any other offense on the citation. This led our team to investigate what proportion of citations for individuals of all races resulted solely from failure to establish financial responsibility.

Black	White	AAPI	Unknown	Other	Unrecorded (N/A)
0.0267934	0.0165184	0.0025123	0.0026984	0.000131	0.0361138

Of all recorded races, the proportion of total citations that mention failure to establish financial responsibility as the sole violation is the highest for individuals of unrecorded race (N/A), followed by black and white individuals and unknown, asian/pacific islanders, and other. It is worth noting that the proportion of financial responsibility citations for unknown race, black, and white individuals were about eighteen, ten, eight times the proportion of these citations for those of AAPI race, respectively. Furthermore, the extremely low proportion of individuals of "Other" race receiving citations solely for failure to establish financial responsibility can be attributed to the extremely low number of citations with race marked as "Other."

After receiving this data, our team ran hypothesis tests at a 95% confidence interval in order to identify which races have a proportion of citations issued solely for failure to establish financial responsibility that is higher than the same proportions for other races. The proportion of these citations for unrecorded race (N/A) was greater than the proportion for all other races by a statistically significant amount. Additionally, the proportion for black individuals was greater than those of white, unknown, and AAPI race by a statistically significant amount. White individuals also had a proportion that was greater than that of unknown and AAPI race.

Excluding unrecorded race, these findings raise suspicion for bias against black individuals as compared to all other races and for white individuals against individuals of AAPI and unknown race. A high proportion of individuals with unrecorded race receiving citations solely for failure to establish financial responsibility

can also be suspicious, as this indicates that officers failed to, whether consciously or unconsciously, record the race of these individuals.

Next, we wanted to investigate if different races and sexes received speeding citations equally, or if certain groups were given citations with different frequencies or for different severity of speeding. We started by averaging the entire dataset to create a baseline.

Mean	Median	SD
14.04268	14	6.198856

Subject Race	Total	Mean	Median	SD
asian/pacific islander	8020	14.63819	14	5.963141
black	54344	14.25130	14	6.072997
other	417	14.87624	14	7.565727
unknown	7056	12.62774	12	6.259297
white	109576	13.89174	14	6.190181
NA	26060	14.41970	14	6.442300

For the breakdown by race, there is not much evidence of bias on this level.

One thing to note is the 'unknown' category. Police officers fill out these citation reports without asking the subject about their race(s) or ethnicity(s). All of these reported races are from the perspective of the police officers and therefore it is difficult to know the true breakdown. 'Unknown' is different from N/A or unrecorded, because the officers did fill out the race section, but itr is an admission that they don't actually know the race of the subject.

Here, the unknown could be evidence of bias. We see that for people the officers are unsure about (and those they didn't ask to clarify), they are given citations for slower speeds (instead of being let off with a warning, for example). This is unproved and could also be explained by the smaller number of unknowns in the sample.

Subject Sex	total	mean	median	sd
female	78617	13.59924	13	6.047648
male	126747	14.31904	14	6.275959

Subject Sex	total	mean	median	sd
NA	109	13.49533	13	5.387571

Here we see that men and women are ticketed for roughly the same speeds, but there are far more men ticketed than women in our dataset. This is not equal to the proportion of men and women in Houston, so this is either evidence that men speed more often than women (due to higher risk tolerances or different driving behaviors), or that officers are more likely to give men citations and let women off with a warning, or that women are simply caught speeding less often.

Subject Race	Subject Sex	Total	Mean	Median	SD
asian/pacific islander	female	2759	13.82355	14	5.573562
asian/pacific islander	male	5259	15.06543	15	6.114859
black	female	24408	13.86948	14	5.531157
black	male	29928	14.56284	14	6.465650
other	female	97	13.95699	14	6.035968
other	male	320	15.15113	14	7.954284
unknown	female	2450	11.79511	10	5.998144
unknown	male	4606	13.06692	13	6.349574
white	female	39745	13.49757	13	6.404608
white	male	69814	14.11699	14	6.052549
NA	female	9158	13.71900	13	5.829660
NA	male	16820	14.80848	14	6.728071

When analyzing both race and gender, the key observation to note is that black women are given speeding citations at a very similar rate as black men, while women of all other races have speeding citation counts of nearly half of their male counterparts. Assuming that women of all races speed roughly the same amount, this break in the pattern could be evidence of bias, as it deviates so much from the norm.

Conclusion

In the end, we found preliminary evidence of bias that can be used as a starting point for future investigation and can potentially guide future policies and interventions. These preliminary biases were seen in the

differences of proportions by race for failure to show financial responsibility as the only citation, particularly for black individuals. Additionally, black women were given speeding citations at roughly the same rate as black men, while the rate of speeding citations for women of all other races was consistently around half of their male counterparts. These peculiarities warrant further investigation, as our analysis unfortunately has some key limitations. Most importantly, our dataset has a lot of messy and missing data. Many fields were free text which is very prone to human error. Many observations have missing data in the race field, and we were not able to discern if there was any pattern to the missingness or if it was missing at random. This is important because if a police officer were to intentionally leave the race category blank when giving a questionable citation, their bias could be obfuscated from our analysis. The only way to definitively diagnose bias is with more data with less missing values in the race category. If that dataset did exist, then the severity of bias could change. It could either become stronger, which would corroborate our findings, or disappear completely, which would mean that what we found was simply an anomaly of the pattern of missingness in this dataset.

Despite this limitation, we believe dedicating time and effort to projects like this is important on chance that any bias is illuminated. Holding our police systems accountable and ensuring accountability is fundamentally important to ensure our societal infrastructure functions properly and fairly for all people. It is an essential step to ensure equal treatment and protection under the law for all individuals, regardless of their demographic.