Investigating Police Citations in Houston

Draft 4, Group 3

Grace Wang, Virginia Baskin, Siddhi Narayan

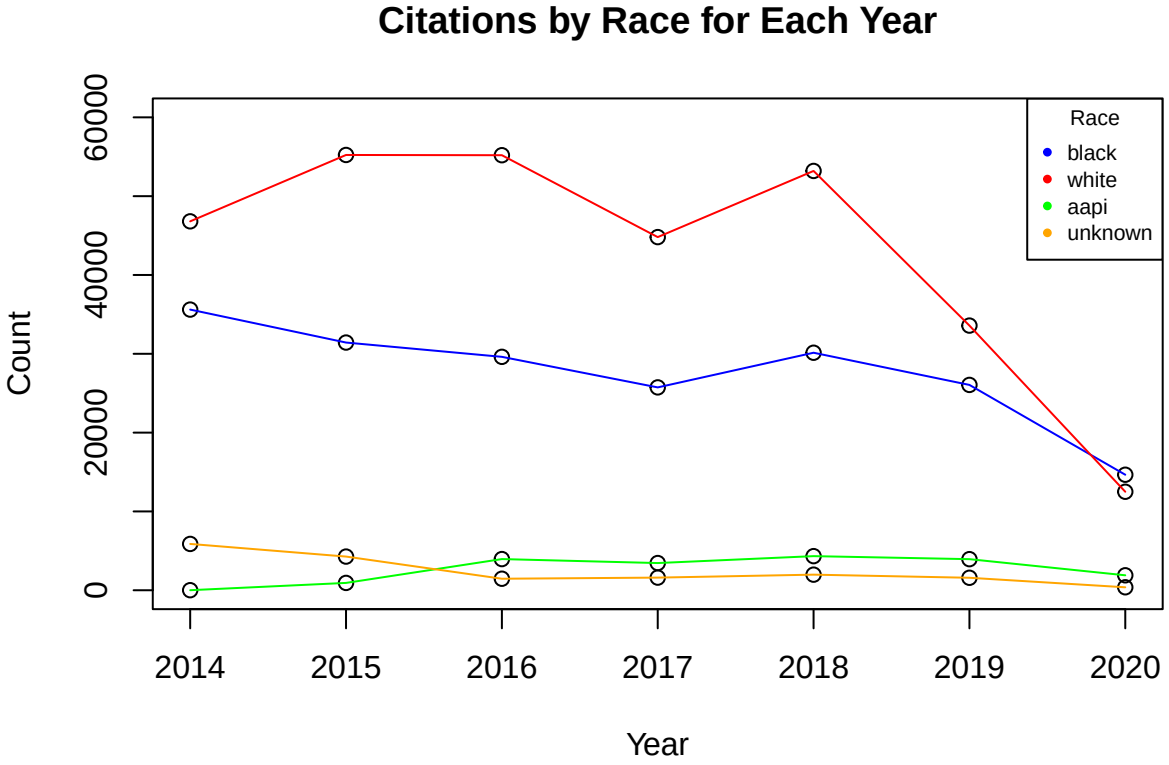
2023-03-30

Introduction

Our team's dataset is sourced from the Stanford Open Policing Project and contains information on police stops that result in citations in Houston, TX from 2014-2020. The data contains a little over 2 million rows. The fields of the dataset include (non-exhaustively): the date, time, location, latitude, longitude of the stop, the police beat of the officer, district, the subject's race, sex, the type of violation they received, and the vehicle make, model, and color. Our team approached this dataset with a special interest in investigating trends regarding subject race, sex, and speeding amounts. Ultimately, we seek to investigate any factors that particularly dispose individuals towards receiving a citation.

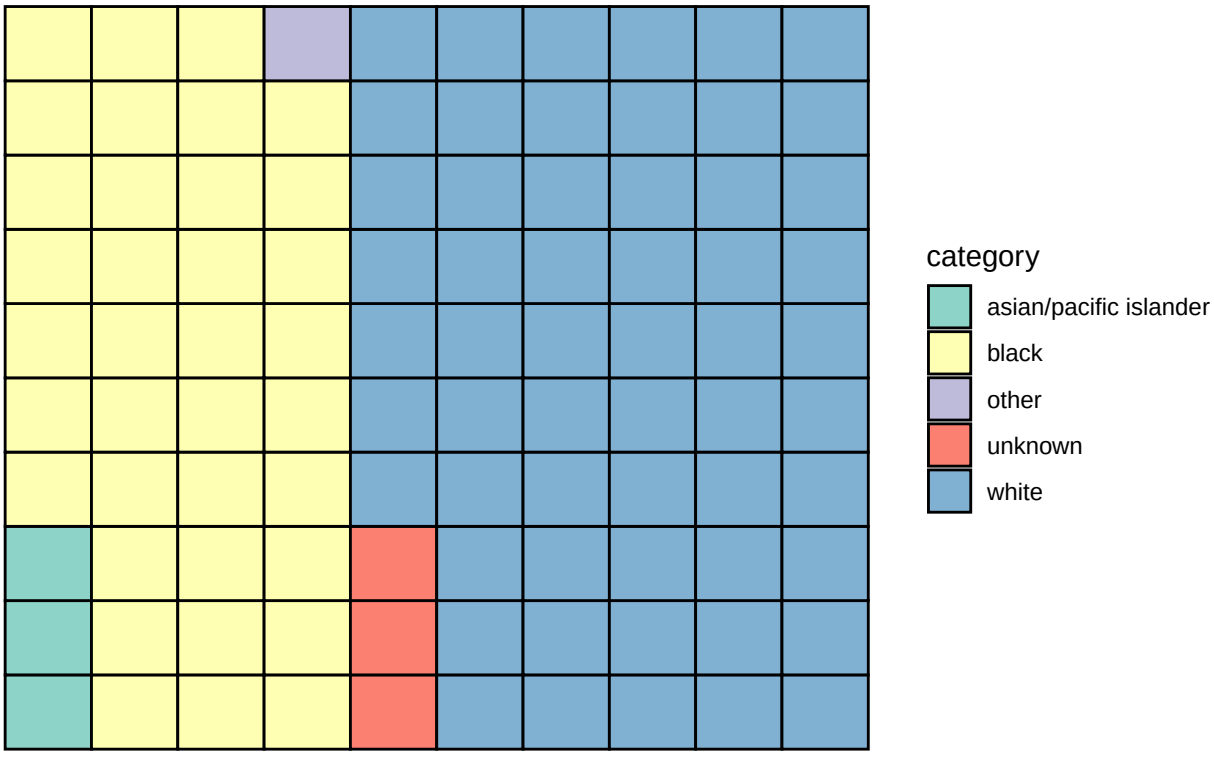
Data Exploration - Graphs

We utilized various visualization techniques in order to comprehensively analyze every feature of the problem space and maximize various perspectives on the dataset. Our dataset has a lot of free text, which is difficult to visualize in the typical way. We visualized every field that was not unstructured free text in some way or another.

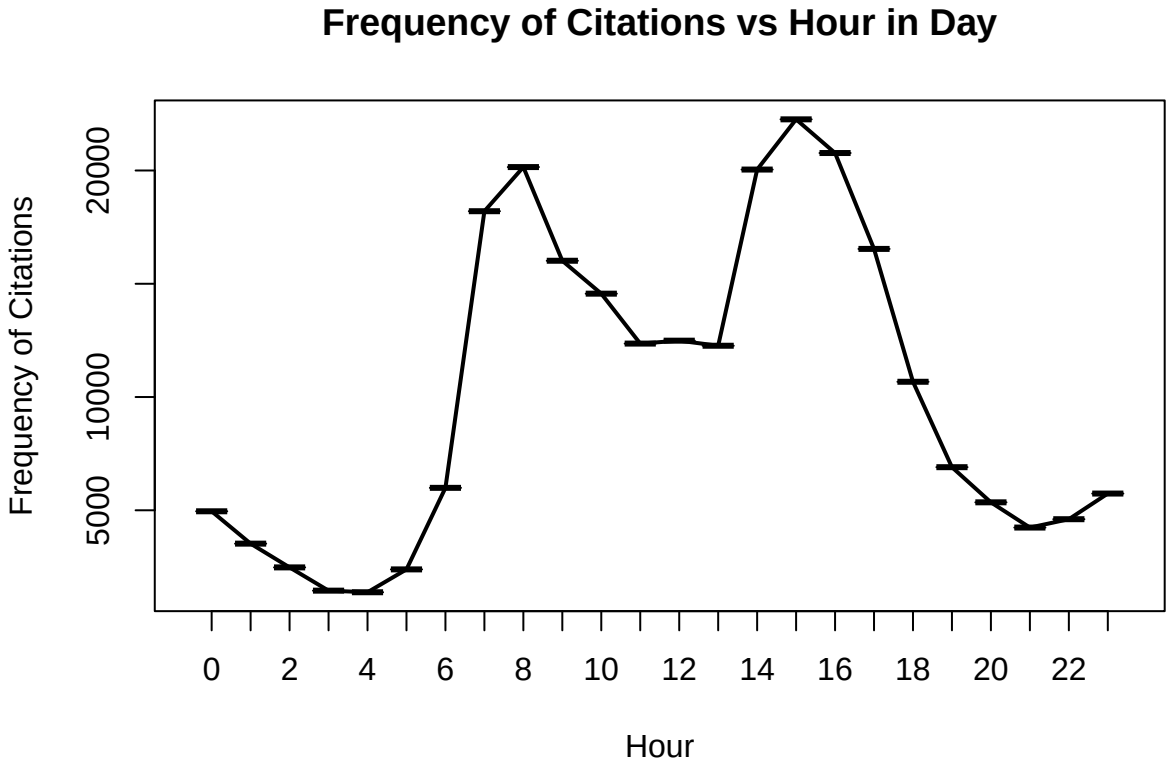


First, we turn our attention to the total individuals of each race issued citations in each year 2014-2020. From 2014-2019, the greatest number of individuals issued citations were white, but in 2020, black individuals narrowly surpassed white individuals as the racial group with the most citations of any racial group. Overall, individuals of unknown race and asian/pacific islanders received significantly fewer citations than white and black individuals. In 2019 and 2020, a decrease in citations issued to white individuals can be observed. This data is not sufficient for drawing conclusions of racial bias in the citation process; more investigation is necessary into confounding factors.

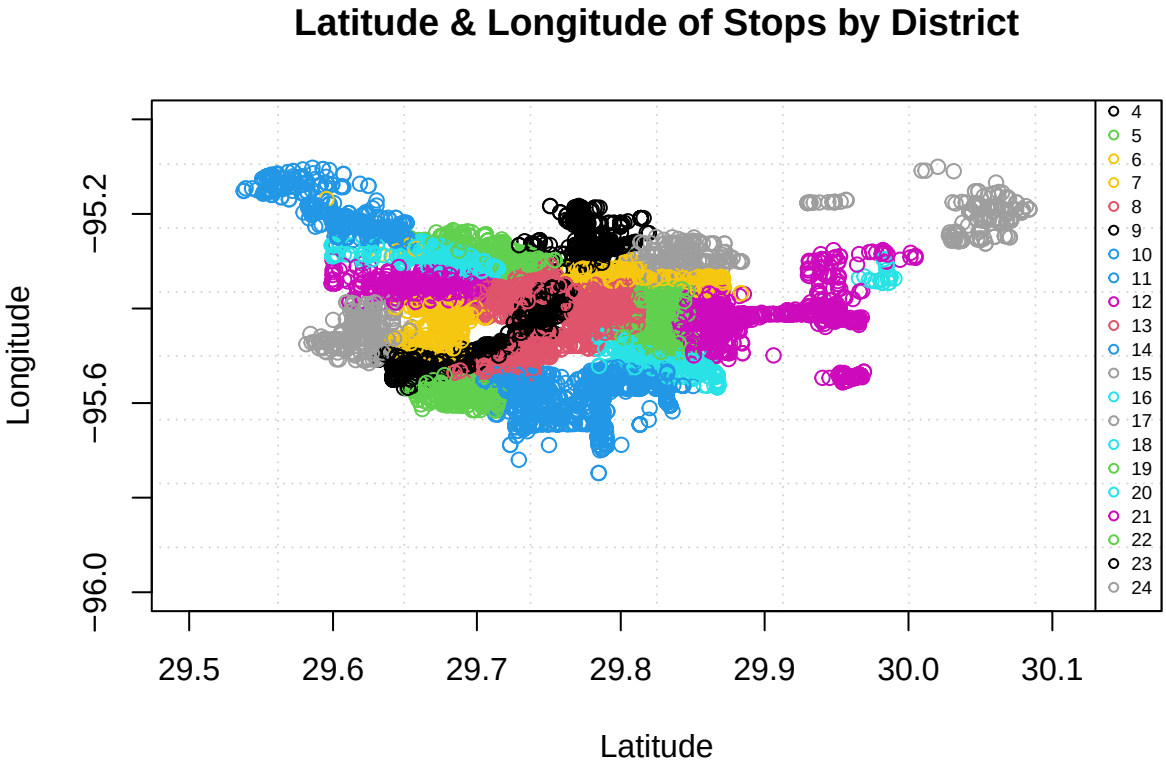
Waffle Chart of Citations per Race



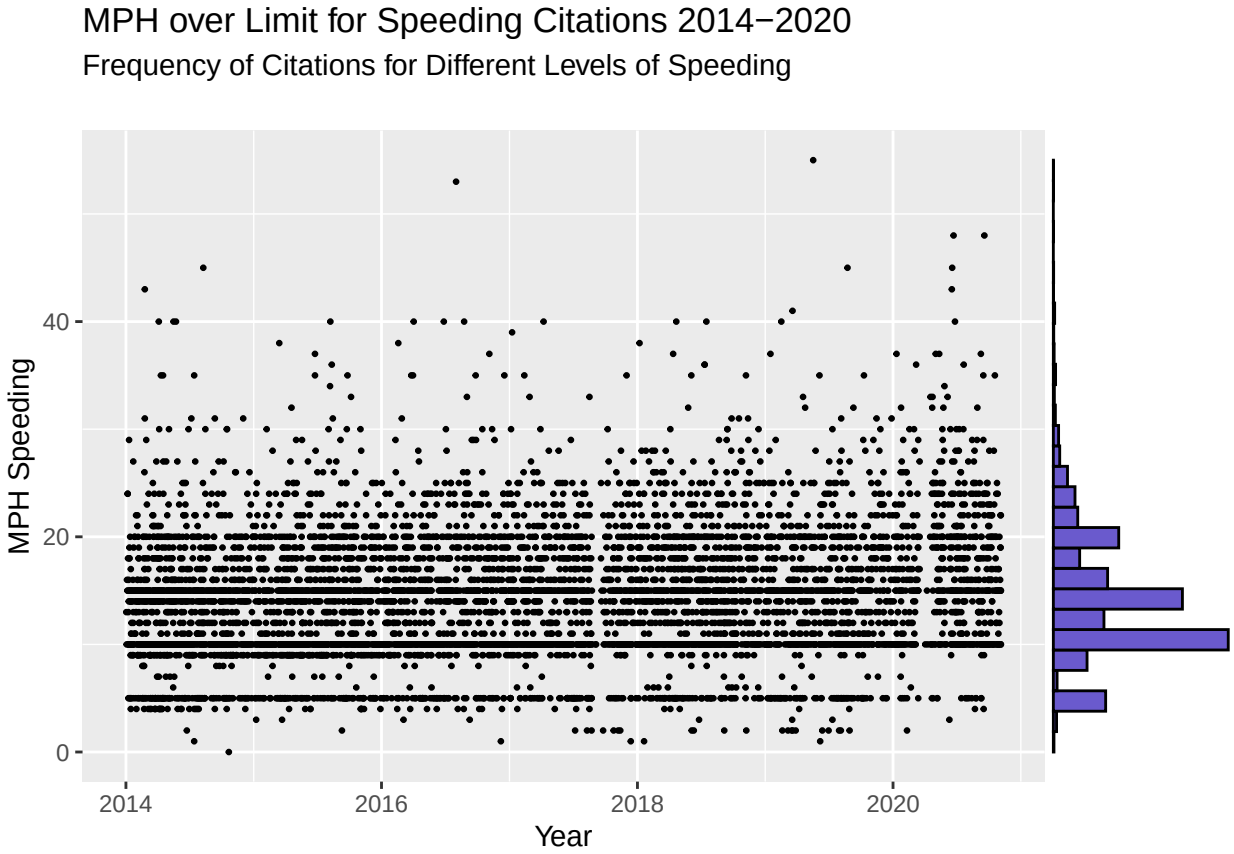
Next, we will look at the % of Citation by Race across all years 2014-2020 in a more geometric representation of the percentages. As we can see, white individuals have the highest percentage of citations, followed by black individuals, and then Asian/pacific islanders, then unknown races. This is somewhat proportional to the actual demographics of Houston (according to the U.S. census) – the White and Black percentages of citations are slightly higher than their population in Houston, while the proportion of citations for Asian individuals is slightly lower than the proportion of Asian people in Houston. This graph is a fun (and new, for us) way to visualize which races make up what proportions in our dataset.



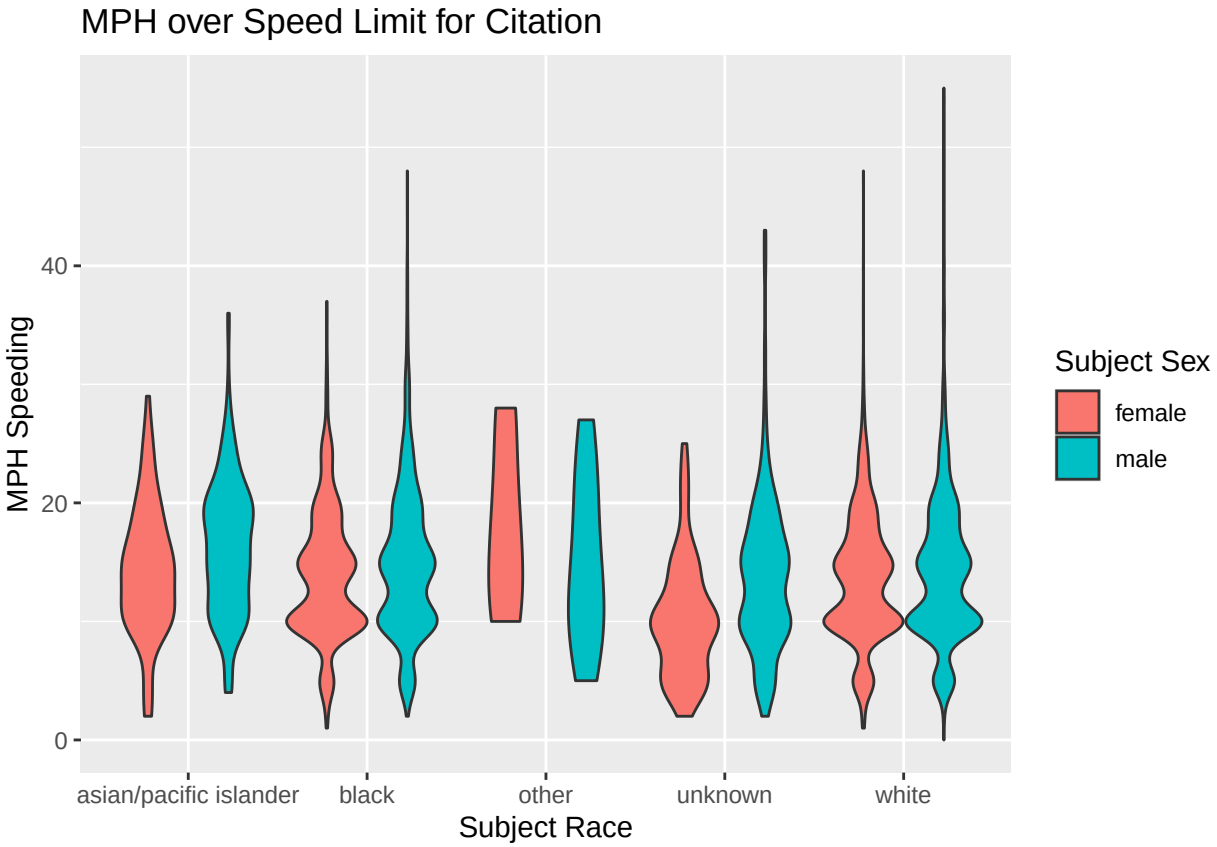
Now we will look at the Frequency of citations over the Hours of the Day. Citations are most common around 8 am, and around 3 pm. There may be a higher proportion of citations around 8 am because of high travel due to work. We are not quite sure why another peak is found at 3 pm, it may be people returning from school and/or work. Children in K-12 grade usually get out of school around 3pm. This peak could be correlated with people speeding to pick up/return home with their children before rush hour. We also see less citations during the super late/super early hours – likely due to less travel as a whole.



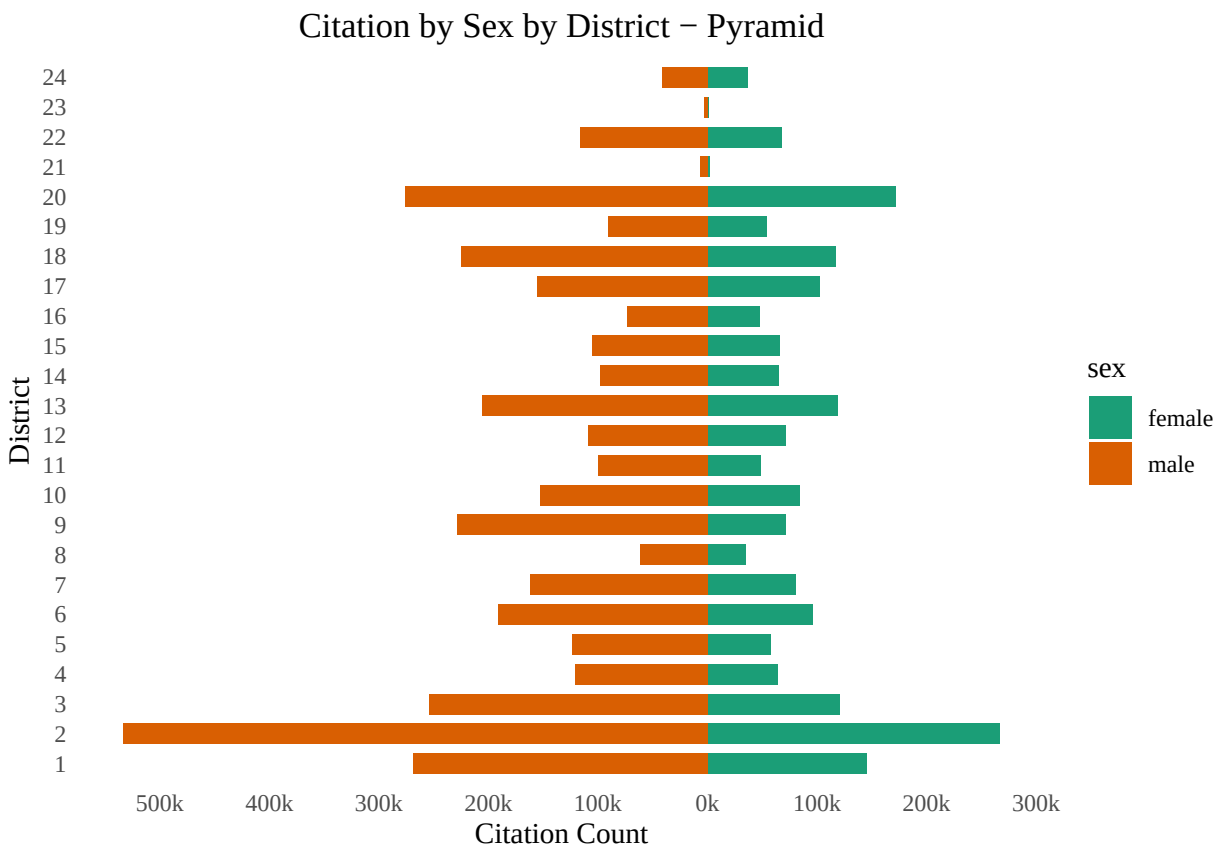
Next, we want to explore the spatial features of our data. We have both the latitude and the longitude of each stop, along with the district and beat. We knew that beat meant a motorized police unit that patrols a specific territory, but "district" in this context is not as clear, as it could mean multiple things. We investigated this by creating a scatterplot of the latitude versus the longitude of stops and colored it by district. We see very clear spatial grouping for each district, which means districts are not a police-defined feature but a location feature. From this plot, we can ascertain that district refers to Houston's subdistricts. [As seen here: https://www.houstontx.gov/police/pdfs/hpd_beat_map.pdf]



This plot investigates citations that are given for speeding. Namely, by how much were people speeding over the limit to have been issued a citation. The y axis is mph over the speed limit (as recorded on the citation). The x axis is the year/general time from the citation occurred in, and there does seem to be some evidence of temporal trends. There is a vertical line before 2018 where it seems like there were less tickets given overall, which could be a number of things that we could look into. It could be something about how the data was collected, if there was some data lost at any point, or maybe HPD deprioritized patrolling for speeding during that time period. There is no way to tell from just looking at the graph. Another interesting feature is that there seems to be less citations given for speeding less than 10mph over the speed limit 2020 and onward. It is also interesting that there are speeding citations given for going as little as 1-5 mph over the speed limit. The most frequent citations were given around 10-15 mph over the speed limit, as seen by the histogram of frequent cited mph over speed limit on the left margin.



Moving forward, we chose to visualize MPH over Speed Limit for Citation by Subject race and sex. We visualized this with violin plots, which demonstrate the density and distribution of the data. We observed that each race except for "unknown" had a fairly similar distribution of speeding amounts, with the greatest peak around 10 MPH. Distributions of speeding amounts for each sex within each race were also very similar, although males of every race reached a greater maximum speeding amount than females.



We created a "population pyramid" of citation count by district and by sex. 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 more citations for men than women for every single districts. This could be because men are more likely to speed, or that women speed the same but are not given citations as often. The core reason for the dependency is unknown, but the unequal count by sex is clear to see. Also, as found before in other graphs, district 21 and 23 have the least amount of citations as they are both airports (IAH and Hobby), and people are simply unable to speed in the same capacity in airport roads as they can on a highway. District 2 has the most citations, and this corresponds to the Houston's Greater Heights district, roughly.

Questions

Investigating aspects of our data with supplemental data as well.

1- How does the racial breakdown of citations compare to ther racial breakdown of Houston?

From our main dataset, we visualized the breakdown of citations per race. Our group wanted to see how the breakdown of citations per race compared to the breakdown of race in Houston.

race	citation_prop	population_prop
asian/pacific islander	0.0348390	0.069
black	0.3638231	0.226
other	0.0017566	0.108
white	0.5672991	0.468

From this table, we can see that the population proportion of White people is lower than their citation proportion, the population, the population proportion of Black people is lower than their citation proportion, and the population proportion of Asian/Pacific Islander people is higher than the citation proportion.The 'other' population proportion is significantly higher than the citation proportion, but this may be because we signified mixed-race individuals as others, while for citation purposes, people giving citations may mark them as a single race.

2- What is the race breakdown of each type of citation?

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.

Speeding Citations by Race:

subject_race	count
white	120836
black	61548
NA	31988
asian/pacific islander	8311
unknown	7620
other	443

Invalid License Citations by Race:

subject_race	count
white	74202
black	58105
NA	51950
unknown	3395
asian/pacific islander	2086
other	127

Failure to Establish Financial Responsibility Citations by Race:

subject_race	count
black	39280
white	36750
NA	26702
unknown	1768
asian/pacific islander	1155
other	78

Seat Belt Citations by Race:

subject_race	count
white	18368

subject_race	count
black	11719
NA	5387
asian/pacific islander	549
unknown	532
other	40

Running a Red Light/Stop Sign Citations by Race:

subject_race	count
white	33214
black	17689
NA	14812
asian/pacific islander	2928
unknown	1823
other	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.

3- What proportion of citations for individuals of each race resulted solely from failure to establish financial responsibility?

```
## Warning: Closing open result set, pending rows
```

subject_race	fin_fail_percentage
asian/pacific islander	0.0209144
black	0.0455865
other	0.0214362
unknown	0.0240289
white	0.0239488

Of all recorded races, the proportion of total citations that mention failure to establish financial responsibility as the sole violation is the highest for black individuals. This proportion is about 0.02 higher for black individuals, while all other races share a similar proportion around 0.02.

4- What is the average speeding amount that results in a citation? For each racial group? For each gender?

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.

All People (baseline):

avg_diff	std_diff
14.04268	6.198856

By Race:

subject_race	count	avg_diff	std_diff
unknown	7056	12.62774	6.259297
white	109576	13.89174	6.190181
black	54344	14.25130	6.072997
NA	26060	14.41970	6.442300
asian/pacific islander	8020	14.63819	5.963141
other	417	14.87624	7.565727

One thing to note is the 'unknown' category. Police officers fill out the reports without asking the subjects they pulled over about their race or ethnicity. All of these reported races are from the perspective of the police officers and therefore it is difficult to know the true breakdown.

Here, the unknown could honestly be evidence of bias. We see that for people the officers are unsure about their race (and that they do not ask to clarify), they are given citations for slower speeds (instead of being let off with a warning).

By Gender:

subject_sex	count	avg_diff	std_diff
NA	109	13.49533	5.387571
female	78617	13.59924	6.047648
male	126747	14.31904	6.275959

Men and women are ticketed for roughly the same speeds, bit 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, 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, but speed just as frequently.

By Race AND Gender:

subject_race	n	avg_diff	std_diff
asian/pacific islander	2	10.50000	0.7071068
unknown	2450	11.79511	5.9981438
NA	82	13.03750	4.8220059
unknown	4606	13.06692	6.3495743
white	39745	13.49757	6.4046078
NA	9158	13.71900	5.8296601

subject_race	n	avg_diff	std_diff
asian/pacific islander	2759	13.82355	5.5735618
black	24408	13.86948	5.5311567
other	97	13.95699	6.0359682
white	69814	14.11699	6.0525494
white	17	14.47059	6.9112015
black	29928	14.56284	6.4656499
NA	16820	14.80848	6.7280708
asian/pacific islander	5259	15.06543	6.1148588
other	320	15.15113	7.9542841
black	8	16.75000	6.9641941

Unknown race females have the lowest average amount speeding of 11.8mph and median of 10mph, the lowest of any group.

Since all of the race and sex data is reported by the officers, if there were bias, it could be easily hidden in the "unknowns" or "NAs" in the data. We are able to see that men get ticketed with a much greater frequency than women, but between race there are not many conclusions we can make. There is one exception, however. For most races it is true that men are ticketed two times as much as women of the same race, with the exception being Black people. Black women are ticketed almost as much as black men (24.4k and 29.9k, respectively). This could be evidence of bias against black women.

5- What time of day do citation happen?

Our group decided to look at the frequency of citations by hour of day.

The lowest number of citations occur around 04 hr.

The highest number of citations occur around 15 hr.

The highest number of citations occur at 15, 16, 08, 14, 07 hrs.

We can see that the highest number of numbers of citations occur, in descending order, at 8am, 3pm, 2pm, 4pm an 1pm. Most of these times are in the afternoon, except for the 8am time. There might be a large number of citations at around 8am because this is the time that most people are travelling to work in the morning. Because of the morning rush and the need to get to work on time, people may be more likely to speed/commit citation-able offences at this time.

The rest of the times where the highest number of citations are given are in the afternoon, between 1pm - 4pm. This may also be attributed to rush hour traffic again, and things like school dismissal – police may be more vigilant around school zones, while there may be more people on the roads due to schools getting out.

The smallest number of citations occur around 4am. This may be due to a lack of people on the road, because it is very late at night/early in the morning.

6- What is the number of citations per square mile for each beat?

Next, our team aimed to understand the geographic distribution of citations by calculating the number of citations per square mile within each beat. To aid with our investigation, we found the median, standard deviation, and interquartile range of the number of citations per square mile. We also looked into the 5 beats with the greatest amount of citations per square mile and the 5 beats with the least amount of citations per square mile.

Beats	Citations_per_sqm
10H10	615.617916
10H20	842.137995
10H30	4685.895340
10H40	6582.548826
10H50	1158.795842
10H60	833.519443
10H70	647.610375
10H80	604.317245
11H10	1069.608634
11H20	957.234512
11H30	1363.805569
11H40	144.299219
11H50	352.687128
12D10	672.679776
12D20	863.694304
12D30	495.866331
12D40	268.411076
12D50	133.433695
12D60	217.824820
12D70	337.008015
13D10	4477.081804
13D20	919.218454
13D30	555.948556
13D40	256.272850
14D10	963.235550
14D20	680.634565
14D30	605.177761
14D40	505.879549
14D50	244.932318
15E10	483.681746
15E20	1118.351941
15E30	603.369753
15E40	724.126680
16E10	647.752013
16E20	364.940163
16E30	628.020271
16E40	178.199509
17E10	2451.664954
17E20	1234.573216
17E30	1553.318267
17E40	1491.548348
18F10	1121.597323
18F20	1149.749903
18F30	2473.102159
18F40	1676.516949
18F50	2580.999599
18F60	1690.624784
19G10	1863.000144
19G20	903.957485
19G30	957.584422
19G40	483.759268
19G50	834.422535

Beats	Citations_per_sqm
1A10	12162.841309
1A20	3323.597182
1A30	2158.945209
1A40	618.733371
1A50	1609.898022
20G10	2327.782508
20G20	1456.252945
20G30	2768.916462
20G40	705.961597
20G50	843.366402
20G60	1446.695246
20G70	804.083934
20G80	219.842977
21I20	1576.738624
21I30	169.437679
21I40	758.262625
21I50	56.395863
21I70	2879.010114
22B10	2129.993215
22B20	934.169525
22B30	1352.571909
22B40	254.656360
23J40	5.922349
23J50	212.816555
24C10	284.351490
24C20	371.933413
24C30	128.124500
24C40	8.352124
24C50	193.005562
24C60	149.746316
2A10	10174.399912
2A20	1915.613313
2A30	1663.167136
2A40	10326.628732
2A50	1566.597958
2A60	2123.661645
3B10	1605.703183
3B30	562.295824
3B40	1710.520011
3B50	4631.942681
4F10	1596.445050
4F20	688.412885
4F30	303.380461
5F10	1539.058496
5F20	1050.695094
5F30	793.968514
6B10	2196.933892
6B20	761.662993
6B30	553.798349
6B40	1452.020122
7C10	1149.992604
7C20	825.600494

Beats	Citations_per_sqm
7C30	2246.082434
7C40	1269.943917
8C10	1269.519313
8C20	448.008663
8C30	294.866264
8C40	91.977928
8C50	236.932896
8C60	165.604809
9C10	394.485175
9C20	1022.907327
9C30	992.679977
9C40	1176.442991

Statistics on number of citations/square mile:

median	standard deviation	IQR
853.5304	1877.273	1094.37

Top 5 Beats ranked by citations/square mile:

```
## Warning: Closing open result set, pending rows
```

beats_c	citations_per_sqm
1A10	12162.841
2A40	10326.629
2A10	10174.400
10H40	6582.549
10H30	4685.895

Upon comparison to a map of Houston's police beats, we saw that all five of these beats are adjacent, small, and located in the 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.

Bottom 5 Beats ranked by citations/square mile:

beats_c	citations_per_sqm
23J40	5.922349
24C40	8.352124
21I50	56.395863
8C40	91.977928
24C30	128.124500

Referencing a map of Houston's police 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.

Map Reference: https://www.houstontx.gov/police/pdfs/hpd_beat_map.pdf

Conclusion

In our investigation, we scanned for evidence of racial bias but did not found statistically significant instances of bias. For example, each type of citation had a racial breakdown similar to the overall population and citation breakdown. We also found that beats with the greatest proportion of citations to square mileage are those located in the heart of Houston that are densely populated with police stations, while the beats with the lowest ratio were suburbs with large square mileage at a greater distance from stations. There are interesting observations like black men and women are given speeding citations approximately at the same rates, while all other races men are given citations almost 2x as much. However, it is a stretch to confidently draw the conclusion of bias from a one off finding like that. We are excited to hopefully receive more data from HPD via FOIA request to bolster this analysis from beyond what is available publicly.