Investigating Police Citations in Houston

Draft 3

Grace Wang, Virginia Baskin, Siddhi Narayan

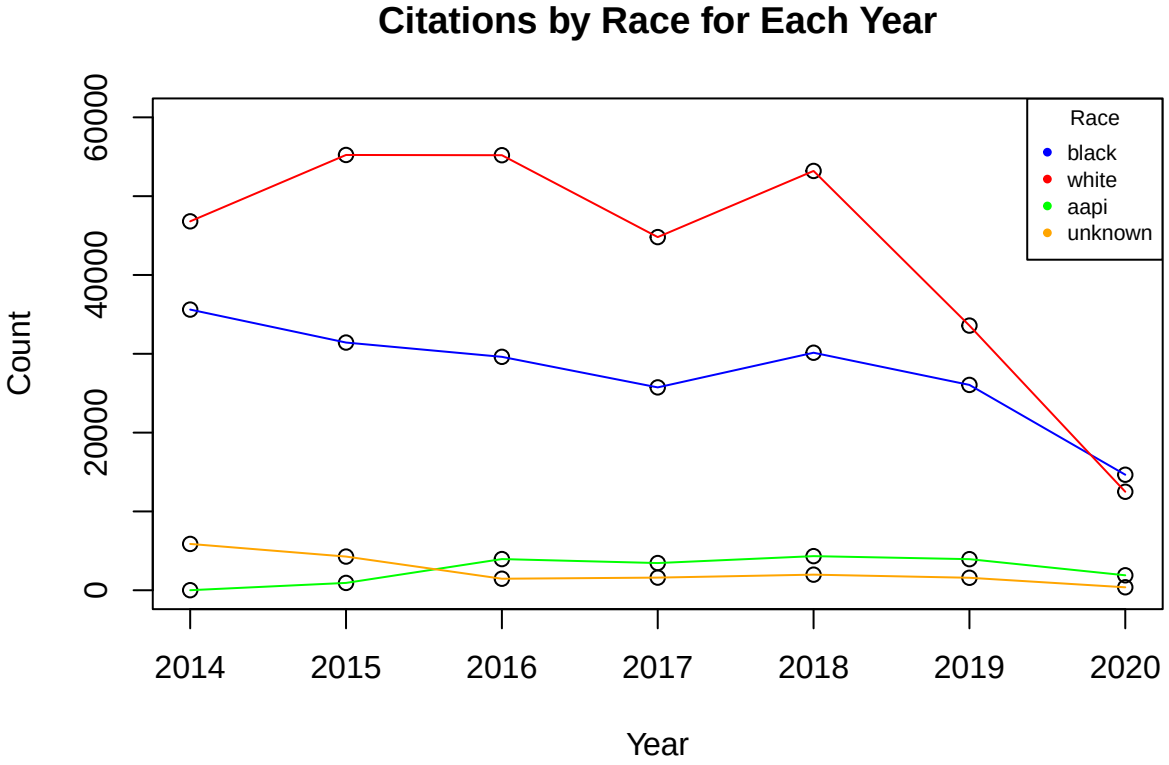
2023-03-07

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.

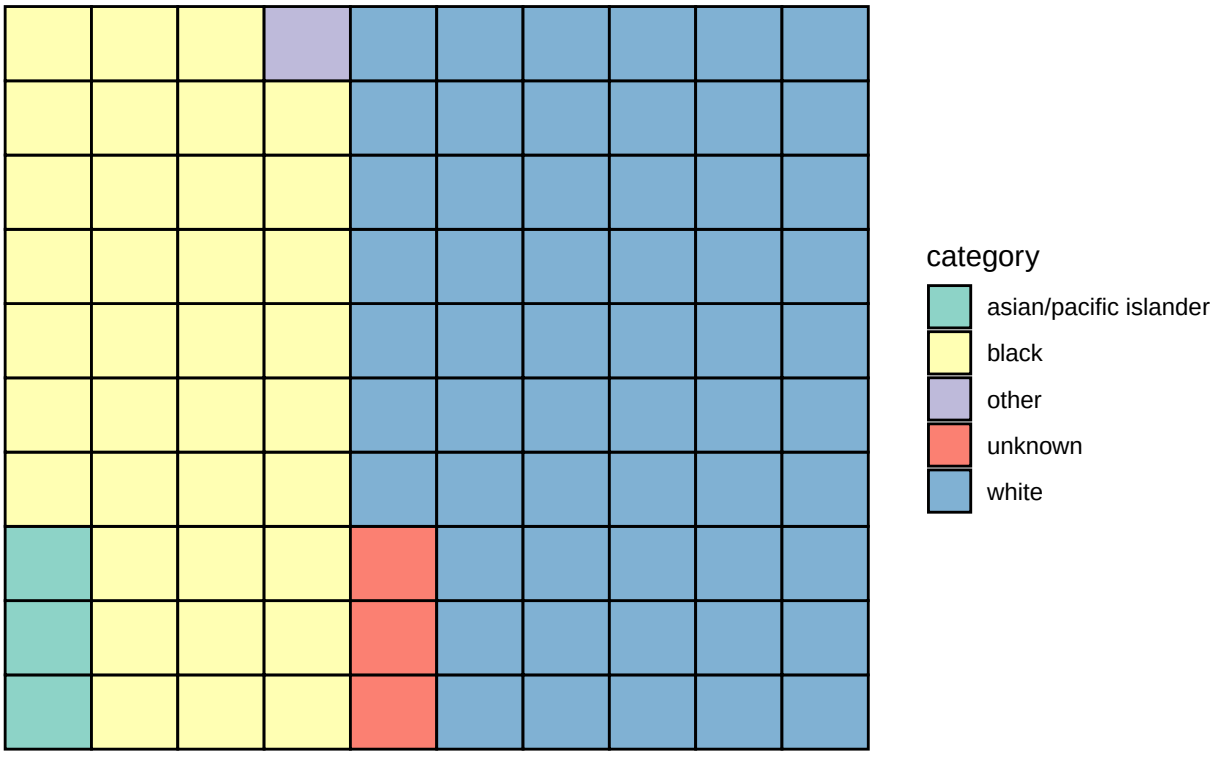
Data Exploration - Graphs

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.

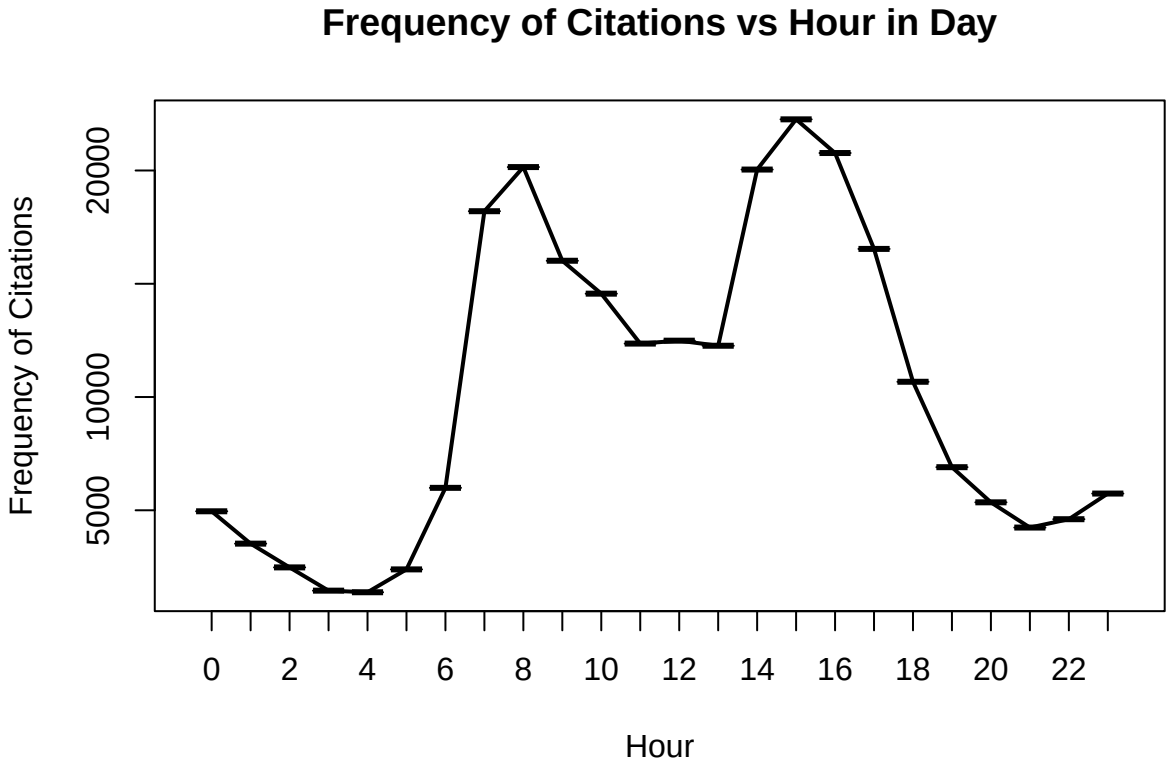


First, we turn our attention to the total individuals of each race issued citations in each year 2014-2020. From 2014-2019, the greatest number of individuals issued citations were white, but in 2020, black individuals narrowly surpassed white individuals as the racial group with the most citations of 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.

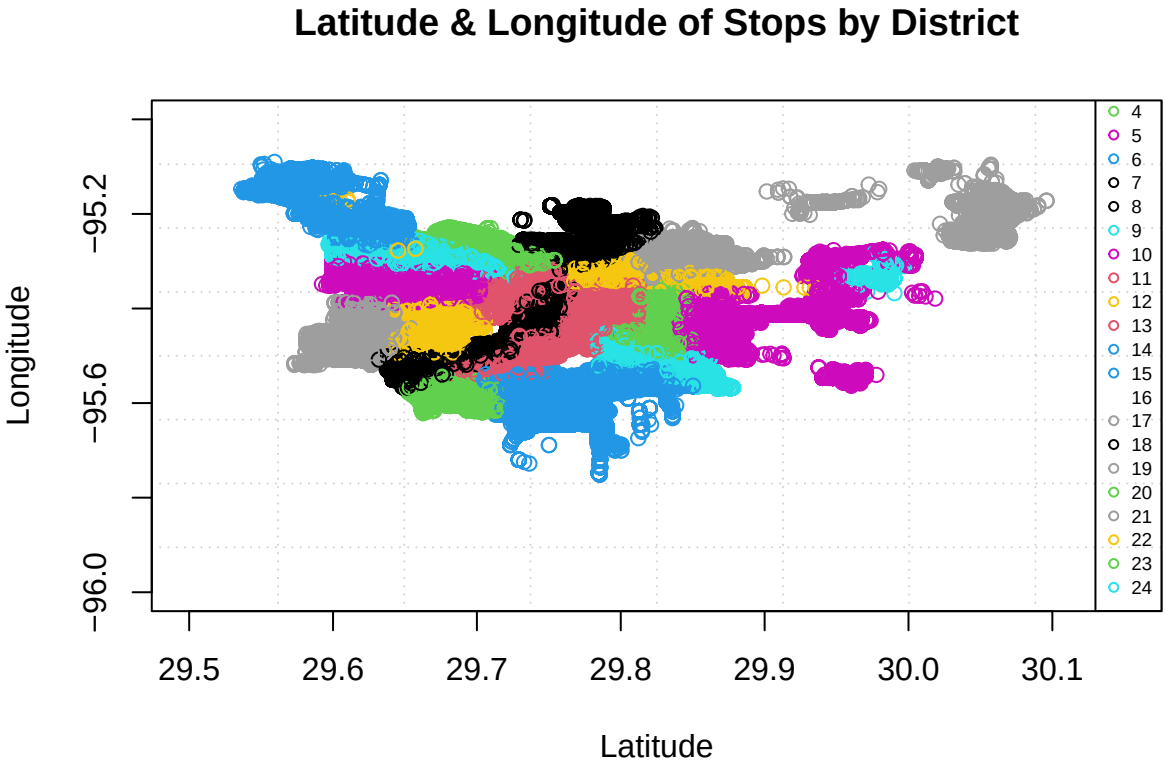
Waffle Chart of Citations per Race



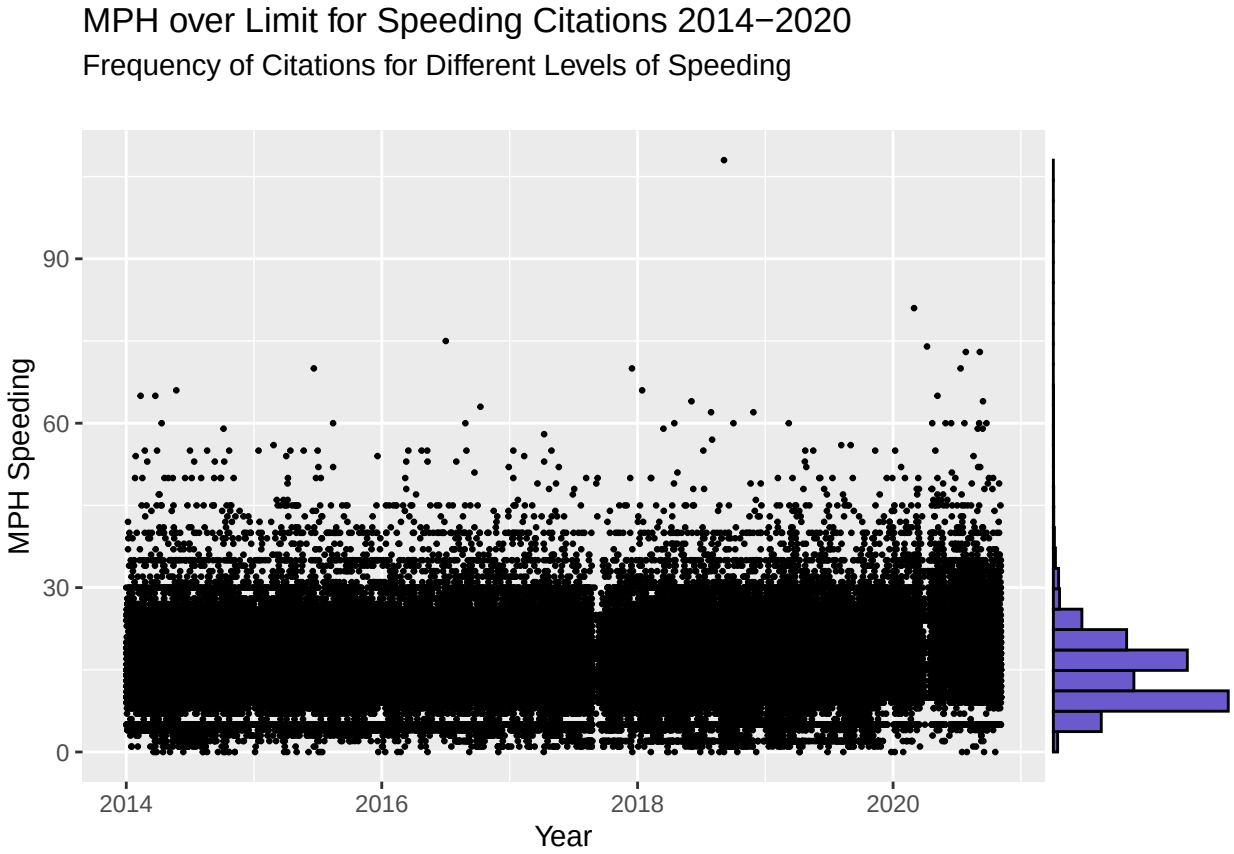
Next, we will look at the % of Citation by Race across all years 2014-2020 in a more geometric representation of the percentages. As we can see, white individuals have the highest percentage of citations, followed by black individuals, and then Asian/pacific islanders, then unknown races. This is somewhat 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.



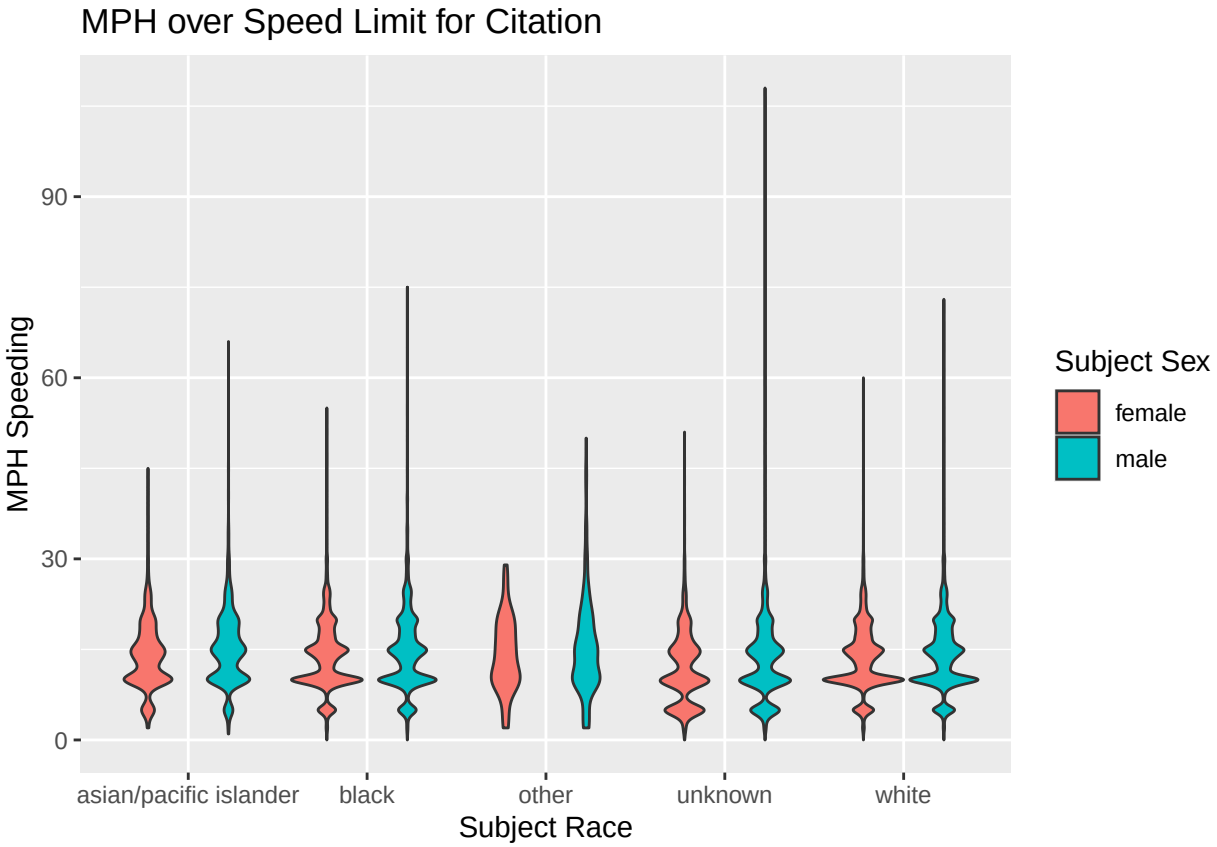
Now we will look at the Frequency of citations over the Hours of the Day. Citations are most common around 8 am, and around 3 pm. There may be a higher proportion of citations around 8 am because of high travel due to work. We are not quite sure why another peak is found at 3 pm, it may be people returning 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.



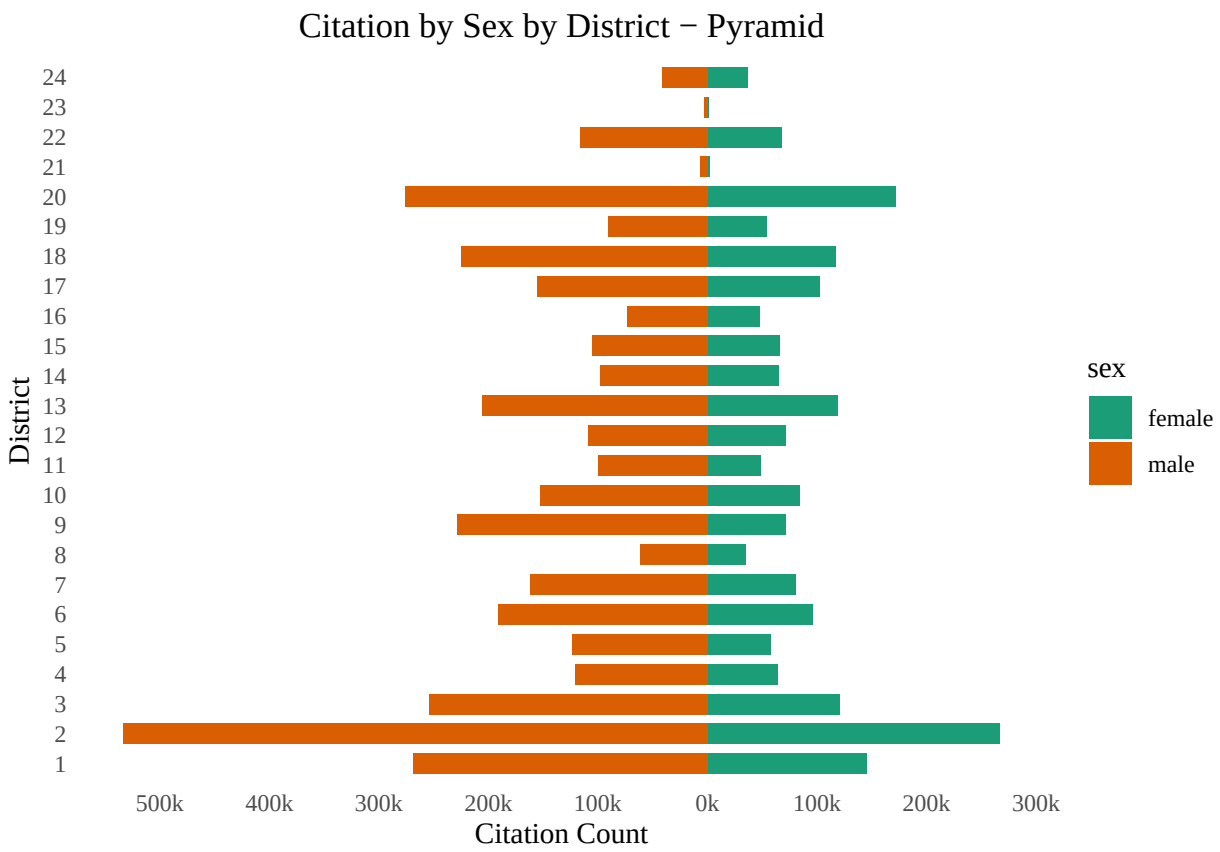
Next, we want to explore the spatial features of our data. We have both the latitude and the longitude of each stop, along with the district and beat. We knew that beat meant a motorized police unit that patrols a specific territory, but "district" in this context is not as clear, as it could mean multiple 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.

```
## # A tibble: 4 x 3
##   race                    population citation_prop
##   <chr>                        <dbl>         <dbl>
## 1 white                      0.441         0.567
## 2 black                      0.412         0.364
## 3 other                      0.00961       0.00176
## 4 asian/pacific islander     0.134         0.0348
```

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:

```
## # A tibble: 6 x 2
##   subject_race           n_individuals
##   <chr>                          <int>
## 1 white                         120836
## 2 black                          61548
## 3 <NA>                           31988
## 4 asian/pacific islander          8311
## 5 unknown                         7620
## 6 other                            443
```

Invalid License Citations by Race:

```
## # A tibble: 6 x 2
##   subject_race           n_individuals
##   <chr>                          <int>
## 1 white                          74202
## 2 black                          58105
## 3 <NA>                           51950
## 4 unknown                         3395
## 5 asian/pacific islander          2086
## 6 other                            127
```

Failure to Establish Financial Responsibility Citations by Race:

```
## # A tibble: 6 x 2
##   subject_race           n_individuals
##   <chr>                          <int>
## 1 black                          39280
## 2 white                          36750
## 3 <NA>                           26702
## 4 unknown                         1768
## 5 asian/pacific islander          1155
## 6 other                             78
```

Seat Belt Citations by Race:

```
## # A tibble: 6 x 2
##   subject_race           n_individuals
##   <chr>                          <int>
## 1 white                          18368
## 2 black                          11719
## 3 <NA>                            5387
## 4 asian/pacific islander           549
## 5 unknown                          532
## 6 other                             40
```

Running a Red Light/Stop Sign Citations by Race:

```
## # A tibble: 6 x 2
##   subject_race           n_individuals
##   <chr>                          <int>
## 1 white                          33214
## 2 black                          17689
## 3 <NA>                           14812
## 4 asian/pacific islander          2928
## 5 unknown                         1823
## 6 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?

```
##       black      white        aapi     unknown       N/A
## 1 0.0267934 0.01651837 0.002512302 0.002698384 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 black individuals. This proportion is about 0.01 higher for black individuals compared to white individuals, while aapi individuals and those of unknown race share a similar proportion around 0.0025. However, the the greatest proportion of total citations that mention failure to establish financial responsibility as the sole violation is attributed to "N/A", which indicates that the officer neglected or failed to record the subject's race.

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):

```
## # A tibble: 1 x 3
##    mean median    sd
##   <dbl>  <dbl> <dbl>
## 1  14.0     14  6.20
```

By Race:

```
## # A tibble: 7 x 5
##   raw_race          total  mean median    sd
##   <chr>             <int> <dbl>  <dbl> <dbl>
## 1 American Indian     417  14.9     14  7.57
## 2 Asian              7715  14.6     14  5.99
## 3 Black             54344  14.3     14  6.07
## 4 Pacific Islander    305  14.4     14  5.23
## 5 Unknown            7056  12.6     12  6.26
## 6 White            109576  13.9     14  6.19
## 7 <NA>              26060  14.4     14  6.44
```

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:

```
## # A tibble: 3 x 5
##   subject_sex  total  mean median    sd
##   <chr>        <int> <dbl>  <dbl> <dbl>
## 1 female       78617  13.6     13  6.05
## 2 male        126747  14.3     14  6.28
## 3 <NA>           109  13.5     13  5.39
```

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:

```
## `summarise()` has grouped output by 'raw_race'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 18 x 6
## # Groups:   raw_race [7]
##    raw_race         subject_sex total  mean median    sd
##    <chr>            <chr>       <int> <dbl>  <dbl> <dbl>
##  1 American Indian  female         97  14.0   14   6.04
##  2 American Indian  male          320  15.2   14   7.95
```

```
##  3 Asian            female       2680  13.8   14   5.60
##  4 Asian            male         5033  15.1   15   6.15
##  5 Asian            <NA>            2  10.5   10.5 0.707
##  6 Black            female      24408  13.9   14   5.53
##  7 Black            male        29928  14.6   14   6.47
##  8 Black            <NA>            8  16.8   16.5 6.96
##  9 Pacific Islander female         79  13.7   14   4.74
## 10 Pacific Islander male          226  14.6   14   5.38
## 11 Unknown          female       2450  11.8   10   6.00
## 12 Unknown          male         4606  13.1   13   6.35
## 13 White            female      39745  13.5   13   6.40
## 14 White            male        69814  14.1   14   6.05
## 15 White            <NA>           17  14.5   15   6.91
## 16 <NA>             female       9158  13.7   13   5.83
## 17 <NA>             male        16820  14.8   14   6.73
## 18 <NA>             <NA>           82  13.0   12.5 4.82
```

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.

```
## [1] "The highest number of citations occur around 16 hr"
```

```
## [1] "The lowest number of citations occur around 5 hr"
```

```
## [1] "The highest number of citations occur at 15, 16, 8, 14, 7 hr"
```

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 race is given the most citations for each district?

Next, we were curious if certain districts ticketed certain races more often than others. We were mainly interested in if the race that was ticketed most was black or white, since the percentage of AAPI people is small compared to the percentage of white and black people.

```
## # A tibble: 25 x 3
## # Groups:   district [25]
##    district subject_race total
##       <dbl> <chr>        <int>
##  1        1 white        21447
##  2        2 white        49067
##  3        3 white        17755
##  4        4 white         9099
##  5        5 white         8039
##  6        6 white        12386
##  7        7 white        11064
##  8        8 black         5030
##  9        9 white        14707
## 10       10 black        10147
## 11       11 white         8741
## 12       12 white         8983
## 13       13 white        15301
## 14       14 black        11350
## 15       15 black         7964
## 16       16 black         5964
## 17       17 black         9516
## 18       18 white        15834
## 19       19 black         4853
## 20       20 white        18820
## 21       21 white          322
## 22       22 white         8616
## 23       23 black          163
## 24       24 white         5944
## 25       NA white        39531
```

In an ideal world, we would have data on the % of each race for each district, and we have been looking everywhere and we have not been able to find it yet. We submitted a Freedom of Information Act (FOIA) request for the data, and we are optimistic HPD will get back to us with this data in a reasonable amount of time. In the next iteration, we will hopefully have data about race/ethnicity break down by race to be able to investigate if the race that is cited the most frequently in each district is the race that is the most prevalent in that district. For now, this is just the simple breakdown of what race is cited the most for what district.

7- 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.

Statistics on number of citations/square mile:

```
##     median standard deviation     IQR
## 1 853.5304           1877.273 1094.37
```

Top 5 Beats ranked by citations/square mile:

```
##   Beats Citations_per_sqm
## 1  1A10         12162.841
## 2  2A40         10326.629
## 3  2A10         10174.400
## 4 10H40          6582.549
## 5 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 Citations_per_sqm
## 1 24C30        128.124500
## 2  8C40         91.977928
## 3 21I50         56.395863
## 4 24C40          8.352124
## 5 23J40          5.922349
```

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.