Crime and COVID-19 in Houston: How do they relate?

Gail Oudekerk, Kelly Zeng, Elijah Sales, Jennifer Jia, Naghmeh Hosseini

Introduction

Our group had previously seen news articles referencing the relationship between COVID-19 and crime in the United States and other countries. Perhaps the COVID-19 recession pushed people into theft, or the pandemic stress led to an increase in suicides and homicides. What if COVID-19 actually led to a decrease in crimes since more people were staying at home due to quarantine? We wanted to see what relationship between crime and COVID-19 was present specifically in Houston, Texas, where us as students are residing. This leads to our research question: Is there a correlation between COVID-19 positivity rates and crime offense rates in Houston, Texas?

Datasets

Main Dataset

Our first main dataset is crime data by streets and neighborhoods in Houston, Texas from the years 2017 to 2021. This dataset was taken directly from the Houston Police Department website. It has 1025169 observations and 15 variables. Some of these variables include:

- Incident
- OccurrenceDate
- OccurenceHour
- NIBRSClass
- NIBRSDescription
- OffenseCount
- Beat
- Premise
- StreetNo
- BlockRange
- StreetName
- StreetType
- Suffix
- City
- ZIPCode

The data provided on the website was sometimes by year and other times by months. Different months or years sometimes had different variables. We ended up standardizing all variables across all months and years through data cleaning, which will be explained in a later section

Secondary Dataset

Our secondary dataset is from the Houston COVID-19 Data Hub, showing the daily COVID-19 positivity rates in Harris County from 2020 to 2022. This dataset has 662 days observed and 20 variables. Some variables include:

- id
- Date
- Positive_Pct
- Postive_Pct_14dayAvg
- Threshold_5PCT
- Status
- Red_Status
- Orange_Status
- Yellow_Status
- Green_Status
- Total
- Negative
- Positive
- Date_Str
- Positive_Pct_Str
- Postive_Pct_14dayAvg_Str
- Threshold_5PCT_Str
- Date_Updated

The status variables are different COVID alert levels. Red means avoid all gatherings, orange means minimize contact, yellow means stay vigilant, and green means resume normal contact.

Freetext Dataset

Our last dataset is composed of freetext and is from Github user MickeysClubhouse. This dataset compiles many news headlines tha tmention COVID-19. Our group mainly used this dataset to see if there were common mentions of COVID-19 with acts of crime.

Data Cleaning Process

The biggest challenge we face in cleaning data sets of Houston Police Department's records of crime counts is that the dataset lacks a uniform standardization. First of all, for the NIBRS Description Differences, which specifies the type of crime committed, 2020 and 2021 data use the same descriptions, while 2019 data is more generalized: theft, robbery, and auto theft are three different crime types for 2020 & 2021, but are all categorized as theft & robbery in 2019 data. To deal with the problem, we researched crime definitions online and finally decided on creating 7 standard crime types for our research purposes:

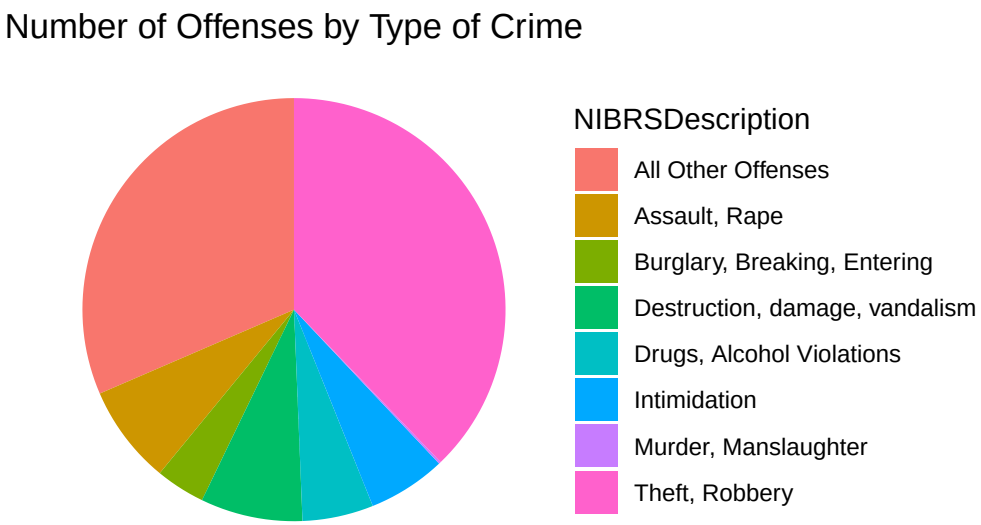
- Theft, Robbery
- Murder, Manslaughter
- Intimidation
- Drugs, Alcohol Violations
- Destruction, Damage, Vandalism
- Burglary, Breaking, Entering
- Assault, Rape
- All Other Offenses

For all other crimes that can't be generalized, we categorized them under "All Other Offenses". Another step we have taken to clean the database also include unifying address types: for example, since street address is recorded differently across the years, we decided to employ "broad range" instead of street address to better visualize where crimes occurred.

Initial Visualizations

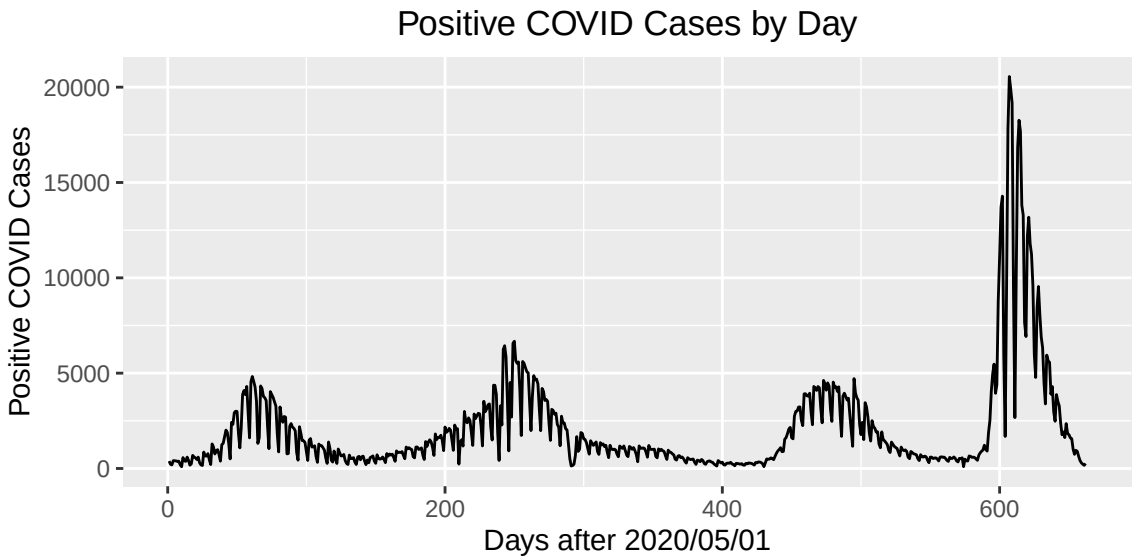
The next few plots are our group's first few visualizations to get an idea of what our datasets looked like in order to decide what we wanted to look into further to answer our research question.

Plot 1



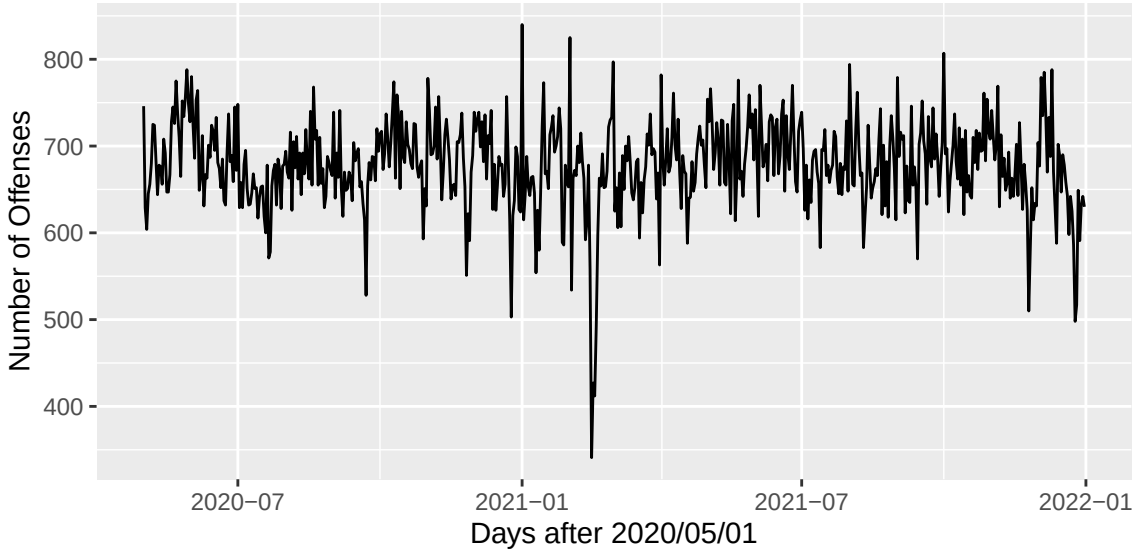
This first plot is a simple pie chart showing the composition of each type of crime. We can clearly see that theft and robbery make up a large percentage of all crimes compared to other types. Although the all other offenses category looks like a lot, it must be kept in mind that this category consists of over 30 other types of crimes that only have a few offenses each. Seeing this plot pushed our group into looking at why theft and robbery was the most occurred crime and if this was consistent throughout the years.

Plot 2



This plot shows the number of COVID cases per day since May 1, 2020, which is the first observation of our COVID dataset. From this graph, we can see clearly when COVID cases peaked. Peaks occurred around the 60th, 250th, 475th, and 600th days after May 1 of 2020, which translates respectively to late summer of 2020, early February of 2021, late summer of 2021, December of 2021. This is a clear indication of how increased summer traveling and Delta/Omicron waves have influence on COVID cases rising rapidly.

Plot 3

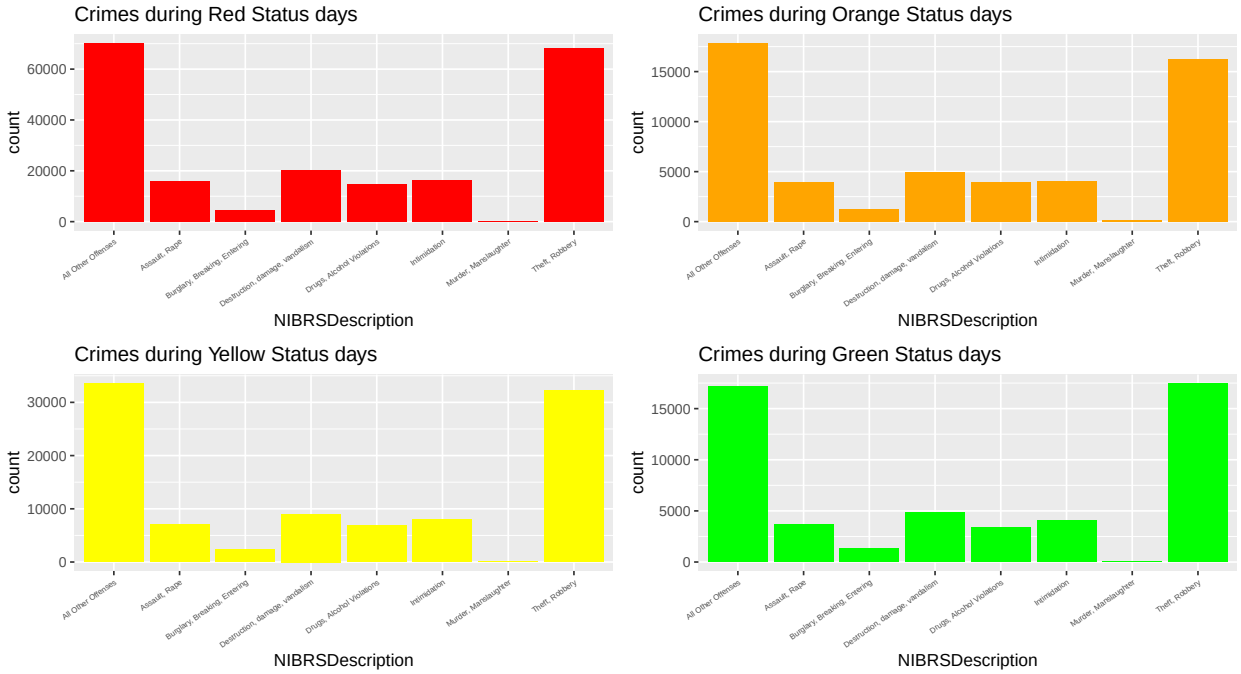


This graph shows crime offense peaks around January of 2021 and a low around February or March of 2021. Otherwise, it seems that crime offenses per day oscillated pack and forth from May 2020 to January 2022, which shows a pretty stable crime offense counts over the past two years.

Crime and COVID in News Headlines

For our freetext dataset, we searched for different keywords related to crime, including "crime," "theft," "jail," and more. However, out of the 4128 different news headlines, only around 20 of them mentioned one of the keywords. This lack of headlines mentioning both COVID-19 and crime-related keywords shows that we are not able to find a correlation between the two with this specific dataset. Now, we will utilize our other datasets to determine whether or not there is a relationship between crime and COVID-19 in Houston, Texas.

Crime Counts for Each Alert Level

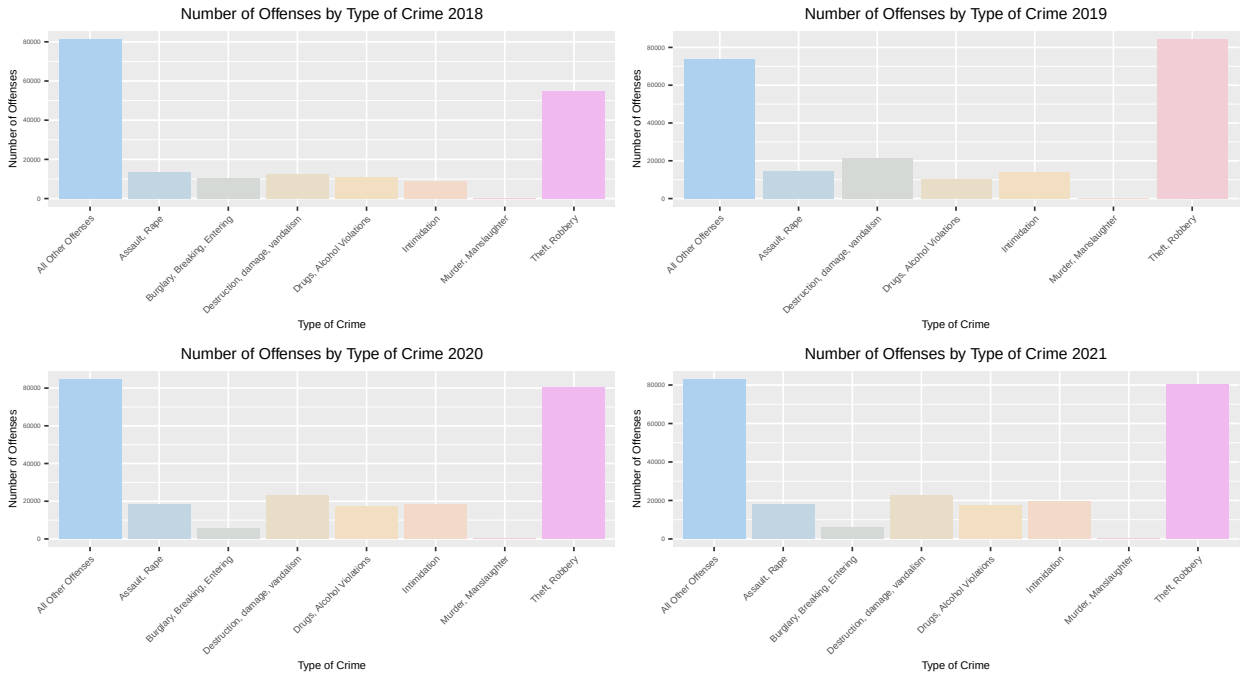


One significant part of our analysis involved further investigating the different COVID-19 "alert levels" found in our secondary dataset and how those different alert levels may have affected the occurrence of crimes across Harris County. Quantitatively, based on exploratory data analysis, an alert level of "green" signifies that less than 5 percent of the Harris County population had tested positive for COVID-19 on a particular day; an alert level of "yellow" signifies that between approximately 5-8 percent of the Harris County population had tested positive; an alert level of "orange" signifies that around 8-10 percent of the Harris County population had tested positive; finally, an alert level of "red" signifies that over around 10 percent of the Harris County population had tested positive for COVID-19. Our group had initially proposed that the total number of crimes across all crime types would be higher for worse COVID-19 alert levels.

To create the plots above, the format of the column containing dates in our secondary dataset was changed to match that of the column containing dates in the main dataset. The main and secondary datasets were then merged and subsets of the data for each COVID-19 alert level were created. Since the secondary dataset only contained dates from May 1, 2020 to December 31, 2021, the merged dataset contained crimes only within that time period. Using ggplot, the plots were generated for each alert level, with the offense type as the x-variable and the total number of offenses during the time period mentioned as the y-variable. The bars are colored based on the alert level to produce a more visually appealing output.

The four bar plots above, show the total number of occurrences of crime for each crime type during certain COVID-19 alert levels from May 1, 2020 to December 31, 2021. When comparing the total number of occurrences across different crime types, the plots show a fairly similar distribution regardless of the alert level. However, it is clear that there is a much higher total number of occurrences of crime on days associated with an alert level of red than on days with other alert levels. Days with an alert level of "yellow" present the next highest numbers of total occurrences for each crime type and days with an alert level of "orange" and "green" show fairly similar values across all crime types. Based on the plots above, it appears that there is generally a higher total number of crimes as COVID-19 alert levels worsen. It should be noted, however, that this observation may not necessarily correlate with the idea that more crimes occur on days with worse COVID-19 levels. This is explained by the fact that there may have been more days with an alert level of "red" compared to other alert levels and not necessarily a higher frequency of crimes on worse COVID-19 alert levels.

Most Prevalent Type of Crime



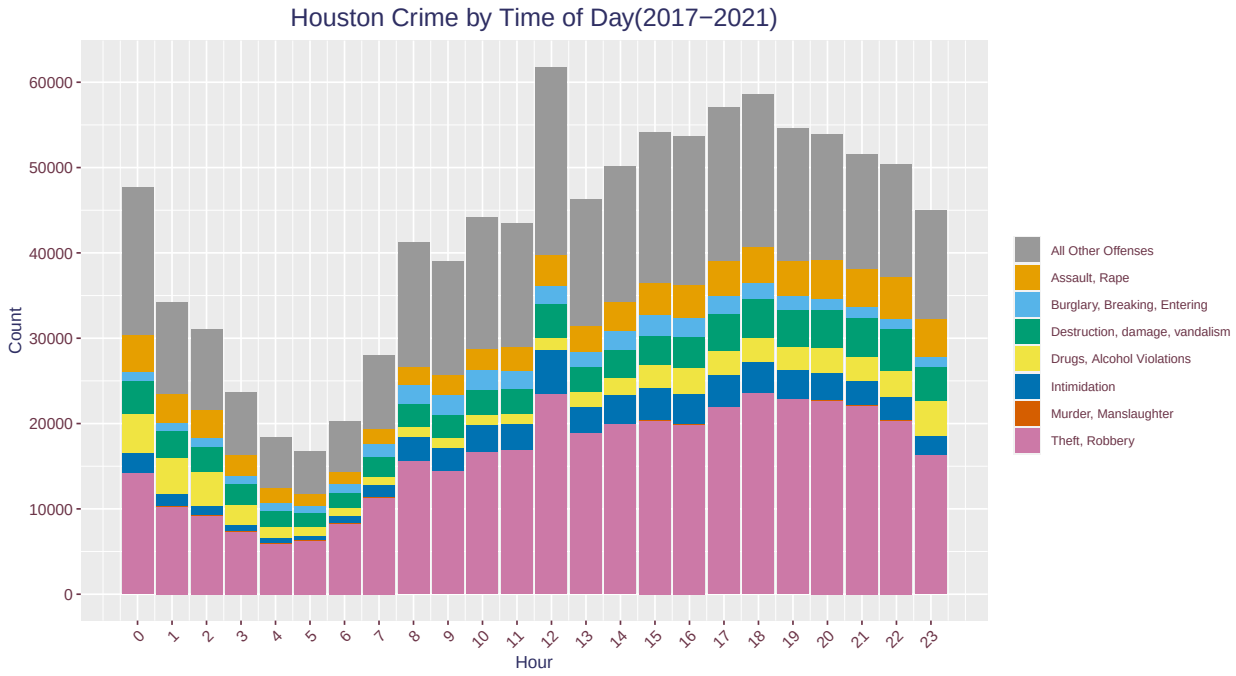
Once we had a general idea that crime cases in general increase due to worsening COVID levels, we decided to dig deeper into each specific type of crime and explore if there's a correlation between certain crime types and COVID cases. Our intuition is, since COVID-19 pandemic contributes to an extremely variable financial market, unemployment rate continued to rise, which in turn led to petty crimes like theft and robbery on the rise. Therefore, our initial hypothesis is theft and robbery cases would be greater for years with COVID cases than those without.

The four plots shown above showed different types of crime cases over a 1 year interval in 2018-2021. Here, we excluded 2017 data set because 2017 was the first year Houston Police Department started to keep track of the crime data; therefore, 2017 database lacked a significant number of data to be considered valid. Our thought process behind creating this set of graphs is to stack all crimes of a certain type together and see if theft and robbery cases, in particular, have increased over the years. Therefore, we used ggplot function to create bar plots and fill the different types of crimes with different colors for better visual presentation. Here, comparing 2019 crime cases with 2020 crime cases, we can actually see a gradual decrease in petty

crimes (2019 theft&robbery cases were well above 80000 while 2020 theft&ribbery cases were under 80000). If COVID cases propel petty crime growth, then we should see the opposite effect on the two graphs. Moreover, theft&robbery cases are also stagnant if 2020 plot and 2021 plot are compared together. Because 2020 and 2021 have different COVID cases and similar petty crime cases, we rejected our initial hypothesis and concluded that a rise in COVID cases did not add to petty crime increase.

While creating the graph for "most prevalent type of crime", we realized there are several issues that also need to be taken into consideration in the future. To start with, our categorization of crime types had significantly influenced the way the graphs turn out to be; in order to make sure the conclusion is still valid, we should consult Houston Police Department to understand better the logic behind their initial categorization of crime types. On the other hand, the graphs we've created are more general and only compared covid cases over 1 year intervals. This could have easily missed other factors that influence crime case counts like lack of police patrol over certain areas and time periods. A better visualization could be created to contrast a peak COVID positive day's crime cases in theft & robbery with that of a lower COVID positive day.

Crime by Time of Day

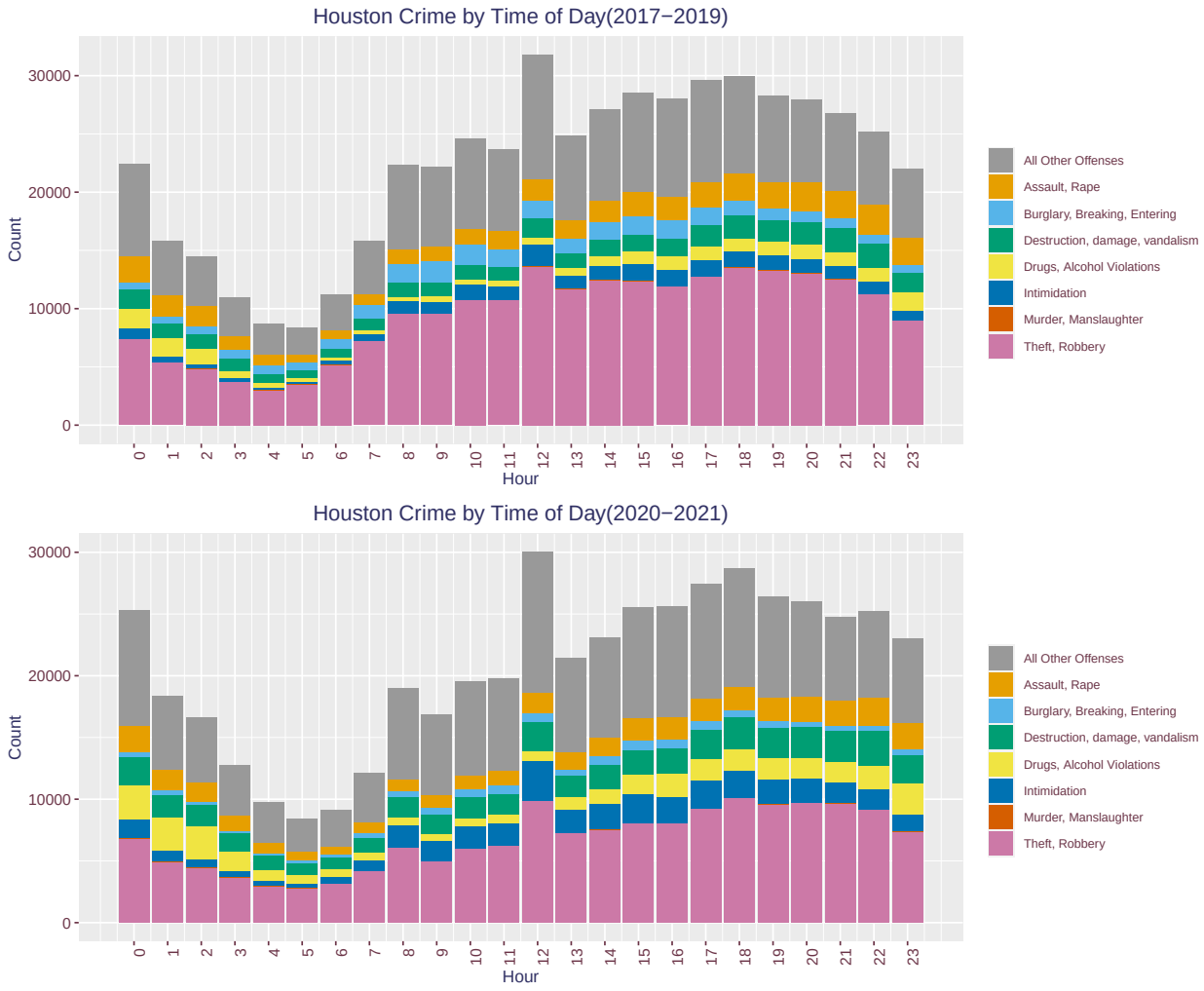


Our group decided to analyze the impact of covid 19 pandemic on time of the crime occurrence since the same location might be attractive for a crime during the day but unattractive at night. For example, homes are normally occupied in the evening and night but vacant during office and school hours. However, the pattern of occupancy versus time was different during pandemic due to working from home or online education.

The original data was cleaned to have the same expression of time (24-hour clock), and a time conversion was made to have time zero as the midnight hour. Also, the data was filtered and sorted to have two time periods, before and during the pandemic. The ggplot was used to create stacked bar plots to identify different types of crimes during each hour.

The first graph shows a trend for Houston crime over time between 2017-2021. The general trend shows that the number of violent crimes such as drug and alcohol, intimidation, and robbery, increases hourly from noon through the afternoon and evening hours.It peaks at 12 p.m., and then drops to a low point at 5

a.m. It looks people are more exposed to the criminals because they are not at home or in a workplace at noon.Theft and rubbery was a dominant crime that mainly occurred during the afternoon and continued in the evening. The number of murder cases compared to other crimes was insignificant, so it doesn't show in this plot.



The above plots depict the Houston crime counts based on time of the day between two time periods, before the pandemic (2017-2019) and the COVID period (2020-2021). Comparing these two plots reveals that both periods have the same general trend with a peak at noon.

However, the trend of some crimes in Houston changed over time due to the pandemic. During the COVID period, some crimes, such as drugs and alcohol violations, showed an hourly increase throughout the afternoon and evening hours and peaked between midnight and 2 a.m. At the same time, Robbery decreased during working hours, probably due to the work-from-home schedule, but still, it is a dominant crime compared to other violations. The same analysis shows that the COVID impacted the intimidation violation's time of occurrence. The intimidation violation started earlier in the morning at about 8 a.m. and had an ascending trend throughout the day.

Killer Plot

Initial Planning

In initial discussions of the "Killer Plot", we had several major goals. First, of course, we wanted to satisfy the requirements of the project, building an original plot using solely the grid primitives that could be displayed in shiny. Secondly, we wished to answer several of our main research questions in one plot. Perhaps most importantly, we aimed to create a plot that was as original and memorable as possible without compromising readability or faithfulness to the original data. The "killer plot" should stick in the mind, helping promote the import of its subject matter, but it should not be complicated, fancy, or whimsical to the detriment of its intended purpose or message.

To these ends, we considered many options, including a time series graph of crime occurrences with shaded-in background columns representing the daily COVID-19 alert level. However, many of our ideas seemed to fall into the category of "an existing plot with something added on," which did not strike us as sufficiently original for this project.

During this brainstorming process, while reviewing our free text source (containing headlines about COVID-19) and popular parlance of the last two years, we began to notice that commonly-used language refers to COVID-19 in nautical terms. The pandemic occurs in "waves", mimicking the rise and fall of case counts; we face a "rising tide" of positive cases. We felt that these semantic features mimicked many of our own feelings about the pandemic, as something inconceivably larger than ourselves whose currents defined our lives and movement in almost every way. These meaningful links led us to the idea of representing the daily positive case count as a "water level" that would rise or fall with the "waves" of the pandemic. Accordingly, we conceived of the representations of daily life in our data–here, the crimes logged by the Houston Police Department–as a boat on that tide, with counts perhaps even rising and falling in tandem with the positivity rate.

This idea underwent multiple iterations. We considered placing a multitude of small boats, one per day, on the time-series wave, but dismissed the idea out of concerns about legibility. We considered representing the body of the "boat" as a semicircle functioning as a pie chart, to compare the relative frequencies of types of crime; however, since bar graphs offer much greater ease of comparison than a pie chart, we began to lean towards including a bar graph instead. Since the boat now hosted vertical bars instead of a sail, we decided to give the "water" (a single bar representing COVID-19 positivity rate) a color matching the daily COVID-19 alert level rather than filling in the "sail" with that color. Throughout this process, we attempted to be wary of overcomplication, and to prioritize elegance and ease of communication over too-whimsical and unclear implementation.

The finalized first draft of our plot could be constructed using grid primitives and implemented in shiny; it involved, directly or indirectly, all three of our datasets; and it was original. One concern we discussed was a mismatch between the more fun, lighthearted nature of our plot and the more serious subject matter it depicted. Above all, we wanted to be respectful to the issues at hand. In the end, we decided that the meaningful links between the commonly-used language surrounding the pandemic and the construction of our plot justified its form; however, we are always open to feedback to the contrary.

Implementation Process

Data Cleaning and Combination

To retrieve and combine the data, we used SQL queries through RStudio. We selected the Occurrence Date, the NIBRS Description (crime type), and the count of each NIBRS Description from our crime dataset after grouping by Occurrence Date and NIBRS Description. Then, we performed an inner join on that table with a part of our COVID-19 table, after selecting Occurrence Date, positivity rate, and COVID daily status for each day. We inner joined on Occurrence Date. This gave us a dataframe in R with each block of rows corresponding to an Occurrence Date between 2020-05-01 and 2021-12-31, and each row corresponding to a

type of crime on the given day with date, crime description, crime count (number of times it occurred on that day), the COVID-19 positivity rate, and the COVID-19 alert level ("Green_Status", "Orange_Status", "Yellow_Status", or "Red_Status").

Given that dataframe, we performed string processing to remove the "_Status" from the end of each alert color name. Then, we replaced the alert color name with the color name in R that we wished to use in the plot itself.

Working in grid

Constructing the plot in grid was an exercise in viewports. We created a function containing our grid code so that we could initialize the plot with a single call (inputting the dataframe and the desired date).

Within that function, we first created and pushed our main viewport (the largest viewport containing all the elements of the plot). We extracted some values from the dataframe corresponding to the inputted date (such as that day's subset of the dataframe, the alert color in string form, and the positivity rate). We then initialized the main bar (the "water") with height corresponding to the positivity rate and color to the alert color, along with its y-axis.

We then moved on to the bars (the "smokestacks" or "funnels"). This demanded a new "bars" viewport, then a second, right-hand y-axis to display the frequency of the types of crime. We took the number of types of crime on the given day ("num_bars") and used it to create "spacing", equal to 1/num_bars. This corresponds to the proportion of the viewport that "belongs" to each bar. We then created a "spacing vector" with the x-axis placement of each bar. In a for loop with number of iterations equal to the number of bars, we initialized a rectangle with height corresponding to frequency, a label (tilted at an angle to mimic steam emitted as the ship moves), and a new viewport and grey semicircle to aesthetically mimic the top of the smokestack. We popped the new viewport at the end of each iteration. After the for loop, we popped the "bars" viewport.

We created a new "inner" viewport for the body of the ship with clip "on"; this served to create our desired body shape by "clipping" (removing from view) any portions of the shape outside the bounds of the viewport. We created a circle within that viewport, which turned into the lower third (or so) of a circle thanks to the clip mechanism. We stored the total crime count in a variable and labeled it in text on the body of the ship. Then we popped the "inner" viewport. We did the same (pushing a viewport with clip set to "on", then creating a circle, then popping the viewport) to mimic a wake behind the ship.

Finally, we created the the y-axis labels, the title, and the subtitle. Due to the variable height of the barplot's crime type labels, we decided to place the title and subtitle below the plot to avoid overlaps. The subtitle's color matches the water, to create an easy mental link; the green and yellow text was too light to be easily legible against a white background, so we created a check for color and then an appropriately shaded rectangle to "highlight" their black subtitle text instead. We popped the main viewport.

The implementation in shiny was relatively simple: we selected the date input as the manipulation point, inputted our date parameters and format, and then enclosed the function call into a shiny "renderPlot" call.

Challenges and Revisions

One major challenge was the proportions of the plot. Due to our variety of viewports and code design, our shapes scaled differently with different ratios of width to height, and text sizes that were previously perfectly scaled became absurdly large or small. In practice, our solution was manual font size adjustment for each type of implementation (presentation in shiny, rmd to pdf, etc.) and a set viewing ratio (1.37 units of width to every one unit of height). In future work, it would be useful to construct these plots in such a way that they are robust to changes in aspect ratio.

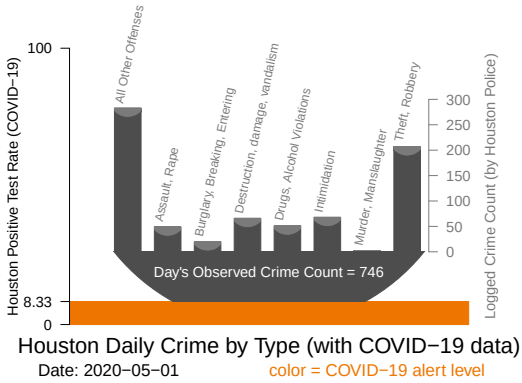
An additional concern, brought up in the Q&A section of the presentation, was that the "wake" behind the ship made the positivity rate bar's meaning (the "water level") less clear. After some consideration, we decided to prioritize legibility over whimsy and removed the "wake".

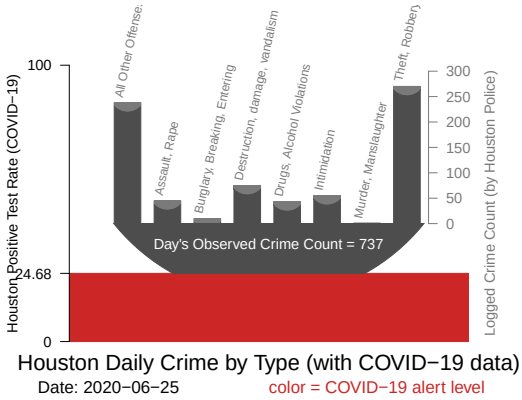
Implications

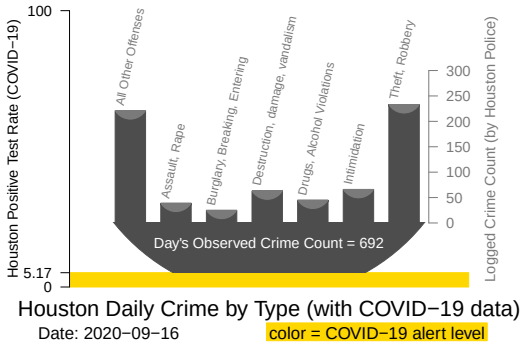
We recognize that the plot is quite specific to our data. In our search for originality, we have also constructed a plot that is, upon first glance, much less clear than its counterparts. However, we are satisfied with the meaning, the effect, and the overall implementation of the plot, as well as the experience and learning about grid and shiny systems that we gained in its construction.

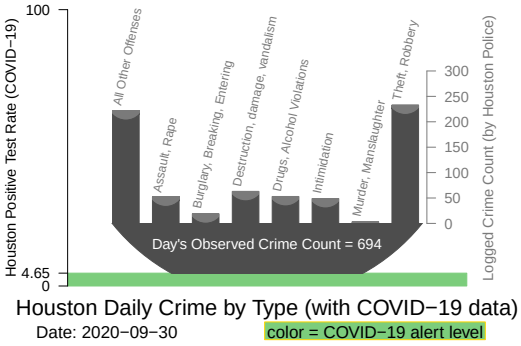
Killer plot examples

To discuss our plot in this pdf submission, we have static (single-day) versions of the "killer plot" rather than the interactive version in shiny, where the user selects a date from a calendar format and the plot displays the output for that date. Please see the presentation file for the interactive version of the "killer plot".









Conclusions

In summary, to answer the question "is there a relationship between COVID-19 positivity rate and crime offense rates in Houston, Texas" we combined and contrasted data bases of Houston crime offenses and Harris County COVID positivity rate. Based on the comparison of graphs we were able to arrive at the following conclusions:

1. When we use daily positivity rate in Harris County as the indicator of COVID severity levels, with the exception of orange status days, crime seems to be more frequent on days with large numbers of COVID cases. However, this doesn't necessarily mean COVID severity is positively correlated with crime cases: there may have been more days with an alert level of "red" compared to other alert levels and not necessarily a higher frequency of crimes on worse COVID-19 alert levels.

2. Even though COVID cases continued to increase over the years, the most common crime cases (theft & robbery) remained stable, which is likely due to a work-from-home life schedule detering theft & robbery cases happening on the street.

3. COVID-19 changed the time of occurrence in certain crime types, but the general trend remained the same: similar across all types of crimes, peak of crime cases is at 12 p.m., while the lowest crime counts occur at 5 a.m.

In general, we deployed ggplot and grid to make some initial observations and see how different aspects of influences COVID cases have on crimes. However, a few slight improvements that could potentially be included are the following:

1. the crime data set from the Houston Police Department is solely based on the greater Houston Area, and does not extend to areas like Katy, Sugar Land, the Woodlands, which are areas that are covered by the Harris County regions in the COVID data base. To arrive at a more precise conclusion about COVID & crime relationship, we should choose a COVID data base for Houston only, but unfortunately such data base is not available from our search.

2. COVID positive rate included in the COVID-19 Data Hub does not represent the actual COVID positive rate in Houston regions since many people may have not been tested but were still COVID-positive. Similarly, the Houston Police Department crime data also only recorded the cases reported, and does not necessarily translate into all crimes committed. More time could be dedicated to devising a more precise metric to analyze the relationship between crimes and COVID cases.