AC221 Final Project
Amy Wickett, Jessica Wijaya

Introduction

There are large disparities in the criminal justice system based on race. Many point to racial discrimination by police, judges, and prosecutors as one of the main drivers of the disparate outcomes. Scholars have long attempted to detect racial bias amongst various officials in the criminal justice system (Arnold et. al 2018, Tuttle 2019, Ouss and Stevenson 2020). One of the ways that scholars attempt to quantify discrimination by police officers is by looking at differential rates speed discounting (Goncalves and Mellow (2017)).

Speed discounting is when police officers ticket an individual for a different, lower speed than the speed an individual was actually driving, normally to lower the fine amount for the driver. In Florida, speed ticket fines increase at certain thresholds (>9mph, >14mph, >19mph, >29mph). So, for example, an individual going 9 miles per hour over the speed limit has a ticket range around $120, while an individual going 10 miles per hour over the speed limit has a ticket range around $190. Other aspects about driver behavior, driving through a school or construction zone, also influence the fine amount. In general, however, speed discounting can substantially decrease the overall fine a driver will pay for a given traffic stop. The decision to speed discount is largely left up to an individual officer i.e. for each stop, a police officer can choose whether to discount the speed or not. Goncalves and Mellow (2017) find differential rates of speed discounting for women and men and for minorities and non-minorities. On average, whites were more likely to benefit from speed discounting than minorities. They also found that minority and female officers were more likely to speed-discount. We will validate these results and then

extend this methodology by evaluating ticketing behavior by political party affiliation for officers in Hillsborough county from 2017-2019.

In order to get officer party affiliation, we are going to link officer names with voting rolls for Hillsborough county and for nearby Polk county, with the assumption that officers that patrol Hillsborough county live in close vicinity to the county. For each county, we have all voters that are registered as of 2019. We will detail more about the data sources in the data section. We will attempt to match officers based on name. If we are able to positively link an officer, we will have information on his or her gender, race, party affiliation, voting history and address.

One of the central tensions in this course is between transparency and privacy. Given the impact a police stop can have, ensuring that officers are targeting the most dangerous drivers rather than the drivers of a certain race is a public imperative. At the same time, releasing vast amounts of information about drivers and police officers does not seem to be the optimal strategy to ensure that police officers are not unfairly targeting certain types of individuals. Additionally, being able to find where police officers live may pose a safety hazard for police officers.

In this paper, we will detail our data, matching algorithms, and analysis of speed discounting by officer covariates. We will conclude by analyzing the tensions between privacy and transparency and a discussion of potential considerations for the Hillsborough police department data releases.

Data Collection

We have 2 main datasets, Police Stop data and Voter Registration data. As a note, all data was provided through online requests and was obtained without fee.

Police Stop Data

Hillsborough county police departments lists all civil stops that occur in Hillsborough county; we utilize online records for the time period from 2017-2019. There are 5,943 unique police officers represented over this time period with over 412,641 unique police stops. Most of the officers are from the Hillsborough County Sheriff's Office (2,514) and the Tampa Police Department (2,130). However, officers from the Florida Highway Patrol and other smaller police forces (including smaller cities, college campuses, etc.) are also represented. Of these cases of police stops, 81,510 were for speed infractions and listed both the "actual" speed (as recorded by the officer), as well as the posted speed limit where the driver was pulled over. For each of the cases, we have information on:

- Officer: officer's first, last name, agency affiliation and sometimes middle name or middle initial (middle name information varied by agency)
- Offender: first and last name, address, date of birth, race, gender, and drivers license number
- Offenses: statute broken, posted speed, actual speed, disposition outcome, disposition date, amount paid, and the judge's name if the case went to trial

Voter Registration Data

We have voter roles for republicans and democrats in Hillsborough and Polk counties for all registered voters as of March 2019. This information includes full name, address, race, sex, birthdate, party affilation, telephone number, and voting history. We have 1,035,342 voter records in total in the dataset, of which 1,014,552 are unique (based on the name). We are assuming that police officers who work in Hillsborough also live either in Hillsborough county or in neighboring Polk county.

Matching Mechanisms:

We attempted to match individuals through various matching algorithms. We began with cleaning the data and removing any corrupted files. As Jim would have predicted, we were not able to use the Pandas package through Python and thus used the suggested approach Jim provided, using csv-reader to read the files, collect the key in a set (to identify duplicates), and assemble them in the form of a dictionary. We used the voter's full name [first name, middle name (if any), last name, and suffix (if any)] to eliminate any potential duplicate records for voters and used Case ID to remove any duplicate police officer stops.

We began by linking based on first and last name alone. We do this through two different methods. First, we drop all individuals in the voter rolls who have the same first and last name as another individual on the voter roll. Through this method we are able to uniquely link 1,560 unique officer names, about 26.14% of the sample. Some of the officer names can have more than one match with the corresponding names in the voter data. For these duplicated data, we do not know which records these officers' should be matched to. However, if these records have the same covariates (name, gender, race, and political party), then it does not matter if we are able to make an exact match because the covariates for all individuals match. To illustrate this point, let's say there are 3 Michelle Obamas in the voter rolls and one police officer named Michelle Obama. If all of the Michelle Obamas are black women who are democrats, it does not matter if we can figure out exactly which Michelle Obama is the police officer in question. When we merge the data, we only care about the officer gender, party, and race, so in this particular case, linking to any of the Michelle Obamas will result in the same officer characteristics. If we add back in the duplicates that have all the same covariates (gender, race, sex, political party) we are able to match 42.2% of officers.

Next, we link based on first, middle and last name. In the first method of linking we are able to link 578 individual officers. This is only 9.68 percent of the overall sample. The low rates of linking are partially due to how names are reported in the police officer data. Some officers have their full middle names listed and some only have a first initial. If an officer does not have a middle name listed, it is unable to be linked in this mechanism. When we add back in the duplicates with the same covariates, we do not see a large increase in the number of stops and our match rate only goes up to 10.37 percent.

Next, we link based on first and last name as well as middle initial. In the first method, we are able to match about 20.52% of individuals and when we add back in individuals with the same name and same covariates, the match rate only increases back to 22.43%. Individuals who only listed a middle initial were added to the sample above, those listing their first, middle and last name. Finally, we add on suffix (Sr., Jr. III). We add suffix data with first, middle and last name as well as with first name, last name and middle initial. In both matchings, we are only able to link around 1% of the police officers.

Our various matching algorithms suffer from different issues. The first and last name matching algorithm seems to suffer from deleting common names out of the voter rolls. Adding back in duplicate records with consistent covariates seems to partially solve this issue. However, it should be noted that the dataset we have created is a subset of all the officer stops. In particular, our dataset is made up of officers with very distinct names as well as those with more common, stereotypical racial and gendered names. It is unclear what exactly this will do to our findings or what bias it may introduce.

Additionally, our "finer" data matches that use middle initial/names and suffixes seem to suffer from the fact that not all police officers have their full names listed in the policing data.

While a big portion of the data has the listed officer names using the *last name, first name and middle initial, suffix (if any)* format, given that there are so many different police forces represented in the data, the naming convention is not entirely standardized. This prevents us from matching more officers and being more confident in our match quality. However, despite these limitations, we will see below that regardless of the strategy we took in matching the names, the general trends of speed discounting seem to be consistent.

Analysis

The first plot below, Figure 1, shows the entire distribution of the share of differences in actual speed reported and the posted speed. We have no reason to believe that individuals are going exactly 9 miles per hour over the speed limit (especially as compared to the levels at 8 miles per hour and 10 miles per hour). In the economics field, this is referred to as bunching, where samples are disproportionately on one side of a threshold, which indicates some form of manipulation or cutoff. We will analyze how bunching behavior varies by demographics of the drivers as well as police officers. The following figures (Figure 2-5) in the next few pages will show a finer breakdown (by gender, race, and political affiliation) of the distribution created using each type of name matching.

Figure 1: Overall Distribution in Differences in Posted - Actual Speed

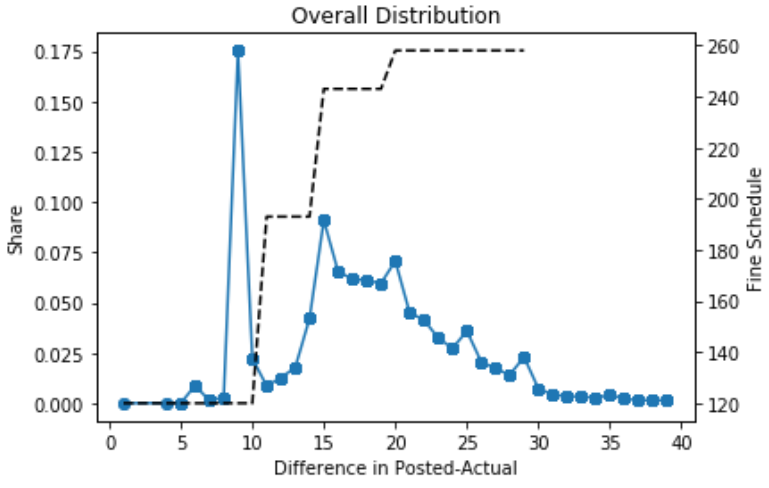
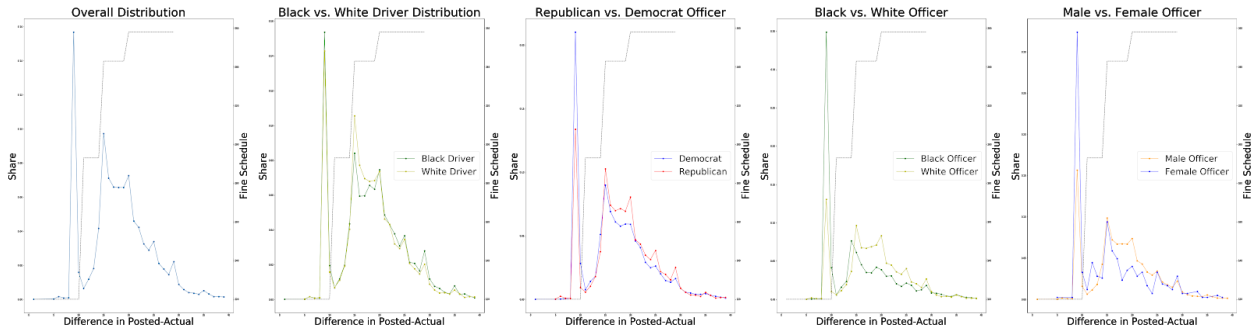


Figure 2 below was created using the dataset that was created by merging the voter data and police stop data using the first and last name only. This figure shows the distribution broken down by race, gender, and political affiliation using 5 separate panels. The first panel shows the overall distribution for the stops that were linked using the first and last officer name. The second shows the distribution of these stops by black vs white drivers, the third panel shows the distribution by republican vs. democrat officers, the fourth panel shows the distribution by white vs black officers and finally the fifth panel shows the distribution for male vs. female officers. There are 30,211 stops that were able to be matched based on the police officer's first and last name, which constitute 35.5 percent of all stops.

Figure 2: Distribution by Police Officer and Driver Demographics for those matched on First and Last Name



In the next page, we show figures 3-5. These figures all show the same panel analysis but using the dataset that were created using different name matching criteria when merging the voter and police stop data. Specifically, the data for figure 3 was created by merging based on the first, last, and middle initial, figure 4 based on the first, last, and middle name; and figure 5 based on the first, last, middle initial, and suffix. There are 15,200 stops shown in figure 3, representing 17.86 percent of all traffic stops over this period. There are 7,373 stops in figure 4,

representing 8.66 percent of all traffic stops and 2,082 traffic stops shown in figure 5, representing 2.45 percent of all traffic stops. As a note, all stops where the difference between posted and actual speed was over forty miles per hour are removed from this analysis because it is unclear what payment schedule fines over 40 mph are bound by. The 40+mph stops represent a small overall fraction of the stops.

Figure 3: Distribution by Police Officer and Driver Demographics for those matched on First and Last Name and Middle Initial

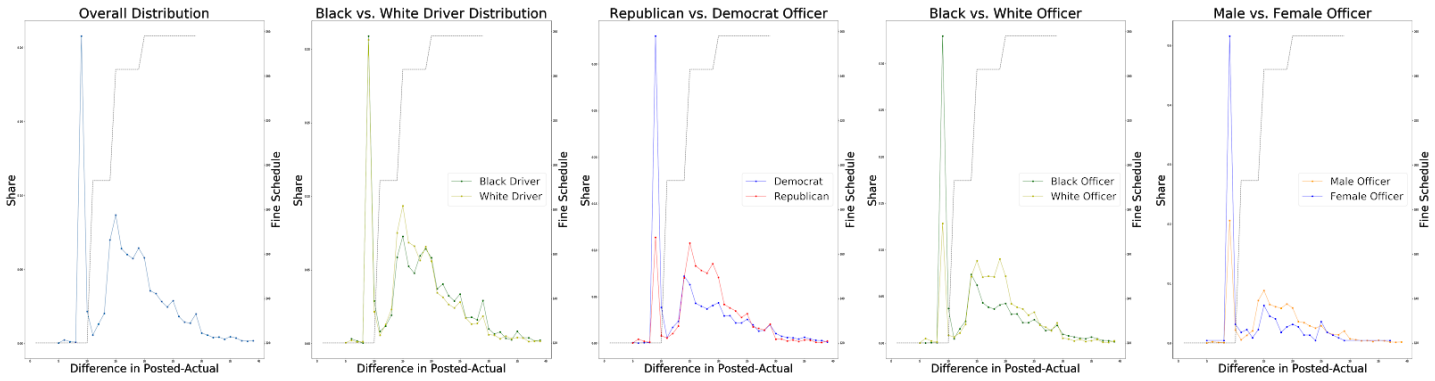


Figure 4: Distribution by Police Officer and Driver Demographics for those matched on First, Middle, and Last Name

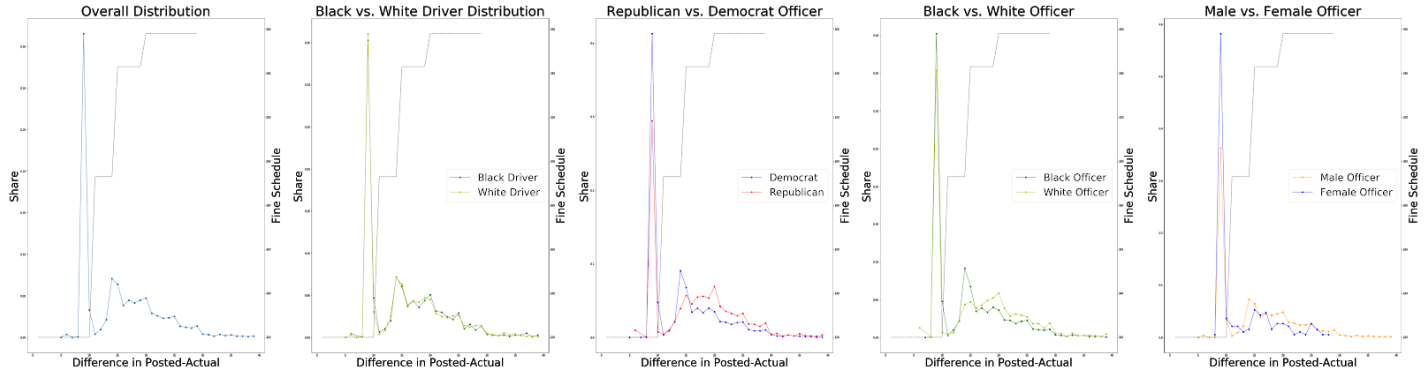
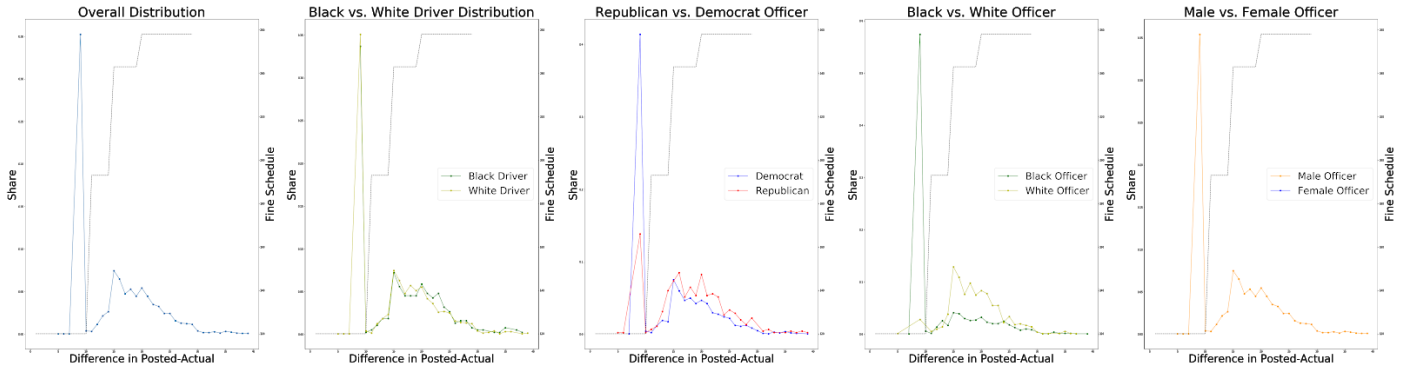


Figure 5: Distribution by Police Officer and Driver Demographics for those matched on First and Last Name, Middle Initial and Suffix



In Figure 2 - 5, we can see that the overall distribution looks similar to the distribution of all speed related traffic stops. It should be noted in all of the following analyses that we have not controlled for any covariates of the traffic stops; this will limit the conclusivity of our findings. In panel 2, we find that black drivers and white drivers seem to get a roughly similar number of tickets at exactly 9 miles per hour. It is impossible, from this analysis, whether to detect that this is because driver race does not impact police behavior, which is unlikely, as we have not controlled for the number of stops overall by race or by driving characteristics by race. For example, if whites/African Americans are more likely to speed then it may be that while the trends look equitable but really whites/African Americans benefit from higher rates of speed discounting that are not shown in this graph. Without additional information, we cannot make strong claims about equity.

Additionally, in the analysis by officer type, we have not controlled for police patrolling area or precinct differences. It is possible, for example, that for some reason black officers are more likely to patrol highways, where speed discounting is more common or be under police

management where speed discounting is more encouraged/discouraged. Given the size of the difference by covariates, the systematic differences would need to be large and be similar across race, gender, and political party. We are not able to directly rule out this explanation; however, it seems unlikely that this trend is driven solely by systematic differences in policing norms rather than differences in individual preferences .

Figure 2, panel 3 shows that democratic officers have approximately 22% of their sample at exactly 9 mph, while republicans only have about 13% of their sample at 9mph. At the 14mph cutoff, republican officers give out slightly more tickets than democrats. This may indicate that republican officers are speed discounting, but are not quite as generous with how much they decrease the speed amount as compared to democrats. These differences are even more pronounced in examining the difference between black and white officers. Black officers have around 35% of the share of tickets at 9mph, compared to only about 13% for white officers. Similar to panel 3, white officers seem to write more tickets at the next threshold of 14 mph than black officers. Finally, panel 5 shows the breakdown by gender. Women officers have about 33% of the share of their tickets at 9mph compared to about 15% for male officers. Unlike the previous panels, men and women officers seem to have the same share of tickets at 14mph but then men have a much higher share of tickets written for speed differences of greater than 15mph.

The trends in Figure 2 are largely echoed in Figures 3-5. In each subsequent figure, the sample size drops, due to difficulties discussed in the matching sections. In Figure 5 which matches names using suffix, there are no women officers who are linked in the sample, likely because women are less likely to have suffixes in their name. As a result, a comparison based on gender is not able to be made. As sample size decreases, the effect of one individual officer's

stopping behavior can have more influence on the distribution. So while we may be more confident in the quality of our matches as the level of specificity in name increases, robustness to outliers suffers as our sample size decreases. Despite these caveats, in all of the figures, the differences in ticketing behavior based on officer covariates are quite sizable and suggest that officer characteristics may play a role in officer-civilian interactions.

Discussion of Transparency

Even without the linkage, the dataset of police stops provided by the Hillsborough police allows individuals to look up the full name of the individual who pulled them over. If a disgruntled individual was upset about being pulled over, he or she can easily find the name of the officer and likely additional information about the officer through internet searches. It appears that through our linking, about a third of officers could be identified with just first and last name. One inelegant solution to this problem may be to not release any information on police stops, thus limiting the potential danger to any police officers. However, having information about individual police officers enables citizens to have some accountability about the police officers ticketing behavior. Simply removing names would remove the level of transparency that the dataset is supposed to provide.

Another potential solution is to take away names of the officers and replace them with a unique code that links officers across stops. The data could also provide demographics about police officers, such as race and gender. However, simply limiting the data to police agencies, race, and gender may not be enough to ensure anonymity. Using our dataset matched on first and last name, while admittedly a smaller sample, there are only 2 Asian American men in the

Hillsborough Sheriff's office and only one individual with unknown gender. It's unclear whether this person failed to fill out the gender category or identifies as a gender other than male/female. Adding in the full sample may mitigate these concerns, however, it is something that should be considered before releasing "anonymized" data.

Finally, there is the troubling issue of whether we should be providing privacy to drivers as well as officers. As above, we could consider not including any information about driver characteristics. However, this may actually limit the ability of the data to speak to claims of racial bias. We, for example, would want to know if all police stops were of black men because that is likely indicating some form of bias on the part of the police. Additionally, researchers and activists may be interested in the distribution of individuals stopped by race. As a result, eliminating all driver characteristics does not seem like the optimal solution. We could consider implementing something similar to the police officers, where instead of name, address, and driver's license number, the information could be only on race, sex, and town, etc.. We may think that police officers should be allowed special provisions because disgruntled individuals may target them and thus we need to be less worried about releasing information on citizens or even that having one's name posted online with all their information is a form of punishment and potential deterrence. However, punishing individuals, especially those committing minor traffic fines with both a loss of privacy and a fine seems like an unreasonable punishment. Just like in the police setting, analyzing some form of k-anonymity is important. Using the 2017-2019 police stop data, if we collapsed based on race, gender, and town we have many (9642) one person bins. If only race and gender are provided, the dataset is 216 k-anonymous.

Conclusion

While traffic tickets may seem like low stakes interactions between the police and the public, they are one of the most common ways civilians and police officers interact. We have seen the catastrophic results when these interactions go very badly (i.e. Sandra Bland) but even when these interactions are not life and death, they can play a large role in shaping views on the legitimacy of policing and police officers. We considered whether police covariates play a role in speed discounting and found that officers who are women, minority, and democrats seem to write more tickets at exactly 9mph as compared to their male, white, and republican colleges, respectively. In addition to this analysis, we also critically considered our ability to make a linkage between datasets. In Latanya Sweeney's work, we saw the power and potential harms of linking public databases. In our setting, linking voter rolls with police officer names may allow disgruntled individuals to target police officers. We considered potential remedies that would make the type of linking we have done if not impossible, much more difficult to do. Finally, we discussed improving privacy for drivers in this dataset and potentially reducing the amount of sensitive information released. For both police officers and drivers, we investigated what level of k-anonymity would be generated with the amount of information provided.

Abstract:

While traffic tickets may seem like low stakes interactions between the police and the public, they are one of the most common ways civilians and police officers interact. As a result, they can play a large role in shaping views on the legitimacy of policing and police officers. We considered whether police covariates play a role in speed discounting and found that officers who are women, minority, and democrats seem to write more tickets at exactly 9mph as compared to their male, white, and republican colleges, respectively. In addition to this analysis, we also critically considered our ability to make a linkage between datasets. In Latanya Sweeney's work, we saw the power, and potential harms, of linking public databases. In our setting, linking voter rolls with police officer names may allow disgruntled individuals to target police officers. We considered potential remedies that would make the type of linking we have done if not impossible, much more difficult to do. Finally, we discussed improving privacy for drivers in this dataset and potentially reducing the amount of sensitive information released. For both police officers and drivers, we investigated what level of k-anonymity would be generated with the amount of information provided.

Works Cited

Arnold, David, Will Dobbie, and Crystal S. Yang. "Racial bias in bail decisions." *The Quarterly Journal of Economics* 133, no. 4 (2018): 1885-1932.

Goncalves, Felipe, and Steven Mello. *A Few Bad Apples?: Racial Bias in Policing*. Industrial Relations Section, Princeton University, 2017.

Ouss, Aurelie, and Megan T. Stevenson. "Bail, Jail, and Pretrial Misconduct: The Influence of Prosecutors." *George Mason Legal Studies Research Paper No. LS* (2020): 19-08.

Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 05 (2002): 557-570.

Sweeney, Latanya. "Simple demographics often identify people uniquely." *Health (San Francisco)* 671 (2000): 1-34.

Tuttle, Cody. "Racial Disparities in Federal Sentencing: Evidence from Drug Mandatory Minimums." Available at SSRN 3080463 (2019).