Exploring Missingness and its Implications in Traffic Stop Data

December 2020

Abstract

As traffic stop data has become increasingly available, so has scholarship for analyzing the data for evidence of discriminatory policing. However, relatively few studies address missingness (NA values) in the data despite all the data being conditional on recording. This project develops a framework for studying missingness through the stop missingness rate (SMR) and presents exploratory data analysis of SMR on data from the Stanford Open Policing Project. Using the SMR, we observe trends in the SMR across variables date and day/night; such trends provide descriptive evidence of missingness as a confounding variable. We run several logistic regressions for data grouped by distinct missingness patterns and observe changes in the significance and magnitude of some race and sex variables. The possibility of missingness as a confounding variable calls for further research in its trends and impacts.

Introduction

Every year, more than 20 million traffic stops are conducted in the United States, representing the most common way in which drivers interact with the police (Davis, Whyde, and Langton 2018). Unfortunately, not all traffic stops are conducted equally – in a 2019 survey by the Pew Research Center, 59% of Black men and 31% of Black women say that they have been unfairly stopped by the police (Anderson 2020). The relationship between belonging to a minority race, such as Black, Latinx, and/or Asian Pacific Islander Desi American, and experiencing discriminatory policing is significant historically and in the present. Since the late 90's, popular concern over racial profiling has led to federal and state mandates requiring the collection of traffic stop data (Russell 2001).

In tandem with the increasing availability of such data is the growing interdisciplinary scholarship analyzing such data for evidence of racial profiling (Baumgartner et al. 2017; Grogger and Ridgeway 2006; Pierson et al. 2020; Smith and Petrocelli 2001). For example, the Stanford Open Policing Project (HYPER LINK THIS https://openpolicing.stanford.edu/), whose data are the focus of this study, has several dozen datasets representing over 100 million separate traffic stops. Scholars have developed statistical methods to test for racial bias, although racial bias is challenging to prove because of the unknown baseline describing how traffic stops would be conducted in the absence of any bias.

Much of the work on discriminatory policing has thus been focused on circumventing this unknown baseline. However, few studies mention how the results of their statistical analysis are impacted by the quality of traffic stop data or the number of observations excluded due to variables that are left unrecorded. For example, a model that uses **time** as a covariate will exclude traffic stop observations for which **time** is NA, or missing. If we assume that traffic stop data is the product of a myriad of human factors – recording practices, state and department policy, human error, resource constraints, individual officer compliance, and racial bias – we begin to distinguish between how a traffic stop is conducted and how it is recorded. By framing traffic stop data as dependent on *data collection practices* as opposed to the actual traffic stop, we study it limitations in describing the phenomenon of traffic stops and, as a result, racial profiling.

This project is about missingness in traffic stop data – unfortunately, not all traffic stops are *recorded* equally. Missingness refers to a variable that is left unrecorded; it is important because traffic stops that have unrecorded variables will be excluded by most statistical models models that test for racial bias. For example, if I were pulled over for a routine traffic stop but the officer failed to record my race, then a logistic regression modeling being searched as a function of driver race and other covariates would altogether exclude my traffic stop.

By framing traffic stop data as the by-product of a myriad of human factors and noting the necessary condition of full data coverage before modeling, we can pose the following questions: are there fundamental differences between the traffic stop observations with high and low missingness? Does missingness have a certain trend with respect to pertinent variables like race and time? Most importantly, are the missingness trends drastic enough to render two subsets of the same dataset incomparable?

With such guiding questions, we conduct a descriptive study exploring trends of missingness across the traffic stop data available through the Stanford Open Policing Project. The benefit of using such data is that this study is not limited to just one dataset; we analyze several dozen datasets to broadly explore differential missingness. This project is also concerned with how missingness could potentially impact the quality of analyses using such data – we attempt to study if and how such trends introduce bias to statistical models.

This project incorporates both data exploration and modeling. We begin by developing a metric for missingness and use it to explore the data: while the data do not provide evidence that traffic data are recorded differentially depending on the race of the driver, we uncover one problematic example of interaction between the recording of the **race** variable and arrests being made in Seattle, Washington. Visualizations for the missingness by **date** of the year and the **time** of the day (represented by a **day/night** variable) demonstrate several trends of temporal missingness that we discuss. Finally, we identify datasets with distinct patterns of missingness and run the same logistic regression model for subsets of the datasets. This final portion of the study demonstrates how the recording of traffic stop data should be treated as a confounding variable and investigated more deeply.

Literature review

This exploratory analysis is motivated by Chanin and Welsh's mixed method review of San Diego Police Department traffic stop data (2020). They both analyze missingness in traffic stop data and conduct officer interviews to study the quality of the data and officers' attitude toward its collection. They point to the lack of meaningful consideration of data quality in existing traffic stop literature: of the 100 papers reviewed, the authors find only 19 papers that address missingness and recording rates (pp. 4). In a simple univariate analysis of the missingness by race, they find that stops involving drivers identified as either Black or Latinx were more likely to missing data (pp. 12). Summarizing the psychological, cultural, and organizational barriers to compliance, "an officer who does not see driver race as affecting their own decision-making... data collection only redounds to their detriment" (pp. 7).

The fundamental problem behind testing for discrimination is the lack of an appropriate benchmark which would relay how traffic stops are conducted in the absence of racial profiling (Ridgeway and MacDonald 2010). For example, one inappropriate but perhaps intuitive external benchmark is US Census data that details the racial makeup of the residents in a city. The assumption that the residents, drivers, and drivers committing traffic violations are identical and racially indistinguishable populations is problematic: not all residents drive, and especially for cities near highways, not all drivers are residents. Furthermore, comparing stop rates or search rates by race with the residential racial composition obscures a range of confounding variables (pp. 3-5). Demonstrating causality between the race of the driver with a punitive traffic stop outcome is thus difficult. The lack of a benchmark has motivated a range of statistical methods, which we briefly review here, heeding the mentions of data sourcing, quality, and missingness.

Traffic stop studies often use logistic regression to explain trends in post-stop outcomes, such as a **search** or an **arrest**, with covariates from the data such as driver **race**, the **time** of the stop, and variables for officer characteristics (Baumgartner et al. 2017; Smith and Petrocelli 2001). Smith and Petrocelli (2001) use data collected over a six-week period in Richmond, Virginia; the authors note that the data had a 64% response rate, meaning that 36% of traffic stops were unrecorded. However, whether or not the 64% of traffic stops that were recorded contained missing values is not explicitly addressed. The authors suggest caution given the "potentially different pool of traffic stops" of the 36% unrecorded traffic stop observations and recommend for future research to compare results (p. 9). They find evidence that being of minority race increases the probability that a driver receives a **warning** as opposed to a **citation**, which is a less punitive outcome.

In contrast, Baumgartner et al. (2017) compile 55 million traffic stop observations from approximately one dozen states in the United States; they find strong evidence of racial disparities in the post-stop outcome of **search**. The study includes a summary of the state traffic data policies and an appendix tallying the observations that are excluded due to missingness and other reasons. Unfortunately, the implication of missingness on the performance of the model is not discussed.

Ridgeway (2006) point out the incompatibility of logistic regression with traffic stop studies and propose propensity score matching as an alternative method. As a generalized linear model, logistic regression estimates are sensitive to the model specification. Importantly, the traffic stops involving minority and white drivers can be quite different, so significant **race** coefficients could result from confounding variables. Estimating propensity scores for an Oakland, California dataset through generalized boosted models, the researchers find that Black drivers are treated equitably in terms of citations and consent searches. Missingness is not discussed in this study.

Developed by Grogger and Ridgeway (2006), the veil of darkness (VOD) method is an instrumental variables approach that uses the natural variation of daylight (and darkness) resulting from daylight savings (DST) as an instrumental variable. The researchers address both missingness and recording rates: they omit about 1,000 observations due to missingness out of 7,600 original observations. While they do not mention the implications of missing data, they analyze the recording rate with respect to the VOD method – as opposed to the standard logistic regression models, the VOD method is more robust for data with a sizable amount of unrecorded traffic stops because it depends on a weak assumption that race-specific reporting rates do not vary between daylight and darkness hours.

A recent analysis of over 20 million traffic stops applies the VOD test and threshold test Pierson et al. (2020).

The threshold test is an extension of the outcome test, incorporating both the rate of searches and rate of successful searches to determine if the police officers use a lower standard for searching POC drivers. The researchers find evidence of both racial profiling and preferentially searches for minority race drivers. Although the discussion primarily focuses on standardizing and improving the quality of traffic stop data collection, it does not mention how missingness plays a role in the analyses.

We extend the data quality review from Chanin and Welsh (2020) in two ways: first, we examine the presence of missingness across multiple datasets as opposed to just one; second, we examine missingness trends across particular variables through the SMRs by **race**, **week**, and **day/night**. For other traffic stop researchers, this project provides an approach for how to explore and conceptualize missingness in one dataset or many.

Data and Methods

Defining Missingness

We define missingness as occurring when a variable is left unrecorded in a dataset. If a variable takes on the value NA, then we deem it to be missing. In contrast to Chanin and Welsh (2020), we refer to an unrecorded variable as "missing" rather than an "error" in order to prevent confusion regarding whether an "error" is due to variables not recorded or recorded incorrectly.

To quantify missingness, we define the stop missingness rate (SMR) to be the percent of all variables that are unrecorded for a single traffic stop, represented by one row of the dataset. A higher SMR means that a larger percentage of the information is not recorded or equivalently, that less information is available.

$$\text{SMR} = \frac{\text{number of unrecorded variables}}{\text{total number of variables}}$$

$$\text{dataset SMR} = \frac{1}{T} \sum_{i=1}^{T} \text{SMR}_i \qquad\qquad T = \text{total number of traffic stops} \qquad (1)$$

As an explanatory example, we consider an abridged traffic stop dataset from Oakland, California that records four variables (Table 1). For a given row, if an officer were to leave one of the variables blank, then the SMR for that row would be 25%. The SMR for an entire dataset can be found by taking the average of the SMR over the traffic stops. We chose the mean as opposed to other measures of central tendency because taking the average row-wise SMR (meaning the average of the SMR per traffic stop) is equivalent to the percentage of missing values for the whole dataset, which is an important aspect for the dataset SMR to capture. With this definition, we can see how the dataset SMR for Table 1 is 50%.

Table 1: SMR for Oakland, CA

race	age	date	outcome	SMR
black	NA	2014-09-17	NA	0.50
hispanic	NA	2016-05-24	citation	0.25
hispanic	NA	2014-06-07	citation	0.25
black	65	2017-02-10	citation	0.00
white	NA	2014-03-01	citation	0.25

The dataset SMR is not the only missingness metric that can describe a dataset's data coverage. We can also determine the missingness across the levels of a categorical variable. The SMR by a variable of interest

enables comparison across datasets of the missingness trends of that variable. We provide a general definition of SMR by a variable first before walking through an example of SMR by race.

Let V denote a variable in a dataset that takes on $k \in \mathbb{Z}$ distinct values. Let V_j denote the set of observations for which $V = j$ for $j = 1, 2, ..., k$. For an observation i such that $V = j$, we say that i is an element of V_j. Then, let $T = \sum_{i=1}^{k} n_j$ be the total number of observations in the dataset, where n_j denotes the number of observations for in V_j. The follow equation gives the SMR by V for each value of $j = 1, 2, ..., k$:

$$\text{SMR by } V = \frac{1}{n_j} \sum_{i \in V_j} \text{SMR}_i \tag{2}$$

The SMR by race, for example, is the average SMR of observations belonging to each racial group. Let R denote the variable for race with k distinct values. Let R_j be the set of n_j traffic stops involving individuals who are recorded as race j. The SMR by race is defined as:

$$\text{SMR by race } = \frac{1}{n_j} \sum_{i \in R_j} \text{SMR}_i \tag{3}$$

Notably, the calculation of SMR by race will result in an SMR for drivers of missing race, whom we refer to as NA-race drivers. An SMR for NA-race drivers is calculated because NA race is indeed a level of driver **race**. For this reason, we do not include **race** in the calculation of SMR by race so as not to artificially inflate the race SMR for missing race drivers. Continuing our example with the abridged Oakland data, Table 2 details the SMR for different racial groups.

Table 2: SMR by race for Oakland, CA

race	SMR by race
black	0.25
hispanic	0.25
white	0.25

The SMR by **race**, **week**, and **day/night** for a dataset can be determined only if the dataset includes records of each of those respective variables. A filtering step to isolate comparable datasets is thus required prior to computing SMRs. Moreover, we encourage a stricter pre-processing step before computing any SMR to enable fairer and more accurate comparisons among datasets.

We caution against comparing the quality of data collection by directly comparing the dataset SMR or race SMRs across datasets. Each dataset has a unique set of variables, with some cities or state patroles recording more variables than others. An average SMR for a dataset containing three variables cannot be compared with an average SMR containing three dozen variables.

In addition, traffic stop data from Los Angeles, California are collected with only x variables, of which driver **race** is not included. In comparison with the Oakland data collected with y total variables including driver race, the Los Angeles data cannot be used to study racial profiling, despite having a close to zero average SMR. A low SMR for the Los Angeles data does not necessarily indicate that the dataset has greater coverage and potential for analysis compared to a dataset that collects more variables, albeit to some degree of missingness.

In an effort to more fairly compare SMRs across datasets, we pre-process the data to remove uncommon variables, like **vehicle type** and **officer age** so that the SMR is determined only from the frequently recorded variables, which we explain in the next section. The average SMRs for datasets that are more similar can thus be more fairly compared with each other.

To summarize, the SMR by **race** can be calculated such that it is more fairly comparable across datasets. First, we filter for datasets that contain traffic stops that record **race**. Next, we remove uncommon variables in each dataset, selecting only for frequently recorded variables. Lastly, we exclude race from the numerator of NA values in (3) to calculate the SMR by **race**. This method can be repeated for SMR by **week** and **day/night**.

For the remainder of this paper, all mentions of dataset SMR or SMR by a certain variable have been computed using the process detailed above; that is, all SMR statistics have been calculated after filtering and pre-processing the datasets.

About the Data

The data used here come from the Stanford Open Policing Project which includes information on vehicular traffic stops from 1999 to 2020; the most frequently recorded variables relating to driver demographic are **race**, **sex**, and **age**; those relating to situational information are **time** of day, **date**, **location** (and **latitude** and **longitude**), and **outcome of the stop**. A "dataset" refers to the data collected by a municipal police department or a statewide patrol agency. We study a total of 66 datasets, all of which contain at least one of the aforementioned frequently recorded variables. Figure 1 provides an overview of how many observations make up these datasets, and Figure 2 relays the distribution of SMR for each dataset.

Figure 1: Number of observations in traffic stop datasets

Figure 2: Dataset SMR of traffic stop datasets
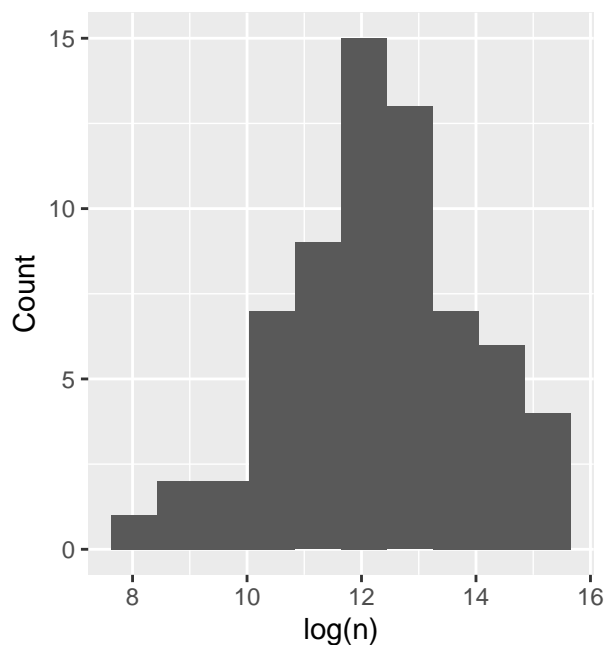
A histogram for the log number of observations in 66 datasets. The y–axis counts the number of datasets.
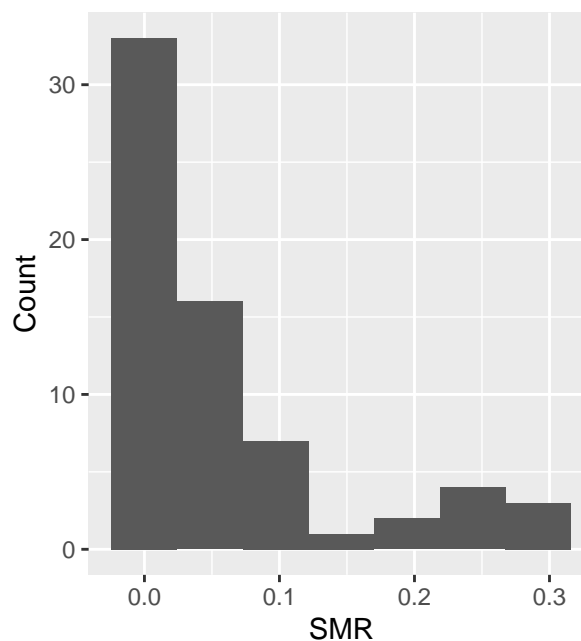
A histogram of the dataset SMR for 66 datasets. The y–axis counts the number of datasets.

Of the total 66 datasets analyzed in this study, 32 of the datasets record **search** conducted (which indicates whether or not a search was conducted), and 33 record **arrest** made. As the dependent variable for most traffic stop studies, the **search** and **arrest** variables are also relevant outcomes for this project because they represent harsh stop outcomes. The driver characteristic and outcome variables can be understood as the minimum covariates needed to run a logistic regression for **search** or **arrest**.

To contextualize the datasets, we present in Table 3 a survey of state legislation requiring the collection of traffic stop data. This overview is helpful in understanding how the data are collected and the variables that

are deemed important from lawmakers' points of view.

Table 3: State Mandates for Traffic Stop Data Collection

State	Outlawing.racial.profiling	Use.officer.perception	Bill.number
CA	Yes	Yes	AB 953
NC	No	Does not mention	GS 143B
NY	Yes	Yes	A03949
TN	Yes	Yes	HB 2167
WA	Yes	Does not mention	RCW 43.101.410

One notable trend of these policies is that usually the **race**, **gender** identity, and **age** is recorded based on officers' perception. The Stanford Open Policing Project's own analysis of its data points out this detail Pierson et al. (2020). We can see evidence of this in how the age density plots of North Carolina datasets spike at multiples of five in Figure 3, but we can also see evidence of officer perceptions in the extensive exclusion of non-binary gender identities and mixed-race identities and the common inclusion of the Hispanic, which is an ethnicity, as a level for race. Another important feature of such state mandates are the different variables required for each state. We are able to compare data quality and missingness across datasets with different variables through the pre-processing step noted in the previous section, but datasets that collect neither **search** nor **arrest** are outside the scope of most statistical models testing for discriminatory policing.

Figure 3: Age density for North Carolina datasets
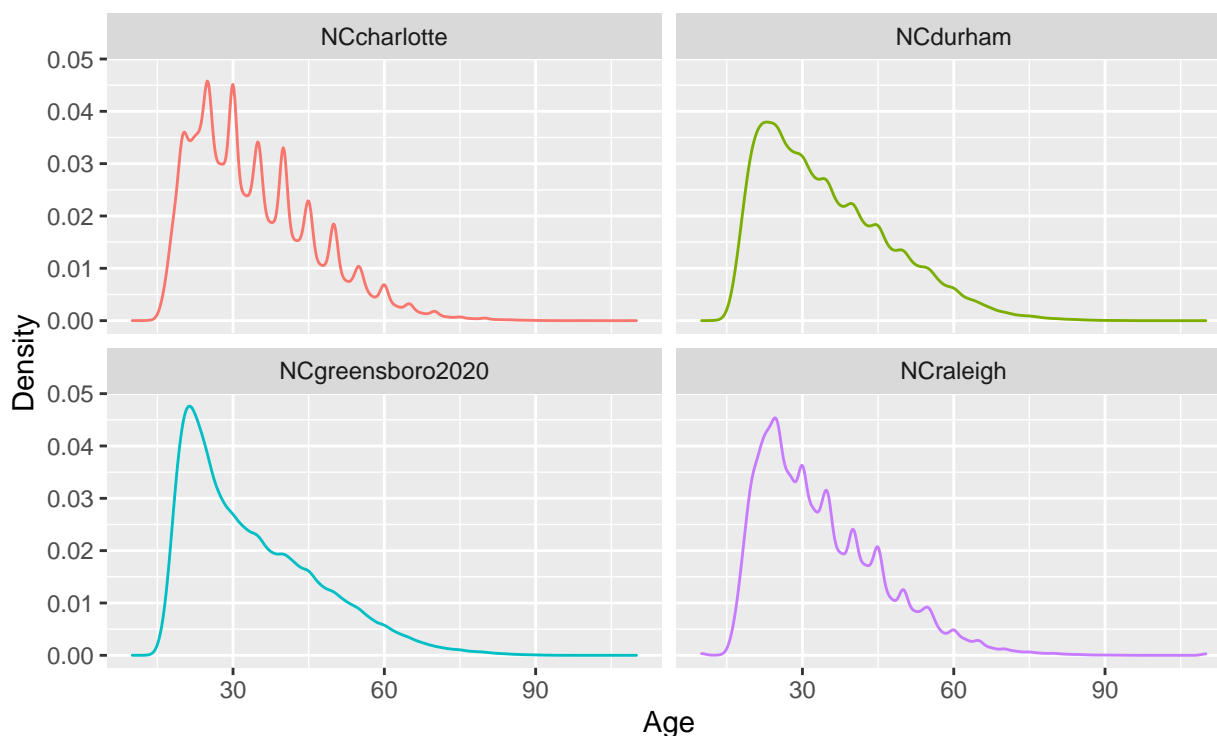
Density plots of driver age for the North Carolina datasets. Each plot consists of more than 130,000 observations.

Methods

The primary method we use to explore trends in missingness is visualization; in particular we consider the SMRs by **race**, **week**, and **day/night**, along with the **search** and **arrest** rates by **race**. Due to the

large amount of data, the SMR statistics and visualizations are created with a 30% random sample of the data. The data are stored in a mySQL database, from which we query, wrangle, and visualize the data entirely in R. The complete set of visualization and the code for my work can be found on my Github: https://github.com/Amber-Patricia-Lee/US-Traffic-Stops/.

Through data visualization, we explore how SMR varies with another variable on both individual dataset and aggregate of all datasets levels. Since the aggregate plots combine information from all datasets into one visualization while the individual dataset plots do not, this paper presents a few selected individual-level plots; however, all the visualizations and code are Github. This project thus takes quite a descriptive approach towards investing missingness on a dataset and across datasets scope. The visualizing strategies developed here can be taken as a framework for exploring and detecting trends in missingness.

To see if and how missingness affects the relationship between race and punitive stop outcomes, we use logistic regression. Despite its limitations and the difficulty of interpreting log-odds, logistic regression on the post-stop outcome is one of the most commonly used models. Furthermore, the purpose of this study is to investigate missingness as a confounding variable, not to provide a test for proving racial profiling. We use a model similar to that in Baumgartner et al. (2017).

Logistic regression is a statistical model that relates a set of covariates with a binary dependent variable (such as **search** vs. no search or **arrest** vs. no arrest) using the logit function. The estimated coefficients describe how the logarithm of the odds (log-odds) of the response variable changes when increasing the respective covariate. Our model is as follows:

$$\text{outcome} \sim \text{race} + \text{sex} + \text{age} + \text{day/night} * + \text{day of the week}$$

where **age** is the only continuous variable; the other variables are factors. We mark day/night with an asterisk because it will not always be used as a predictor variable.

To see how missingness influences the regression results, we first identify a dataset with distinct patterns of missingness. Then, we divide that dataset into groups so that the observations in each group exhibit the same missingness pattern. Then we run the same model over the distinct groups and also the combined data.

The purpose of these regressions is not to test for discrimination or racial profiling; rather, we pose the question of how missingness may potentially interfere with the models that employ data with differential missingness.

Results

Race

We begin with investigating how missingness varies by **race**. A total of 57 datasets out of 66 record the **race** of the driver; these 57 datasets were used to plot the distributions of SMR by race and plot minority race SMR against white SMR.

Figure 4: SMR by race
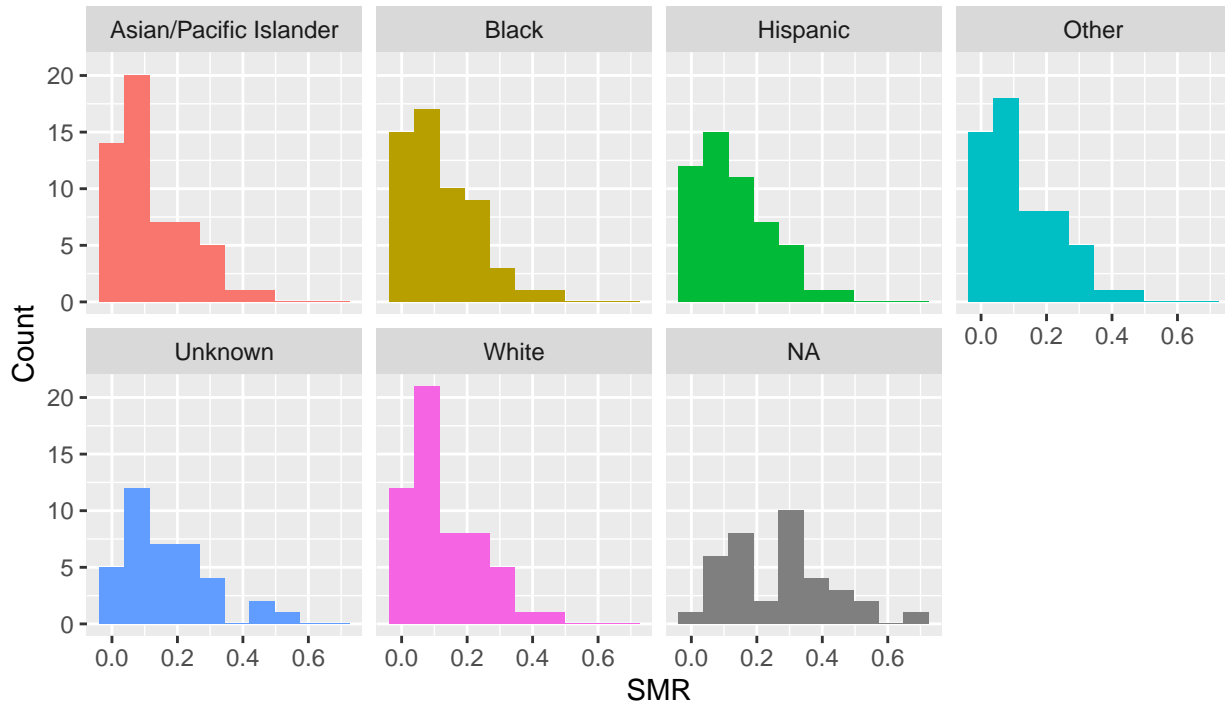
Histograms of the SMR by race for 57 datasets. The y−axis counts the number of datasets, and the scaling for the x− and y−axes is the same for all histograms.

Figure 5: SMR for minority race and white drivers across datasets
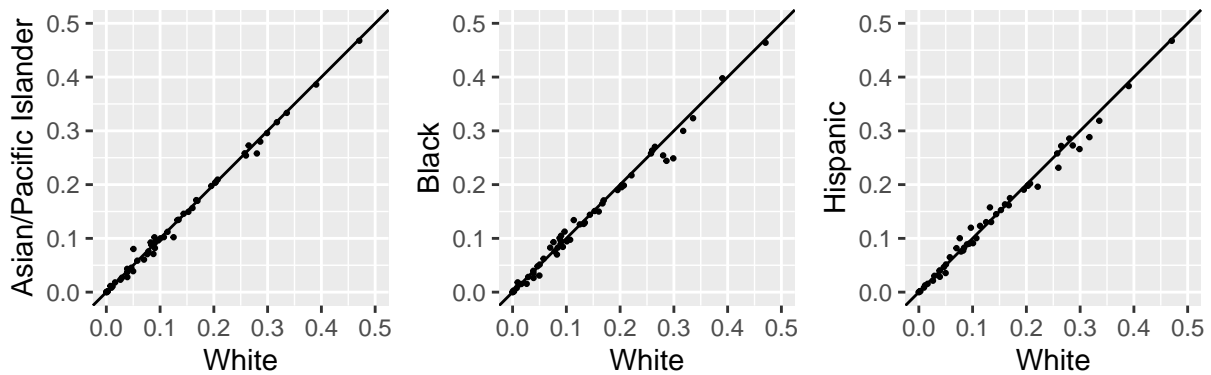
Scatterplots of SMR by race for 57 datasets, with minority race (APIDA, Black, and Hispanic) SMR plotted against white SMR. Each point represents one dataset.

As seen in Figure 4, the distribution of SMR by **race** is comparable with one another, with the exception of drivers whose **race** is not recorded. The distribution likely differs for NA-race drivers because of the fewer data points that make up the SMR calculation; additionally, some datasets have no missing recordings of **race**, so the SMR for NA-race drivers does not exist in those datasets.

In Figure 5, we see that SMRs are quite comparable with each other across **race**, implying that the data collection quality for drivers identified as APIDA, Black, and Hispanic are comparable to that for white drivers. Thus, we do not find different SMR trends between drivers identified as a minority race and as white. While Chanin and Welsh (2020) find that in San Diego, California, drivers identified to minority race (Black and Latinx) are more likely to have missing data, Figure 5 suggests that the overall occurrence of missing

data is comparable among race groups.

Race and Outcome

Although the previous analysis did not provide evidence of a trend between missingness and certain **race** values, we continue how officers record traffic stop data involving NA-race drivers. This is relevant because NA-race drivers are altogether excluded from most racial profiling studies.

This section concerns the stop **outcome** of drivers by **race**. The outcomes of interest are **search** and **arrest** because they are more punitive than **warning** and **citation**. Also, the decision to search a driver involves a degree of officer discretion, which can be susceptible to bias. We cannot discern how NA-race drivers actually experience traffic stops and if they are subjected to harsher post-stop outcomes than white or minority race drivers due to confounding variables. However, we use the available data and descriptive statistics to understand the rates at which drivers of different races are either **search**ed or **arrest**ed.

We do not use SMR as a predictor variable in modeling **search** or **arrest** rates; SMR is used as a tool for understanding changes in models across **time**, **race**, and **day/night**. The search rate by **race** is the number of stops involving drivers from a race group who are searched divided by the total number of stops involving drivers of that race group. The arrest rate is calculated similarly.

Both search and arrest rates are conditional on the driver being involved in a stop – a 10% **search** rate for white drivers should be interpreted as 10% of white drivers who are pulled over and recorded experience a search, which is distinct from the interpretation that 10% of all white drivers are searched.

To compare the post-stop recorded outcomes of drivers of different races, we compute the **search** and **arrest** rates by the different race levels. Then, we plot the Black and white search rates against the NA-race search rate; we do the same for **arrest**.

Figure 6: Post–stop outcome rates for Black, white, and NA–race drivers
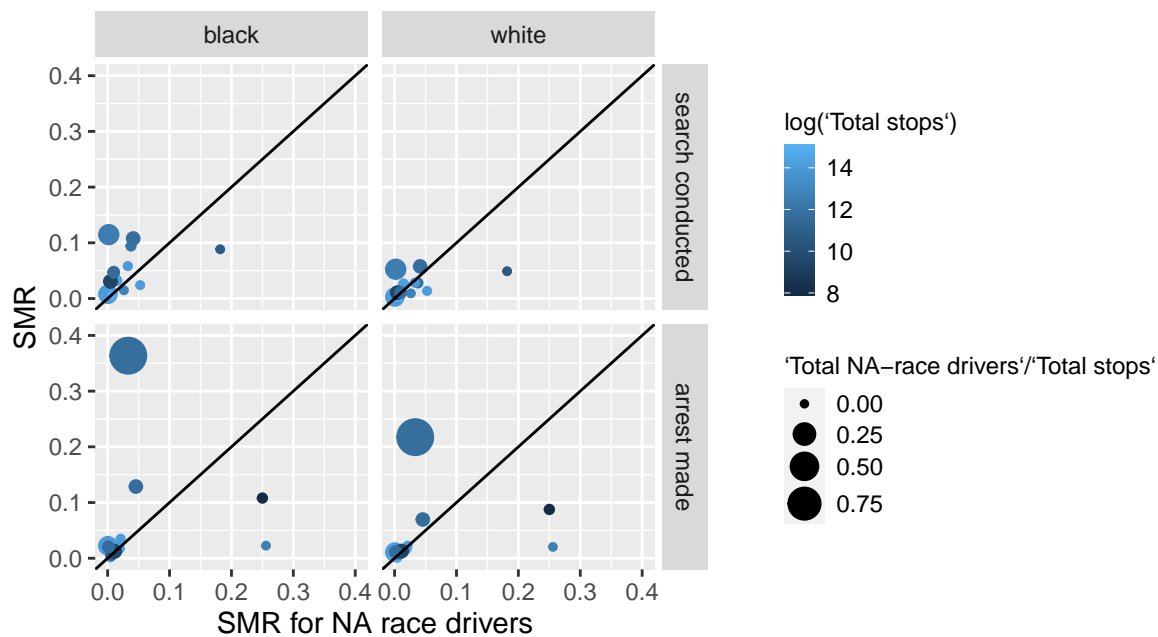
Scatterplots for the search and arrest rates of Black and white drivers compared to NA–race drivers.
The search rate plots represents 32 datasets, while the arrest rate plots represent 33.

We read the scatterplots in Figure 6 from left to right, comparing the search rate relationship between Black and white drivers (on the y-axis) with the search rate of NA-race drivers, on the x-axis. The points in the right column (representing the search rate of white drivers) look more closely bunched to the y = x line than the points in the left column (that represent the search rate for Black drivers). The Black search rates that are above the y = x line provide descriptive evidence that drivers who are stopped and identified as Black experience search and arrest rates higher than do NA-race drivers.

Next, we discuss the size of the scatterplots that relay the percentage of traffic stops that involve NA-race drivers. Consider the traffic stop data from Seattle, Washington that is represented by the largest point in the **arrest** scatterplots. The large point provides evidence of interaction between the *recording* of **race** and *outcome* that a stop results in an **arrest**. Table 4 presents the full consideration of the data including a summarized count of the arrests and non-arrests by race for a 30% random sample of the data. Out of the drivers who have been arrested, just over 0.3% of the individuals have **race** recorded as Black or white; in contrast, a much lower percentage of drivers (0.04%) are identified as white, and an even lower percentage of drivers (0.02%) are identified as Black. While the actual counts are quite small, the differences across both **arrest** and **race** are important and worth understanding prior to any analyses associated with **race** on this dataset.

Table 4: Driver Race and Arrests in Seattle, Washington

Driver race	No arrest	Arrest
Asian/Pacific Islander	14	4
Black	21	12
Hispanic	5	1
Unknown	1	0
White	54	15
NA	120289	4146

The number of NA-race drivers who are not arrested is orders of magnitude larger than the other counts; the denominator of the arrest rate is thus biased by the missingness, unable to accurately represent the amount of non-arrests in each race group. As a result, the arrest rates by race do not indicate the true percentage of stopped drivers of a certain race who are arrested. Missingness by race and outcome is important because a large presence of NA-race drivers can introduce bias into the outcome rates. Traffic stop data in Seattle, Washington provide a cautionary example of how informative trends regarding race cannot be captured if the *recording* of race is linked to an outcome variable.

Week

Next, we address how the weekly SMR varies over the years. To compute SMR by **week**, we first categorize each traffic stop into a one-week period based on its recorded **date.** Then, we apply the necessary filtering and selecting steps before applying equation (2) to calculate the average SMR by **week** for each dataset. We plot the weekly SMR value against the **date**. All of the 66 datasets in our analysis record **date** so we generate 66 visualizations; the plots shown in Figure 7 were selected to illustrate the common trends observed throughout all the plots, the complete set of which can all be found on my Github.

Figure 7: Weekly SMR of Selected Datasets
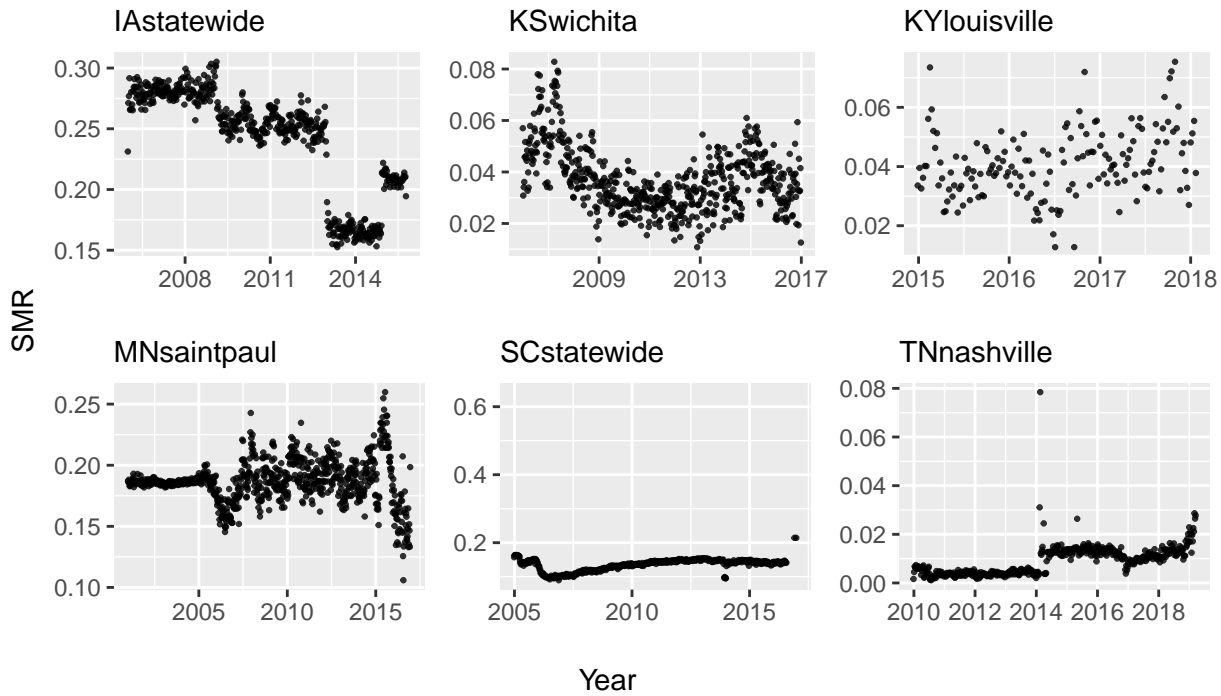
Scatterplots of the weekly SMR plotted by year. Note that the scaling of the y–axes is different so the details for each trend are clear.

We first discuss the plots for Iowa statewide and Nashville, Tennessee, which have roughly constant weekly SMRs with a few level changes. Such level changes are fairly common, with increasing or decreasing jumps identified in 15 of the 66 scatterplots. The trend of constant SMR plots with distinct level changes is possibly indicative of changes in department policy regarding variable collection: if the level drops, then the department has likely adopted a new variable to record, while a level increase might reflect a phasing out of one or more traffic stop variables. For example, the jump in 2014 for Nashville, Tennessee reflects the increasing missingness in **latitude** and **longitude**. Abrupt changes in weekly SMR is evidence that the recording of data can change with respect to the **date** of the year.

We also observe how the variance of the weekly SMR through time is different across datasets and, occasionally, different within a dataset. Regardless of the functional relationship between time of **year** and the weekly SMR, the data collection practices of some departments are more regular than others. Within a level set, the SMR values calculated for the Iowa dataset are more spread out than the SMR values in Nashville, TN; the plots for Wichita, Kansas and Louisville, Kentucky have similar spread (note the different axes labeling). Traffic data from Saint Paul, Minnesota has low variability until approximately 2005, then it begins to increase. Such a display of SMR value is quite rare across the datasets we examine, but we include this plot to convey the idea that some datasets exhibit high spread, while others exhibit low spread. Understanding the missingness associated with a particular dataset can be paramount to correct model assessments.

Consider the suite of plots given in Figure 7: they demonstrate that the trends of missingness (as measured by SMR) can be functional, varied, and difficult to interpret. For example, both linear and concave relationships that can be observed in the total 66 plots. The plots for Wichita and South Carolina statewide provide examples of concave relationships that sometimes change through the years. The unique, identifiable, but difficult to explain pattern in South Carolina is rare, but we include the plot to demonstrate how unique these functional relationships can be. In contrast, the plot for Louisville, Kentucky presents a common and relatively mild form of the relationship between weekly SMR and **date.**

Understanding the impact of missingness in a dataset is a difficult task, but it is incredibly important for any

model interpretation. The implication of a functional form existing between SMR by **week** and **date** of the year is that the data, which are conditional on their collection, may not be comparable even within the same dataset. Furthermore, missingness as a confounding variable can further obscure methods to demonstrate racial profiling.

One final trend to report is not visualized in Figure 7, but it is still indicative of the importance of understanding data missingness and recording practices. Some departments periodically update their traffic data and enter the date of the traffic stop as the date that the data are entered. For example, the Oregon statewide dataset that spans five years has a weekly SMR plot with only five points (i.e., five recorded values on the x-axis) because the department updates its traffic data annually. This finding is immediately relevant for models involving variables like **time**, **date**, and **sunset/sunrise** times as covariates. For such models, the Oregon data would be inappropriate to use.
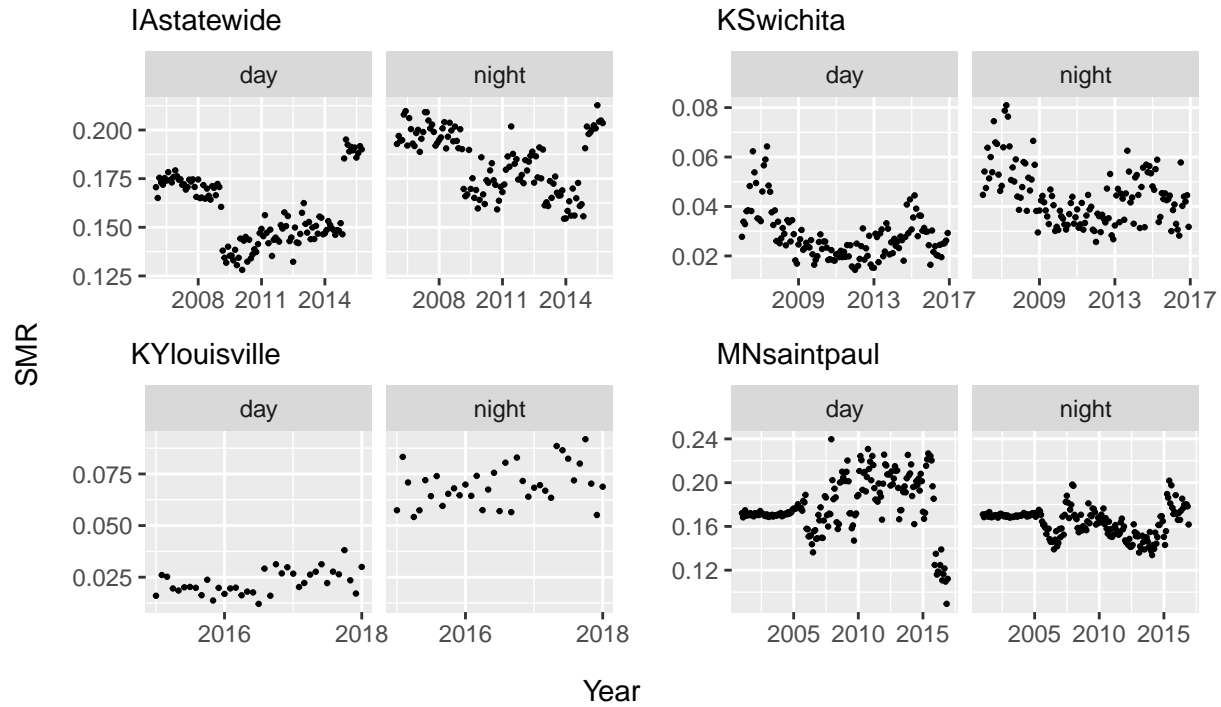
Day and night

For the last section of our exploratory data analysis, we extend on the weekly SMR by considering how missingness varies between stops occurring in the day compared to stops occurring at night. To categorize traffic stops as occurring during the day or night, we find the **latitude** and vlongitude** coordinates for each dataset; then, we find the sunrise and sunset times for each date of the particular location. To simplify the comparison, we exclude stops which occur in the times between dawn and sunrise and between sunset and dusk.

Rather than calculate SMR by**day/night**, which would result in only two statistics per dataset, we determine the monthly SMR for all stops occurring during the day for that month; we apply the same method for nighttime stops. After categorizing and filtering for traffic stops during the day and night, we filter the datasets, select the relevant variables, and apply equation (2) to calculate the monthly SMR for **day** and **night** stops.

We visualize the resulting statistics by plotting the monthly SMR throughout the year, distinguishing between daytime and nighttime SMRs. These visualizations relate three variables: SMR by **month**, the **year**, and **day/night**, so we can glean the relationships between each of the three variables; we refer to these as **day/night** SMR plots for brevity. 66 datasets record date and the time of the stop, so we generate 66 total plots. The plots shown in Figure 8 were selected to summarize common trends and extend from those in Figure 7; the remaining plots can be viewed through my Github.

Figure 8: Monthly SMR in Daytime and Nighttime Stops for Selected Datasets

Scatterplots of the monthly SMR for stops occurring during the day and night, plotted by year.
Note that the scaling of the y–axes are different for ease of interpretation.

Across all of the datasets (not shown here), the most common difference between the **day** and **night** monthly SMRs is simply no difference. The **day/night** plot for Nashville, Tennessee (not displayed here) looks like two side-by-side copies of the Nashville plot of Figure 7. Day/night SMR plots that are similar imply that recording practices between stops occurring during the night and day are similar and not a cause for concern. However, there are some datasets where the missingness across day and night vary in systematic ways that could impact model interpretability.

Level changes are a common trend and can be seen in Wichita, Kansas and Louisville, Kentucky (see Figure 8). Distinct differences in the monthly SMR from **day** to **night** can occur regardless of the underlying functional form of the missingness, whether it is a constant error (as in Louisville) or more complex (as in Wichita). The functional form doesn't change between day and night; rather, the day and night SMR are vertical shifts of one another. We also note that the level change is not always a change so that the SMR during the nighttime is higher, as one may intuitively expect.

We can also identify complicated functional relationships between missingness, **date** of the year, and **day/night**. The Iowa statewide and Saint Paul, Minnesota plots in Figure 8 illustrate two different functional forms of monthly SMR observed between day and night. In Iowa, the quantity of missing variables changes between day and night even within the same year. In Saint Paul, Minnesota, the variability and shape of the missingness between **day** and **night** are noticeably different.

The previous three plots are indicative of the benefit of disaggregating the missingness measures. The most telling example is that of Louisville, Kentucky: the plot in Figure 7 presents a fairly mild, potentially concave relationship between missingness and the year. Only after plotting the monthly SMR by day and night in Figure 8 can we notice the fairly stark difference between the recording of daytime and nighttime stops. Furthermore, the level changes in the Iowa plot in Figure 7 are complicated by how missingness is different between day and night, even during the same year. Lastly, the disaggregated missingness in Saint Paul, Minnesota in Figure 8 sheds light on the erratic trends found in Figure 7. Although we do not know why the shape and spread of monthly SMR is different between night and day, we see how aggregating those

conflicting trends results in what seems to be a sudden introduction of highly variable missingness after 2005.

These trends in differential missingness between day and night imply that in some cases, traffic stops are recorded differently. Whether or not this obscures trends in racial profiling or biases the results of models that use the data is unclear, but we find evidence that the available information to analyze can be different based on the time, week, and outcome of the stop.

The exploration presented thus so far provides evidence of some trends in missingness. Furthermore, until such trends are more fully explained and understood, the question of if and to what extent this missingness results in bias for the models relying on the data remains unanswered. We identify missingness as another potential confounding variable.

Logistic Regression

While we find descriptive evidence of some patterns of missingness from Figures 3-8, we have not yet considered if differential SMR can impact statistical models predicting post-stop outcomes like **search**. Trends in missingness can impact model results for a variety of reasons: missingness can altogether exclude a non-random group of observations from the model; missingness can also interact with the recording of other pertinent variables. In this section, we investigate how missingness can affect the significance, magnitude, and possibly the sign of the estimated coefficients from a logistic regression.

We choose datasets that have distinct patterns of missingness and run the same minimal regression on the observations from each missingness pattern. We select the Nashville and Louisville data because, as found in Figures 5 and 6, missingness behaves in only two patterns within each dataset. In the Nashville data, missingness remains constant at approximately 0% before 2014 and jumps to about 1% after 2014; In the Louisville data, missingness remains relatively constant for day and night stops, at approximately 2.5% and 7%, respectively.

The minimal logistic regression that we run is for **search** and uses the predictor variables **race**, **age**, **sex**, and **day** of the week. For the Nashville, Tennessee model, we also include the predictor variable **day/night**. For each dataset, the same logistic model is run three times: once for each subset of the data with a distinct pattern of missingness, only the day stops and a third time on the combined dataset. The purpose of the minimal model is not to find evidence of racial profiling but rather to explore if and the extent to which missingness is a confounding variable.

The model is run on a 30% random sample of the Nashville, Tennessee dataset representing over 1.2 million traffic stops. Due to the smaller amount of the data for Louisvillege, Kentucky, we use the entire dataset consisting of about 110,000 observations to run the logistic regressions. The baseline group in the model is a female driver identified as Asian/Pacific Islander. We present the estimated logarithm of the odds ratios (log-odds) for only the predictor variables relating to driver demographics.

Table 5: Logistic Regression of Searches in Nashville, Tennessee by Before and After 2014

	Before 2014	After 2014	Combined
Intercept	-3.816 (0.000)	-4.013 (0.000)	-3.892 (0.000)
Black	1.099 (0.000)	1.282 (0.000)	1.189 (0.000)
Hispanic	1.141 (0.000)	0.940 (0.000)	1.049 (0.000)
Missing Race (NA)	1.127 (0.001)	0.321 (0.274)	0.574 (0.008)
Other Race	-0.170 (0.217)	0.697 (0.003)	0.046 (0.693)
Unknown	-0.297 (0.014)	-0.384 (0.002)	-0.338 (0.000)
White	0.502 (0.000)	0.454 (0.000)	0.488 (0.000)
Age	-0.035 (0.000)	-0.032 (0.000)	-0.034 (0.000)
Male	0.992 (0.000)	0.805 (0.000)	0.904 (0.000)
Missing Sex (NA)	1.806 (0.000)	0.766 (0.000)	0.808 (0.000)

In comparing the before and after 2014 coefficients in Table 5, most of the coefficients are quite similar. However, the log-odds for drivers identified as "Other Race" does change: it is insignificant before 2014, significant after 2014, and insignificant in 2014. The other main difference among the three models is the magnitude of the log-odds for drivers whose sex is not recorded ("Missing Sex (NA)"). Statistically significant log-odds of 1.806, 0.766, and 0.808 translate into the probabilities 8.59%, 6.83%, and 6.92%. Compared to the baseline group, NA-sex drivers before 2014 were 8.59% more likely to be searched; the probability drops to 6.83% after 2014.

The question of whether these changes in search behavior are due to policing practices, driving behavior, or data collection and missingness is ambiguous. We thus see how missingness is another confounding variable that should be considered during model interpretation.

Table 6: Logistic Regression of Searches in Louisville, Kentucky by Day/Night

	Day	Night	Combined
Intercept	-4.530 (0.000)	-3.088 (0.000)	-3.634 (0.000)
Black	2.070 (0.000)	1.404 (0.000)	1.692 (0.000)
Hispanic	1.080 (0.039)	0.552 (0.153)	0.732 (0.013)
Missing Race (NA)	3.282 (0.015)	3.324 (0.023)	3.452 (0.000)
Other Race	2.495 (0.000)	1.067 (0.077)	1.754 (0.000)
Unknown	1.437 (0.004)	0.816 (0.025)	1.009 (0.000)
White	-0.037 (0.000)	-0.022 (0.000)	-0.036 (0.000)
Age	1.036 (0.000)	0.874 (0.000)	1.034 (0.000)
Male	3.282 (0.000)	-8.708 (0.950)	2.288 (0.007)
Missing Sex (NA)	-0.152 (0.155)	0.147 (0.137)	-0.194 (0.006)

We notice more differences in the estimated log-odds among the day, night, and combined models in Table 6. The coefficient for Hispanic is significant at 1.080 for daytime stops, insignificant for nighttimes stops, and is significant again in the combined model, albeit at a lower value at 0.732. A similar trend is seen for drivers identified as "Other Race" and "Unknown"; the coefficients are insignificant for the night model, but significant in the day and combined model. For these three levels of **race**, the magnitude of the log-odds is greatest for daytime stops and lower in the combined model. Lastly, the coefficient for drivers identified as male exhibits a similar pattern of being significant in the day and combined variables, with the log-odds becoming quite negative (but insignificant) for nighttime stops.

As with the models run for Nashville, Tennessee, the models for Louisville, Kentucky cannot be interpreted to make any conclusions regarding police bias towards race or gender; driving behavior of different races and genders; or the impact of missingness. Whether due to missingness or another confounding variable, the reason for the changes in significance and the magnitude of the log-odds cannot be isolated.

The three-part models run on traffic stop data from Nashville, Tennessee and Louisville, Kentucky neither prove nor disprove the impact of missingness on models that test for racial profiling. Missingness as a trend and as a confounding variable should thus be treated with caution and incorporated in model interpretation. These datasets and regressions demonstrate the need for further research into the more complex functional forms of missingness, especially those that do not result in distinct groups of observations that have the same missingness pattern. Furthermore, we can apply the SMR framework towards more sophisticated statistical tests to see if and how missingness affects results.

Discussion

This project explores patterns in missingness by defining the SMR statistic and visualizing it by different variables across several dozen datasets. By analyzing missingness from different angles, we see how traffic

stop data potentially contain another confounding variable: the recording practices of state patrol agencies and municipal departments.

We do not find evidence of large-scale recording practices changing by the race of the driver; however, we find one instance in Seattle, Washington of interaction between the stop outcome arrest and the recording of race. Arrest rates by race groups are thus biased due to the absence of non-arrested drivers' race information. With regards to time of the year and time of the day, we find several different trends reflecting changes in department collection practices. Weekly SMR plots demonstrate trends such as constant SMR with distinct jumps; different degrees of variability; and complex functional forms like concavity. The day/night plots usually relay no difference between the missingness in stops recorded during the day and night, but a few datasets have level changes and functional forms that are different. Until we have a stronger understanding of what these missingness trends entail and why they occur, the impact of differential SMR on models that use traffic stop data will not be known.

To begin understanding the impact of differential SMR, we run logistic regressions on the two datasets Nashville, Tennessee and Louisville, Kentucky. Both of the datasets distinct patterns of missingness. Separating those patterns of missingness and running the same logistic regression on them, we find slight differences in the significance and sign of coefficients for certain values of driver **race** (unknown race, other race) and **sex** (male, NA sex). The log-odds of **race** and **sex** variables (Hispanic, "Other Race," "Unknown Race", and male) from the regressions on Louisville, Kentucky also change. We notice that the magnitude of the log-odds for the day model are greater than those in the combined model. The reason for the slight differences is, again, ambiguous – it could result from the variable by which we divide the dataset, missingness, some other confounding variable, or a combination of these.

Future avenues for research on traffic stop missingness are plenty. Due to the frequent recording of the variables **latitude**, **longitude**, and **location**, spatial analyses of missingness would involve the SMR by location for several dozen datasets. Also, more exploration can be done regarding the interaction between the recording of race and other variables, like the string variable **violation**. Lastly, further research into the common temporal trends of missingness can be pursued and would greatly inform how missingness impacts model interpretation and results.

References

Anderson, Monica. 2020. "For Black Americans, Experiences of Racial Discrimination Vary by Education Level, Gender." *Pew Research Center.* Pew Research Center. https://www.pewresearch.org/fact-tank/2019/05/02/for-black-americans-experiences-of-racial-discrimination-vary-by-education-level-gender/.

Baumgartner, Frank R, Leah Christiani, Derek A Epp, Kevin Roach, and Kelsey Shoub. 2017. "Racial Disparities in Traffic Stop Outcomes." *Duke FL & Soc. Change* 9: 21.

Chanin, Joshua, and Megan Welsh. 2020. "Examining the Validity of Traffic Stop Data: A Mixed-Methods Analysis of Police Officer Compliance." *Police Quarterly*, 1098611120933644.

Davis, Elizabeth, Anthony Whyde, and Lynn Langton. 2018. "Contacts Between Police and the Public, 2015." *US Department of Justice Office of Justice Programs Bureau of Justice Statistics Special Report*, 1–33.

Grogger, Jeffrey, and Greg Ridgeway. 2006. "Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness." *Journal of the American Statistical Association* 101 (475): 878–87.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, et al. 2020. "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States." *Nature Human Behaviour*, 1–10.

Ridgeway, Greg. 2006. "Assessing the Effect of Race Bias in Post-Traffic Stop Outcomes Using Propensity Scores." *Journal of Quantitative Criminology* 22 (1): 1–29.

Ridgeway, Greg, and John MacDonald. 2010. "Methods for Assessing Racially Biased Policing." *Race, Ethnicity, and Policing: New and Essential Readings*, 180–204.

Russell, Katheryn K. 2001. "Racial Profiling: A Status Report of the Legal, Legislative, and Empirical Literature." *Rutgers Race & L. Rev.* 3: 61.

Smith, Michael R, and Matthew Petrocelli. 2001. "Racial Profiling? A Multivariate Analysis of Police Traffic Stop Data." *Police Quarterly* 4 (1): 4–27.