Report on Racial Profiling by Metro Nashville Police Department

Under the Guidance of
Dr. Niamh Cahill
Assistant Professor
Department of Mathematics and Statistics
National University of Ireland, Maynooth

Submitted By
Raghav Chadha
MSc. Data Science and Analytics
19250266
Batch 2019-20

Submitted in Partial Fulfilment of the Requirement for the Degree of Master of Science

CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 4

CHAPTER 5

CHAPTER 6

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 4

APPENDIX

STATEMENT OF ORIGINALITY

I have read and understood the Department policy on plagiarism and I certify that work demonstrated in this thesis titled is my own and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education.

I confirm that:

1. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.
2. Sources of all figures or tables has been provided in the document which are not my work.

Raghav

Raghav Chadha
19250266
MSc. Data Science & Analytics
National University of Ireland, Maynooth

ACKNOWLEDGEMENT

I feel very grateful to have the opportunity to work under **Dr. Niamh Cahill** for my thesis in MSc. Data Science and Analytics, whose expertise and proficiency helped me to carry out the analysis using Bayesian Models in particular. I was not very confident initially as the topic and background was unfamiliar for me and I would like to thank her for continuous guidance and support. She helped me throughout the project, be it making me understand the topic or clearing my last-minute doubts. She has not only been a great mentor but a motivator. I really enjoyed working with her and hope to work with her in the coming future!

I would also want to thank **Dr. Catherine Hurley** for providing me the knowledge of the R programming language and Machine learning.

At last I would like to express my gratitude towards National University of Ireland, Maynooth for providing me all the necessary facilities which facilitated this research work.

Ragh

Raghav Chadha

ABSTRACT

Metro Nashville Police Department’s (MNPD) over-policing of predominantly black drivers raise serious concerns about the effectiveness, legitimacy, and constitutionality of MNPD’s traffic stop and search regime. With this report, I will show that Black drivers are subjected to racial biasing by analysing the stops made by the MNPD with a series of finding using the Stanford Open Policing Project. However, my findings show that traffic enforcement targets and impacts entire communities, not just people who commit crimes, and regardless of the area, black people are searched at much higher rates than white people, that is, once stopped minority drivers were searched about 2 to 3 times more than white drivers, while they have a lower hit rate, means they were less likely than their white peers to possess drugs, firearms, or other illegal contrabands.

Keywords: Metro Nashville Police Department (MNPD), racial biasing, hit rate

CHAPTER 1

INTRODUCTION

The United States of America has been largely reserved for white people through the intentional exclusion and oppression of people of colour, but only now increasing numbers of people becoming aware of it. The term “systemic racism” (Feagin & Feagin, 1978) refers to the institutional, broad and inevitable hierarchical system in the US which is formed and supported by whites and directed at people of colour. Systemic racism is a ‘material, social, and ideological reality that is well embedded in major US institutions’ (Feagin & lias, 2013). In a recent study conducted by the Massachusetts Institute of Technology, the researchers applied for job positions in Chicago and Boston using typical white and black names. Brenden, Emily, Gregg and Anne received 50% more responses to their applications by the employers than Tamika and Tyrone with African-America family names associated with them like Jackson, Jones and Robinson (Bertrand & Mullainathan, 2004). There is a strong opinion of American people that law enforcement agencies practice racial bias, with most Americans believing that the judicial system is biased by favouring “Whites over Blacks” (Sparks, 2020).

The research done in the past few years on racial profiling by the police reveals there were various extra-legal factors that affect the action taken by the police after an encounter with a citizen (Lundman & Kaufman, 2003), for example even in late 1940’s police looked at the race, nationality and social class of an offender when deciding to respond formally or not (Goldman, 1963, pp. 21-22). There are a number of incidents that have come to light in which the rights of African Americans were abused by the law enforcement agencies. Approximately half of the blacks under the age of 65 say that there was at least once that they have felt to be in danger because of their race, suggesting that many do understand the social and economic benefit of being White, sometimes called as “White Privilege” (DiJulio, et al., 2015) or even called as “Leukology”, referring to the study which is focused on the privilege and power of White people and their oppression on Blacks (Feagin, 2013).

In a recent event of deadly use of force by the now-former Minneapolis police officer who pressed his knee into the neck of George Floyd, an unarmed Black man has reinvigorated a very public debate about police brutality and racism. By understanding the history of racism in the justice system, American citizens can comprehend that death like of George Floyd's are symptomatic of a larger failure of American justice showing the long existing systematic racism which is deep rooted in the country. This sentiment of "White Privilege" is embodied in activist movements such as the "Black Lives Matter" campaign and in phrases such as "driving while black" (Harris, 1999).

Given the history of racism by law enforcement agencies, black drivers have a higher probability of being stopped in the United States as compared to whites. Similarly, the probability of being searched is higher for blacks as compared to other races and as Anwar & Hanming (2006) have pointed out that the data on search conducted by the police gives evidence that a higher proportion of searches are being conducted on minority motorists resulting in "Racial Profiling", a term that refers to the treatment conducted by the police which uses race or ethnicity as one of the criterias to search a vehicle.

From all of these findings it is very much clear that the minority population in America is facing racial bias by the police. Therefore, through the relevant facts and statistics, I examined how public officials indulge in "Preference-based Discrimination" while stopping vehicles, resulting in biasing of race, gender, and other demographic characteristics when performing their duties in Nashville, Tennessee under the Metropolitan Nashville Police Department (MNPD) stop and search regime. The study focuses on exploring the probability of search conducted by the MNPD that resulted in contraband being found. This report is an effort towards amplifying and deepening the claim black Nashvillians (constitute only 28% of the total population as seen from Table 1.1) have been making for decades: *MNPD engages in racial profiling every single day!*

Table 1.1: Racial Composition of Nashville

Census Demographics	Percent
White	62.35%
Black	28.04%
Asian/Pacific Islander	3.52%
Hispanic	10.4%
Other	3.38%

CHAPTER 2

DATA

Police Departments tend to stop traffic for several reasons, one of them being to prevent potential loss of life by focusing on crime prevention and detection. Traffic stops are necessary considering everyone's safety on the road as well as pedestrians and in the detection of contrabands such as drugs or weapons. Keeping this in mind, the Stanford Open Policing Project with the Data from 21 state patrol agencies and 29 municipal police departments facilitate rigorous statistical analysis to help researchers, journalists, and policymakers investigate and improve interactions between police and the public.

My focus on the whole is on Nashville, Tennessee and the data is provided by Metropolitan Nashville Police Department (MNPD). This dataset includes detailed information on Date, Time of the stop extended by information about the officer, the race of the driver along with the outcome of the stop like citation, arrest or warning and further if a search for contraband was performed on the driver or not. I chose this dataset as I found it interesting as it has a large number of variables, 42, both categorical and continuous to explore and perform analysis on. The dataset is also large with 3,092,351 observations with the date range of Jan 2010 - Mar 2019. All these things contribute to the dataset making it useful to find concrete, data-driven insights to improve both the equity and efficacy of the MNPD's policing strategies.

After breaking down some basic summary statistics by race/ethnicity as shown in Table 2.1, we find blacks are searched more often than white drivers. If minorities also happen to carry contraband at higher rates, these higher search rates may stem from appropriate police work. Disentangling discrimination from effective policing is challenging and requires more subtle statistical analysis, as done below.

Table 2.1: Summary Statistics

subject_race	n_stops	n_searches	n_hits	search_rate	hit_rate
Asian/ Pacific Islander	41668	833	144	0.0200	0.173
Black	1165871	67985	14752	0.0583	0.217
Hispanic	164814	10165	1231	0.0617	0.121
Other	10397	243	50	0.0234	0.206
Unknown	36878	590	97	0.0160	0.164
White	1670873	47826	9971	0.0286	0.208

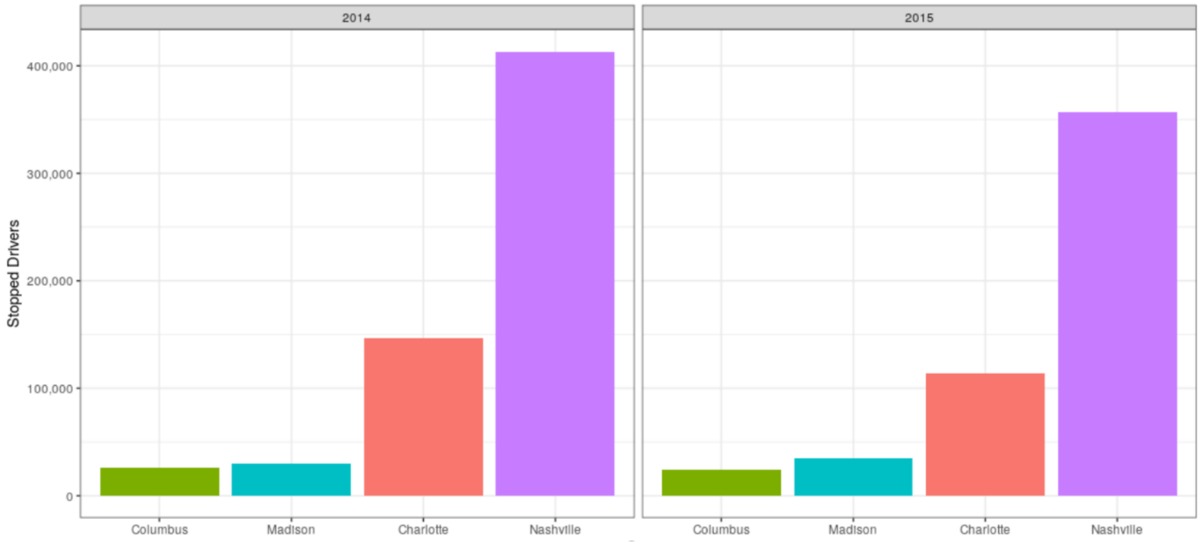
2.1 Preliminary Findings

The data on search conducted by the police across different states in America continues to show that they search minority population at a higher rate than white motorists. However we cannot reach to the conclusion of racial profiling based on average stop rates and searches which is referred "Statistical Discrimination" (Knowles, et al., 2001). The findings of my report analyse traffic stops of drivers and through these findings I will demonstrate in detail about the policing practices used by the MNPD showing racial bias and how blacks are subjected to more traffic stops, arrests and their vehicles being searched more often compared to whites. These finding cannot completely rule-out "Statistical Discrimination" but they provide support to my claim that police officers indulge in "Preference-based Discrimination" (Becker, 2010). My research on the stops by MNDP resulted in the findings, which are elaborated below:

Finding #1: In 2014 and 2015, MNPD conducted nearly 0.8 million traffic stops, a number 3 to 15 times greater than comparable cities.

Nashville Metro Police Department (MNPD) for many years employed a strategy of making large numbers of traffic stops as a way to address crime and violence. When compared Nashville's Traffic Stops to its neighboring cities of Charlotte, Columbus and Madison for both the years, we can clearly see that Nashville has the highest stops as seen from Figure 2.1. Another important thing to consider here is, Charlotte has a higher population when compared to Nashville, still more traffic stops are being conducted in Nashville. This gives us the required evidence of over-policing under the MNPD's stop and search regime. Also, a report by Policing Project (2018) suggests that increasing traffic stops has not successfully reduced crime in Nashville.

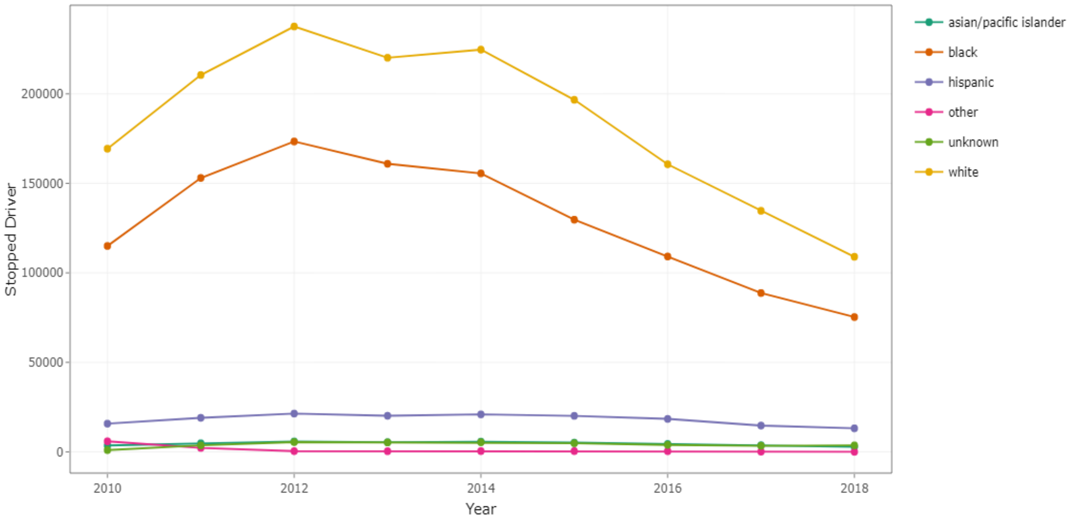
Figure 2.1: Traffic stops in neighboring cities of Nashville for years 2014 and 2015



Finding #2: Across all years, black drivers are arrested at rates disproportionately higher than other drivers.

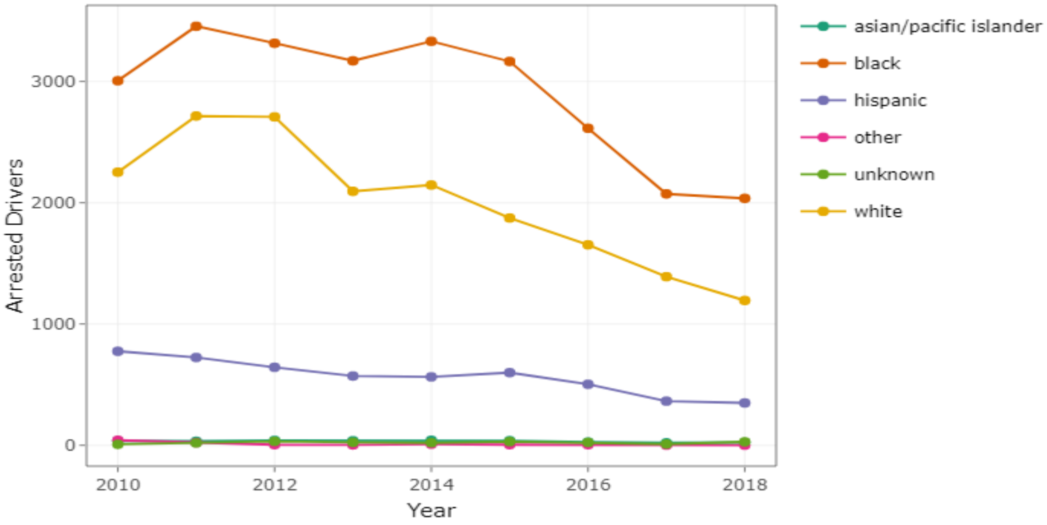
Moving our focus back to Nashville, to have a more concrete evidence of racial disparities I analysed the traffic stops by race/ethnicity which will help in gaining a better perspective. Figure 2 shows the number of stops conducted from 2010-2018 for all the races and we can see that given the population of whites in Nashville is the highest, hence they are being stopped the most. In fact, White drivers in Nashville are stopped 1.43 times higher than Black drivers.

Figure 2.2: Drivers Stopped by Race/Ethnicity (2010-2018)



But when we compare the outcome of these stops, that is, how many of these stops resulted in an arrest and what race did the driver belong to? The findings of Figure 2.3 shows us that across all the years blacks have been arrested the most even though whites were stopped the most. Although blacks make up 28% of the population but the arrested rate is 1.45 times higher than that of whites and this gap between the number of arrests is increasing each year.

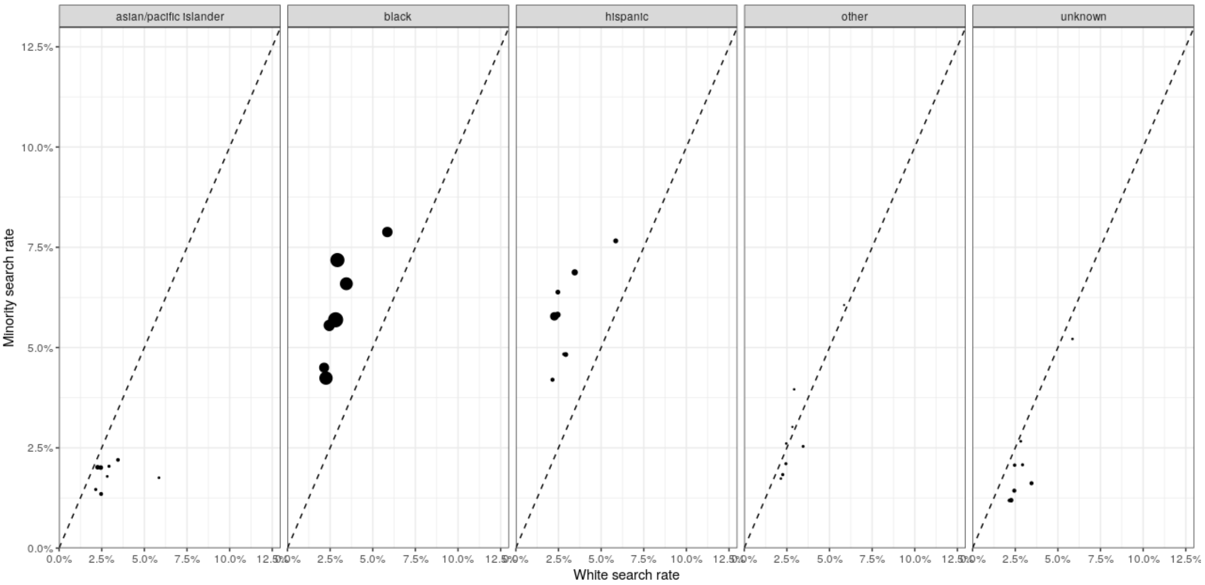
Figure 2.3: Drivers Arrested by Race/Ethnicity (2010 - 2018)



Finding #3: MNPD officers conduct probable cause and consent searches of Black and Hispanic drivers at higher rates than White drivers.

After stopping a vehicle the police officer can search it or the driver on suspision of a serious criminal offence. From Figure 2.4, we see that points lie above the diagnol line for black and hispanic drivers indicating that the search rates are higher for minorities within the same precinct. From the number of stopped drivers the total number of probable cause searches of black and hispanic drivers exceeds the number of probable cause searches of white drivers.

Figure 2.4: Comparing search rates for minority and white drivers within the same precinct



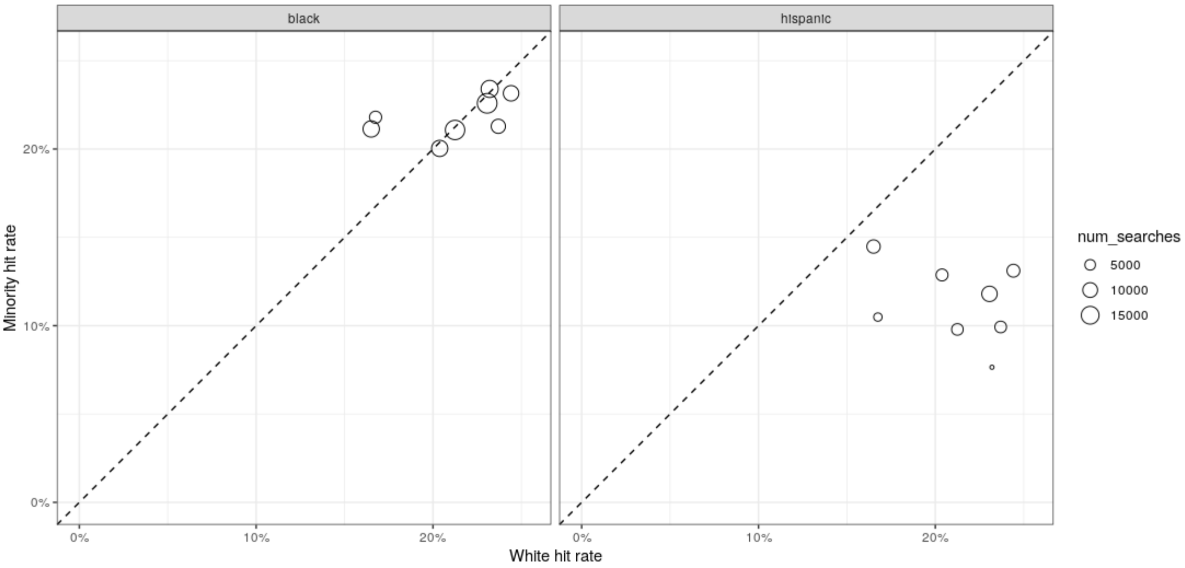
The results of the above plot were also evident from Table 2.1, in which the search rates in Nashville for blacks were 5.83% (95% CI 5.78% - 5.87%), 6.17% for hispanics (95% CI 6.05% - 6.28%) and for white drivers were 2.86% (95% CI 2.83% - 2.88%). It is evident that black and hispanic drivers are much more likely to be searched, almost 2 - 3 times higher than the search rate for whites.

The Fourth Amendment to the U.S. Constitution generally prohibits law enforcement officers from conducting routine searches during traffic stops. However, in traffic stops the normal requirement of obtaining a warrant before conducting a search is waived. Hence, a successful search will be the one in which a contraband is found, which is known as "hit rate". Keeping this in mind one would expect MNPD to make use of this legal framework and engage in racial bias when deciding to search a vehicle or not.

As Figure 2.5 demonstrates, the hit rates of both the minority groups, blacks and hispanics, is lower than that of whites and it seems that with the help of our analysis that nearly every precinct has a lower hit rate of minorities than of whites. We can confidently say that racial disparities in the frequency of probable cause searches are both significant and growing by each year even when the the probability of a successful search of contraband being found in vehicles driven by minorities is less.

Police targets the minority thinking that only in their vehicles illegal material would be found and thus, strengthening the argument of the report that systematic racial profiling exists in the state. In section 4 and 5 of this report various methods and its findings will be used to predict the trend in the coming years of contraband being found or not based on the subject race and year are presented.

Figure 2.5: Hit rates for Minorities in each Precinct



CHAPTER 3

METHODS

3.1 LOGISTIC REGRESSION

In the year 1972 Logistic Regression was adapted on the exponential family of distributions as an alternate option to overcome the limitations with ordinary least squares method. The method was then named as Generalized Linear Models (GLM), proposed by Nelder and Wedderbrun and was available in various statistical packages by 1980's (Hilbe, 2011), of which Logistic Regression is a part. It has become the most used algorithm to calculate odds ratio with the use of one or more predictors.

The algorithm is similar to multiple linear regression except for the fact that the response in binomial, that is, 0/1 with 0 indicating failure and 1 indicating success. In the case of a Linear regression model, it assumes normal or Gaussian distribution for the response variable and the error term, the variance, that is, σ^2 to be constant, and observations are independent and identically distributed (iid). The first two assumptions mentioned above are violated in the case of Logistic Regression. Analogical to Linear regression model which is based on Gaussian probability distribution function (pdf), a Logistic Regression model is derived from a Bernoulli distribution, which is a subset of the binomial pdf with the binomial denominator taking the value of 1 (Peng & So, 2002).

The Logistic Regression Model has the following components associated with it:

1.1.1 Random component:

$Y \sim$ Bernoulli (π).

E[Y] = π

1.1.2 Systematic component:

$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

1.1.3 Link function: logit link

$g(\pi) = log(\frac{\pi}{1-\pi}) = \eta$

Now, since we are interested in the parameter π:

$$log(\frac{\pi}{1-\pi}) = \eta$$

$$\frac{\pi}{1-\pi} = e^{\eta}$$

$$\pi = \frac{e^{\eta}}{1+e^{\eta}}$$

This is called the **logistic function.**

Notice that $0 < \frac{e^{\eta}}{1+e^{\eta}} < 1$

This gives us the **logistic regression function**:

$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$

The Logistic Regression Model used here is:

$logit(\pi) \sim \beta_0 + \beta_1 * subject\_raceblack + \beta_2 * subject\_racehispanic + \beta_3 * subject\_raceother + \beta_4 * subject\_racewhite + \beta_5 * year$

Where π is the probability of contraband_found which is based on the number of vehicle searches.

β_0 is the intercept and acts as the baseline for the model which is $subject\_raceasian$

3.2 Generalized Additive Model (GAM)

Generalized Additive Model (Hastie & Tibshirani, 1990) is a natural addition to the binary regression method by the inclusion of covariate terms in the regression approach. We can think of Generalized Additive Model (GAM) as a Genrelaized Linear Model (GLM) with the added functionality to include arbitrary smooth functions in addition to linear terms in the linear predictor (McCullagh & Nelder, 1989). GAMs are data-driven rather than model-driven meaning that the results obtained after fitting a model are not derived from a priori model. The reason behind this approach of fitting a nonparametric model is due to the fact that first, the structure of the data should be examined before fitting a priori determined model.

With the use of GAMs, a constraint in the relationship of x_k and η to be linear as in GLMs, is no longer required and the relationship is merely constrained to be smooth (Hastie & Tibshirani, 1986). This helps to identify patterns that the linear model would not capture and it also makes use of the smoothing function that prevents over-fitting.

The linear regression model is:

g(E(Y)) = β_0 + $\beta_1 x_1$ + ... + $\beta_k x_k$ + ε

A generalised additive model (GAM) extends this by allowing a function f_k for every predictor x_k.

The GAM is then:

g(E(Y)) = β_0 + $f_1 x_1 + \ f_2 x_2 + \cdots + \ f_k x_k$ + ε

Where Y is the dependent variable (i.e., what we are trying to predict), E(Y) denotes the expected value, and g(Y) denotes the *link function* that links the expected value to the predictor variables $x_1, \dots, x_k$.

We fit the Generalized Additive Model using the same predictors of contraband_found we used with the Logistic Regression Model because we have a time component in the data and we want to use a non-linear function (e.g., a spline) to capture the relationship. We will use splines for predictor year. Subject_race is a factor and takes just a few values, so a spline is not appropriate.

The model is:

$$g\big(E(Y)\big) \sim \ \beta_0 + \ \beta_1 * subject\_raceblack + \beta_2 * subject\_racehispanic \ + \ \beta_3 * subject\_raceother + \ \beta_4 * subject\_racewhite \ + f_5 * year + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0,\sigma^2)$$

Where $g\big(E(Y)\big)$ is the probability of contraband_found which is based on the number of vehicles searched.

β_0 is the intercept and acts as the baseline for the model which is $subject\_raceasian$.

3.3 BAYESIAN HIERARCHAL MODEL

In a Bayesian approach the parameters (eg. θ) are observed to be random and the information that is related to the random variables is summarized in a prior probability distribution. The information is then updated through the data to obtain the posterior distribution for the parameters which is based on Bayes' rule:

$$p(\theta|\mathrm{y}) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Where,

3.3.1 $\boldsymbol{p(\theta|\mathrm{y})}$ is known as the likelihood which represents the information in the data.

3.3.2 $\boldsymbol{p(\theta)}$ is known as the prior distribution which represents information about θ before observing any data.

3.3.3. $\boldsymbol{p(y|\theta)}$ is known as the posterior which represents the updated information about θ after observing the data.

3.3.4 $\boldsymbol{p(y)}$ is known as the marginal likelihood.

Note if $\boldsymbol{y}$ fixed, marginal likelihood is constant with respect to θ, so we can write the equation above in terms of the posterior distribution: $p(\theta|\mathrm{y}) \propto p(y|\theta)p(\theta)$ (Cahill, 2020).

Bayesian models are structured to integrate the prior information (information external to the data) rationally into the statistical analysis process. While doing the same, they provide answers to a number of vexing problems that are encountered. The probability distributions are different for both Bayesian and frequentist methods. The frequentist approach is only dependant on the given data, that is, the likelihood for both observed and unobserved data while the Bayesian approach is dependant the prior belief as well as the given data that is the likelihood of the observed data (Orloff & Bloom, 2014).

Hierarchal Models involve multiple parameters such that the resulting values of some of these parameters are meaningfully dependant on the values obtained by the other parameters. In the case of the hierarchical Bayesian model, the prior distribution of some of the model parameters depends on other parameters, which are also assigned a prior (Kruschke, 2014)

Hierarchical Bayesian model is of the form:

$$\underbrace{p(\alpha,\theta|y)}_{\text{posterior}} \propto \underbrace{p(y|\theta,\alpha)\, p(\theta|\alpha)}_{\text{likelihood}}\ \underbrace{p(\alpha)}_{\text{prior}}$$

Hierarchical Bayes model is a combination of two things: i) a model written in hierarchical form that is ii) estimated using Bayesian methods. A hierarchical model is one that is written modularly, or in terms of sub-models.

We use the JAGS package (Just Another Gibbs Sampler; Plummer, 2003) to fit the model via Gibbs sampling. After the model has been specified, the next step is to setup the data file, number of chains, burn-in period, number of iterations after the burn-in and save the parameters we want the results of. The logistic regression model assumes that a sequence of independent and identically distributed random variables $\{Y\}_1^n$ has a Binomial distribution, denoted by Y_i ~Binomial(p_i), in the form of:

$$\mathrm{f}(Y_i|p_i\mathrm{i}) = \binom{n}{Y_i} p_i^{Y_i}(1-p_i)^{n-Y_i}$$

Where, $Y_i \in \{0,1\}, \log(\frac{p_i}{1-p_i}) = X_i\beta, X_i = (1, x_{i1}, \dots)$ is the line vector of covariates associated to the individual i and β is the vector of unknown parameters. The function $\log(\frac{p_i}{1-p_i})$ is called logit and provides the link between the random variable and its deterministic components (covariates) (Wundervald, 2018).

Let the observations be indexed by i = 1, . . . , n_obs (where n_obs = 40) and let the heirarcihal priors with race specific paramters be indexed by j = 1, . . . , n_race (where n_race = 5), consider the model:

$$mu\_i \sim alpha\_j[i] + beta\_j[i] * year$$

Here $alpha\_j[i]$ and $beta\_j[i]$ represent the intercept and slope for j^{th} race respectively which are associated with observation i.

The priors in the model are:

$$alpha\_j \sim Normal(mu\_alpha, sigma\_alpha^{-2})$$

$$beta\_j \sim Normal(mu\_beta, sigma\_beta^{-2})$$

CHAPTER 4

RESULTS

4.1 Logistic Regression

We begin the analysis by fitting a Logistic Model, a two predictor logistic model was fit to the data to test the research hypothesis regarding the relationship between the response, contraband_found that is dependent on total_searches, with the predictors, subject_race and year. The results are showed in Table 4.1 followed by interpretation of the model coefficients.

Table 4.1: Logistic Regression Model Summary Table (N = 40)

Predictors	Estimate	Std.Error	Z Value	Pr(>|Z|)	Odds Ratio
Intercept	-160.59389	6.86618	-23.389	< 2e-16	1.798697e-70
year	0.07900	0.00341	23.167	< 2e-16	1.082203
subject_raceblack	0.22560	0.09479	2.380	0.0173	1.253072
subject_racehispanic	-0.49838	0.09975	-4.996	5.84e-07	0.6075146
subject_raceother	0.32793	0.18623	1.761	0.0783	1.388097
subject_racewhite	0.19262	0.09501	2.027	0.0426	1.212422

*Reference category is subject_raceasian

The coefficients are in log-odds terms and the predicted logit of (contraband_found) = -160.59 + (0.079)*year + (0.22560)*subject_raceblack – (-0.49838)*subject_racehispanic + (0.32793)*subject_raceother + (0.19262)*subject_racewhite

The interpretation of the model coefficients could be as follows:

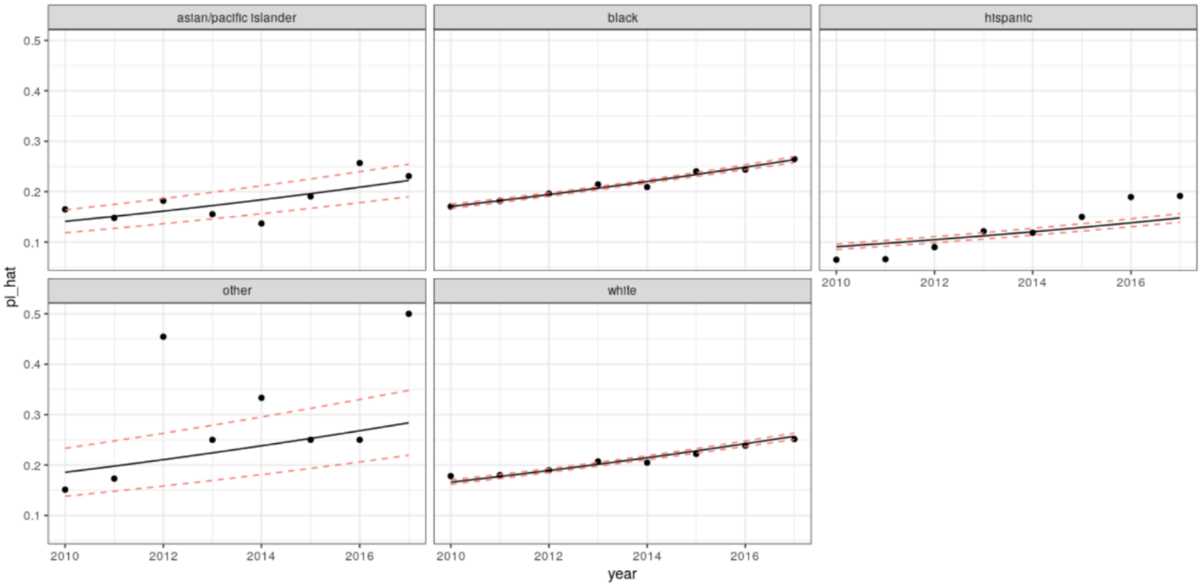
4.1.1 The intercept = -160.59, which is interpreted as log odds of a driver with subject_race as asian having contraband_found when searched.

4.1.2 The coefficient for year indicates that one unit increase in the year will increase the odds of contraband_found by exp (0.07900) 1.08 times i.e. 8%.

4.1.3 The coefficient for subject_raceblack indicates increase the odds of contraband_found by exp (0.22560) = 25% increase in subject_raceblack compared to subject_raceasian.

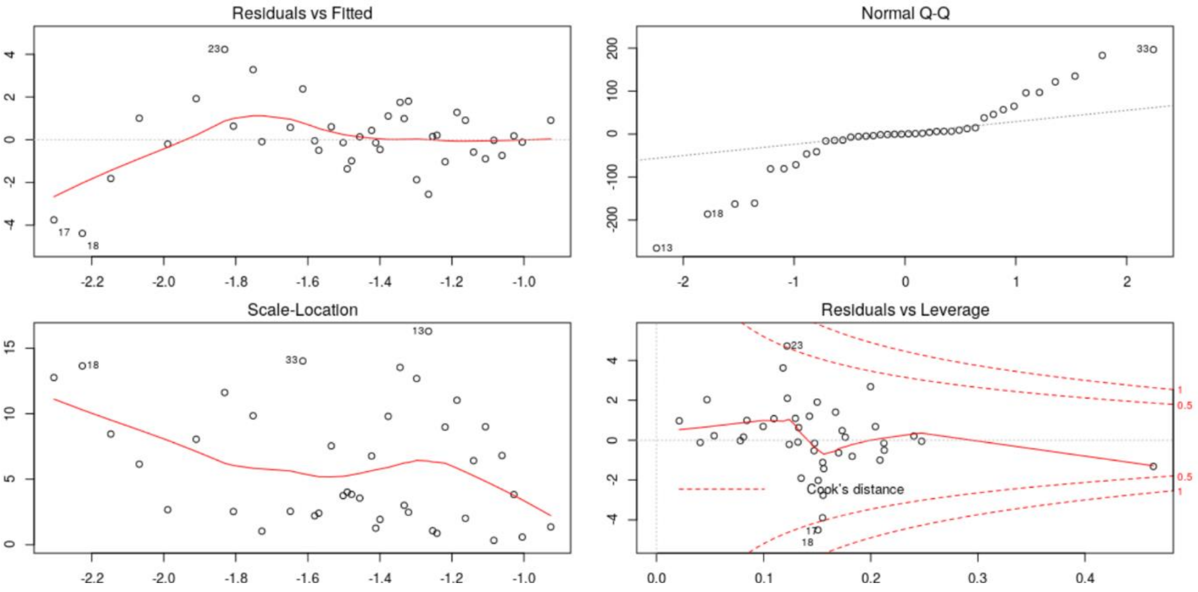
After fitting a binary logistic regression model, the next step is to check how well the model performs on the actual data that we have by overlaying predictions which can be seen in Figure 4.1. The points represent the actual data of the probability of contraband_found from the year 2010 to 2017 of different subject_race, the line represents the model predictions and the dashed red lines represent the 95% Confidence Interval. This 0.95 confidence interval is the probability that contraband_found will lie within the confidence interval of the regression model fit to our data. We can see there is variability within our observations as the model does not fit quite well.

Figure 4.1: Plot of contraband_found with Overlaid Predictions from Logistic Regression for each Race & Year



From Figure 4.2, we can interpret for Residual vs Fitted plot that for any fitted value residuals are not centred on the horizontal line with curvature, therefore we can assume that linearity assumption might be violated. Also, we see there is non-constant variance. From the Normal Q-Q plot we can see that there are discrepancies or points far away from the line, the normality assumption might be violated due to presence of outliers

Figure 4.2: Diagnostic Plots of Logistic Regression Model



To test the accuracy of the Logistic Regression Model, we have removed the year 2018 from the data and predict the values of the same year. If the predicted values lie within the 95% Confidence interval, we can say the model fits well. The results of the prediction are shown in Table 4.2.

Table 4.2: Testing Logistic Regression Model Accuracy through Pi_hat and Confidence Interval

Subject_race	Year	Pi_hat	95% Confidence Interval	
Asian/Pacific Islander	2018	0.1489362	0.2022513	0.2698104
Black	2018	0.2826443	0.2718600	0.2863254
Hispanic	2018	0.2263868	0.1485605	0.1675037
Other	2018	0.5	0.2334590	0.3668212
White	2018	0.2838309	0.2647913	0.2802209

*Pi_hat refers to the probability of contraband_found from raw data

4.2 Generalized Additive Models (GAMS)

The results from GAMS are shown in Table 4.3. As with the Logistic Regression, the intercept represents the contraband_found for subject_raceasian when searched and the intercept is -1.53006. In this case, the parametric components, the model intercept and subject_race both are significant.

Table 4.3: Generalized Additive Model Summary Table for Parametric Coefficients

Predictors	Estimate	Std. Error	Z Value	Pr(>|Z|)
Intercept	-1.53006	0.09429	-16.228	<2e-16
subject_raceblack	0.22560	0.09479	2.380	0.0173
subject_racehispanic	-0.49838	0.09975	-4.997	5.84e-07
subject_raceother	0.32793	0.18623	1.761	0.0783
subject_racewhite	0.19262	0.09501	2.027	0.0426

*Reference category is subject_raceasian

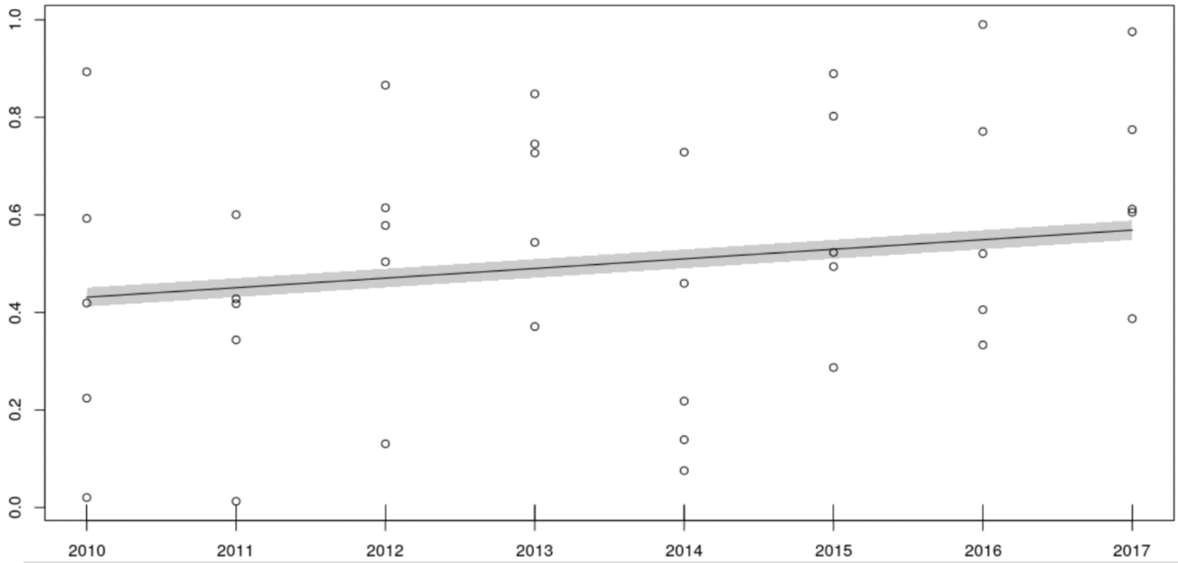
The next section of the GAMS summary shows the smooth term for which the coefficients are not printed and instead Effective Degrees of Freedom (EDF) are presented in the second coloumn of Table 4.4 which shows the output the for smooth term. In this case an EDF value of 1, represnts a straight line (Wood, 2006). In this model, the year term is linear but significant which can also be seen from Figure 4.3.

Table 4.4: Generalized Additive Model Summary Table for Smooth Coefficient

Predictors	EDF	Ref.df	Chi.Sq	P-Value
s(year)	1	1.001	536.7	<2e-16

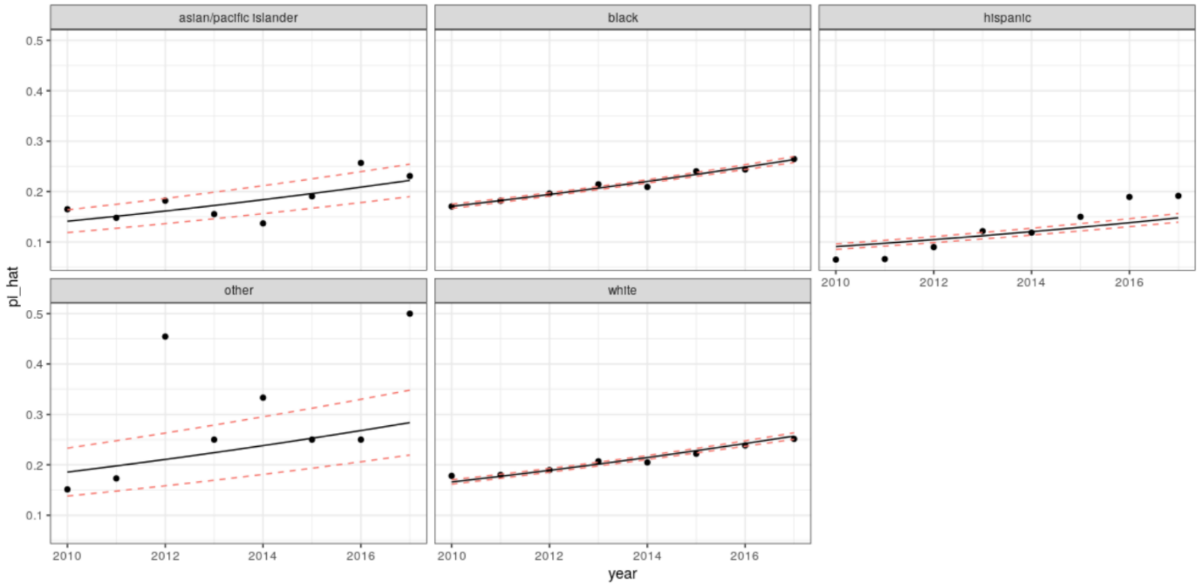
*EDF: Effective Degrees of Freedom

Figure 4.3: Partial contributions of Year with confidence bands



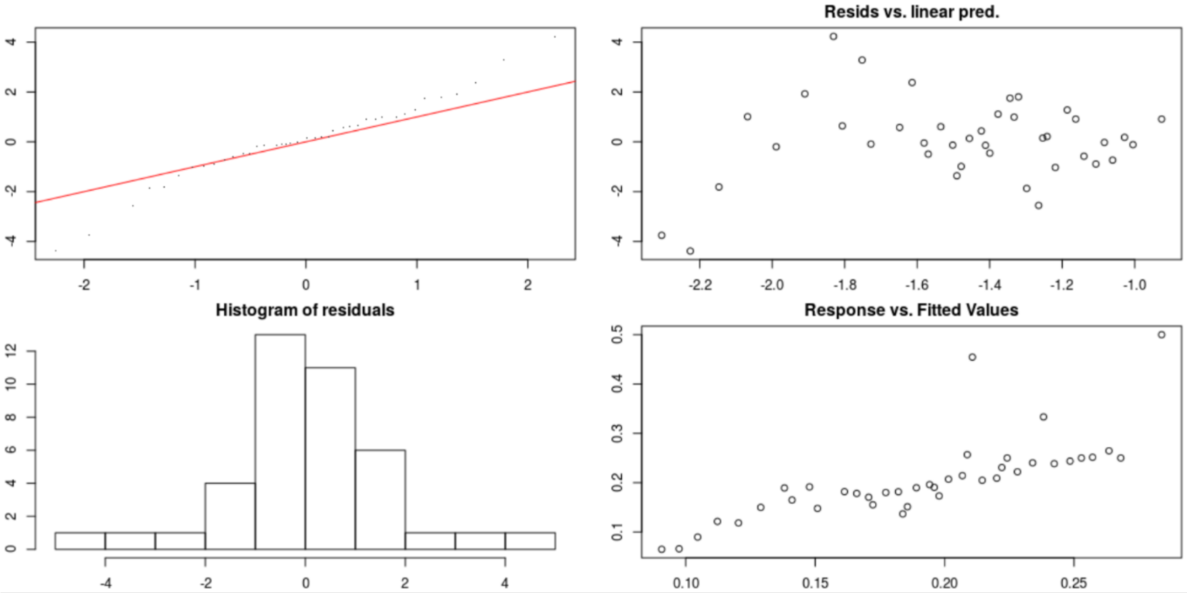
After fitting the Generalized Additive Model, the next step is to check how well the model performs on the actual data that we have by overlaying predictions which can be seen in Figure 4.4. It shows similar results as seen in Figure 4.1.

Figure 4.4: Plot of contraband_found with Overlaid Predictions from Generalized Additive Model for each Race & Year



From Figure 4.5, we can interpret by looking at the Q-Q plot, on the top left that there are discrepancies or points far away from the line, the normality assumption might be violated due to presence of outliers. On the bottom left, there is a Histogram of Residuals which is not bell shaped. On top right, a plot of residual values which are not evenly distributed around zero. Finally, on the bottom right, a plot of response against fitted values which does not form a straight line.

Figure 4.5: Diagnostic Plots of Generalized Additive Model



To test the accuracy of the Generalized Additive Model, I have removed the year 2018 from the data and predict the values of the same year. If the predicted values lie within the 95% Confidence interval, we can say the model fits well. The results of the prediction are shown in Table 4.5.

Table 4.5: Testing Generalized Additive Model Accuracy through Pi_hat and Confidence Interval

Subject_race	Year	Pi_hat	95% Confidence Interval	
Asian/Pacific Islander	2018	0.1489362	0.2022513	0.2698104
Black	2018	0.286443	0.2718600	0.2863254
Hispanic	2018	0.2263868	0.1485605	0.1675037
Other	2018	0.5	0.2334590	0.3668212
White	2018	0.2838309	0.2647913	0.2802209

*Pi_hat refers to the probability of contraband_found from raw data

4.3 Bayesian Hierarchal Model

After fitting the Bayesian Model, the next step is to check how well the model performs on the actual data that we have by overlaying predictions which can be seen in Figure 4.6.

Figure 4.6: Plot of contraband_found with Overlaid Predictions from Bayes Model for each Race & Year

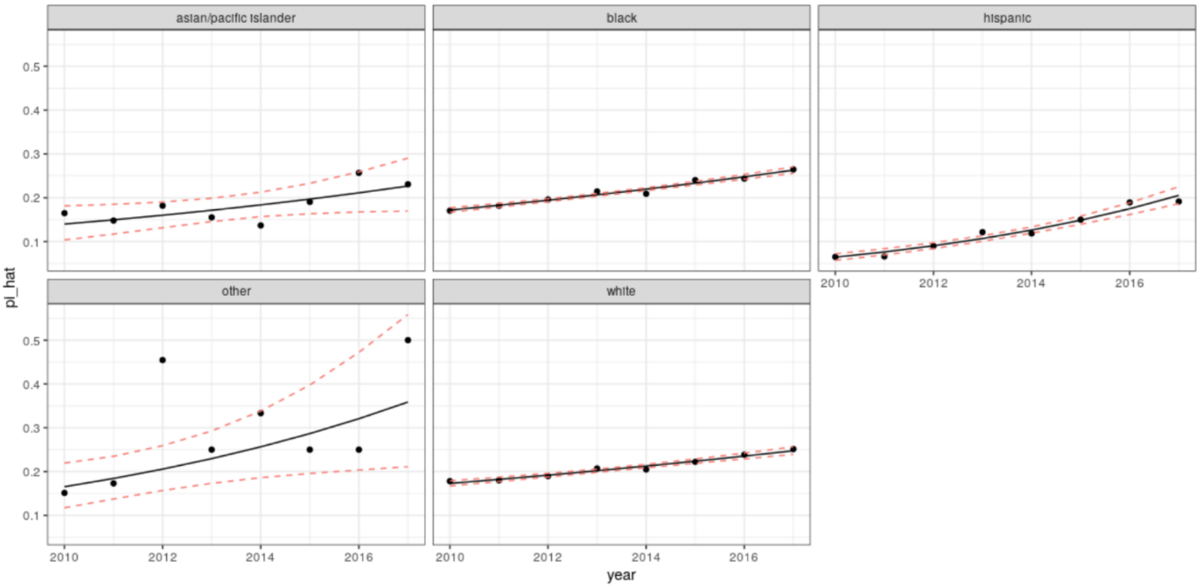


Figure 4.7: Trace Plots for α[1]

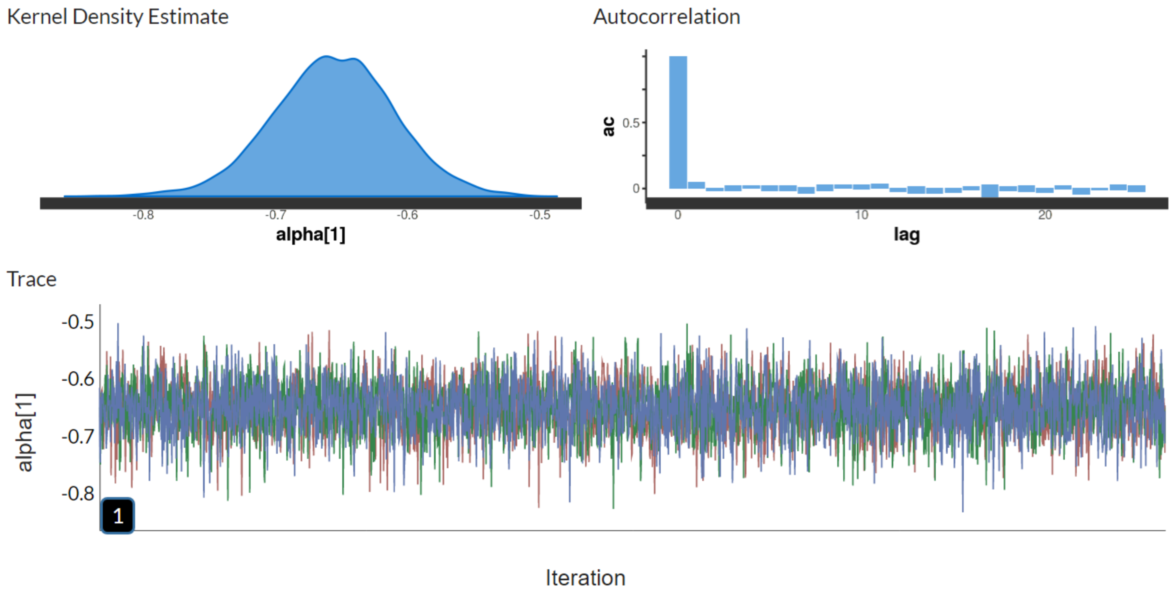


Figure 4.8: Trace Plots for β[1]

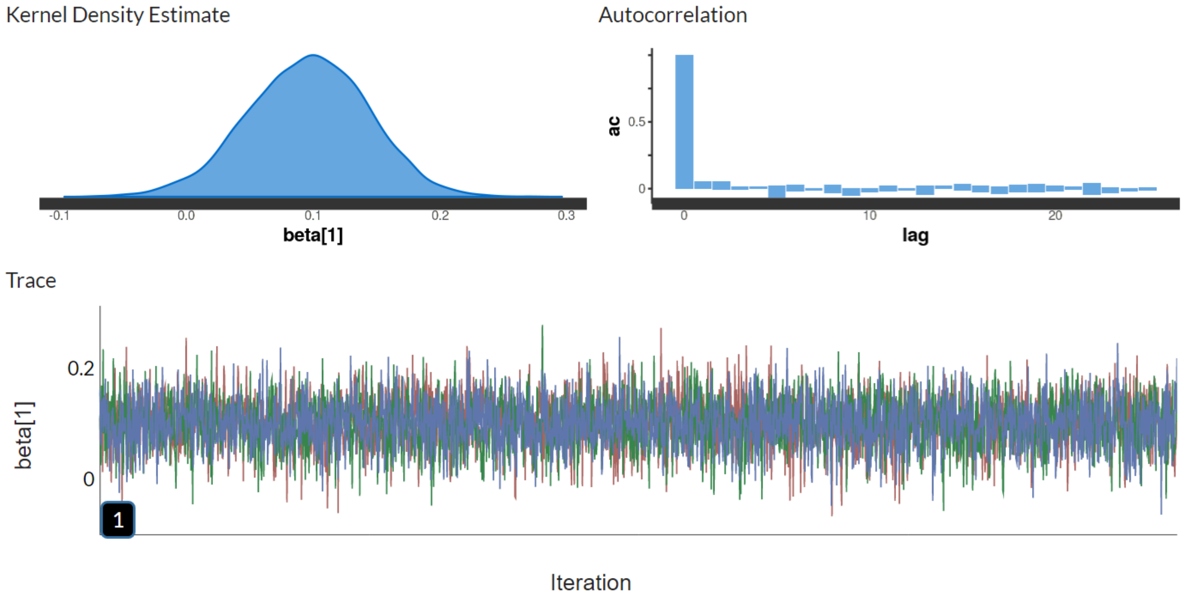


Figure 4.9: Trace Plots for mu_alpha

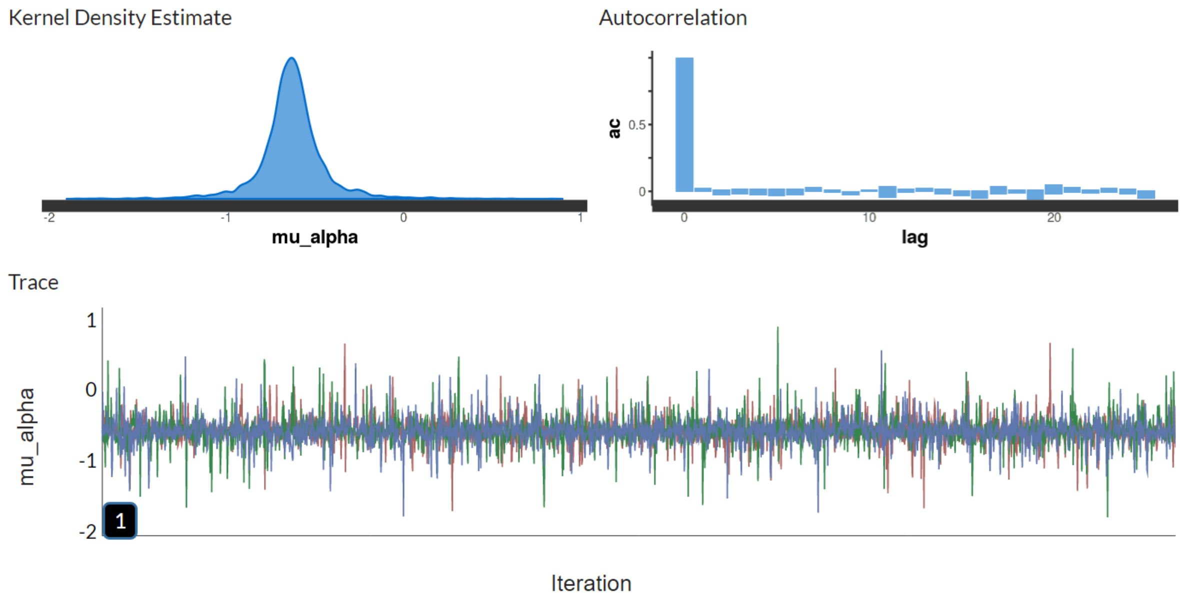
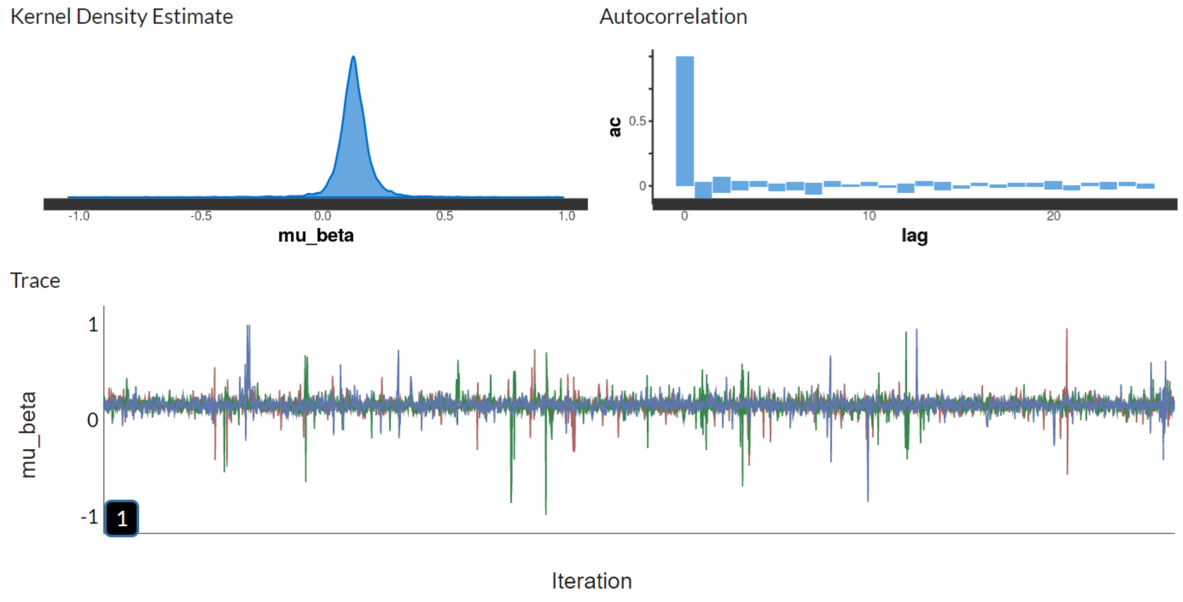


Figure 4.10: Trace Plots for mu_beta



From all the trace plots above in Figure 4.7, 4.8, 4.9 and 4.10 we can see that the Markov Chain has mixed well, it has covered enough sample space and it is not getting stuck at one place for a particular parameter. They chains starts to converge towards the target distribution quickly. There is no autocorrelation as all the lags are approximately at 0, giving indication that we are getting independant Markov Chain like draws from our posteior. Thus we have high effect of our sample size.

As the rule of thumb states that if chains are mixed well, $\hat{\mathrm{R}}$ is close to 1.1 and Effective Sample Size (ESS) > 100 which helps to achieve the same level of precision as through Markov Chain samples (Cahill, 2020). The model fit is achieving all the criteria's and there are no flagged values.

To test the accuracy of the Bayesian Hierarchal Model, I have removed the year 2018 from the data and predict the values of the same year. If the predicted values lie within the 95% Credible Interval, we can say the model fits well. The results of the prediction are shown in Table 4.6 which are better than the previous models.

Table 4.6: Testing Bayesian Hierarchal Model Accuracy through Pi_hat and Credible Interval

Subject_race	**Year**	**Pi_hat**	**95% Confidence Interval**	
Asian/Pacific Islander	2018	0.1489362	0.161	0.284
Black	2018	0.286443	0.274	0.288
Hispanic	2018	0.2263868	0.215	0.255
Other	2018	0.5	0.216	0.619
White	2018	0.2838309	0.258	0.277

*Pi_hat refers to the probability of contraband_found from raw data

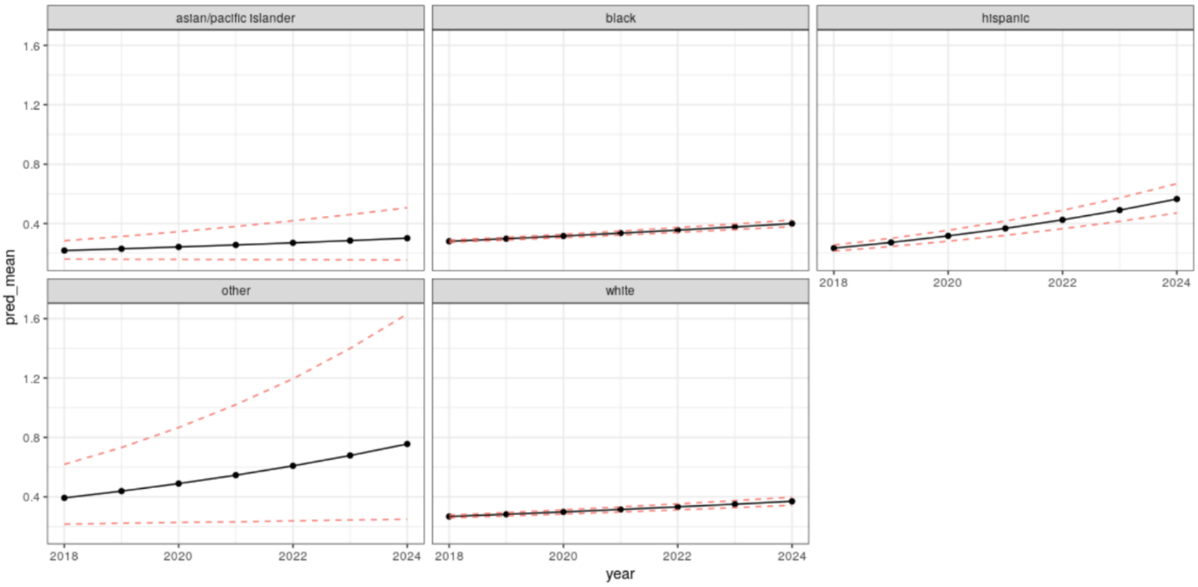
CHAPTER 5

DISCUSSION

In the period from 2010 to 2017, the MNPD's records indicate that they were stopping and searching Blacks and Hispanics more often than Whites, so I developed models to predict the likelihood of contraband being found in a vehicle when searched for different races in the coming years to see if the trend followed by the MNPD to stop and search minorities continues to increase or not. Both, Logistic Regression and Generalized Additive Model were built with a response as the probability of contrand_found and contraband_notfound as a matrix to analyse if we could build an accurate model that can give us useful predictions. In the case of Bayesian Hierarchal Model the response is contraband_found and is structured to integrate the prior information (information external to the data) rationally into the statistical analysis process.

The best model, Bayesian Hierarchal Model predicts contraband_found most accurately using variables subject_race and year. When we compare the predictions of the model with the actual probability of contraband being found (Pi_hat) for the year 2018 approximately all the values lie within the 95% Credible Interval and it can be seen in Table 4.6. This level of accuracy was not achieved in the case of Logistic Regression and GAMs which can be seen from Tables 4.2 and 4.5 respectively. When Bayesian Model was used to predict contraband_found for the future years the results of the same are shown in Figure 5.1.

Figure 5.1: Probability of contraband_found for the years 2018-2024



From the above Figure we can infer that contraband_found for minorities continues to increase which aligns with MNPD's track record of gross ethnic and racial disparities. As we can see for Hispanics it increases at a faster rate than Whites. We do not have much information as to what subject_race "other" has been categorized by the police officers, by looking at the plot they have the highest probability of contranband_found by the year 2024. While for other races it increases at a constant rate, from which we can conclude through our results that in the coming years, this trend of stopping and searching shall continue.

Our model lacks information about police discretion that leads to their decision of searching a vehicle. As sometimes that discretion can lead to effective and efficient outcomes (Tillyer & Klahm, 2011) while others argue that it can lead to unfair and unequal treatment of the minority population due to unbridled discretion, which is supported by our results. While the results of this report reflect the reality of racial biasing, these findings should not be viewed as the main or primary explanation for stopping and searching behaviour of MNPD.

Indeed, all the statistical models with their restricted strength indicates that there may be some unmeasured factors that could prove to be much stronger predictors of contraband_found than those we have used in our models. Furthermore, subject_race is likely to act as a proxy for other unmeasured predictors that affect the behaviour of MNPD.

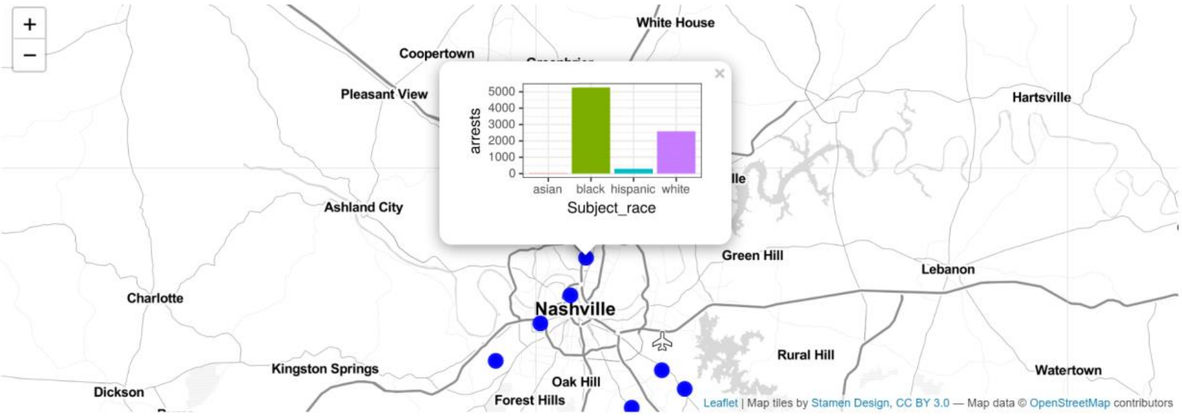
CHAPTER 6

CONCLUSION

Racial oppression is foundational to and deeply ingrained in US history and is operational throughout societal levels group relations, institutions, organizations, power structures. Majority of Americans expect at least some change to the policing practises currently, with 55% of Americans wanting a substantial reform or complete redesigning of the system. Only about 7% would want to keep it the same (Jackson & Machi, 2020). From the abolition of slavery, to the black freedom movements of the 1950s, 60s, and 70s, to the critical mass emerging in opposition to mass incarceration and racialized state and police violence today that we all have seen with the Black Lives Matter movement.

To briefly summarise our findings, Metro Nashville Police Department has employed stop and search strategy as a way to address crime and violence, due to which Black drivers are arrested at 1.45 times more than White drivers out of the total stops made while they only represent 28% of the population in Nashville, Tennessee. When we look at these values in each precinct, Black and Hispanic drivers are much more likely to be searched, almost 2 - 3 times higher than the total searches for whites which can also be seen from Figure 6.1 below and previously seen in Figure 2.4.

Figure 6.1: Spatial Analysis of MNPD Searches in Each Precinct By Race



The findings in Section 2.1 gives us the evidence that MNPD officers significantly arrest minority drivers at a higher rate. In addition, I find that the vast majority of MNPD's consent and pat down searches are against innocent civilians who have no contraband or weapons when their vehicle is searched, proved by the significantly higher search rates than white drivers and this bias seems to be systematic throughout Nashville. The bias in searches is a compelling evidence of racial discrimination by an officer as Blacks and Hispanics have a lower hit-rate when compared with Whites. Indeed, the constant differences in hit-rates supports the hypothesis of unbridled discretion when stopping Black and Hispanic drivers such that MNPD is always less accurate at determining when a search for Black or Hispanic driver is likely to result in the finding of contraband such as weapon, marijuana or other drugs.

To reduce the damage caused from these tactics imposed on Nashville's Black and Hispanic drivers, MNPD should try to strengthen the relation between its officers and the people it serves. Many new policies are being explored as the previous strategy to stop and search failed to make an impact to reduce crime. The findings of my report support the 2016 study of "Driving While Black" which is published by Gideon's Army, a local non-profit and it reflects the harm caused to minority communities of Nashville by such policing practices.

CHAPTER 7

REFERENCES

Anwar, S. & Hanming, F., 2006. An Alternative Test of Racial Prejudice in Motor Vehicle. *The American Economic Review,* March, 96(1), pp. 127-151.

Becker, G., 2010. *The Economics of Discrimination.* Chicago: University of Chicago Press.

Bertrand, M. & Mullainathan, S., 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review,* September, 94(4), pp. 991-1,013.

Cahill, N., 2020. ST466: Advanced Statistical Modeling. *Maynooth University.*

Cahill, N., Kemp, A. C., Horton, B. P. & Parnell, A. C., 2016. A Bayesian hierarchical model for reconstructing relative sea. *Climate of the Past,* 12(2), pp. 525-542.

DiJulio, B., Norton, M., Jackson, S. & Brodie, M., 2015. *Survey of Americans on Race,* s.l.: Kaiser Family Foundation/CNN.

Feagin, J., 2013. *Systemic Racism: A Theory of Oppression.* s.l.:s.n.

Feagin, J. & Feagin, C., 1978. Discrimination American style: Institutional racism and sexism. *Prentice Hall,* Volume 510.

Feagin, J. & lias, S., 2013. Rethinking racial formation theory: a systemic. *Ethnic and Racial Studies,* 36(6), pp. 931-960.

Goldman, N., 1963. The differential selection of juvenile offenders for court appearance. *National Research and Information Center, National Council on Crime and Delinquency..*

Harris, D., 1999. Driving while Black: Racial profiling on our nation's highways. *American Civil Liberties Union.*

Hastie, T. J. & Tibshirani, R. J., 1990. *Generalized Additive Models.* London: Chapman and Hall.

Hastie, T. & Tibshirani, R., 1986. Generalized Additive Models. *Statistical Science,* 1(3), pp. 297-310.

Hilbe, J., 2011. Logistic Regression. *International encyclopedia of statistical science,* Volume 1, pp. 15-32.

Jackson, C. & Machi, S., 2020. America’s hidden common ground on police reform and racism in the United States. *Public Agenda/USA Today/Ipsos Hidden Common Ground poll,* 29 June.

Knowles, J., Persico, N. & Todd, P., 2001. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy,* 109(1), p. 203–229.

Knowles, J., Persico, N. & Todd, P., 2001. Racial Bias in Motor Vehicle Searches: Theory and Evidence. *Journal of Political Economy,* 109(1), pp. 203-229.

Kruschke, J., 2014. *Doing Bayesian Data Analysis : A Tutorial with R, JAGS, and Stan.* s.l.:Elsevier Science & Technology.

Liao, T., 1994. *Interpreting probability models: Logit, probit, and other generalized linear models.* s.l.:Sage.

Long, J. S., 1997. Regression Models for Categorical and Limited Dependent Variables.. *Sage.*

Lundman, R. J. & Kaufman, R. L., 2003. Driving while black: Effects of race, ethnicity, and gender on citizen self-reports of traffic stops and police actions. *Criminology,* February, 41(1), p. 195.

McCullagh, P. & Nelder, J., 1989. *Generalized Linear Models.* London: Chapman and Hall.

Morris, A., 2020. We'll Never Fix Systemic Racism by Being Polite. *Scientific American,* 3 August.

Orloff, J. & Bloom, J., 2014. Comparison of frequentist and Bayesian inference.. *MIT OpenCourseWare.*

Peng, C. & So, T., 2002. Logistic regression analysis and reporting: A primer. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences,* 1(1), pp. 31-70.

Pierson, E. et al., 2017. A large-scale analysis of racial disparities in police stops across the United States. *Stanford University working paper,* August.

Project, P., 2018. *An Assessment of Traffic Stops and Policing Strategies in Nashville,* New York: New York University School of Law.

Sparks, G., 2020. Polling highlights stark gap in trust of police between black and white Americans. *CNN.*

Tillyer, R. & Klahm, C., 2011. Searching for Contraband: Assessing the Use of Discretion by Police Officers. *Police Quarterly,* 14(2), p. 166 –185.

Wood, S. N., 2006. Generalized Additive Models: An Introduction with R. *CRC Press,* Volume 66.

Wundervald, B., 2018. JAGS: Just Another Gibbs Sampler.

APPENDIX A

Table A.1: Description of Dataset

Column name	Column meaning	Example value
raw_row_number	An number used to join clean data back to the raw data	38299
date	The date of the stop, in YYYY-MM-DD format. Some states do not provide the exact stop date: for example, they only provide the year or quarter in which the stop occurred. For these states, stop_date is set to the date at the beginning of the period: for example, January 1 if only year is provided.	"2017-02-02"
time	The 24-hour time of the stop, in HH:MM format.	20:15
location	The freeform text of the location. Occasionally, this represents the concatenation of several raw fields, i.e. street_number, street_name	"248 Stockton Rd."
lat	The latitude of the stop. If not provided by the department, we attempt to geocode any provided address or location using Google Maps. Google Maps returns a "best effort" response, which may not be completely accurate if the provided location was malformed or underspecified. To protect against spurious responses, geocodes more than 4 standard deviations from the median stop lat/lng are set to NA.	72.23545
lng	The longitude of the stop. If not provided by the department, we attempt to geocode any provided address or location using Google Maps. Google Maps returns a "best effort" response, which may not be completely accurate if the provided location was malformed or underspecified. To protect against suprious responses, geocodes more than 4 standard deviations from the median stop lat/lng are set to NA.	115.2808

precinct	Police precinct. If not provided, but we have retrieved police department shapefiles and the location of the stop, we geocode the stop and find the precinct using the shapefiles.	8
reporing_area	Police reporting area. If not provided, but we have retrieved police department shapefiles and the location of the stop, we geocode the stop and find the reporting area using the shapefiles.	8
zone	Police zone. If not provided, but we have retrieved police department shapefiles and the location of the stop, we geocode the stop and find the zone using the shapefiles.	8
subject_age	The age of the stopped subject. When date of birth is given, we calculate the age based on the stop date. Values outside the range of 10-110 are coerced to NA.	54.23
subject_race	The race of the stopped subject. Values are standardized to White, Black, Hispanic, Asian/pacific islander, and other/unknown	"hispanic"
subject_sex	The recorded sex of the stopped subject.	"female"
officer_id*	Officer badge number or other form of identification provided by the department.	8
type	Type of stop: vehicular or pedestrian.	"vehicular"
violation	Specific violation of stop where provided. What is recorded here varies widely across police departments.	"SPEEDING 15-20 OVER"
arrest_made	Indicates whether an arrest made.	FALSE
citation_issued	Indicates whether a citation was issued.	TRUE
warning_issued	Indicates whether a warning was issued.	TRUE
outcome	The strictest action taken among arrest, citation, warning, and summons.	"citation"
contraband_found	Indicates whether contraband was found. When search_conducted is NA, this is coerced to NA under the assumption that contraband_found shouldn't be discovered when no search occurred and likely represents a data error.	FALSE

contraband_drugs	Indicates whether drugs were found. This is only defined when contraband_found is true.	TRUE
contraband_weapons	Indicates whether weapons were found. This is only defined when contraband_found is true.	TRUE
frisk_performed	Indicates whether a frisk was performed. This is technically different from a search, but departments will sometimes include frisks as a search type.	TRUE
search_conducted	Indicates whether any type of search was conducted, i.e. driver, passenger, vehicle. Frisks are excluded where the department has provided resolution on both.	TRUE
search_person	Indicates whether a search of a person has occurred. This is only defined when search_conducted is TRUE.	TRUE
search_vehicle	Indicates whether a search of a vehicle has occurred. This is only defined when search_conducted is TRUE.	TRUE
search_basis	This provides the reason for the search where provided and is categorized into k9, plain view, consent, probable cause, and other. If a search occurred but the reason wasn't listed, we assume probable cause.	"consent"
reason_for_stop	A freeform text field indicating the reason for the stop where provided.	"EQUIPMENT MALFUNCTION

(Source: Stanford Open Policing Project, https://github.com/5harad/openpolicing)

APPENDIX B: R CODE

```
## Loading Required Libraries
library(lubridate)
library(tidyverse)
library(ggplot2)
library(mgcv)
library(R2jags)
library(rjags)
library(plotly)
library(RColorBrewer)
library(shinystan)

## Read in the data
Nashville File link: https://maynoothuniversity-
my.sharepoint.com/:x:/g/personal/raghav_chadha_2020_mumail_ie/EfRm-
Emvda9Li2yz95kcciUBvFZZSMjfZZcyqhMQaYZYyA?e=zNelmb
tn_nv <- read_csv("tn_nashville_2020_04_01.csv")

## Computing summary statistics broken down by the race of the driver (TABLE 2.1)-------------
summary_stats <- function(search_conducted, contraband_found) {
  n_stops     = length(search_conducted)
  n_searches  = sum(search_conducted, na.rm = T)
  n_hits      = sum(contraband_found, na.rm = T)
  search_rate = n_searches / n_stops
  hit_rate    = n_hits / n_searches
  return(data.frame(n_stops, n_searches, n_hits, search_rate, hit_rate))
}
## Pull out data relevant for searches
tn_search_raw <- tn_nv %>%
  select(subject_race,search_conducted, contraband_found,date) %>%
  mutate(year = lubridate::year(date))
basic_summary_statistics_by_race = tn_search_raw %>%
 group_by(subject_race) %>%
  do(summary_stats(.$search_conducted, .$contraband_found)) %>%
  drop_na()
basic_summary_statistics_by_race
## Comparing drivers stopped in Nashville with neighbouring cities----------------------------------
## Nashville
tn_nv_new <- tn_nv %>%
  select(subject_race, subject_sex, outcome, date, time) %>%
  filter(year(date) >  2013 & year(date) < 2016)
tn_nv_new$city <- "Nashville"
## Madison file link: https://maynoothuniversity-
my.sharepoint.com/:x:/g/personal/raghav_chadha_2020_mumail_ie/EZX2ZQlatw5MjMvSWcOK
vXUBMv_Tl6z6QYU7c2WDMrumLg?e=foDExV
wi_m <- read_csv("wi_madison_2019_12_17.csv")
wi_m <- wi_m %>%
  select(subject_race, subject_sex, outcome, date, time) %>%
  filter(year(date) >  2013 & year(date) < 2016)
wi_m$city <- "Madison"
```

```
## Columbus file link: https://maynoothuniversity-my.sharepoint.com/:x:/g/personal/raghav_chadha_2020_mumail_ie/EfRm-Emvda9Li2yz95kcciUBvFZZSMjfZZcyqhMQaYZYyA?e=zNelmb
oh_c <- read_csv("oh_columbus_2019_12_17.csv")
oh_c <- oh_c %>%
  select(subject_race, subject_sex, outcome, date, time) %>%
  filter(year(date) >  2013 & year(date) < 2016)
oh_c$city <- "Columbus

## Charotte file link: https://maynoothuniversity-my.sharepoint.com/:x:/g/personal/raghav_chadha_2020_mumail_ie/EfRm-Emvda9Li2yz95kcciUBvFZZSMjfZZcyqhMQaYZYyA?e=zNelmb
nc_c <- nc_c %>%
  select(subject_race, subject_sex, outcome, date, time) %>%
  filter(year(date) >  2013 & year(date) < 2016)
nc_c$city <- "Charlotte"

## Creating a dataframe for all the cities to plot
df <- rbind(oh_c, wi_m,tn_nv_new, nc_c)
df1 <- data.frame(df)

## Summarise the results
stops_byyearcity <- df1 %>%
 group_by(city, year = year(date)) %>%
 summarise(stops = n())

## Plottting the results (FIGURE 2.1)--------------------------------------------------------------------------
plot1 <- ggplot(data = stops_byyearcity, mapping = (aes(x=reorder(city,stops), y= stops, fill = city)))+
   geom_bar(stat="identity") +
   ggtitle("Drivers Stopped in Cities by Year") +
   theme_bw() +
   scale_y_continuous(labels = scales::comma) +
   xlab("City") +
   ylab("Stopped Drivers") +
   theme(legend.position = "none") +
   facet_wrap(~year) +
   theme_bw() +
   theme(legend.position = "none")

ggplotly(plot1)
## Calculating stops made by MNPD each year
stops_made <- tn_nv %>%
   group_by(year = year(date), subject_race) %>%
   summarise(stops = n()) %>%
   filter(year != 2019) %>%
   drop_na()
## Plotting the results of stop_made (FIGURE 2.2)---------------------------------------------------------
plot2 <- ggplot(data = stops_made, mapping = aes(x = year, y = stops, colour = subject_race))+
  scale_color_brewer(palette = "Dark2") +
  geom_line() +
  geom_point() +
  theme_bw() +
  xlab("Year") +
  ylab("Stopped Driver") +
  ggtitle("Stopped Drivers from 2010-2018 according to Race") +
  theme(legend.title = element_blank())

ggplotly(plot2)

## Calculating Arrests made by MNPD each year
arrests_byyear <- tn_nv %>%
    group_by(year = year(date), arrest_made, subject_race) %>%
    summarise(arrest = sum(arrest_made)) %>%
    filter(arrest_made == "TRUE") %>%
    filter(year != 2019) %>%
    drop_na()

## Plotting the results of arrests_byyear (FIGURE 2.3)----------------------------------------------------
plot3 <- ggplot(data = arrests_byyear, mapping = aes(x = year, y = arrest, colour =
subject_race)) +
  scale_color_brewer(palette = "Dark2") +
  geom_line() +
  geom_point() +
  theme_bw() +
  xlab("Year") +
  ylab("Arrested Drivers") +
  ggtitle("Arrested Drivers from 2010-2018 according to Race") +
  theme(legend.title = element_blank())

ggplotly(plot3)

## Scatterplot which compares search rates and hit rates or minority and white drivers within
the same precinct-----------------------------------------------------------------------------------------------------
basic_summary_statistics_by_race_and_precinct = tn_nv %>%
  filter(!is.na(precinct)) %>%
  group_by(subject_race, precinct) %>%
  do(summary_stats(.$search_conducted, .$contraband_found))

basic_summary_statistics_by_race_and_precinct

data_for_plot <- basic_summary_statistics_by_race_and_precinct %>%
  filter(subject_race == 'white') %>%
  right_join(basic_summary_statistics_by_race_and_precinct %>%
  filter(subject_race != 'white'), by='precinct')

## Plot search rates (FIGURE 2.4)------------------------------------------------------------------------------
max_val = max(basic_summary_statistics_by_race_and_precinct$search_rate) * 1.05

search_plot <- ggplot(data_for_plot) +
geom_point(aes(x = search_rate.x, y = search_rate.y, size = n_stops.y)) + # specify data we  want
to plot
facet_grid(.~subject_race.y) +   # make one subplot for each minority race group
geom_abline(slope = 1, intercept = 0, linetype='dashed') +   # add a diagonal line to indicate
parity
```

```
scale_x_continuous('White search rate', limits=c(0, max_val), labels = scales::percent,
expand=c(0,0)) +
scale_y_continuous('Minority search rate', limits=c(0, max_val), labels = scales::percent,
expand=c(0,0)) +
theme_bw() +
theme(legend.position="none") +
scale_size_area(max_size=5)
 search_plot

## Now let's compute hit rates by race and district (FIGURE 2.5)----------------------------------------
hit_rates <-  tn_nv %>%
  filter(search_conducted) %>%
  group_by(subject_race, precinct) %>%
  summarize(hit_rate = mean(contraband_found, na.rm = T)) %>%
  drop_na()

hit_rates

## Reshape table to show hit rates of minorities vs white drivers
hit_rates <-
  hit_rates %>%
  filter(subject_race %in% c("black", "white", "hispanic")) %>%
  spread(subject_race, hit_rate, fill = 0) %>%
  rename(white_hit_rate = white) %>%
  gather(minority_race, minority_hit_rate, c(black, hispanic)) %>%
  arrange(precinct)

hit_rates

## Get corresponding number of searches (to size points). For each district we want to know the
number of white+black searches and white+Hispanic searches----------------------------------------
search_counts <- tn_nv %>%
  filter(
    search_conducted,
    subject_race %in% c("black", "white", "hispanic")
  ) %>%
  count(precinct, subject_race) %>%
  spread(subject_race, n, fill = 0) %>%
  rename(num_white_searches = white) %>%
  gather(minority_race, num_minority_searches, c(black, hispanic)) %>%
  mutate(num_searches = num_minority_searches + num_white_searches) %>%
  select(precinct, minority_race, num_searches) %>%
  drop_na()

## We'll use this just to make our axes' limits nice and even
max_hit_rate <- hit_rates %>%
  select(ends_with("hit_rate")) %>%
  max()

hit_rates %>%
  ggplot(aes(
    x = white_hit_rate,
    y = minority_hit_rate
  )) +
  geom_point() +
  # This sets a diagonal reference line (line of equal hit rates)
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  # These next few lines just make the axes pretty and even
  scale_x_continuous("White hit rate",
               limits = c(0, max_hit_rate + 0.01),
               labels = scales::percent  ) +
  scale_y_continuous("Minority hit rate",
               limits = c(0, max_hit_rate + 0.01),
               labels = scales::percent
  ) +
  # This makes sure that 1% on the x-axis is the same as 1% on the y-axis
  coord_fixed() +
  # This allows us to compare black v. white and Hispanic v. white side by side, in panels
  facet_grid(. ~ minority_race)

hit_rates %>%
  left_join(
    search_counts,
    by = c("precinct", "minority_race")
  ) %>%
  ggplot(aes(
    x = white_hit_rate,
    y = minority_hit_rate
  )) +
  geom_point(aes(size = num_searches), pch = 21) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  scale_x_continuous("White hit rate",
               limits = c(0, max_hit_rate + 0.01),
               labels = scales::percent
  ) +
  scale_y_continuous("Minority hit rate",
               limits = c(0, max_hit_rate + 0.01),
               labels = scales::percent
  ) +
  coord_fixed() +
  facet_grid(. ~ minority_race) +
  theme_bw()

## Prob contraband when searched (within race groups)
tn_contraband <- tn_search_raw %>%
  select(subject_race,year,contraband_found)%>%
  drop_na() %>%
  group_by(subject_race,year) %>%
  summarise(total_searches = n(),
          contraband_found = sum(contraband_found),
          contraband_notfound = total_searches - contraband_found,
          pi_hat = sum(contraband_found)/n()) %>%
  ungroup()

## LOGISTSIC REGRESSION---------------------------------------------------------------------------------
## Removing year 2019 due to lack of data and subject_race “unknown”
tn_contraband_new <- tn_contraband %>%
  filter(year!=2019) %>%
  filter(subject_race!="unknown")

## Removing 2018 to check accuracy of model on latest year using CI
tn_contraband_new_2018 <- tn_contraband_new %>%
  filter(year!=2018)

## Creating matrix of contraband found and not found to give as response
glm_data <- tn_contraband_new_2018 %>%
  select(contraband_found, contraband_notfound) %>%
  as.matrix()

## Fitting the Logistic Regression Model
fit_lr <- glm(glm_data ~ year + subject_race, data = tn_contraband_new_2018, family =
binomial)

summary(fit_lr)
exp(coef(fit_lr))

anova(fit_lr, test="Chisq")
# It appears all terms are making a significant contribution to the model.
teststat <- sum(residuals(fit_lr, type = "deviance")^2)
teststat
teststat > qchisq(0.95,34)

## Predictions
pred_lr <- as.tibble(predict(fit_lr, type = "response",se.fit = T))
pred_lr$year = tn_contraband_new_2018$year
pred_lr$subject_race = tn_contraband_new_2018$subject_race
pred_lr <- pred_lr %>%
  mutate(upr = fit+(1.96*se.fit),
         lwr = fit-(1.96*se.fit))

## Plot of contraband_found with Overlaid Predictions from Logistic Regression for each Race &
Year (FIGURE 4.1)
ggplot(tn_contraband_new_2018 , aes(x = year, y = pi_hat)) +
  geom_point() +
  geom_line(data = pred_lr, aes(x = year, y = fit))+
  geom_line(data = pred_lr, aes(x= year, y = lwr, colour = "red"), linetype="dashed")+
  geom_line(data = pred_lr, aes(x= year, y = upr, colour = "red"), linetype="dashed")+
  facet_wrap(~subject_race) +
  theme_bw() +
  theme(legend.position = "none")

## Diagnostic plots (FIGURE 4.2)
par(mar=c(1,1,1,1))
plot(fit_lr)
par(mfrow=c(1,1))

## GENERALIZED ADDITIVE MODELS (GAMS) -------------------------------------------------------
fit_g <- mgcv::gam(glm_data ~ s(year, bs = "cr", k = 3) + subject_race, data =
tn_contraband_new_2018, family = binomial(), method = "REML")
summary(fit_g)
```

```
## Predictions
pred_gam <- as.tibble(predict(fit_g, type = "response", se = TRUE))
pred_gam$year = tn_contraband_new_2018$year
pred_gam$subject_race = tn_contraband_new_2018$subject_race
pred_gam <- pred_gam %>%
 mutate(upr = fit+(1.96*se.fit),
        lwr = fit-(1.96*se.fit))

## Plot of contraband_found with Overlaid Predictions from Generalized Additive Model for each
Race & Year (FIGURE 4.4)
ggplot(tn_contraband_new_2018 , aes(x = year, y = pi_hat)) +
 geom_point() +
 geom_line(data = pred_gam, aes(x = year, y = fit))+
 geom_line(data = pred_gam, aes(x= year, y = lwr, colour = "red"), linetype="dashed")+
 geom_line(data = pred_gam, aes(x= year, y = upr, colour = "red"), linetype="dashed")+
 facet_wrap(~subject_race) +
 theme_bw() +
 theme(legend.position = "none")

## Diagnostic Plots (FIGURE 4.5)
par(mfrow=c(2,2))
mgcv::gam.check(fit_g)

## BAYESIAN HEIRARICHAL MODEL------------------------------------------------------------------------
tn_contraband_new_2018$subject_race <- as.factor(tn_contraband_new_2018$subject_race)
as.numeric.factor <- function(x) {as.numeric(levels(x))[x]}
tn_contraband_new_2018$race_index <- as.numeric (as.factor (
tn_contraband_new_2018$subject_race ) )

## Specify the JAGS model
binom_model = "
model{
for( i in 1:n_obs ) {
#Binomial likelihood
y.i[i] ~ dbinom(p.i[i], N.i[i])
}
for( i in 1:n_obs ) {
nu.i[i] <- alpha[race_index[i]] + beta[race_index[i]]*t.i[i] #separate regression for each race
p.i[i] <- exp(nu.i[i])/(1+exp(-nu.i[i]))
}

## Hierarchical priors with race specific parameters centered on overall parameter with cross
race variation
for( j in 1:n_race ){
alpha[j] ~ dnorm(mu_alpha,sigma_alpha^-2)
beta[j] ~ dnorm(mu_beta,sigma_beta^-2)
}
## Priors for overall parameter with cross race variation
mu_alpha ~ dnorm(0,0.001)
mu_beta ~ dnorm(0,0.001)sigma_alpha ~ dt(1,1,10^-2)T(0,)
sigma_beta ~ dt(1,1,10^-2)T(0,)
}
"
```

```
## Specify the JAGS data
jags.data <- list(y.i = tn_contraband_new_2018$contraband_found,
            t.i = scale(tn_contraband_new_2018$year)[,1], #to standardise time to get the model to
work
            N.i = tn_contraband_new_2018$total_searches,
            n_obs = nrow(tn_contraband_new_2018),
            race_index = tn_contraband_new_2018$race_index,
            n_race = max(tn_contraband_new_2018$race_index))

## monitor parameters
parnames <- c("p.i","alpha", "beta","mu_alpha","mu_beta")
mod <- jags(data = jags.data, parameters.to.save=parnames,
        model.file = textConnection(binom_model),
        n.iter = 15000,
        n.burnin = 3000,
        n.thin = 6)
plot(mod)

## Create output objects
mcmc.array <- mod$BUGSoutput$sims.array
dim(mcmc.array)

## Summary output
mod$BUGSoutput$summary

## Plot of contraband_found with Overlaid Predictions from Bayes Model for each Race & Year
(FIGURE 4.6)
p_mean <- as_tibble(mod$BUGSoutput$mean$p.i)
p_mean$year <- tn_contraband_new_2018$year
p_mean$subject_race = tn_contraband_new_2018$subject_race
p_mean$lwr = apply(mod$BUGSoutput$sims.list$p.i,2,quantile,probs = 0.025)
p_mean$upr = apply(mod$BUGSoutput$sims.list$p.i,2,quantile,probs = 0.975)

ggplot(tn_contraband_new_2018 , aes(x = year, y = pi_hat)) +
 geom_point() +
 geom_line(data = p_mean, aes(x= year, y = value)) +
 geom_line(data = p_mean, aes(x= year, y = lwr, colour = "red"), linetype="dashed")+
 geom_line(data = p_mean, aes(x= year, y = upr, colour = "red"), linetype="dashed")+
 facet_wrap(~subject_race) +
 theme_bw() +
 theme(legend.position = "none")

## Trace Plots (FIGURE 4.7, 4.8, 4.9, 4.10)
shiny.array <- as.shinystan(mod$BUGSoutput$sims.array)
launch_shinystan(shiny.array)

## SHINY APP (FIGURE 6.1)
library(shiny)
library(leaflet)
library(leafpop)
library(lattice)

myData <- data.frame(id=rep(c(1,2,3,4,5,6,7,8),each=4),

arrests=c(34,1435,313,2056,24,5256,287,2577,85,2537,1677,1806,13,1596,174,1442,28,3327,
779,3274,9,5044,89,784,7,2370,464,1657,56,2074,723,2545),
          Subject_race=rep(c("asian","black","hispanic","white"),each=1,times=4),
          lng=rep(c(-86.89,-86.77,-86.64,-86.83,-86.67,-86.79,-86.70,-86.71),each=4),
          lat=rep(c(36.11,36.22,36.08,36.15,36.10,36.18,36.28,36.06),each=4))

folder <- tempfile()
dir.create(folder)

chronogramme<- function(dataId){
  dataFiltered<-filter(myData,id==dataId)

  p<- ggplot(dataFiltered,aes(Subject_race, arrests, fill = Subject_race))+
    geom_bar(stat = "identity",position="dodge") +
    theme_bw() +
    theme(legend.position = "none")
  return(p)
}

ui <- fluidPage(
  leafletOutput("map")
)
server <- function(input, output, session) {

  #Sortie map
  output$map <- renderLeaflet({
    leaflet()%>%
      addProviderTiles(providers$Stamen.TonerLite) %>%
      addCircleMarkers(
        layerId=~id,
        data = myData,
        lat = myData$lat,
        lng = myData$lng,
        radius = 5,
        color = 'blue',
        #stroke = FALSE,
        fillOpacity = 1,
        label = ~paste0("Precinct:",myData$id)
      )
  })

  # When map is clicked, show a popup with precinct info
  showPopup <- function(id, lat, lng) {
    chrngr <- chronogramme(id)
    svg(filename= paste(folder,"plot.svg", sep = "/"),
      width = 500 * 0.005, height = 300 * 0.005)
    print(chrngr)
    dev.off()

    content <- paste(readLines(paste(folder,"plot.svg",sep="/")), collapse = "")
    leafletProxy("map") %>% addPopups(lng, lat, content, layerId = id)
  }
  observe({
    leafletProxy("map") %>% clearPopups()
    event <- input$map_marker_click
    if (is.null(event))
      return()

    isolate({
      showPopup(event$id, event$lat, event$lng)
    })
  })
}

# Create Shiny app ----
shinyApp(ui = ui, server = server)
```