UNDERSTANDING RACIAL DISPARITY IN THE VIRGINIA COURT SYSTEM

Gaurav Anand
School of Data Science
University of Virginia
ga7er@virginia.edu

Shannon Paylor
School of Data Science
University of Virginia
sep4y@virginia.edu

Amanda West
School of Data Science
University of Virginia
amawest@umich.edu

June 22, 2021

ABSTRACT

America is comprised of some of the most racially diverse states in the world, with newest estimates indicating that nearly four out of every ten Americans identify as a race other than white. [1]. At the same time, there are currently as many Americans with criminal records as have college degrees [2]. Using data scraped from the past 20 years of Virginia circuit criminal court records, we study the relationship between a person's race and the charge they receive - misdemeanor or felony - in order to identify possible racial disparities for marijuana-related charges. We find that race, gender, county, and charge code are all nonzero factors, with Black men having the highest predicted likelihood of being charged with a felony and white women having the lowest in both a full and reduced logistic regression model.

1 Introduction

In background checks, criminal charges show up if the person under consideration has ever been charged with a crime, regardless of whether or not they were convicted. Because many employers, landlords, and banks run background checks on potential employees, tenants, and lessees, having a criminal record can be very damaging to people's well-being. Charges can be expunged from the record in certain cases so that they no longer appear in background checks. Currently, this is only applicable to charges with no conviction and only after a long, costly bureaucratic process. The Legal Aid Justice Center (LAJC) is advocating the Virginia legislature to expand which records qualify for expungement and to make expungement of certain records automatic. Code for Charlottesville, a civic tech organization, is collaborating with the LAJC to investigate the impact of such legislation, as well as whether the court data reflects any racial disparities in case outcomes. For this project, we worked closely with Code for Charlottesville to answer some of these questions.

In this paper, we apply data analysis and machine learning techniques to Virginia court case data to gain a better understanding of racial disparities within the court system. Specifically, we investigate the role of race and other descriptive variables in whether someone is charged with a misdemeanor or felony for the same crime. With the recent move toward decriminalization and eventual legalization of marijuana, we place a particular focus on these charges in our analysis.

2 Data

Virginia court data is publicly available through the state. These records are searchable by name, case number, or hearing date, but are not directly available for bulk download. A civic tech volunteer named Ben Schoenfeld has webscraped this data and made it readily available for download in anonymized csv format at https://virginiacourtdata.org. We use these files for our analysis.

The data includes circuit criminal, district criminal, circuit civil, and district civil court cases dating back as early as 2000, with the most recent year included being 2020. For criminal record expungement, we are only concerned with criminal courts. We choose to use circuit instead of district data because more years are available (circuit available back to 2000; district available back to 2009). The data contains features related to the charge (charge code, type, and description), features related to the court outcome (sentence, fine, parole, etc.), and features related to the person charged (locality, race, and gender).

3 Exploratory Data Analysis

We now explore data from circuit criminal court cases from 2019 to identify interesting patterns in the data. We chose to use 2019 data for EDA because it is the most recent data available, with the most granular corresponding Census data which allowed for quick comparisons. We make the assumption that all years will be relatively similar in terms of race and gender distributions.

3.1 Gender Distribution

We organize all criminal court cases by gender and find that the number of criminal cases filed on men is almost three times as many as the number of cases filed on women.

Table 1: Gender Distribution across 2019 Criminal Court Case Data

Sex	Count	Percentage
Male	21837	28%
Female	56283	72%

3.2 Race Distribution

We also group criminal cases by race, and find that Black and white populations account for over 99% of all criminal court cases in Virginia in 2019. The remaining populations make up the remaining 1% of the criminal cases.

Table 2: Race Distribution across 2019 Criminal Court Case Data

Race	Count	Proportion
American Indian or Alaskan Native	38	0.1%
White	43890	57.0%
Asian or Pacific Islander	260	0.3%
Black	32401	42.1%
Hispanic	368	0.5%

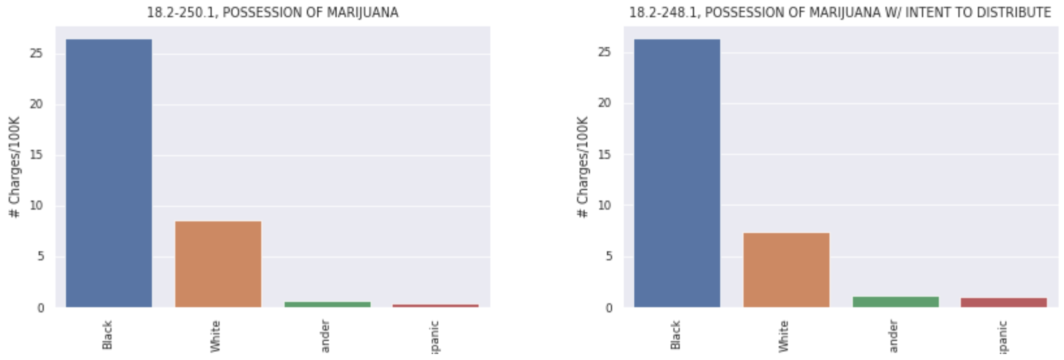
3.3 Racial Disparities in Marijuana-related Charges

Next, we shift our focus to specifically marijuana-related possession charges, in light of the recent decriminalization and future legalization of marijuana in the state of Virginia.

Since population sizes differ in the state of Virginia, we normalize the number of charges of each race by their population as given in the 2019 Census data. In Figure 1, *#Charges/100K* refers to the number of charges per 100,000 Virginia residents that identify with that race.

From Figure 1, we see that there is a large racial disparity in number of marijuana-related charges, especially between the two majority races in Virginia, white and Black. Relative to population size, Black people are charged with these crimes at roughly three times the rate of white people.

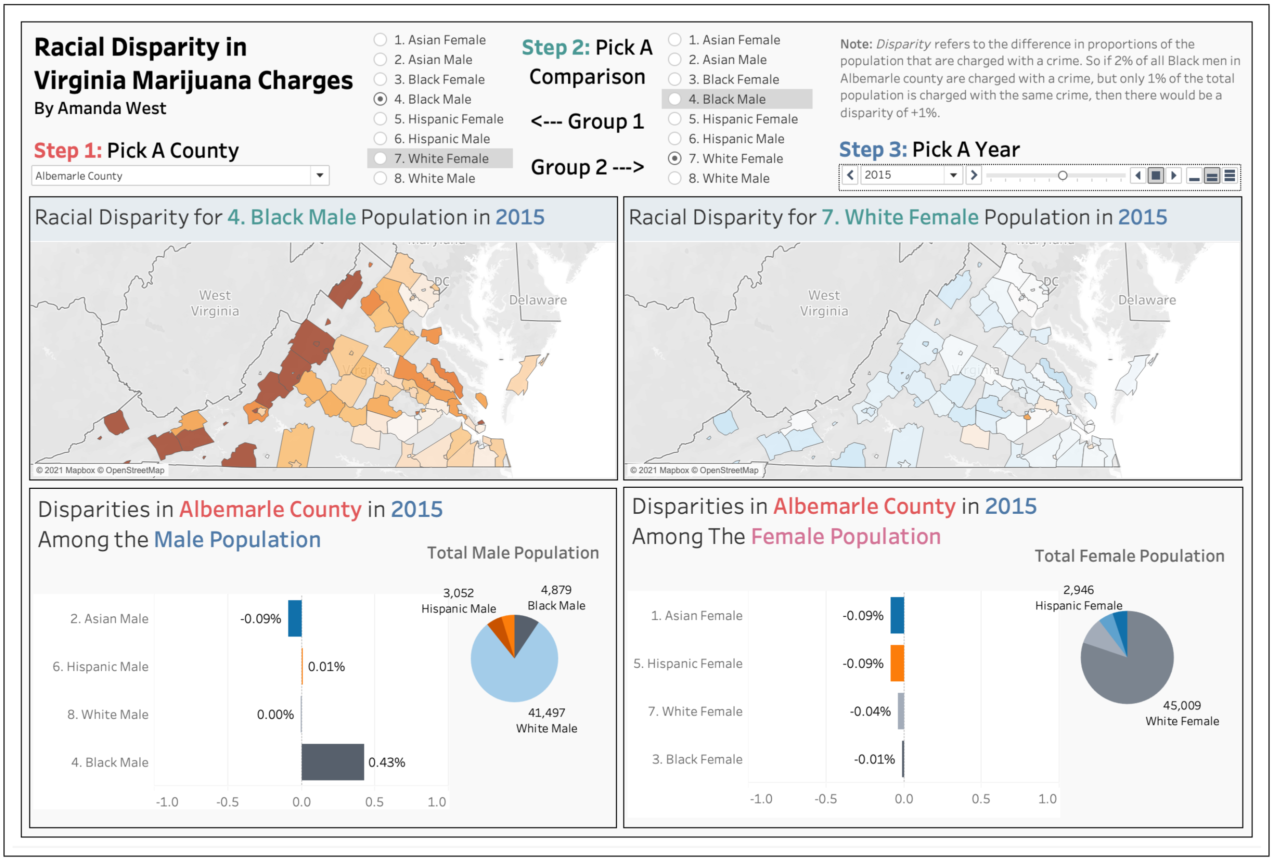
Finally, we create an interactive dashboard using Tableau in order to allow non-technical users to view the racial disparities within each county, organized by race and gender. We calculate disparity by subtracting the percentage of race x and gender y charged with a marijuana-related crime by the percentage of the overall population charged with a marijuana-related crime. Theoretically then, if 1% of the overall population was charged with a marijuana-related crime in that year, but 2% of all Black men and 0.5% of all white women were charged with a marijuana-related crime, then we would calculate the disparity at +1% for Black men and -0.5% for white women in Virginia.



(a) Possession of Marijuana

(b) Possession of Marijuana with intent to distribute

Figure 1: *Analyzing race disparities in marijuana-related criminal court cases.* (a) refers to possession of marijuana charges that are not related to distribution while (b) solely tackles the marijuana possession charges that involve intent to distribute.



Racial Disparity in Virginia Marijuana Charges

By Amanda West

Step 1: Pick A County

Albemarle County

- ○ 1. Asian Female
- ○ 2. Asian Male
- ○ 3. Black Female
- ● 4. Black Male
- ○ 5. Hispanic Female
- ○ 6. Hispanic Male
- ○ 7. White Female
- ○ 8. White Male

Step 2: Pick A Comparison

<--- Group 1

Group 2 --->

- ○ 1. Asian Female
- ○ 2. Asian Male
- ○ 3. Black Female
- ○ 4. Black Male
- ○ 5. Hispanic Female
- ○ 6. Hispanic Male
- ● 7. White Female
- ○ 8. White Male

Note: *Disparity* refers to the difference in proportions of the population that are charged with a crime. So if 2% of all Black men in Albemarle county are charged with a crime, but only 1% of the total population is charged with the same crime, then there would be a disparity of +1%.

Step 3: Pick A Year

2015

Racial Disparity for 4. Black Male Population in 2015

Racial Disparity for 7. White Female Population in 2015

Disparities in Albemarle County in 2015 Among the Male Population

Disparities in Albemarle County in 2015 Among The Female Population

Figure 2: *Interactive dashboard in Tableau.* Created by one of the authors.

3.4 Racial Disparities in Infraction Convictions

We also want to understand how race affects infractions, which are less serious than criminal court cases. To this end, we investigate the 2019 district court cases and group them by race.

In Figure 3, we see that a Black person that is charged with a particular offense is consistently more likely to be found guilty of an infraction for the five most common offenses across Virginia compared to white people charged with the same offense.

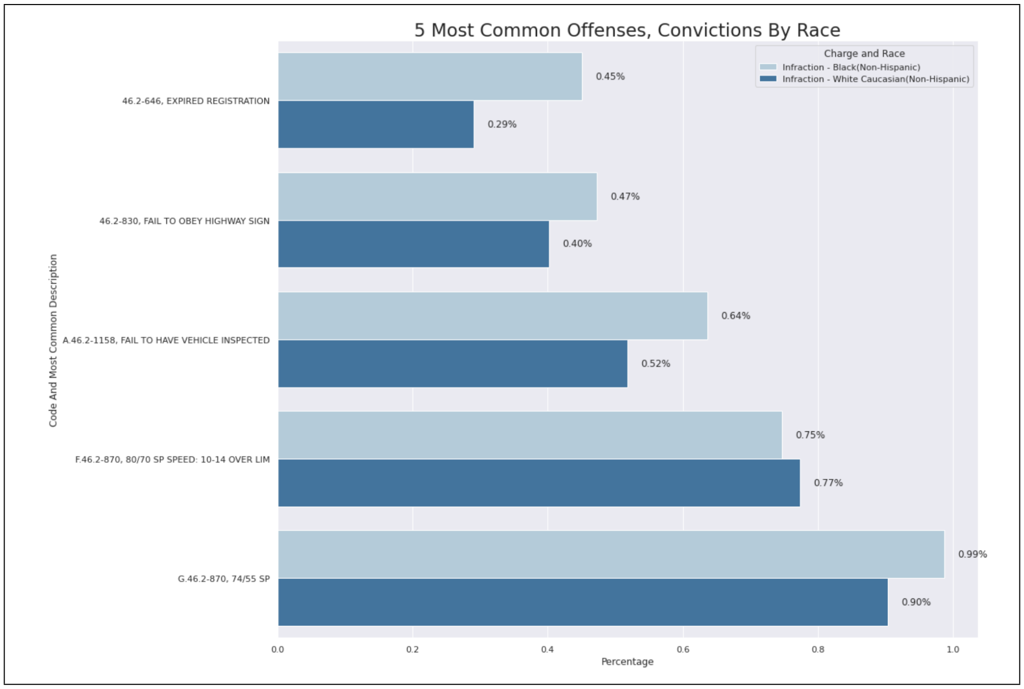


Figure 3: *Infraction Convictions As Percentage of Population*

4 Methods

Generally when creating machine learning models, we must take care to exclude variables that may make algorithms discriminate on race, gender, or other demographic characteristic. This also includes other variables such as zip code that are proxies for race and socioeconomic status. However, in this case our goal is not to generate a model to be used for predictive decision-making, but rather to make inferences about the impact of these demographic variables in past outcomes. For this reason, we choose to include race, gender, zip code, and charge code (an alphanumeric code indicating the law a person is charged with breaking) in our models. Additionally, due to our focus on inference, we chose to use relatively simple, interpretable models in our analysis rather than potentially more accurate but less explainable ones.

Since all of our predictor variables are categorical, we use one-hot encoding on each, and we convert the response variable to one for felonies and zero for misdemeanors. We remove charges below misdemeanors, such as traffic infractions, before modeling.

4.1 Random Forest Feature Importance

For our first model, we implement a random forest classifier with race, gender, county, and charge code as our predictor variables and felony or misdemeanor classification as our binary response variable. We fit the model and extract the importance score for each feature.

We see that charge code is by far the most influential, with sixty-five out of the top seventy most important features being one-hot encoded charge code values. The other five out of the top seventy most important features are particular counties. White is the only race with non-zero feature importance, and is the 73rd most important. Gender does have a non-zero feature importance, but it is less important than most other features previously mentioned. Charge code being the most important makes sense intuitively, since some minor crimes will never be felonies and other, more serious crimes will frequently be felonies, regardless of any other factors.

4.2 Logistic Regression Modeling

Next, we create a logistic regression model using the same four predictors and response variable. We use a lasso penalty to encourage the coefficients of unnecessary predictor variables to be equal to zero. We split our data into 90% for training and 10% for testing. This places a bit more data in the training set than is customary, which we choose since we are focused more on interpretation than on creating the most accurate model possible.

We also create a second, more limited logistic regression model by filtering the data to only include charges of marijuana possession with intent to distribute, since this particular charge is a focus of our investigation. Additionally, we remove county as a predictor, leaving only race and gender. We choose to exclude county because sometimes geographic indicators like county or zip code are highly correlated with race and income, and we want to see how the model changes with this predictor removed.

5 Results

In both of our logistic regression models, we find that the regression coefficients for all one-hot encoded race variables are non-zero. Both models have high accuracy, precision, and recall, though the reduced model has a significantly lower AUC. These metrics are shown in table 3.

Table 3: Performance Metrics for Logistic Regression Models

	Accuracy	Precision	Recall	AUC
Full Model	91.8%	91.7%	91.8%	96%
Reduced Model	88.3%	88.3%	77.9%	53%

Since we are especially interested in marijuana charges, we look at predictions for charge code 18.2-248.1, which is marijuana possession with intent to distribute. This charge can be either a felony or misdemeanor, while marijuana possession is always a misdemeanor.

Since our first logistic regression model includes county as a predictor, we select predictions for different demographic groups in Albemarle County to compare the predicted likelihood of being charged with a felony for this crime. The results, shown in table 4, reveal very slight predicted differences, with Black men having the highest predicted likelihood of being charged with a felony for this crime and white women having the lowest.

Table 4: Full logistic model predicted felony likelihood for marijuana possession w/ intent, Albemarle County

	Black	White
Male	92.4%	92.2%
Female	91.8%	91.6%

In our second logistic regression model, county is not a predictor, so our resulting predictions are location-agnostic within the state. For that model, we see a slightly wider discrepancy among different races and genders. However, we see the same pattern as in the previous model, where Black men have the highest predicted likelihood of being charged with a felony and white women have the lowest predicted likelihood. It is not clear whether the larger difference compared to the first model arises from the removal of multicollinearity between county and race or from the model overreaching in trying to predict the response from so few predictor variables.

Table 5: Reduced logistic model predicted felony likelihood for marijuana possession w/ intent

	Black	White
Male	89.8%	88.3%
Female	89.1%	87.4%

6 Conclusions

In this study, we sought to measure the role race and other variables play in whether someone is charged with a misdemeanor or a felony for the same crime. With the move towards decriminalization and eventual legalization of marijuana, we specifically studied the role of race and other variables in whether someone is charged with a misdemeanor or a felony for an identical crime for marijuana possession-related charges.

As described in tables 3 and 4, we found that both our full logistic model and reduced logistic model predicted that Black men had the highest predicted likelihood of being charged with a felony as opposed to a misdemeanor, with white women having the lowest likelihood. Further, the first model predicted that white men had a higher likelihood of being charged with a felony than Black women, while the second model suggested the opposite. This coincides with our findings in the exploratory data analysis, which suggested that Black people were charged with felonies at disproportionately high rates relative to their white counterparts of the same gender. Both models had high accuracy, precision, and recall, though the full model had better predictive performance than the reduced one.

All data is publicly available for the last twenty years at Virginia Court Data, and our own work, code, and data samples can be found on GitHub. As stated, these results are not conclusive, and are meant to be inferential rather than predictive. However, we hope that our work opens the door to future research regarding current racial disparities that may exist in the Virginia court system.

References

[1] William Frey. The nation is diversifying even faster than predicted, according to new census data. 2020.

[2] Gary Fields and John Emshwiller. As arrest records rise, americans find consequences can last a lifetime. 2014.