Identifying County Outliers in the South Carolina Incident-Based Reporting System (SCIBRS) 2011–2015:

Sensitivity Analysis by Varying Operational Definitions of Intimate-Partner Violence (IPV)

Prepared for the South Carolina Statistical Analysis Center

September 2018

Prepared by Stonewall Analytics

Authors:

Todd C. Leroux, PhD, and Chad A. Smith, PhD

Stonewall Analytics, LLC
(301) 547-5691
www.stonewallanalytics.com
info@stonewallanalytics.com

Disclaimer:

The authors are grateful for funding received through the United States Department of Justice (DOJ), Office of Justice Programs (OJP), Bureau of Justice Statistics (BJS) Grant Number 2016-BJ-CX-K022. Any views expressed in this research—including those related to statistical, methodological, technical, or operational issues—are solely those of the authors and do not necessarily reflect the official position or policies of the DOJ, OJP, BJS, or the South Carolina Statistical Analysis Center located in the South Carolina Department of Public Safety's Office of Highway Safety and Justice Programs.

EXECUTIVE SUMMARY

In 1987, South Carolina served as the pilot state for the FBI's NIBRS—which is the National Incident-Based Reporting System (NIBRS). In 1991, the South Carolina Incident-Based Reporting System (SCIBRS) was the first uniform crime reporting program in the nation to become NIBRS-certified. The outcome of this project gives South Carolina another opportunity to advance uniform crime reporting; its results will be shared with other uniform crime reporting programs. As the FBI moves all states to NIBRS-compatibility in 2021, South Carolina leads the way in data integrity, while also improving the lives of domestic violence survivors.

The South Carolina Governor's Domestic Violence Task Force identified the SCIBRS as the primary source for domestic violence data. It is of paramount importance that the integrity of the SCIBRS data be ensured. Quality data are best placed to guide policy and the distribution of resources for criminal justice agencies, government institutions, and nonprofit organizations in their mission to aid domestic violence victims. An essential element of this endeavor must be the assurance that the quality of SCIBRS reporting is consistent across jurisdictions. Because limited resources render it infeasible for the South Carolina Law Enforcement Division (SLED) to visit all 275 reporting agencies, the South Carolina Statistical Analysis Center (SC SAC) designed a three-phase research project to aid SLED in its effort to ensure integrity of the SCIBRS data.

SCIBRS, which is managed by SLED, contains information about crime in South Carolina: it results from South Carolina's approximately 275 law enforcement agencies reporting information about victim, offense, offender, and arrestee (if applicable) for all criminal incidents known to police. The result is a rich set of crime data unparalleled in criminal justice data collection. SLED provides support to law enforcement agencies through auditing, training, and guidance on coding individual incidents. SLED stores every incident submitted by law enforcement agencies on a state repository and submits those same incidents to the FBI.

As of 2015, 14 of the SCIBRS offense categories require reporting the victim-to-offender relationship, which can include the following five intimate relationships: spouse, ex-spouse, common-law spouse, (ex-)boy/girlfriend, same-sex relationship. Information about offenses, coupled with victim-to-offender intimate relationships, means that the SCIBRS can be used to study domestic violence. SCIBRS itself does not have a domestic violence indicator; rather, the category of domestic violence is constructed from available victim-to-offender relationships and offenses required by the SCIBRS to record those relationships. As a NIBRS-certified system, its crimes are categorized by general definitions; the SCIBRS provides a unique opportunity to study domestic violence across jurisdictions—independent of statutory differences.

The SC SAC project focuses on a specific kind of domestic violence: intimate-partner violence (IPV) as constructed from SCIBRS data extracts. The first phase of this project focused on a subset of IPV: intimate-partner violent victimization (IPVV), which the SC SAC defines as any crime in which the victim recorded in the SCIBRS is an intimate partner of the offender (as defined above), and the offense recorded in the SCIBRS is a crime included in the FBI's Violent Crime Index (i.e., murder, sexual battery, robbery, aggravated assault). IPVV provides an ideal subset of IPV to develop the modeling methodology because (1) it is the most serious subset of IPV, and so is of deep concern to stakeholders, and (2) the more violent crimes have a better chance of being recorded accurately in police incident data. The aim of Phase 1 was to develop a methodology to detect county outliers for victim counts (recorded on incident reports) of IPVV. Phase 2 extends the methodology by applying it to different definitions of IPV, including the most inclusive category possible: all crimes for which the SCIBRS requires a victim-to-offender relationship to be reported (and thus able to be studied for IPV, in which the relationship is intimate) that occur within the context of domestic violence.

The initial phase of the project focused on identifying counties as likely outliers for IPVV. During this phase, 11 counties were classified as outliers. Stonewall Analytics established socio-economic profiles for all counties in South Carolina[1] and combined these profiles with SCIBRS data. A supervised machine learning model (random forest model) was used to establish predicted values for IPVV counts at the county level (by year, from 2011–2015) across all South Carolina counties. Any county where actual counts fell outside one standard deviation of the predicted count for three or more years was flagged as an outlier.

Phase 2 extends this effort by applying the same machine learning methods used in Phase 1 on varying definitions of IPV—as they are available in the data and as subsets of interest identified by the victim community. Five different models of IPV were established, using different categories of variables constructed from SCIBRS data. The first category covers all SCIBRS IPV crimes, which includes crimes from the other categories, in addition to kidnapping/abduction, fondling, and sexual exposure. The second category is the same IPVV used in Phase 1: murder, sexual battery, robbery, and aggravated assault. The remaining three categories are for aggravated assault, simple assault, and intimidation, which cover crimes that range from more serious (but less numerous) to less serious (but more numerous). Each IPV definition gives a different view of IPV, with no definition being intrinsically better or worse. This phase evaluates the conditions under which a county is deemed an outlier when the dependent variable (i.e., the definition of IPV being used) is changed. This approach is a form of sensitivity analysis.

[1] Using data from the American Community Survey.

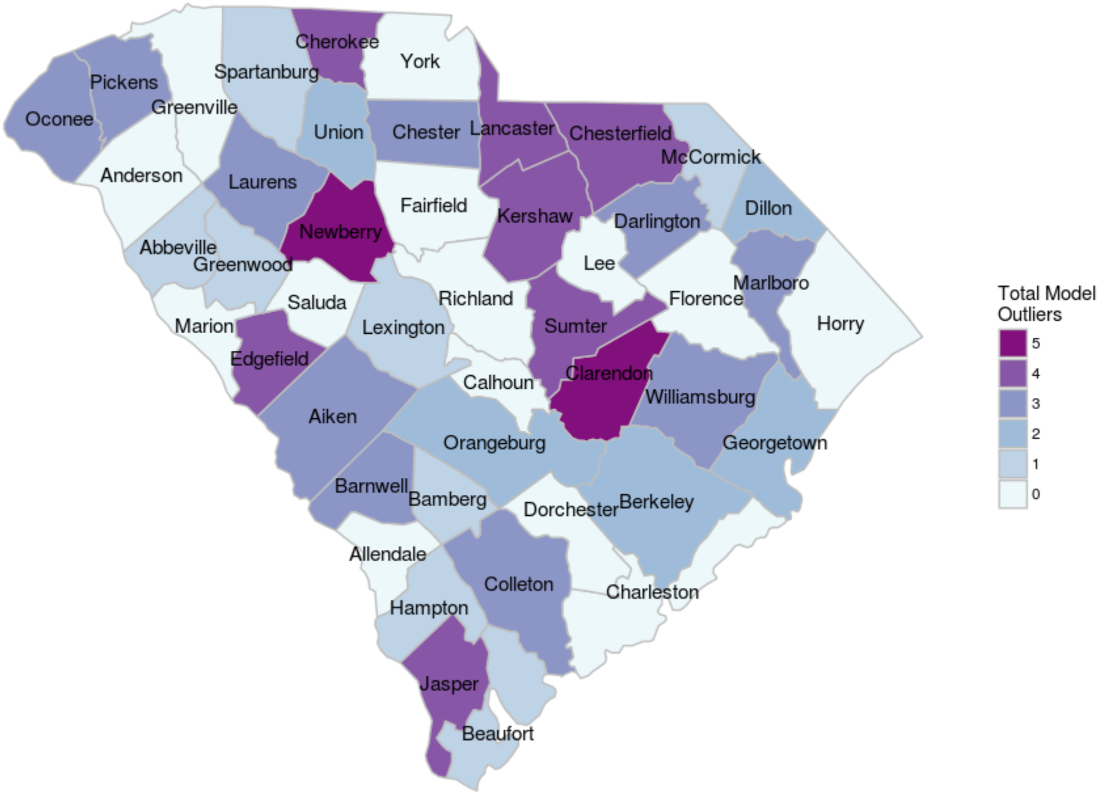
County-Level Summary of Outliers[2]

Total Model Outliers	County	All SCIBRS IPV Crimes (20 Outliers)	Violent Crimes (11 Outliers)	Aggravated Assault (17 Outliers)	Simple Assault (14 Outliers)	Intimidation (24 Outliers)
1	Abbeville					□
3	Aiken	□		□	-	□
0	Allendale					
0	Anderson	-	-	-	-	
1	Bamberg					□
3	Barnwell			□	□	□
1	Beaufort		-	-	□	
2	Berkeley	-	□	□	-	-
0	Calhoun					
0	Charleston	-	-	-	-	-
4	Cherokee	□	□	□		□
3	Chester	□		□		□
4	Chesterfield	□	□	□		□
5	Clarendon	□	□	□	□	□
3	Colleton	□		□	□	
3	Darlington		□		□	□
2	Dillon	□				□
0	Dorchester	-			-	-
4	Edgefield	□	□	□		□
0	Fairfield					
0	Florence	-			-	-
2	Georgetown	□				□
0	Greenville	-	-	-	-	-
1	Greenwood				□	-
1	Hampton	□				
0	Horry	-	-	-	-	-
4	Jasper	□	□	□		□
4	Kershaw	□		□	□	□
4	Lancaster	□		□	□	□
3	Laurens	□	□		□	-
0	Lee					
1	Lexington	-	-	-	-	□
3	Marion	□		□		□
1	Marlboro	□				
0	McCormick					
5	Newberry	□	□	□	□	□
3	Oconee	□	□			□
2	Orangeburg				□	□
3	Pickens	□			□	□
0	Richland	-	-	-	-	-
0	Saluda					
1	Spartanburg	-	-	-	-	□
4	Sumter		□	□	□	□
2	Union			□	□	
3	Williamsburg	□		□		□
0	York	-	-	-	-	-

[2] A □ symbol indicates that the county is classified as an outlier. A - symbol indicates that the county was excluded from that model's analysis. Although the magnitude of difference was useful in Phase 1 to triage counties, such an interpretation in this phase is not warranted. Counties should be viewed in a binary way (outlier or not an outlier).

Given the dependent variable (IPV definition) being examined, there is significant variation in counties classified as outliers. For comparison of consistency in the models, only two counties were classified as outliers across all five models. The figure below provides an overview of the results of the sensitivity analysis, demonstrating the number of times a county was identified as an outlier. There are several geographic clusters of counties classified with an equal count of outliers. For instance, Lancaster, Chesterfield, Kershaw, and Sumter all exhibit a count of four for outlier status. Aiken and Barnwell were outliers in three of the five models. Additionally, Orangeburg, Berkeley, and Georgetown were outliers in two of five models. Some counties that are outliers are also geographically isolated. Newberry and Jasper fall in this category, where the counties surrounding them are either not outliers, or they have a much smaller number of outliers.

Outlier Results from Sensitivity Analysis



“Total Model Outliers” refers to the number of times a county met outlier criteria depending on varied definitions of Intimate-Partner Violence.

Source: Stonewall Analytics

CONTENTS

INTRODUCTION AND BACKGROUND

In 1987, South Carolina served as the pilot state for the FBI's NIBRS—which is the National Incident-Based Reporting System (NIBRS). In 1991, the South Carolina Incident-Based Reporting System (SCIBRS) was the first uniform crime reporting program in the nation to become NIBRS-certified. The outcome of this project gives South Carolina another opportunity to advance uniform crime reporting; its results will be shared with other uniform crime reporting programs. As the FBI moves all states to NIBRS-compatibility in 2021, South Carolina leads the way in data integrity, while also improving the lives of domestic violence survivors.

The South Carolina Governor's Domestic Violence Task Force identified the SCIBRS as the primary source for domestic violence data. As such, it is of paramount importance that the integrity of the SCIBRS data be ensured. Quality data are best placed to guide policy and the distribution of resources for criminal justice agencies, government institutions, and nonprofit organizations in their mission to aid domestic violence victims. An essential element of this endeavor must be the assurance that the quality of SCIBRS reporting is consistent across jurisdictions. Because limited resources render it infeasible for the South Carolina Law Enforcement Division (SLED) to visit all 275 reporting agencies, the South Carolina Statistical Analysis Center (SC SAC) designed a three-phase research project to aid SLED in its effort to ensure integrity of the SCIBRS data.

SCIBRS, which is managed by SLED, contains information about crime in South Carolina: it results from South Carolina's approximately 275 law enforcement agencies reporting information about victim, offense, offender, and arrestee (if applicable) for all criminal incidents known to police. The result is a rich set of crime data unparalleled in criminal justice data collection. SLED provides support to law enforcement agencies through auditing, training, and guidance on coding individual incidents. SLED also stores every incident submitted by the law enforcement agencies on a state repository and submits those same incidents to the FBI.

As of 2015, 14 of the SCIBRS offense categories require reporting the victim-to-offender relationship, which can include the following five intimate relationships: spouse, ex-spouse, common-law spouse, (ex-)boy/girlfriend, same-sex relationship. Information about offenses, coupled with victim-to-offender intimate relationships, means that the SCIBRS can be used to study domestic violence. SCIBRS itself does not have a domestic violence indicator; rather, the category of domestic violence must be constructed from available victim-to-offender relationships and offenses required by the SCIBRS to record those relationships. Further, as a NIBRS-certified system, its crimes are categorized by general definitions; thus, the SCIBRS provides a unique opportunity to study domestic violence across jurisdictions—independent of statutory differences.

The SC SAC project focuses on a specific kind of domestic violence: intimate-partner violence (IPV) as constructed from SCIBRS data extracts. The first phase of this project focused on a subset of IPV: intimate-partner violent victimization (IPVV), which the SC SAC defines as any crime in which the victim recorded in the SCIBRS is an intimate partner of the offender (as defined above), and the offense recorded in the SCIBRS is a crime included in the FBI's Violent Crime Index (i.e., murder, sexual battery, robbery, and aggravated assault). IPVV provides an ideal subset of IPV to develop the modeling methodology because (1) it is the most serious subset of IPV, and so is of deep concern to stakeholders, and (2) the more violent crimes have a better chance of being recorded accurately in police incident data. The aim of Phase 1 was to develop a methodology to detect county outliers for victim counts (as recorded on individual incident reports) of IPVV. Phase 2 extends the methodology by applying it to different definitions of IPV, including the most inclusive category possible: all crimes for which the SCIBRS requires a victim-to-offender relationship to be reported (and thus able to be studied for IPV, in which the relationship is intimate) that occur within the context of domestic violence.

The initial phase of the project focused on identifying counties as likely outliers for IPVV. During this phase, 11 counties were classified as outliers. Stonewall Analytics established socio-economic profiles for all counties in South Carolina[3] and combined these profiles with SCIBRS data. A supervised machine learning model (random forest model) was used to establish predicted values for IPVV counts at the county level (by year, from 2011–2015) across all South Carolina counties. Any county where actual counts fell outside one standard deviation of the predicted count for three or more years was flagged as an outlier.

Phase 2 extends this effort by applying the same machine learning methods used in Phase 1 on variations of the definitions of the crimes—as they are available in the data and as subsets of interest identified by the victim community. This report documents the second phase of the two-phase project, which established a methodology by which uniform crime reporting programs could readily identify counties that may be more likely to have data quality issues. While the first phase encompassed a methodology that identified potential county outliers, this phase involves sensitivity analysis of determining outlier counties by varying the operational definition for the dependent variable in the model.

[3] Using data from the American Community Survey.

An outlier is an observation that deviates from other observations in the sample from which it exists (Hodge & Austin, 2004). Outliers in data are problematic, as they can misinform decisions made with data, draw into question the validity of the entire data, and can bias results of quantitative analyses. Outliers are caused by mechanical faults, fraudulent behavior, human error, instrument error, or can be naturally occurring in the population (Hodge & Austin, 2004). Any findings based on analyses conducted with data containing outliers can have skewed coefficient estimation, prediction, and hypothesis tests (Jensen & Ramirez, 2015). The intent of this project is to identify county outliers without considering the reason or cause.

RELEVANT LITERATURE

The literature review provides an overview of intimate-partner violence, with a focus on prior literature that uses quantitative-based models to analyze associations with IPV and community-based measures. This literature review is not an exhaustive collection of prior work, but rather a robust primer to orient the reader to prior work and to set the stage for the methods and findings in this project.

INTIMATE-PARTNER VIOLENCE ANALYSES

IPV is a public health epidemic that crosses state and international borders (Abramsky, et al., 2011). Victims of IPV often suffer from a large magnitude of mental disorders as a result of their ordeal (Golding, 1999). While existing theories for intimate-partner violence attempt to provide a conceptual understanding of the magnitude of its impact to its victims, very few of the theories are backed by explanatory value from quantitative-based studies (Bell & Naugle, 2008). IPV has several subtypes, which include physical, sexual, psychological, and stalking; all subtypes negatively affect its victims to varying degrees (Basile, Aria, Desai, & Thompson, 2004). Men and women both are exposed to all subtypes of IPV; however, women tend to experience greater rates of IPV as compared to men (Carbone-López, Kruttschnitt, & Macmillan, 2006). Prior research suggests that victims of IPV tend to experience greater rates of physical or sexual violence as compared to nonphysical subtypes of IPV (Coker, Smith, McKeown, & King, 2000). IPV is associated with community-level factors, such as neighborhood-level income (Bonomi, Trabert, Anderson, Kernic, & Holt, 2014; Capaldi, Knoble, Shortt, & Kim, 2012) and individual-level factors such as self-rated health status (Decker, et al., 2014), and marital status (Li, et al., 2010). Poverty is a factor that is significantly associated with IPV (Jewkes, 2002).

OBJECTIVE OF THIS WORK

The objective of this work is to assess how varying dependent variables in a statistical model for SCIBRS data affect counties identified as an outlier. This is the second phase of a project that evaluates the integrity of SCIBRS data. This report, in supplement with the first-phase report, provides the framework for monitoring the quality and integrity of the data going forward–in South Carolina and potentially in other states. Prior analyses have applied quantitative measures to predict repeat incidents of IPV, but no prior studies to our knowledge were aimed at addressing the quality or integrity of the data itself (Roehl, O'Sullivan, Webster, & Campbell, 2005).

METHODOLOGICAL APPROACH AND DATA

This section describes the overall methodological approach, the data used, and the quantitative models employed in this analysis.

DATA

This project used SCIBRS data and contextual data gathered from the United States Census Bureau (American Community Survey). National, county-level arrest data for 2011 from the Inter-University Consortium for Political and Social Research (ICPSR) were used for training, testing, and validation of the statistical models.

SCIBRS DATA

The SCIBRS data included variables related to all crimes of SCIBRS intimate-partner violence (IPV), reports of IPVV, cases of aggravated assault, cases of simple assault, and cases of intimidation for all 46 counties in South Carolina from 2011–2015. These variables serve as the outcome (or dependent) variables for the statistical models. Five different models of IPV were established, using five different categories of variables constructed from SCIBRS data. The first and most inclusive category covers all SCIBRS IPV crimes, which includes crimes from all of the other categories, in addition to kidnapping/abduction, fondling, and sexual exposure. The second category concerns violent crimes and is the same IPVV used in Phase 1: murder, sexual battery, robbery, and aggravated assault. The remaining three categories are for aggravated assault, simple assault, and intimidation, which cover crimes that range from more serious (but less numerous) to less serious (but more numerous). Each IPV definition gives a different view of IPV, with no definition being intrinsically better or worse.

The unit of analysis for the SCIBRS data was county-year (i.e., each variable was provided for the specific year, ranging from 2011–2015). Each model was run for the given definition and year, ranging from 2011–2015. All variables corresponding to crimes were reported as raw numbers, rather than normalized rates. As the distribution of the dependent variables resembled a Poisson distribution (in which the mean and standard deviation are equal), a relatively low proportion of the data were exhibited in the right-hand tail of the data distributions for all variables. Consistent with Phase 1, a model cutoff was employed to improve the accuracy of the models–the cutoff was established at the 75th percentile for each of the dependent variables. During model testing and refinement, the cutoff significantly improved model accuracy, but it did result in occasionally excluding some county-years from analysis.

CONTEXTUAL DATA

Contextual data involving the county-level data from the American Community Survey (ACS) were obtained through the US Census Bureau. The ACS contains county-level information on gender, age, education, race, ethnicity, income, labor force status, and residence, among others (American Community Survey: Information Guide, 2013). The ACS is published in 1-year, 5-year, and supplemental estimates. The typical cutoff for a 1-year estimate is a population area of at least 65,000 individuals. Since some counties in South Carolina have a population less than 65,000, use of the 5-year estimates were used. The ACS has an annual data release for the five-year estimates, which include data for the previous five years.

To maintain consistency with the methodology and independent variables used in Phase 1, the following independent variables were included in the analysis.

- Median household income in the past 12 months (inflation-adjusted dollars) [**median_hh_inc]**
- Average household size of occupied housing units by tenure **[avg_house_size]**
- Gini Index **[gini_index]**
- Hispanic or Latino origin [**hispanic**]
- Poverty status in the past 12 months of families by household type by tenure **[poverty_status_population]**
- Presence of own children under 18 years of age for females 20 to 64 years **[fem_20to64_wchild]**
- Race (White alone) **[white]**
- Race (Black or African American alone) **[black]**
- Receipt of Supplemental Security Income (SSI) **[receiving ssi]**
- Sex (gender) **[male] [female]**
- Total population **[total_pop]**
- Bachelor's degree or higher **[educ_bach]**
- Drive to work **[drive]**
- Women who had a birth in the past 12 months **[birth_last_yr]**
- Total working population **[tot_working_pop]**
- Male worked in the past 12 months (16–64 years) **[male_worked]**
- Female worked in the past 12 months (16–64 years) **[female_worked]**
- Marital Status **[married]**
- Divorced **[divorced]**

While the variables above are straightforward in their interpretation, the Gini Index is an economical measure. This index provides a measure of wealth inequality across a community and ranges between zero and one. An index of zero represents complete equality, whereas an index of one represents maximum inequality.

TRAINING/TESTING DATA

To train and test the statistical models, it became necessary to incorporate data that could be used as a proxy for the dependent variables in each model. A proxy variable served as a bridge to the training and testing datasets. Through exploratory data analysis, proxy variables were selected for each of the five dependent variables used in the model.[4]

SUPERVISED MACHINE LEARNING MODELS

The overall approach with machine learning is to develop a robust model using a training, or practice, dataset. The model developed with the training dataset is then evaluated with a testing dataset. In most cases, a dataset is statistically split at random into a testing dataset and training dataset. For this project, the machine learning model was developed with the training dataset, evaluated with the testing dataset,[5] and then applied to the SCIBRS dataset. As in Phase 1 of the project, a random forest model was used as the basis of the statistical analysis. Random forest models tend to avoid model overfitting and are easy to interpret.

CRITERION FOR DETERMINING OUTLIER COUNTIES

In addition to running the machine learning model on the county-level data for 2011, 2012, 2013, 2014, and 2015, a criterion was needed to classify a county as a potential outlier: in one year the county might fall within the expected range of the model, but in other years it might not. A county was flagged as a potential outlier if three or more of its county-years fell outside of one standard deviation of the predicted result, either higher or lower than expected.

[4] Table 1 in the Results section contains information on the proxy variables, and the summary statistics for each variable.

[5] Training and testing were done using the ICPSR data.

RESULTS

This section presents the descriptive statistics of the SCIBRS data, the contextual data, and the training and testing data for the random forest model, along with findings of potential outlier counties.

COMPARISON OF SCIBRS AND ICPSR TRAINING/TESTING DATA

The training/testing data acquired from ICPSR for all counties throughout the United States in 2011 would be of little use if the data did not share similar behavior and correspond well with the SCIBRS data. Through exploratory data analysis, the following variables were selected as proxies for the five dependent variables in the model. Table 1 provides an overview of summary statistics comparing the training/testing and the SCIBRS data for the outcome variables of interest.[6]

Table 1. Comparison of SCIBRS and South Carolina Proxy Variables

Variable	Variable Class	Min	25th Percentile	Median	Mean (SD)	75th Percentile	Maximum
All SCIBRS IPV Crimes	SCIBRS	31	212	442	747 (774)	1,052	2,964
Violent Crimes – Reports	Proxy	38	269	650	1,063 (1150)	1,456	5,273
Violent Crimes	SCIBRS	7	31	59	119 (145)	141	679
Aggravated Assaults – Arrests	Proxy	8	41	69	138 (156)	173	699
Aggravated Assault	SCIBRS	6	28	51	111 (138)	133	641
Aggravated Assaults – Arrests	Proxy	8	41	69	138 (156)	173	699
Simple Assault	SCIBRS	21	167	318	535 (566)	777	2,163
Aggravated Assaults – Reports	Proxy	27	122	278	461 (520)	563	2,574
Intimidation	SCIBRS	0	18	48	90 (105)	147	423
Total Violent Crime – Arrests	Proxy	0	42	73	112 (116)	141	478

Source: Stonewall Analytics. SD = Standard deviation. Note: Some statistics presented above were rounded.

[6] For ease of interpretation, the column heading identified as Variable Class references whether the variable was constructed from SCIBRS data, or is a proxy variable.

MODEL RESULTS

Five random forest models were run for each county annually from 2011–2015. Table 2 presents the proportion of variability of the dependent variable accounted for in each model.

Table 2. Proportion of Variability Explained by Model

Model	% Variation Explained
All SCIBRS IPV Crimes	82.3
Violent Crimes	49.5
Aggravated Assault	51.9
Simple Assault	73.6
Intimidation	59.1

Source: Stonewall Analytics

The supervised machine learning model accounted for 82.3% of the variability in the dependent variable for the model of all SCIBRS IPV crimes for the training/testing dataset. The proportion of variation in the model for violent crimes was 49.5%, which represents the least amount of explanation across all five models.[7] Figure 1 presents the most influential variables for each of the five models used in development of the machine learning models with the training data.[8]

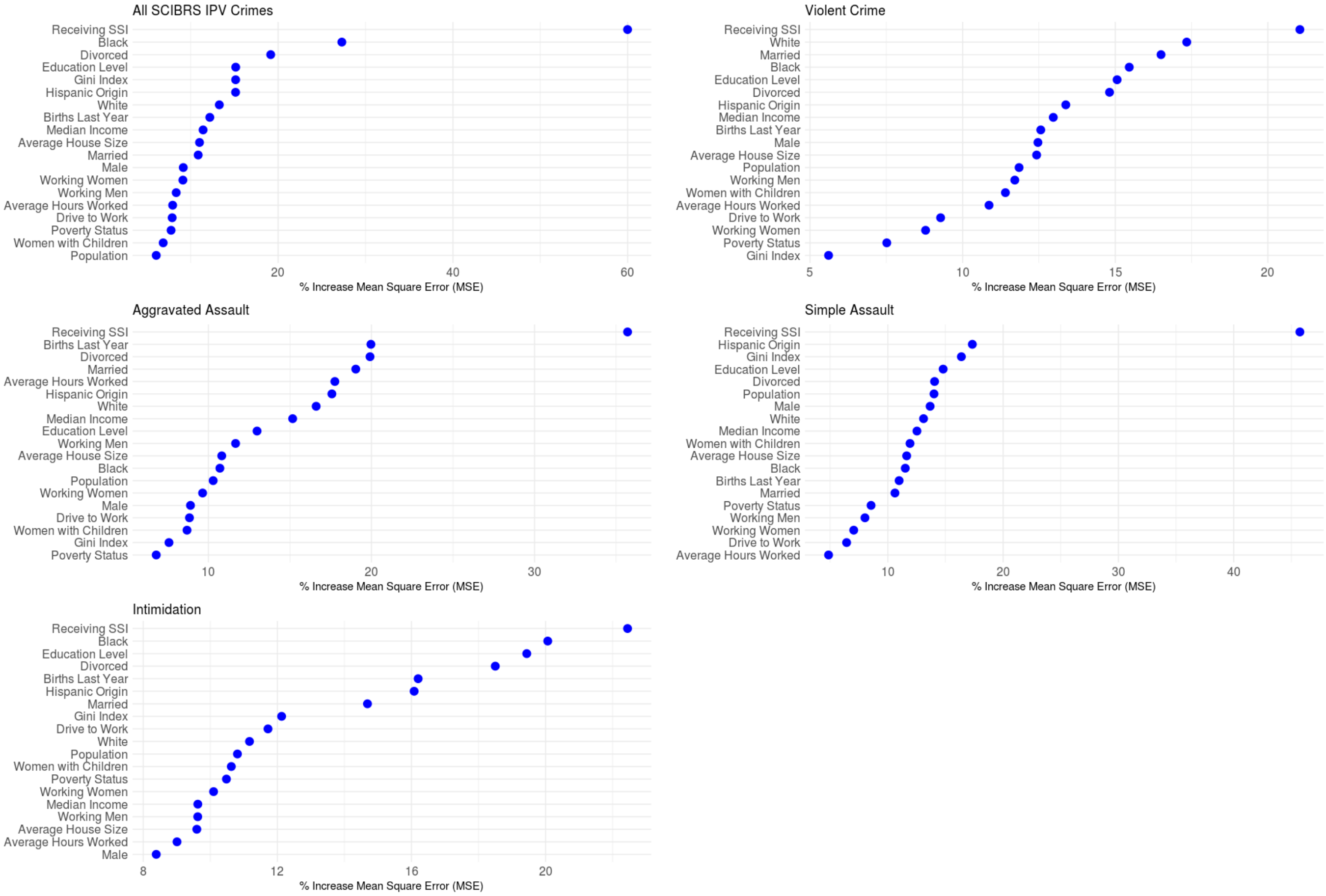
Across all five models, the most important (and consistent) variables include the proportion of county residents receiving SSI, race and ethnicity, and marriage. Aside from these variables, there is variation in the order of variables most important based upon the model. For instance, while the Gini Index is the fifth most important variable when evaluating all SCIBRS IPV crimes (top-left plot in Figure 1), this variable is least important for violent crime (top-right plot in Figure 1). Although the scale of the x-axis has no interpretable value, it is important to note that the x-axis in each plot is free-scaled and not consistent across the plots.

[7] This percentage is slightly higher than in the Phase 1, as the training and testing data were randomly selected. As such, results will vary slightly.

[8] These variables were determined most important through the difference in the calculation of the percent increase in mean square error between the specific variable and random permutated variables.

Figure 1. Importance of Variables

Source: Stonewall Analytics



POTENTIAL COUNTY OUTLIERS

In the first phase of the project, 11 counties were identified as outliers. With five differing dependent variables in this phase of the analysis, there is a wide range of counties identified as outliers. For all SCIBRS IPV crimes, there were 20 outliers. With regard to violent crime, 11 counties were identified as outliers (this is the same dependent variable used in the first phase of the project). For aggravated assault, 17 counties were classified as outliers. When simple assault was the dependent variable, there were 14 outliers. Lastly, there were 24 outliers when the dependent variable was intimidation. Table 3 lists the county along with a symbol (□) to identify whether the county is considered an outlier. Those counties without the symbol are not considered an outlier. Counties with a (-) symbol indicate they were excluded from the particular model, as they exceeded the 75th percentile cutoff criterion.

Table 3. County-Level Summary of Outliers[9]

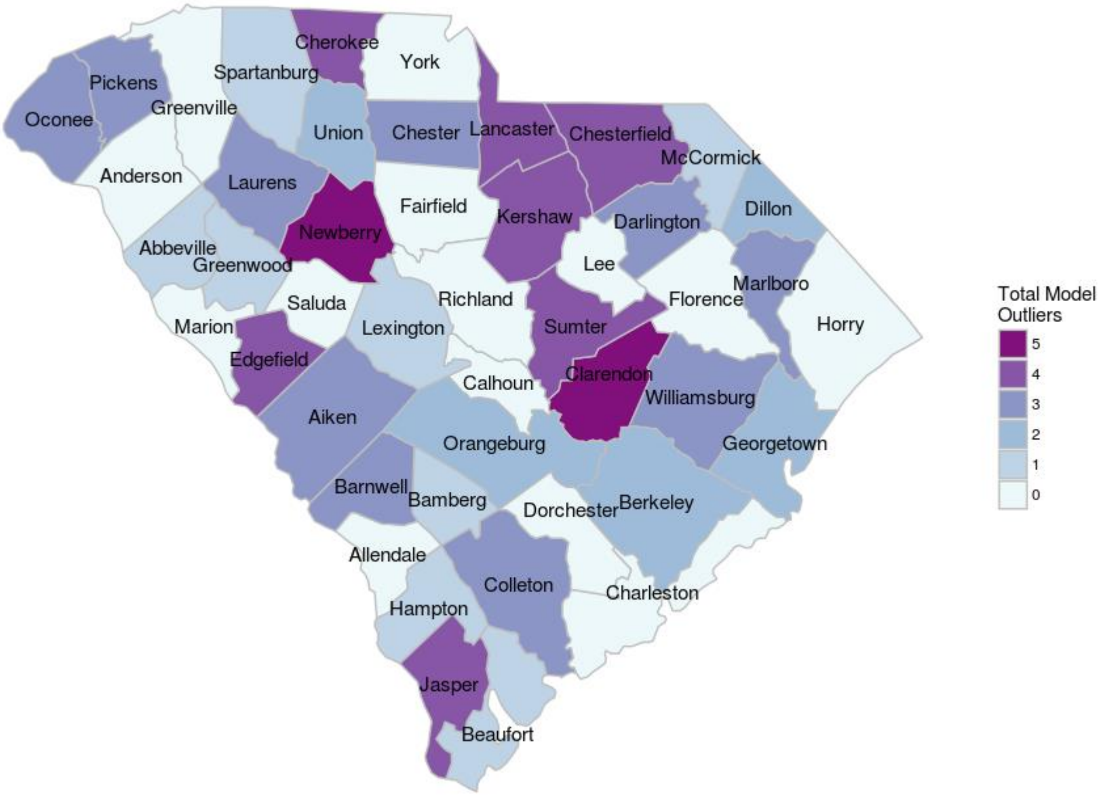
Total Model Outliers	County	All SCIBRS IPV Crimes (20 Outliers)	Violent Crimes (11 Outliers)	Aggravated Assault (17 Outliers)	Simple Assault (14 Outliers)	Intimidation (24 Outliers)
1	Abbeville					□
3	Aiken	□		□	-	□
0	Allendale					
0	Anderson	-	-	-	-	
1	Bamberg					□
3	Barnwell			□	□	□
1	Beaufort		-	-	□	
2	Berkeley	-	□	□	-	-
0	Calhoun					
0	Charleston	-	-	-	-	-
4	Cherokee	□	□	□		□
3	Chester	□		□		□
4	Chesterfield	□	□	□		□
5	Clarendon	□	□	□	□	□
3	Colleton	□		□	□	
3	Darlington		□		□	□
2	Dillon	□				□
0	Dorchester	-			-	-
4	Edgefield	□	□	□		□
0	Fairfield					
0	Florence	-			-	-
2	Georgetown	□				□
0	Greenville	-	-	-	-	-
1	Greenwood				□	-
1	Hampton	□				
0	Horry	-	-	-	-	-
4	Jasper	□	□	□		□
4	Kershaw	□		□	□	□
4	Lancaster	□		□	□	□
3	Laurens	□	□		□	-
0	Lee					
1	Lexington	-	-	-	-	□
3	Marion	□		□		□
1	Marlboro	□				
0	McCormick					
5	Newberry	□	□	□	□	□
3	Oconee	□	□			□
2	Orangeburg				□	□
3	Pickens	□			□	□
0	Richland	-	-	-	-	-
0	Saluda					
1	Spartanburg	-	-	-	-	□
4	Sumter		□	□	□	□
2	Union			□	□	
3	Williamsburg	□		□		□
0	York	-	-	-	-	-

Source: Stonewall Analytics

[9] A □ symbol indicates the county is classified as an outlier. A - symbol indicates the county was excluded from that model's analysis. Although the magnitude of difference was useful in Phase 1 to triage counties, such an interpretation in this phase is not warranted. Counties should be viewed in a binary way (outlier or not an outlier).

For all 46 counties in South Carolina, two counties were considered an outlier by all five models. Seven counties were considered an outlier with four models, and 10 counties were considered an outlier by three models. Figure 2 contains a map of counties in South Carolina by outlier status from the five models. Across the models, the number of counties classified as an outlier ranged from 11 to 24.

Figure 2. Outlier Results from Sensitivity Analysis



“Total Model Outliers” refers to the number of times a county met outlier criteria depending on varied definitions of Intimate Partner Violence.

Source: Stonewall Analytics

There are several geographic clusters of counties classified with an equal count of outliers. For instance, Lancaster, Chesterfield, Kershaw, and Sumter all exhibit a count of four for outlier status. Aiken and Barnwell were outliers in three of the five models. Additionally, Orangeburg, Berkeley, and Georgetown were outliers in two of five models. Some counties that are outliers are also geographically isolated. Newberry and Jasper fall in this category, where the counties surrounding them are either not outliers, or they exhibit a smaller number of outliers.

DISCUSSION

This report documents the second phase of a project that assesses the integrity and quality of SCIBRS data at the county level. Performing sensitivity analysis is critical to any robust analysis. As demonstrated in this report, there is significant variation in counties deemed an outlier based upon the dependent variable selected. Whereas 11 counties were classified as outliers in the first phase of the project, with the addition of new dependent variables, counties classified as an outlier ranged between 11 and 24. When counties classified as outliers were examined geographically, clusters of outliers were also evident. There were small groupings of outlier counties with neighboring borders that demonstrated the same count of outliers. In a smaller number of cases, certain counties classified as an outlier were also geographically isolated from other outlier counties.

When accounting for the results of this phase of the project in terms of next steps, a logical next step in the analysis would be to focus on counties classified as an outlier by multiple models to then assess at the ground-level the integrity of the reporting. A supervised machine learning model was used to predict various crime elements in South Carolina from 2011–2015. The model was then applied to contextual data to create predicted values at the county level, by calendar year. Where large differences between reported and predicted counts existed for each year, over three or more years, counties were flagged as potential outliers by denoting whether the reported values were higher or lower than the predicted values.

Like any study, this research has both strengths and weaknesses. Perhaps the most significant weakness is the model's inability to evaluate counties with a higher volume of cases/counts. If training/testing data were available with similar incident reporting volume, these counties could be included in future analyses. This study is limited in the contextual data presented here, as prior studies have assessed health-related measures. Further, since this study's variables do not capture health-related aspects, the study likely suffers from some omitted variable bias (Decker, et al., 2014; Jiang, DeBare, Shea, & Viner-Brown, 2017). Another limitation is not controlling for the factors associated with county-level policing practices and policies (Leisenring, 2008; Maxwell, Garner, & Fagan, 2002). Additionally, the machine learning model used to train and test data at the county level for aggravated assault arrests in 2011 may not hold constant through the remaining years (2012–2015) in the timeline of the analysis.

To the knowledge of this study's authors, this is the first project aimed at assessing the quality and integrity of county-level incident reporting of IPV in any state. This framework and methodology have the ability to scale to other states and time periods, and it can be used to assess the quality and integrity of incident reporting going forward.

REFERENCES

Abramsky, T., Watts, C., Garcia-Moreno, C., Devries, K., Kiss, L., Ellsberg, M., . . . Heise, L. (2011). What factors are associated with recent intimate partner violence? Findings from the WHO multi-country study on women's health and domestic violence. *BMC Public Health, 11*(1), 109.

American Community Survey: Information Guide. (2013). US Census Bureau.

Basile, K. C., Aria, I., Desai, S. & Thompson, M. P. (2004). The differential association of intimate partner physical, sexual, psychological, and stalking violence and posttraumatic stress symptoms in a nationally representative sample of women. *Journal of Traumatic Stress, 17*(5), 413–421.

Bell, K., & Naugle, A. (2008). Intimate partner violence theoretical considerations: Moving towards a contextual framework. *Clinical Psychology Review, 28(7)*, 1096–1107.

Bonomi, A., Trabert, B., Anderson, M., Kernic, M., & Holt, V. (2014). Intimate Partner Violence and Neighborhood Income: A Longitudinal Analysis. *Violence Against Women, 20*(1), 42–58.

Capaldi, D., Knoble, N., Shortt, J., & Kim, H. (2012). A Systematic Review of Risk Factors for Intimate Partner Violence. *Partner Abuse, 3*(2), 231–280.

Carbone-López, K., Kruttschnitt, C., & Macmillan, R. (2006). Patterns of Intimate Partner Violence and Their Associations with Physical Health, Psychological Distress, and Substance Use. *Public Health Reports, 121*(4), 382–392.

Coker, A., Smith, P., McKeown, R., & King, M. (2000). Frequency and correlates of intimate partner violence by type: physical, sexual, and psychological battering. *American Journal of Public Health, 90*(4), 553–559.

Decker, M., Peitzmeier, S., Olumide, A., Acharya, R., Ojengbede, O., Covarrubias, L., . . . Brahmbhatt, H. (2014). Prevalence and Health Impact of Intimate Partner Violence and Non-partner Sexual Violence Among Female Adolescents Aged 15-19 Years in Vulnerable Urban Environments: A Multi-Country Study. *Journal of Adolescent Health, 55*(6), S58–S67.

Golding, J. (1999). Intimate partner violence as a risk factor for mental disorders: A meta-analysis. *Journal of Family Violence, 14*(2), 99–132.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85–126.

Jensen, D., & Ramirez, D. (2015). Shift outliers in linear inference. *Journal of Multivariate Analysis, 136*, 95–107.

Jewkes, R. (2002). Intimate partner violence: Causes and prevention. *The Lancet, 359*(9315), 1423–1429.

Jiang, Y., DeBare, D., Shea, L., & Viner-Brown, S. (2017). Violence against women: Injuries and deaths in Rhode Island. *Rhode Island Medical Journal, 100*(12), 24–24.

Leisenring, A. (2008). Controversies surrounding mandatory arrest policies and the police response to intimate partner violence. *Sociology Compass, 2*(2), 451–466.

Li, Q., Kirby, R., Sigler, R., Hwang, S.-S., LaGory, M., & Goldenberg, R. (2010). A Multilevel Analysis of Individual, Household, and Neighborhood Correlates of Intimate Partner Violence Among Low-Income Pregnant Women in Jefferson County, Alabama. *American Journal of Public Health, 100*(3), 531–539.

Maxwell, C., Garner, J., & Fagan, J. (2002). The preventive effects of arrest on Intimate Partner Violence: research, policy and theory. *Criminology and Public Policy, 2*(1), 51–80.

Roehl, J., O'Sullivan, C., Webster, D., & Campbell, J. (2005). *Intimate Partner Violence Risk Assessment Validation Study, Final Report.* National Institute of Justice, Washington DC.

APPENDIX

R Syntax for Project Analysis and Related Figures[10]

Set up the working directory as appropriate. The following code will evaluate the current working directory. One could place the data files in this default location, or set the working directory with the ‘setwd()’ command. Please type ‘help(setwd)’ within R for more information.

```
setwd(INSERT PATH TO WORKING DIRECTORY HERE)

icpsr <- read.csv(file = 'icpsr.csv', header = TRUE)
socar <- read.csv(file = 'socar.csv', header = TRUE)
crimes <- read.csv(file = 'phase_2_crimes.csv', header = TRUE) # in long format

## 1. all scibrs ipv crimes --> icpsr(violent_crime_report
summary(crimes$count[crimes$measure == 'all_ipv_crimes' & crimes$year == 2011])
sd(crimes$count[crimes$measure == 'all_ipv_crimes' & crimes$year == 2011])

## 2. ipvv crimes --> icpsr(agg_assault_arrest)
summary(crimes$count[crimes$measure == 'ipvv_crimes' & crimes$year == 2011])
sd(crimes$count[crimes$measure == 'ipvv_crimes' & crimes$year == 2011])

## 3. aggravated assault --> icpsr(agg_assault_arrest)
summary(crimes$count[crimes$measure == 'aggravated_assault' & crimes$year == 2011])
sd(crimes$count[crimes$measure == 'aggravated_assault' & crimes$year == 2011])

## 4. simple assault --> icpsr(agg_assault_report)
summary(crimes$count[crimes$measure == 'simple_assault' & crimes$year == 2011])
sd(crimes$count[crimes$measure == 'simple_assault' & crimes$year == 2011])

## 5. intimidation --> icpsr(total_violent_crime_arrest)
summary(crimes$count[crimes$measure == 'intimidation' & crimes$year == 2011])
sd(crimes$count[crimes$measure == 'intimidation' & crimes$year == 2011])

## looking at proxy variables
lapply(icpsr_sc[ , 24:35], summary)
lapply(icpsr_sc[ , 24:35], sd)


library(randomForest)
library(ggplot2)

## 1. violent crimes
icpsr_all_scibrs <- icpsr[icpsr$violent_crime_report <= 1051 , -c(1,2,3,23:33,35)]
set.seed(8675309)

# splitting into 75-25
train <- sample(x = 1:nrow(icpsr_all_scibrs), size = nrow(icpsr_all_scibrs) * 0.75)
icpsr_all_scibrs_test <- icpsr_all_scibrs[-train, 'violent_crime_report']

train_rf_all_scibrs <- randomForest(formula = violent_crime_report ~ . ,
                                    data = icpsr_all_scibrs,
```

[10] This portion of the appendix contains code for the project that was conducted in R. A number of external packages were used in this analysis, so in cases where syntax containing ‘library’ is displayed, it may be required to first install the package using the install.packages() command. For more information, please type the command, ‘help(install.packages)’ within R. It is recommended that the syntax from this section not be directly copy and pasted into R, as this code is no longer in a plain text format. On occasion, error messages may occur with code copied and pasted directly from a word processing document directly in R. It is advisable to type the syntax above in lieu of copying and pasting.

```
                                     subset = train,
                                     mtry = 19,
                                     n.trees = 100,
                                     importance = TRUE)
train_rf_all_scibrs
yhat_rf_all_scibrs <- predict(object = train_rf_all_scibrs, newdata = icpsr_all_scibrs[-
train , ])

mean((yhat_rf_all_scibrs - icpsr_all_scibrs_test)^2) # mean square error
sqrt(mean((yhat_rf_all_scibrs - icpsr_all_scibrs_test)^2)) # standard deviation

out <- as.data.frame(importance(train_rf_all_scibrs))
out2 <- cbind(rownames(out), data.frame(out, row.names = NULL))
names(out2) <- c('var', 'mse', 'purity')
out2
variable_full <- c('Average House Size', 'Gini Index', 'Hispanic Origin', 'Average Hours
                    Worked','Drive to Work', 'Poverty Status', 'Women with Children',
                    'White', 'Black','Receiving SSI', 'Male', 'Population', 'Median
                     Income', 'Education Level','Births Last Year', 'Working Men',
                    'Working Women', 'Married', 'Divorced')
out3 <- cbind(out2, variable_full)
out3

j <- ggplot(data = out3, aes(x = mse, y = reorder(variable_full, mse)))
j <- j + geom_point(color = 'blue', size = 3.5)
j <- j + theme_minimal()
j <- j + labs(x = '% Increase Mean Square Error (MSE)', y = '', title = 'All IPV Crimes')
j <- j + theme(axis.text = element_text(size = 12))
j

## 2. ipvv crimes --> icpsr(agg_assault_arrest)
icpsr_ipvv_crimes <- icpsr[icpsr$agg_assault_arrest <= 141, -c(1,2,3,23,24,26:35)]
set.seed(8675309)
train <- sample(x = 1:nrow(icpsr_ipvv_crimes), size = nrow(icpsr_ipvv_crimes) * 0.75) #
splitting into 75-25
icpsr_ipvv_crimes_test <- icpsr_ipvv_crimes[-train, 'agg_assault_arrest']

train_rf_ipvv_crimes <- randomForest(formula = agg_assault_arrest ~ . ,
                                     data = icpsr_ipvv_crimes,
                                     subset = train,
                                     mtry = 19,
                                     n.trees = 100,
                                     importance = TRUE)
train_rf_ipvv_crimes
yhat_rf_ipvv_crimes <- predict(object = train_rf_ipvv_crimes, newdata =
icpsr_ipvv_crimes[-train , ])

mean((yhat_rf_ipvv_crimes - icpsr_ipvv_crimes_test)^2) # mean square error
sqrt(mean((yhat_rf_ipvv_crimes - icpsr_ipvv_crimes_test)^2)) # standard deviation

summary(icpsr_ipvv_crimes$agg_assault_arrest)
sd(icpsr_ipvv_crimes$agg_assault_arrest)

out <- as.data.frame(importance(train_rf_ipvv_crimes))
out2 <- cbind(rownames(out), data.frame(out, row.names = NULL))
names(out2) <- c('var', 'mse', 'purity')
out2
variable_full <- c('Average House Size', 'Gini Index', 'Hispanic Origin', 'Average Hours
                    Worked','Drive to Work', 'Poverty Status', 'Women with Children',
                    'White', 'Black','Receiving SSI', 'Male', 'Population', 'Median
                     Income', 'Education Level','Births Last Year', 'Working Men',
                    'Working Women', 'Married', 'Divorced')
out3 <- cbind(out2, variable_full)
out3
```

```
j <- ggplot(data = out3, aes(x = mse, y = reorder(variable_full, mse)))
j <- j + geom_point(color = 'blue', size = 3.5)
j <- j + theme_minimal()
j <- j + labs(x = '% Increase Mean Square Error (MSE)', y = '', title = 'IPVV Crimes')
j <- j + theme(axis.text = element_text(size = 12))
j

## 3. aggravated assault --> icpsr(agg_assault_arrest)
icpsr_agg_assault <- icpsr[icpsr$agg_assault_arrest <= 133, -c(1,2,3,23,24,26:35)]
set.seed(8675309)
train <- sample(x = 1:nrow(icpsr_agg_assault), size = nrow(icpsr_agg_assault) * 0.75)
icpsr_agg_assault_test <- icpsr_agg_assault[-train, 'agg_assault_arrest']

train_rf_agg_assault <- randomForest(formula = agg_assault_arrest ~ . ,
                                     data = icpsr_agg_assault,
                                     subset = train,
                                     mtry = 19,
                                     n.trees = 100,
                                     importance = TRUE)
train_rf_agg_assault
yhat_rf_agg_assault <- predict(object = train_rf_agg_assault, newdata =
icpsr_agg_assault[-train , ])

mean((yhat_rf_agg_assault - icpsr_agg_assault_test)^2) # mean square error
sqrt(mean((yhat_rf_agg_assault - icpsr_agg_assault_test)^2)) # standard deviation

summary(icpsr_agg_assault$agg_assault_arrest)
sd(icpsr_agg_assault$agg_assault_arrest)

out <- as.data.frame(importance(train_rf_agg_assault))
out2 <- cbind(rownames(out), data.frame(out, row.names = NULL))
names(out2) <- c('var', 'mse', 'purity')
out2
variable_full <- c('Average House Size', 'Gini Index', 'Hispanic Origin', 'Average Hours
                    Worked','Drive to Work', 'Poverty Status', 'Women with Children',
                    'White', 'Black','Receiving SSI', 'Male', 'Population', 'Median
                     Income', 'Education Level','Births Last Year', 'Working Men',
                    'Working Women', 'Married', 'Divorced')

out3 <- cbind(out2, variable_full)
out3

j <- ggplot(data = out3, aes(x = mse, y = reorder(variable_full, mse)))
j <- j + geom_point(color = 'blue', size = 3.5)
j <- j + theme_minimal()
j <- j + labs(x = '% Increase Mean Square Error (MSE)', y = '', title = 'Aggravated
Assault')
j <- j + theme(axis.text = element_text(size = 12))
j

## 4. simple assault --> icpsr(agg_assault_report)
summary(crimes$count[crimes$measure == 'simple_assault'])
sd(crimes$count[crimes$measure == 'simple_assault'])

icpsr_simple_assault <- icpsr[icpsr$agg_assault_report <= 777 , -c(1,2,3,23:30,32:35)]
set.seed(8675309)
train <- sample(x = 1:nrow(icpsr_simple_assault), size = nrow(icpsr_simple_assault) *
0.75)
icpsr_simple_assault_test <- icpsr_simple_assault[-train, 'agg_assault_report']

library(randomForest)
train_rf_simple_assault <- randomForest(formula = agg_assault_report ~ . ,
                                        data = icpsr_simple_assault,
                                    subset = train,
                                    mtry = 19,
                                    n.trees = 100,
                                    importance = TRUE)
train_rf_simple_assault
yhat_rf_simple_assault <- predict(object = train_rf_simple_assault, newdata =
icpsr_simple_assault[-train , ])

mean((yhat_rf_simple_assault - icpsr_simple_assault_test)^2) # mean square error
sqrt(mean((yhat_rf_simple_assault - icpsr_simple_assault_test)^2)) # standard deviation

summary(icpsr_simple_assault$agg_assault_report)
sd(icpsr_simple_assault$agg_assault_report)

out <- as.data.frame(importance(train_rf_simple_assault))
out2 <- cbind(rownames(out), data.frame(out, row.names = NULL))
names(out2) <- c('var', 'mse', 'purity')
out2
variable_full <- c('Average House Size', 'Gini Index', 'Hispanic Origin', 'Average Hours
                    Worked','Drive to Work', 'Poverty Status', 'Women with Children',
                    'White', 'Black','Receiving SSI', 'Male', 'Population', 'Median
                     Income', 'Education Level','Births Last Year', 'Working Men',
                    'Working Women', 'Married', 'Divorced')

out3 <- cbind(out2, variable_full)
out3

j <- ggplot(data = out3, aes(x = mse, y = reorder(variable_full, mse)))
j <- j + geom_point(color = 'blue', size = 3.5)
j <- j + theme_minimal()
j <- j + labs(x = '% Increase Mean Square Error (MSE)', y = '', title = 'Simple Assault')
j <- j + theme(axis.text = element_text(size = 12))
j

## 5. intimidation --> icpsr(total_violent_crime_arrest)
summary(crimes$count[crimes$measure == 'intimidation'])
sd(crimes$count[crimes$measure == 'intimidation'])

icpsr_intimidation <- icpsr[icpsr$total_violent_crime_arrest <= 147 , -
c(1,2,3,23:27,29:35)]
set.seed(8675309)
train <- sample(x = 1:nrow(icpsr_intimidation), size = nrow(icpsr_intimidation) * 0.75)
icpsr_intimidation_test <- icpsr_intimidation[-train, 'total_violent_crime_arrest']

library(randomForest)
train_rf_intimidation <- randomForest(formula = total_violent_crime_arrest ~ . ,
                                      data = icpsr_intimidation,
                                      subset = train,
                                      mtry = 19,
                                      n.trees = 100,
                                      importance = TRUE)
train_rf_intimidation
yhat_rf_intimidation <- predict(object = train_rf_intimidation, newdata =
icpsr_intimidation[-train , ])

mean((yhat_rf_intimidation - icpsr_intimidation_test)^2) # mean square error
sqrt(mean((yhat_rf_intimidation - icpsr_intimidation_test)^2)) # standard deviation

summary(icpsr_intimidation$total_violent_crime_arrest)
sd(icpsr_intimidation$total_violent_crime_arrest)

out <- as.data.frame(importance(train_rf_intimidation))
out2 <- cbind(rownames(out), data.frame(out, row.names = NULL))
names(out2) <- c('var', 'mse', 'purity')
```

```
out2

variable_full <- c('Average House Size', 'Gini Index', 'Hispanic Origin', 'Average Hours
                    Worked','Drive to Work', 'Poverty Status', 'Women with Children',
                    'White', 'Black','Receiving SSI', 'Male', 'Population', 'Median
                     Income', 'Education Level','Births Last Year', 'Working Men',
                    'Working Women', 'Married', 'Divorced')

out3 <- cbind(out2, variable_full)
out3

j <- ggplot(data = out3, aes(x = mse, y = reorder(variable_full, mse)))
j <- j + geom_point(color = 'blue', size = 3.5)
j <- j + theme_minimal()
j <- j + labs(x = '% Increase Mean Square Error (MSE)', y = '', title = 'Intimidation')
j <- j + theme(axis.text = element_text(size = 12))
j

################# APPLYING TO SC DATA #################
newdata <- merge(x = crimes, y = socar, by.x = c('county', 'year'), by.y =
c('county_name', 'year'), all.x = TRUE)
newdata <- newdata[order(newdata$measure, newdata$year, newdata$county) , ]

results <- data.frame(county = newdata$county,
                      year = newdata$year,
                      measure = newdata$measure,
                      reported = newdata$count,
                      predicted = NA,
                      st_dev = NA,
                      model_cutoff = NA,
                      excluded = NA,
                      outlier = NA)

results <- results[order(results$measure, results$year, results$county) , ]

## 75th percentile from the south carolina data from 2011
results$model_cutoff[results$measure == 'aggravated_assault'] <- 133
results$model_cutoff[results$measure == 'all_ipv_crimes'] <- 1051
results$model_cutoff[results$measure == 'intimidation'] <- 147
results$model_cutoff[results$measure == 'ipvv_crimes'] <- 141
results$model_cutoff[results$measure == 'simple_assault'] <- 777

## standard deviation from the model
results$st_dev[results$measure == 'aggravated_assault'] <- 17.38808
results$st_dev[results$measure == 'all_ipv_crimes'] <- 91.75058
results$st_dev[results$measure == 'intimidation'] <- 17.90890
results$st_dev[results$measure == 'ipvv_crimes'] <- 17.38808
results$st_dev[results$measure == 'simple_assault'] <- 71.36005

results$excluded <- 0
results$excluded[results$reported > results$model_cutoff] <- 1

for (i in 1:dim(newdata)[1]) {
 rf_model = NULL
 if (newdata$measure[i] == 'aggravated_assault')
    {rf_model <- get('train_rf_agg_assault')}
 if (newdata$measure[i] == 'all_ipv_crimes')
    {rf_model <- get('train_rf_all_scibrs')}
 if (newdata$measure[i] == 'intimidation')
    {rf_model <- get('train_rf_intimidation')}
 if (newdata$measure[i] == 'ipvv_crimes')
    {rf_model <- get('train_rf_ipvv_crimes')}
 if (newdata$measure[i] == 'simple_assault')
    {rf_model <- get('train_rf_simple_assault')}
 results$predicted[i] <- predict(object = rf_model,
                                 newdata = newdata[i, 6:24])
    }

## setting up the flag component
results$outlier <- ifelse(test = abs(results$predicted - results$reported) > (1 *
                                 results$st_dev),
                          yes = 1,
                          no = 0)
results$outlier[results$excluded == 1] <- 0

table(results$outlier)
table(outlier = results$outlier, excluded = results$excluded)
prop.table(table(outlier = results$outlier, excluded = results$excluded))

library(dplyr)
summary_results <- data.frame(results %>%
                                group_by(county, measure) %>%
                                  summarise(outlier_status = if_else(condition =
                                   sum(outlier) >= 3,
                                                                  true = 1,
                                                                  false = 0)))
library(tidyr)
summary_results_wide <- summary_results %>% spread(measure, outlier_status)
write.csv(x = summary_results_wide, file = 'output_summary_results.csv', row.names =
FALSE)

dta <- read.csv(file = 'output_summary_results.csv', header = TRUE)
head(dta)

## count of outliers by category
apply(dta[-1],2,sum)

## outlier by county
summary(apply(dta[-1],1,sum))
table(apply(dta[-1],1,sum))
```