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

2011–2015 Intimate-Partner Violent Victimization Counts

Prepared for the South Carolina Statistical Analysis Center

March 2018

Prepared by Stonewall Analytics

Authors:

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

Contributor:

Kristina Lugo, 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

The South Carolina Law Enforcement Division (SLED) manages the South Carolina Incident-Based Reporting System (SCIBRS)—which is National Incident-Based Reporting System (NIBRS)-certified by the Federal Bureau of Investigation (FBI). Basically, the SCIBRS 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.

Currently, fourteen of the SCIBRS offense categories require reporting the victim-to-offender relationship, which (as of 2015) 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. 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 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 is 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 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. In the first phase, the SC SAC's project focuses on a specific kind of domestic violence: intimate-partner violent victimization (IPVV) as recorded in the SCIBRS. The SC SAC defines SCIBRS IPVV 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 recognized by the FBI as 'violent' (i.e., murder, sexual battery, robbery, aggravated assault).

During this (current) phase, two methods are used to identify counties with SCIBRS IPVV that is outside expectation (i.e., either higher or lower than expected). These counties are deemed "outliers" and will guide SLED during the second phase of the project in its drilldown to an agency-level review. The first method for outlier identification is st: the SC SAC uses SCIBRS IPVV rates of victims per 10,000 and descriptive statistics to identify outlier counties. The second method is more complex: Stonewall Analytics (having been contracted by the SC SAC) uses advanced statistical techniques to analyze IPVV counts of victims along with socioeconomic contextual data. As SLED reviews agencies, the result will be a real-world comparison of methods to see which was more successful in pointing toward data issues. During the third phase, reporting agencies in South Carolina will benefit from SLED's resulting educational outreach. This report documents the analysis and findings of Stonewall Analytics.

Using data from the American Community Survey, Stonewall Analytics established socio-economic profiles for all counties in South Carolina, which it then combined with SCIBRS data. This became the foundation for the application of a supervised machine learning model (random forest model) to predict the expected values for IPVV counts of victims for all counties. After testing and refining the model, predicted values for IPVV counts of victims at the county level (by year, from 2011–2015) were established 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.

The model identified the following 11 counties as outliers.

- Berkeley
- Cherokee
- Chesterfield
- Clarendon
- Colleton
- Edgefield
- Florence
- Jasper
- Newberry
- Pickens
- Sumter

In 1987, South Carolina served as the pilot state for the FBI's NIBRS. In 1991, the SCIBRS was the first uniform crime reporting program in the nation to become NIBRS-certified. The project at-hand gives South Carolina another opportunity to advance uniform crime reporting. The outcome of the statistically-guided agency review 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.

CONTENTS

INTRODUCTION AND BACKGROUND

The South Carolina Law Enforcement Division (SLED) manages the South Carolina Incident-Based Reporting System (SCIBRS)—which is National Incident-Based Reporting System (NIBRS)-certified by the Federal Bureau of Investigation (FBI). Basically, the SCIBRS 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.

Currently, fourteen of the SCIBRS offense categories require reporting the victim-to-offender relationship, which (as of 2015) 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. 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 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 is 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 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. In the first phase, the SC SAC's project focuses on a specific kind of domestic violence: intimate-partner violent victimization (IPVV) as recorded in the SCIBRS. The SC SAC defines SCIBRS IPVV 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 recognized by the FBI as 'violent' (i.e., murder, sexual battery, robbery, aggravated assault).

During this (current) phase, two methods are used to identify counties with SCIBRS IPVV that is outside expectation (i.e., either higher or lower than expected). These counties are deemed "outliers" and will guide SLED during the second phase of the project in its drilldown to an agency-level review. The first method for outlier identification is relatively simple: the SC SAC uses SCIBRS IPVV rates of victims per 10,000 and descriptive statistics to identify outlier counties. The second method is relatively more complex: Stonewall Analytics (having been contracted by the SC SAC) uses advanced statistical techniques to analyze IPVV counts of victims along with socioeconomic contextual data. As SLED reviews agencies, the result will be a real-world comparison of methods to see which was more successful in pointing toward data issues. During the third phase, reporting agencies in South Carolina will benefit from SLED's resulting educational outreach. This report documents the analysis and findings of Stonewall Analytics.

An outlier is an observation that deviates from other observations in the sample from which it exists (1). 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 (1). Any findings based on analyses conducted with data containing outliers can have skewed coefficient estimation, prediction, and hypothesis tests (2). The intent of this project is to identify county outliers without considering the reason or cause.

This project represents the initial phase of an attempt to establish a methodology by which uniform crime reporting programs could readily identify counties that may be more likely to have data quality issues. If successful, this methodology could be used by all such programs to allow for a more focused and efficient assessment of data quality associated with IPVV.

In 1987, South Carolina served as the pilot state for the FBI's NIBRS. In 1991, the SCIBRS was the first uniform crime reporting program in the nation to become NIBRS-certified. The project at-hand gives South Carolina another opportunity to advance uniform crime reporting. The outcome of the statistically-guided agency review 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.

RELEVANT LITERATURE

The literature review covers two domains—the first is on analyses related to IPVV, while the second covers methodologies related to detecting outliers. 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 VIOLENT VICTIMIZATION ANALYSES

Prior studies have examined IPVV at the individual level (3), the county-level (4, 5), the neighborhood-level (6), and a hybrid mix between individual- and structural-level factors (7). Contextual data at the county level or census-tract level have been used to examine the factors associated with IPVV (8). IPVV primarily affects young women in their adolescence to early adulthood, usually within the realm of cohabitation with others and involves physical, sexual, or emotional abuse (9). In counties with women in powerful positions (positions of leadership in business or government), there is a marked decrease in the risk of IPVV (10). Other factors statistically associated with IPVV can include poverty-related measures, race, and ethnicity (5). Among perpetrators, factors such as employment and educational level have been shown to be associated with higher IPVV rates (11). For perpetrators that are Hispanic young adults, the degree of acculturation with community is significantly associated with attitudes related to IPVV (12).

DETECTING DATA OUTLIERS

Prior analyses have applied quantitative measures to predict repeat incidents of IPVV, but no prior studies to our knowledge were aimed at addressing the quality or integrity of the data itself (13). Some standard methodologies for examining outliers involve examining the dispersion (i.e., standard deviation) around the mean and identifying observations falling outside a specified range or confidence interval (14). The application of machine learning principles is another approach that is gaining popularity in outlier detection analysis—specifically the use of supervised machine learning (15, 16). Although popularity is growing in machine learning principles, as with all quantitative methodologies, they are not without their drawbacks. These tools and techniques, while robust and popular, have limitations, especially in the order in which they are employed, and in developing training and testing data splits for modelling (17).

OBJECTIVE OF THIS WORK

The objective of this work is two-fold. The first objective is to develop a quantitatively based model that can be applied to identify counties that are outliers in terms of IPVV count reporting. The second objective is to apply the model and criterion to identify counties that are outliers for IPVV count reporting. This report can provide the framework for monitoring the quality and integrity of this data going forward—in South Carolina and potentially in other states.

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[1] and contextual data gathered from the United States Census Bureau (American Community Survey), and the United States Department of Agriculture. National, county-level arrest data for 2011 from the Inter-University Consortium for Political and Social Research (ICPSR) was used for training, testing, and validation of the statistical models.

SCIBRS DATA

The SCIBRS data included variables related to IPVV for all 46 counties in South Carolina from 2011–2015. 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). For the statistical models, the outcome (or dependent) variable was the count of SCIBRS IPVV victims. All variables corresponding to crimes were reported as raw numbers, rather than normalized rates.

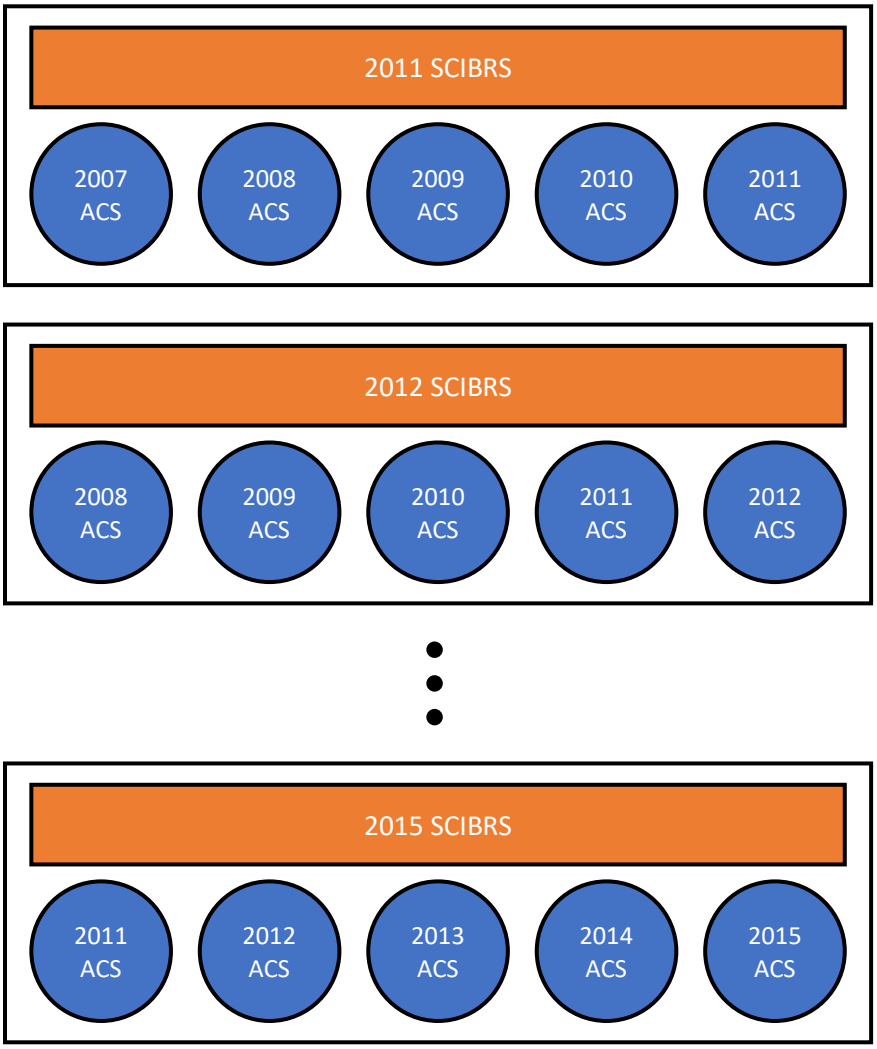
CONTEXTUAL DATA

Contextual data involving the county-level data from the American Community Survey (ACS) was 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 (18). The ACS is

[1] SCIBRS data downloaded as of July 14, 2017.

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. Figure 1 provides a schematic for how the data and the 5-year estimates were aligned for the 2011–2015 time period.

Figure 1: Aligning ACS and SCIBRS Data



Source: Stonewall Analytics

The literature indicated specific variables that were associated (statistically) with IPVV at the county level. We chose to include the following variables from the ACS:[2]

[2] The bold, square brackets indicate the shortened variable name used in the analysis with the statistical software (R). Of note, all variable names were in lowercase format for ease of typing within the statistical software—here, the names are presented in lowercase format for consistency.

- 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 equality, whereas an index of one represents maximum inequality. The list above contains 19 variables from ACS.[3] In addition to ACS data, prior research had incorporated Urban Rural Continuum Codes from the US Department of Agriculture (4), however because the variable does not change during the time-period of the analysis it was not included.

[3] In the initial stages of analysis (involving exploratory data analysis), several other variables were included, but as the statistical models were developed, they were excluded from the final phases of analysis.

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 SCIBRS IPVV count. As IPVV counts were not available in the training dataset, a proxy variable served as a bridge to the training and testing datasets. Through exploratory data analysis, the aggravated assault arrest variable from the national ICPSR data was identified as a close match to the SCIBRS IPVV.[4] Table 1 and Figure 3 (see the Results section), respectively, outline the descriptive statistics and distribution for these two variables.

SUPERVISED MACHINE LEARNING MODELS

As the goal of this phase of the project is to identify outlier counties, the use of a supervised machine learning model became the most suitable approach, since testing and training data were readily available. Machine learning can be divided into two domains—the first is supervised machine learning, and the second is unsupervised machine learning. For supervised machine learning, the outcome of interest is known to the researcher, whereas with unsupervised machine learning, the outcome is unknown to the researcher. Classifying the species of a flower based upon flower characteristics is an example of supervised machine learning. Researchers already know the species of flowers, but would use supervised machine learning to correctly classify any additional flowers whose species had not been identified. In an unsupervised machine learning model, researchers would not know each flower's species, and would employ the model to cluster flowers into separate groups based upon their known characteristics.

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.

Three separate supervised machine learning models were initially created and evaluated with respect to the training and testing data. These models included a decision tree regression model, a random forest

[4] The Results section presents the findings that compares the suitability of these two datasets for substitution/proxy.
[5] Training and testing were done using the ICPSR data.

regression model, and a generalized boosted regression model.[6] All syntax for the analysis is contained in an appendix to this document. Data for the training and testing models, and data that combine SCIBRS variables along with contextual variables, are available for download.[7] After testing and review, the random forest model was used as the final model in identifying potential counties for outliers. Random forest models tend to avoid model overfitting, are easy to interpret, and are completed quickly.

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, because in one year the county might fall within the expected range of the model, whereas 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. This criterion was tested for its sensitivity (i.e., a county was also flagged if it exceeded two standard deviations).

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.

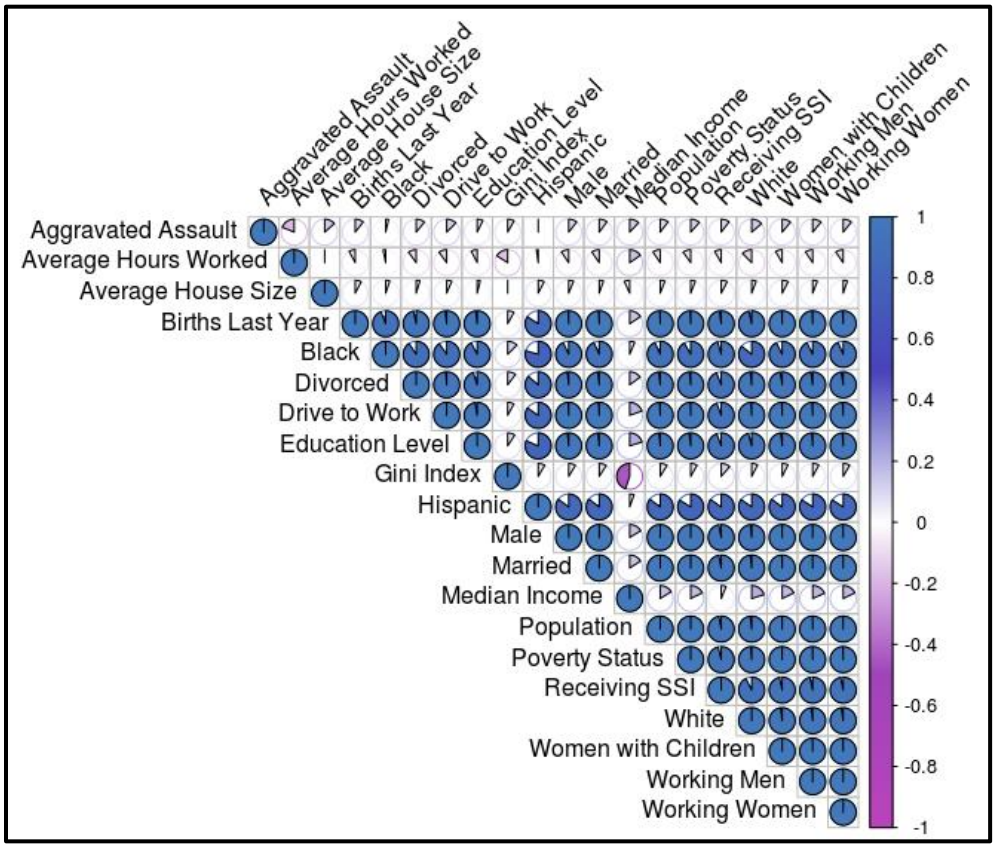
CONTEXTUAL DATA

Prior published literature has incorporated county-level contextual data from the ACS with research involving IPVV. A number of iterations involving exploratory data analysis and descriptive statistics were performed to examine the behavior of the contextual variables. Figure 2 presents a correlation matrix in a graphical form, referred to as a "correlogram" (19). This correlogram includes the contextual ACS data, along with the proxy variable (Aggravated Assault) from the training/testing ICPSR dataset. Performing a correlation analysis of variables from separate data sources is an important component to exploratory data analysis to examine the linear dependence between two variables. Here, the correlation coefficients are presented as pies. The color of the pie indicates the direction of the relationship (blue for positive, purple for negative). The size of the pie indicates the magnitude of the correlation

[6] All models and statistical analysis were performed using R, and specific packages within R include tree, randomForest, and gbm.
[7] https://www.stonewallanalytics.com/southcarolina

coefficient. For example, looking at the variable labels along the diagonal axis, Median Income has eight pies associated with other variables (these other variables correspond to the horizontal axis). When two variables intersect across the diagonal and horizontal axes, the pie represents the correlation coefficient for these two variables. Returning to Median Income, the correlation coefficient is 1 with itself, and the remaining variables (Population, Poverty Status, Receiving SSI, White, Women with Children, Working Men, and Working Women) are positive and have corresponding correlation coefficients less than 0.25. What is readily apparent is the relatively strong degree of correlation among ACS variables, and the contrast of the same strength of relationship among the training/testing outcome variable.

Figure 2: Correlogram for Variables of Interest



Source: Stonewall Analytics

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 aggravated assault arrest variables from the training/testing data was a close match to the smart total variable from the SCIBRS data. Table 1

provides an overview of descriptive statistics comparing the training/testing and the SCIBRS data for the outcome variables of interest.

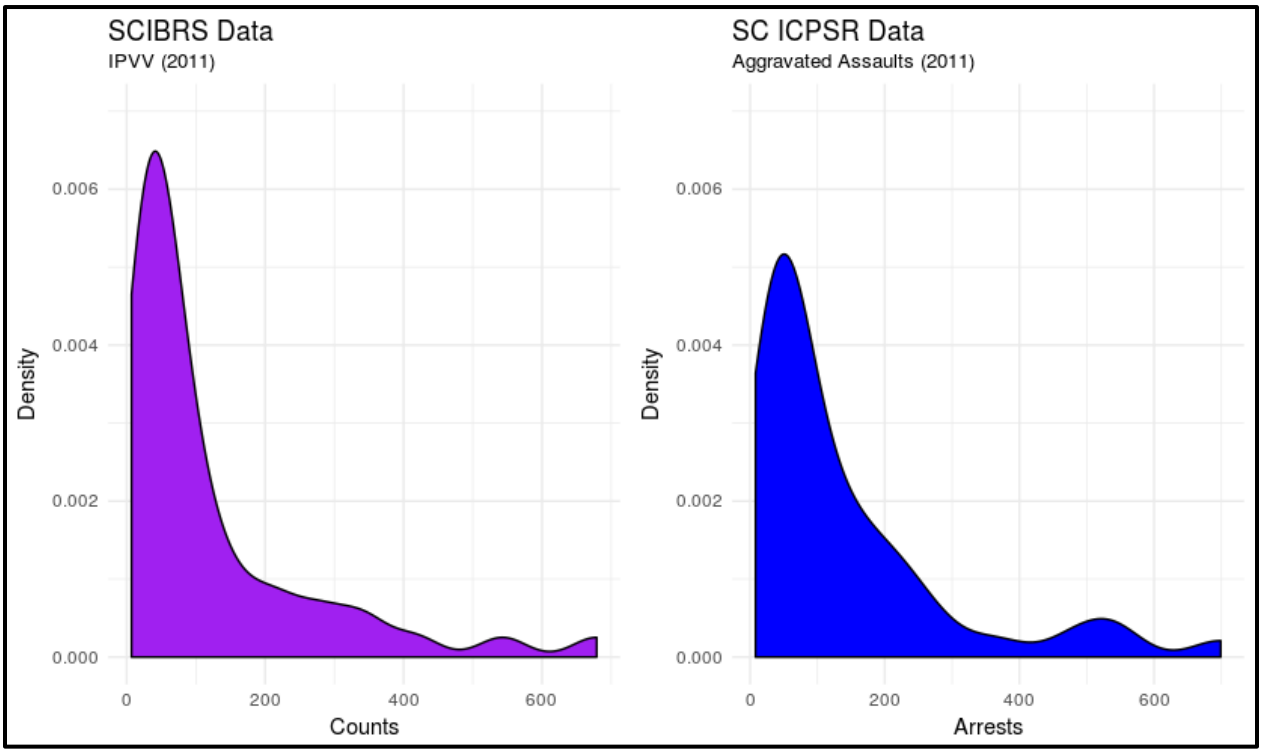
Table 1: Comparison of SCIBRS and the ICPSR Training/Testing Data (2011)

	SCIBRS	SC ICPSR
Minimum	7	8
25th Percentile	31	41
Median	59	69
Mean (SD)	119 (145)	139 (156)
75th Percentile	141	174
Maximum	679	699

Source: Stonewall Analytics

Figure 3 is a side-by-side density plot of the outcome variable from SCIBRS (2011 IPVV count), and the training/testing data from ICPSR (aggravated assault arrests in 2011 for South Carolina). Figure 3 demonstrates the distribution of these variables are similar in nature, although the SCIBRS data appears to have a greater density of counts between 0 and 200 when compared to the ICPSR data.

Figure 3: Distributions for Variables of Interest



Source: Stonewall Analytics

Apparent in Figure 3 is the low proportion of counts in the ICPSR data that extend beyond 200 counts. This behavior is apparent in the training/testing data. Due to the low proportion of counts, during model selection and refinement it became necessary to exclude some county-years from predictions that were higher in nature in order to arrive at a model with predictive abilities. The cutoff for including counties was set at or below the 75th percentile of IPVV counts of victims reported by counties in South Carolina in 2011, which was 141. Due to this cutoff at the 75th percentile of SCIBRS data, the following nine counties were excluded in the analysis, since their reported counts exceeded this cutoff for three or more years in the available data.

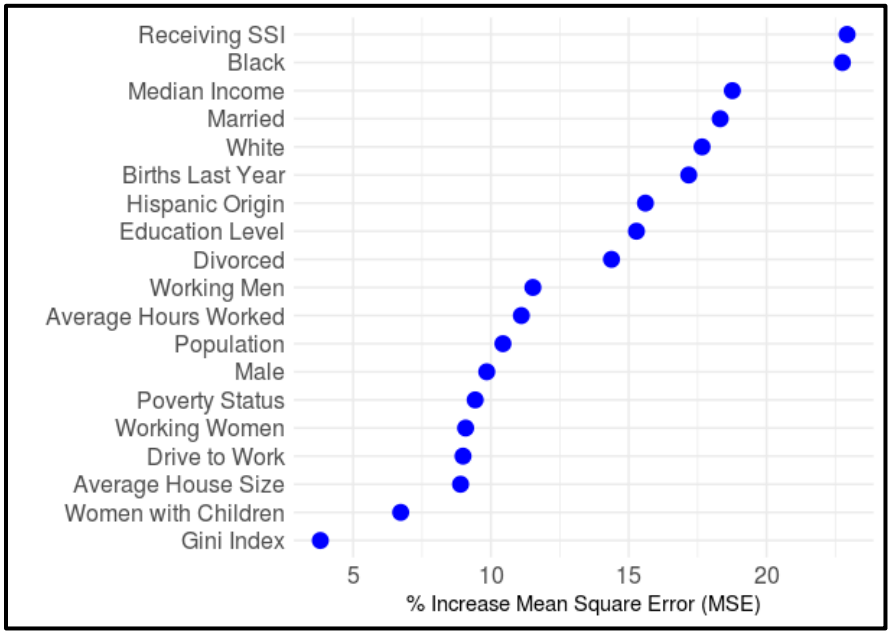
- Anderson
- Beaufort
- Charleston
- Greenville
- Horry
- Lexington
- Richland
- Spartanburg
- York

MODEL RESULTS

The supervised machine learning model was able to account for 45.6% of the variability in the dependent variable, in this case, the aggravated assault arrests for the training/testing dataset. The five most influential variables (in order) related to the number of arrests for aggravated assault with the training data included the number of residents receiving Supplemental Security Income (SSI) benefits, the number of black residents in the county, the median income for county residents, the number of married residents in the county, and the number of white residents in the county.[8] Figure 4 displays the variables of importance used in development of the machine learning model with the training data.

[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 4: Importance of Variables



Source: Stonewall Analytics

POTENTIAL COUNTY OUTLIERS

Our model identified 11 counties as potential outliers for incident reporting. Across all counties in the model, the average number of outlying years was 1.8, the standard deviation was 1.6, and the median was 2. The potential number of outliers across all counties for all years ranged from zero to five. Table 2 presents the county with the corresponding number of outlying years and the average difference (in absolute value) between the reported and predicted number of IPVV counts of victims.

Table 2: County-Level Summary Model Performance[9]

County	Number of Outlying Years	Average Difference
Abbeville	0	13
Aiken	2	19
Allendale	0	9
Bamberg	0	11
Barnwell	1	16
Berkeley	4	38
Calhoun	0	3
Cherokee	4	32
Chester	0	16
Chesterfield	4	36
Clarendon	4	26
Colleton	4	27
Darlington	1	22
Dillon	0	11
Dorchester	1	13
Edgefield	3	26
Fairfield	0	9
Florence	3	34
Georgetown	2	26
Greenwood	0	14
Hampton	0	9
Jasper	4	31
Kershaw	2	15
Lancaster	2	21
Laurens	2	18
Lee	0	16
Marion	2	17
Marlboro	0	9
McCormick	0	4
Newberry	5	35
Oconee	2	17
Orangeburg	2	20
Pickens	5	31
Saluda	0	9
Sumter	4	54
Union	0	6
Williamsburg	1	20

Source: Stonewall Analytics

[9] The average difference column in this table is rounded to the nearest whole number for the county-years included in the analysis. Some counties did not have all county-years included, as they fell outside of the 75th percentile cutoff.

Table 3 presents the counties that were identified as potential outliers. With the model and the outlier criterion, 11 counties were flagged as potential outliers. Seven of the 11 counties, on average, had lower-than-expected reported counts. The remaining four counties had higher-than-expected reported counts. Across the five years stratified by county, the average difference between the reported and predicted counts ranged between 26 and 54 counts.

Table 3: Potential County Outliers[10]

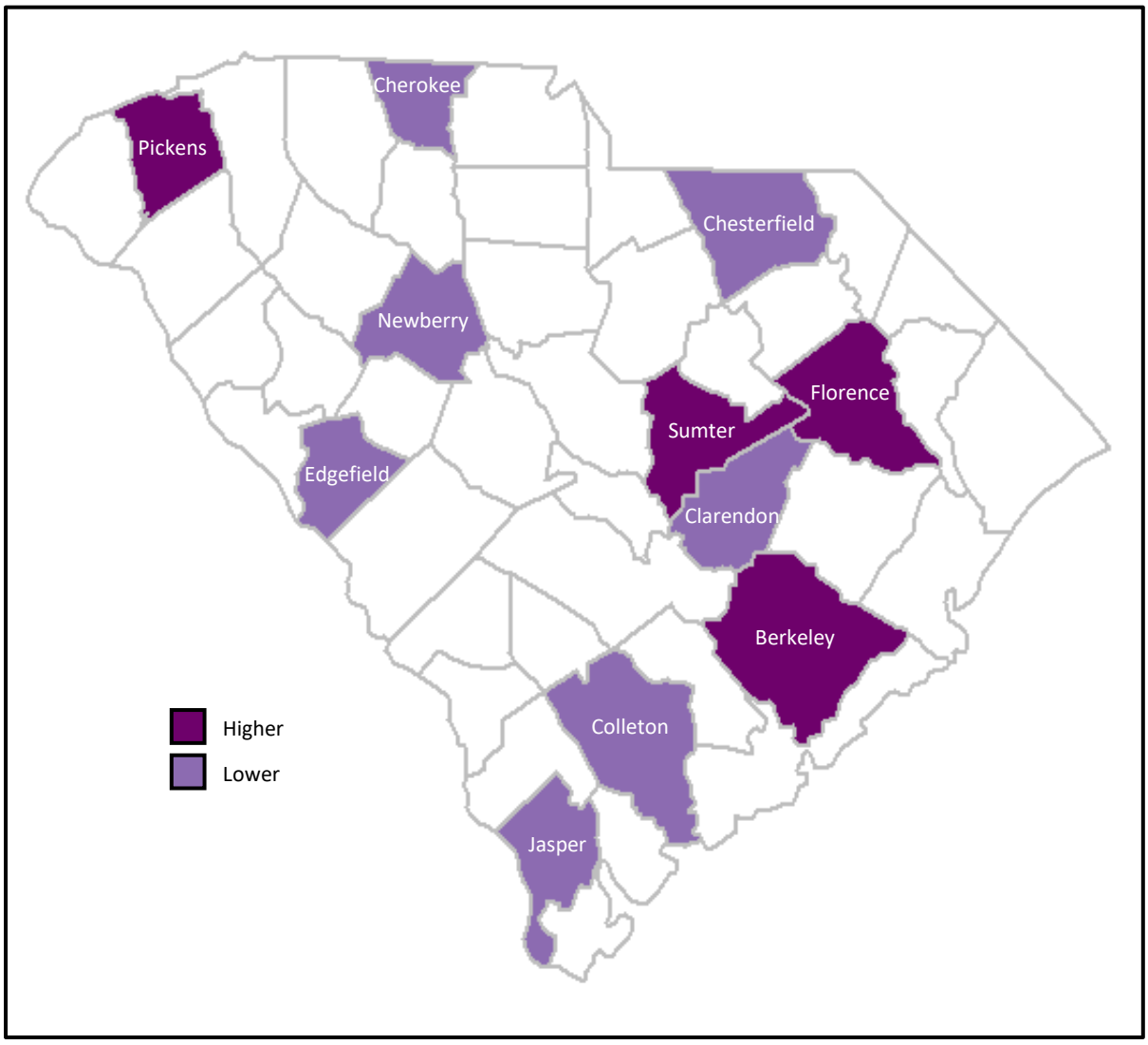
County	Average Difference	Direction
Berkeley	38	Higher
Cherokee	32	Lower
Chesterfield	36	Lower
Clarendon	26	Lower
Colleton	27	Lower
Edgefield	26	Lower
Florence	34	Higher
Jasper	31	Lower
Newberry	35	Lower
Pickens	31	Higher
Sumter	54	Higher

Source: Stonewall Analytics

[10] The average difference column is rounded to the nearest whole number. Due to rounding, minor differences may appear when comparing two or more tables in this report.

Figure 5 displays the counties flagged as outliers. The color coding for the flagged counties corresponds the relation of reported values to predicted values: the lighter color purple indicates a lower reported value, and darker color purple indicates a higher reported value.

Figure 5: County-Level Map of Outlying Counties



Source: Stonewall Analytics

DISCUSSION

This report documents the framework for developing a machine learning model that was applied to SCIBRS IPVV victim data to assess data integrity and quality at the county level. A supervised machine learning model was created to predict IPVV counts of victims in South Carolina from 2011–2015. The model was then applied to contextual data to create predicted values of IPVV counts 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. Eleven counties were identified as outliers. SLED plans to review agencies to determine whether the county truly has reporting inconsistencies. Counties that are potential outliers have either higher-than-reported or lower-than-reported predictions for victim counts. One recommendation is to incorporate these machine learning principles to monitor ongoing data collection with SCIBRS (20).

As this analysis is novel in assessing the quality of IPVV count data, there are several recommended future studies. Future studies should evaluate the sensitivity to the operational definition used for determining what constitutes domestic violence. Some states and counties throughout the United States, to include South Carolina, have domestic violence resources available to residents. Future studies could examine county-level changes based upon the availability of domestic violence resources to determine if there is any associated effect on IPVV rates (9).

While this study presents a number of strengths, it is necessary to cover the limitations, too. Perhaps the most significant is the model's inability to evaluate counties with a higher volume of victim 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 IPVV with health-related measures. Since this study's variables do not capture health-related aspects, the study likely suffers from some omitted variable bias (3, 21). Another limitation is not controlling for the factors associated with county-level policing practices and policies (22). 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 IPVV in any state. This framework and methodology has the ability to scale to other states, time periods, and can be used to assess the quality and integrity of incident reporting going forward.

REFERENCES

1. Hodge V, Austin J. A survey of outlier detection methodologies. *Artificial Intelligence Review*. 2004;22(2): 85-126.
2. Jensen DR, Ramirez DE. Shift outliers in linear inference. *Journal of Multivariate Analysis*. 2015;136: 95-107.
3. Decker MR, Peitzmeier S, Olumide A, Acharya R, Ojengbede O, Covarrubias L, et al. 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*. 2014;55(6): S58-S67.
4. Madkour AS, Martin SL, Halpern CT, Schoenbach VJ. Area disadvantage and intimate partner homicide: An ecological analysis of North Carolina counties, 2004-2006. *Violence and Victims*. 2010;25(3): 363-77.
5. Tiefenthaler J, Farmer A, Sambira A. Services and intimate partner violence in the United States: A county-level analysis. *Journal of Marriage and Family*. 2005;67(3): 565-78.
6. Beyer KMM, Layde PM, Hamberger LK, Laud PW. Does neighborhood environment differentiate intimate partner femicides from other femicides? *Violence Against Women*. 2014;21(1): 49-64.
7. Copp JE, Kuhl DC, Giordano PC, Longmore MA, Manning WD. Intimate partner violence in neighborhood context: The roles of structural disadvantage, subjective disorder, and emotional distress. *Social Science Research*. 2015;53: 59-72.
8. Cunradi CB, Caetano R, Clark C, Schafer J. Neighborhood poverty as a predictor of intimate partner violence among White, Black, and Hispanic couples in the United States. *Annals of Epidemiology*. 2000;10(5): 297-308.
9. Lundgren R, Amin A. Addressing intimate partner violence and sexual violence among adolescents: Emerging evidence of effectiveness. *Journal of Adolescent Health*. 2015;56(1): S42-S50.
10. Whitaker MP. Linking community protective factors to intimate partner violence perpetration. *Violence Against Women*. 2014;20(11): 1338-59.
11. Peitzmeier SM, Kågesten A, Acharya R, Cheng Y, Delany-Moretlwe S, Olumide A, et al. Intimate partner violence perpetration among adolescent males in disadvantaged neighborhoods globally. *Journal of Adolescent Health*. 2016;59(6): 696-702.

12. Grest CV, Amaro H, Unger J. Longitudinal predictors of intimate partner violence perpetration and victimization in Latino emerging adults. *Journal of Youth and Adolescence*. 2017.
13. Roehl J, O'Sullivan C, Webster D, Campbell J. Intimate partner violence risk assessment validation study, final report. (Document No. 209731). Washington DC: National Institute of Justice. 2005.
14. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*. 2013;49(4): 764-6.
15. Aggarwal CC. Outlier Analysis. 2 ed. New York: Springer; 2017.
16. Tang B, He H. A local density-based approach for outlier detection. *Neurocomputing*. 2017;241: 171-80.
17. Abedjan Z, Chu X, Dong D, Fernandez RC, Ilyas IF, Ouzzani M, et al. Detecting data errors: Where are we and what needs to be done? 42nd International Conference on Very Large Data Bases, VLDB 2016; 9/9/20162016. p. 993-1004.
18. US Census Bureau. American Community Survey: Information Guide. 2013.
19. Wei T, Simko V. corrplot: Visualization of a Correlation Matrix. 0.77 ed2016.
20. Hazen BT, Boone CA, Ezell JD, Jones-Farmer LA. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*. 2014;154: 72-80.
21. Jiang Y, DeBare D, Shea LM, Viner-Brown S. Violence against women: Injuries and deaths in Rhode Island. *Rhode Island Medical Journal*. 2017;100(12): 24-8.
22. Leisenring A. Controversies surrounding mandatory arrest policies and the police response to intimate partner violence. *Sociology Compass*. 2008;2(2): 451-66.

Table 4: Model Results of All Counties by Year[11]

County	Year	Reported	Predicted	Difference	Flag	Direction
Abbeville	2011	42	21	21	0	-
Abbeville	2012	39	20	19	0	-
Abbeville	2013	14	20	6	0	-
Abbeville	2014	14	23	9	0	-
Abbeville	2015	15	23	8	0	-
Aiken	2011	93	80	13	0	-
Aiken	2012	86	80	6	0	-
Aiken	2013	88	76	12	0	-
Aiken	2014	102	79	23	1	Higher
Aiken	2015	124	84	40	1	Higher
Allendale	2011	37	16	21	0	-
Allendale	2012	6	14	8	0	-
Allendale	2013	19	15	4	0	-
Allendale	2014	5	12	7	0	-
Allendale	2015	9	12	3	0	-
Bamberg	2011	10	23	13	0	-
Bamberg	2012	15	21	6	0	-
Bamberg	2013	12	21	9	0	-
Bamberg	2014	8	21	13	0	-
Bamberg	2015	8	21	13	0	-
Barnwell	2011	30	46	16	0	-
Barnwell	2012	33	43	10	0	-
Barnwell	2013	22	46	24	1	Lower
Barnwell	2014	28	41	13	0	-
Barnwell	2015	22	41	19	0	-
Beaufort	2015	87	50	37	1	Higher
Berkeley	2011	103	74	29	1	Higher
Berkeley	2012	130	65	65	1	Higher
Berkeley	2013	78	64	14	0	-
Berkeley	2014	89	64	25	1	Higher
Berkeley	2015	118	61	57	1	Higher
Calhoun	2011	20	16	4	0	-
Calhoun	2012	14	15	1	0	-
Calhoun	2013	9	13	4	0	-

[11] Some counties are omitted from this list, and some counties are not listed for all five years. The model did not assess counties when the reported violent crime exceeded the 75th percentile (141 counts) due to balancing overall model performance. The predicted column includes values that are rounded to the nearest whole number. Due to rounding, minor differences may appear when comparing two or more tables in this report.

County	Year	Reported	Predicted	Difference	Flag	Direction
Calhoun	2014	11	13	2	0	-
Calhoun	2015	9	13	4	0	-
Cherokee	2011	21	53	32	1	Lower
Cherokee	2012	26	61	35	1	Lower
Cherokee	2013	6	63	57	1	Lower
Cherokee	2014	49	64	15	0	-
Cherokee	2015	45	67	22	1	Lower
Chester	2011	43	55	12	0	-
Chester	2012	33	51	18	0	-
Chester	2013	28	50	22	0	-
Chester	2014	38	48	10	0	-
Chester	2015	26	46	20	0	-
Chesterfield	2011	48	70	22	0	-
Chesterfield	2012	55	79	24	1	Lower
Chesterfield	2013	24	73	49	1	Lower
Chesterfield	2014	32	71	39	1	Lower
Chesterfield	2015	28	73	45	1	Lower
Clarendon	2011	53	55	2	0	-
Clarendon	2012	31	55	24	1	Lower
Clarendon	2013	15	52	37	1	Lower
Clarendon	2014	16	56	40	1	Lower
Clarendon	2015	28	53	25	1	Lower
Colleton	2011	35	61	26	1	Lower
Colleton	2012	36	65	29	1	Lower
Colleton	2013	26	66	40	1	Lower
Colleton	2014	43	59	16	0	-
Colleton	2015	38	61	23	1	Lower
Darlington	2011	129	113	16	0	-
Darlington	2012	105	87	18	0	-
Darlington	2013	63	77	14	0	-
Darlington	2014	64	79	15	0	-
Darlington	2015	27	75	48	1	Lower
Dillon	2011	73	63	10	0	-
Dillon	2012	48	63	15	0	-
Dillon	2013	44	58	14	0	-
Dillon	2014	38	52	14	0	-
Dillon	2015	52	50	2	0	-
Dorchester	2011	83	76	7	0	-
Dorchester	2012	83	83	0	0	-
Dorchester	2013	73	88	15	0	-
Dorchester	2014	76	93	17	0	-
Dorchester	2015	70	96	26	1	Lower
Edgefield	2011	13	31	18	0	-
Edgefield	2012	10	43	33	1	Lower

County	Year	Reported	Predicted	Difference	Flag	Direction
Edgefield	2013	8	43	35	1	Lower
Edgefield	2014	8	32	24	1	Lower
Edgefield	2015	9	30	21	0	-
Fairfield	2011	33	38	5	0	-
Fairfield	2012	29	43	14	0	-
Fairfield	2013	38	48	10	0	-
Fairfield	2014	41	46	5	0	-
Fairfield	2015	39	48	9	0	-
Florence	2012	107	68	39	1	Higher
Florence	2013	83	70	13	0	-
Florence	2014	109	69	40	1	Higher
Florence	2015	113	70	43	1	Higher
Georgetown	2011	66	73	7	0	-
Georgetown	2012	55	76	21	0	-
Georgetown	2013	53	73	20	0	-
Georgetown	2014	39	79	40	1	Lower
Georgetown	2015	36	78	42	1	Lower
Greenwood	2013	78	69	9	0	-
Greenwood	2014	83	71	12	0	-
Greenwood	2015	94	73	21	0	-
Hampton	2011	41	39	2	0	-
Hampton	2012	30	37	7	0	-
Hampton	2013	20	35	15	0	-
Hampton	2014	24	35	11	0	-
Hampton	2015	26	37	11	0	-
Jasper	2011	7	26	19	0	-
Jasper	2012	7	34	27	1	Lower
Jasper	2013	6	47	41	1	Lower
Jasper	2014	5	44	39	1	Lower
Jasper	2015	10	40	30	1	Lower
Kershaw	2011	56	62	6	0	-
Kershaw	2012	71	68	3	0	-
Kershaw	2013	43	69	26	1	Lower
Kershaw	2014	37	69	32	1	Lower
Kershaw	2015	62	71	9	0	-
Lancaster	2011	61	96	35	1	Lower
Lancaster	2012	76	80	4	0	-
Lancaster	2013	67	75	8	0	-
Lancaster	2014	62	79	17	0	-
Lancaster	2015	47	87	40	1	Lower
Laurens	2011	101	65	36	1	Higher
Laurens	2012	90	68	22	1	Higher
Laurens	2013	79	59	20	0	-
Laurens	2014	75	63	12	0	-

County	Year	Reported	Predicted	Difference	Flag	Direction
Laurens	2015	63	64	1	0	-
Lee	2011	23	42	19	0	-
Lee	2012	32	44	12	0	-
Lee	2013	22	43	21	0	-
Lee	2014	28	45	17	0	-
Lee	2015	31	40	9	0	-
Marion	2011	46	55	9	0	-
Marion	2012	46	56	10	0	-
Marion	2013	34	54	20	0	-
Marion	2014	34	59	25	1	Lower
Marion	2015	32	55	23	1	Lower
Marlboro	2011	44	41	3	0	-
Marlboro	2012	63	50	13	0	-
Marlboro	2013	45	50	5	0	-
Marlboro	2014	28	47	19	0	-
Marlboro	2015	48	51	3	0	-
McCormick	2011	10	8	2	0	-
McCormick	2012	16	8	8	0	-
McCormick	2013	6	8	2	0	-
McCormick	2014	6	9	3	0	-
McCormick	2015	3	9	6	0	-
Newberry	2011	25	54	29	1	Lower
Newberry	2012	27	56	29	1	Lower
Newberry	2013	19	58	39	1	Lower
Newberry	2014	14	61	47	1	Lower
Newberry	2015	30	61	31	1	Lower
Oconee	2011	109	76	33	1	Higher
Oconee	2012	94	79	15	0	-
Oconee	2013	87	87	0	0	-
Oconee	2014	66	77	11	0	-
Oconee	2015	47	75	28	1	Lower
Orangeburg	2011	72	85	13	0	-
Orangeburg	2012	38	78	40	1	Lower
Orangeburg	2013	109	78	31	1	Higher
Orangeburg	2014	94	80	14	0	-
Orangeburg	2015	83	80	3	0	-
Pickens	2011	65	42	23	1	Higher
Pickens	2012	84	42	42	1	Higher
Pickens	2013	66	39	27	1	Higher
Pickens	2014	66	39	27	1	Higher
Pickens	2015	78	40	38	1	Higher
Saluda	2011	28	30	2	0	-
Saluda	2012	11	27	16	0	-
Saluda	2013	20	27	7	0	-

County	Year	Reported	Predicted	Difference	Flag	Direction
Saluda	2014	14	27	13	0	-
Saluda	2015	17	26	9	0	-
Sumter	2012	133	79	54	1	Higher
Sumter	2013	131	82	49	1	Higher
Sumter	2014	136	81	55	1	Higher
Sumter	2015	137	79	58	1	Higher
Union	2011	29	26	3	0	-
Union	2012	28	32	4	0	-
Union	2013	32	32	0	0	-
Union	2014	19	33	14	0	-
Union	2015	23	33	10	0	-
Williamsburg	2011	28	39	11	0	-
Williamsburg	2012	27	46	19	0	-
Williamsburg	2013	18	46	28	1	Lower
Williamsburg	2014	28	47	19	0	-
Williamsburg	2015	25	46	21	0	-

Source: Stonewall Analytics

R Syntax for Project Analysis and Related Figures[12]

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.

```
getwd()
```

Importing the SCIBRS data.

```
sc_scibrs <- read.csv(file = 'sc_scibrs_data.csv', header = TRUE)
```

Importing the ICPSR data.

```
icpsr <- read.csv(file = 'icpsr.csv', header = TRUE)
```

Extracting 2011 SCIBRS data and South Carolina ICPSR for comparison.

```
sc_scibrs11 <- sc_scibrs[sc_scibrs$year == 2011 , ]
icpsr_sc <- icpsr[icpsr$state == 45 , ]
```

Creating basic density plots in ggplot2 of outcome variables of interest.

```
library(ggplot2)
f <- ggplot(data = sc_scibrs11, aes(x = smart_total))
f <- f + geom_density(fill = 'blue')
f <- f + theme_minimal()
f <- f + ylim(0, 0.007)
f <- f + labs(title = 'SCIBRS Data',
              subtitle = 'Violent Crime Smart Total',
              x = 'Counts',
              y = 'Density')

g <- ggplot(data = icpsr_sc, aes(x = agg_assault_arrest))
g <- g + geom_density(fill = 'red')
g <- g + theme_minimal()
```

[12] 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.

```
g <- g + ylim(0, 0.007)
g <- g + labs(title = 'SC (ICPSR) Data',
              subtitle = 'Aggravated Assaults',
              x = 'Arrests',
              y = 'Density')

## making side-by-side plots
library(gridExtra)
grid.arrange(f, g, ncol = 2)
```

Summary/descriptive statistics of both dependent variables.

```
summary(sc_scibrs11$violent_crime_smart_total)
sd(sc_scibrs1111$violent_crime_smart_total)
summary(icpsr_sc$agg_assault_arrest)
sd(icpsr_sc$agg_assault_arrest)
```

Now creating a training and testing sample from the ICPSR data (75% training, 25% testing).

```
set.seed(8675309)
train <- sample(x = 1:nrow(icpsr), size = nrow(icpsr) * 0.75)
```

Creating an example of a pairs plot with random variables of interest (substitute others you may feel are appropriate).

```
pairs(formula = ~ avg_house_size + gini_index + avg_hours_worked,
      data = icpsr,
       subset = train)
```

Now making a correlation plot (correlelogram).

```
library(corrplot)
c <- cor(x = icpsr[train,])
colnames(c) <- c('Average House Size', 'GINI Index', 'Hispanic', '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', 'Aggravated Assault')
rownames(c) <- c('Average House Size', 'GINI Index', 'Hispanic', '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', 'Aggravated Assault')

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
```

```
par(mfrow = c(1,1))
corrplot(c,
      method = "pie", # visualization method,
      shade.col = NA, # color of shade line
      tl.col = "black", # color of text label
      tl.srt = 45, # text label rotation
      col = col(200), # color of glyphs
      order = "alphabet",
      diag = TRUE,
      type = 'upper')
```

Now creating a standard linear model. On inspection of residuals, apparent systematic behavior is present, leading us to determine that machine learning models seem like a valid approach.

```
fit1 <- lm(formula = agg_assault_arrest ~ ., data = icpsr, subset = train)
summary(fit1)
plot(fit1)

library(MASS)
fit2 <- glm.nb(formula = agg_assault_arrest ~ ., data = icpsr, subset = train)
summary(fit2)

yhat_reg <- predict(object = fit1, newdata = icpsr[-train , ])

icpsr_test <- icpsr[-train, 'agg_assault_arrest'] ## we will call this multiple times in other functions below

plot(x = icpsr_test, y = yhat_reg)
abline(0,1)

mean((yhat_reg - icpsr_test)^2) # mean square error
sqrt(mean((yhat_reg - icpsr_test)^2)) # standard deviation
```

Now moving forward with a regression tree.

```
library(tree)
train_tree <- tree(formula = agg_assault_arrest ~ . , data = icpsr, subset = train)
summary(train_tree)
plot(train_tree)
text(train_tree, pretty = 0)

complex <- cv.tree(train_tree) # cross-fold validation to determine optimal level of complexity
complex
plot(complex$size, complex$dev,type = 'b')

## pruning for interpretation
train_prune <- prune.tree(train_tree, best = 6) # have your best match the above plot for number
summary(train_prune)
```

```
plot(train_prune)
text(train_prune)

## let's look the prediction aspect
yhat_tree <- predict(object = train_prune, newdata = icpsr[-train , ])
plot(x = icpsr_test, y = yhat_tree)
abline(0,1)

mean((yhat_tree - icpsr_test)^2) # mean square error
sqrt(mean((yhat_tree - icpsr_test)^2)) # standard deviation
```

Now a gradient boosted model.

```
library(gbm)
train_boost <- gbm(formula = agg_assault_arrest ~ . ,
                   data = icpsr[train , ], # there is no subset command in this function
                   distribution = 'gaussian',
                   n.trees = 5000,
                   shrinkage = 0.001,
                   interaction.depth = 3)
summary(train_boost)
yhat_boost <- predict(object = train_boost,
                      newdata = icpsr[-train , ],
                      n.trees = 5000,
                      interaction.depth = 3)
plot(x = icpsr_test, y = yhat_boost)
abline(0,1)

mean((yhat_boost - icpsr_test)^2)
sqrt(mean((yhat_boost - icpsr_test)^2))
```

Random forest model with output of figure showing variable importance. The random forest model is what will be applied to the prediction aspect for identifying potential outlier counties.

```
library(randomForest)
train_rf <- randomForest(formula = agg_assault_arrest ~ . ,
                         data = icpsr,
                         subset = train,
                         mtry = 19,
                         n.trees = 100,
                         importance = TRUE)
train_rf
yhat_rf <- predict(object = train_rf, newdata = icpsr[-train , ])
plot(x = icpsr_test, y = yhat_rf)
abline(0,1)

mean((yhat_rf - icpsr_test)^2) # mean square error
```

```
sqrt(mean((yhat_rf - icpsr_test)^2)) # standard deviation

out <- as.data.frame(importance(train_rf))
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', '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

library(ggplot2)
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 = '')
j <- j + theme(axis.text = element_text(size = 12))
j

varImpPlot(x = train_rf, main = 'Variable Importance Plot')
```

Now the random forest model is applied to the SC SCIBRS data to identify potential outlier counties.

```
## setting the number of standard deviations to evaluate
st_dev <- 1 * (sqrt(mean((yhat_rf - icpsr_test)^2)))

## 2011
yhat_rf11 <- predict(object = train_rf,
             newdata = socar[ , -c(1, 2, 22, 23)][socar$year == 2011 &
                                    socar$violent_crime_smart_total <= 141, ] )
summary(yhat_rf11)
summary(socar$violent_crime_smart_total[socar$year == 2011 & socar$violent_crime_smart_total <=
141])

one <- socar[ , c(1,22)][socar$year == 2011 & socar$violent_crime_smart_total <= 141 , ]
two <- yhat_rf11

results11 <- cbind(one, two)
names(results11) <- c('county', 'reported', 'predicted')

## 2012
 yhat_rf12 <- predict(object = train_rf,
              newdata = socar[ , -c(1, 2, 22, 23)][socar$year == 2012 &
                                     socar$violent_crime_smart_total <= 141, ] )
 summary(yhat_rf12)
 summary(socar$violent_crime_smart_total[socar$year == 2012 & socar$violent_crime_smart_total <=
```

```
141])

 one <- socar[ , c(1,22)][socar$year == 2012 & socar$violent_crime_smart_total <= 141 , ]
 two <- yhat_rf12

 results12 <- cbind(one, two)
 names(results12) <- c('county', 'reported', 'predicted')


## 2013
 yhat_rf13 <- predict(object = train_rf,
                 newdata = socar[ , -c(1, 2, 22, 23)][socar$year == 2013 &
                                          socar$violent_crime_smart_total <= 141, ] )
 summary(yhat_rf13)
 summary(socar$violent_crime_smart_total[socar$year == 2013 & socar$violent_crime_smart_total <=
141])

 one <- socar[ , c(1,22)][socar$year == 2013 & socar$violent_crime_smart_total <= 141 , ]
 two <- yhat_rf13

 results13 <- cbind(one, two)
 names(results13) <- c('county', 'reported', 'predicted')


## 2014
 yhat_rf14 <- predict(object = train_rf,
                 newdata = socar[ , -c(1, 2, 22, 23)][socar$year == 2014 &
                                          socar$violent_crime_smart_total <= 141, ] )
 summary(yhat_rf11)
 summary(socar$violent_crime_smart_total[socar$year == 2014 & socar$violent_crime_smart_total <=
141])

 one <- socar[ , c(1,22)][socar$year == 2014 & socar$violent_crime_smart_total <= 141 , ]
 two <- yhat_rf14

 results14 <- cbind(one, two)
 names(results14) <- c('county', 'reported', 'predicted')


## 2015
 yhat_rf15 <- predict(object = train_rf,
                 newdata = socar[ , -c(1, 2, 22, 23)][socar$year == 2015 &
                                          socar$violent_crime_smart_total <= 141, ] )
 summary(yhat_rf15)
 summary(socar$violent_crime_smart_total[socar$year == 2015 & socar$violent_crime_smart_total <=
141])

 one <- socar[ , c(1,22)][socar$year == 2015 & socar$violent_crime_smart_total <= 141 , ]
 two <- yhat_rf15
```

```
 results15 <- cbind(one, two)
 names(results15) <- c('county', 'reported', 'predicted')


## OVERALL
results11$year <- 2011
results12$year <- 2012
results13$year <- 2013
results14$year <- 2014
results15$year <- 2015

dta <- rbind(results11, results12, results13, results14, results15)

## setting up the flag component
dta$flag <- 0
dta$flag[dta$predicted - st_dev > dta$reported |
            dta$predicted + st_dev < dta$reported] <- 1

## clean-up
dta$predicted <- round(dta$predicted,0)

## showing the difference
dta$delta_abs <- round(abs(dta$predicted - dta$reported),0)

## whether higher or lower than expected
dta$reported_direction <- NA
dta$reported_direction[dta$predicted - st_dev > dta$reported] <- 'lower'
dta$reported_direction[dta$predicted + st_dev < dta$reported] <- 'higher'

dta <- dta[order(dta$county, dta$year, dta$flag),]
```

Finally rolling up the data to identify counties by the number of years as an outlier.

```
library(sqldf)
dta2 <- sqldf("select county, sum(flag) as 'count', avg(delta_abs) as 'mean_diff', reported_direction
                  from dta
                  group by county")
dta2$mean_diff <- round(dta2$mean_diff, 0)
dta2
```