|  |  |
| --- | --- |
| Public safety outcomes  Create Safer Communities | Megan Cusey  11/10/2020 |

# Executive Summary

The following report proposes an alternative solution to the Budgeting for Results Unit’s work on program assessments that begins with a top-down approach of identifying key components of already defined Result Areas and Outcomes. With each Outcome, an analysis can be performed that models key factors of what impacts the outcome which can be interpreted to key stakeholders to provide advice on how to each the outcomes desired. In addition, it would provide an iterative and ongoing methodology that provides comparisons over time that would allow us to validate our response and help determine the effectiveness of the solutions and programs the State provides.

It is assumed that key decision makers and domain expertise will be available to assist in the success of this approach. Committees can be established to help guide the modeling process in identifying and collecting meaningful data as it relates to each outcome. A recommended start in gathering these resources is to enlist the assistance of the leadership of each program that is assigned to a particular state outcome.

The text details each step in the CRISP Data-Mining project methodology that includes Business Understanding, Data Understanding, Data Preparation, Modeling, Model Evaluation & Validation, and Deployment. This is meant to be an example of how the process can be applied to our use case and open a conversation about how we may implement this approach. The project was performed with existing resources that are either open sourced, already procured by the state, or apart of a trial license. With additional time, project iterations, leadership support, and perhaps monetary investments, this analysis will only improve and be able to extract more useful information.

Table of Contents

[Executive Summary 1](#_Toc55414303)

[Introduction 3](#_Toc55414304)

[Business Understanding 5](#_Toc55414305)

[Data Understanding 8](#_Toc55414306)

[Data Preparation 9](#_Toc55414307)

[Modeling 10](#_Toc55414308)

[Model Evaluation and Validation 11](#_Toc55414309)

[Model Deployment 12](#_Toc55414310)

[Conclusion 13](#_Toc55414311)

# Introduction

The Budgeting for Results (BFR) Unit at the Governor’s Office of Management and Budget (GOMB) is tasked with the challenge of evaluating programs and awarding budgeted dollars based off merit. Since its creation, the BFR unit has established a catalog of over 400 programs across 70 state agencies, universities, boards, and commissions. In addition, a program evaluation tool titled SPART, derived from the Federal PART (Program Assessment Rating Tool) has been implemented to assess these programs. The SPART features a benefit-cost component that is the result of a collaborative effort between Results-First Pew Research Foundation and Washington State Institute of Public Policy designed for analyzing and monetizing outcomes for state run programs. A second portion of the SPART is a qualitative rating that considers how well a program is executed and managed. The SPART is a proven, effective, and consistent methodology for reviewing state run programs against best practices and outcomes for Illinois.

This project seeks to take a different approach to the objectives of the BFR unit. While the SPART tool provides a great methodology and standardization for program evaluation, there is a disconnect in the BFR unit’s stated goals and the prioritization of work that I believe could be improved with additional data driven analysis.

The table on the next page was taken from the [***2020 Budgeting for Results Annual Report***](https://www2.illinois.gov/sites/budget/Documents/Budgeting%20for%20Results/2020%20BFR%20Annual%20Commission%20Report.pdf). It defines the seven statewide Result Areas and their nine outcomes and definitions. The goal of this report is to suggest a methodology for the Budgeting for Results unit to follow this table as a guide to accomplish several goals:

* Create a process that will iteratively review these Statewide Result Areas, more clearly define their outcomes, and understand what effects the outcome measures.
* Establish sub-committees for each Result Area – Outcome that combines leadership from all the programs that have the corresponding outcome in the Illinois Performance Reporting System (IPRS) to identify additional data that can be included in the model, domain knowledge that may be applied, and other opportunities to better understand these outcomes.
* Apply the scientific methodology to answer questions and validate hypothesis on what approaches may be reasonable to accomplish a particular goal or effect an outcome measure.
* Produce a dashboard that is more inclusive of all result areas and a holistic look at the State of Illinois in terms of BFR defined Result Areas and Outcomes.
* Use analysis from a top-down approach (Result Area focused) to drive bottom-up analysis (SPART analysis). The analysis this report discusses will identify opportunities, strengths, and weaknesses that may be used as a guide to the SPART analysis efforts.

Throughout the remainder of this report, I will focus on one Result Area/Outcome to provide an example of how we can accomplish the above in a first iteration attempt with completely free resources: my time and publicly available data. Keep in mind that with future iterations, better data, and more domain knowledge, the results of this model will continue to improve and provide more insightful knowledge to guide BFR’s efforts and political leaders. The intention is that in future project efforts, we will repeat the same process for each Result Area/Outcome.



Table 1 - This table displays the seven statewide Result Areas along with their nine associated outcome areas and definitions defined in the 2020 Budgeting For Results Annual Report

# Business Understanding

The Result Area/Outcome that will be focused on is Public Safety – Create Safer Communities. This outcome is defined as “reduce incidence of death, violence, injury, exploitation and fraud” in the 2020 Budgeting for Results Annual Report. The definition of the outcome is extremely broad in nature. I encourage a more defined or further hierarchy of outcomes in the future, but for now – I extracted all programs assigned tqo the *Create Safer Communities* outcome in IPRS to better define the outcome. The table on the next page lists each program name sorted by the agency that runs the program.

My interpretation of the goals of the below programs can be summarized as follows:

* Reduce recidivism by addressing mental health and substance use among inmates, providing support to assist in establishing a productive and stable life outside of the correctional facility, and providing technology to assist in parole monitoring.
* Limit the number of preventable accidents or injuries that occur by the inspection and upkeep of certain structures and enforcing safety guidelines.
* Lower the crime rate by proper law enforcement training, community engagement, and prevention programs.
* Shorten, minimize, or prevent the effects of natural or man-made disasters that (may) occur.
* Lessen the number of motor vehicle accidents per year by providing safety outreach programs.

For the purpose of this exercise, I choose to focus on index crime rate proportioned by population due to data availability and the general understanding of what influences crime in any given county. In future iterations of this project, I would work with leaders from the below programs and come up with a compilation of what metrics we can use to support the outcome*, Create Safer Communities.* This would be used to provide an overall analysis of how Illinois is doing in achieving this outcome and if we are doing better or worse over time.

The business objective in this case would be to gain an understanding of what influences the crime index rate per population and to what extent. Investigating what impacts the crime index rate will help us understand the following:

* What populations to target our program efforts in?
* What areas of Illinois should be focused on?
* How can we produce more effective positive outcomes in Illinois?
* Where to best place or utilize our resources to get the maximum gain?
* Which programs may have a larger positive impact if scaled up or in a certain area?
* What opportunities are available to us with the new insights gathered by the analysis?

|  |  |
| --- | --- |
| Agency | Program Name |
| Department of Corrections | Facility Operations |
| Department of Corrections | Sheridan Correctional Center/Southwestern Illinois Correctional Center |
| Department of Corrections | Mental Health Treatment |
| Department of Corrections | Vocational Programming |
| Department of Corrections | Parole Operations |
| Department of Corrections | Educational Programming |
| Department of Corrections | Parole Re-Entry |
| Department of Corrections | Substance Abuse Treatment |
| Department of Corrections | Electronic Monitoring |
| Department of Corrections | GPS Monitoring |
| Department of Financial & Professional Regulation | Regulatory Enforcement |
| Department of Human Services | Domestic Violence Prevention and Intervention |
| Department of Insurance | Property and Casualty Insurance Products |
| Department of Juvenile Justice | Aftercare Services |
| Department of Labor | Amusement Ride and Attraction Safety |
| Department of Military Affairs | Illinois National Guard |
| Department of Transportation | Promote/Enforce Highway Safety |
| Department of Transportation | Promote/Enforce Motor Carrier Safety |
| Department of Transportation | Promote Motorcyclist Safety |
| Illinois Criminal Justice Information Authority | Federal Funding |
| Illinois Criminal Justice Information Authority | Violence Prevention and Reduction |
| Illinois Criminal Justice Information Authority | Adult Redeploy Illinois |
| Illinois Criminal Justice Information Authority | Restore, Reinvest, and Renew (R3) Program |
| Illinois Criminal Justice Information Authority | Research Programs |
| Illinois Emergency Management Agency | Homeland Security Preparedness |
| Illinois Emergency Management Agency | Disaster Assistance |
| Illinois Emergency Management Agency | Mitigation |
| Illinois Emergency Management Agency | Regulation and Licensing of Radioactive Materials and X-Ray Equipment |
| Illinois Emergency Management Agency | Nuclear Evaluation, Monitoring and Response |
| Illinois Emergency Management Agency | Disaster Coordination |
| Illinois Emergency Management Agency | Radiological Emergency Preparedness |
| Illinois Emergency Management Agency | Environmental Monitoring |
| Illinois Emergency Management Agency | Nuclear Facility Inspection |
| Illinois Emergency Management Agency | Escort, Incident Response and Preventive Radiological Nuclear Detection |
| Illinois Emergency Management Agency | Radon Activities |
| Illinois Emergency Management Agency | Hazardous Materials |
| Illinois Emergency Management Agency | Low Level Radioactive Waste (LLRW), Decommissioning and Site Cleanup |
| Illinois Law Enforcement Training Standards Board | Reimbursement of Training Expenses |
| Illinois Law Enforcement Training Standards Board | In-Service Training |
| Illinois Law Enforcement Training Standards Board | Law Enforcement Intern Program |
| Illinois State Police | Public Safety Enforcement |
| Illinois State Police | Support of Law Enforcement Programs |
| Illinois State Police | Forensic Services and Identification |
| Illinois State Police | Internal Investigation |
| Illinois State Police | Adult-Use Cannabis Program |
| Illinois State Police Merit Board | Recruitment and Selection |
| Illinois State Police Merit Board | Promotional Assessments |
| Illinois State Police Merit Board | Disciplinary Hearings |
| Liquor Control Commission | Liquor Control Regulation |
| Office of The State Fire Marshal | Fire Service Education and Grants |
| Office of The State Fire Marshal | Fire Prevention |
| Office of The State Fire Marshal | Boiler and Pressure Vessel Safety |
| Office of The State Fire Marshal | Petroleum and Chemical Safety |
| Office of The State Fire Marshal | Arson Investigation |
| Office of The State Fire Marshal | Elevator Safety |
| Office of The State Fire Marshal | Technical Services |
| Prisoner Review Board | Parole and Mandatory Supervised Release (MSR) Revocation Hearings |
| Prisoner Review Board | Mandatory Supervised Release |
| Prisoner Review Board | Clemency |
| Prisoner Review Board | Juvenile Parole Revocation Hearings |
| Prisoner Review Board | Modification of Release Condition Hearings |
| Prisoner Review Board | Discretionary Parole Consideration Hearings |
| Prisoner Review Board | Victim Notification |
| Prisoner Review Board | Statutory Sentence Credit Review |

Table 2 - IPRS Programs assigned the outcome of Create Safer Communities.

# Data Understanding

The data used for this analysis is all free, publicly available data found on the internet by reputable sources. The below table describes each source, the link to the data, and description of what the data describes. As a result of pulling in these data fields, we get a better understanding of crime, drugs, economy, and family makeup for each county.

|  |  |  |
| --- | --- | --- |
| Data Source | Link | Description |
| Illinois State Police | [Crime in Illinois 2018 Annual Uniform Crime Report](https://isp.illinois.gov/CrimeReporting/Cii2018) | The Illinois State Police reports on crime each year including the crime index rate and drug arrests. |
| Census.gov | [County Population Estimates](https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html) | Census.gov adds information on the racial and age make-up of the population for each county. |
| Census Reporter | [Table B11003: Family Type by Presence and Age of Own Children](https://censusreporter.org/tables/B11003/)  [Table B17001: Poverty Status by Sex by Age](https://censusreporter.org/tables/B17001/)  [Table B19122: Number of Earners in Family](https://censusreporter.org/tables/B19122/) | This data source has many different topics to choose from to pull in the analysis. The data I decided to include in this analysis have to do with the family structures and income of residents for each county. |
| Bureau of Economic Analysis | [GDP by County](https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas)  [Personal Income per Capita by County](https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas) | The Bureau of Economic Analysis provides statistics per county that describe the economic health. |
| Bureau of Labor Statistics | [Local Area Unemployment Statistics](https://beta.bls.gov/dataViewer/view) | BLS adds the unemployment rate per county to this analysis. |

Table 3 - Data Sources included in the analysis

In the following sections, we will review each of the features and describe their distributions.

## About the Target

The target variable, or dependent variable, is Total Index Crime Raters Per 100K. In the following image, you can see that counties such as Vermillion, Massac, and Winnebago have the highest crime index rates per populations of 100K.

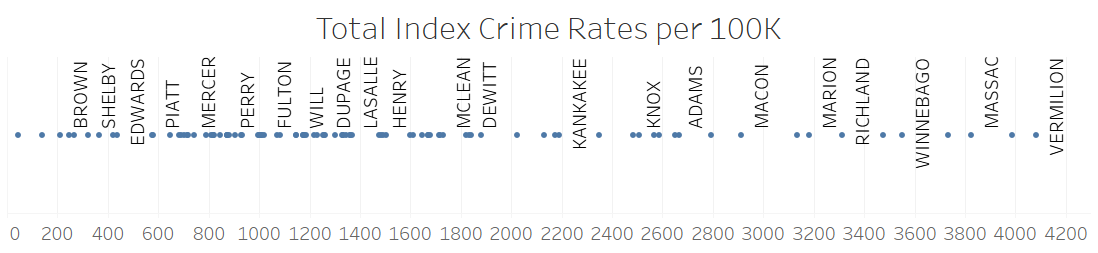


Figure 1 - Distribution of Index Crime Rate per 100K

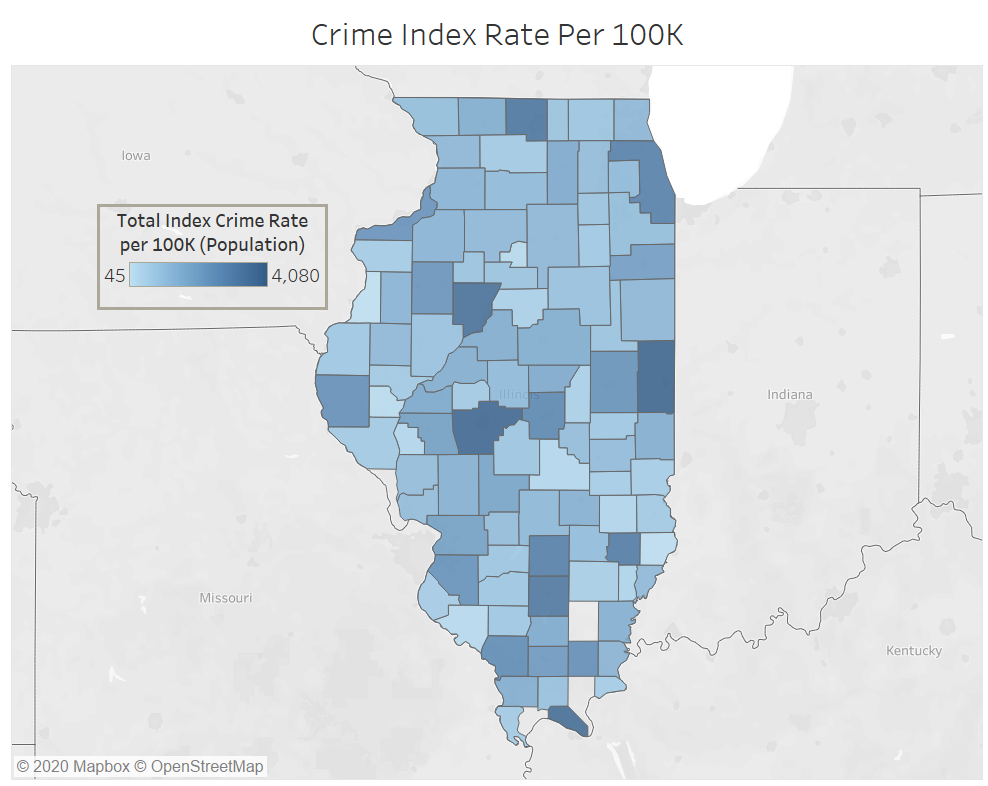


Figure 2 - A map of the State of IL with color coordinating scale of index crime rates per 100K

## Crime Features

The crime features are all proportioned by the population of the counties per 100,000 people. These data describe the following types of activities: arson, battery/assault, burglary, homicides, human trafficking involuntary servitude, human trafficking sexual conduct, motor vehicle theft, robbery, theft, and crime index rate.

This data is all derived from the Illinois State Police Annual Report. Per the distributions below, you can see that arson, homicides, human trafficking, and robbery are among the rarer crimes. Battery/Assault, burglary, motor vehicle theft, and theft make up most of the total crime index rate since they are more popular.

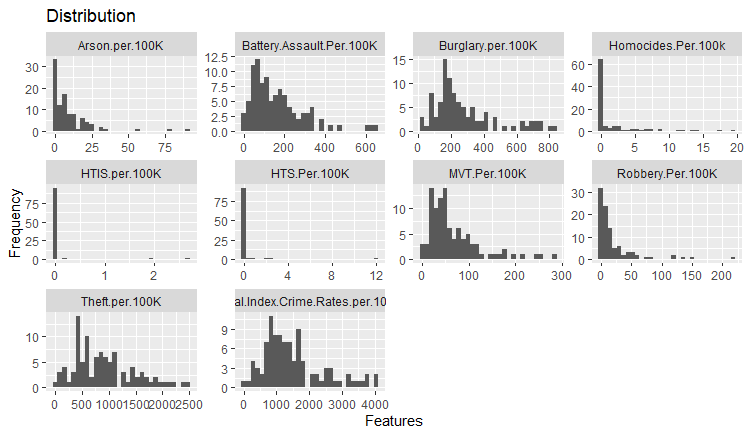


Figure 3 - Distribution of Crime Features

## Drug Features

The below image describes the distribution of drug related arrests per 100K population.

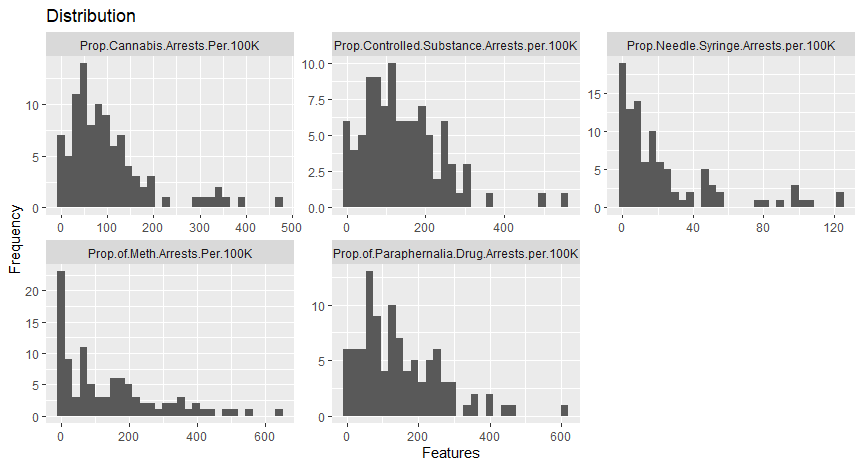


Figure 4 - Distribution of Drug Features

## Economic Features

The below images depict the distribution of economic features included in the analysis.

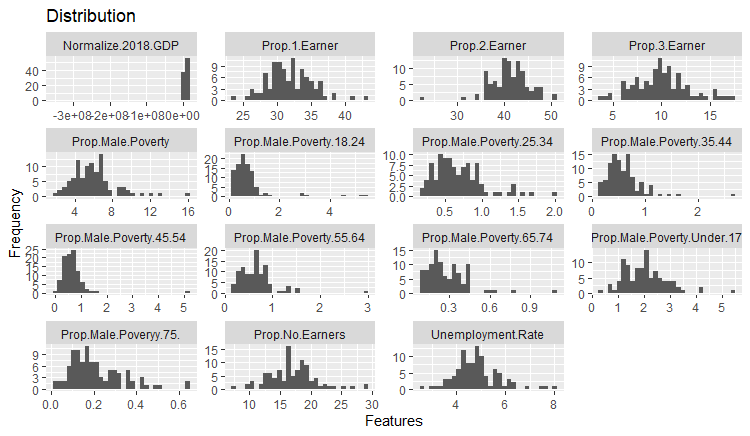


Figure 5 - 1st set of economic features distribution

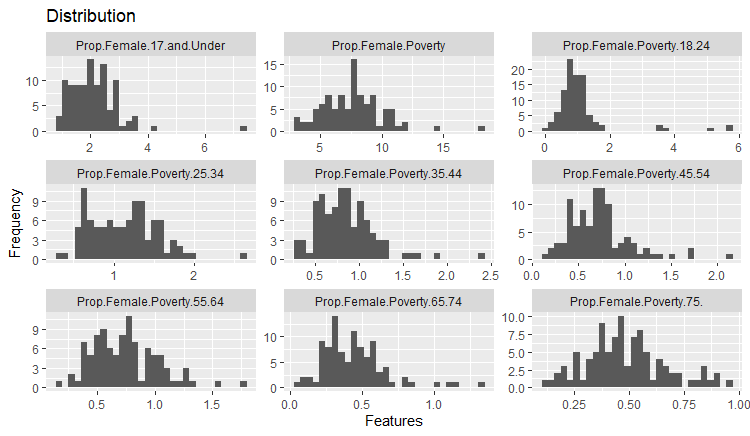


Figure 6 - 2nd set of economic features distribution

## Family Features

Below are the family related features included in the analysis along with the distributions of each data.

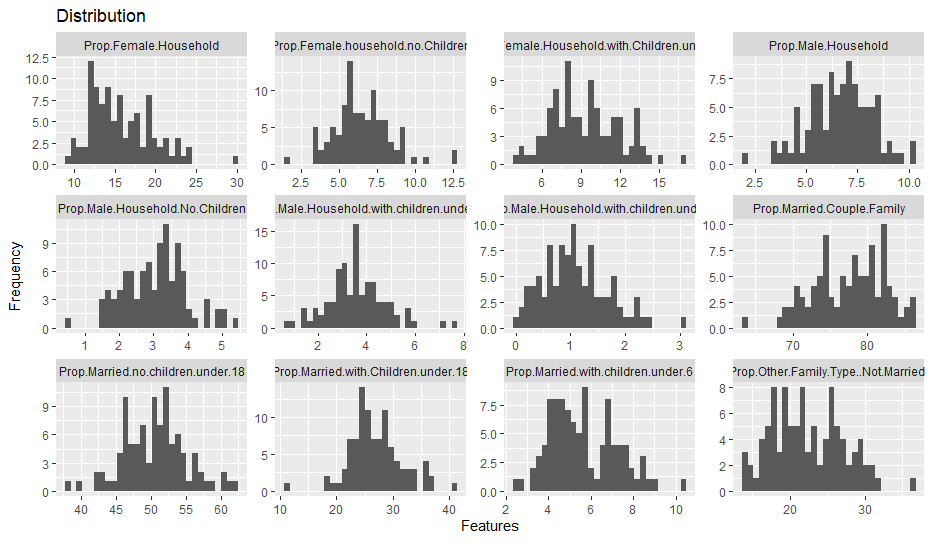


Figure 7 - Distribution of family related features.

# Data Preparation

Many data preparation steps were taken to combine the data from its raw format, to one that can be consumed by modeling algorithms. Each file format had to be reformatted where the level of granularity is one row per county. In addition, the county names are sometimes spelled differently between files. For example, “De Witt” county is sometimes spelled as “DeWitt” instead. Since the files are being joined based on the county names, any discrepancies in spellings had to be resolved.

In addition to combining the raw data points into a single file with each item calculated by county, it was also necessary to scale the features that would put counts in terms of another relative feature. For example, the number of homicides in Cook County is 638 and in Champaign County, 10. From eyeballing this information, the homicide rate in Cook County seems significantly greater than that of Champaign County. However, if you calculate the homicide rate in terms of population per 100,000 people, Cook County has 12.3 per 100K and Champaign County has 4.78 per 100K. This makes the counties easier to compare against each other, not only visually, but for the model algorithm as well. Without this adjustment, the distribution of homicides is extremely skewed and the number of homicides for Cook County would appear as an outlier due to the high population. In the previous sections, you see the features have already been scaled against an appropriate feature.

The next item of business is to address multicollinearity against the independent variables. The goal of this exercise is to find independent variables (everything but the target variable) that have a strong correlation with the dependent variable (the target variable, Crime Index Rate Per 100k). However, the interpretation of the results, specifically the coefficients of the algorithms or variable importance information, are not reliable if there is a high degree of correlation between the dependent variables.

Below is a covariance matrix that shows the correlation between the variables in the algorithm. Some significant covariances are expected such as the Percent of Male Population and the Percent of Female Population would travel together since one is the opposite of the other. It is necessary to review the high correlations and strategically select which variables to include and which to exclude.

The following describes how to read the covariance matrix:

* The correlations range from 1 to -1.
* 0 = no correlation, the closer the cell is to yellow indicates the correlation is close to 0
* 1 = positive correlation, the closer the cell is to red indicates the correlation is close to 1
* -1 = negative correlation, the closer the cell is to green indicates the correlation is close to -1
* The goal is to balance the correlations between independent variables while maximizing the correlations for the dependent variable.

While the below is difficult to read and only illustrates color, you can use the [Correlation’s Spreadsheet](https://d.docs.live.net/886d1fcb0e1fdeb1/Documents/GitHub/DataScienceProjects/MDS%20576%20-%20Public%20Safety%20Outcomes/correlations.csv) to view the original file.

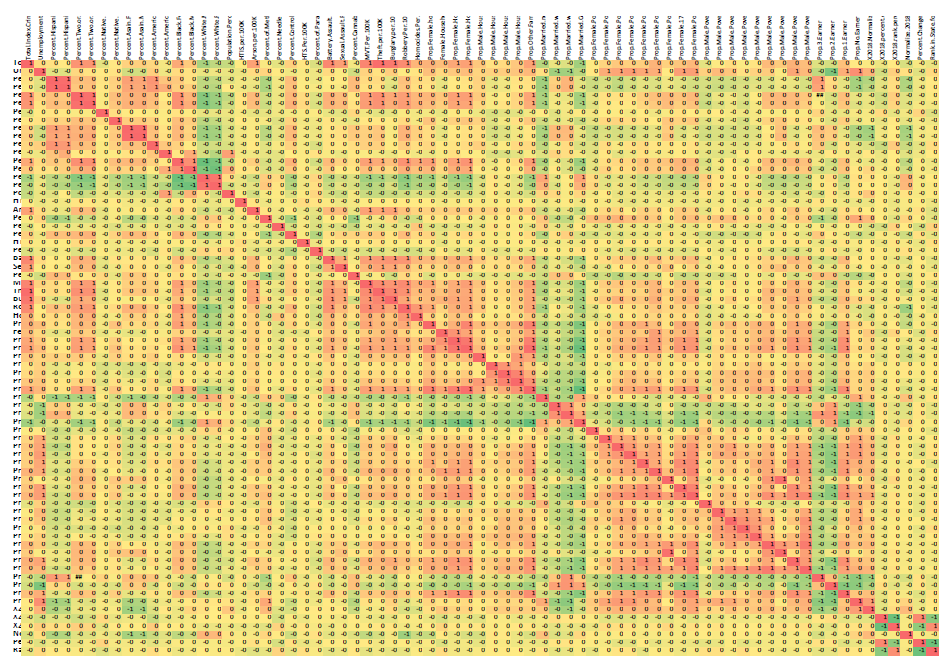


Figure 8 - Covariance Matrix of all features

The covariance matrix was calculated and written to an Excel file with RStudio/R. The heat map was generated by Excel.

Due to the correlations amongst the independent variables, it was decided to remove the following features:

|  |  |
| --- | --- |
| Percent Two or More Races – Female | Percent White – Female |
| Percent American Indian – Female | Percent Asian – Female |
| Percent Native Hawaiian – Female | Percent Hispanic – Female |
| Percent Female | Proportion of “Other” Family Type – Not Married |
| Rank in Percent Change – Personal Income per Capita | Rank in Percent Change – GDP |

Due to data leakage, which is a feature that is indicative of the target (dependent variable), the following features were excluded from the analysis:

|  |  |
| --- | --- |
| Motor Vehicle Theft per 100K | Theft per 100K |
| Burglary per 100K | Robbery per 100K |
| Homicides per 100K | Human Trafficking Involuntary Servitude per 100K |
| Human Trafficking Sexual Offense per 100K | Battery per 100K |
| Sexual Assault per 100K |  |

The data combining and cleaning steps were accomplished by utilizing Tableau Prep Builder.

# Modeling

There are many different modeling approaches that can be used to tackle this objective, however; each come with their own pros and cons. The most significant component of this use case is to be able to gain insights by assessing the importance of what variables impact the crime rate and by how much? Another consideration in choosing the type of model to be applied is the simplicity, transparency, and interpretability of the results.

In addition to the above modeling characteristics, it is important to understand the challenge of this dataset as well. There are only about 100 observations (counties) in the dataset and 60-70 features to consider. With the shape of this dataset, there is often intercorrelation between variables and spurious correlations found for the target (index crime rate per 100K) variable. When modeling and validating the model, the limited observations also make it difficult to trust the predictions as much as we would like. However, the feature importance can still be derived which is the object of this assessment.

Ideally, for future Result Area – Outcome analysis, a more appropriate, longer, dataset can be obtained through a committee of vested stakeholders.

With the above considerations in mind, the following features were found to be the most important as it relates to the crime index rate per 100K in Illinois counties. The score column provides a reference point for how strong the algorithm utilizes the variable in determining the predicted crime index rate. The independent variable with the largest strength will always be scored at 100. The subsequent independent variables scores can be used then to compare strength.

The below results were obtained through MARS Regression Splines. The algorithm allows for nonlinearity in the variables by allowing basic functions to adjust the equation of the line at major points in the independent variable’s values as identified by the modeling program. Similar feature importance results were seen with other model methods used.

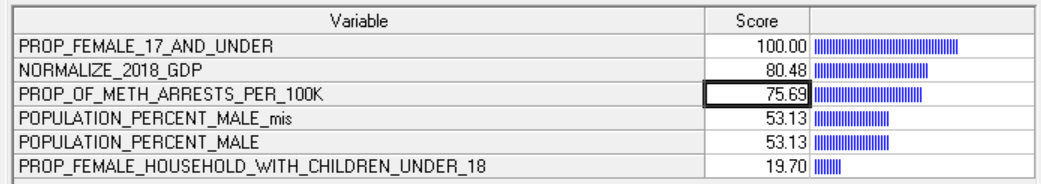


Figure 9 - Feature Importance produced by MARS Regression Splines

The above results are interesting in that it appears Meth arrests increase the crime rate as well as young females & female only (no husband) households with children. Understanding that these groups of people are drivers of the crime rate can help us ensure that we are reaching these groups of people and make sure that we have programs available to them for assistance. Next questions might be, how well are we reaching these people? How effective have those programs been? Are they available in the same areas we see these high crime rates in?

# Model Evaluation and Validation

The MARS Regression Splines algorithm was found preferable over other methods because it provides transparency through the basic functions algorithms that can be imported to a data visualization program such as Tableau. In addition, the model performed well at an R-Squared of .70, meaning the features included in the model account for 70% of the dependent variable/target (crime rate index per 100K). I assume that there is some multicollinearity that can be discovered and identified with other variables not included in this analysis that would provide better clarification on what impacts the crime rate index per 100K, but this is a good start.

The MARS Regression Spines algorithm also limited the dataset to 6 features instead of the 60-70 original features. This helps tremendously in reducing correlations between variables by narrowing the dataset.

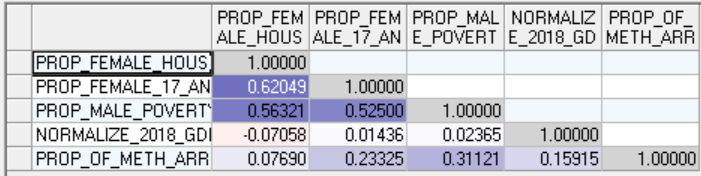


Figure 10 - Covariance Matrix of Features selected by the model.

This algorithm was able to achieve better and more consistent R-Squared results and limited the RMSE compared to other modeling approaches. The learn and test sets divide up the 100 observations to 1.) produce the model and 2.) validate the model. However, with the limitations of the dataset, these results are a bit difficult to rely on. I am pleased to see that in key evaluation measurements, the learn and test set perform similarly and helps to believe that the model does not overfit the data.

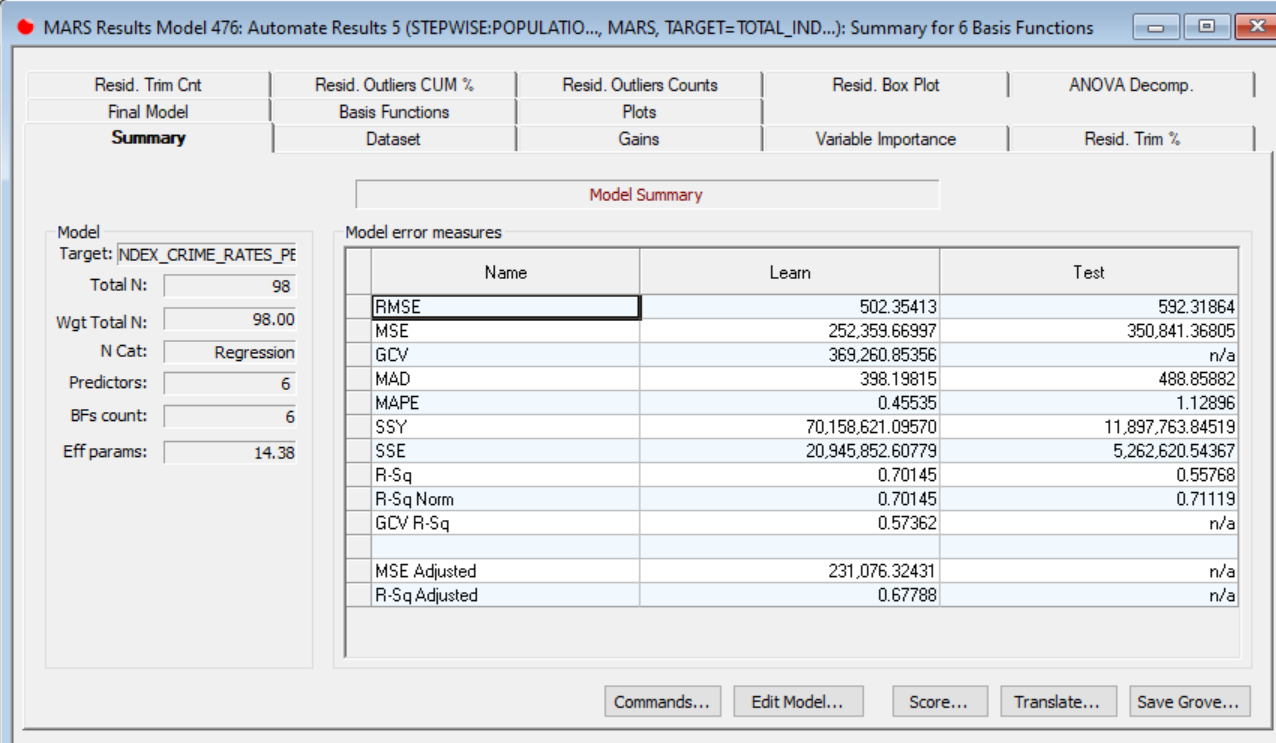


Figure 11 - Summary of model evaluation metrics

# Model Deployment

In deploying the model results, three approaches are recommended.

* With the nature of the MARS Regression Spline, it would be ideal to implement the basic function algorithms in to a data visualization program such as Tableau and allow business leaders to play around with the values of the independent variables to see the impact it has on the dependent variable (crime rate).
* Another deployment method would be to extract the insights from the model and provide a report to the committee invested in the outcomes and legislative leadership in highly impacted areas to review and encourage action.
* Information learned such as the most susceptible population to contribute to the crime rate can be internalized to ensure that we are reaching these people with programs the State offers and are effectively assisting them in their situations.

# Conclusion

To summarize this report, the objective of the analysis was to provide a use case in applying BFR Result Area/Outcomes with data science techniques in understanding what impacts an outcome. This study focused on Public Safety, Creating Safer Communities – specifically crime rates in counties of Illinois. This is an extremely broad topic that would need some work in narrowing down and identifying the best ways to measure outcomes.

The report provides a path of how similar analysis would be conducted in future Result-Area/Outcome studies. The hope is that resources and time would be invested into this approach to drive insights and guide more specific program assessments the BFR team has already implemented.

Even with the Guinea Pig approach to this project, interesting results were still obtained by following the CRISP methodology. The end results discovered that female populations/female only households, and high rates of meth arrests are indicative of higher crime rates. Many things can be done with this information, but the BFR unit can review programs that are designed to assist in rehabilitating IL citizens with Meth addictions and provide resources to assist single mothers and cross reference them with areas of high crime rates. Are the resources available in the correct areas of IL? Can we improve how we reach these people?

Hopefully, in the future we have the ability to further refine our goals and outcomes, gather more meaningful data, and gain additional insights on how to increase of program effectiveness by first analyzing the condition of out Result Areas & Outcomes.