A Machine Learning Approach To Crime In the US

Joshua Fernandes[1] 22169738

[1] Birmingham City University, 15 Bartholomew Row, Birmingham B5 5JU

Abstract.
Violent Crime Rates over the world have been increasing and this is even more so in the US. On the other hand, physicians have reported a burnout rate 3 times higher in 2023 than in 2019 as well as the number of police officers decreasing from 2020.Previous studies showed ways in which crime rate has been predicted based on socioeconomic data but focused a lot on human knowledge for feature selection or used correlation to select features that reflected a racial bias. This project aims to eliminate any racial bias as well as using feature selection methods to use on multiple models to choose the best one. To begin with the data quality was checked using the P value to see if the data had any statistical significance. A linear regression model was then run on all the features that were chosen using human knowledge. Using the linear regression model an MSE value and an R2 score were gathered. Using the highest R2 score the features for the model were chosen. Both classification models and regression models are used in this study with and without best parameters. Each model was run 10 times and averaged out to maintain integrity. The classifier models chosen were Decision Tree Classifier, Random Forest Classifier and Logistic Regression. The Random Forest Classifier and Regressor performed better than their counterparts. The random forest classifier had an accuracy of 67.47% and the random forest regressor had an MSE and R2 score of 0.07 and 0.47 respectively. These 2 models were then deployed on to streamlit to forecast crime using user input with a futuristic User Interface.

Keywords: crime, prediction, data,mining

1 Introduction

Higher crime rates have an adverse effect not only on the people directly affected by it but the people around. Crime imposes larger costs on community through lower property values, higher insurance premiums and reduced investment in high crime areas (Shapiro & Hassett, 2022). Over the past few years, crime has increased drastically in the US especially when it comes to Violent Crimes. From 2019-2021 violent crimes like homicides and aggravated assault went up by 39.42% and 14.9% percent respectively showing that crimes are getting more and more violent (Policy Circle,2023) whereas non-violent crimes such as larceny and burglaries went down. One of the problem the US is having is the number of police officers is reducing going from 696,644 in 2020 to 660,288 in 2021(Statista,2022). This shows us that as violent crime is going up the police force is going down shows us our need for machine learning regarding crime rate prediction as this shows a greater need for resource allocation. An increase in violent crime results in an increased need for physicians. This is a problem because as it stands physicians are already overworked. As shown by a recent study conducted by the national burnout survey series in which it was shown that 63% of physicians in 2021 reported symptoms of burnout going up from 38%in 2020(American Medical Association,2023). This shows that as violent crime is going up, the amount of police officers is going down and the amount of burnout endured by physicians is going up. A machine learning application for crime rate prediction helps to solve both these things.

Many scholars have produced papers on data mining in crime as well as predictive policing. Examples of this would be papers featuring communities and crime dataset and a New York crime dataset. Existing studies feature all kinds of data mining methods for classification and regression. 5 of the 6 studies feature the communities and crime dataset. 3 of which split the target variable into classes and predict those classes. Another study features crime in New York and how to predict any one of 25 classes of crime. What all these papers have in common are that they have to do with crime in the US and feature predictive policing.

1.1 Aim

The aim of this project is to design and develop a system to predict the levels of crime and if possible a ballpark value of Violent Crimes predicted to occur based on the state and other socioeconomic features. This dissertation will have the following objectives:

- Conducting a systematic literature review to see similar projects and help us make early inferences into the field of crime namely crime rate prediction and predictive policing.
- Creating multiple models and compare them to achieve the best system in which to create a system to compute crime rates for law makers and the government to make the right changes and for police forces allocate their resources efficiently.
- Creating a user-friendly UI so that even local watch teams can use it to predict crime rate in a certain area.

1.2 Scope

The scope of this work is nationwide for the United States of America as it includes every state but may be limited when it comes to the community in which is takes place in. It focuses on socioeconomic and crime data. The scope of the crime data is violent crime however some aspects of EDA will consider non-violent crime.

1.3 Organization of the Dissertation

The dissertation will be organized as followed:

Chapter 1 will feature the introduction, research question and our aim. It will look at all our objectives ibn order to obtain the answer to our research question.

Chapter 2 will look at the theoretical background needed to understand the dissertation and all its components.

Chapter 3 will highlight the system methodology of the dissertation and outline each phase and its processes.

Chapter 4 will look at the literature review methodology and the subject area as well as looking at related papers and reveal any gaps in the literature.

Chapter 5 will show how each phase of the methodology was applied to this dissertation.

Chapter 6 will look at discussing the results as seen in the modelling and evaluation phases of the dissertations.

Chapter 7 will discuss the areas of improvement as well as areas in which can be built on for future work and its application in real life.

2 Theoretical Background

2.1 Data Mining and KDD

The definition of Big data is a collection of massive and complex data sets and data volume that include the huge quantities of data. They are large and complex sets of data that cannot be effectively processed or analysed using traditional data processing applications(Gandomi & Haider,2015).This is where data mining comes in.Data mining is a way to extract knowledge out of large datasets.It is a way in which hidden relationships can be discovered among data using artificial intelligence methods(Keyvanpour et al.,2011).Data mining is one of the 5 steps in the knowledge discovery process shown in Figure 1. The steps in KDD are as follows:

Step 1: Data Selection- This dataset is the combination of 3 different datasets. The first dataset is the socioeconomic data gathered from the 1990 Census.The second dataset is the law enforcement data from US LEMAS survey. The third dataset is the crime data from the 1995 FBI UCR .

Step 2 : Preprocessing- This step includes cleaning all the data and removing all the outliers and handling missing values.

Step 3: Transformation- This step includes dimension reduction which is the process of reducing the number of input variables. This is done using feature selection and feature engineering.

Step 4 : Data Mining - Using the preprocessed and transformed data models can be created to mine the data . In this case it will be both classification and regression models.This helps us make insights as to what the data means and how it can help.

Step 5: Interpretation/Evaluation- All the machine learning algorithms are evaluated in this section of the project. The classification algorithms will be evaluated using the accuracy,F1 score, precision and recall for each of the classifier models.The regression algorithms will be evaluated using the MSE and R2 score for each of the regressor models.

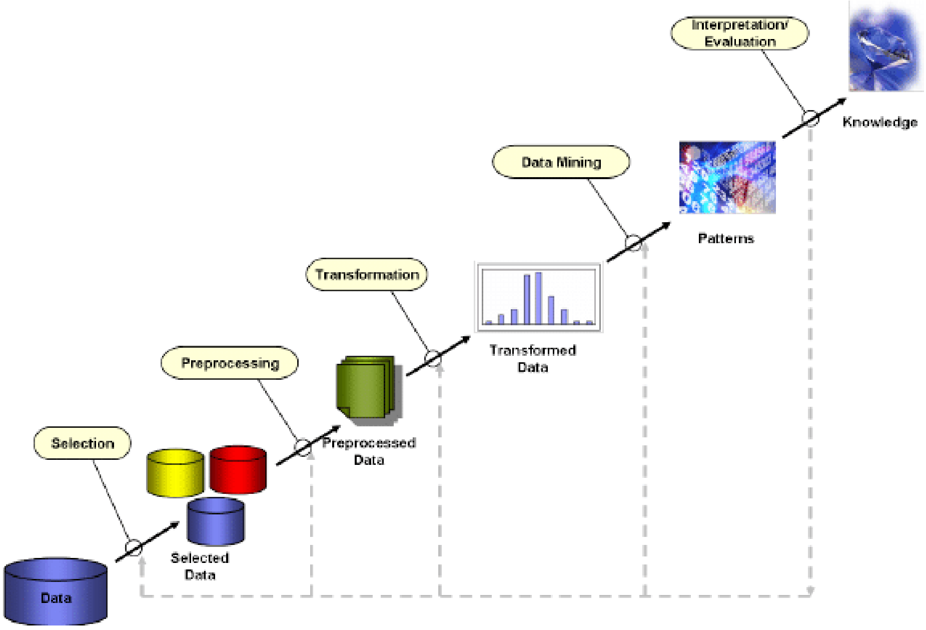


Fig 1.Diagram to show the process of KDD[Guerra-Hernandez,2008]

2.2 Statistical Tests

To understand statistical tests it is important to have an understanding of hypothesis testing. Hypothesis tests are used to assess whether a difference between two samples represents a real difference between the populations from which the samples were taken (Walker,2019). The null hypothesis is taken as a starting point and if the results are not within the limits of the null hypothesis it is stated that the null hypothesis is rejected. This applies to statistical test as they are an example of a hypothesis test. Using different statistic tests a p value is calculated. According to conventions that have been agreed upon a p value of 0.95 or more is considered the null hypothesis(Greenland et al., 2016). To reject the null hypothesis a p value of 0.05 or lower must be obtained. A p value that rejects the null hypothesis shows that it is a value with statistical significance.

2.3 Machine Learning Models

Linear Regression

The first model that is going to be discussed is linear regression because it will not be one to create a predictive model to be deployed but rather will be used to prepare the data. The main features it will be ultilised to do in this project will be statistical tests and feature selection using R2 score.

Linear Regression is when you want to predict a variable based on the value of another variable.The variable you are trying to predict is known as the dependent variable whereas the variable that will be used to affect the dependent variable is known as the independent variable (Schneider, Hommel and Blettner, 2010).It is commonly used in data analysis to see how one factor affects another as an alternative to correlation.

It is denoted by the following formula :

$$y = \alpha + \beta x \quad (1)$$

X denotes the variable which you are trying to predict, y is the response variable , β is the estimated slope and α is the estimated intercept[Schneider, Hommel and Blettner, 2010]. A good example of this is if we are trying to predict life expectancy based on alcohol consumption. X would life expectancy whereas Y would be alcohol consumption.

Linear Regression is an important tool used for feature selection(Brownlee, 2020). Using linear regression accompanied with a best line of fit one can plot the actual values vs predicted values when used in the model. It can also predict the residuals of actual values vs predicted values as will be shown in the methodology section.Using Linear Regression we can also get the p values of which will be used as a statistical test to check quality of the data.In addition to that R2 score can also be calculated to find the best features for a more complex model.

Decision Tree

The decision tree methodology is a commonly used data mining application. It is a non parametric supervised learning method meaning that can be used for both classification and regression (Matzavela and Alepis, 2021) It is formed using a hierarchy of branches. The root node refers to the attribute in which it will use to predict the classification label or continuous variable dependent on whether it is a classifier or regressor. Figure 2 shows an example in which there is a binary target variable which is continuous denoted by Y and the value is between 0 and 1 with 0 meaning no and 1 meaning yes. There are also 2 X variables which are continuous and have been scaled which means that they are between 0 and 1. The decision tree checks the conditions and then proceeds to choose an outcome denoted by $R_{x.}$

Decision Tree Parameters

This paragraph will focus on the parameters that can be optimised to help the model perform better. The first important parameter is criterion which is how the model splits the nodes. The criterion for this in classification and regression differ. In a regressor this is done using the MSE value. The regressor will choose a criteria that minimises the MSE. Classifiers have 2 potential criteria which is Gini and Entropy. Gini criteria is computed by the following formula:

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2 \ (2)$$

Sigma represents a summation operator. j in this case refers to the number of cases.P(i) refers to the probability of a specific class being classified(Dash,2022).

Entropy can be denoted by the following formula:

$$Entropy = - \sum_{i=1}^{n} p_i log_2(p_i)(\ 3)$$

Class Weights refers to the weight of each class with the class weight being between 0 and 1. Max depth refers to the longest path from the tree root to the lowest leaf (Krishnan, 2018). Max features refers to the method in which the max number of features are chosen. Max leaf nodes refer to the highest number of leaf nodes to achieve the best accuracy.Minimum impurity decrease refers to the threshold for splitting the nodes.It is usually used to modify overfitting(Gulati,2022).min_samples_leaf refers to the minimum number of samples required to be at a leaf node.min _samples_split refers to the number of samples a node must contain to consider splitting.Splitting refers to the ways in which nodes are split.

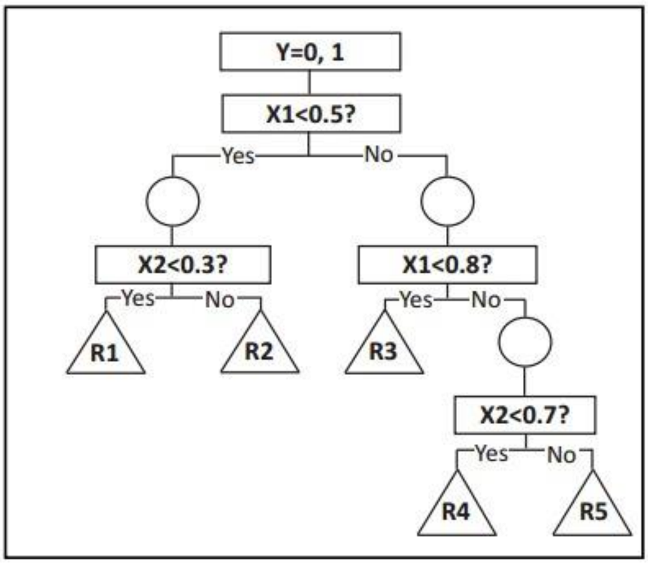


Fig 2. Diagram to show a visual representation of a decision tree [Lu and Song, 2015]

Random Forest

Random forest can be defined as an extension of decision trees. This is known as an ensemble learning method. Ensemble learning methods are made up of multiple individual models to improve their performance. A few examples of ensemble methods are bagging, boosting, stacking and voting.

Random Forest is considered an extension of Bagging as it utilises bagging and feature randomness to create an uncorrelated forest of decision trees (IBM,2023).It is commonly referred to as the random subspace name method and is essentially feature randomness. Essentially random forests randomly selects observations, builds a decision tree and the average result is taken (Donges, 2021).

Random Forest Parameters

The only parameters that differ from decision tree is n_estimators. N_estimators refer to the number of random observations taken before creating the decision tree.

3 Methodology

3.1 System Development Methodology

To succeed in creating an effective product it is important to follow a methodology that is focused on the model itself rather than the deployment and uses an iterative process. This is why the usual software development cycle cannot be used. To create a model focused product the Crisp DM methodology has been used as it uses an iterative process and is focused on creating the best possible model.

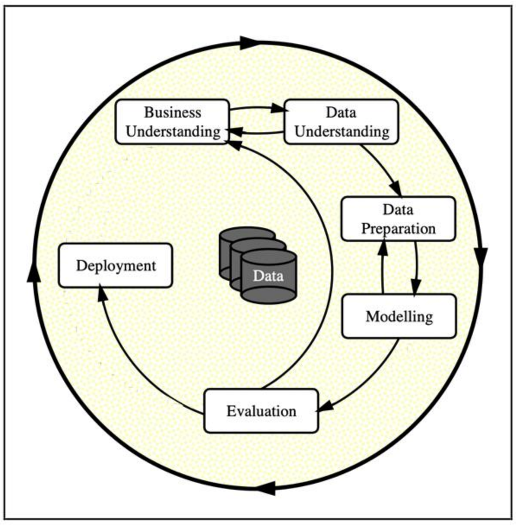
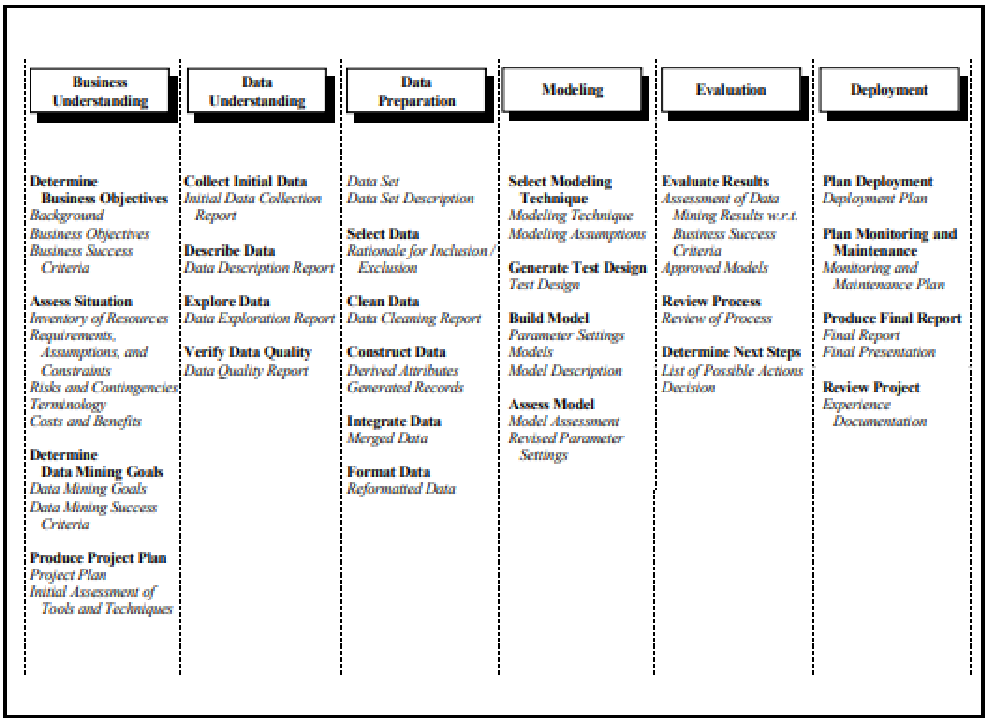


Fig 3 Diagram to show CRISP DM methodology [Wirth and Hipp, 2000]

As shown in figure 3 CRISP DM has 6 iterative phases in it's methodology. The term CRISP DM stands for 'Cross Industry Standard Process for Data Mining' and the reason it is so widely used is because it is an industry approved process used in data mining frequently(IBM,2023).It has many proven advantages such as being able to create large data mining projects that have lower costs, more reliability, more repeatable, more manageable and faster(Hipp&Wirth,2000).



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Fig 4. Diagram to show the processes and deliverables for each phase.[Wirth and Hipp, 2000]

Figure 4 shows all the objectives for each of the phases and the following description shows just how this has been altered to fit this project.

Phase 1: Business Understanding – Business Understanding deals with determining the objectives,creating a project plan,determining how to use data mining to solve these problems and ultimately creating a success criteria so that we know whether or not the project fulfilled the research objectives

Phase 2: Data Understanding -Data understanding will deal with the collection of data, describing our dataset,finding the sources of our dataset and checking the quality of data using various data analysis tools.

Phase 3:Data Preparation-Data Preparation will deal with getting the data ready for modelling . This means that the data will be cleaned, preprocessed, and transformed to get it ready to be used for modelling. Feature Selection and engineering will then be done.

Phase 4:Modelling- This section will deal with choosing an appropriate prediction model and applying it to the data. The data will be split into testing and training with a split of 70 to 30.Hyperparameters will be tuned in order to achieve the best results.

Phase 5:Evaluation-This section will focus on the evaluation metrics of each model that is created.It will also decide whether the best model fulfils the business understanding created earlier on.The performance of the models will be graphed and shown in this section in order to pick the best one for deployment.

Phase 6:Deployment-This section will focus on the deployment of the best regression and classification model. 4 models in total will be deployed. A regression and classification model will be created for the average user and a regression and classification model will created for use by police forces and town councils.

3.2 Limitations

One of the main limitations of CRISP DM is that in some environments it can be hard to ffollow linear and sequential processes when dealing with real world data analysis problems.The other limitation is that it can be hard to follow all steps and van be very tempting to skip some steps when dealing with CRISP DM(Linkedin,2023)

3.3 Literature Review Methodology

This literature review will follow a systematic literature review analysing many different things. The first part will be to research predictive policing.It will look at the history,case studies,modern applications and the ethical considerations.The importance of a lower crime rate will then be researched as well as the main predictors of crime.

It will then look at similar studies regarding crime rate prediction and predictive policing. After researching multiple papers on the use of predictive policing and crime rate prediction. Out of those papers 6 were picked for further analysis and finding gaps in the literature which will be revealed in the comparative analysis section.

4 Literature Review

4.1 What is predictive policing and how does it work in conjunction with other policing methods?

Predictive policing can be described as the approach of using historical data to predict in what geographic areas there is an increase chance for criminal activity(Ratcliffe, 2004).The ideas put forth by Robert Peel about the 3 core ideas about what the police should do prompted the concept of predictive policing.One of the core ideas stated that ‘The goal is preventing crime, not catching criminals. If the police stop crime before it happens, we don’t have to punish citizens or suppress their rights. An effective police department doesn’t have high arrest stats; its community has low crime rates.’ This shows us that crime rate prediction is also essential to determining success of a police force based on the level of crime according to the works of Robert Peel (Law Enforcement Action Partnership, 2022).

Using this information, police forces can deploy forces to prevent the criminal behaviour.

Case studies of this will be provided later. This builds on traditional and conventional policing methods rather than replacing them. The 3 main conventional methods are hotspot policing, intelligence led policing and problem oriented policing(College of Policing, 2022). Hotspot policing is the first process that will be looked at. A hotspot can be defined as a small geographic in which crime frequently occurs. As Sherman et al said an officer should be able to stand in the centre of the hotspot and see most of it with their naked eye.There are two theories that are in effect when talking about hotspot policing which are deterrence theory and crime opportunity theory(College of Policing,2022).These theories heavily influence hot-spot policing. Deterrence theory can be defined as when a potential criminal offender believes that the cost of committing a crime outweighs the potential benefits.This is done by increasing the number of visible police officers and a result deters potential offenders. The other theory is Crime Opportunity theory and can be explained using figure 3(College of Policing,2022).

This figure shows a triangle with 3 sides in which one side must be removed for the offender to get away with the crime. One of the sides is the potential offender who looks for the lowest effort and risk crime that will attain the highest possible reward. The next side is the target which can be a victim or a property depending on the crime. The last side of the triangle is what is referred to as a capable guardian(College of Policing, 2022). An example of a capable guardian would be the owner of a house as if the owner is at the house, then the chances of it getting robbed are much lower. To make the community safer the police can be used as a capable guardian in cases where the owner for example can’t be. Both these theories underpin hot spot policing. Hot spot policing has found the most success in crimes like drug offences, disorder offences, property offences and serious violence.

The second method of policing is problem-oriented policing which is also known as community policing as it alters policies to reduce crime rates. A good example of this is a case study that was conducted in which 24 violent neighbourhoods all undertook policy changes(Braga et al., 1999). The policy changes included aggressive order maintenance, the enforcement of drug law, requiring store owners to clean store fronts, removing trash from the street, investigating robberies using police resources, increased lighting in the neighbourhoods, enforcement of housing codes, erecting fences around vacant lot and cleaning vacant lot, surveillance of place using videotape, code investigation of tavern, destructions of abandoned buildings, adding trash cans, removing weapons belonging to drug dealers and helping homeless people find shelter and helping them deal with substance abuse(Braga et al,1999) . The program at the end of the day was successful and did not result in any spatial displacement meaning that the crime rate did not transfer to areas nearby. The last example of policing is intelligence led policing.

The main idea behind intelligence led policing was for police to go from reactive to proactive (Ratcliffe,2016). Reactive policing would be when an incident occurs, and the police would have to go help the people in need whereas proactive policing is when the police would leverage information to prevent an incident from happening. A good example of this is how police forces deal with counter terrorism.

4.2 Benefits and Limitations

Predictive policing especially as it stands now has positive and negative effects on society. One of the main benefits of predictive is that it has a larger long term impact on the crime rate as it is a more permanent solution depending on the strategy that is employed using predictive policing.

Another benefit is that resources can be deployed more accurately in place and time. More examples of it’s benefits can be found in the case studies section of the literature review.

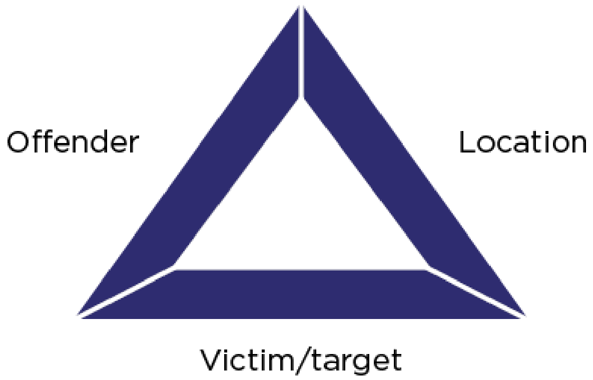


Fig 5. Deterrence Theory Diagram[College Of Policing,2022]

4.3 History of Predictive Policing

Before the age of digitisation there were theories in which crime could be predicted based on things that weren't do with research the materialisation of a crime but rather using things such as socioeconomic theories. The 3 people who shaped the idea of crime prediction were Adolphe Quetelet, André-Michel Guerry and Burgess . Guerry wrote a paper which can be translated to 'An Essay on the moral statistics of France'.The essay is considered one of the first examples of the stastical analysis of the relationship between crime and social behaviours. The paper focuses on analysing crime rates and suicide rates. Similarly to Guerry, Quetelet worked on statistical analysis of crime rates and he believed that crime rates followed predictable patterns which is one of the foundations upon which crime forecasting used for. Lastly was Burgess who laid the foundation for social disorganisation theory.

4.4 Case Studies of Predictive policing

One case of predictive policing occurred in Richmond, New York. Every year in Richmond New year's eve would see an increased amount of gunfire. To reduce this predictive policing have been put into effect. Using data gathered from previous years the police force was able to anticipate the time, location, and nature of future incidents. On New Year's Eve in 2003 police officers were placed at those locations. This ended up in resulting in a 47% decrease in random gunfire and a 246% increase in weapons seized. The department saved over$15,000 in personnel costs(Meijer & Wessels,2019). This doesn't include things like hospital fees and the hospital hours that would have to be put in by medical professionals.

Another example of this would be in London where data analytics was used to reduce crime. London was chosen as it was a city with high crime rates. Other cities included San Francisco, Chicago, and New York. When used in London it resulted in a 12% decrease in robberies, a 21% decrease in burglaries and a 32% decrease in vehicle theft(Dixon,2021).

4.5 Modern Predictive Policing Applications

The main predictive policing applications used in the US are Compstat,PredPol and ShotSpotter.

The first application is known as PredPol.PredPol is a predictive policing software that analyses historical crime data including the nature,time and location of the crimes.Predpol has had many success stories. When deployed in LA crimes reduced by 13% whereas other cities saw an increase of 0.4%. This also resulted in a 20% decrease in predicted crimes for the next year. During the initial launch in Atlanta aggregate crime decreased by 8% and 9% respectively in the 2 areas in Atlanta.As a result of it’s success it was implemented citywide and it saw an aggregate crime drop of 19%(PredPol, 2020).

The second application that will be looked at is Compstat. Compstat is an application that relies on four core concepts;timely and accurate information or intelligence,rapid deployment of resources,effective tactics and relentless follow up(Bureau of Justice Assistance,2013).Using data such as crime report data and 911 data. Using this data digital maps are created as well as dashboards to aid in predictive policing. The honus of compstat is on the police force as they are the ones who allocate the resources based on this information.

The last application is Shotspotter but has since rebranded to SoundThinking and is mainly used for being able to spot whether a gun shot has gone off based on sound.The way it works is if a gun goes SoundThinking detects it and predicts the latitude and longitude of the gun shot.SoundThinking is used in more than 85 cities in the US.This application is a way for the community to help the officers as it is an app for the public as well as the police. The one problem with it is that gun crime in the US is vastly underreported with only 1 in 5 shooting incidents being reported to the police. This creates problems for the stats regarding SoundThinking. An example of this was in New York city when only 16% of the shots caught by SoundThinking were reported(ShotSpotterFaq,2018).

4.6 The importance of having a low crime rate

It has now been shown how machine learning models can reduce crime rate and how one of the core ideas of a police force should be to have low rates of crime but that still leaves the question about why a low crime rate is needed.

A report released by the Centre for American Progress explored the direct and indirect costs of homicide in 8 metropolitan areas meaning that these were 8 of some of the more densely populated states in the US. The states were Seattle, Milwaukee, Houston, Dallas, Boston, Philadelphia, Chicago and Jacksonville and they resulted and the direct costs of violent crime for them were an average of $320 per person calculated from a total of 3.7 billion for the year for all 8 states. This has a significant impact on the housing market. The study indicates that a 10% reduction results in a 0.83% increase in housing values(Hassett & Shapiro,2012).A similar report was conducted by the Federal Reserve bank on different favelas in Rio de Janeiro in which they increased the amount of police intervention in low income neighbourhoods. Homicides went down between 10-25% depending on the area and robberies went down between 10-20% and because of this the selling price of properties went up by 5-10% (Frischtak and Mandel, 2012).

Another reason why a low crime rate is important is because it affects how well a child will do when it come to academics and progression later in life.The main reason for this is because if a student gets involved in things such as crime or is impacted by someone in their community being affected by crime the more likely they are to drop out, get expelled or go to prison. This severely hinders their ability to progress academically.(Iresearchnet)

More importantly low crime rate has many economic benefits. The study by Center for American Progress states that the estimated savings for municipal budgets from a 25 percent reduction in violent crime would result in $6 million per year in Seattle to $12 million per year in Boston and Milwaukee, to $42 million per year in Philadelphia and $59 million for Chicago(Shapiro and Hassett, 2012). It would also result in lower out pocket medical costs as many of the victims wouldn't be injured and therefore not need it.In addition to that taxes could potentially be lowered as the government would need to spend loss on law enforcement as there would be less violent crime happening. A study conducted in India reported that a 1% in homicides results in a 0.25% in economic growth(Kalluru,2023).

4.7 Main Predictors of crime

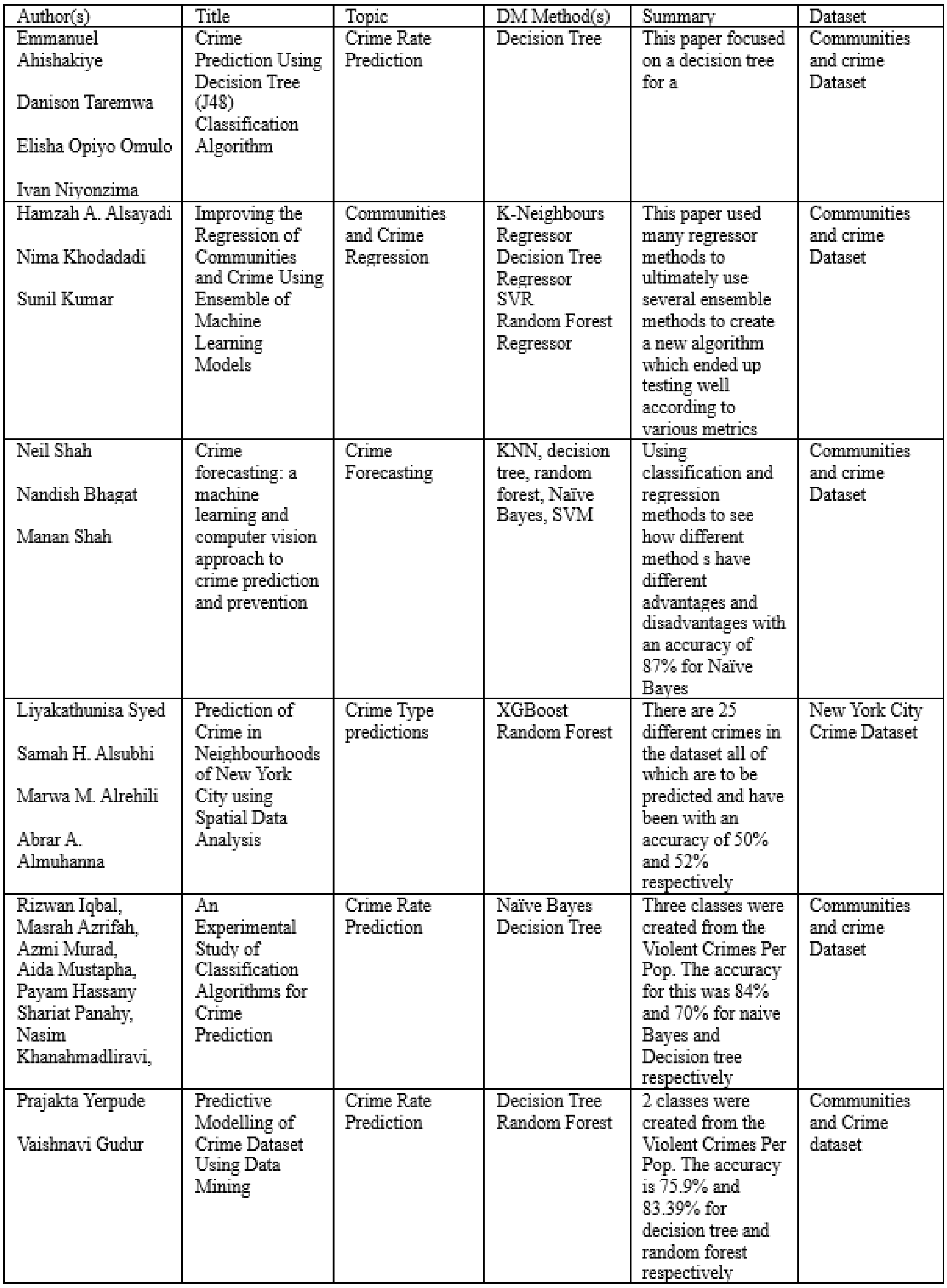
Age is an important variable to keep track of as it is believed that one of early predictors of criminal behaviour is early aggressive behaviour. In addition to that a paper published by Jeffrey Ulmer stated that age is a consistent predictor of crime both in the aggregate and for the individuals. Quetelet also proved that the proportion of the population in France involved in crime tends to peak in adolescent and early adulthood and then goes on to decline as age increases (Ulmer and Steffensmeier, 2014).It also saw that in the modern era there was a long-term trend toward younger-age crime distributions as shown by the FBI UCR report. In this report the peak age-crime involvement is for the age groups younger than 25. The median age was also shown to be younger than 30.

Another predictor of crime is education level because it is a fact that a person's lack of education often increases the likelihood in which they will be involved in crime or anti-social behaviour. This is supported by the fact that in individuals aged 15-24 that an increase in their school leaving age the arrest rate reduces by 6% showing us that the higher the school leaving age the less likely they are to commit a crime(Costa et al.,2018).In addition to that a policy brief conducted by the national bureau of economic research showed that there was sustained crime reductions between the ages 22-30 when nearly all youth are finished with their schooling (E. Jason Baron, Hyman and Vasquez, 2022). It also states that students who attended better funded schools were 15% likely to get arrested up until the age of 30 and states that the reason for this was that more funding results in more opportunities for the individuals later in life meaning that they won't delve into crime.

The next predictor of crime is unemployment. Unemployment is a big predictor of crime however when it is not as big a deal for Violent Crime as it is for Non-Violent Crime. The reason for this being that as income is low there is more reliance on things like burglaries and larceny to make money or even to survive. On the other its effect on violent crime is lower because a lot of violent crimes comes because of drug or alcohol consumption. Statistics show that roughly an average of 40% of inmates who are incarcerated for violent crimes were under the influence of alcohol at the time of the crime and had a blood alcohol level of 3 times higher than the limit. A study in 1983 featured in Sociology of Deviant Behaviour(Clinard&Meier,2011) also revealed that in the US perpetrators of Violent Crime were intoxicated 54% of the time. An international study showed that internationally this number jumps to 64% of the time. This shows that unemployment is not as heavily affected for non-violent crime as it is for violent crime as they don't have access to money that they can use for alcohol but which leads them to commit non-violent crimes such as larceny and burglaries (Alcoholchange).

4.8 Ethical Considerations

The most widespread criticism of place-based predictive policing is that it discriminates against people of color (Lum and Isaac 2016).The main reason for this is because before the use of predictive policing or any machine learning methods were used for policing, policing would be done the old fashioned way. This means that since policing back then was already racially biased as proved by the disparity of black prison population the datasets would also show this. If datasets are showing racial bias that means that the machine learning models would reflect this.



Author(s)	Title	Topic	DM Method(s)	Summary	Dataset
Emmanuel Ahishakiye Danison Taremwa Elisha Opiyo Omulo Ivan Niyonzima	Crime Prediction Using Decision Tree (J48) Classification Algorithm	Crime Rate Prediction	Decision Tree	This paper focused on a decision tree for a	Communities and crime Dataset
Hamzah A. Alsayadi Nima Khodadadi Sunil Kumar	Improving the Regression of Communities and Crime Using Ensemble of Machine Learning Models	Communities and Crime Regression	K-Neighbours Regressor Decision Tree Regressor SVR Random Forest Regressor	This paper used many regressor methods to ultimately use several ensemble methods to create a new algorithm which ended up testing well according to various metrics	Communities and crime Dataset
Neil Shah Nandish Bhagat Manan Shah	Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention	Crime Forecasting	KNN, decision tree, random forest, Naïve Bayes, SVM	Using classification and regression methods to see how different method s have different advantages and disadvantages with an accuracy of 87% for Naïve Bayes	Communities and crime Dataset
Liyakathunisa Syed Samah H. Alsubhi Marwa M. Alrehili Abrar A. Almuhanna	Prediction of Crime in Neighbourhoods of New York City using Spatial Data Analysis	Crime Type predictions	XGBoost Random Forest	There are 25 different crimes in the dataset all of which are to be predicted and have been with an accuracy of 50% and 52% respectively	New York City Crime Dataset
Rizwan Iqbal, Masrah Azrifah, Azmi Murad, Aida Mustapha, Payam Hassany Shariat Panahy, Nasim Khanahmadliravi,	An Experimental Study of Classification Algorithms for Crime Prediction	Crime Rate Prediction	Naïve Bayes Decision Tree	Three classes were created from the Violent Crimes Per Pop. The accuracy for this was 84% and 70% for naive Bayes and Decision tree respectively	Communities and crime Dataset
Prajakta Yerpude Vaishnavi Gudur	Predictive Modelling of Crime Dataset Using Data Mining	Crime Rate Prediction	Decision Tree Random Forest	2 classes were created from the Violent Crimes Per Pop. The accuracy is 75.9% and 83.39% for decision tree and random forest respectively	Communities and Crime dataset

Table 1. Literature Review for related papers[Source:Self]

4.9 Comparative Analysis

Out of all the papers researched, search terms based on crime rate prediction and predictive policing. Out of all those papers 6 papers were chosen to be further analyzed and has been done so in table 1. Out of these 6 papers 5 of them used the communities and crime dataset. Out of these datasets 3 of them were used as a classification algorithm after splitting the Violent Crimes per Pop column into classes. 2 of them split the column into 3 classes whereas one paper split them into 2.

‘Predictive Modelling of Crime Dataset Using Data Mining’ split the target variable into 2 classes high and low. Although it achieved a high accuracy there were only 2 classes, which indicates that it wouldn’t be as hard to predict. Another problem with this is the ethical implications as the features they used had to do with race, sex and how well they spoke English. This shows that there is racial bias and could affect systems in the future.

The other 2 papers didn’t use any racial features so didn’t risk having any racial bias. However, in Crime Prediction Using Decision Tree (J48) Classification Algorithm didn’t use any feature selection other than human knowledge. Both papers featured 3 classes, however the cutoff was 25% and 40% respectively. This resulted in imbalanced classes.

Improving the Regression of Communities and Crime Using Ensemble of Machine Learning Models was a good paper that created a model on its own based on an ensemble of different machine learning models and it performed well. The only problems with it were that the machine learning models were never deployed and since it is a regression model there are no real metrics as visible as accuracy.

5 Implementation

5.1 Phase 1: Business Understanding

Objectives

1. Improve Resource Allocation of police officers in the US
2. Gaining trust of the community that it is being treated as an issue and solutions are being proposed
3. Changing from reactive policing to proactive policing
4. Reducing crime in the US

Guiding Questions

The guiding questions will deal with the initial data and will be used to give us our initial thoughts about our data and will guide the data. This will also be the guidance for the exploratory data analysis.

1. Which states have the highest amount of Violent crime?
2. Which communities have the highest amount of violent crime?
3. What fields have the most effect on Violent Crime?
4. Which states have the most effect on the most important features in the dataset?
5. Which communities have the most effect on the most important features in the dataset?

5.2 Phase 2- Data Understanding

Dataset Description

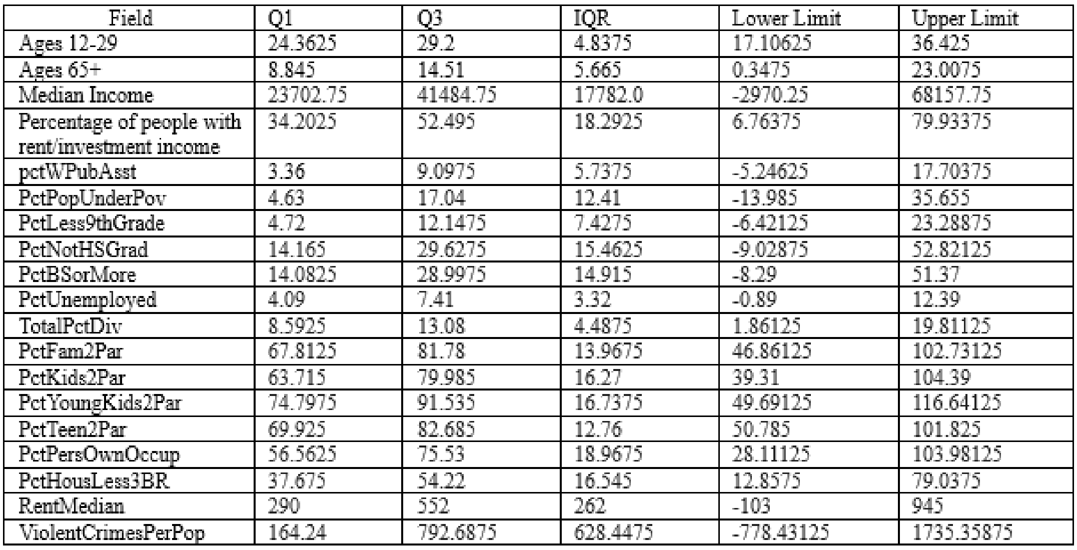
The dataset used for this project consists of socio-economic data from the 1990 US census , law enforcement data from the 1990 US LEMAS survey and crime data from the 1995 FBI UCR. Every 10 years the US conducts a census in which socio-economic data is collected from all over the US.The 1990 US LEMAS survey is used for law enforcement such as the ampunt of police budget, number of active policers etc. LEMAS stands for Law Enforcement Management and Administrative Statistics. 1995 FBI UCR is the uniform crime data collected by the FBI. It collects data for all types of crimes by state and community.The source for this dataset is the UCI Machine Learning Repository authored by Michael Redmond(Redmond,2009).

Field Name	Description
pctWInvInc	Percent of population with a 2nd income from rent or investments
pctWPubAsst	Percentage of population that receive public assistance from the government
PctPopUnderPov	Percentage of population in poverty
PctUnemployed	Percentage of population unemployed
TotalPctDiv	Percentage of population divorced
PctFam2Par	Percentage of families with 2 parents
PctKids2Par	Percentage of kids with 2 parents
PctYoungKids2Par	Percentage of young kids with 2 parents
PctTeen2Par	Percentage of teens with 2 parents
PctPersOwnOccup	Percentage of population with their own house
State	2 letter abbreviation for the states in the US
Community Name	Name of the community
medIncome	Median Income of the community
PctLess9thGrade	Percentage of people who didn’t make it past the 9th grade
PctNotHSGrad	Percentage of people who didn’t make it past high school
PctBSorMore	Percentage of people with a bachelors degree or more
agePct12t29	Percentage of people aged between 12 and 29
agePct65up	Percentage of people aged 65 and over
PctHousLess3BR	Percentage of housing units with less than 3 bedrooms
RentMedian	Median Rent in the community
ViolentCrimesPerPop	Violent Crimes per 100k population
NonViolentCrimesPerPop	Non Violent Crimes per 100k population
murdPerPop	Murders per 100k population
rapesPerPop	Rapes per 100k population
robbbPerPop	Robberies per 100k population
assaultPerPop	Assault per 100k population
burglPerPop	Burglaries per 100k population
larcPerPop	Larcenies per 100k population
autoTheftPerPop	Autotheft per 100k population
assaultPerPop	Assault per 100k population
arsonsPerPop	Arsons per 100k population

Table 2 Table showing all fields and description from communities and crime dataset[Redmond,2009]

In order to prepare the data the correlation between the columns and violent crimes per 100k population were calculated. From that and the literature review talking about the main predictors of crime 18 features were chosen to calculate the quality of data. The way we check the quality of the data is via IQR.The IQR has been calculated to spot outliers.

Table 2 refers to Q1 and Q3 where Q1 is 25th percentile and Q3 is the 75% percentile.IQR refers to Q3-Q1 which is what is shown by the values in the IQR column. Lower limit refers to the formula Q1-IQR*1.5 whereas the upper limit refers to Q3+IQR*1.5. Both these columns have been shown in the table.In order to clean the data any data points that are not between the lower and upper limit will be removed. The data points that aren’t between the lower and upper limit are the outliers. The next way in which the quality is calculated is using a histogram for frequency for all predictor variables as well as the one target variable.



Field	Q1	Q3	IQR	Lower Limit	Upper Limit
Ages 12-29	24.3625	29.2	4.8375	17.10625	36.425
Ages 65+	8.845	14.51	5.665	0.3475	23.0075
Median Income	23702.75	41484.75	17782.0	-2970.25	68157.75
Percentage of people with rent/investment income	34.2025	52.495	18.2925	6.76375	79.93375
pctWPubAsst	3.36	9.0975	5.7375	-5.24625	17.70375
PctPopUnderPov	4.63	17.04	12.41	-13.985	35.655
PctLess9thGrade	4.72	12.1475	7.4275	-6.42125	23.28875
PctNotHSGrad	14.165	29.6275	15.4625	-9.02875	52.82125
PctBSorMore	14.0825	28.9975	14.915	-8.29	51.37
PctUnemployed	4.09	7.41	3.32	-0.89	12.39
TotalPctDiv	8.5925	13.08	4.4875	1.86125	19.81125
PctFam2Par	67.8125	81.78	13.9675	46.86125	102.73125
PctKids2Par	63.715	79.985	16.27	39.31	104.39
PctYoungKids2Par	74.7975	91.535	16.7375	49.69125	116.64125
PctTeen2Par	69.925	82.685	12.76	50.785	101.825
PctPersOwnOccup	56.5625	75.53	18.9675	28.11125	103.98125
PctHousLess3BR	37.675	54.22	16.545	12.8575	79.0375
RentMedian	290	552	262	-103	945
ViolentCrimesPerPop	164.24	792.6875	628.4475	-778.43125	1735.35875

Table 3 Table that shows IQR values for each feature from communities and crime dataset[Redmond,2009**]**

Normal Distribution

Normal distribution can be defined as a hypothetical symmetrical distribution used to make comparisons.

In addition to that we can make an inference on the skweness using a frequency histogram and a q-q plot. Skewness is a way to check the distribution of data. A long tail on the right shows us that the data is skewed rightward or positive whereas if the long tail is on the left it is skewed leftward or negative.Skewness is important as it shows which way the outliers are.Figure 6 shows a frequency distribution histogram for Violent Crimes Per Pop. As can be seen in the figure it is a normal distribution as it maintains the shape of a bell curve however it is a right skewed distribution as the tail is rightward.

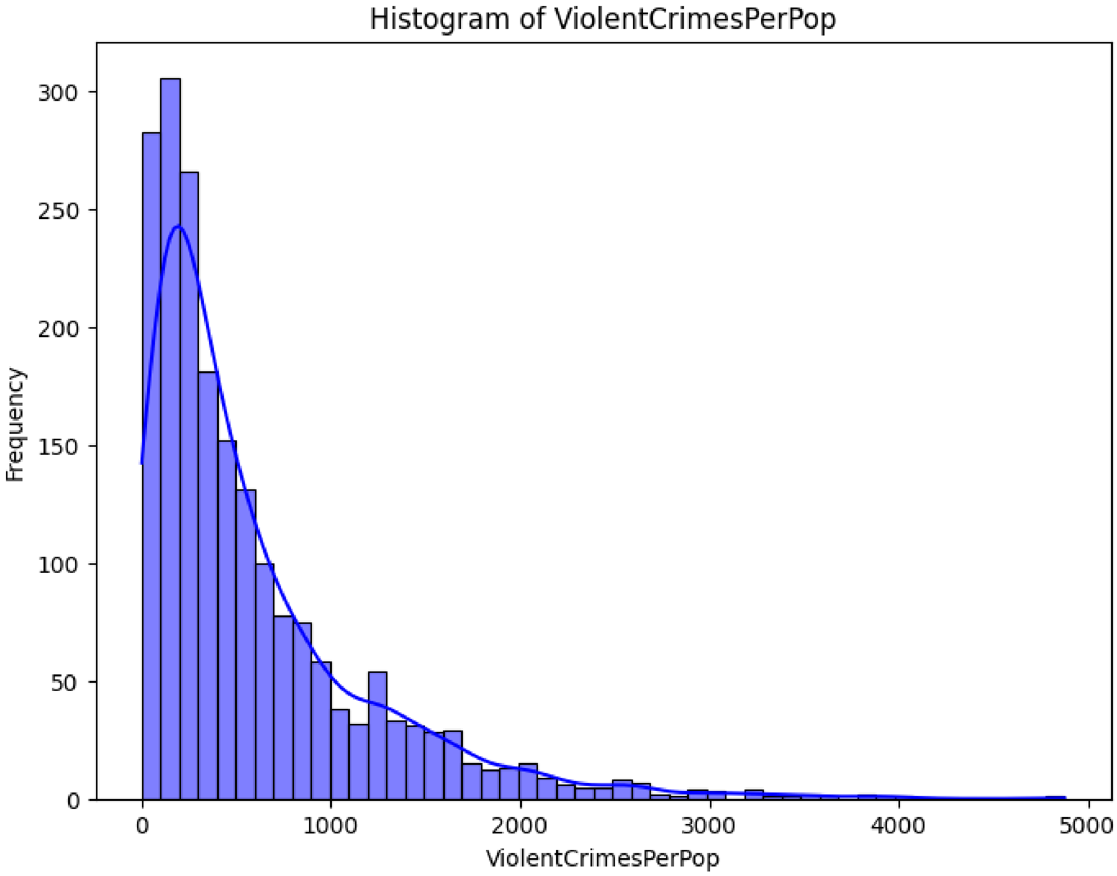


Fig6.** Histogram to show Frequency Distribution based on communities and crime dataset[Redmond,2009**]

A Q-Q plot has also been calculated for the data and using the graph you can see just how good the data is by seeing where the points land on the 45 degree angle line.The 45 degree line shows normal distribution and the more points that are on the line the closer the data is to normal distribution.Figure 7 shows the Q-Q plot for the Violent Crimes column.The dotted lines don't fit the line perfectly but you can see that there are quite a few data points close to the line.

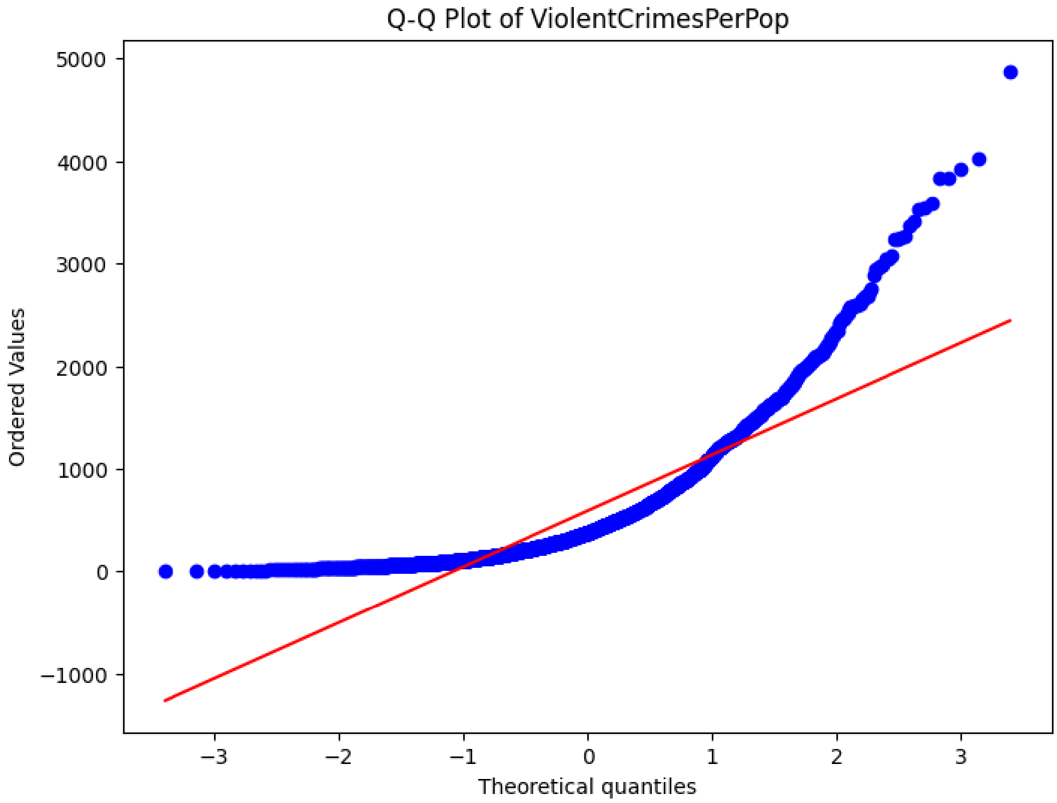


Fig7. Q-Q plot for Violent Crimes Per Pop from communities and crime dataset[Redmond,2009**]**

Linear Regression

The next step would be to use linear regression to get the R2 value, MSE value and p value. All these values have been displayed in table 4. In Layman’s terms R2 score shows us how well the variable fits the regression model. The higher the R2 score the better it does in fitting the regression model. This value will later be used to do feature selection.

The MSE value is used to determine accuracy of the model. A low MSE value means that the gap between actual predicted is low whereas a high MSE means that the gap between actual and predicted is high. The MSE has been graphed using a horizontal bar chart with the feature names on the y axis and the MSE score on the x axis. Figure 8 shows us this bar to show which features will be good for the model and which won’t.

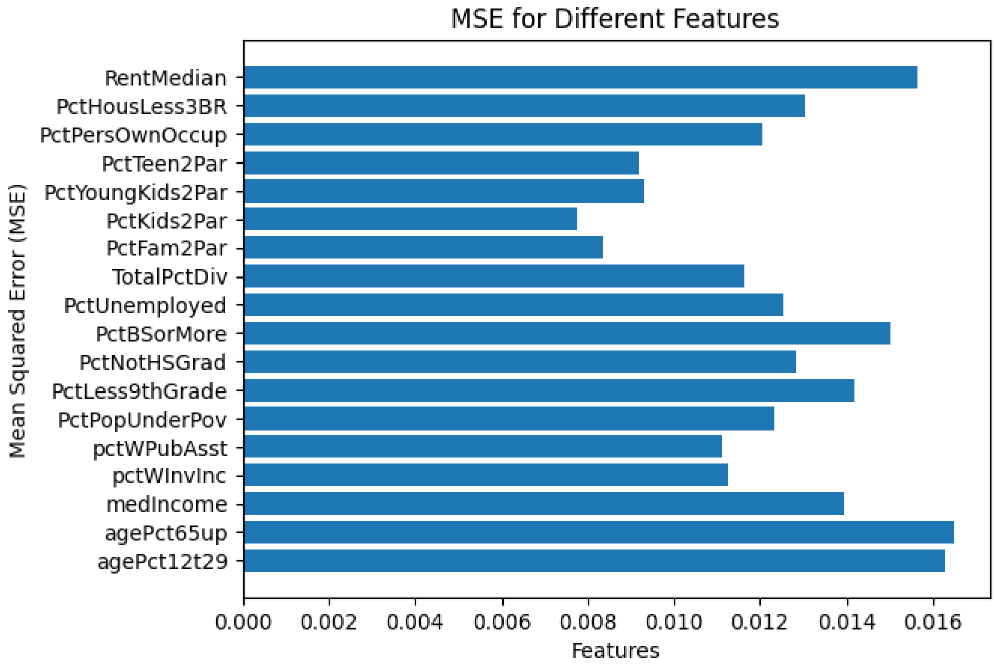


Fig 8. Horizontal Bar Chart showing the MSE for each predictor feature based on communities and crime dataset[Redmond,2009]

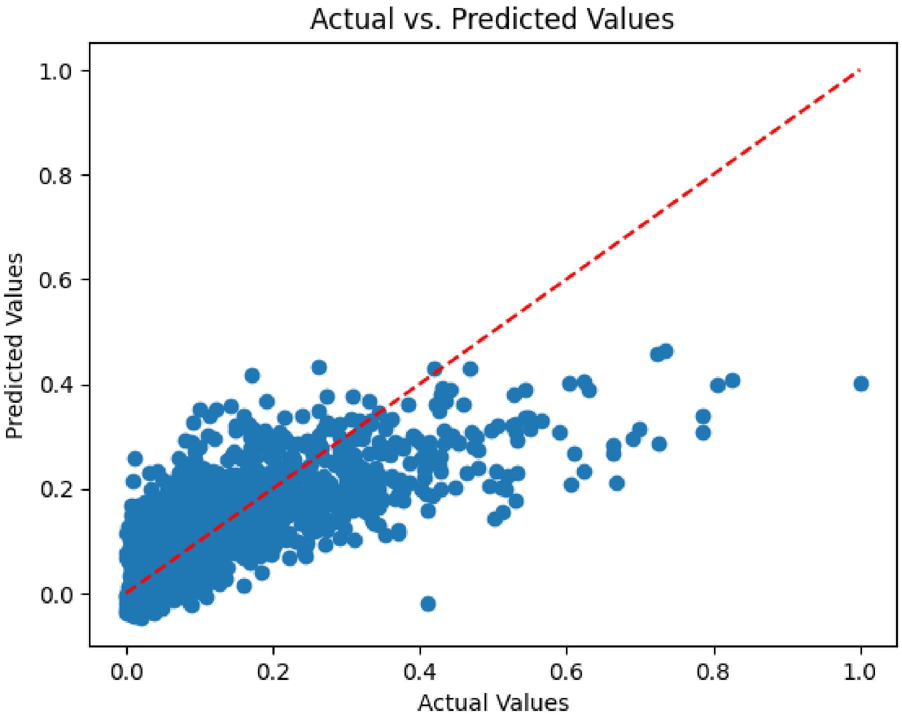


Fig 9. Actual vs Predicted Values for pctkids2par based on communities and crime dataset[Redmond,2009]

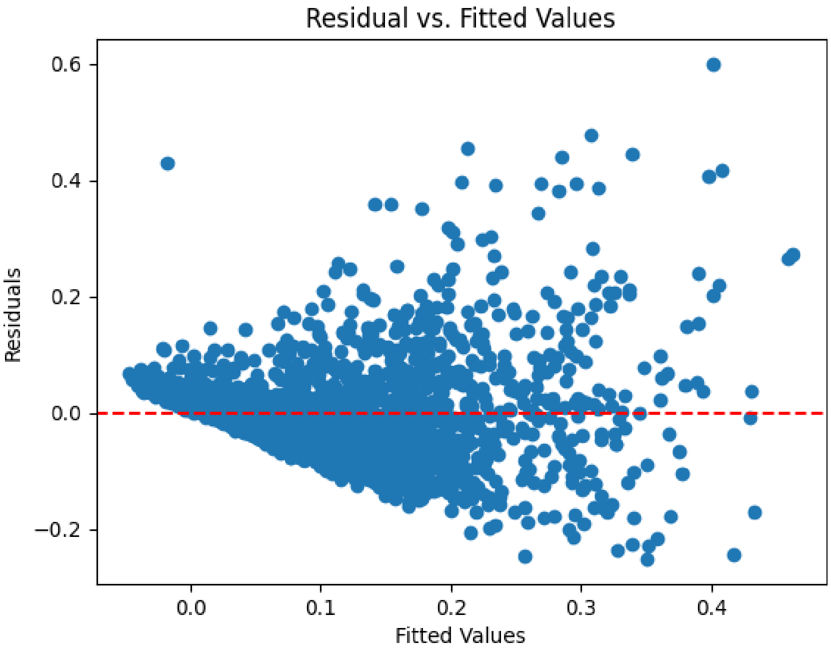


Fig 10. Residual plot for pctkids2par based on communities and crime dataset[Redmond,2009]

An Actual vs Predicted graph shows actual values on the x axis and predicted values on the y axis. It then plots a 45-degree line on the graph to show that if they are the same the dot will be placed somewhere on that line. The closer the dots are to that line the better the regression model is doing. This project uses linear regression to see how well the values predicted lie on that 45-degree line for each of the features. Figure 9 shows the Actual vs Predicted values for pctKids2par which was one of the best performing features.

A residual plot is a graph in which you compare the fitted response vs the observed response for the data points. For this project it has been done for the linear regression model so it will measure how much the regression line missed a data point. The identity line is graphed as y=0 and the higher the number of data points that are hitting this line the better. Figure 10 shows the residual vs fitted values for pctKids2par which was one of the best performing features.

Lastly using the linear regression model a p-value was able to be extracted of which can be used to see if a variable is useful to the model or whether it is just chance that it could perform well as a feature. It is conventionally accept that the null hypothesis is a p-value higher than 0.05.Regarding it can be observed that the value of p for every variable rejects the null hypothesis showing us that the data is statistically significant.

The final part of the data understanding is the EDA section in which we can answer most of the guiding questions set in the business understanding section.

Question 1 : What states have the highest levels of violent crime ?

This question can be answered using figure 11 . As can be seen on the graph the x axis Is the 2 letter abbreviation for the state and on the y axis is the Violent Crime per 100k of population for the state.In addition to that figure 12 shows an interactive map using folium which shows the violent Crime per 100k population by state when hovered on.

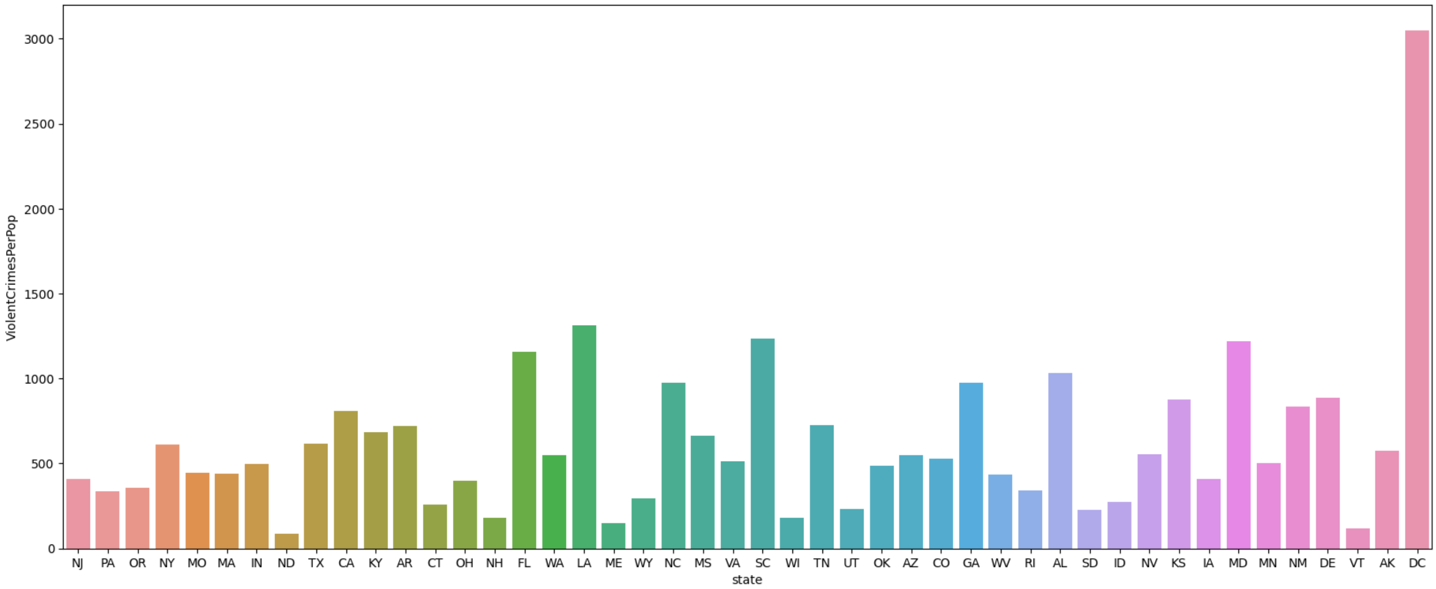


Fig 11. Graph showing Violent Crimes per Pop by State based on data analysis for communities and crime dataset[Redmond,2009]

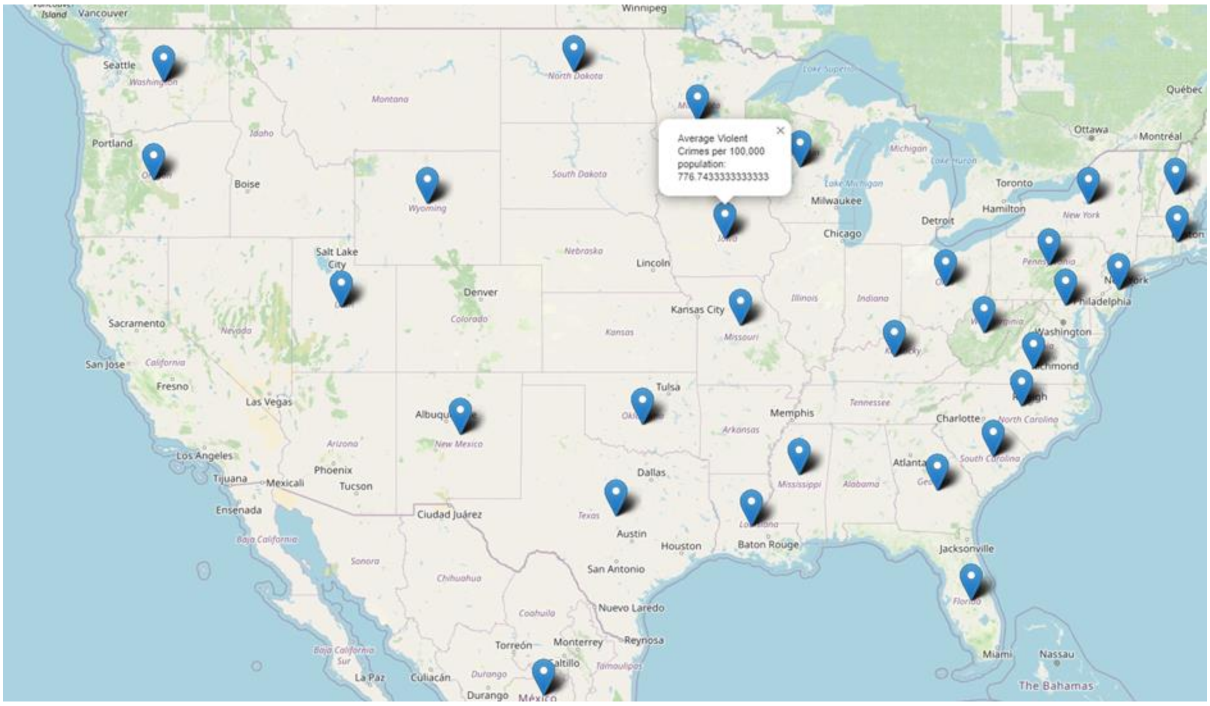


Fig12. Interactive Map showing Violent Crimes Per Pop By State based on data analysis for communities and crime dataset[Redmond,2009]

Graphs have also been created for all the other types of crimes by state and graphed in the same way as figure 11 . These graphs have been placed in the appendices.

Question 2- Which communities have the highest amount of violent crime ?

This question is answered by figure 13 showing the top 10 communities with the highest amounts of Violent Crimes.

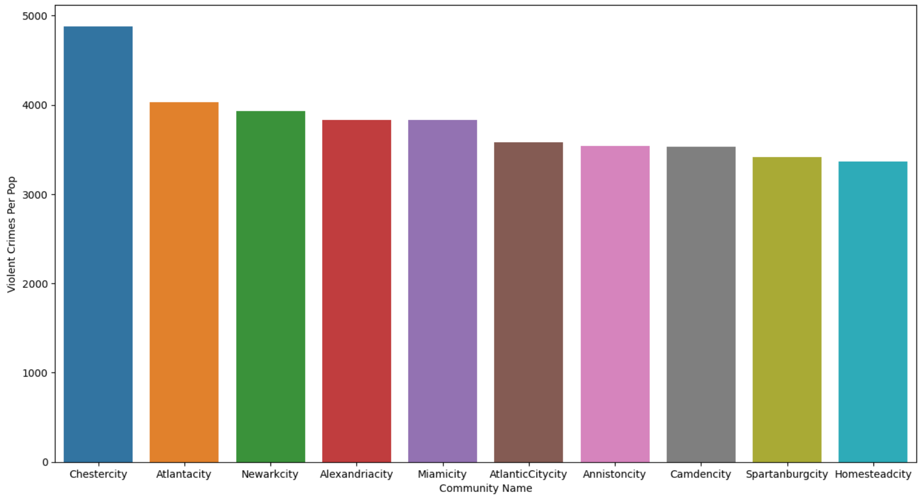


Fig13. Bar chart to show the communities with the most violent crime based on data analysis for communities and crime dataset[Redmond,2009]

The appendices also contain graphs which show the communities with highest levels of a particular crime.

Question 3 – What features have the most affect on Violent Crime?

This question will be answered using a linear regression model. Using a linear Regression model an R2 score and the MSE is outputted. The highest R2 scores will be the factors that affect Violent Crime the most.Table 4 shows the MSE and R2 score. The highlighted features are the 10 features with the highest R2 score.

Field	MSE	R2 score	p-value
Age 12 to 29	0.016297385160901668	0.014046746776201546	1.0651032257866213^{23}
agePct65up	0.016499947842029204	0.0017921836995085094	0.019044200888472133
medIncome	0.013927456146637993	0.15742184642890322	4.656233706420385^{60}
pctWInvInc	0.01125074797014979	0.3193563597562715	5.544481925676516^{120}
pctWPubAsst	0.011100360290160864	0.32845445867618794	7.668945710632465^{114}
PctPopUnderPov	0.012345774859697371	0.2531098230598754	$1.5269350833837685^{-130}$
PctLess9thGrade	0.014183249486470305	0.14194695440967975	9.420404337137267^{-51}
PctNotHSGrad	0.012829009244998371	0.22387528576734428	4.748329210159761^{-66}
PctBSorMore	0.015032154899000299	0.09059018490955151	2.1407141254872104^{-20}
PctUnemployed	0.012547793553300981	0.2408881699416594	1.0413954998216403^{-74}
TotalPctDiv	0.011640128538292025	0.2957997564047724	8.542585835992524^{109}
PctFam2Par	0.008359866169217887	0.4942478707670225	$2.3801858982072216^{-188}$
PctKids2Par	0.007749914611643792	0.5311484972516842	3.084108696524363^{-222}
PctYoungKids2Par	0.009302380551402	0.43722798002074326	5.471462522400436^{-154}
PctTeen2Par	0.009204729045782338	0.4431356651307994	$1.1506646496000673^{-155}$
PctPersOwnOccup	0.012063283890416918	0.27019985851148864	1.4511110979106658^{-98}
PctHousLess3BR	0.01303767116456287	0.21125173319204527	1.0314060004845306^{-76}
RentMedian	0.015669935259514215	0.05200598167187365	$1.2.0332806107249665^{-16}$

Table4. Table showing MSE,R2 score and p-value based on communities and crime dataset[Redmond,2009]

5.3 Phase 3- Data Preparation

Data Cleaning

The first step of data preparation is to upload the CSV file and to make sure the separators are commas and to choose what rows you want in your dataframe as the dataset has 128 columns which is more than you need. Each iteration of data preparation will have a different number of columns as some will be removed after the data understanding section and some will be added during the feature engineering step. The next step after choosing what columns you will need is to check the data types of your column. In this dataframe all the crime data is shown as a string despit the data being numbers. To rectify this the column will be changed to a numeric data type. After this all the null values will be removed and the dataframe will go from 2215 rows to 1994 rows.

The next step is to get rid of outliers as calculated in the data preparation section. The IQR was calculated for all variables and using the following formulas the upper and lower limit was calculated for each variable and subsequently any data that didn’t fall within those limits were removed.

Data Preprocessing

Feature Selection

The idea for this project was to ran 2 models side by side and see how the features affected. In this case 2 iterations of feature selections were undergone. The first was to pick the features that literature review had shown us affected the prediction of crime rate. The other iteration of feature selection was doing using the results for the linear regression model shown in table 4. The 10 features with the highest R2 score were chosen and have been highlighted in yellow in table 4. State was also used but this time was used as an encoded variable as it is a categorical variable. The 2 iterations of feature selection have been shown in table 4.The first iteration includes all the variables in that table and the second iteration includes only the ones highlighted in yellow.

Feature Engineering

The first feature that has been engineered is state. Using a label encoder a number can be assigned to each state so that it can be used as a feature in the models. Latitude and Longitude has also been engineered as a feature. This has been done by saving each abbreviation of state into a series in pandas.Using nominatim and a geolocator a function has been written in which each state has been given a set of coordinates which correspond to it’s latitude and longitude. The latitude and longitude has been mapped to the state in the original communities and crime Dataframe.

In order to create a classifier model the violent crimes per pop needs to be split into categories. The categories will be split into low,moderate and high crime rate. The threshold chosen for each category has been done so that each split has a similar number of values. If Violent Crime per Pop is lower than 200 then it will be classified as low crime rate.If Violent crime Per Pop is higher than 200 but lower than 600 it will be classified as high crime rate.

5.4 Phase 4 : Modelling

The first iteration of models were derived from our theoretical knowledge of the domain and what an end user would look for as variables.The variables are depicted in table 4 but have not been highlighted.

Following the literature review there were certain features that were just taken without running a feature importance algorithm such as correlation or checking the r2 score for the linear regression which has been done later. The first iteration followed solely on the paragraph in the literature review in which it stated many of the known predictors of crime. The predictors are shown in table 4 but only the features highlighted have been used as they had the highest R2 score and lowest MSE. These were the features used for the first iteration of the models.

Each classifier model has 4 iterations of each model. The first iteration is done using the unfiltered data using features from the literature review. The second iteration is done using features from the literature review on the unfiltered data.The third iteration is done using the features selected using the highest R2 scores for the unfiltered data.The fourth iteration is done using the features selected using the highest R2 scores for the filtered data.

Data splitting is key when it comes to implementing a model. For this project a 70/30 split has been used in which 70 percent of the data is used as training data and 30 percent is used as testing data.

Using param grid each model created will have the main parameters added to the grid and using a grid checker the ideal parameters will be identified. Another feature of grid checker is the fact that you can perform cross validation by setting the cv value which was set to 5 for this project.Each different model has been checked for their best parameters.In addition to that a confusion matrix has been created for all classifier models.

The literature review showed us many studies in which crime prediction using machine machine learning was used . In most of these studies decision tree and random forest was the base model used in many cases and why it has been used for this project.It also includes logistic regression as there were a few experiments.

Decision Tree Parameters

The best parameters were checked using the grid check function in python and by checking the following parameters the model were rerun and attained the results in the 3rd column of table.The list of parameters are as follows:

class_weight= None,
criterion ='entropy',
max_depth= 20,
max_features= 'auto',
max_leaf_nodes= 20,
min_impurity_decrease= 0.0,
min_samples_leaf= 1,
min_samples_split= 10,
splitter= 'best'

The meaning of what each of these do is discussed in the theoretical background.

Random Forest Parameters

criterion= 'gini',
max_depth = None,
max_leaf_nodes = None,
min_samples_leaf= 2,
min_samples_split= 5,
n_estimators= 100

5.5 Phase 5: Evaluation

Each model has been run 20 times altogether.10 times it was using default parameters and 10 times it was calculated using the best parameters calculated using param grid. 3 confusion matrices have been provided. One for each model using the best parameter. It has only been done on the first instance of each model to demonstrate how it can be used.

As classification is going to be used for the public and the law enforcement agencies there is more emphasis on classification which is why the model has been run on the filtered and unfiltered data. On the other hand, since Regression will most likely only be of value to the law enforcement agencies it has only been done on the uncleaned data as when cleaned too many rows were removed and as a result affect performance of the models.

The criteria for which the modelling has been done differs for classification and regression. Classification for the unfiltered data only has accuracy computed. However classification has 4 specific values outputted and are as follows as well as their formulas:

$$Accuracy = \frac{(TP + TN)}{(TP+FP+TN+FN)} \quad (4)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

$$F1\ Score = \frac{2(Precision*Recall)}{Precision+Recall} \quad (7)$$

These are the formulas for which TP is True Positives,TN is True Negatives,FP is False Positives and FN is False Negatives.True Positives refers to the predictions made that are observed as positive and are positive. TN refers to predictions that were observed as negative and ended up being negative. FP refers to predictions that were observed as negative but ended up being positive.FN refers to predictions that were observed as negative but ended up being negative.

The averages for the metrics have been added to the comparison table so that we can compare all the metrics for both classification and regression. In addition to that standard deviation has also been calculated so that it can be evaluated as to how the data appears to look.

Tables 5-8 show the Decision Tree classifier results for Accuracy, F1 score,precision and recall respectively. It shows the values for both standard and best parameters for each.

Tables 9-12 show the Random Forest classifier results for Accuracy, F1 score,precision and recall respectively. It shows the values for both standard and best parameters for each.

Table 13 shows the Logistic Regression results for Accuracy, F1 score,precision and recall respectively.

Table14 show the Decision Tree Regressor results for Mean Squared Error using standard and best parameters for each for the features chosen from the highest R2 score.

Table 15 show the random Forest Regressor results for Mean Squared Error using standard and best parameters for each for the features chosen from the highest R2 score.

Tables 16 and 17 shows the averages from the 10 runs of the models for both classification and regression respectively.

Figure 14 shows the 3x3 confusion matrix for the decision tree classifier model.

Figure 15 shows the 3x3 confusion matrix for the random forest classifier model.

Figure 16 shows the 3x3 confusion matrix for the logistic regression model.

Runs	Accuracy	Accuracy using best parameters
Run 1	0.6062992125984252	0.6272965879265092
Run 2	0.5984251968503937	0.6561679790026247
Run 3	0.6115485564304461	0.6272965879265092
Run 4	0.5879265091863517	0.6771653543307087
Run 5	0.6010498687664042	0.6535433070866141
Run 6	0.5984251968503937	0.6220472440944882
Run 7	0.5905511811023622	0.6404199475065617
Run 8	0.6036745406824147	0.6561679790026247
Run 9	0.5958005249343832	0.6325459317585301
Run 10	0.6062992125984252	0.6509186351706037

Table 5 Table showing Decision Tree Classifier showing accuracy using features from highest R2 score for uncleaned data for communities and crime dataset[Redmond,2009]

Runs	F1 Score	F1 Score using best parameters
Run 1	0.599422224879485	0.637243844431509
Run 2	0.6073398837321177	0.6414628914628916
Run 3	0.5955146424263024	0.6592676045846874
Run 4	0.609133220949215	0.6416487705514332
Run 5	0.592776430043246	0.6376851248943628
Run 6	0.569638463976032	0.6547619047619048
Run 7	0.5951613702751349	0.6637768974241687
Run 8	0.610929269253869	0.6544072816544413
Run 9	0.5998090907941892	0.6710684921962096
Run 10	0.61060568398376	0.599693309098318

Table 6 Table showing Decision Tree Classifier showing F1 Score using features from highest R2 score for uncleaned data for communities and crime dataset[Redmond,2009]

Runs	Precision	Precision using best parameters
Run 1	0.600894766343499	0.6479799065932506
Run 2	0.6105323820778804	0.6430752688172042
Run 3	0.5985726507429917	0.6631461461023328
Run 4	0.608641975308642	0.6518773148563667
Run 5	0.5943377706685932	0.6420243712893469
Run 6	0.572035142831603	0.6811203366957158
Run 7	0.5996763233315376	0.69311607224218
Run 8	0.5920300649394904	0.6566585394171601
Run 9	0.6015369620632778	0.6791656361081603
Run 10	0.6125219646891474	0.6215265601818848

Table 7 Table showing Decision Tree Classifier showing Precision using features from highest R2 score for uncleaned data for communities and crime dataset[Redmond,2009]

Runs	Recall	Recall using best parameters
Run 1	0.5981365762416191	0.631354279597752
Run 2	0.6050949924030131	0.6406779845448709
Run 3	0.5935769523511715	0.6564415175188865
Run 4	0.6096546162934607	0.6362412067669022
Run 5	0.5913963622660269	0.6344623721498414
Run 6	0.5677540695280362	0.6453730541490774
Run 7	0.5920300649394904	0.653258400068267
Run 8	0.61047903111357	0.6536338109154928
Run 9	0.5983547784274208	0.6662147991480379
Run 10	0.6093476441895507	0.5923364643342514

Table 8 Table showing Decision Tree Classifier showing Recall using features from highest R2 score for uncleaned data for communities and crime dataset[Redmond,2009]

Runs	Accuracy	Accuracy using best parameters
Run 1	0.6666666666666666	0.681704260651629
Run 2	0.6766917293233082	0.6842105263157895
Run 3	0.6867167919799498	0.6791979949874687
Run 4	0.6616541353383458	0.681704260651629
Run 5	0.6716791979949874	0.6766917293233082
Run 6	0.6842105263157895	0.6741854636591479
Run 7	0.6741854636591479	0.6691729323308271
Run 8	0.6791979949874687	0.6917293233082706
Run 9	0.6716791979949874	0.6741854636591479
Run 10	0.6741854636591479	0.6766917293233082

Table 9 Table showing Decision Tree Classifier showing Accuracy using features from highest R2 score for uncleaned data for communities and crime dataset[Redmond,2009]

Runs	Precision	Precision using best parameters
Run 1	0.6825455010938882	0.675332413906189
Run 2	0.6799855873781144	0.6917582417582416
Run 3	0.6878158336388358	0.6897643584787069
Run 4	0.6728541657132854	0.6912146760675048
Run 5	0.6809070379520255	0.6858820424858161
Run 6	0.7014989940686536	0.7034108554488988
Run 7	0.6806077694235588	0.692639017167319
Run 8	0.69585346215781	0.6944706906302809
Run 9	0.6852039773092405	0.6909372814884627
Run 10	0.6919431445747235	0.6913773542092126

Table 10 Table showing Decision Tree Classifier showing Precision using features from highest R2 score for uncleaned data for communities and crime dataset[Redmond,2009]

Runs	Recall	Recall using best parameters
Run 1	0.6764648611438032	0.6665008673680063
Run 2	0.6831031328908583	0.6810310711894654
Run 3	0.6668144256271309	0.6801994975049926
Run 4	0.6777053491607234	0.6764648611438032
Run 5	0.6746997715463204	0.6931690688951925
Run 6	0.6858017384013512	0.68351863337852
Run 7	0.678216205721708	0.68351863337852
Run 8	0.684241105970092	0.6807175129303408
Run 9	0.6899524479593514	0.6845546642292165
Run 10	0.6734592835294003	0.6764648611438032

Table 11 Table showing Decision Tree Classifier showing Recall using features from highest R2 score for uncleaned data for communities and crime dataset[Redmond,2009]

Runs	F1 score	F1 score using best parameters
Run 1	0.679095554248553	0.670227006297981
Run 2	0.6852489897608848	0.6846549174455969
Run 3	0.6694748078226919	0.6845806976381504
Run 4	0.6791981027143402	0.6803898748382963
Run 5	0.6772910485270035	0.6972916469636935
Run 6	0.6898886853284246	0.6872867120630025
Run 7	0.6811350830692958	0.6879646427016542
Run 8	0.687496784489471	0.6848847382316804
Run 9	0.6924454944615879	0.6875295545083168
Run 10	0.6779209970531413	0.6799015195553118

Table 12 Table showing Decision Tree Classifier showing F1 score using features from highest R2 score for uncleaned data for communities and crime dataset[Redmond,2009]

Metrics	Logistic Regression
Accuracy	0.6892230576441103
F1 Score	0.6766917293233082
Precision	0.6867167919799498
Recall	0.6616541353383458

Table 13 Table showing metrics for logistic regression for communities and crime dataset[Redmond,2009]

Runs	MSE	R2 Score
Run 1	0.0068872660259445394	0.4679538187049421
Run 2	0.006976283808170726	0.4610771319438406
Run 3	0.006756969026439077	0.47801935425416286
Run 4	0.006807880139491992	0.47408643469170564
Run 5	0.006902393386109898	0.46678521941187934
Run 6	0.00667071661640339	0.4846824140538153
Run 7	0.00690881592960331	0.4662890733755569
Run 8	0.006905434905442523	0.466550259888004
Run 9	0.006966346543789439	0.46994659544250594
Run 10	0.006992823865454401	0.45979940079725334

Table 14 Table showing MSE and R2 Score for Random Forest Regressor for communities and crime dataset[Redmond,2009]

Runs	MSE	R2 score
Run 1	0.013213620758994406	-0.020761567713645457
Run 2	0.013835703322099954	-0.06881788656407894
Run 3	0.012314548659610005	0.04869238910818341
Run 4	0.014466534137273047	-0.11755001408634413
Run 5	0.01265963553043859	0.02203418378436195
Run 6	0.013018280417517937	-0.005671387902974612
Run 7	0.012914333787430303	0.002358562935033004
Run 8	0.01456342968089332	-0.12503526349786287
Run 9	0.011908706874159332	0.08004395464167613
Run 10	0.014191426370946425	-0.09629774417731007

Table 15 Table showing MSE and R2 Score for Decision Tree Regressor for communities and crime dataset[Redmond,2009]

	Average MSE	Average R2 Score
Decision Tree Regressor	0.013308621953936334	0.05872629544114706
Random Forest Regressor	0.006877493024684929	0.4695189702563666

Table 16 Table for Average MSE and R2 score for regression models for communities and crime dataset[Redmond,2009]

	Average Value	Average value using best parameters
Decision Tree Classifier Accuracy	0.5999999999999999	0.6443569553805775
Random Forest Classifier Accuracy	0.67468671679198	0.6789473684210526
Decision Tree Classifier F1 score	0.5990330280313352	0.6461016121059926
Random Forest Classifier F1 score	0.6819195547475394	0.6844711310243684
Decision Tree Classifier Precision	0.5990780002996662	0.6579690152303602
Random Forest Classifier Precision	0.6859215473310135	0.6906786931640632
Decision Tree Classifier Recall	0.597582508775336	0.6806139671161862
Random Forest Classifier Recall	0.6790458321950739	0.6806139671161862
Logistic Regression Accuracy	0.6466165413533834	0.6466165413533834
Logistic Regression F1 Score	0.652912684635075	0.652912684635075
Logistic Regression Precision	0.649621212121212	0.649621212121212
Logistic Regression Recall	0.6571547177889441	0.6571547177889441

Table 17 Metrics for classification models for communities and crime dataset[Redmond,2009]

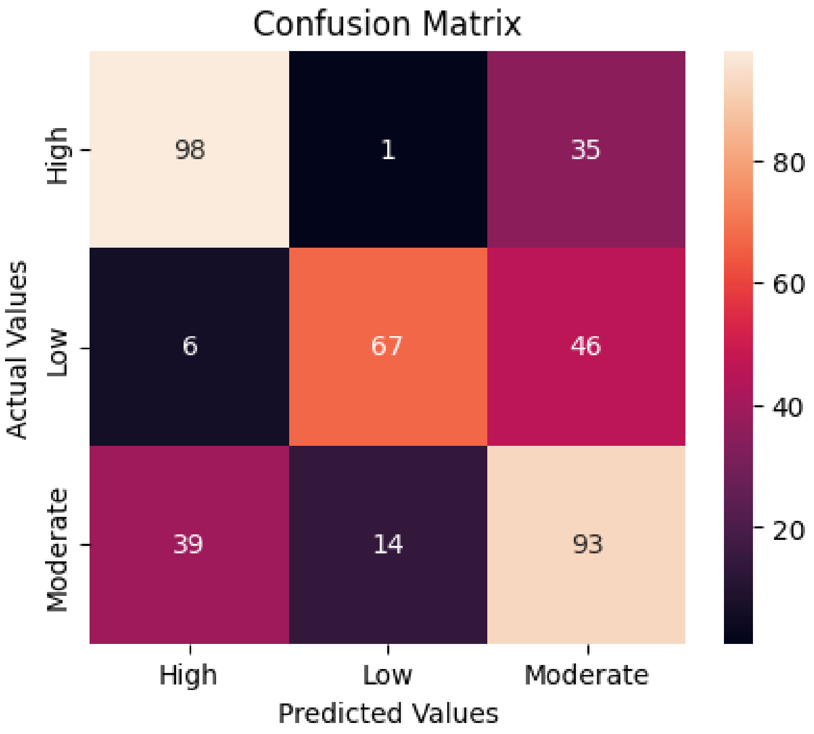


Fig 14. Confusion Matrix for Decision Tree classifier[Redmond,2009]

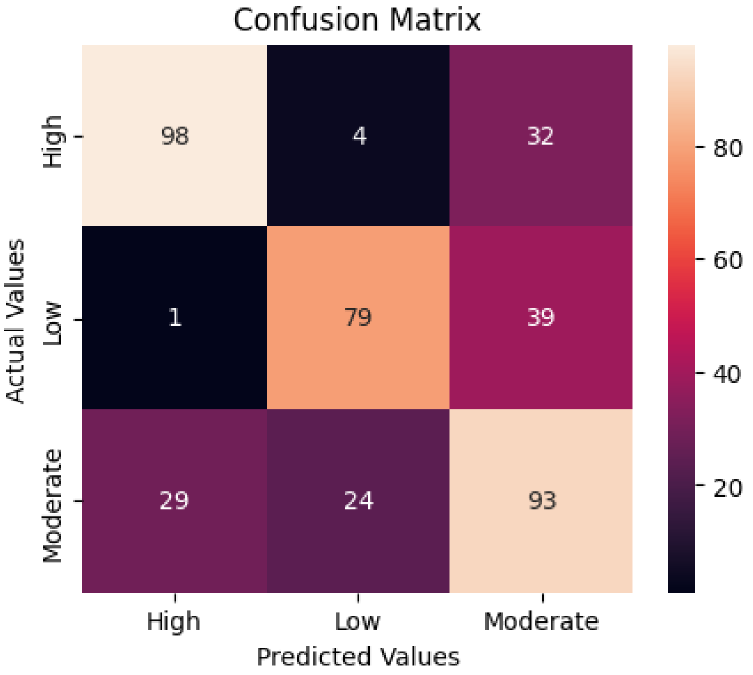


Fig 15. Confusion Matrix for Random Forest Classifier[Redmond,2009]

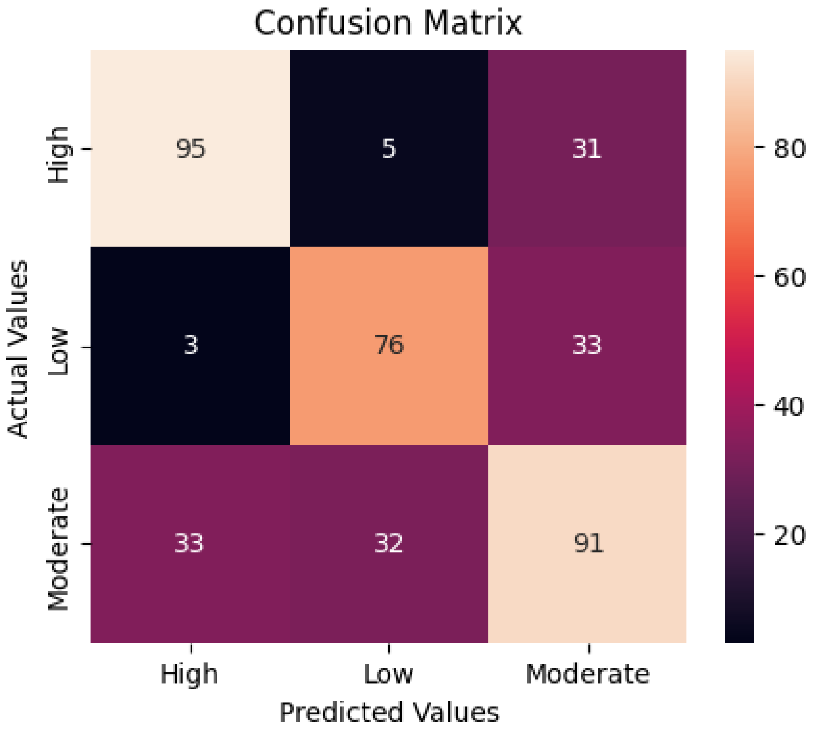


Fig 16. Confusion Matrix for Logistic Regression[Redmond,2009]

5.6 Phase 6 : Deployment

Using Streamlit in conjunction with pickle and the models that have been created 2 web applications have been created.One application has been created for classification and the other for regression. Both applications take in the same input which are the 10 features chosen in the feature selection part of the project.

Streamlit is a free and open source framework to rapidly build and share machine learning applications. It is a python based libraries designed specifically for machine learning. Pickle is the process in which complex python objects is deserialised into a byte stream when saving the model and then when loading the model the byte stream is serialised into an object again.

Figure 17 shows how a model is first trained then tested. It then shows how the model is saved using an action known as dumping. The file is then processed by pickle using the process mentioned and then loaded to complete forecasting using user input via streamlit.

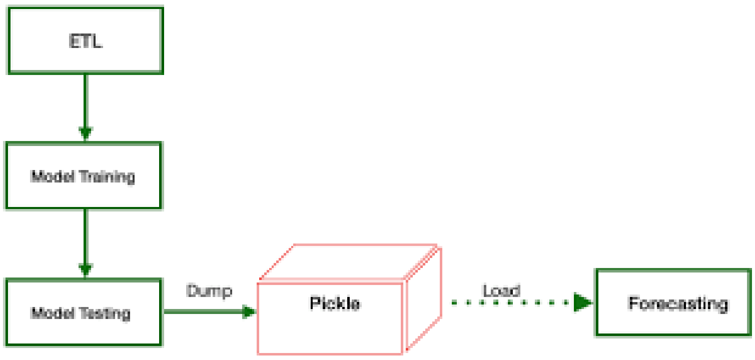
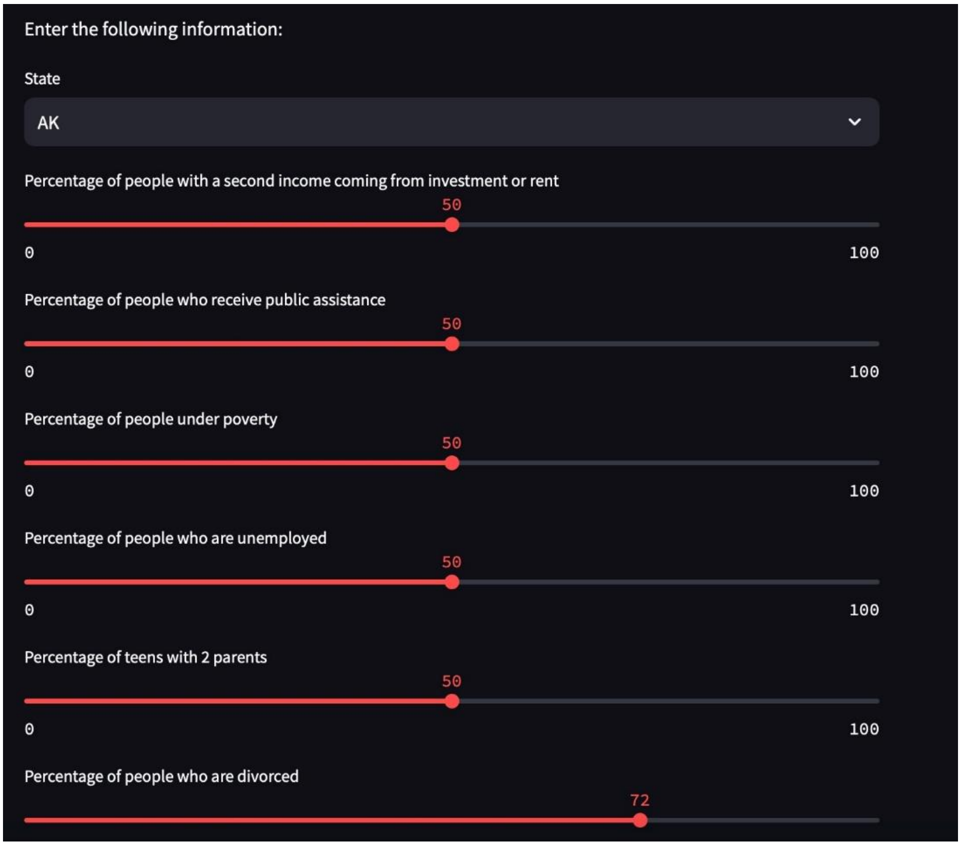


Fig 17. Figure showing a model is serialised and deserialised to deploy a model[Pisa,2021]

Figure 18 shows the first web application and as you can see there is a drop box in which the state you want to look at is chosen. It then has 9 sliders in which you set the socioeconomic variables according to the things you look to be using for the specific search. At the bottom of the screen based on all 10 user inputs you will receive an answer stating whether the model believes crime rate is low,medium or high based on your input. Figure 19 shows this.

Figure 20 on the other hand shows a similar web app however instead of the return variable being a class of low,medium or high it will show you the predicted violent crimes per 100k population but does not take state into account.The return is shown in figure 21.



Enter the following information:

State

AK

Percentage of people with a second income coming from investment or rent

50

Percentage of people who receive public assistance

50

Percentage of people under poverty

50

Percentage of people who are unemployed

50

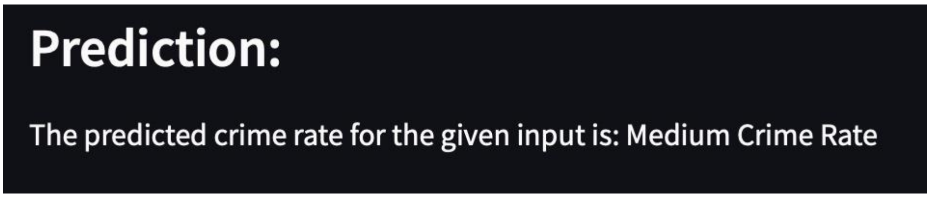
Percentage of teens with 2 parents

50

Percentage of people who are divorced

72

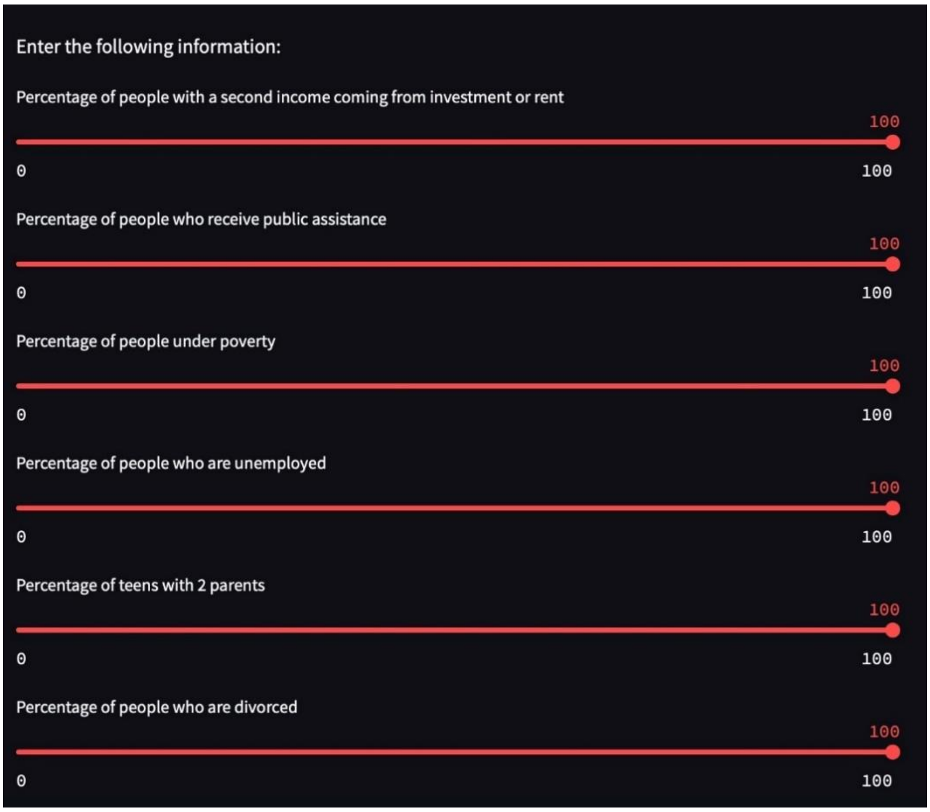
Fig 18. Figure showing screenshot of the deployment[Redmond,2009]



Prediction:

The predicted crime rate for the given input is: Medium Crime Rate

Fig 19. Figure showing screenshot of the return of the prediction[Redmond,2009]



Enter the following information:

Percentage of people with a second income coming from investment or rent

100

0 100

Percentage of people who receive public assistance

100

0 100

Percentage of people under poverty

100

0 100

Percentage of people who are unemployed

100

0 100

Percentage of teens with 2 parents

100

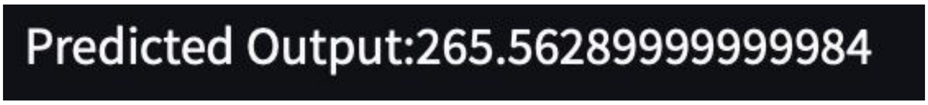
0 100

Percentage of people who are divorced

100

0 100

Fig 20. Figure showing deployment of the model[Redmond,2009]



Predicted Output:265.56289999999984

Fig 21. Figure showing screenshot of the return of the prediction[Redmond,2009]

6 Discussion of Results

When comparing all 3 classification models it is obvious that the random forest classifier performed the best as can be seen from table with an accuracy of 67.8% when compared to decision tree and logistic regression with accuracies of . It is also evident that the model answered the research question and more. It also showed us that there are multiple models that can be used to achieve similar results. Overall the model didn’t do better than some of the benchmarks set but the reasons for that is the fact that all the other examples that used this dataset and created model for crime rate prediction had class imbalance.Some chose only 2 classes whereas others chose classes where 25% of the highest Violent Crime Per Pop was considered to be low however a lot of these would be classified as low if the 25% value was used which is why for this experiment we used 10% and we got a split of 732 for moderate,670 for high and 592 for low showing that they were balanced pretty equally and still achieved an accuracy of 67.8% showing that the classification model did better than could be expected.

There were no benchmarks regarding the regression models but we could see that the random forest regressor did quite well with an average R2 score of and an average MSE score of . The decision tree regressor on the other hand did not fare well as can be seen by the extremely low R2 score.Overall this turned out okay but there is a lot of room for improvement when using it in the real world as when you predict a continuous variable for something as volatile as crime it has to be close to an exact science or it can risk casualties in the real world. However if touched up can serve the real world well as it can provide an exact number of people affected by Violent Crime per 100k population and can serve as a benchmark to see if modern applications such as predpol can help reduce the predicted number lower than what is outputted.

We also saw how much of an effect using the best parameters had on the accuracy and Mean Squared Error showing us why there is a need for the best parameters and why there is a need for cross validation.

The linear regression model also helped us to achieve feature selection by using a regression line to see how well each of the features did by computing their R2 score,MSE and p value.The R2 score and MSE proved to be vital in feature selection whereas the p value was extremely useful when doing hypothesis testing. The choice of features is very important as it helps improve the accuracy.

To check the quality of the data it is important to use methods to check the normal distribution and to do the correct hypothesis testing to find out if the data really is statistically significant. The normal distribution was checked in 2 ways. The first way was using a frequency distribution histogram and the other way was to use a Q-Q plot. As is seen by the graphs all the graphs had sufficiently fit normal distribution as it followed a bell curve shape. The Q-Q plot was checked by plotting the ordered values vs the theoretical quantities and by adding a 45-degree angle line we could check how well it was following normal distribution. As can be seen by the graphs they all follow the diagonal line to a certain extent especially for the 10 features chosen for the models.

The statistical test used in this case was computing the p value using a linear regression model. Using this method all the values were below 0.05 and thus rejecting the null hypothesis thus showing that the features were statistically significant and for the 10 features used in the models that passed the initial testing the p value was considerably below 0.05.

The reason in which only accuracy was shown for the filtered data for the features based on the literature review is because after many initial tests it was seen that it was performing worse than the unfiltered data for the features chosen using feature selection based on R2 score according to every metric.

When calculating the accuracy for the models it is important to note that not only was the accuracy computed but so was precision, recall and f1 score and on all accounts the random forest classifier performed the best and was the reason it made it to deployment.

The confusion matrix also shows us where the system was making mistakes and the way to read the confusion matrix is by first checking the top left square, the middle square, and the bottom left square. These squares are the the correctly predicted squares.The other squares are what have been predicted wrongly. For example the second square on the first row is where the system predicted that crime rate is low but in actuality it was high.

7 Conclusions and Future Work

Crime rate in the US has increased ever since the end of the COVID pandemic and the police force since then has lowered drastically in numbers calling for another way to to deal with crime in the US. The answer to this is data mining and predicting what locations require the most help and basing it off socioeconomic status to know the places in which more police is required. This machine learning application does just that by predicting the level of crime based on the state and the socioeconomic status. In this case it works to predict violent crime but making a few subtle changes can be used to predict any crime type in the US. This makes it very easy to apply in the real world whether you are dealing with violent crime or non violent crime as the need for crime rate prediction will always be there.

To begin the project we had to assess the quality of the data.The reason for this being. We did this in multiple ways. The first way was to check normal distribution put by using a histogram to show frequency distribution and using a Q-Q plot. The other way to check data quality was to check the P values using a linear regression model. P values shows us if our data is statistically signidicant.

The next step was to perform the EDA.This was done using the guiding questions as well as using our theoretical knowledge to see what inferences we could make from the data.This was done in many ways using bar charts,horizontal barcharts and even an interactive map.

Then came the Data preparation which consisted of data cleaning and data preprocessing.Data cleaning consisted of whereas data preprocessing.To start data cleaning all the columns that weren't being used had to be removed from the dataframe. The data types then had to be checked and changed to the correct data type. The last step was to remove any null values.

After that the linear regression model was run in order to check the p values,R2 score, and MSE. The R2 score and MSE was used for feature selection. The next step was to prepare the models.

The models were then run for the best features and for the features using the literature review. Both these models were run on clean and uncleaned data but it was evident that the uncleaned data was performing better for all the models under every metric. This is why the models were performed on uncleaned data using the features from the linear regression model.

Once the best model for classification and regression was chosen it was deployed to streamlit to create a machine learning application.

In the future however there are a few things that would be done differently. The first thing would be to use one hot encoder for state instead of a label encoder as the number for state could have affected the value when put into the model but if using a one hot encoder it would use either a 1 or 0. The other thing would be to see how to incorporate a time variable in order to not only get location but also to get the time in which a crime can be committed. The last change that could have been made would be to incorporate the latitude and longitude into the model that was extracted for EDA and getting the coordinates for landmarks such as schools,police stations and hospitals and see how the distance to those would affect crime rate in that city.

Overall the project was a success and could easily be implemented in the real world using socioeconomic variables to affect policy changes as well as making new laws in addition to allocating resources more efficiently not only for police officers but also for physicians.

References

1. Ahishakiye, E., Taremwa, D., Omulo, E. and Niyonzima, I. (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. International Journal of Computer and Information Technology, [online] 06(03), p.3. Available at: https://www.ijcit.com/archives/volume6/issue3/Paper060308.pdf [Accessed 21 Sep. 2023].
2. Alcohol Change UK (2020). Alcohol Statistics. [online] Alcohol Change UK. Available at: https://alcoholchange.org.uk/alcohol-facts/fact-sheets/alcohol-statistics [Accessed 24 Sep. 2023].
3. Almuhanna, A.A., Alrehili, M.M., Alsubhi, S.H. and Syed, L. (2021). Prediction of Crime in Neighbourhoods of New York City Using Spatial Data Analysis. [online] IEEE Xplore. doi:https://doi.org/10.1109/CAIDA51941.2021.9425120.
4. Alsayadi, H.A., Nima Khodadadi and Kumar, S. (2022). Improving the Regression of Communities and Crime Using Ensemble of Machine Learning Models. [online] 1(1), pp.27–34. doi:https://doi.org/10.54216/jaim.010103.
5. American Medical Association. (2023). Measuring and addressing physician burnout. [online] Available at: https://www.ama-assn.org/practice-management/physician-health/measuring-and-addressing-physician-burnout#:~:text=Physician%20burnout%20rates%20spike%20to%2063%25%20in%202021 [Accessed 10 Sep. 2023].
6. Bandekar, S.R. and Vijayalakshmi, C. (2020). Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India. Procedia Computer Science, [online] 172, pp.122–127. doi:https://doi.org/10.1016/j.procs.2020.05.018.
7. Braga, A., Papachristos, A. and Hureau, D. (2012). Hot Spots Policing Effects on Crime. Campbell Systematic Reviews, [online] 8(1), pp.1–96. doi:https://doi.org/10.4073/csr.2012.8.
8. Braga, A.A., Turchan, B., Papachristos, A.V. and Hureau, D.M. (2019). Hot spots policing of small geographic areas effects on crime. Campbell Systematic Reviews, [online] 15(3). doi:https://doi.org/10.1002/cl2.1046.
9. Braga, Anthony.A., Weisburd, David.L., Waring, Elin.J., Mazerolle, Lorraine.G., Spellman, W. and Gajewski, F. (1999). Problem-Oriented Policing in Violent Crime Places: a Randomized Controlled Experiment. [online] National Institute of Justice. Available at: https://nij.ojp.gov/library/publications/problem-oriented-policing-violent-crime-places-randomized-controlled [Accessed Sep. 2023].
10. Brownlee, J. (2020). How to Perform Feature Selection for Regression Data. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/feature-selection-for-regression-data/ [Accessed 11 Sep. 2023].
11. Bureau of Justice Assistance (2013). Compstat:Its Origins,Evolution and Future in Law Enforcement Agencies. [online] Bureau of Justice Assistance. Available at: https://bja.ojp.gov/sites/g/files/xyckuh186/files/Publications/PERF-Compstat.pdf [Accessed 20 Sep. 2023].
12. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y. and Chau, M. (2004). Crime data mining: a general framework and some examples. Computer, [online] 37(4), pp.50–56. doi:https://doi.org/10.1109/mc.2004.1297301.
13. College of Policing (2022). Hot Spot Policing. [online] College of Policing. Available at: https://www.college.police.uk/guidance/hot-spots-policing [Accessed 15 Sep. 2023].
14. Communications and November (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data Knowledge Discovery in Databases creates the context

for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information. Communications of the ACM, [online] 39(11), pp.27–34. Available at: https://sceweb.uhcl.edu/boetticher/ML_DataMining/p27-fayyad.pdf [Accessed 21 Sep. 2023].

15. Dash, S. (2022). Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning.. [online] Medium. Available at: https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c [Accessed 14 Sep. 2023].
16. Donges, N. (2021). A Complete Guide to the Random Forest Algorithm. [online] Built in. Available at: https://builtin.com/data-science/random-forest-algorithm [Accessed 15 Sep. 2023].
17. E. Jason Baron, Hyman, J.E. and Vasquez, B. (2022). Public School Funding, School Quality, and Adult Crime. National Bureau of Economic Research. [online] doi:https://doi.org/10.3386/w29855.
18. Frischtak, C. and Mandel, B.R. (2012). Crime, House Prices, and Inequality: The Effect of UPPs in Rio. SSRN Electronic Journal. [online] doi:https://doi.org/10.2139/ssrn.1995795.
19. Gandomi, A. and Haider, M. (2015). Beyond the Hype: Big Data Concepts, Methods, and Analytics. International Journal of Information Management, [online] 35(2), pp.137–144. doi:https://doi.org/10.1016/j.ijinfomgt.2014.10.007.
20. GLESS, S. (2018). PREDICTIVE POLICING.: IN DEFENCE OF ‘TRUE POSITIVES’. [online] JSTOR. Available at: https://www.jstor.org/stable/j.ctvhrd092.14 [Accessed 15 Sep. 2023].
21. Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. and Altman, D.G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology, [online] 31(4), pp.337–350. doi:https://doi.org/10.1007/s10654-016-0149-3.
22. Guerra-Hernandez, A. (2008). Steps in the KDD process[. [Image] Available at: https://www.researchgate.net/figure/Steps-in-the-KDD-process_fig1_236373188 [Accessed 10 Sep. 2023].
23. Gulati, H. (2022). Hyperparameter Tuning in Decision Trees and Random Forests. [online] Engineering Education (EngEd) Program | Section. Available at: https://www.section.io/engineering-education/hyperparmeter-tuning/#:~:text=min_impurity_decrease [Accessed 27 Sep. 2023].
24. Huber, S., Wiemer, H., Schneider, D. and Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. Procedia CIRP, [online] 79, pp.403–408. doi:https://doi.org/10.1016/j.procir.2019.02.106.
25. IBM (2023). What is Random Forest? | IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/random-forest [Accessed 15 Sep. 2023].
26. Ishwarappa and Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science, [online] 48(48), pp.319–324. doi:https://doi.org/10.1016/j.procs.2015.04.188.
27. Jason Baron, E., Hyman, J., Vasquez, B., Arcidiacono, P., Bayer, P., Brunner, E., Bulman, G., Chyn, E., Edmonds, E., Goldstein, E., Gross, M., Hendren, N., Jackson, K., Jacob, B., Kapustin, M., Lafortune, J., Carlos, J. and Serrato, S. (2022). NBER WORKING PAPER SERIES PUBLIC SCHOOL FUNDING, SCHOOL QUALITY, AND ADULT CRIME We received valuable feedback from. [online] Available at: https://www.nber.org/system/files/working_papers/w29855/w29855.pdf [Accessed 25 Sep. 2023].

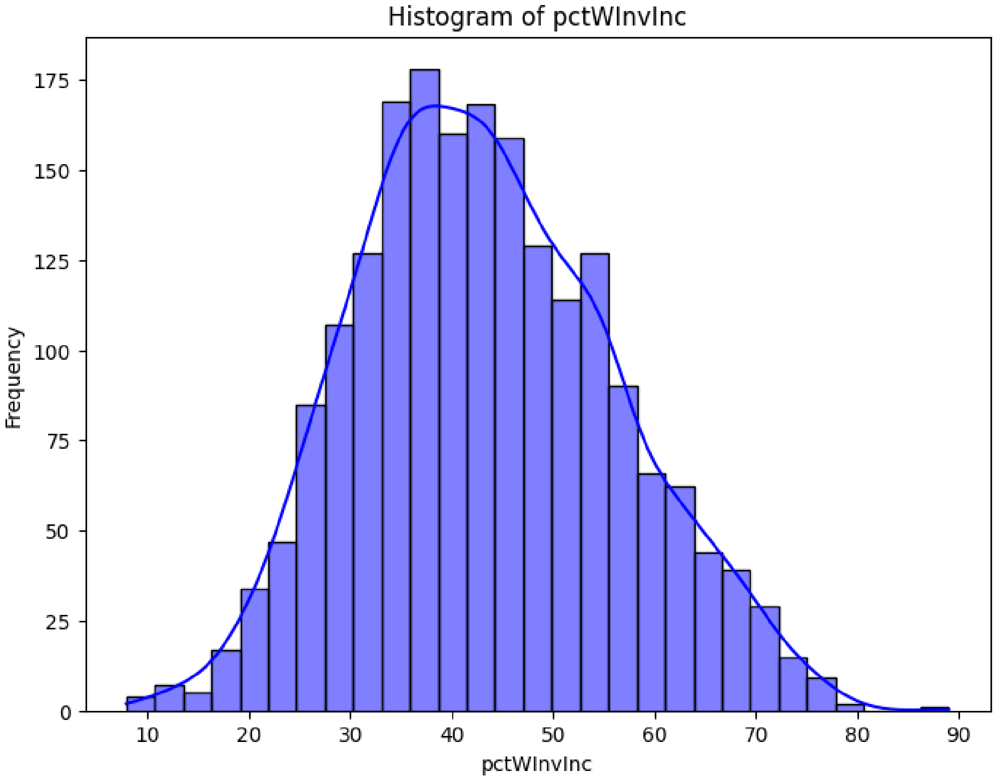
28. Jiang, T., Gradus, J.L. and Rosellini, A.J. (2020). Supervised Machine Learning: A Brief Primer. Behavior Therapy, [online] 51(5), pp.675–687. doi:https://doi.org/10.1016/j.beth.2020.05.002.
29. Keyvanpour, M.R., Javideh, M. and Ebrahimi, M.R. (2011). Detecting and investigating crime by means of data mining: a general crime matching framework. Procedia Computer Science, [online] 3, pp.872–880. doi:https://doi.org/10.1016/j.procs.2010.12.143.
30. Kirk, D.S. and Sampson, R.J. (2012). Juvenile Arrest and Collateral Educational Damage in the Transition to Adulthood. Sociology of Education, [online] 86(1), pp.36–62. doi:https://doi.org/10.1177/0038040712448862.
31. Krishnan, H. (2018). Maximum Depth of a Binary Tree. [online] Medium. Available at: https://medium.com/@harycane/maximum-depth-of-a-binary-tree-609d129fa571#:~:text=its%20maximum%20depth.- [Accessed 14 Sep. 2023].
32. Law Enforcement Action Partnership (2022). Sir Robert Peel's Policing Principles. [online] Law Enforcement Action Partnership. Available at: https://lawenforcementactionpartnership.org/peel-policing-principles/ [Accessed 15 Sep. 2023].
33. Lu, Y. and Song, Y.-Y. (2015). Decision tree methods: applications for classification and prediction. [Image] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/ [Accessed 9 Spring 2023].
34. Matzavela, V. and Alepis, E. (2021). Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. Computers and Education: Artificial Intelligence, [online] 2, p.100035. doi:https://doi.org/10.1016/j.caeai.2021.100035.
35. McClendon, L. and Meghanathan, N. (2015). Using Machine Learning Algorithms to Analyze Crime Data. Machine Learning and Applications: an International Journal, [online] 2(1), pp.1–12. doi:https://doi.org/10.5121/mlaij.2015.2101.
36. Meijer, A. and Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks. International Journal of Public Administration, [online] 42(12), pp.1031–1039. doi:https://doi.org/10.1080/01900692.2019.1575664.
37. PredPol. (2020). Proven Results of our Predictive Policing Software. [online] Available at: https://www.predpol.com/results/#:~:text=During%20Atlanta [Accessed 20 Sep. 2023].
38. Raphael, S. and Winter-Ebmer, R. (2001). Identifying the Effect of Unemployment on Crime. The Journal of Law and Economics, [online] 44(1), pp.259–283. doi:https://doi.org/10.1086/320275.
39. Ratcliffe, J. (2004). The Hotspot Matrix: a Framework for the Spatio-Temporal Targeting of Crime Reduction. Police Practice and Research, [online] 5(1), pp.5–23. doi:https://doi.org/10.1080/1561426042000191305.
40. Redmond, M. (2009). UCI Machine Learning Repository. [online] archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/dataset/183/communities+and+crime [Accessed 27 Sep. 2023].
41. Schneider, A., Hommel, G. and Blettner, M. (2010). Linear Regression Analysis. Deutsches Aerzteblatt Online, [online] 107(44), pp.776–782. doi:https://doi.org/10.3238/arztebl.2010.0776.
42. Schröer, C., Kruse, F. and Gómez, J.M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, [online] 181, pp.526–534. doi:https://doi.org/10.1016/j.procs.2021.01.199.
43. Shah, N., Bhagat, N. and Shah, M. (2021). Crime forecasting: a Machine Learning and Computer Vision Approach to Crime Prediction and Prevention. Visual Computing for In-

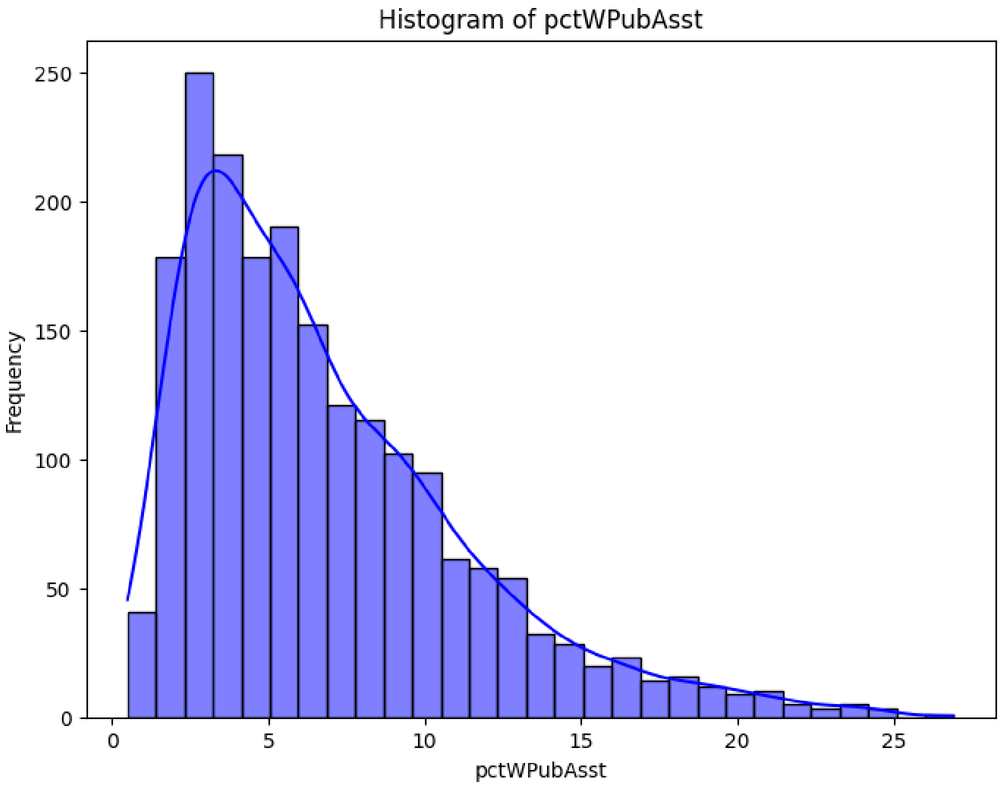
dustry, Biomedicine, and Art, [online] 4(1). doi:https://doi.org/10.1186/s42492-021-00075-z.

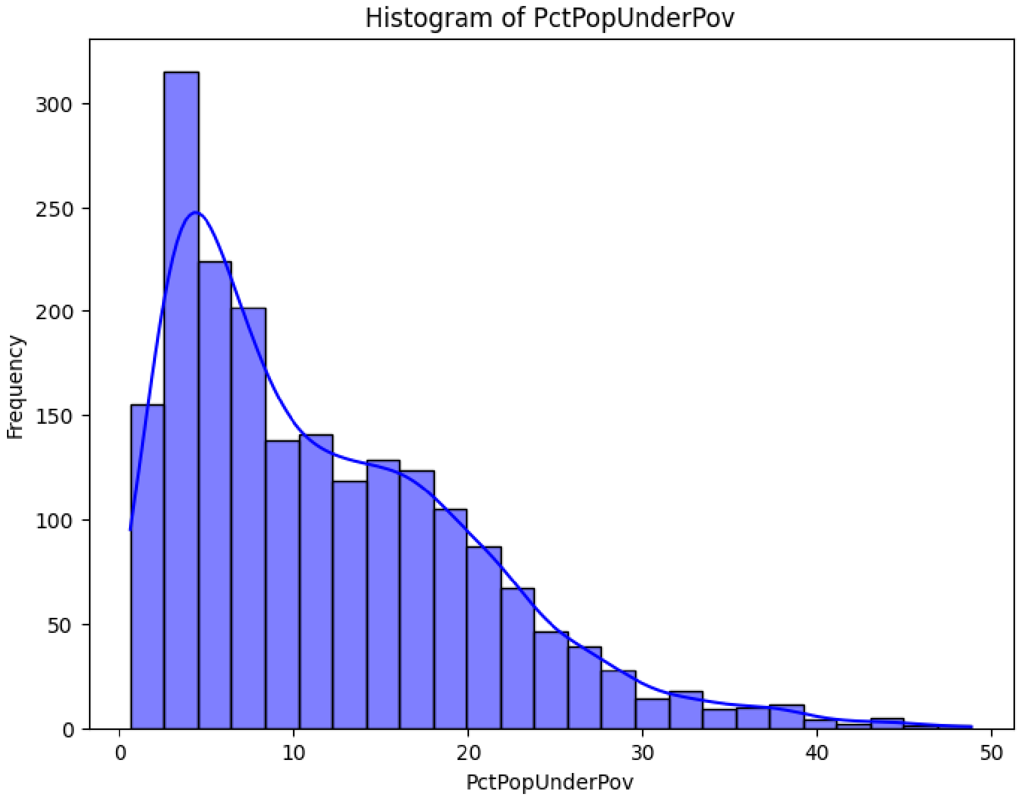
44. Shapiro, Robert.J. and Hassett, Kevin.A. (2012). The Economic Benefits of Reducing Violent Crime. [online] Center for American Progress. Available at: https://www.americanprogress.org/article/the-economic-benefits-of-reducing-violent-crime/ [Accessed 15 Sep. 2023].
45. ShotSpotter Frequently Asked Questions. (2018). Available at: https://www.shotspotter.com/system/content-uploads/SST_FAQ_January_2018.pdf [Accessed 15 Sep. 2023].
46. Taylor, S. (2020). R-Squared. [online] Corporate Finance Institute. Available at: https://corporatefinanceinstitute.com/resources/data-science/r-squared/ [Accessed 23 Sep. 2023].
47. Travis, J., Sherman, L., Gottfredson, D., Mackenzie, D., Eck, J., Reuter, P. and Bushway, S. (1998). Preventing Crime: What Works, What Doesn't, What's Promising. [online] Available at: https://www.ojp.gov/pdffiles/171676.pdf [Accessed 19 Sep. 2023].
48. Ulmer, J.T. and Steffensmeier, D.J. (2014). The age and crime relationship: Social variation, social explanations. The Nurture versus Biosocial Debate in Criminology: on the Origins of Criminal Behavior and Criminality, [online] pp.377–396. doi:https://doi.org/10.4135/9781483349114.n23.
49. Walker, J. (2019). Hypothesis tests. BJA Education, [online] 19(7), pp.227–231. doi:https://doi.org/10.1016/j.bjae.2019.03.006.
50. Wesolowski, B. and Thompson, D.J.M. (2018). Normal Distribution. The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. [online] doi:https://doi.org/10.4135/9781506326139.n476.
51. Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. [Image] Available at: http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf [Accessed 15 Sep. 2023].
52. www.ibm.com. (2021). CRISP-DM Help Overview. [online] Available at: https://www.ibm.com/docs/no/spss-modeler/18.0.0?topic=dm-crisp-help-overview [Accessed 20 Sep. 2023].
53. www.linkedin.com. (2023). What are the benefits and challenges of using a framework like CRISP-DM for data analysis? [online] Available at: https://www.linkedin.com/advice/0/what-benefits-challenges-using-framework-like#:~:text=However%2C%20using%20a%20framework%20like [Accessed 28 Sep. 2023].
54. Yang, F. (2019). Predictive Policing. Oxford Research Encyclopedia of Criminology and Criminal Justice. [online] doi:https://doi.org/10.1093/acrefore/9780190264079.013.508.
55. Yerpude, P. (2020). Predictive Modelling of Crime Data Set Using Data Mining. [online] Social Science Research Network. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3656953 [Accessed 25 Sep. 2023].
56. Yerpude, P. and Gudur, V. (2017). Predictive Modelling of Crime Dataset Using Data Mining. International Journal of Data Mining & Knowledge Management Process, [online] 7(4), pp.43–58. doi:https://doi.org/10.5121/ijdkp.2017.7404.

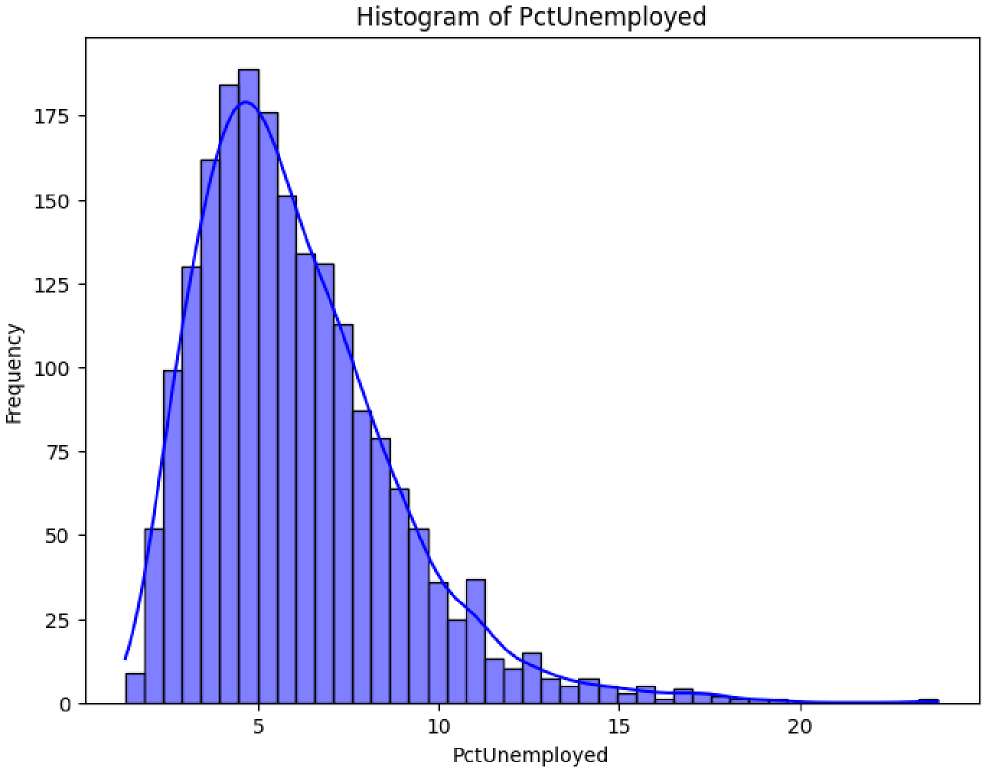
    Zhou Wang and Bovik, A.C. (2009). Mean Squared error: Love It or Leave it? a New Look at Signal Fidelity Measures. IEEE Signal Processing Magazine, [online] 26(1), pp.98–117. doi:https://doi.org/10.1109/msp.2008.930649.

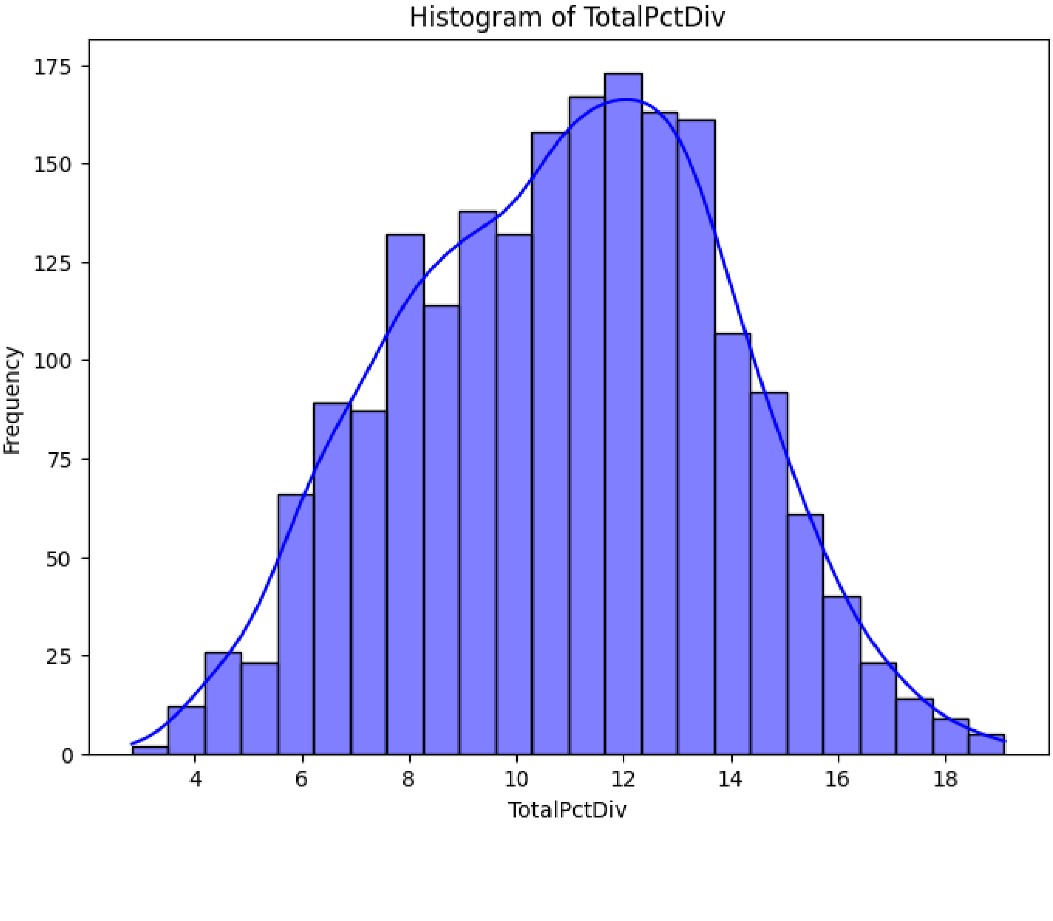
Appendices

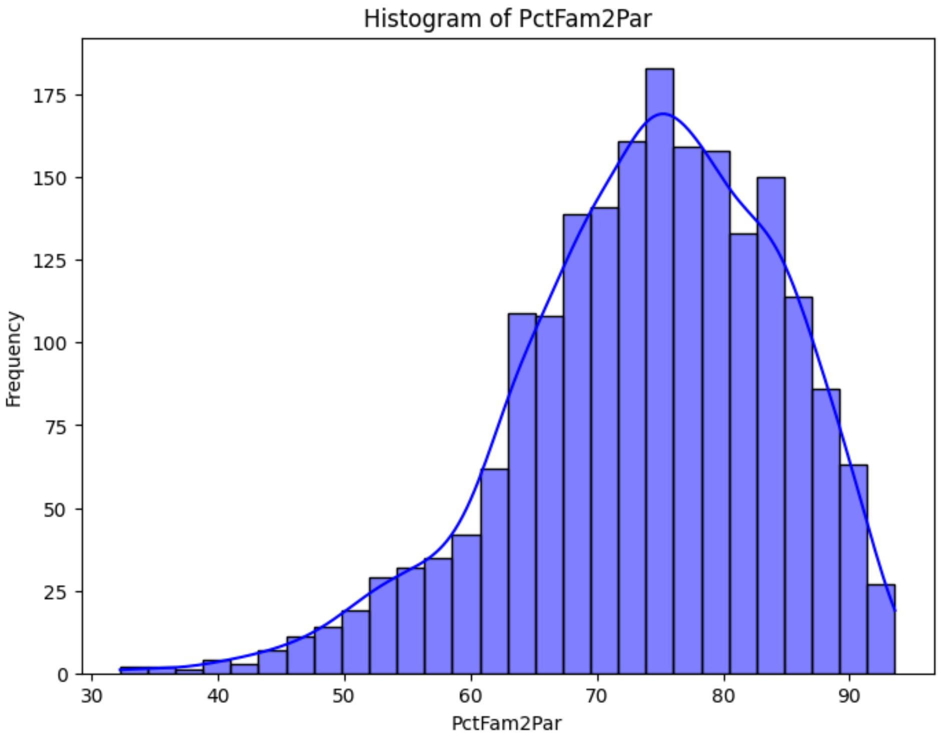

Histogram of pctWInvInc
Frequency
0
25
50
75
100
125
150
175
10
20
30
40
50
60
70
80
90
pctWInvInc

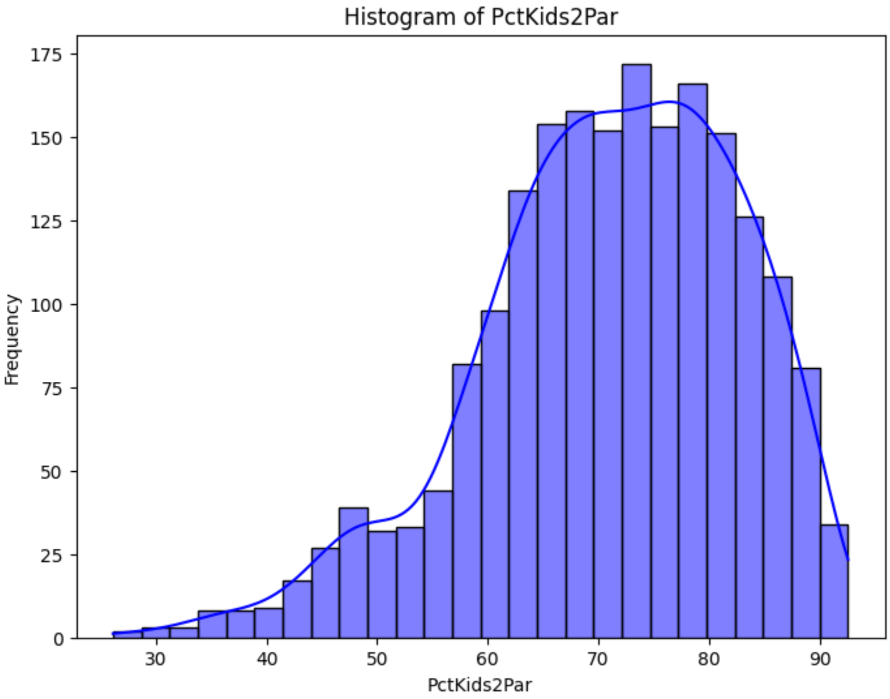

Histogram of pctWPubAsst
Frequency
0
50
100
150
200
250
0
5
10
15
20
25
pctWPubAsst

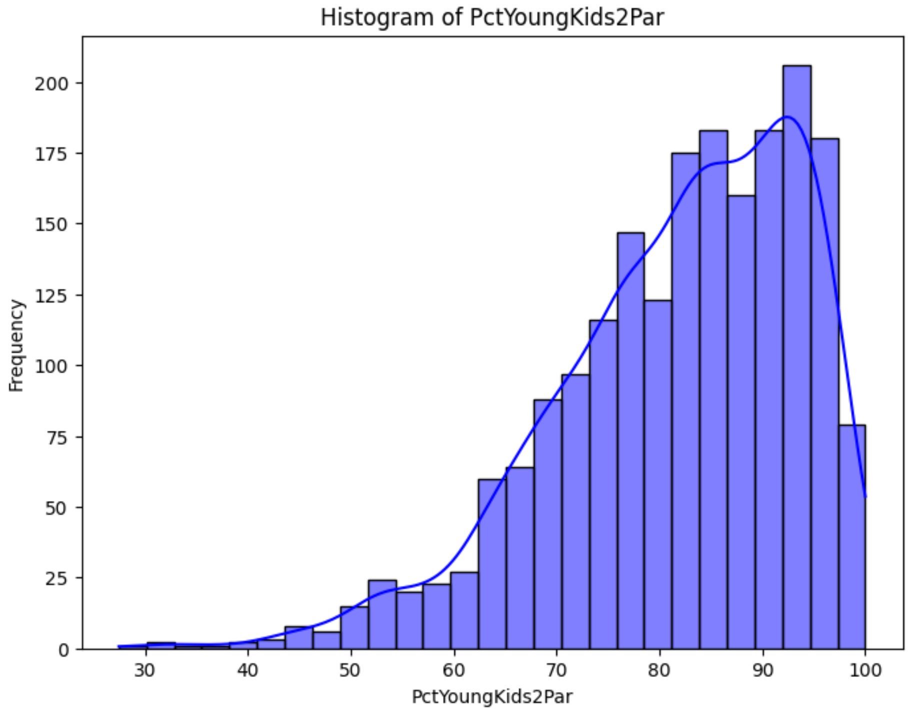

Histogram of PctPopUnderPov
Frequency
0
50
100
150
200
250
300
0
10
20
30
40
50
PctPopUnderPov

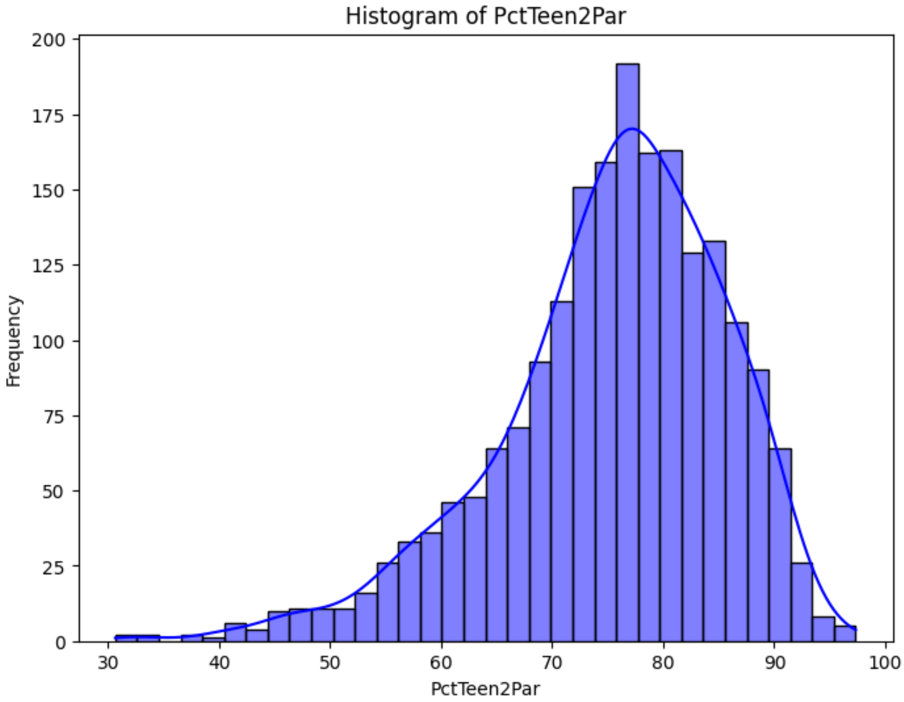

Histogram of PctUnemployed
Frequency
0
25
50
75
100
125
150
175
5
10
15
20
PctUnemployed

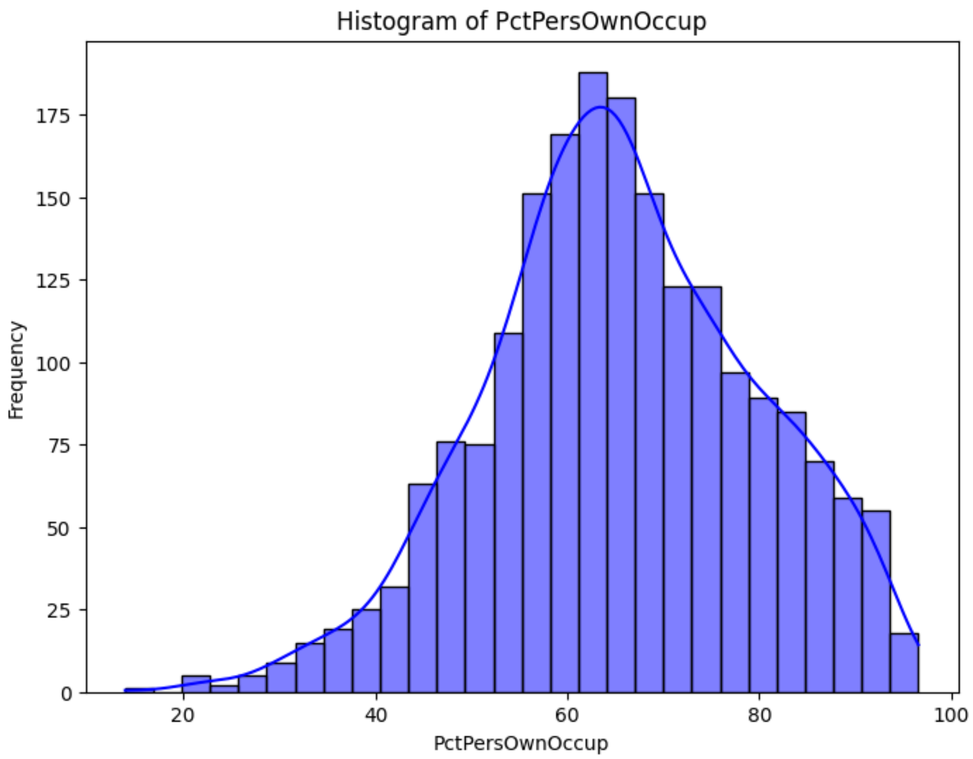

Histogram of TotalPctDiv
Frequency
0
25
50
75
100
125
150
175
4
6
8
10
12
14
16
18
TotalPctDiv

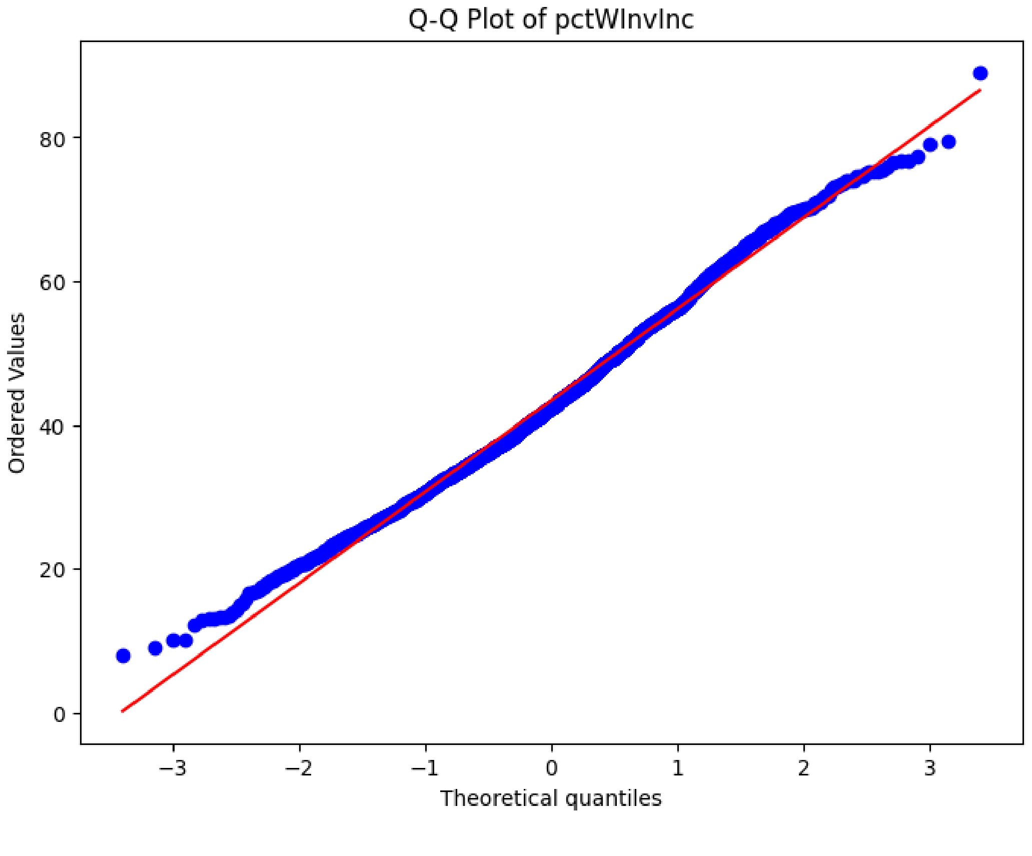

Histogram of PctFam2Par
Frequency
0
25
50
75
100
125
150
175
30
40
50
60
70
80
90
PctFam2Par

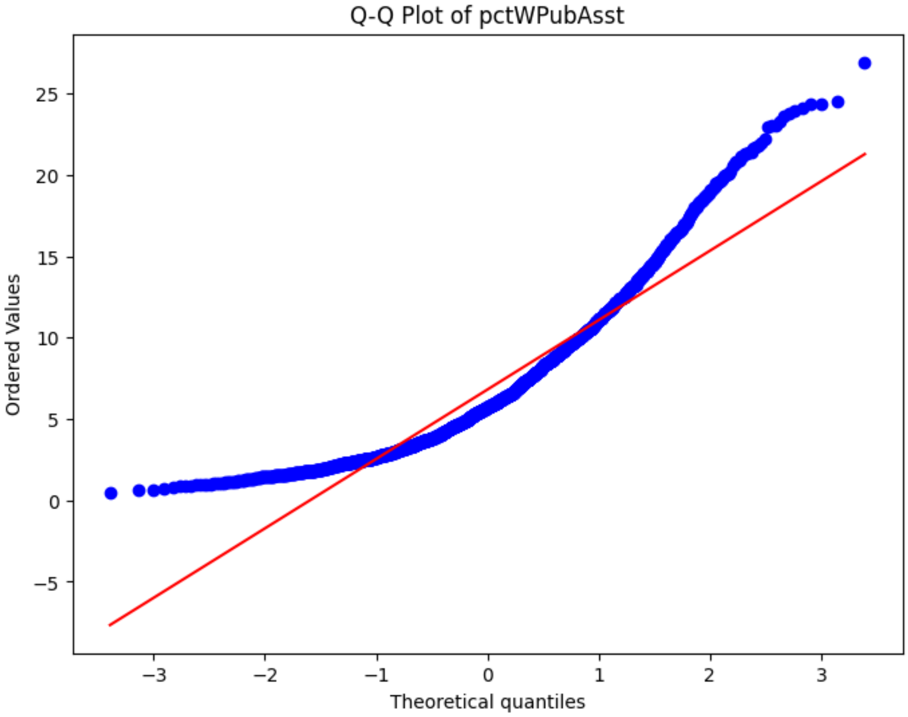

Histogram of PctKids2Par
Frequency
0
25
50
75
100
125
150
175
30
40
50
60
70
80
90
PctKids2Par

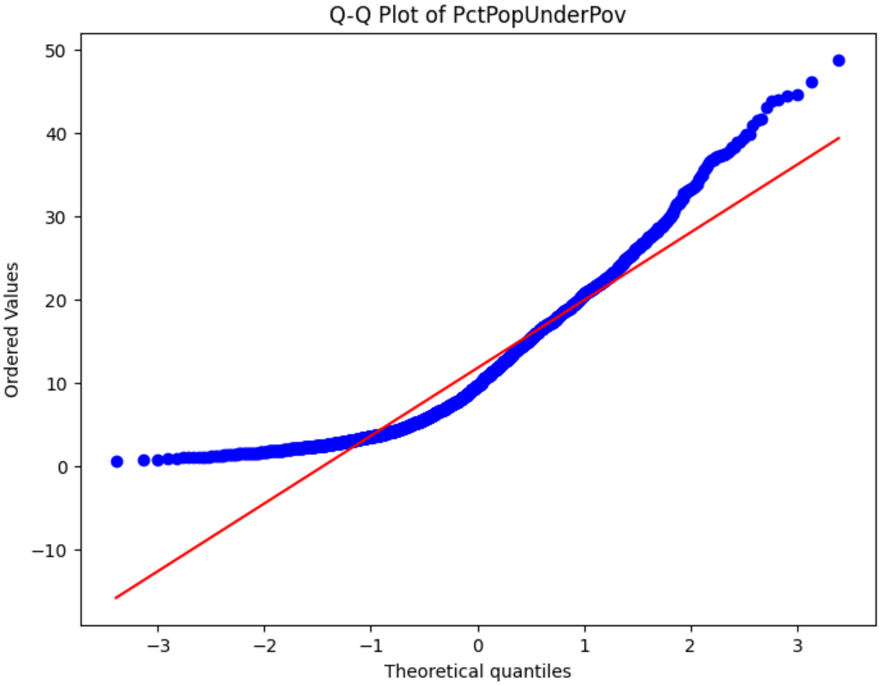

Histogram of PctYoungKids2Par
Frequency
0
25
50
75
100
125
150
175
200
30
40
50
60
70
80
90
100
PctYoungKids2Par

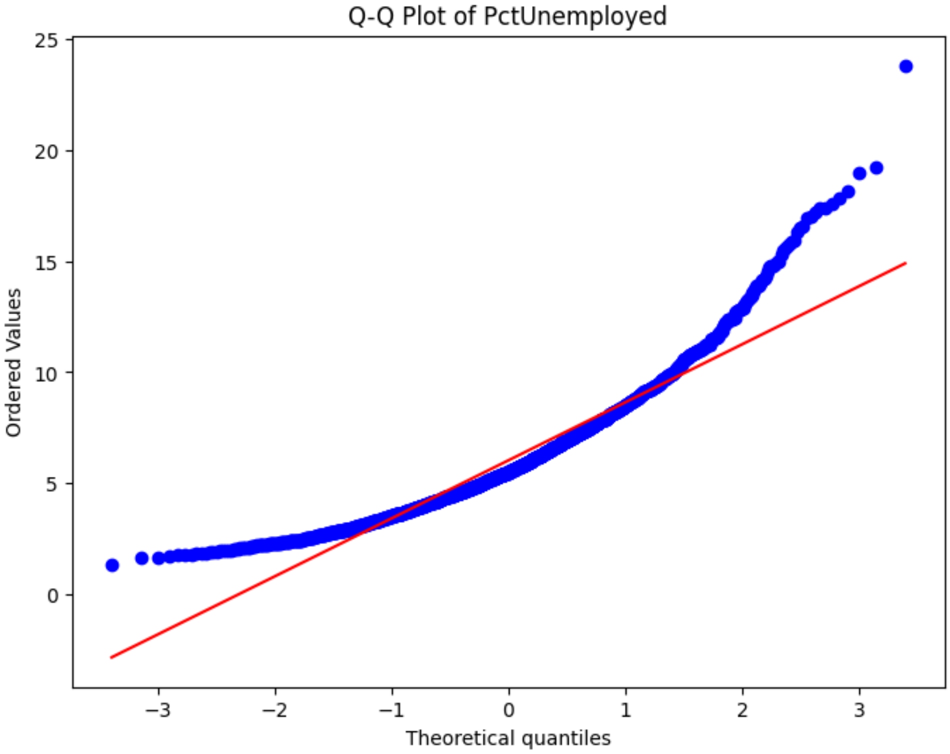

Histogram of PctTeen2Par
Frequency
0
25
50
75
100
125
150
175
200
30
40
50
60
70
80
90
100
PctTeen2Par

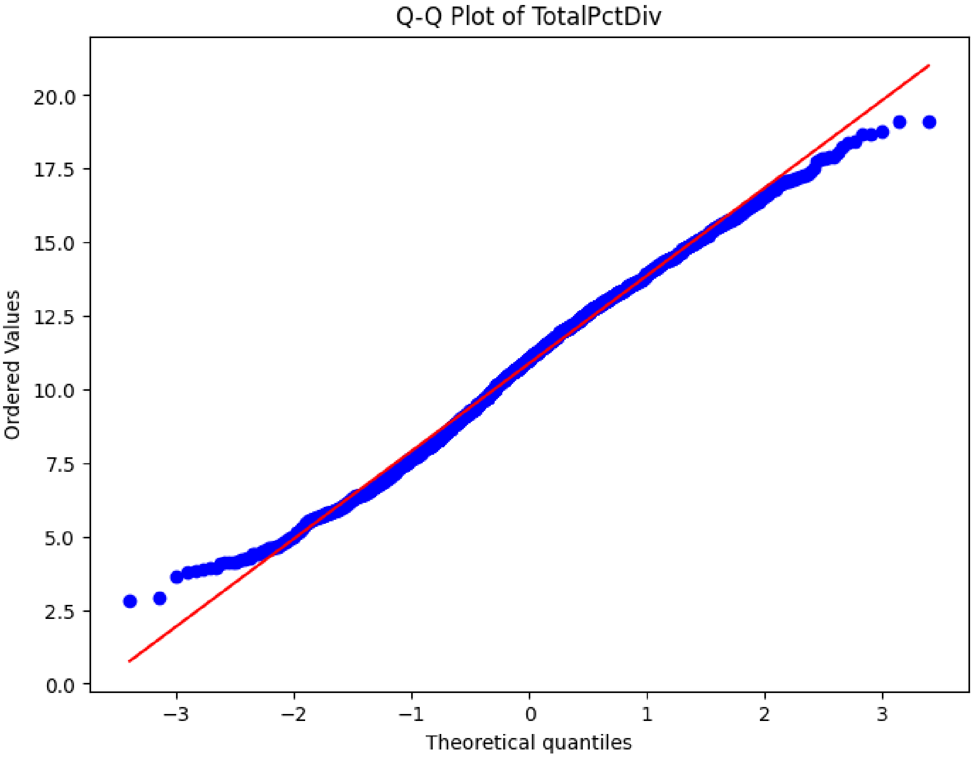

Histogram of PctPersOwnOccup
Frequency
0
25
50
75
100
125
150
175
20
40
60
80
100
PctPersOwnOccup

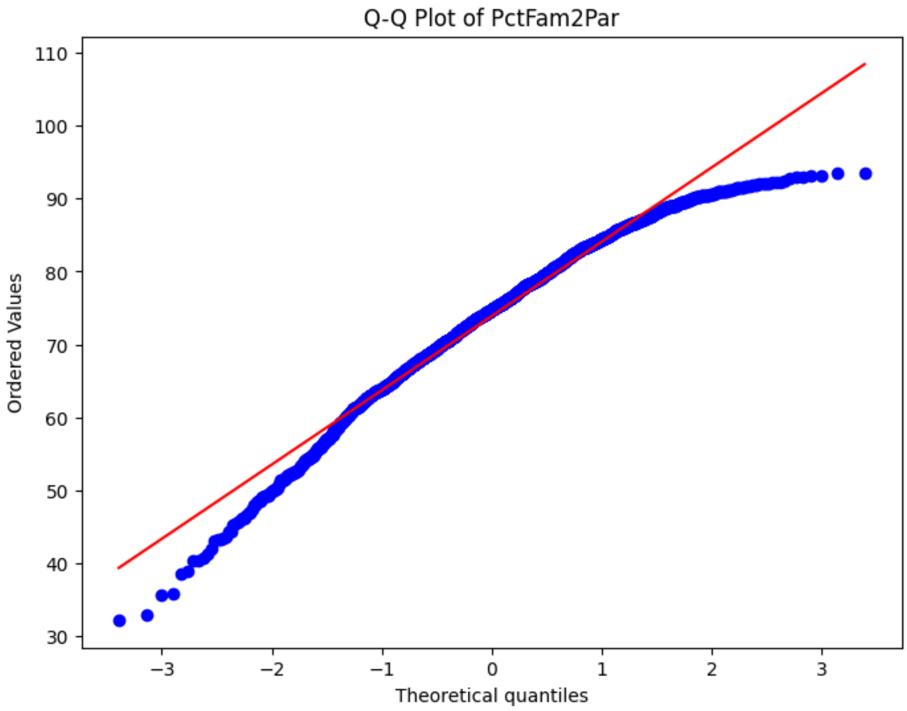

Q-Q Plot of pctWInvInc
Ordered Values
0
20
40
60
80
−3
−2
−1
0
1
2
3
Theoretical quantiles

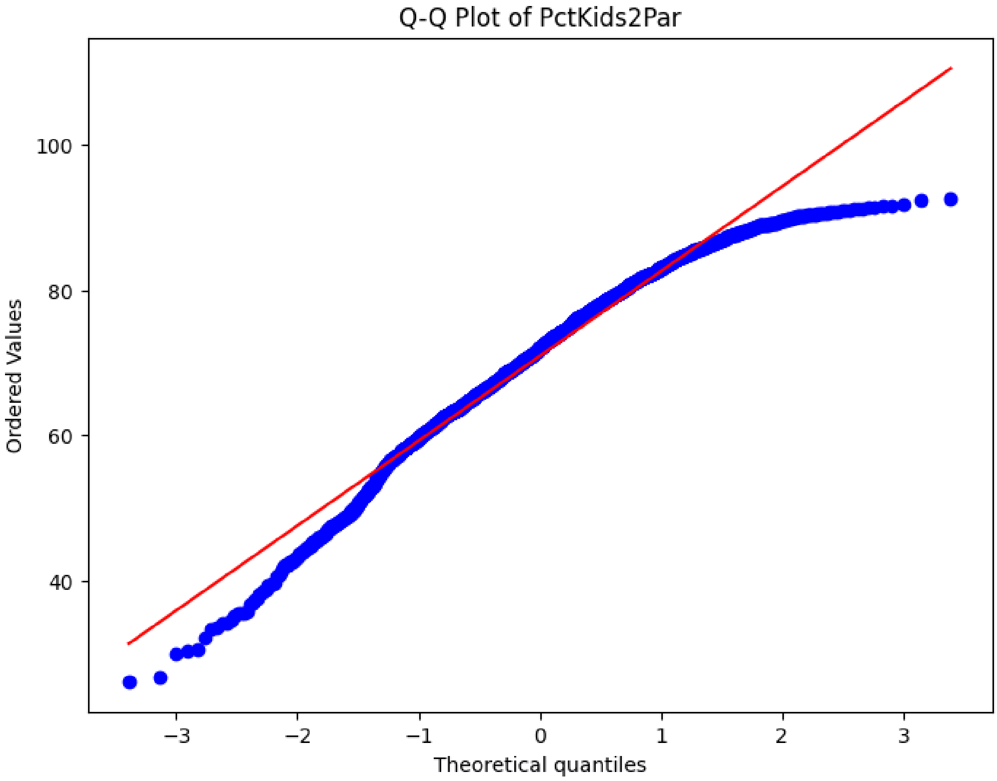

Q-Q Plot of pctWPubAsst
Ordered Values
Theoretical quantiles
25
20
15
10
5
0
−5
−3
−2
−1
0
1
2
3

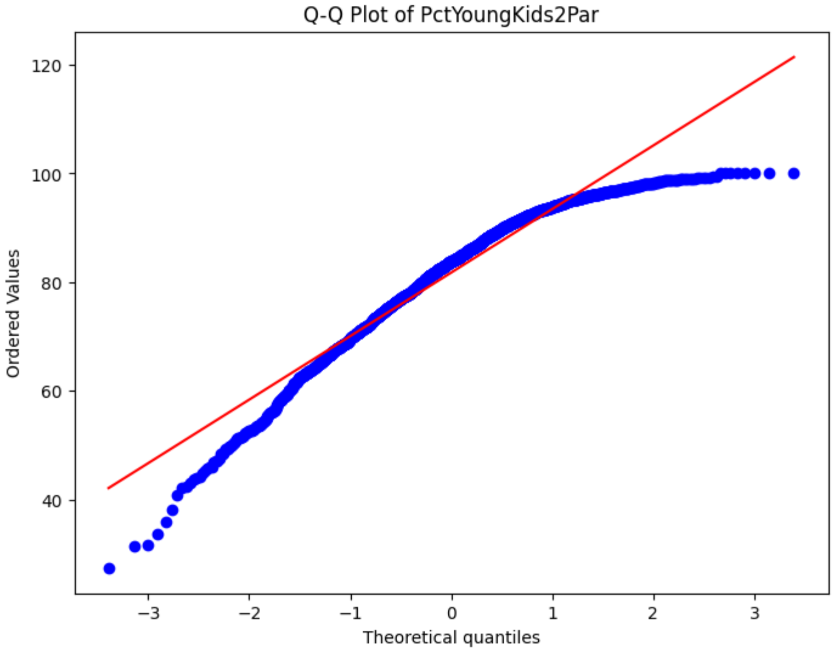

Q-Q Plot of PctPopUnderPov
Ordered Values
Theoretical quantiles
50
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3

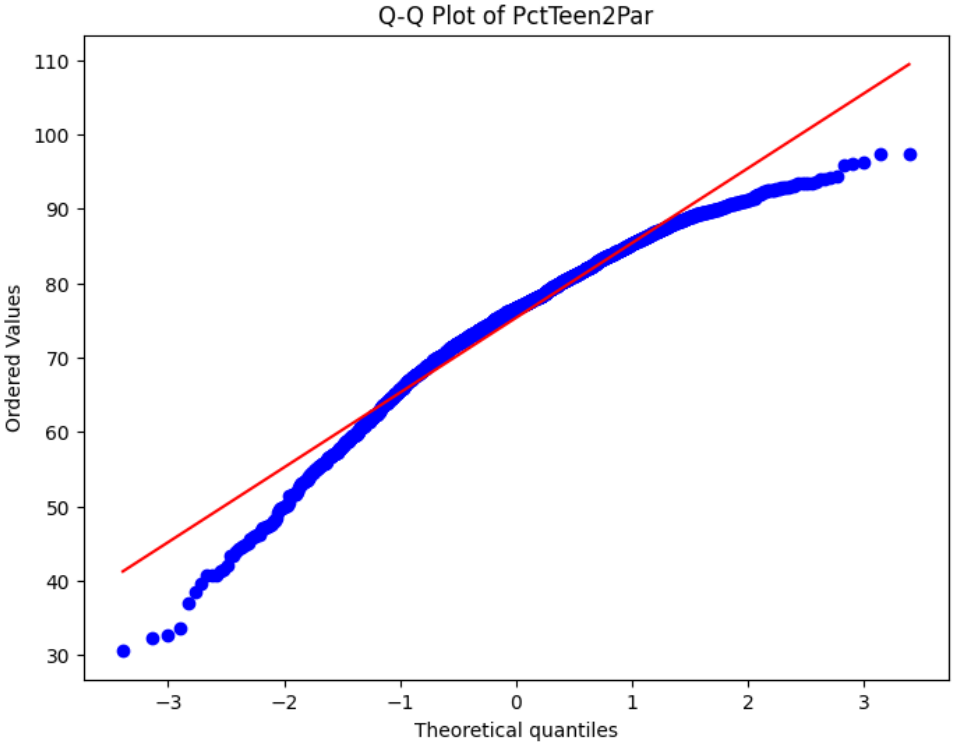

Q-Q Plot of PctUnemployed
Ordered Values
Theoretical quantiles
25
20
15
10
5
0
−3
−2
−1
0
1
2
3

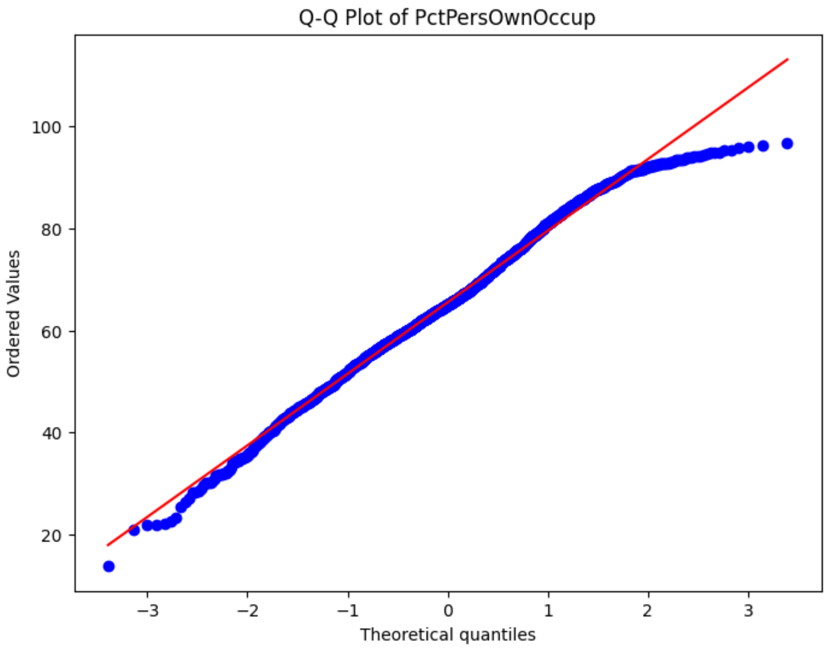

Q-Q Plot of TotalPctDiv
Ordered Values
Theoretical quantiles
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−3
−2
−1
0
1
2
3

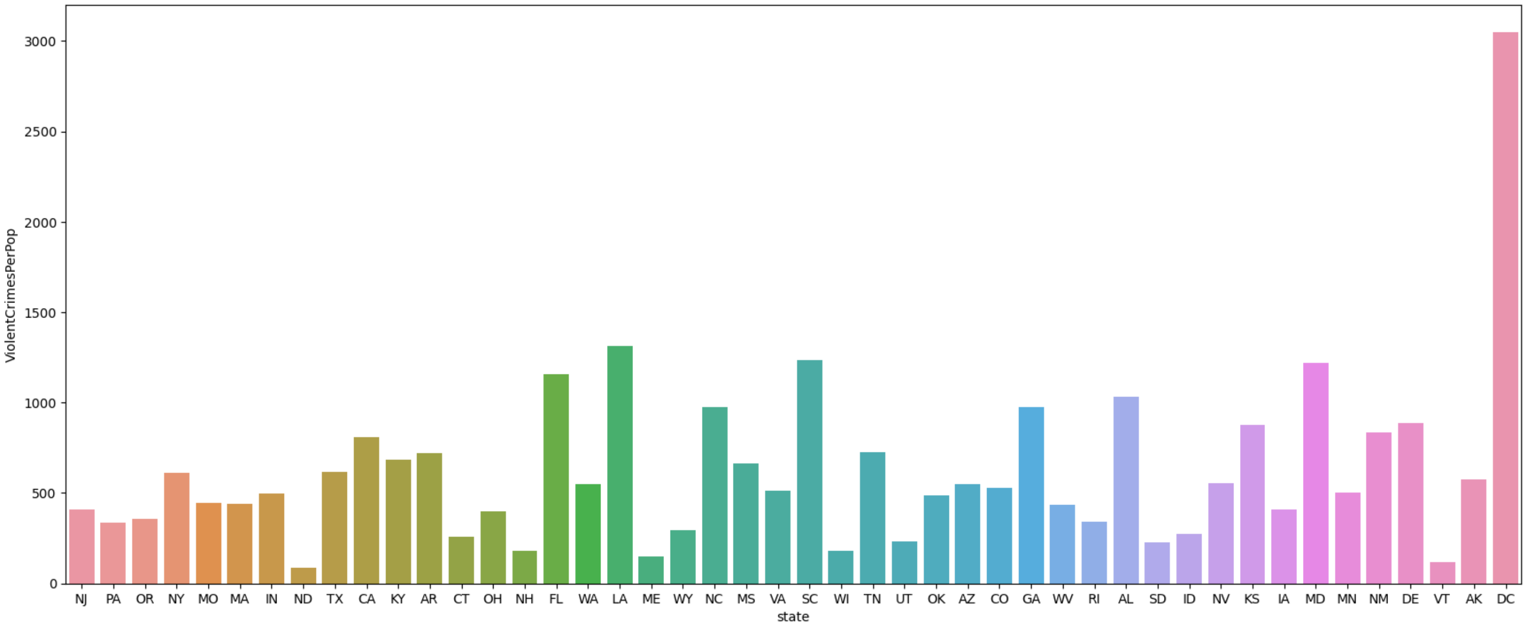

Q-Q Plot of PctFam2Par
Ordered Values
Theoretical quantiles
110
100
90
80
70
60
50
40
30
−3
−2
−1
0
1
2
3

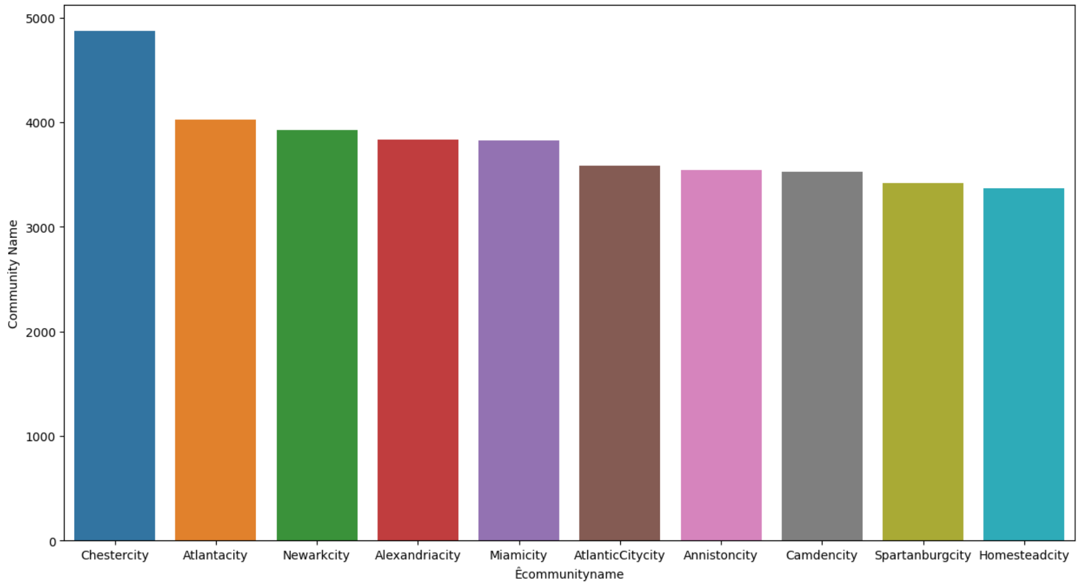

Q-Q Plot of PctKids2Par
Ordered Values
Theoretical quantiles
100
80
60
40
−3
−2
−1
0
1
2
3

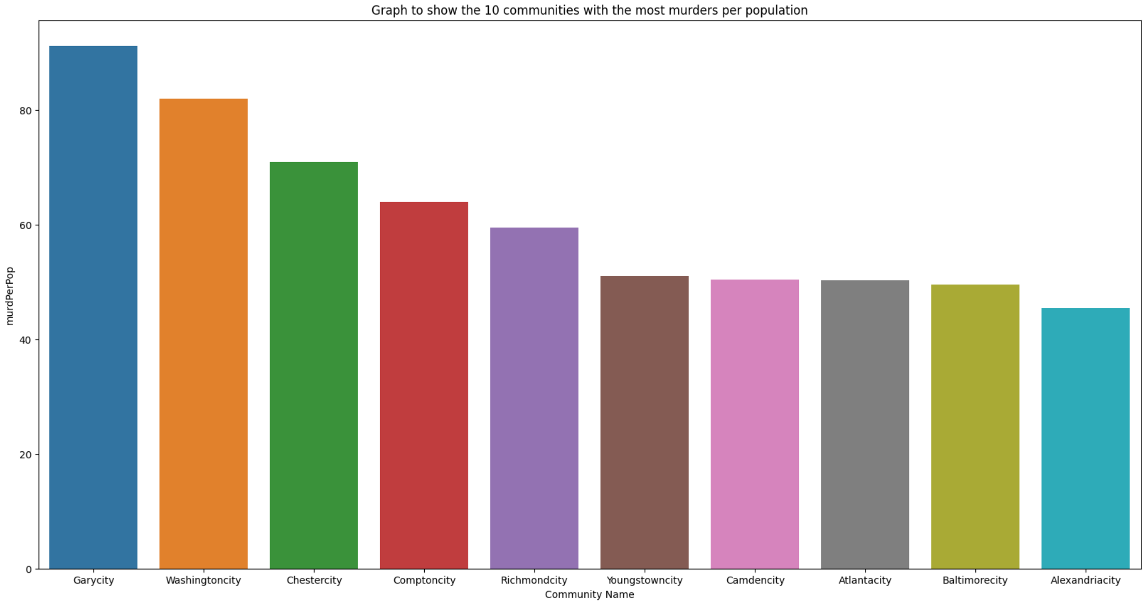

Q-Q Plot of PctYoungKids2Par
120
100
80
60
40
Ordered Values
−3
−2
−1
0
1
2
3
Theoretical quantiles

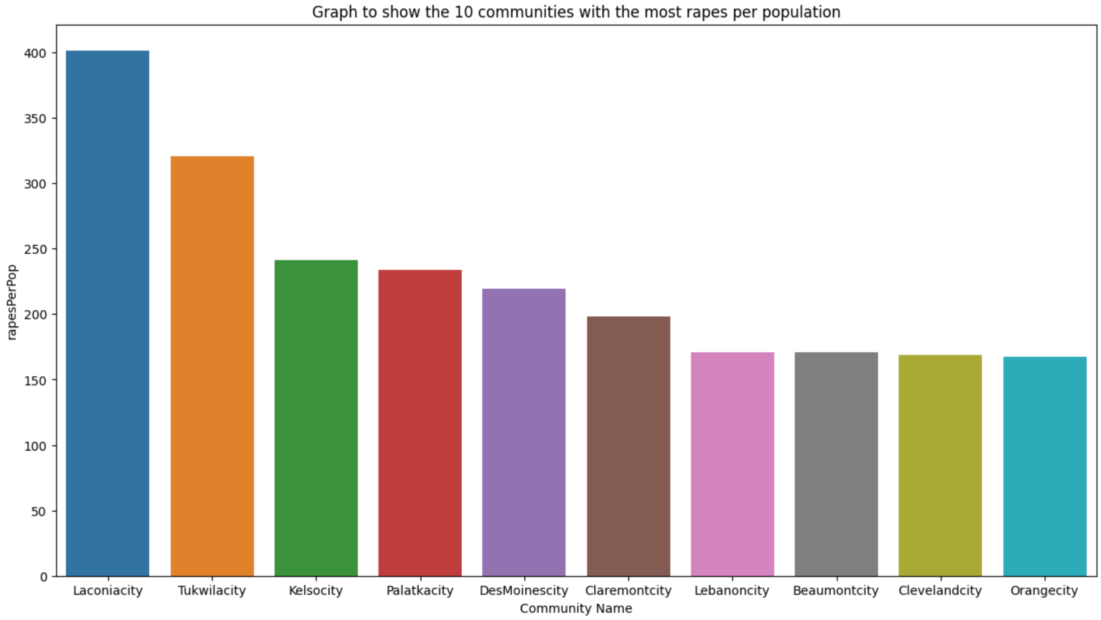

Q-Q Plot of PctTeen2Par
110
100
90
80
70
60
50
40
30
Ordered Values
−3
−2
−1
0
1
2
3
Theoretical quantiles

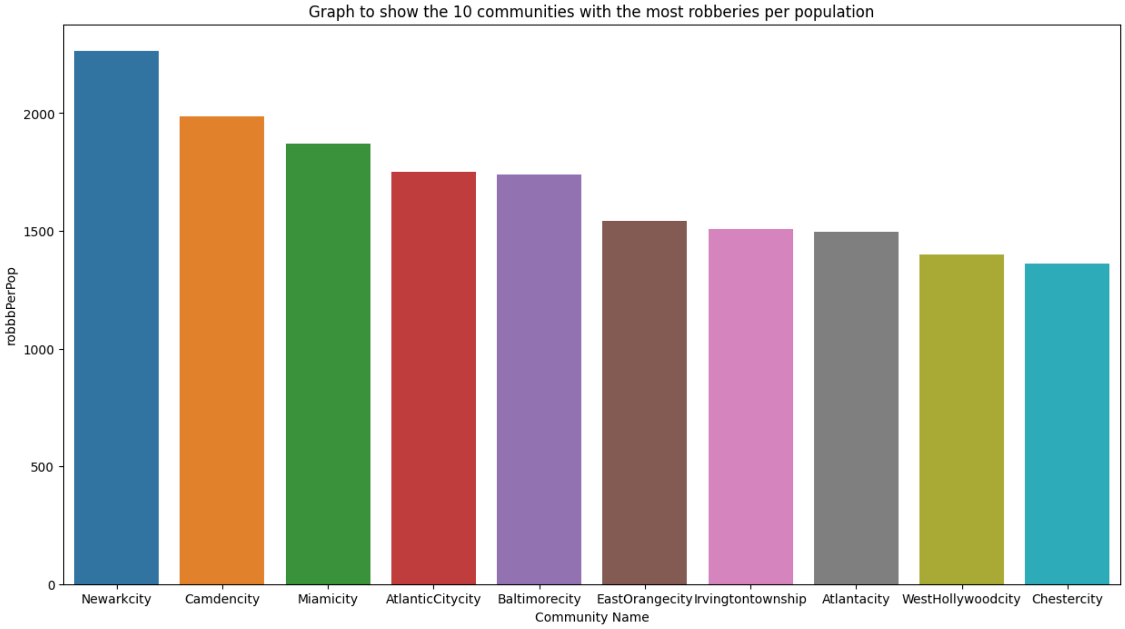

Q-Q Plot of PctPersOwnOccup
Ordered Values
20
40
60
80
100
Theoretical quantiles
−3
−2
−1
0
1
2
3

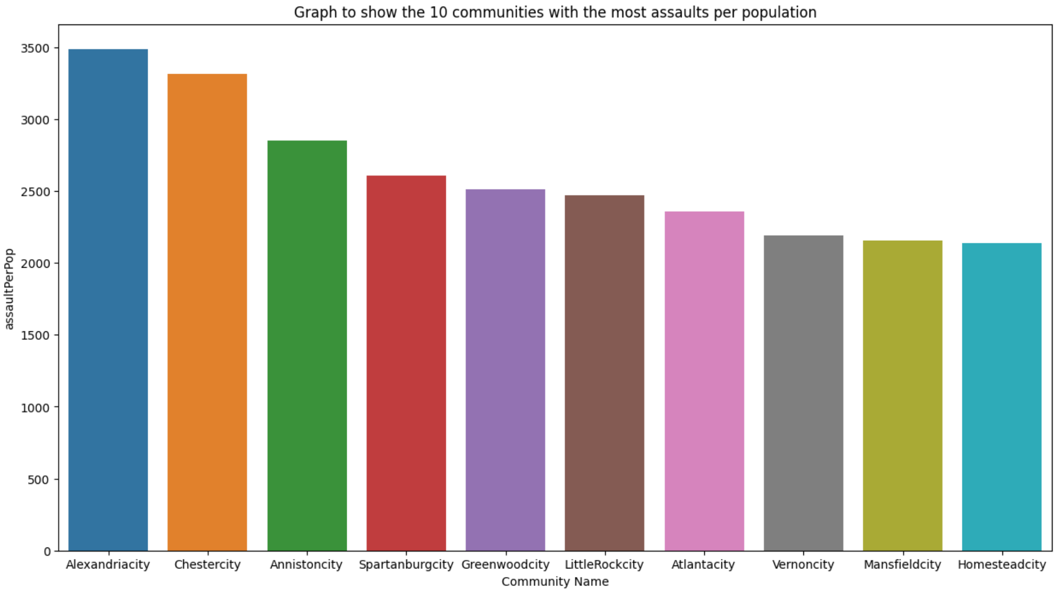

ViolentCrimesPerPop
0
500
1000
1500
2000
2500
3000
state
NJ
PA
OR
NY
MO
MA
IN
ND
TX
CA
KY
AR
CT
OH
NH
FL
WA
LA
ME
WY
NC
MS
VA
SC
WI
TN
UT
OK
AZ
CO
GA
WV
RI
AL
SD
ID
NV
KS
IA
MD
MN
NM
DE
VT
AK
DC

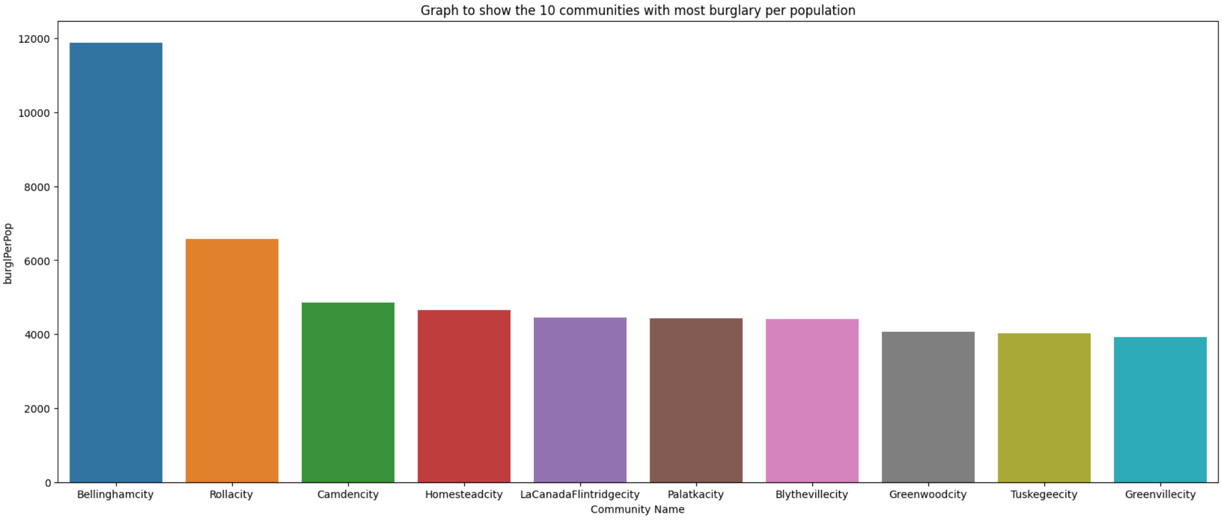

5000
4000
3000
2000
1000
0
Community Name
Chestercity
Atlantacity
Newarkcity
Alexandriacity
Miamicity
AtlanticCitycity
Annistoncity
Camdencity
Spartanburgcity
Homesteadcity
Êcommunityname

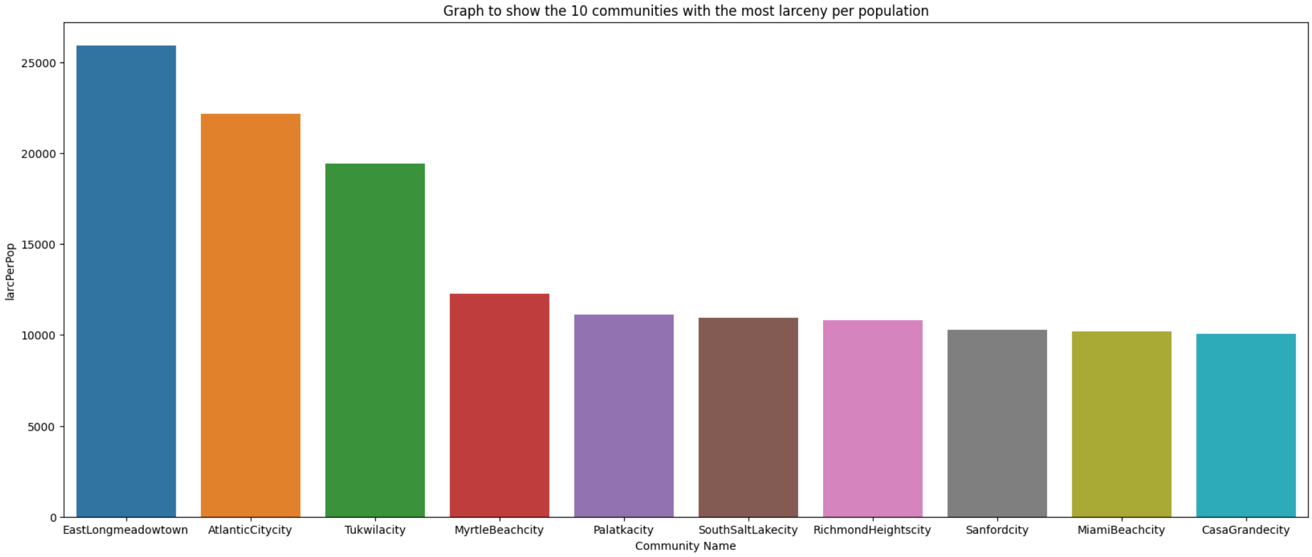

Graph to show the 10 communities with the most murders per population
80
60
40
20
0
murdPerPop
Garycity
Washingtoncity
Chestercity
Comptoncity
Richmondcity
Youngstowncity
Camdencity
Atlantacity
Baltimorecity
Alexandriacity
Community Name

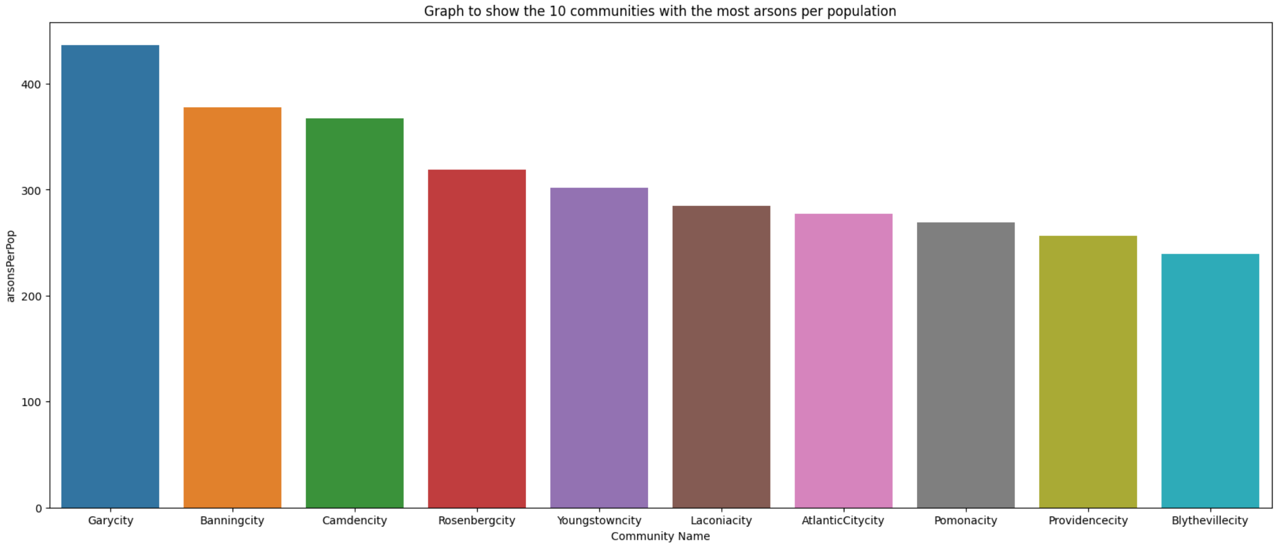

Graph to show the 10 communities with the most rapes per population
rapesPerPop
400
350
300
250
200
150
100
50
0
Laconiacity
Tukwilacity
Kelsocity
Palatkacity
DesMoinescity
Claremontcity
Lebanoncity
Beaumontcity
Clevelandcity
Orangecity
Community Name

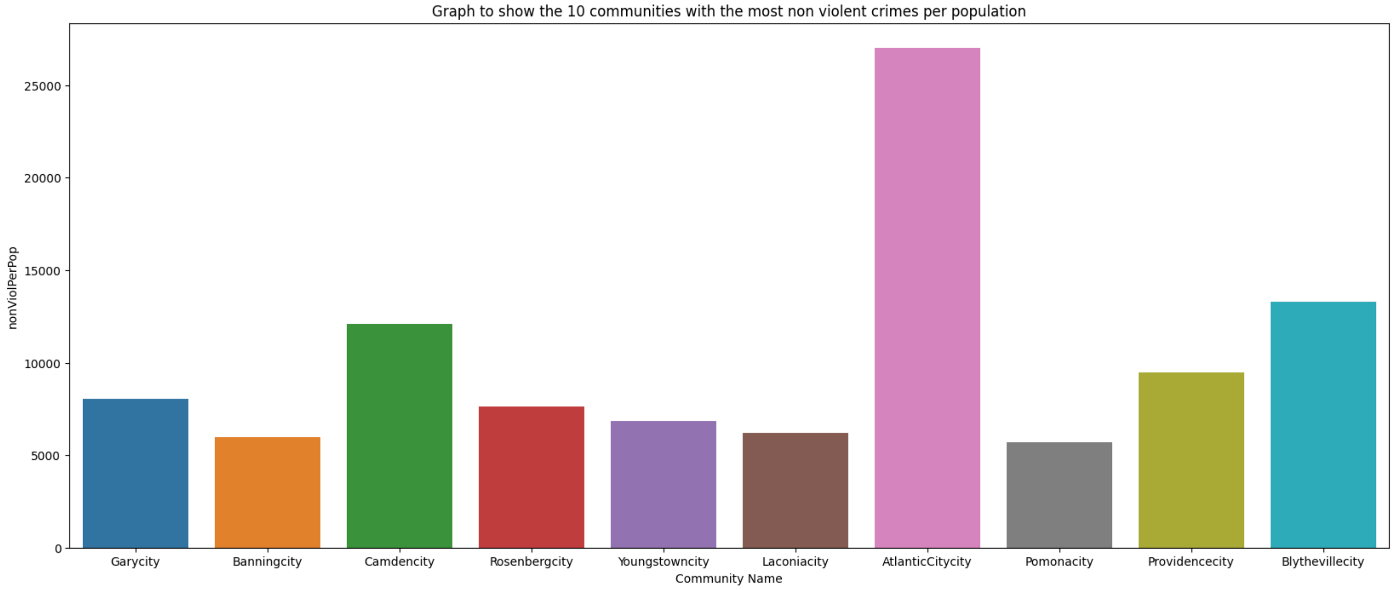

Graph to show the 10 communities with the most robberies per population
robbbPerPop
2000
1500
1000
500
0
Newarkcity
Camdencity
Miamicity
AtlanticCitycity
Baltimorecity
EastOrangecity
Irvingtontownship
Atlantacity
WestHollywoodcity
Chestercity
Community Name

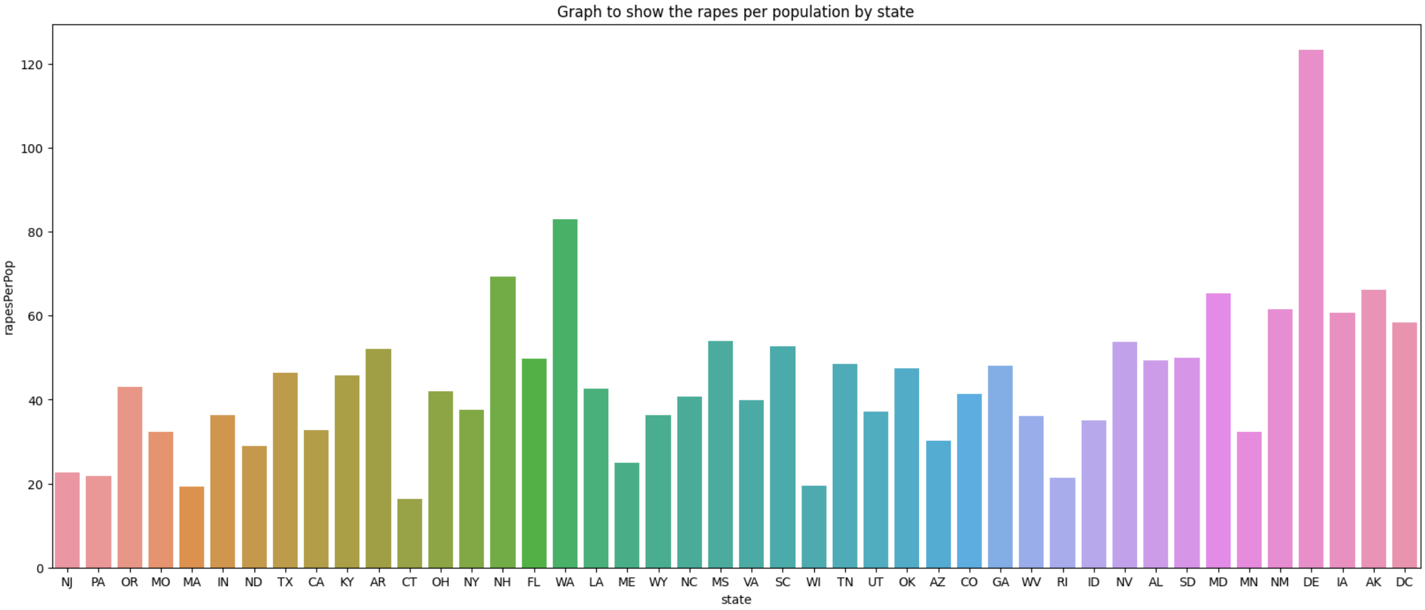

Graph to show the 10 communities with the most assaults per population
assaultPerPop
0
500
1000
1500
2000
2500
3000
3500
Alexandriacity
Chestercity
Annistoncity
Spartanburgcity
Greenwoodcity
LittleRockcity
Atlantacity
Vernoncity
Mansfieldcity
Homesteadcity
Community Name

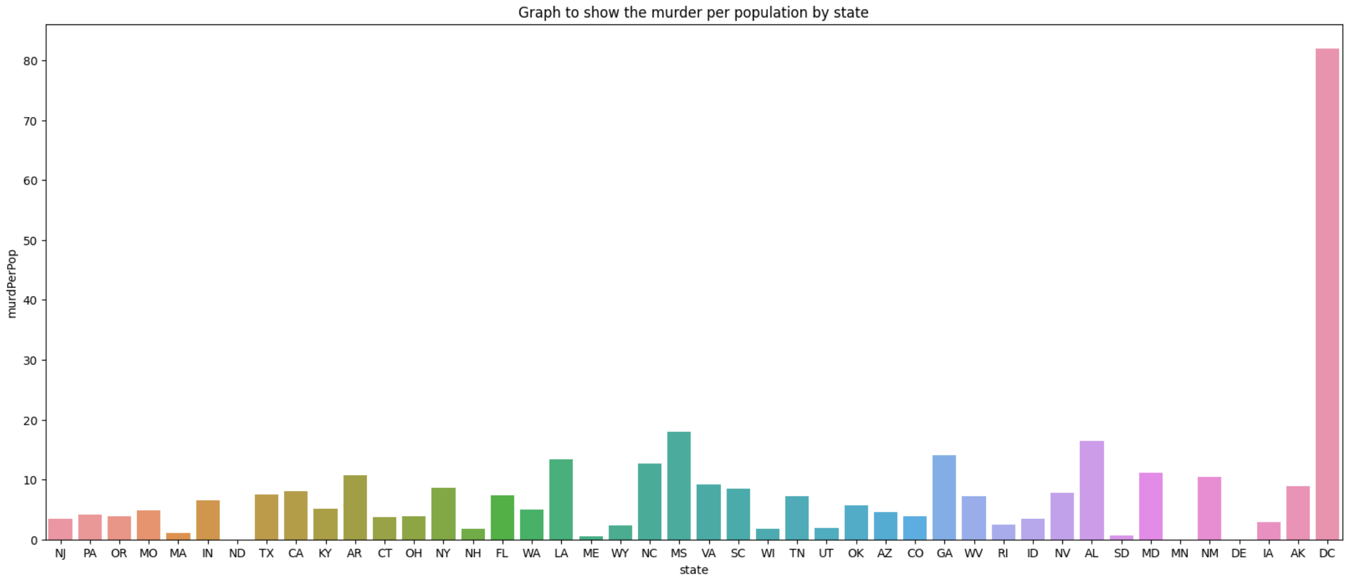

Graph to show the 10 communities with most burglary per population
burglPerPop
0
2000
4000
6000
8000
10000
12000
Bellinghamcity
Rollacity
Camdencity
Homesteadcity
LaCanadaFlintridgecity
Palatkacity
Blythevillecity
Greenwoodcity
Tuskegeecity
Greenvillecity
Community Name


Graph to show the 10 communities with the most larceny per population
larcPerPop
0
5000
10000
15000
20000
25000
EastLongmeadowtown
AtlanticCitycity
Tukwilacity
MyrtleBeachcity
Palatkacity
SouthSaltLakecity
RichmondHeightscity
Sanfordcity
MiamiBeachcity
CasaGrandecity
Community Name

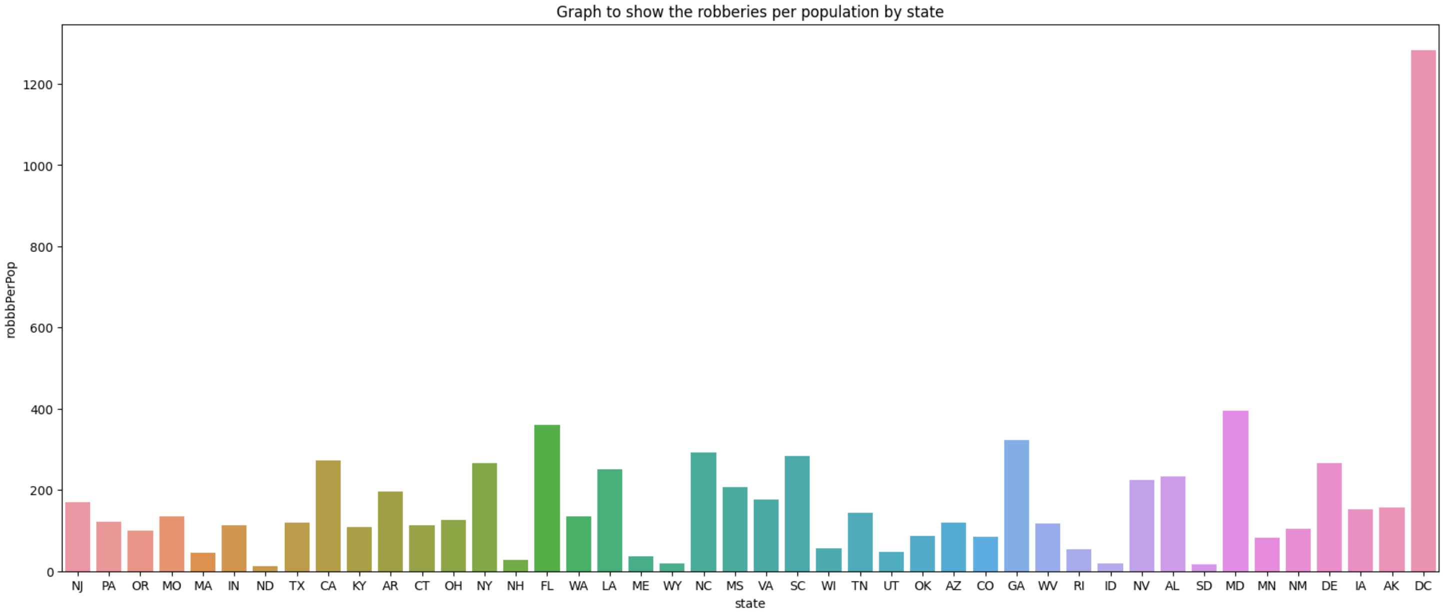

Graph to show the 10 communities with the most arsons per population
arsonsPerPop
0
100
200
300
400
Garycity
Banningcity
Camdencity
Rosenbergcity
Youngstowncity
Laconiacity
AtlanticCitycity
Pomonacity
Providencecity
Blythevillecity
Community Name


Graph to show the 10 communities with the most non violent crimes per population
nonViolPerPop
25000
20000
15000
10000
5000
0
Garycity
Banningcity
Camdencity
Rosenbergcity
Youngstowncity
Laconiacity
AtlanticCitycity
Pomonacity
Providencecity
Blythevillecity
Community Name

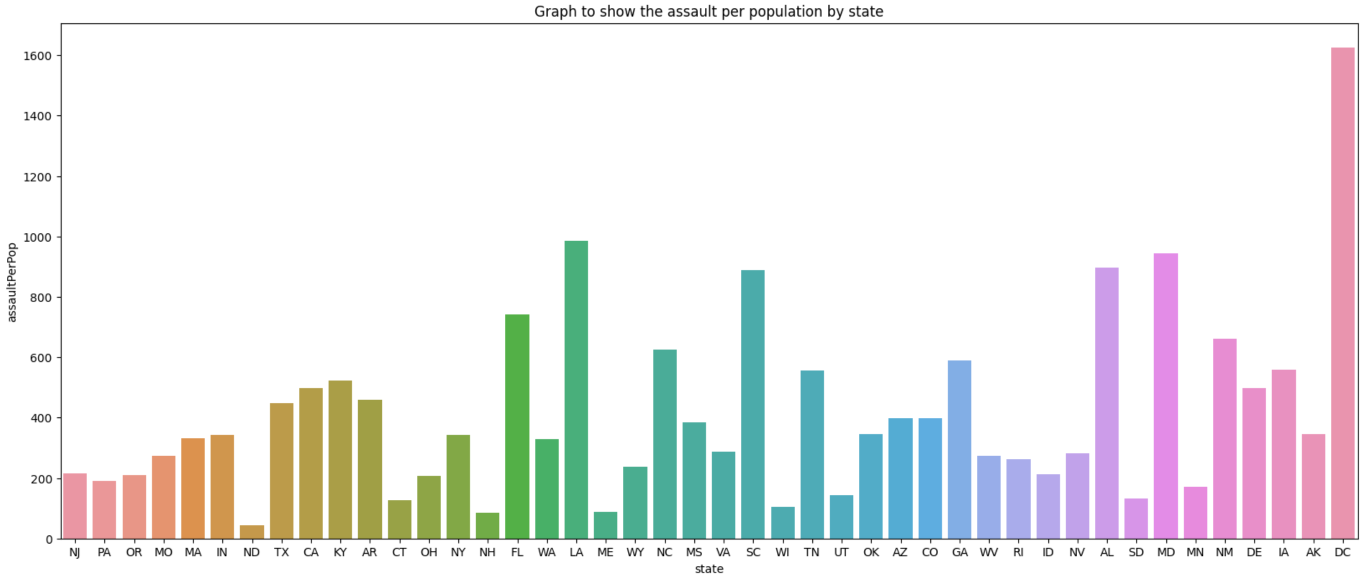

Graph to show the rapes per population by state
rapesPerPop
120
100
80
60
40
20
0
NJ
PA
OR
MO
MA
IN
ND
TX
CA
KY
AR
CT
OH
NY
NH
FL
WA
LA
ME
WY
NC
MS
VA
SC
WI
TN
UT
OK
AZ
CO
GA
WV
RI
ID
NV
AL
SD
MD
MN
NM
DE
IA
AK
DC
state



Graph to show the murder per population by state

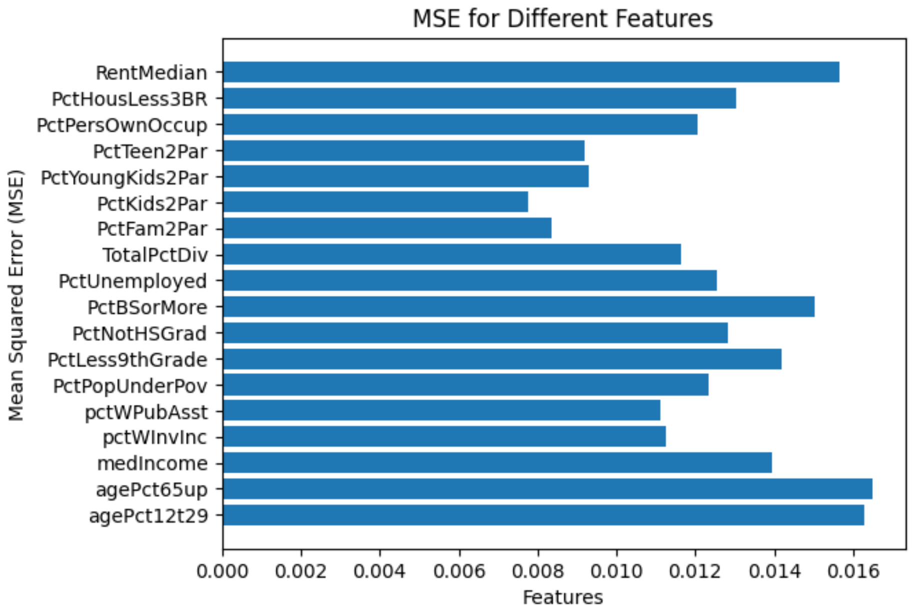


Graph to show the robberies per population by state

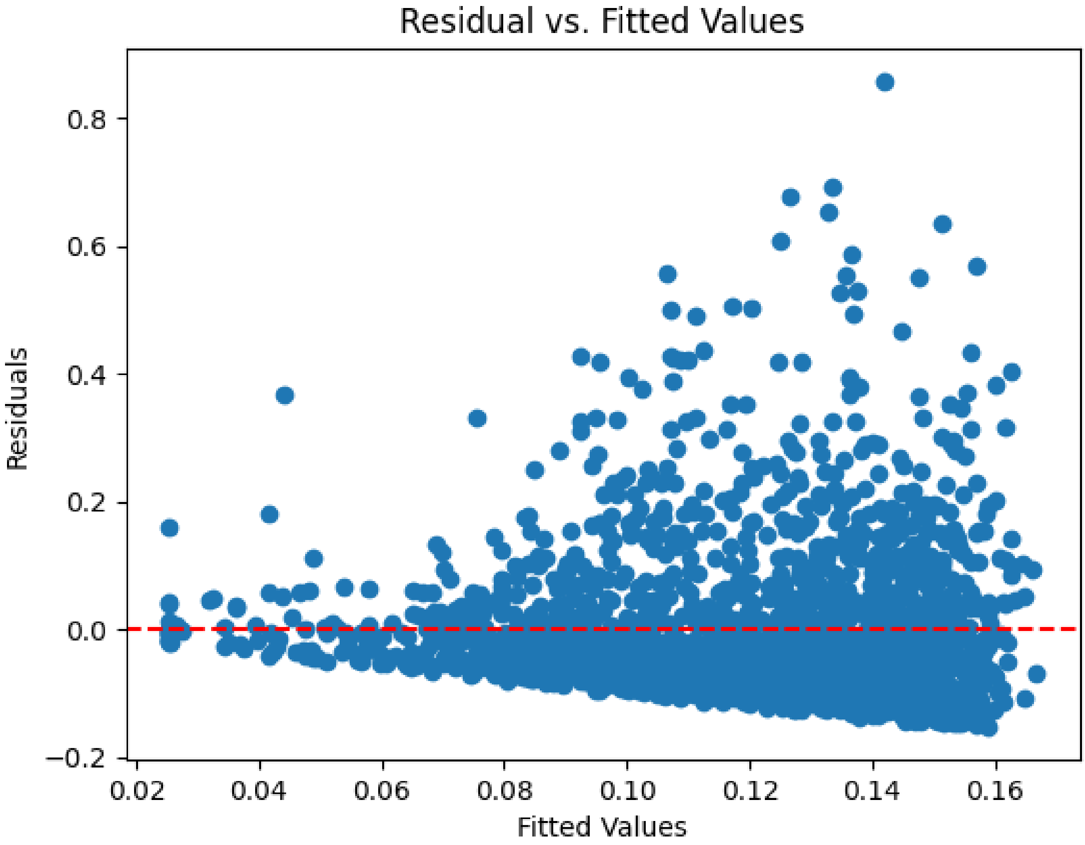


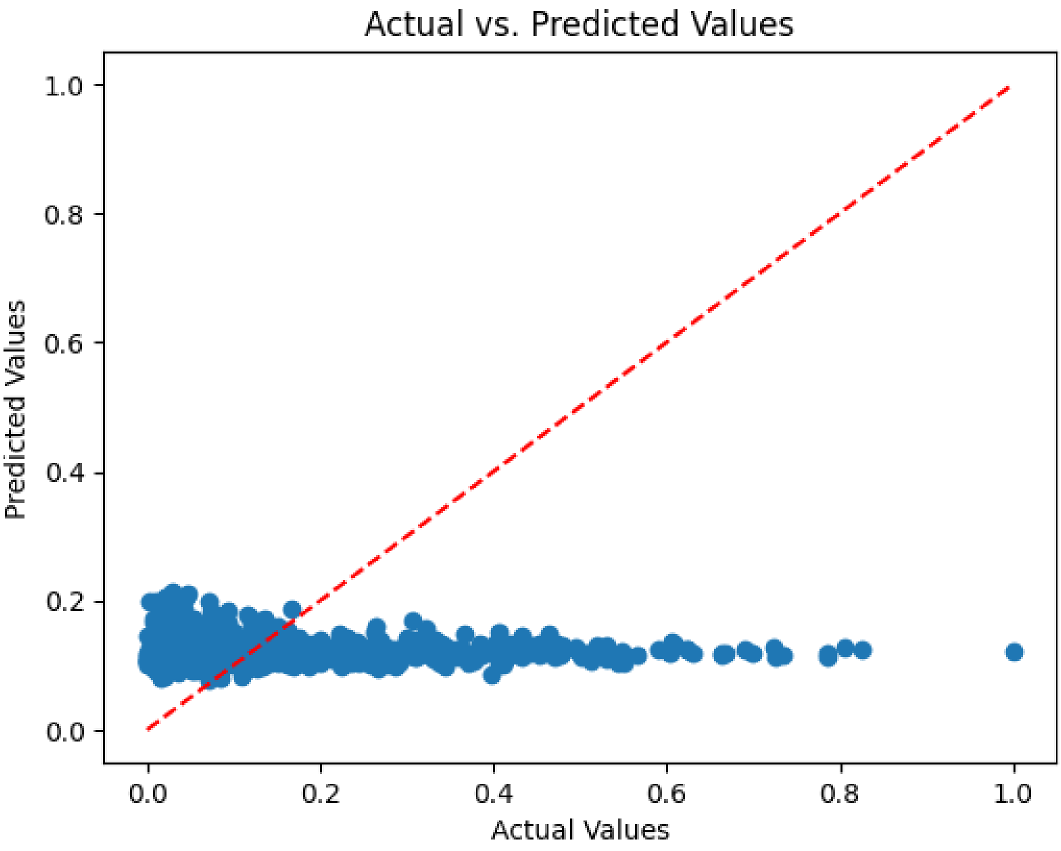
Field	MSE	R2 score
Age 12 to 29	0.017606658399656576	0.06483085676854172
agePct65up	0.018758484249943513	0.00365218395543776
medIncome	0.015760041429352542	0.1629130237985773
pctWInvInc	0.013121883981721963	0.30303748035408706
pctWPubAsst	0.013370721877888564	0.2898205758808239
PctPopUnderPov	0.012704703952857502	0.3251957957658328
PctLess9thGrade	0.01621417049800407	0.13879217801357402
PctNotHSGrad	0.015474270495892	0.1780916086804173
PctBSorMore	0.017783939823420174	0.05541463970971083
PctUnemployed	0.015070256371348463p	0.1995506234534965
TotalPctDiv	0.013578737199927447	0.27877194268340166
PctFam2Par	0.010644508899622094	0.43462205934696896
PctKids2Par	0.009595492303231274	0.49034006837595767
PctYoungKids2Par	0.011825094822217861	0.37191580920659373
PctTeen2Par	0.011764551305661156	0.37513154879904576
PctPersOwnOccup	0.014010306439017491	0.37513154879904576
PctHousLess3BR	0.014978236129227716	0.20443823409662043
RentMedian	0.01799930764204712	0.04397548221296976

Field	P value
Age 12 to 29	1.0651032257866213^{23}
agePct65up	0.019044200888472133
medIncome	4.656233706420385^{60}
pctWInvInc	5.544481925676516^{120}
pctWPubAsst	7.668945710632465^{114}
PctPopUnderPov	$1.5269350833837685^{-130}$
PctLess9thGrade	9.420404337137267^{-51}
PctNotHSGrad	4.748329210159761^{-66}
PctBSorMore	2.1407141254872104^{-20}
PctUnemployed	1.0413954998216403^{-74}
TotalPctDiv	8.542585835992524^{109}
PctFam2Par	$2.3801858982072216^{-188}$
PctKids2Par	3.084108696524363^{-222}
PctYoungKids2Par	5.471462522400436^{-154}
PctTeen2Par	$1.1506646496000673^{-155}$
PctPersOwnOccup	1.4511110979106658^{-98}
PctHousLess3BR	1.0314060004845306^{-76}
RentMedian	$1.2.0332806107249665^{-16}$

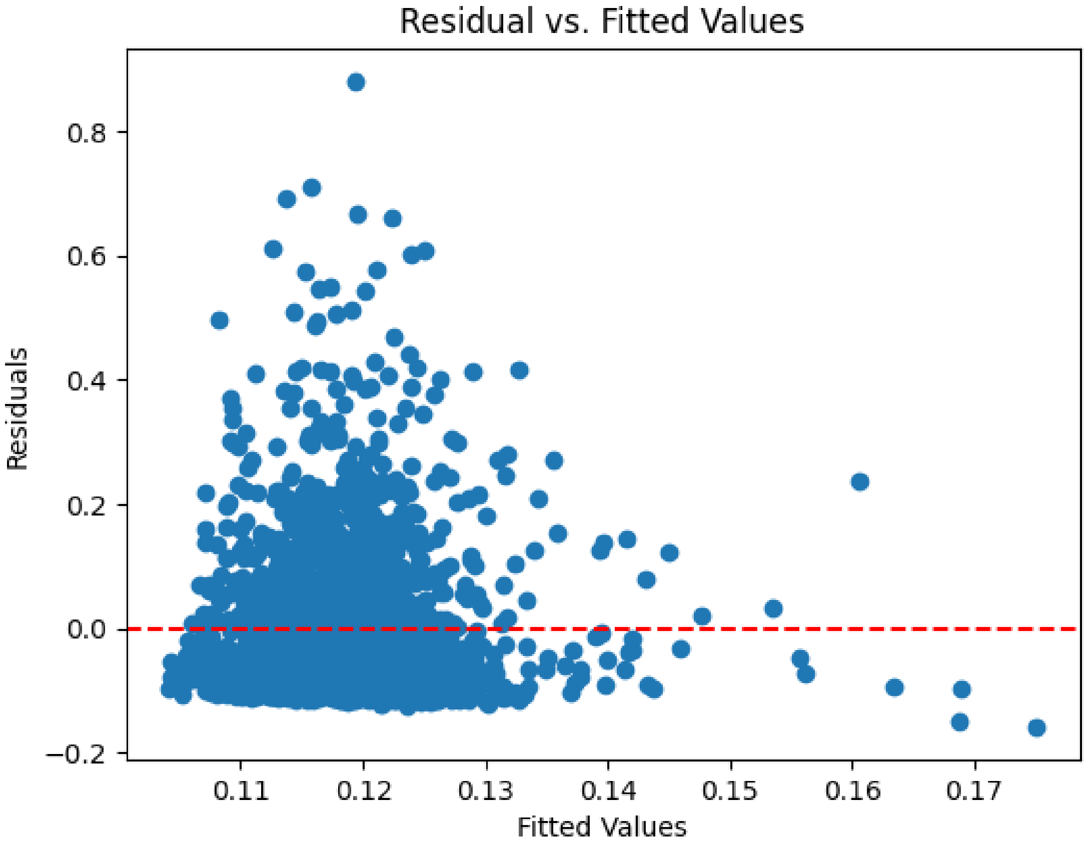


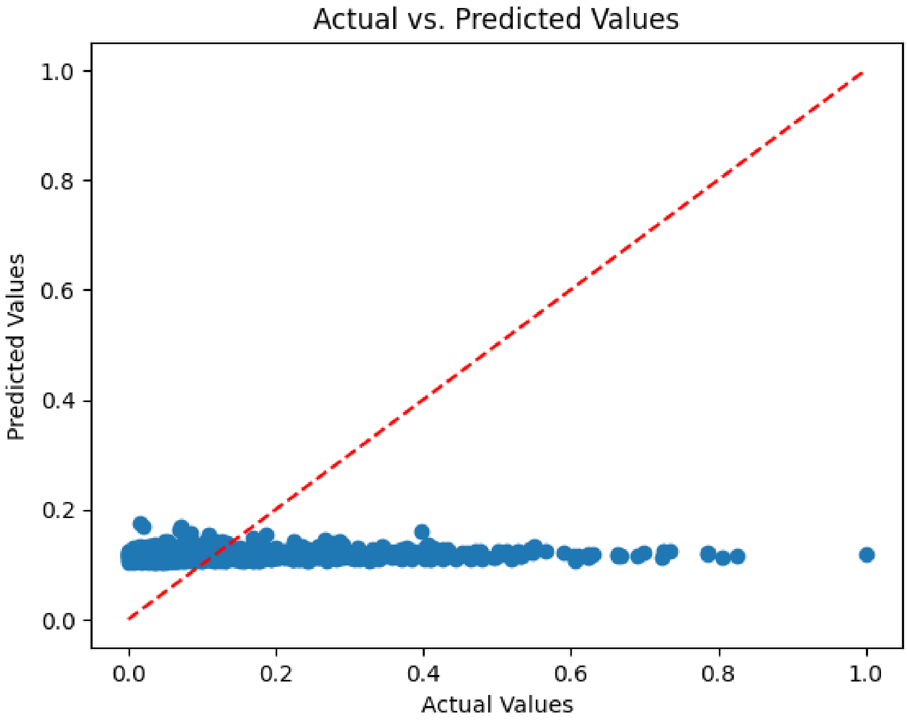
Age 12 to 29 residual plot



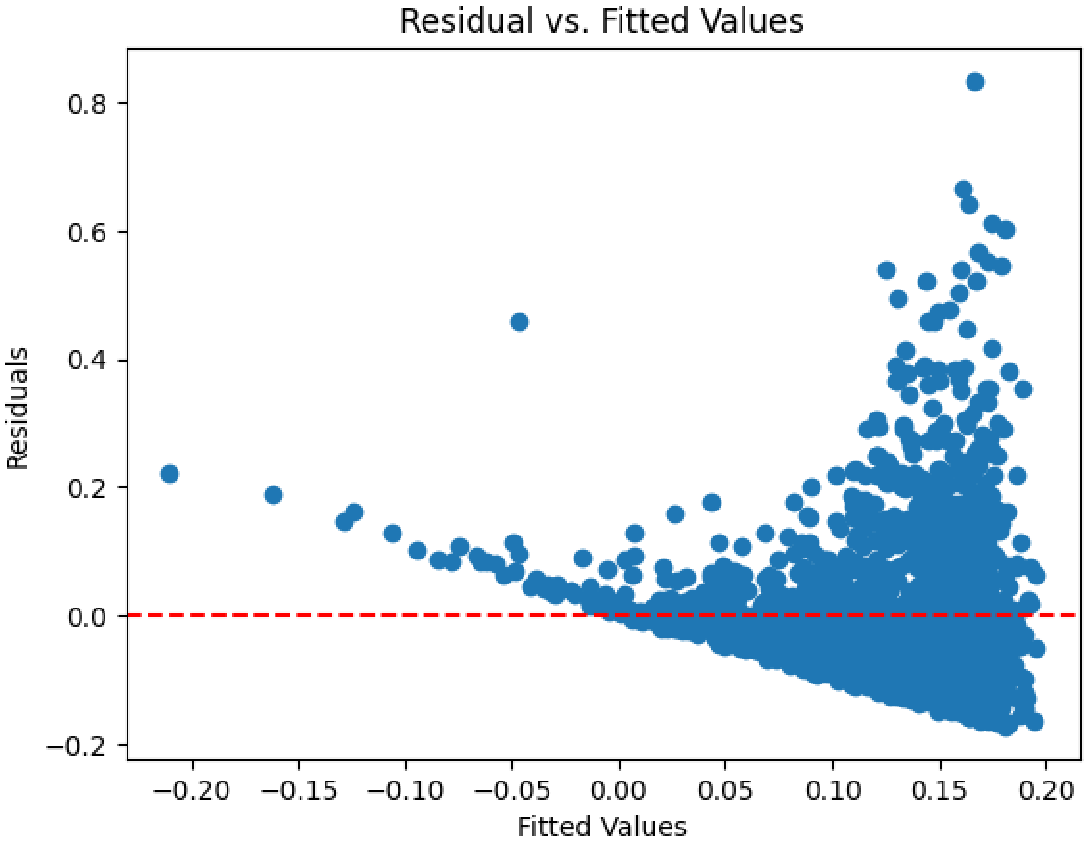


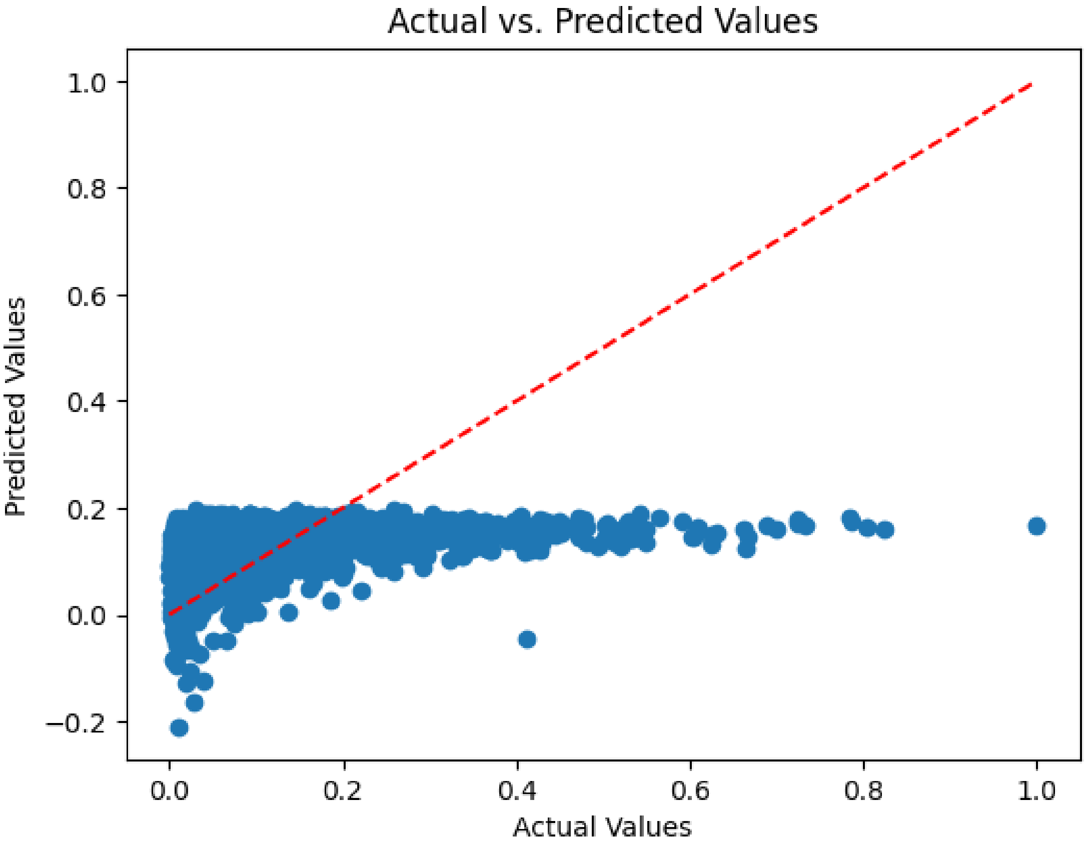
agePct65up Residual Plot for partially cleaned data


Residual vs. Fitted Values
Residuals
0.8
0.6
0.4
0.2
0.0
−0.2
0.11
0.12
0.13
0.14
0.15
0.16
0.17
Fitted Values

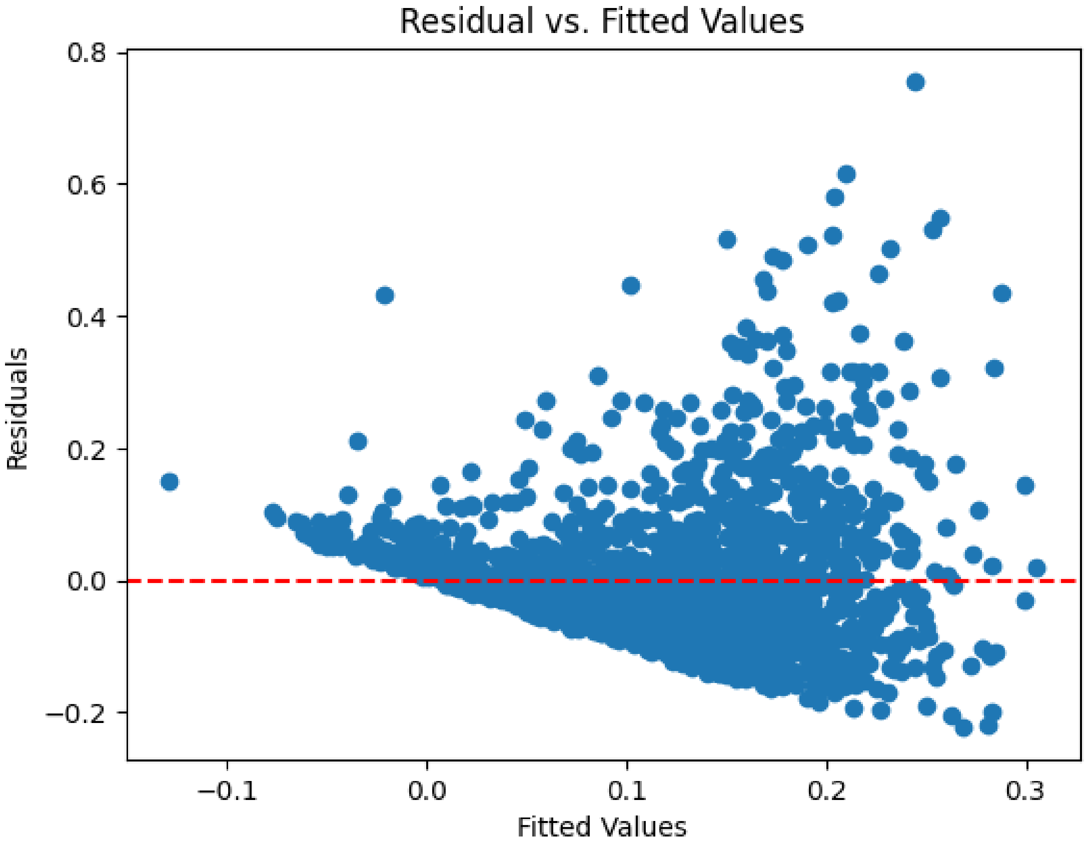

Actual vs. Predicted Values
Predicted Values
Actual Values
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

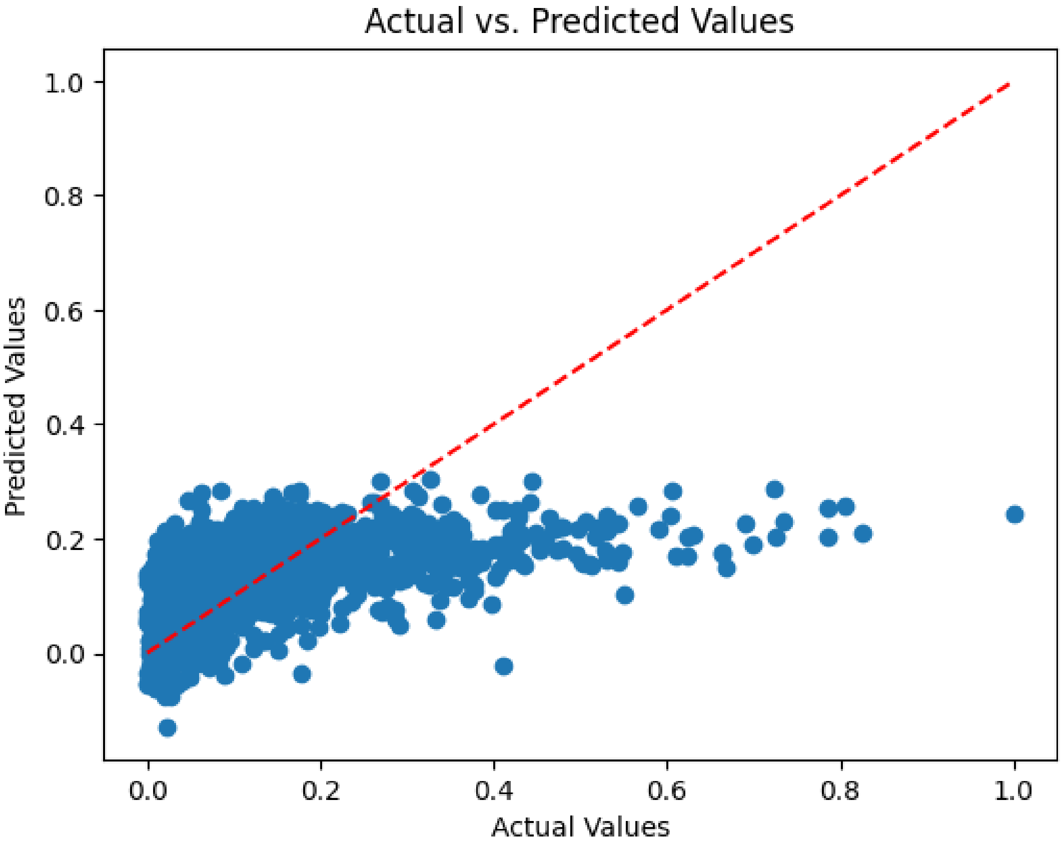
medIncome


Residual vs. Fitted Values
Residuals
Fitted Values
0.8
0.6
0.4
0.2
0.0
−0.2
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20

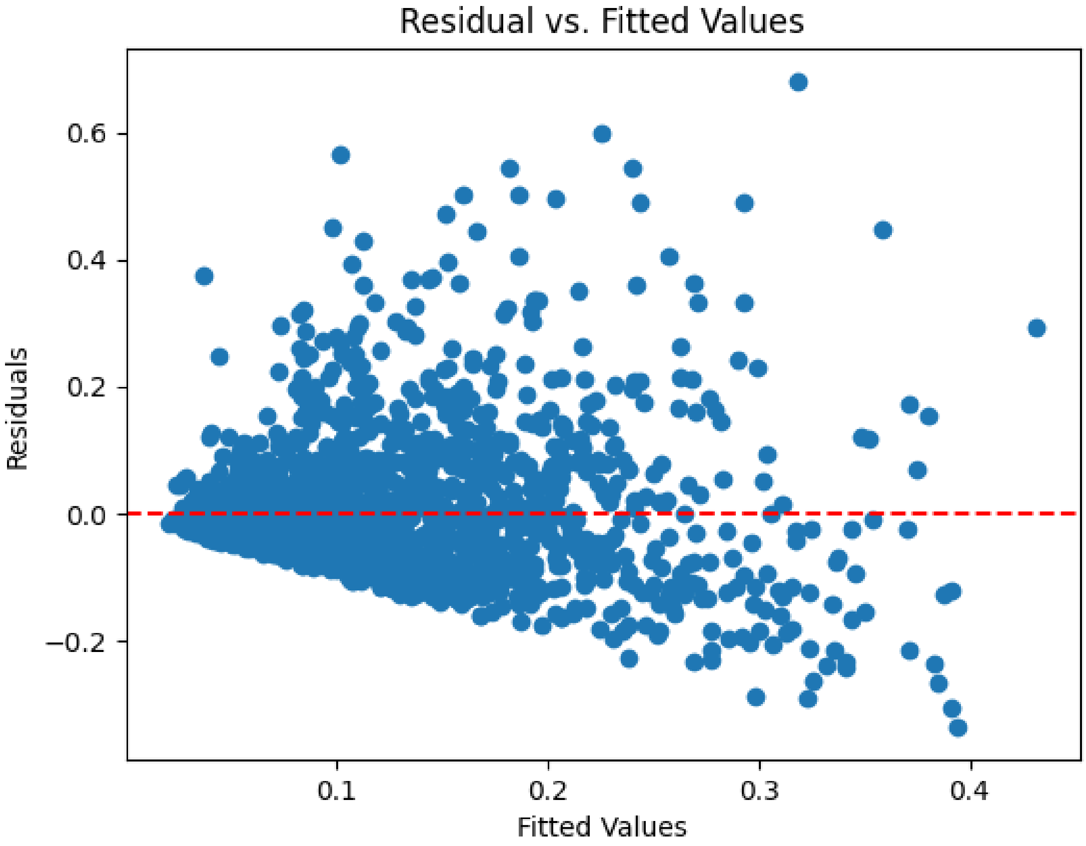

Actual vs. Predicted Values
Predicted Values
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Actual Values

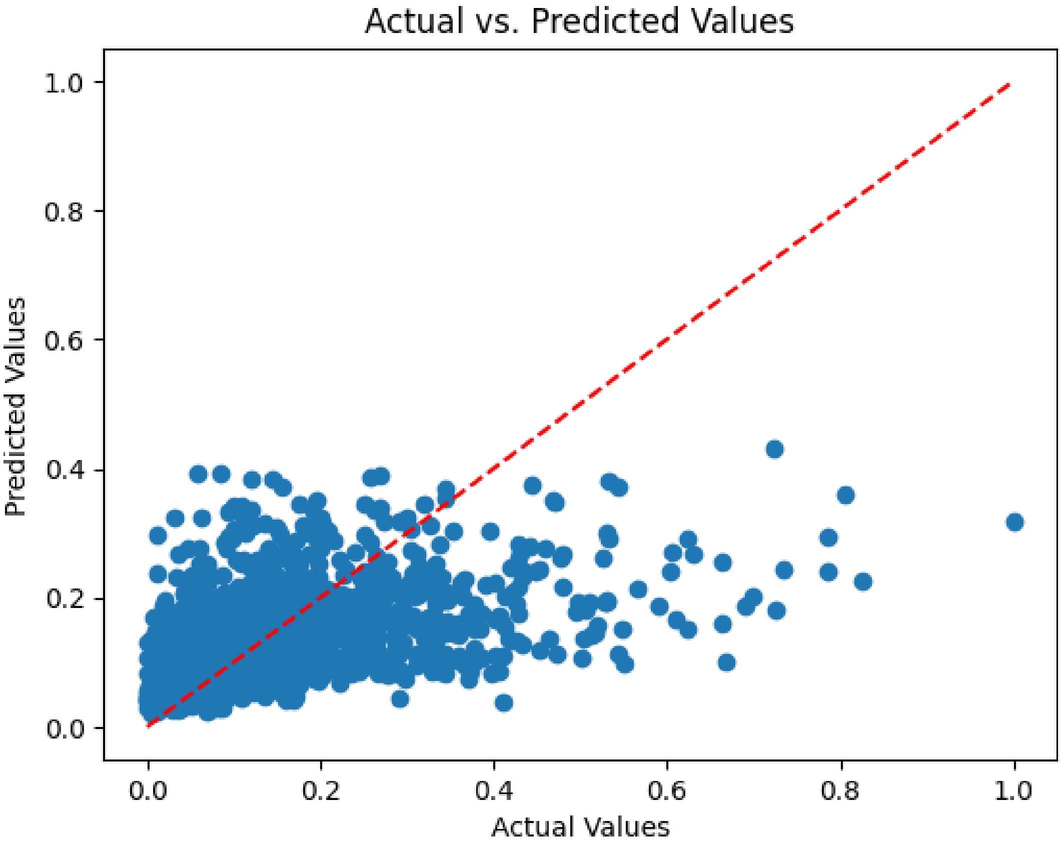
pctWInvInc


Residual vs. Fitted Values
Residuals
0.8
0.6
0.4
0.2
0.0
−0.2
−0.1
0.0
0.1
0.2
0.3
Fitted Values

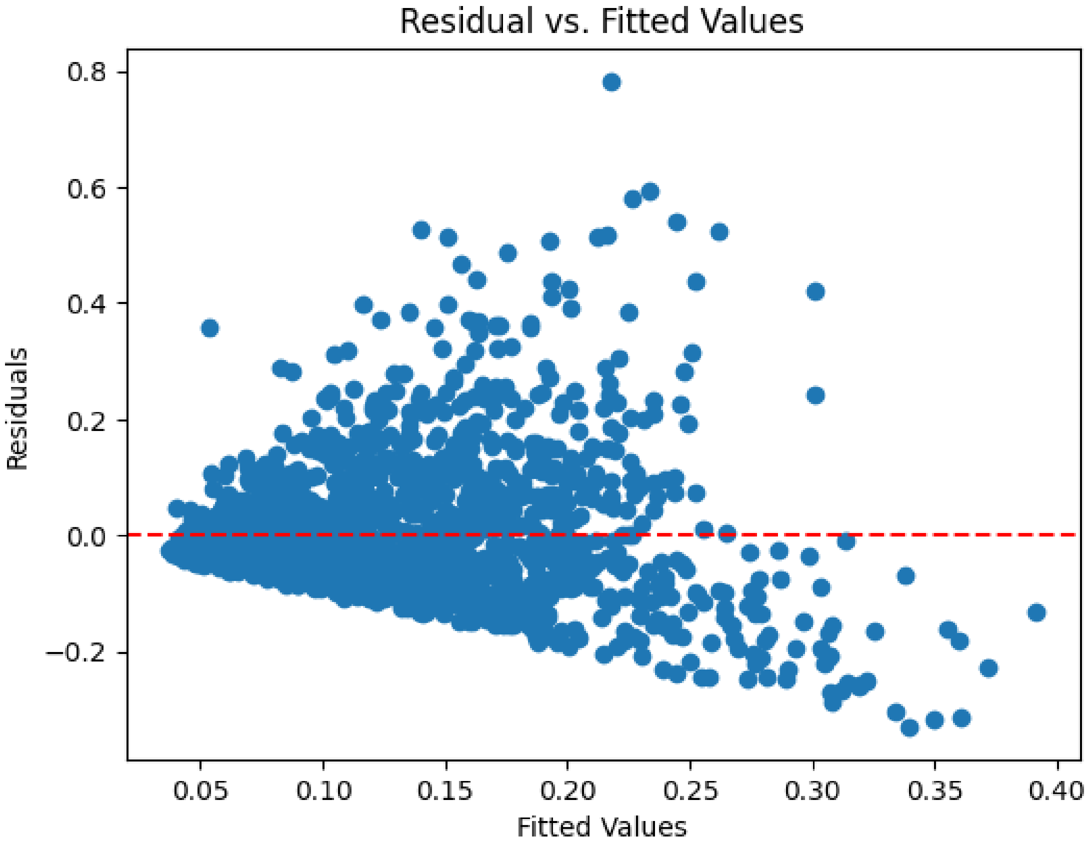


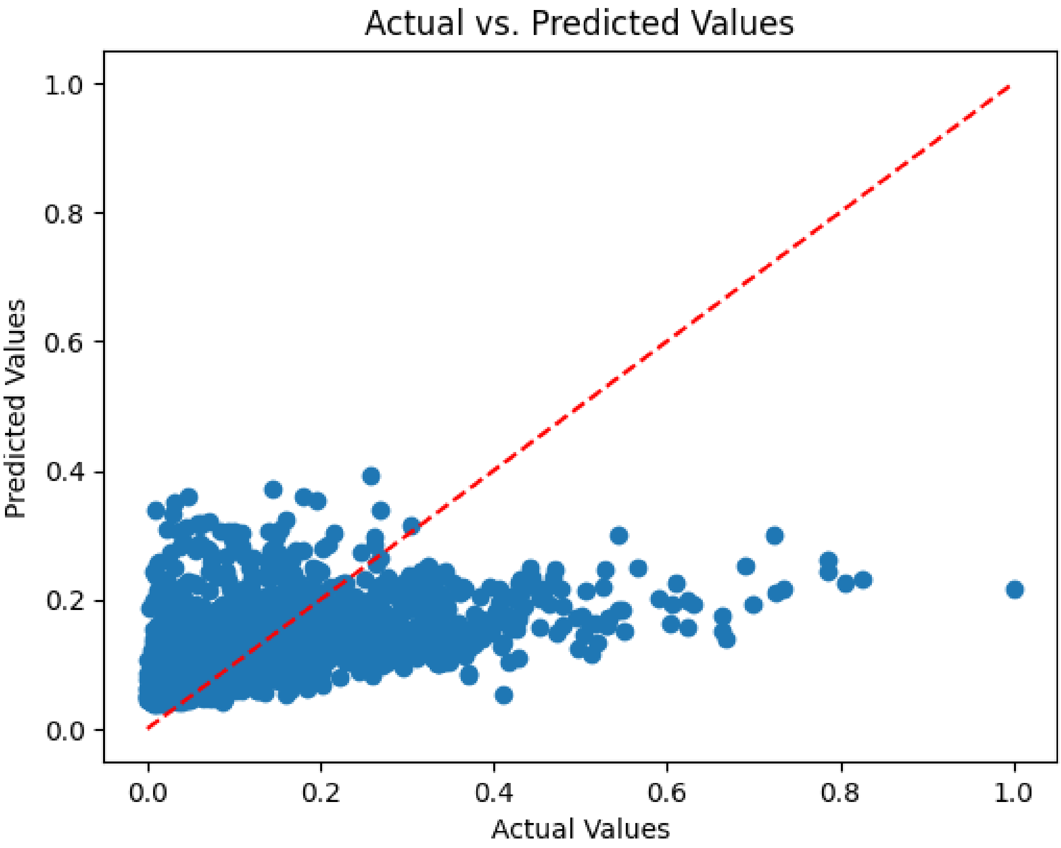
pctWPubAsst


Residual vs. Fitted Values
Residuals
0.6
0.4
0.2
0.0
−0.2
0.1
0.2
0.3
0.4
Fitted Values

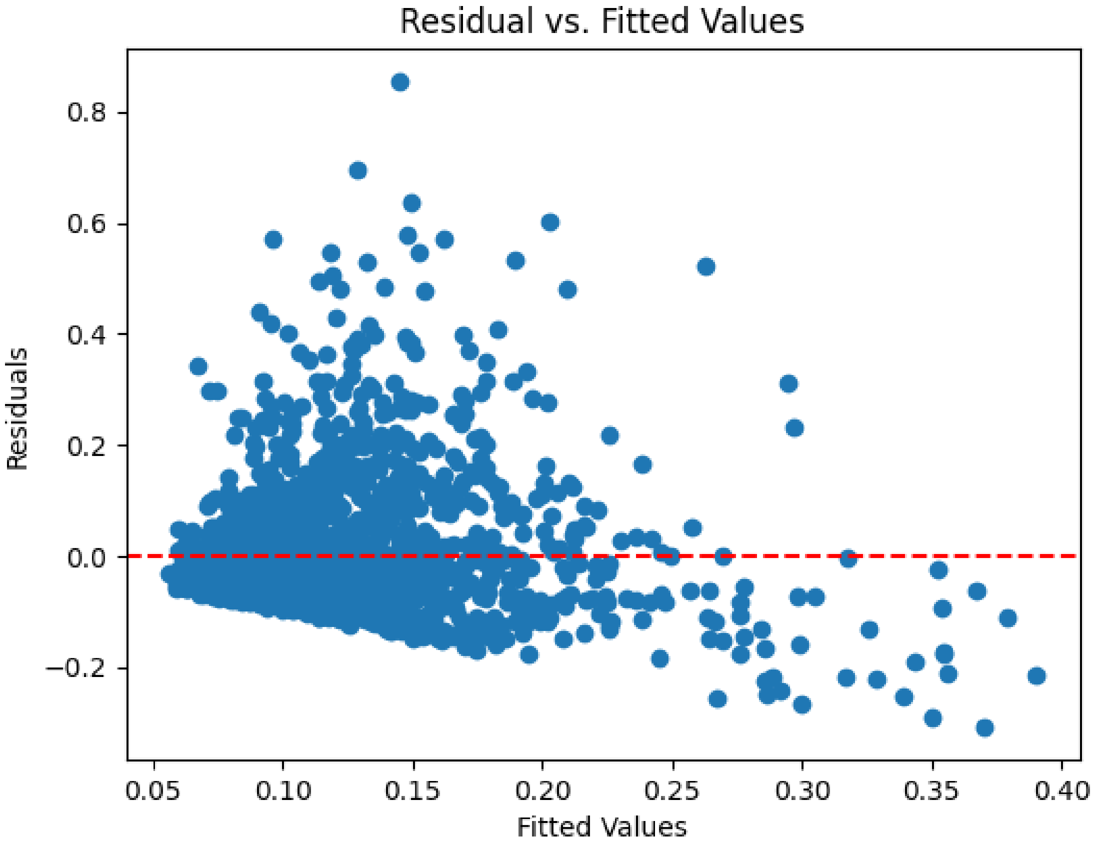


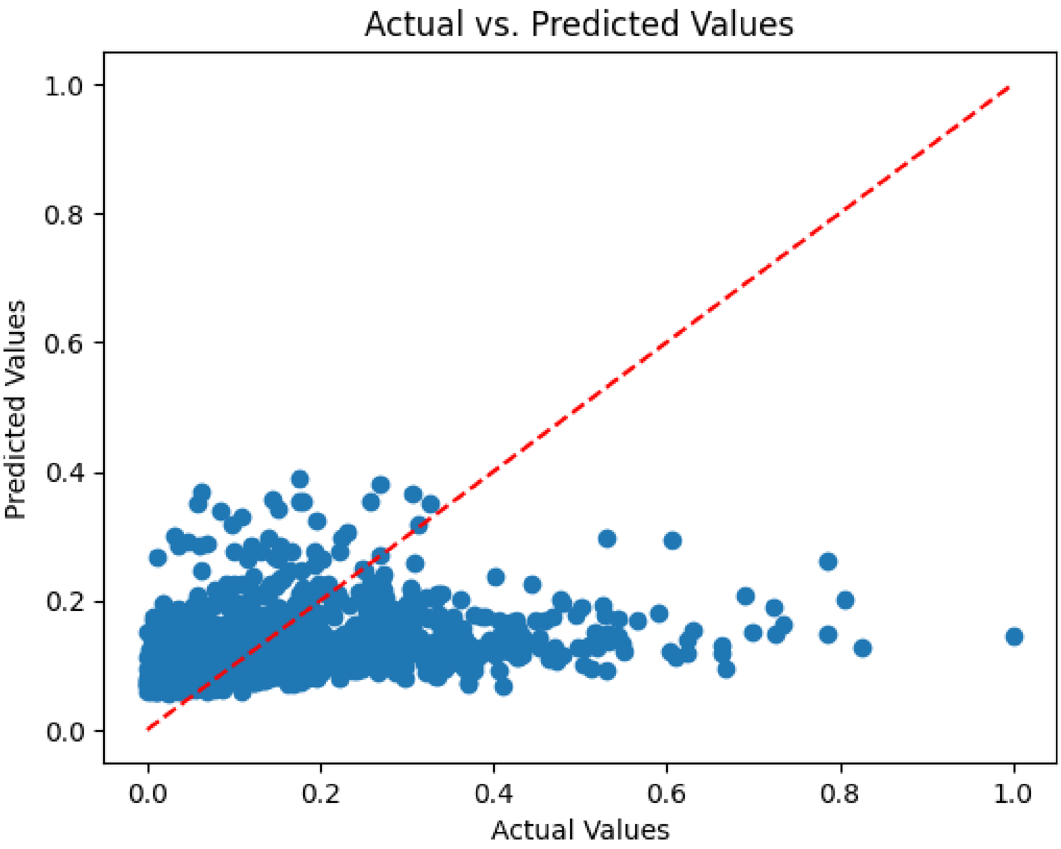
PctPopUnderPov


Residual vs. Fitted Values
Residuals
0.8
0.6
0.4
0.2
0.0
−0.2
Fitted Values
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

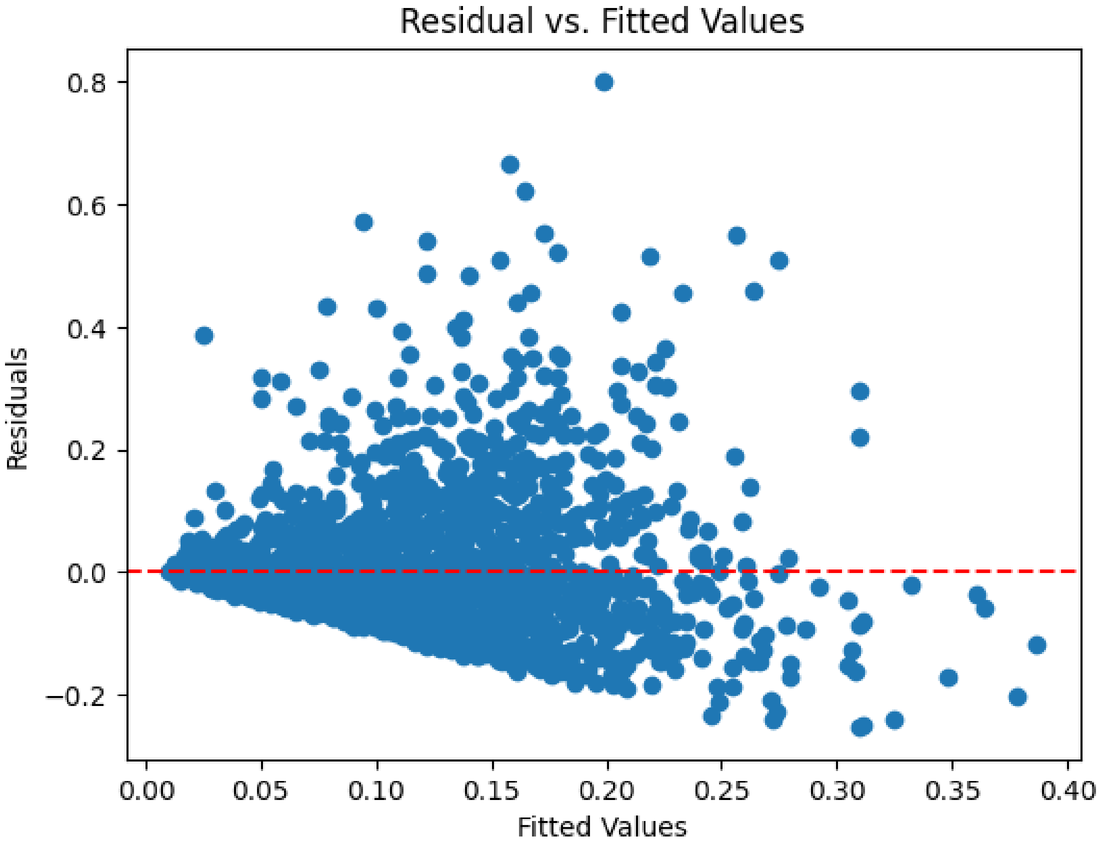


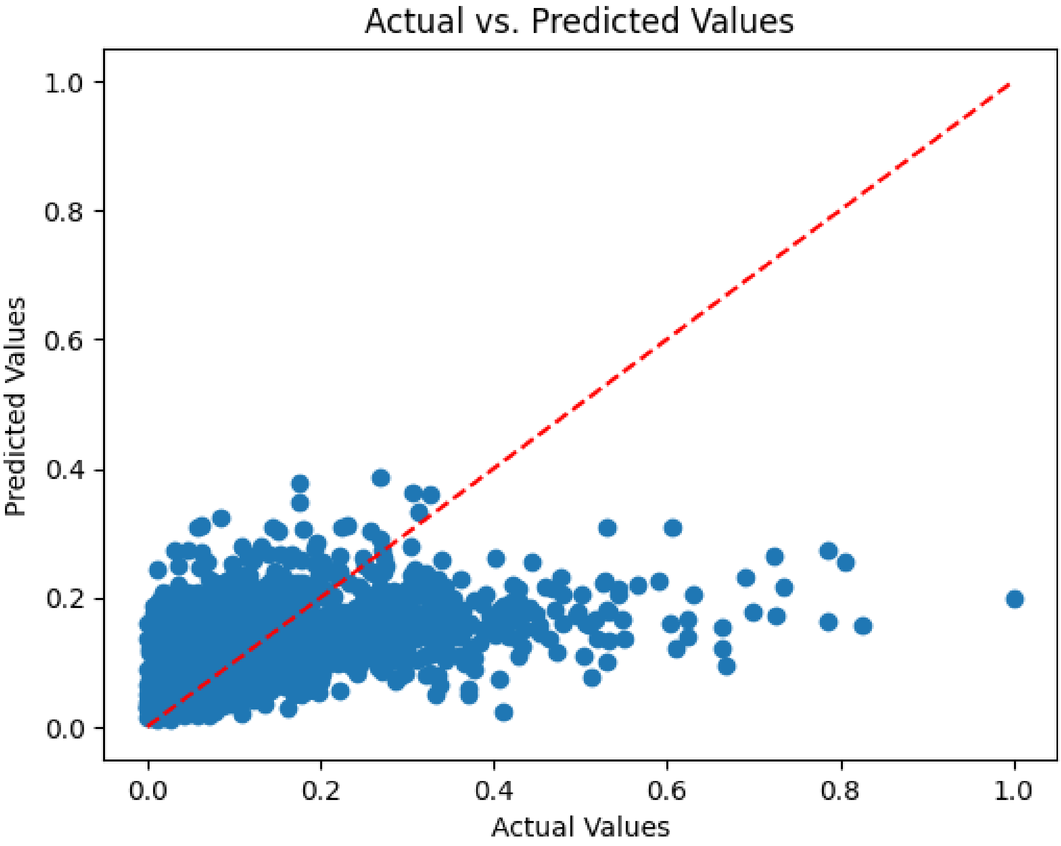
PctLess9thGrade


Residual vs. Fitted Values
Residuals
0.8
0.6
0.4
0.2
0.0
−0.2
Fitted Values
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

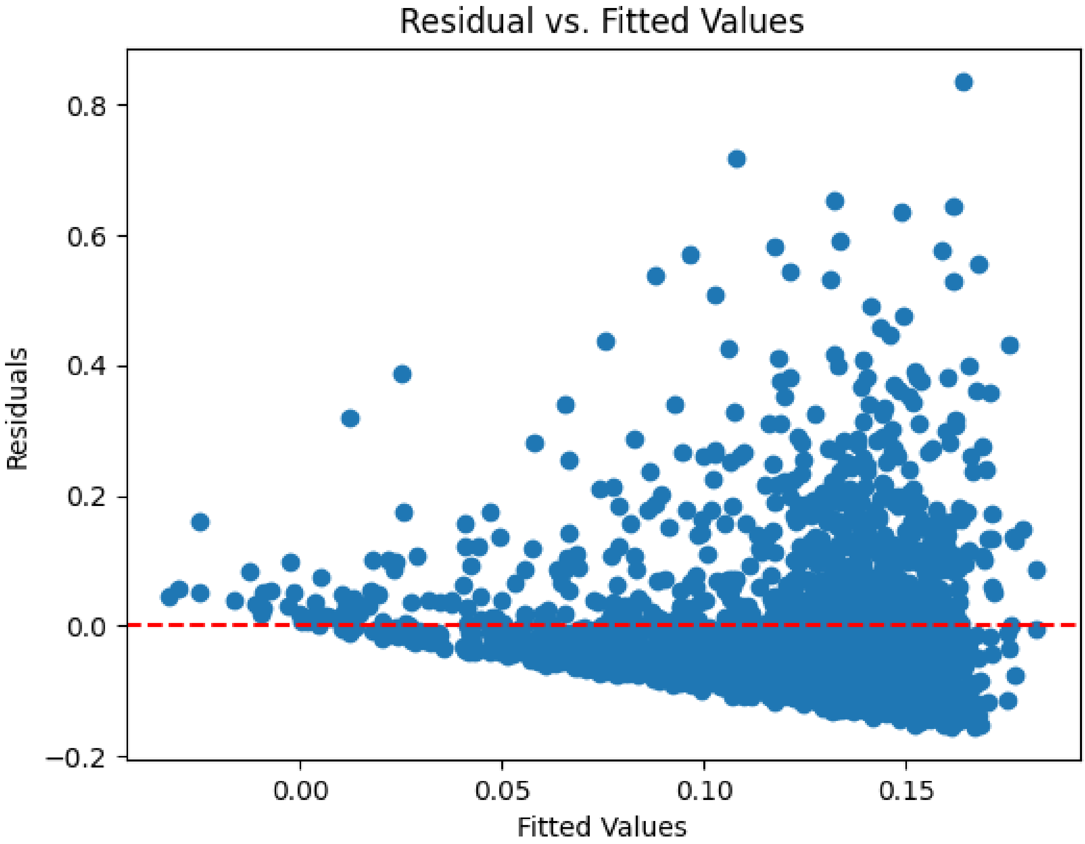


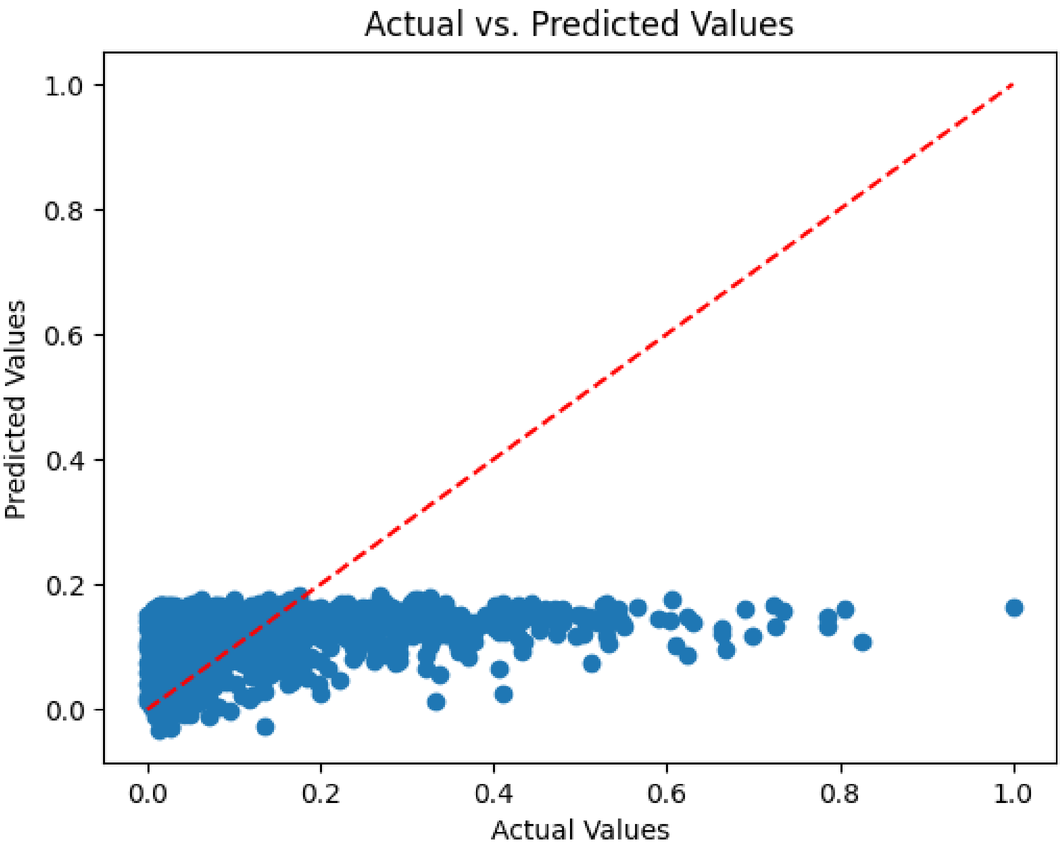
PctNotHSGrad


Residual vs. Fitted Values
Residuals
Fitted Values
0.8
0.6
0.4
0.2
0.0
−0.2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

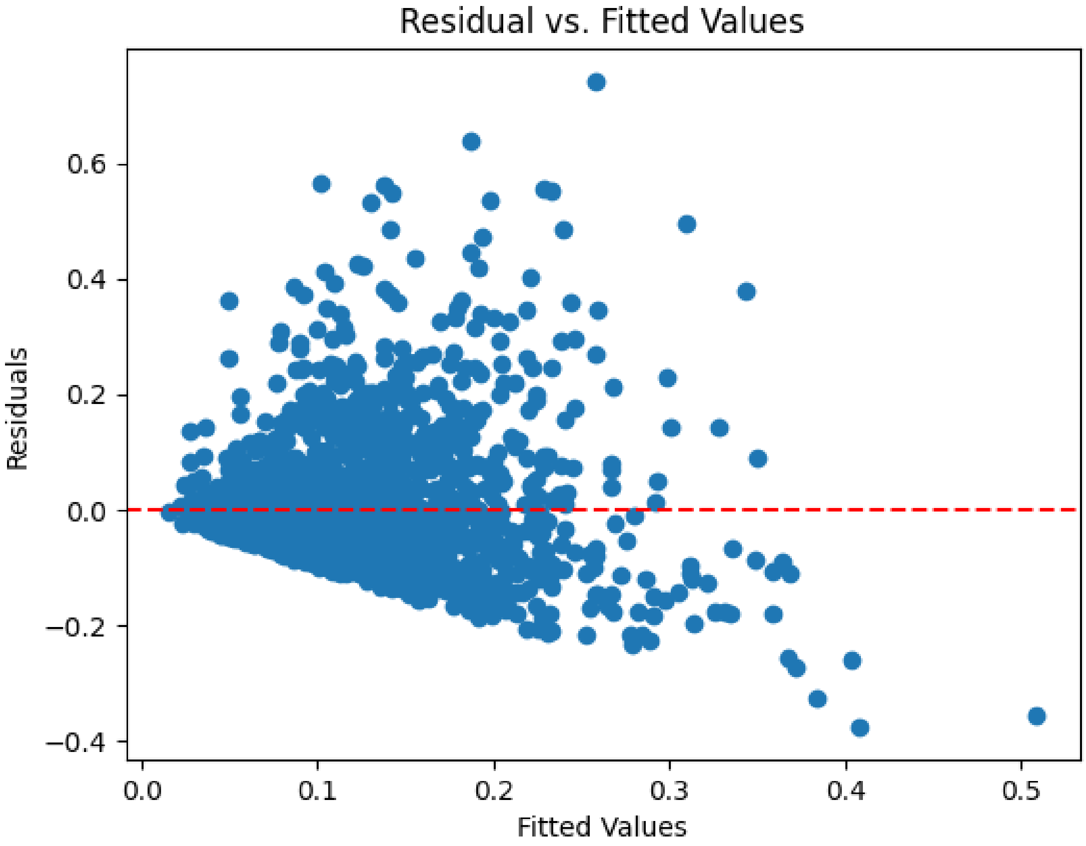


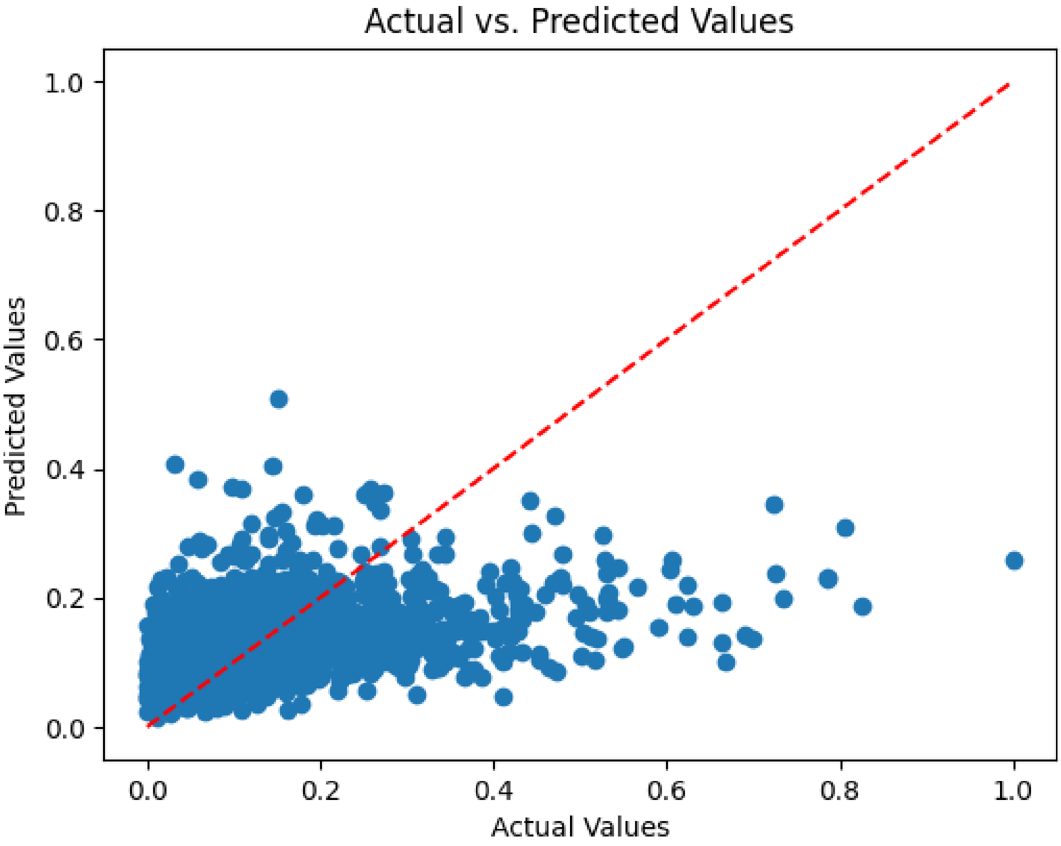
PctBSorMore


Residual vs. Fitted Values
Residuals
Fitted Values
0.8
0.6
0.4
0.2
0.0
−0.2
0.00
0.05
0.10
0.15

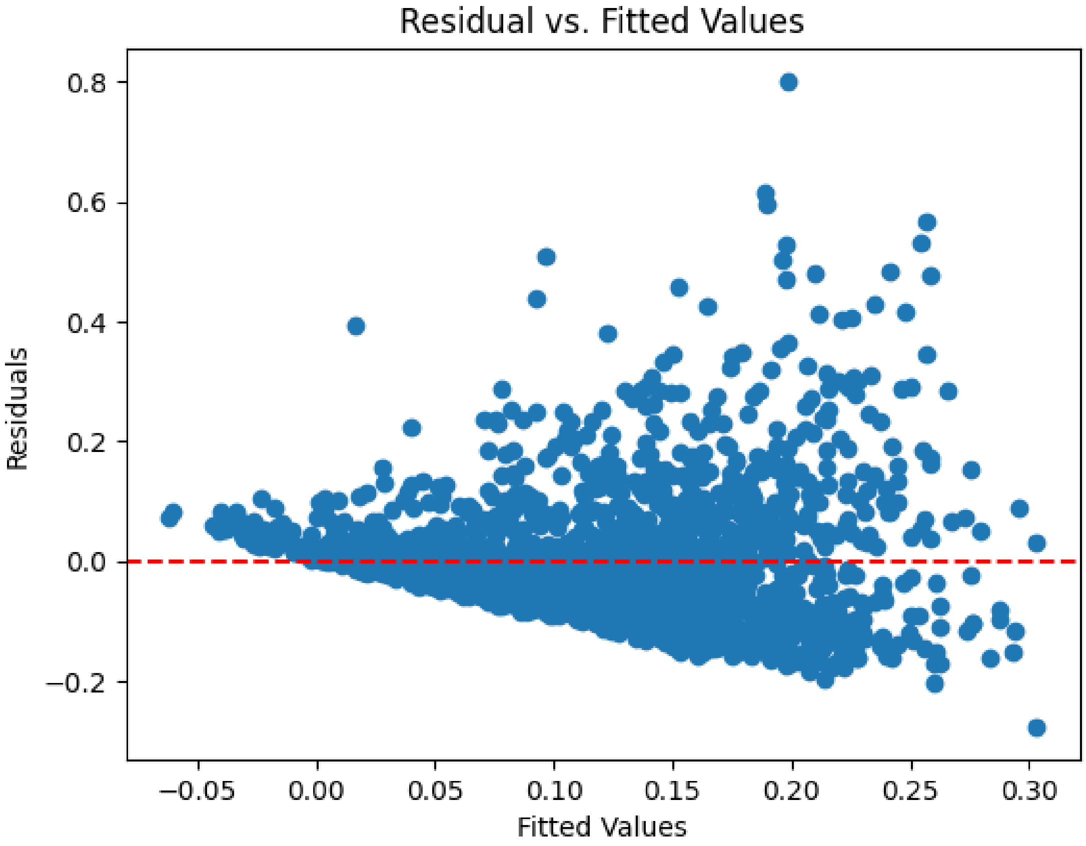


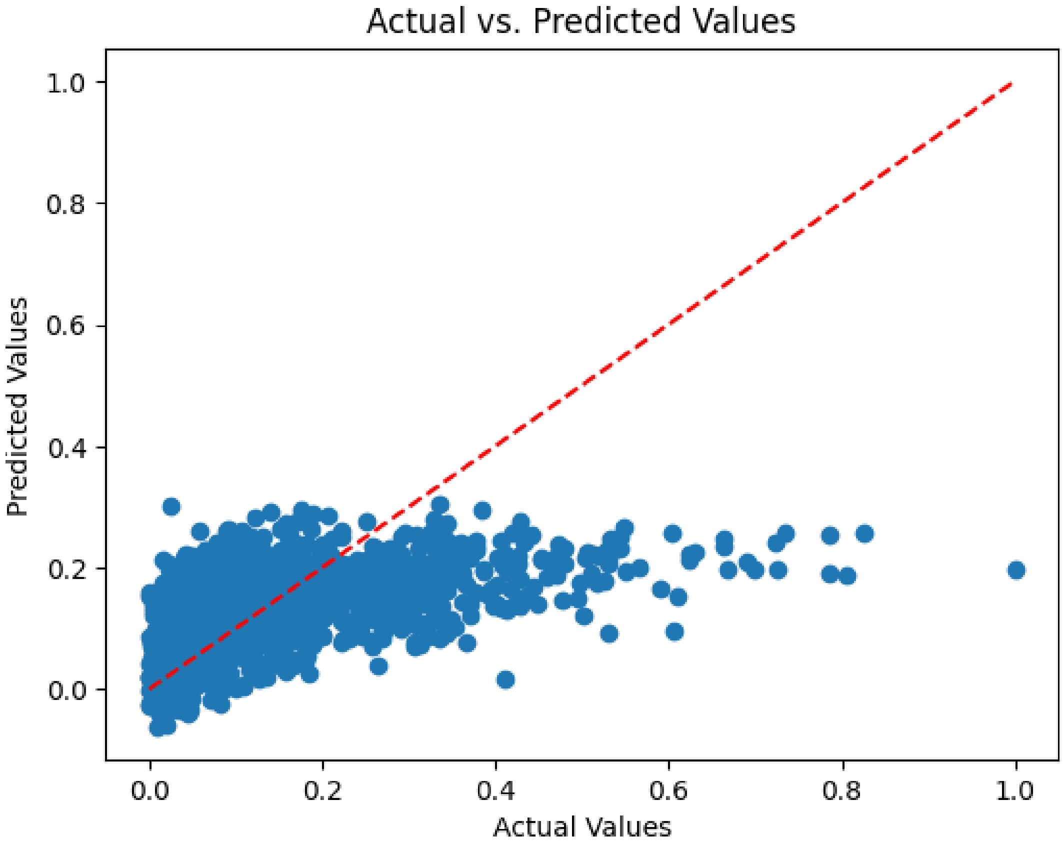
PctUnemployed


Residual vs. Fitted Values
Residuals
0.6
0.4
0.2
0.0
−0.2
−0.4
Fitted Values
0.0
0.1
0.2
0.3
0.4
0.5

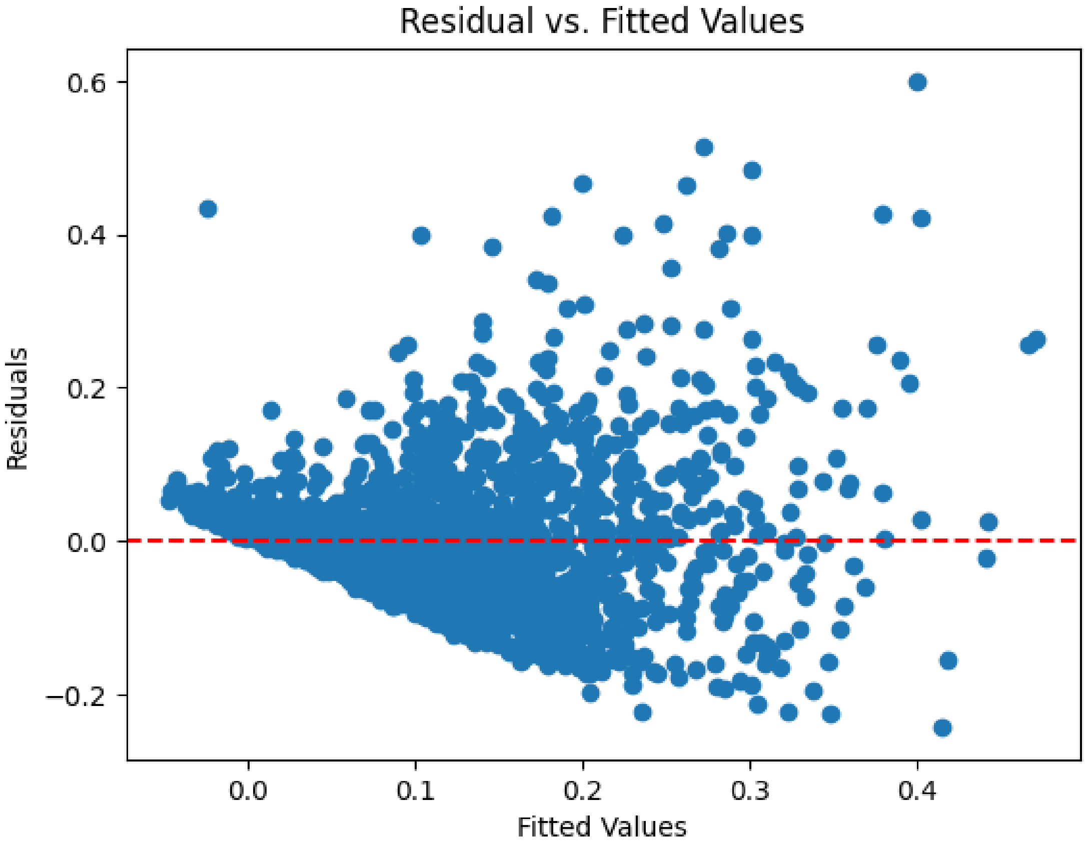


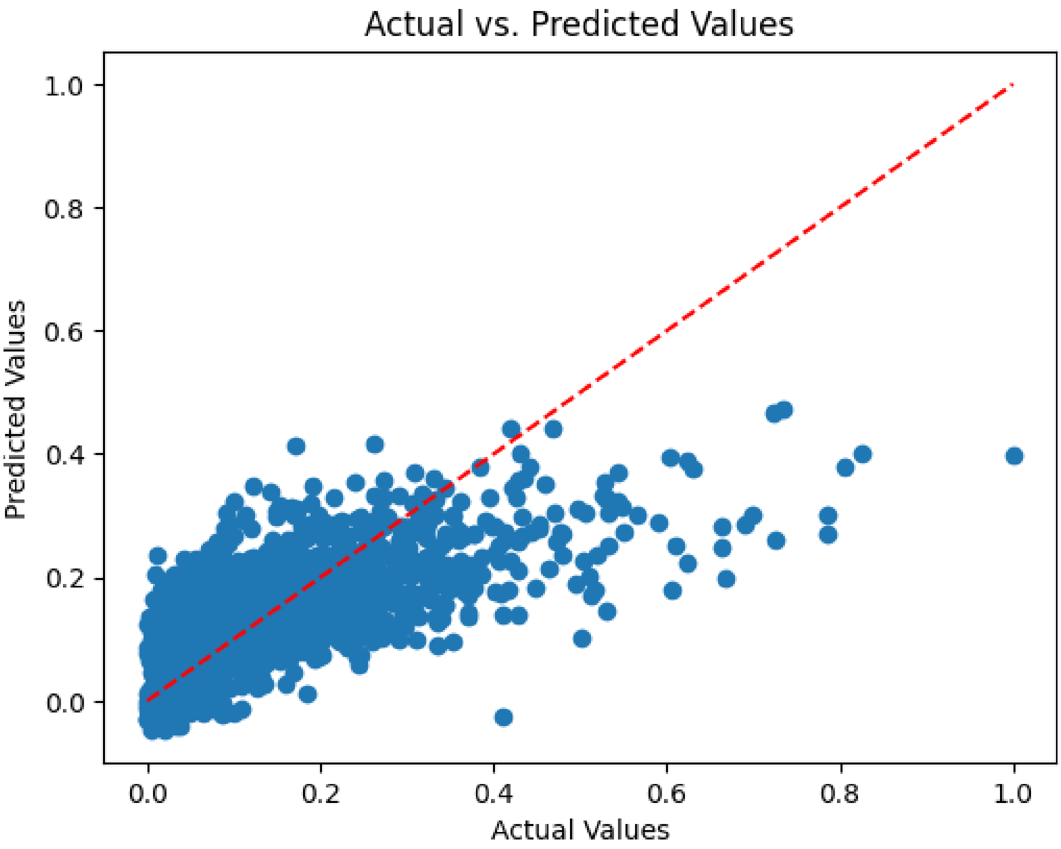
TotalPctDiv


Residual vs. Fitted Values
Residuals
Fitted Values
0.8
0.6
0.4
0.2
0.0
−0.2
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30

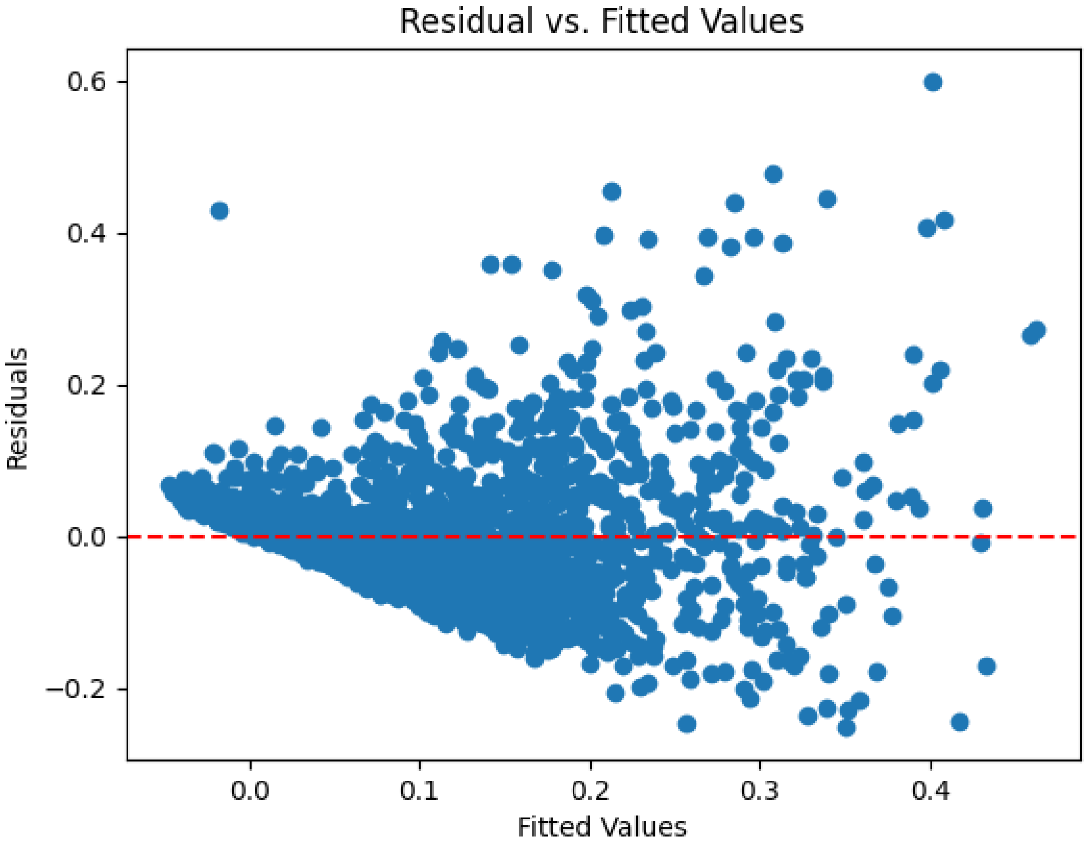

Actual vs. Predicted Values
Predicted Values
Actual Values
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

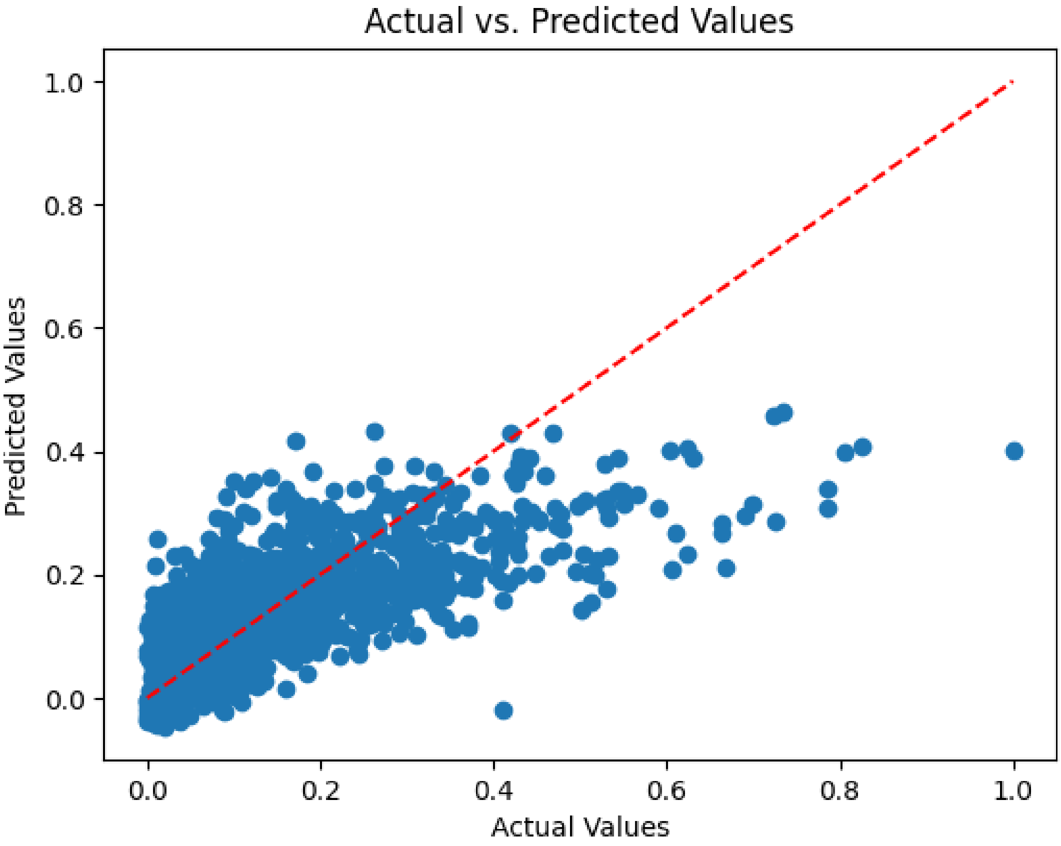
PctFam2Par


Residual vs. Fitted Values
Residuals
Fitted Values
0.6
0.4
0.2
0.0
−0.2
0.0
0.1
0.2
0.3
0.4

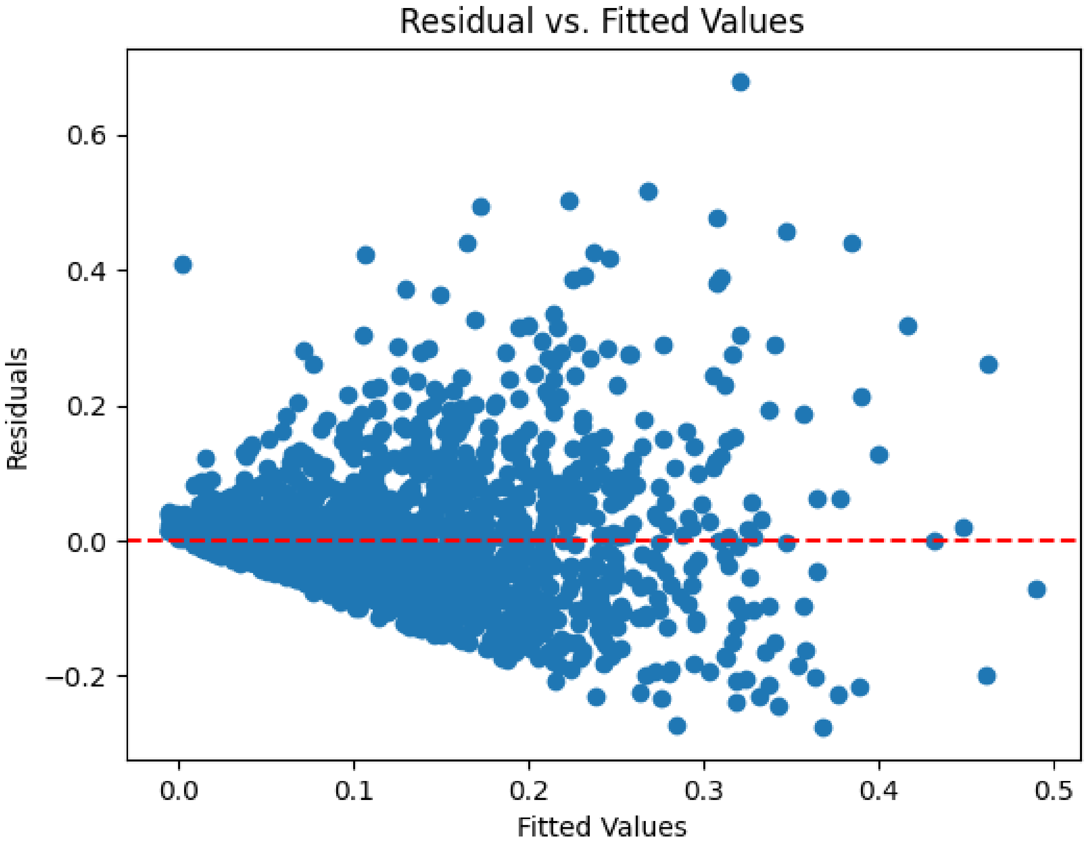


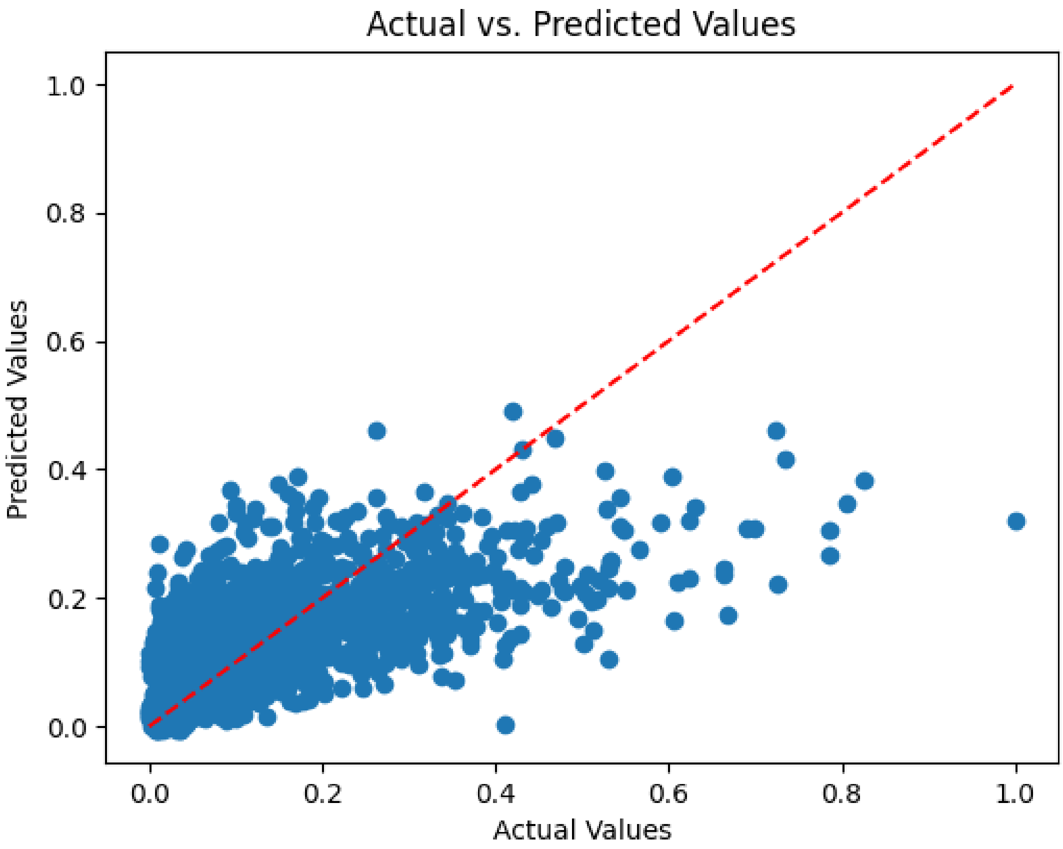
PctKids2Par


Residual vs. Fitted Values
Residuals
0.6
0.4
0.2
0.0
−0.2
Fitted Values
0.0
0.1
0.2
0.3
0.4

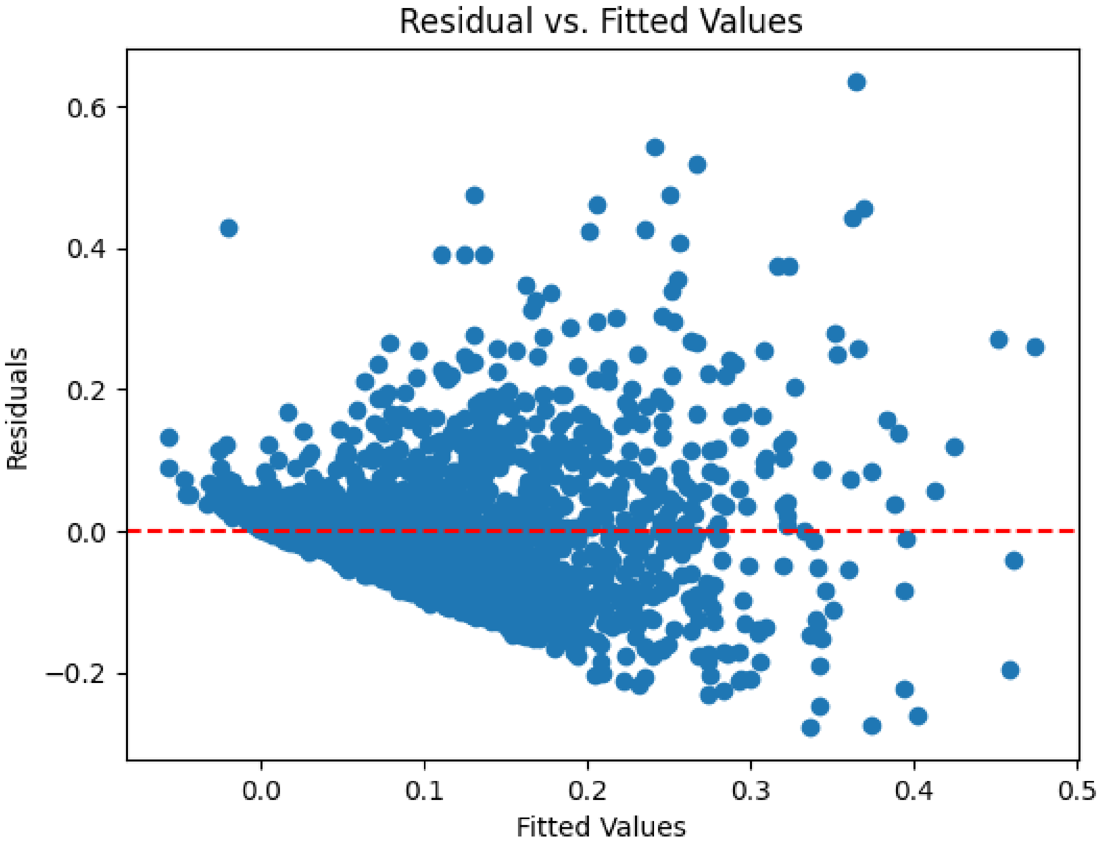


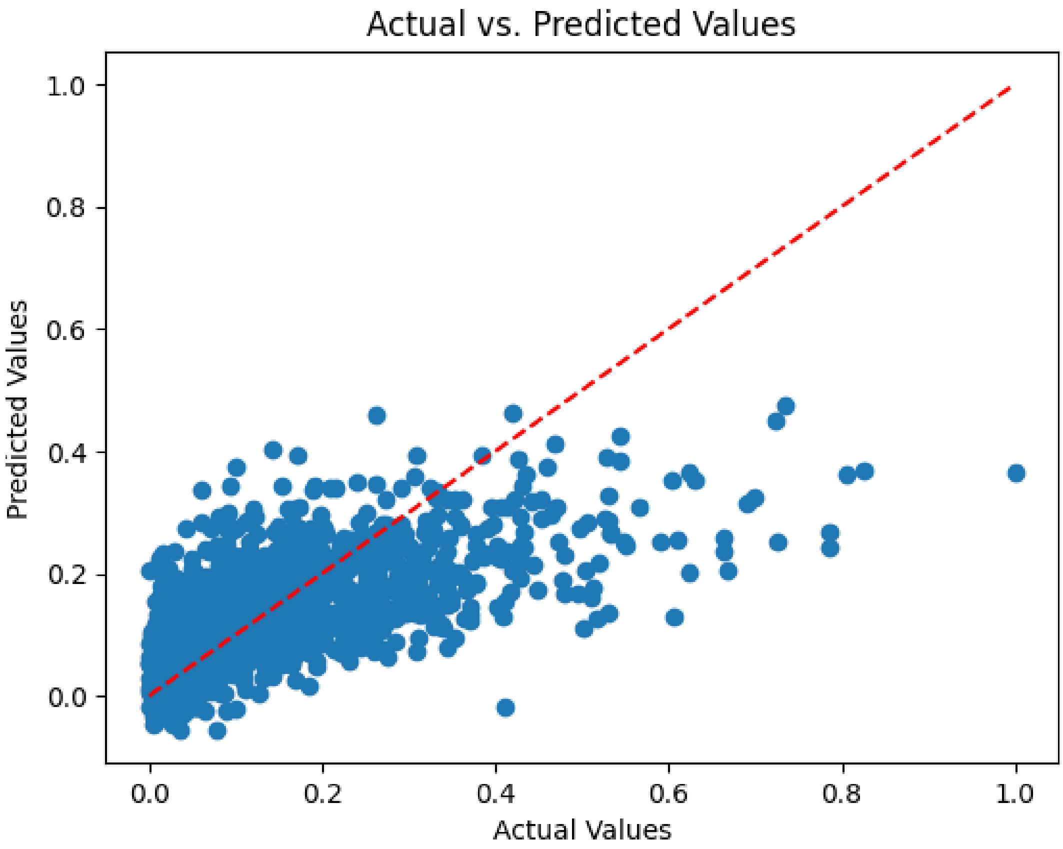
PctYoungKids2Par


Residual vs. Fitted Values
Residuals
0.6
0.4
0.2
0.0
−0.2
Fitted Values
0.0
0.1
0.2
0.3
0.4
0.5

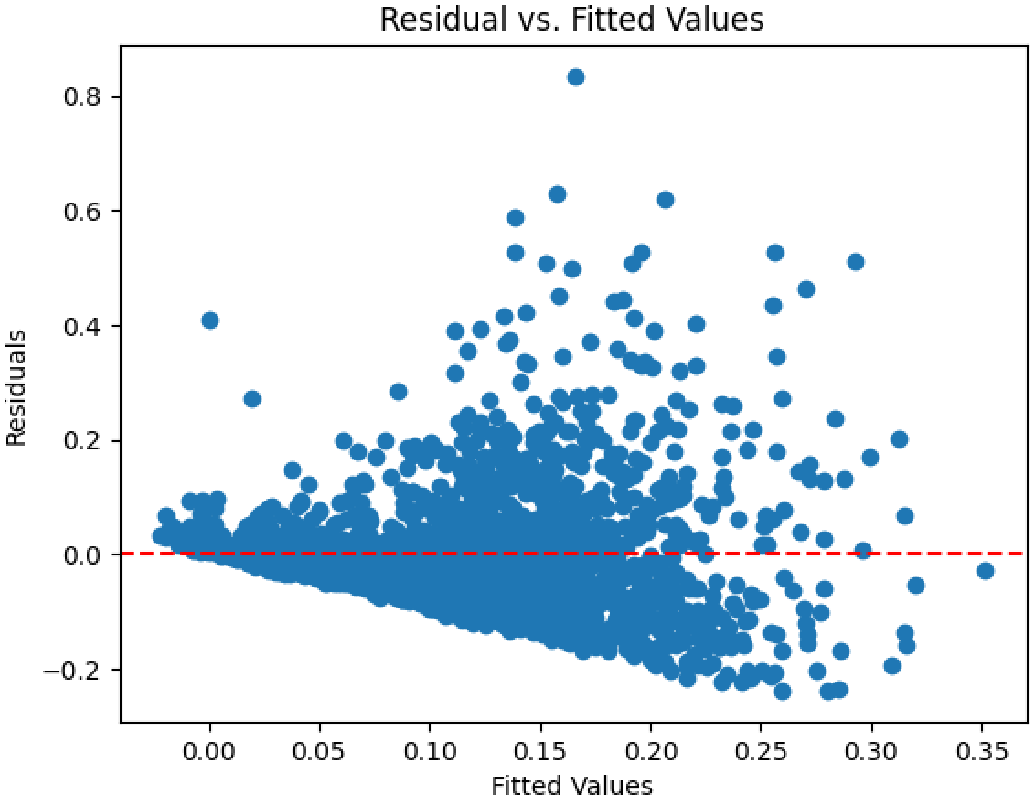


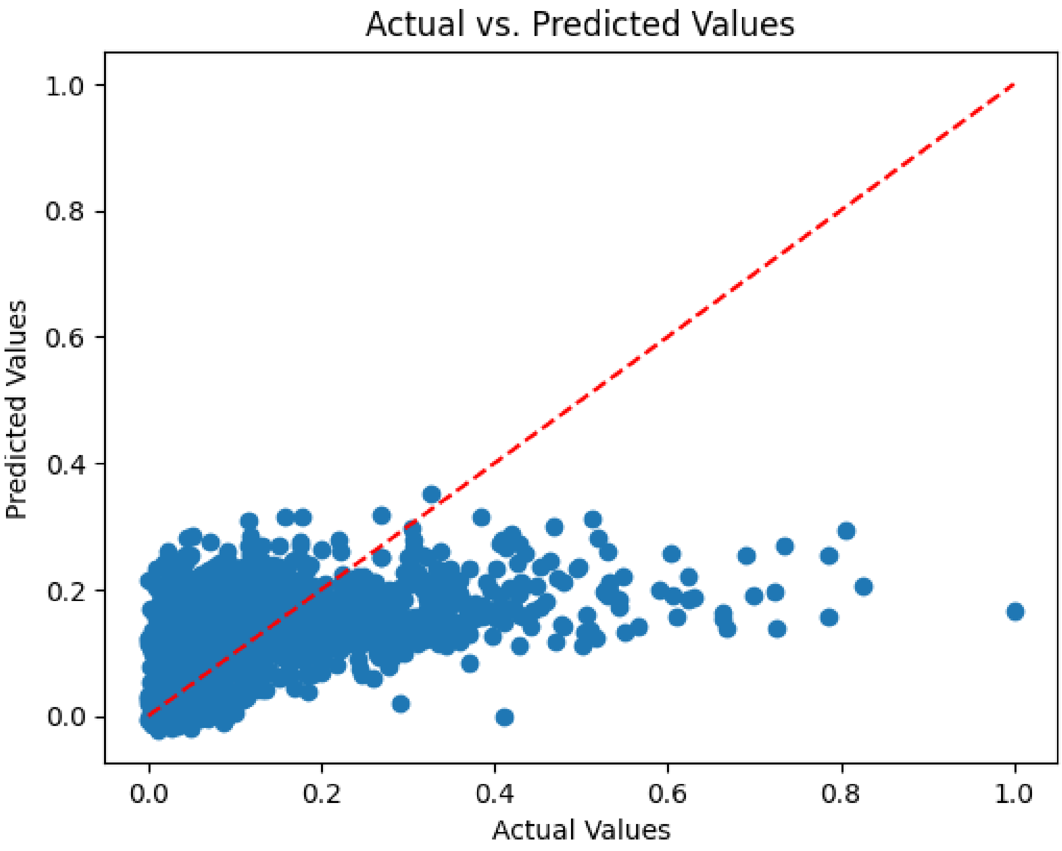
PctTeen2Par


Residual vs. Fitted Values
Residuals
0.6
0.4
0.2
0.0
−0.2
Fitted Values
0.0
0.1
0.2
0.3
0.4
0.5

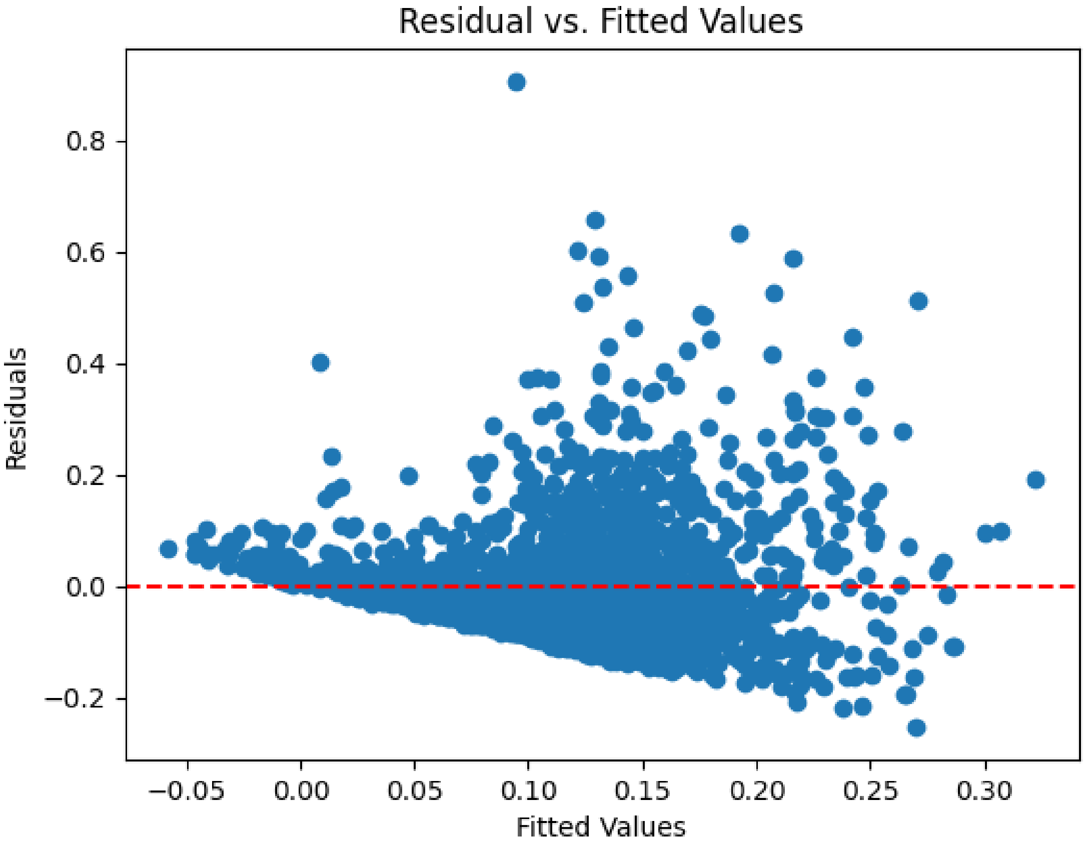


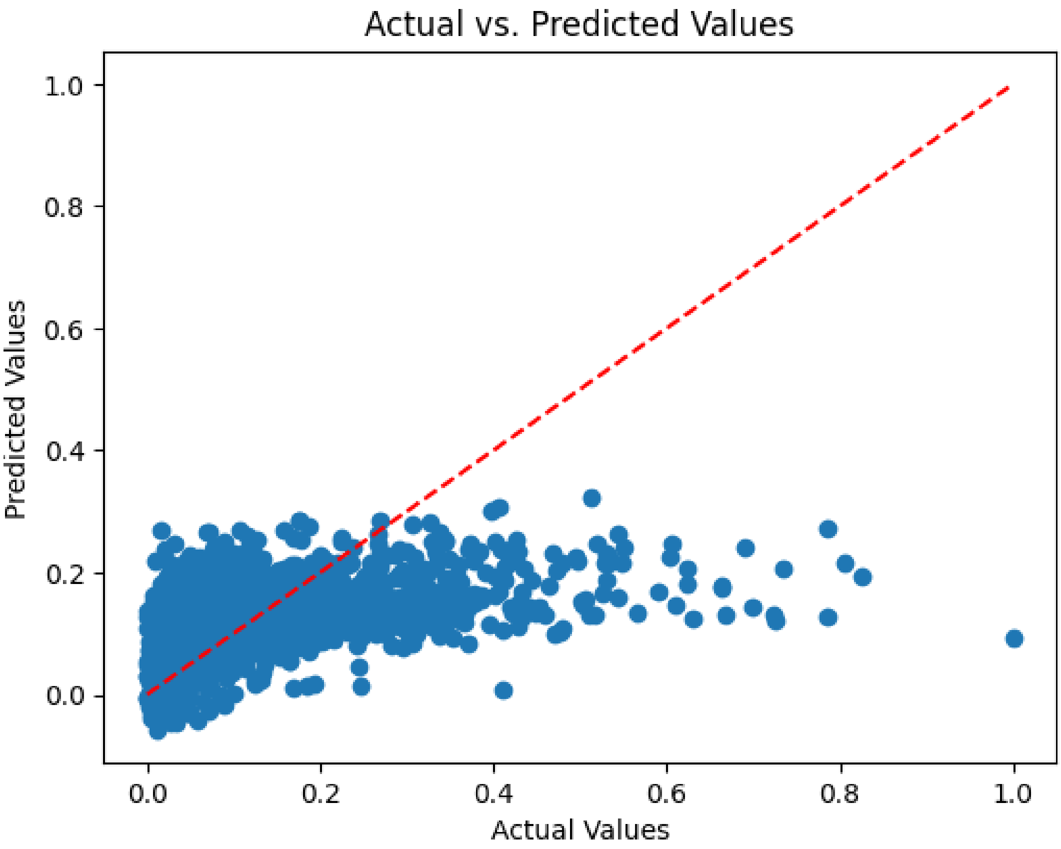
PctPersOwnOccup


Residual vs. Fitted Values
Residuals
0.8
0.6
0.4
0.2
0.0
−0.2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Fitted Values

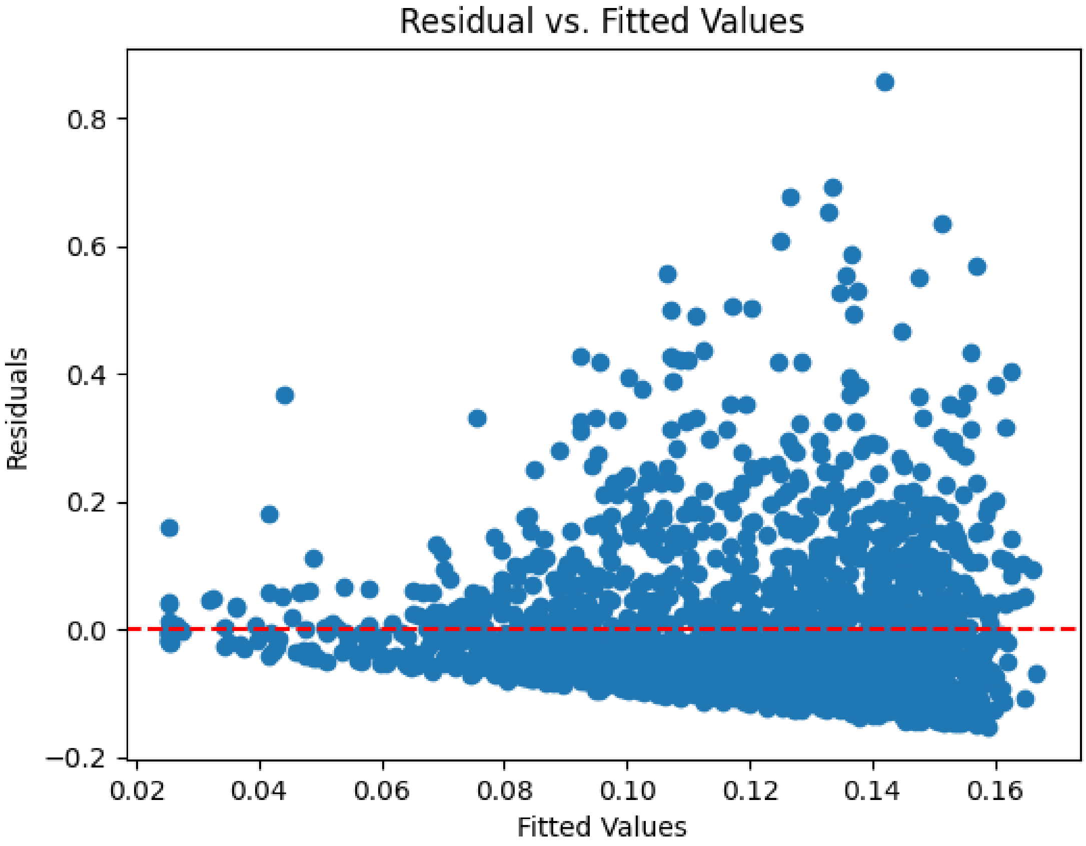


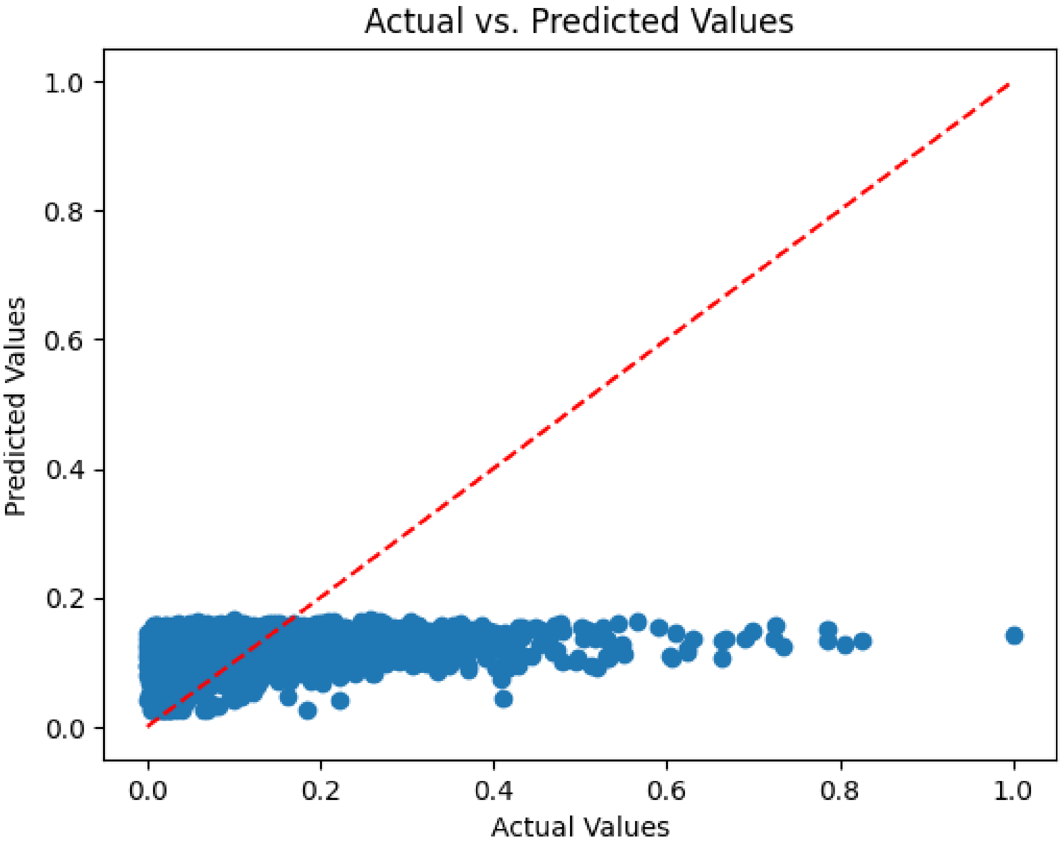
PctHousLess3BR


Residual vs. Fitted Values
Residuals
0.8
0.6
0.4
0.2
0.0
−0.2
Fitted Values
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30


Actual vs. Predicted Values
Predicted Values
1.0
0.8
0.6
0.4
0.2
0.0
Actual Values
0.0
0.2
0.4
0.6
0.8
1.0

RentMedian


Residual vs. Fitted Values
Residuals
Fitted Values
0.8
0.6
0.4
0.2
0.0
−0.2
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16


Actual vs. Predicted Values
Predicted Values
0.0
0.2
0.4
0.6
0.8
1.0
Actual Values
0.0
0.2
0.4
0.6
0.8
1.0

Runs	Accuracy	Accuracy using best parameters
Run 1	0.5590551181102362	0.6089238845144357
Run 2	0.5616797900262467	0.6115485564304461
Run 3	0.5511811023622047	0.6220472440944882
Run 4	0.5695538057742782	0.6456692913385826
Run 5	0.5485564304461942	0.6666666666666666
Run 6	0.5721784776902887	0.6325459317585301
Run 7	0.5695538057742782	0.6456692913385826
Run 8	0.5485564304461942	0.6430446194225722
Run 9	0.5538057742782152	0.6509186351706037
Run 10	0.5748031496062992	0.6404199475065617

Table showing accuracy for partially cleaned data

Runs	Accuracy	Accuracy using best parameters
Run 1	0.6062992125984252	0.6272965879265092
Run 2	0.5984251968503937	0.6561679790026247
Run 3	0.6115485564304461	0.6272965879265092
Run 4	0.5879265091863517	0.6771653543307087
Run 5	0.6010498687664042	0.6535433070866141
Run 6	0.5984251968503937	0.6220472440944882
Run 7	0.5905511811023622	0.6404199475065617
Run 8	0.6036745406824147	0.6561679790026247
Run 9	0.5958005249343832	0.6325459317585301
Run 10	0.6062992125984252	0.6509186351706037

Table showing accuracy using the 10 highest R2 values

Runs	Accuracy	Accuracy using best parameters
Run 1	0.5514950166112956	0.5980066445182725
Run 2	0.5415282392026578	0.6112956810631229
Run 3	0.5415282392026578	0.6345514950166113
Run 4	0.5481727574750831	0.5980066445182725
Run 5	0.5448504983388704	0.5913621262458472
Run 6	0.5514950166112956	0.6312292358803987
Run 7	0.5282392026578073	0.5780730897009967
Run 8	0.5548172757475083	0.6345514950166113
Run 9	0.53156146179402	0.6046511627906976
Run 10	0.5382059800664452	0.6013289036544851

Table showing accuracy using highest R2 score for cleaned data

Runs	Accuracy	Accuracy using best parameters
Run 1	0.6392405063291139	0.6107594936708861
Run 2	0.6265822784810127	0.6139240506329114
Run 3	0.6392405063291139	0.6075949367088608
Run 4	0.6329113924050633	0.6170886075949367
Run 5	0.6392405063291139	0.6139240506329114
Run 6	0.6234177215189873	0.6170886075949367
Run 7	0.6360759493670886	0.6044303797468354
Run 8	0.620253164556962	0.6234177215189873
Run 9	0.629746835443038	0.6170886075949367
Run 10	0.6455696202531646	0.6107594936708861

Random forest classifier for fully cleaned data

Runs	Accuracy	Accuracy using best parameters
Run 1	0.5448504983388704	0.5980066445182725
Run 2	0.5382059800664452	0.5946843853820598
Run 3	0.5481727574750831	0.6146179401993356
Run 4	0.5249169435215947	0.6212624584717608
Run 5	0.53156146179402	0.6146179401993356
Run 6	0.53156146179402	0.6013289036544851
Run 7	0.5382059800664452	0.5913621262458472
Run 8	0.521594684385382	0.6411960132890365
Run 9	0.521594684385382	0.627906976744186
Run 10	0.5182724252491694	0.6079734219269103

Decision Tree Classifier for clean Data on features chosen by literature review

Runs	Accuracy	Accuracy using best parameters
Run 1	0.5514950166112956	0.5980066445182725
Run 2	0.5415282392026578	0.6112956810631229
Run 3	0.5415282392026578	0.6345514950166113
Run 4	0.5481727574750831	0.5980066445182725
Run 5	0.5448504983388704	0.5913621262458472
Run 6	0.5514950166112956	0.6312292358803987
Run 7	0.5282392026578073	0.5780730897009967
Run 8	0.5548172757475083	0.6345514950166113
Run 9	0.53156146179402	0.6046511627906976
Run 10	0.5382059800664452	0.6013289036544851

Random Forest Classifier for cleaned Data on features chosen by literature review