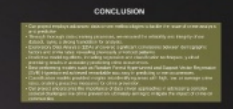
EXPLORING CRIME DATA FOR ANALYSIS & PREDICTION

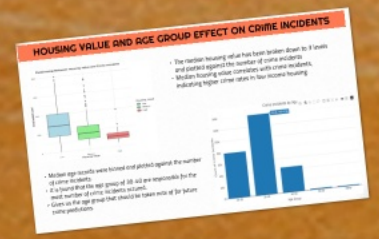
PREDICTIVE MODELING

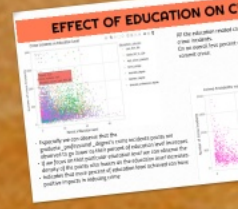


What can we do?

BY:

- Jason Rayen
- Madhuri Muppa
- Abhinav Chandoli





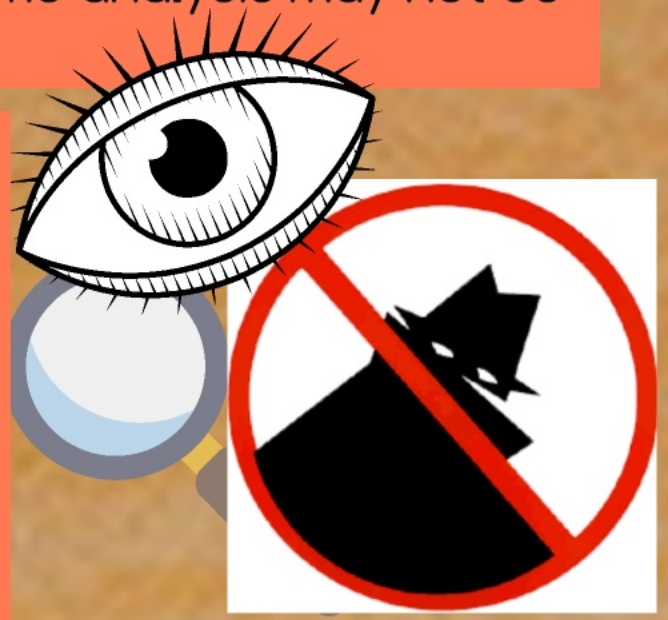
V


CONTENTS
1. Introduction
2. Previous & Current Work
3. Approach framework
4. EDA results
5. Modeling Results
6. Future work
7. Conclusion
Percent of population of arts, entertainm
Percent of population of agriculture, forestry, fi
Percent of population not in labor force and unemp
Percent of population not in labor force
Percent of population in labor force

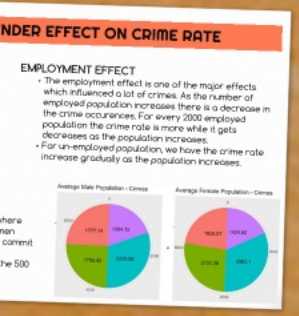
Problem and Motivation

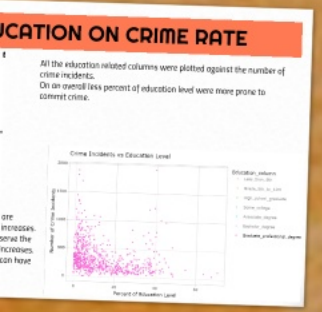
- Crime remains a significant concern for public safety.
- Traditional methods for crime analysis may not be sufficient.

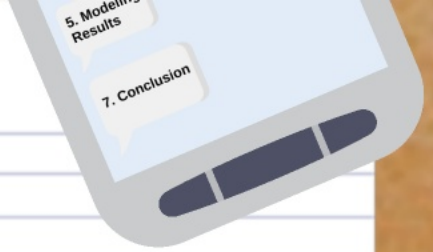
- There's a need for innovative data-driven approaches to understand and prevent crime effectively.
- Our main motivation is to help decision makers to act upon insights we will gain from this research and take suitable decisions.



What can we do?







Crime prediction

Current approaches:

- Existing methods for crime analysis may lack depth and effectiveness.
- Limited use of data analysis techniques for understanding crime patterns.
- Prediction of the number of crimes resulting in a lower accuracy and a higher RMSE value.
- No detail explanation on what the data is about.

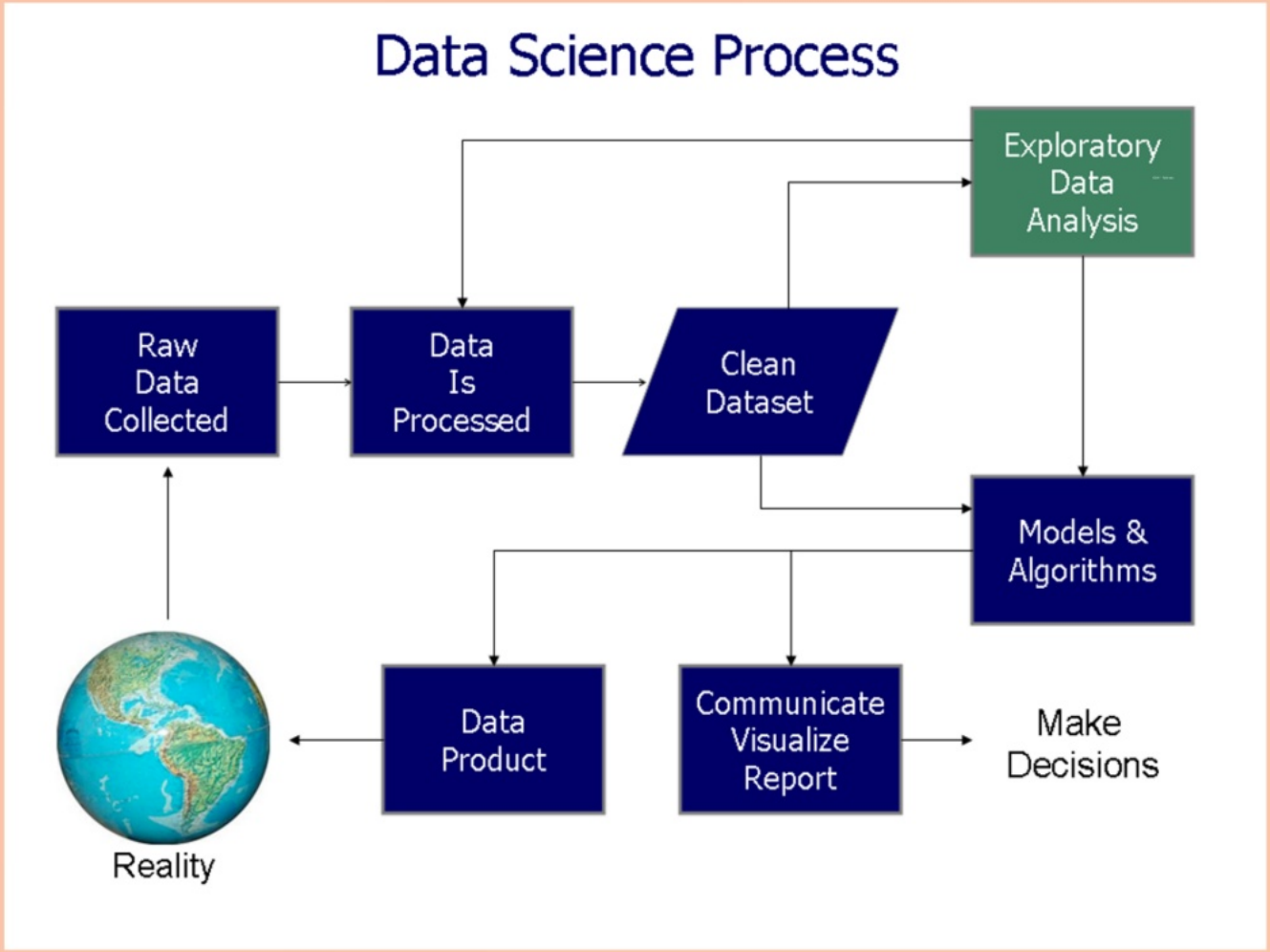
Our approaches:

- Leveraging advanced data analysis techniques, including EDA and predictive modeling.
- Focus on extracting insights from comprehensive crime data to inform proactive measures.
- Emphasis on employing machine learning algorithms for accurate prediction and hotspot identification.
- In-depth Predictive Analytics is performed with many algorithms to get the best useful results.

Problem and

- Crime remains a significa safety.
- Traditional methods for cr sufficient.
- There's a need for innovative data-driven approaches to understand and prevent crime effectively.
- Our main motivation is to help decision makers to act upon insights we will gain from this research and take suitable decisions.

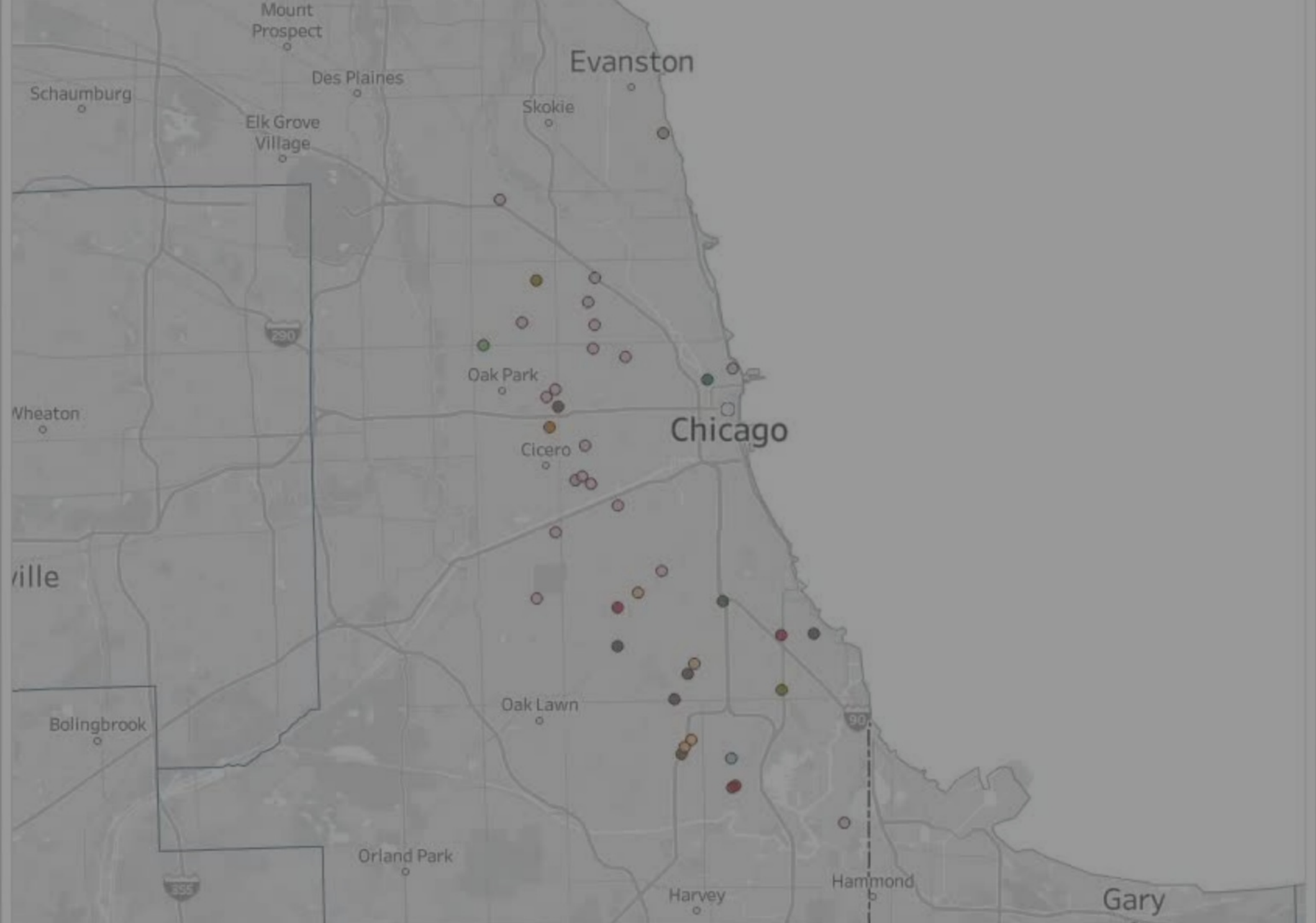
PLAN OF APPROACH



REFERENCES

DIKW paradigm and project guidance chapter of AIT664 by Dr. Ebrima Ceesay.

Image: https://commons.wikimedia.org/wiki/File:Data_visualization_process_v1.png



BURGLARY
CONCEALED CARRY LI..
CRIM SEXUAL ASSAULT
CRIMINAL DAMAGE
CRIMINAL SEXUAL AS..
CRIMINAL TRESPASS
DECEPTIVE PRACTICE
GAMBLING
HOMICIDE
INTERFERENCE WITH ..
INTIMIDATION
KIDNAPPING
LIQUOR LAW VIOLATI..
MOTOR VEHICLE THEFT
NARCOTICS
OBSCENITY
OFFENSE INVOLVING ..
OTHER OFFENSE
PROSTITUTION
PUBLIC PEACE VIOLAT..
ROBBERY
SEX OFFENSE
STALKING
THEFT
WEAPONS VIOLATION

Year

2013

DATA CLEANING

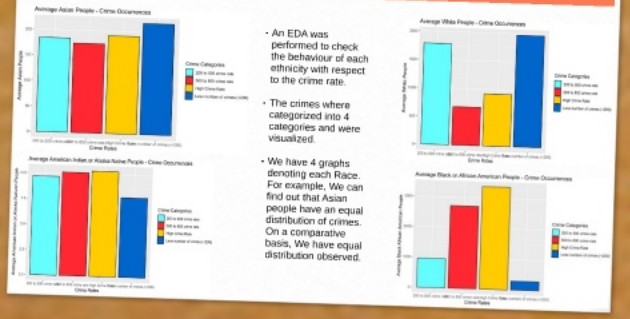
The whole dataset had very less null values in certain columns which were fixed by taking the mean of the column and filling them inplace of the Null values. Columns having Null values are displayed on the right.

Variable Name	Number of Null Values
Percent of vacant housing units	3
Percent of unemployed population	3
Percent of under 16 years	3
Percent of some college	3
Percent of population of wholesale trade	4
Percent of population of transportation, warehousing, and utilities	4
Percent of population of retail trade	4
Percent of population of public administration	4
Percent of population of professional, scientific, management, administrative, and waste management services	4
Percent of population of other services	4
Percent of population of manufacturing	4
Percent of population of information	4
Percent of population of finance, insurance, real estate, rental, and leasing	4
Percent of population of educational services, health care, and social assistance	4
Percent of population of construction	4
Percent of population of arts, entertainment, recreation, accommodation, and food services	4
Percent of population of agriculture, forestry, fishing, hunting, and mining	4
Percent of population not in labor force and unemployed population	3
Percent of population not in labor force	3
Percent of population in labor force	3

Variable Name	Number of Null Values
Percent of population not in labor force	3
Percent of population in labor force	3
Percent of over 65 years	3
Percent of over 16 years	3
Percent of occupied housing units	3
Percent of male population	3
Percent of less than high school graduate	3
Percent of less than bachelor's degree	3
Percent of less than 9th grade	3
Percent of high school graduate or higher	3
Percent of high school graduate	3
Percent of graduate or professional degree	3
Percent of female population	3
Percent of employed population	3
Percent of bachelor's degree of higher	3
Percent of bachelor's degree	3
Percent of associate's degree	3
Percent of 9th to 12th grade	3
Median income	5
Median housing value	9
Median age	3
Mean income	3

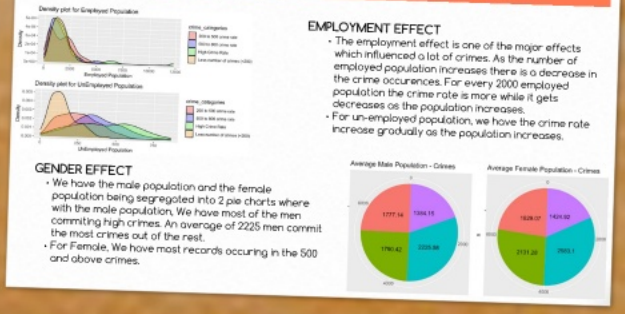
EXPLORATORY DATA ANALYSIS

RACE AGAINST CRIME OCCURENCES:



- An EDA was performed to check the behaviour of each ethnicity with respect to the crime rate.
- The crimes where categorized into 4 categories and were visualized.
- We have 4 graphs denoting each Race. For example, We can find out that Asian people have an equal distribution of crimes. On a comparative basis, We have equal distribution observed.

EMPLOYMENT AND GENDER EFFECT ON CRIME RATE



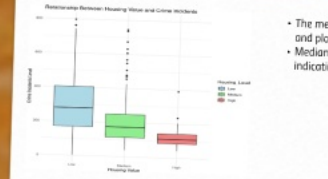
EMPLOYMENT EFFECT

- The employment effect is one of the major effects which influenced a lot of crimes. As the number of employed population increases there is a decrease in the crime occurences. For every 2000 employed population the crime rate is more while it gets decreases as the population increases.
- For un-employed population, we have the crime rate increase gradually as the population increases.

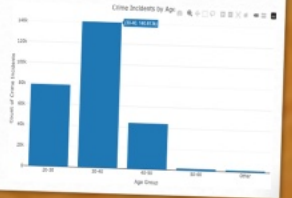
GENDER EFFECT

- We have the male population and the female population being segregated into 2 pie charts where with the male population, We have most of the men commiting high crimes. An average of 2225 men commit the most crimes out of the rest.
- For Female, We have most records occuring in the 500 and above crimes.

HOUSING VALUE AND AGE GROUP EFFECT ON CRIME INCIDENTS



- The median housing value has been broken down to 3 levels and plotted against the number of crime incidents
- Median housing value correlates with crime incidents, indicating higher crime rates in low income housing.



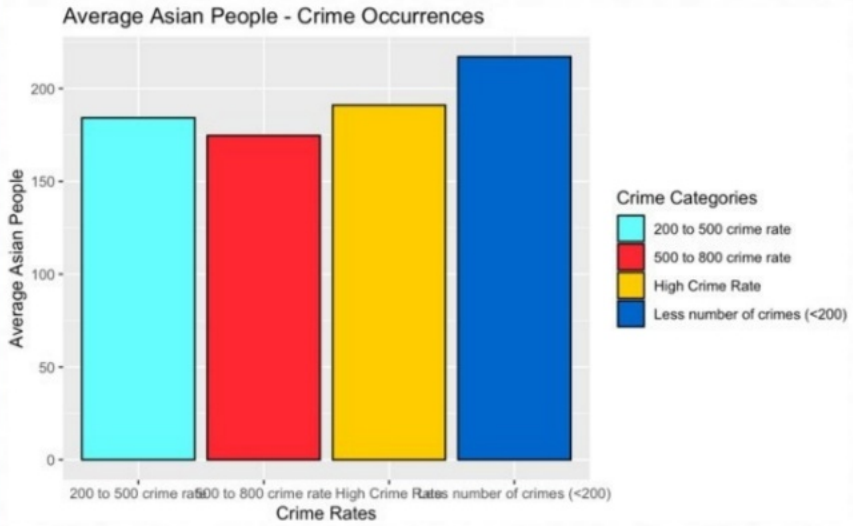
- Median age records were binned and plotted against the number of crime incidents.
- It is found that the age group of 30-40 are responsible for the most number of crime incidents occured.
- Gives us the age group that should be taken note of for future crime predictions.

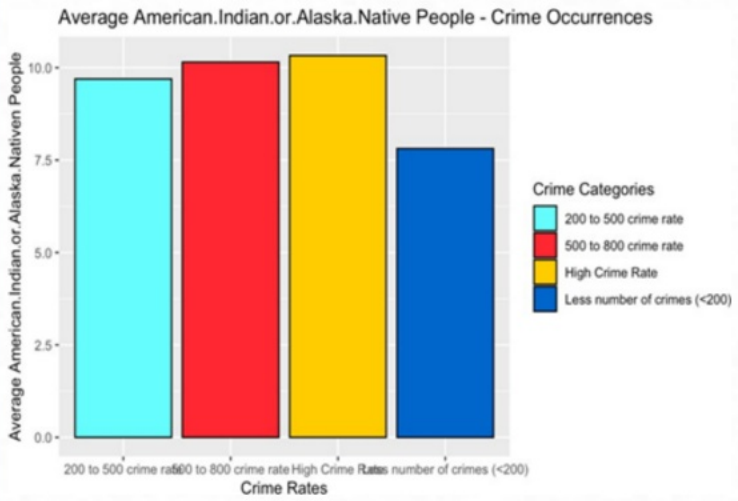
EFFECT OF EDUCATION ON CRIME RATE

All the education related columns were plotted against the number of crime incidents.
On an overall less percent of education level were more prone to commit crime.

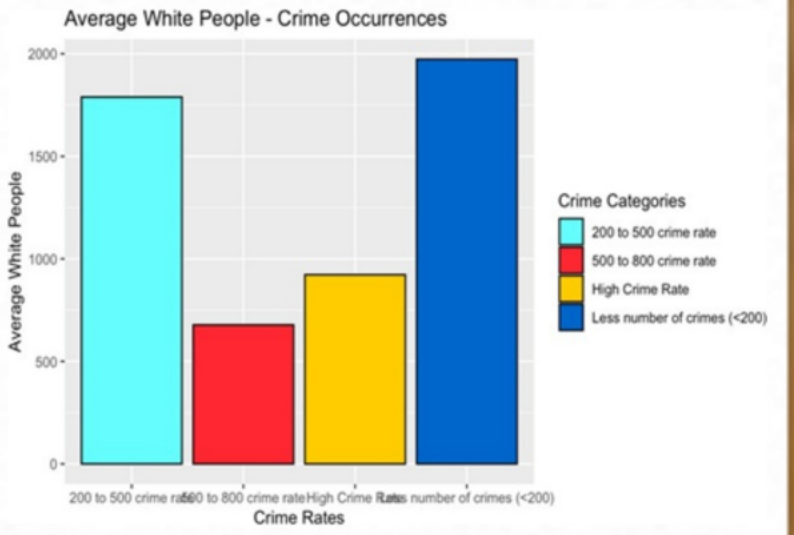
- Especially we can observe that the graduate_professional_degree's crime incidents points are observed to go lower as their percent of education level increases.
- If we focus on that particular education level we can observe the density of the points also lowers as the education level increases.
- Indicates that more percent of education level achieved can have positive impacts in reducing crime

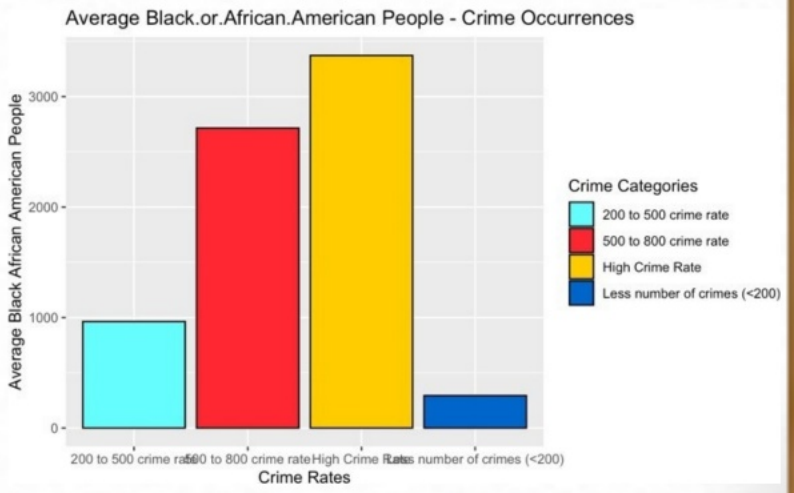
RACE AGAINST CRIME OCCURENCES:



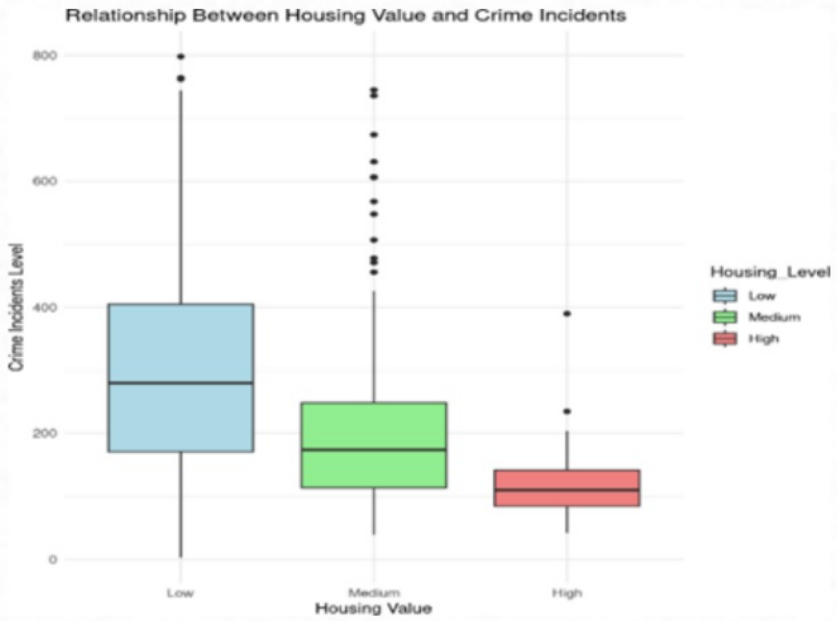


- An EDA was performed to check the behaviour of each ethnicity with respect to the crime rate.
- The crimes where categorized into 4 categories and were visualized.
- We have 4 graphs denoting each Race. For example, We can find out that Asian people have an equal distribution of crimes. On a comparative basis, We have equal distribution observed.



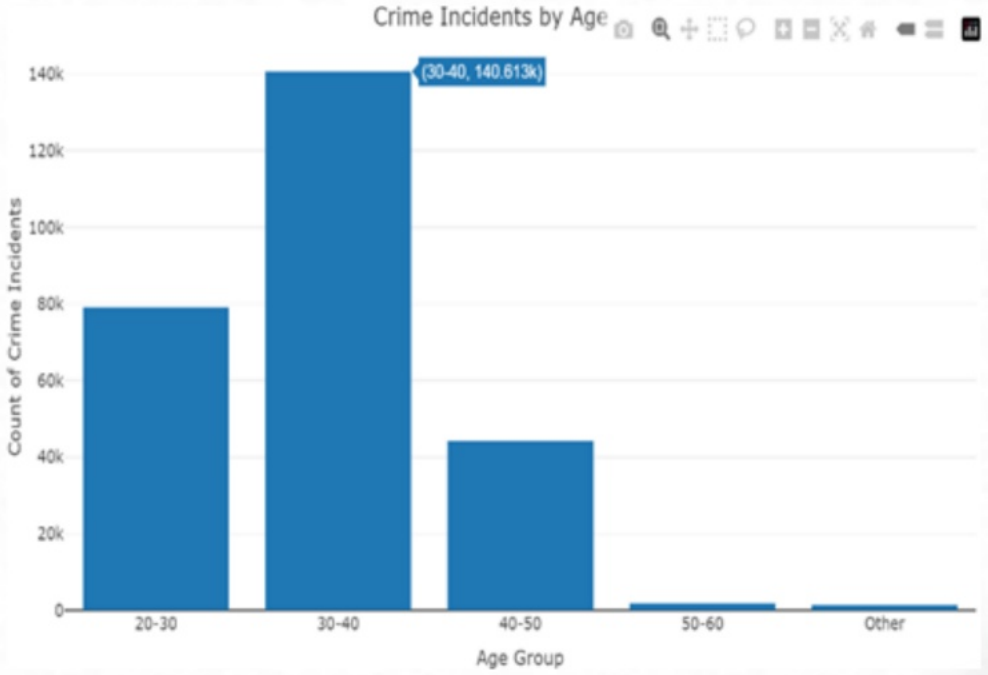


HOUSING VALUE AND AGE GROUP EFFECT ON CRIME INCIDENTS

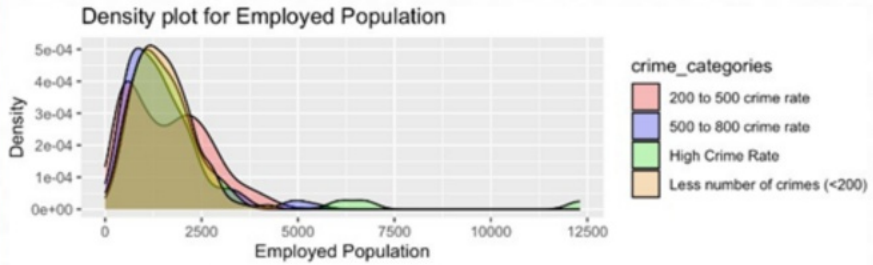


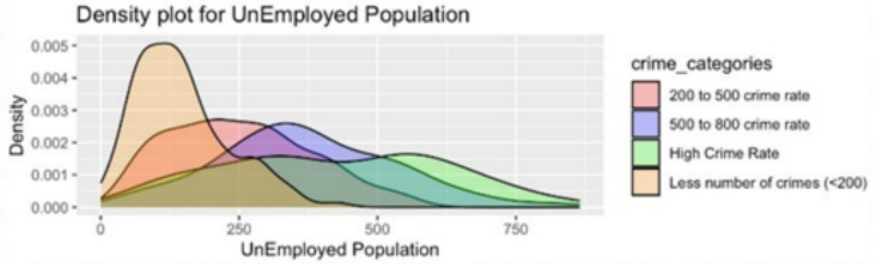
- The median housing value has been broken down to 3 levels and plotted against the number of crime incidents
- Median housing value correlates with crime incidents, indicating higher crime rates in low income housing.

- Median age records were binned and plotted against the number of crime incidents.
- It is found that the age group of 30-40 are responsible for the most number of crime incidents occured.
- Gives us the age group that should be taken note of for future crime predictions.



EMPLOYMENT AND GENDER EFFECT ON CRIME RATE



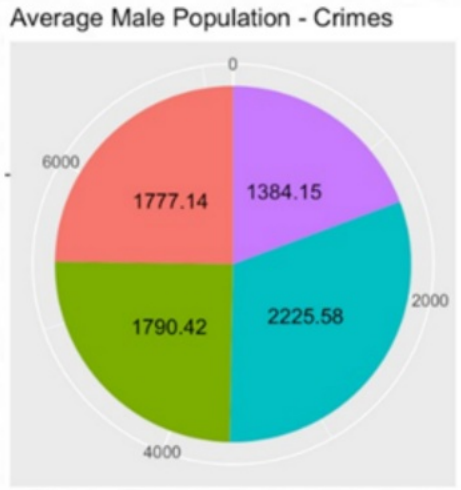


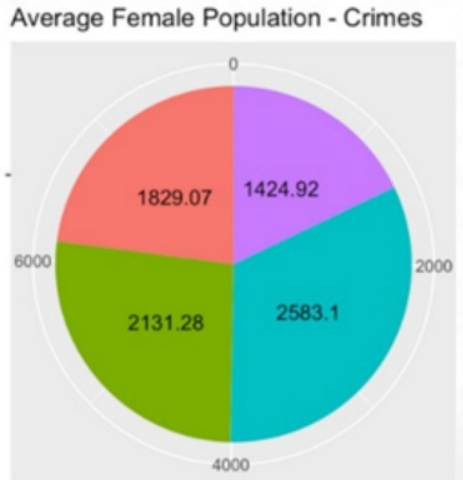
EMPLOYMENT EFFECT

- The employment effect is one of the major effects which influenced a lot of crimes. As the number of employed population increases there is a decrease in the crime occurences. For every 2000 employed population the crime rate is more while it gets decreases as the population increases.
- For un-employed population, we have the crime rate increase gradually as the population increases.

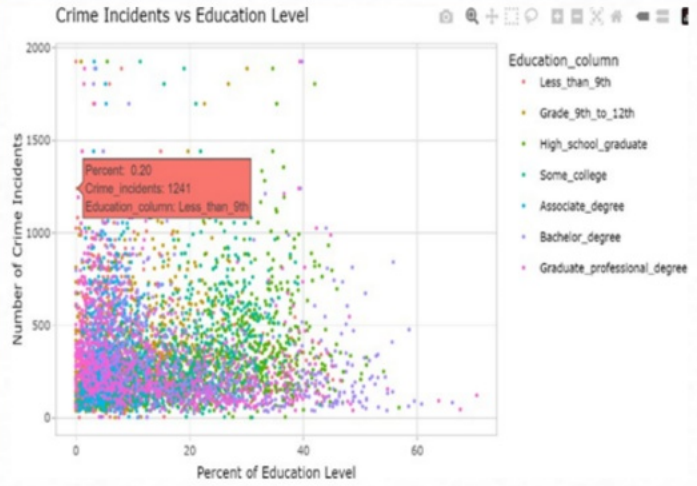
GENDER EFFECT

- We have the male population and the female population being segregated into 2 pie charts where with the male population, We have most of the men commiting high crimes. An average of 2225 men commit the most crimes out of the rest.
- For Female, We have most records occuring in the 500 and above crimes.



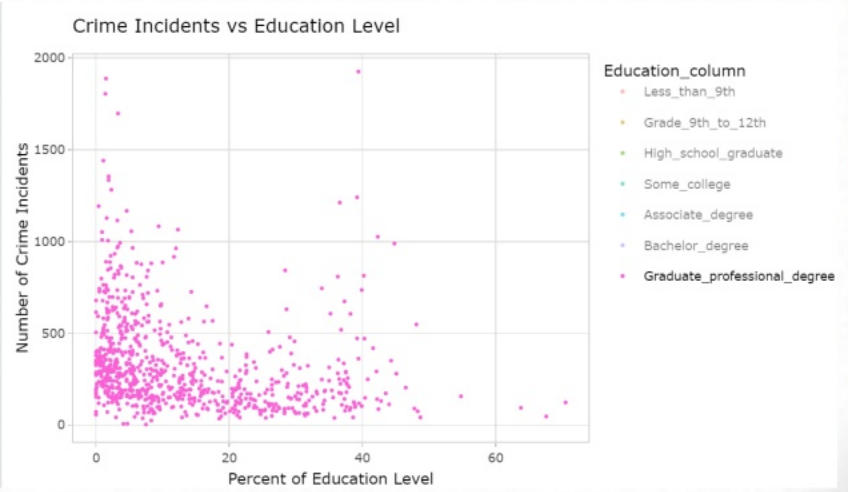


EFFECT OF EDUCATION ON CRIME RATE



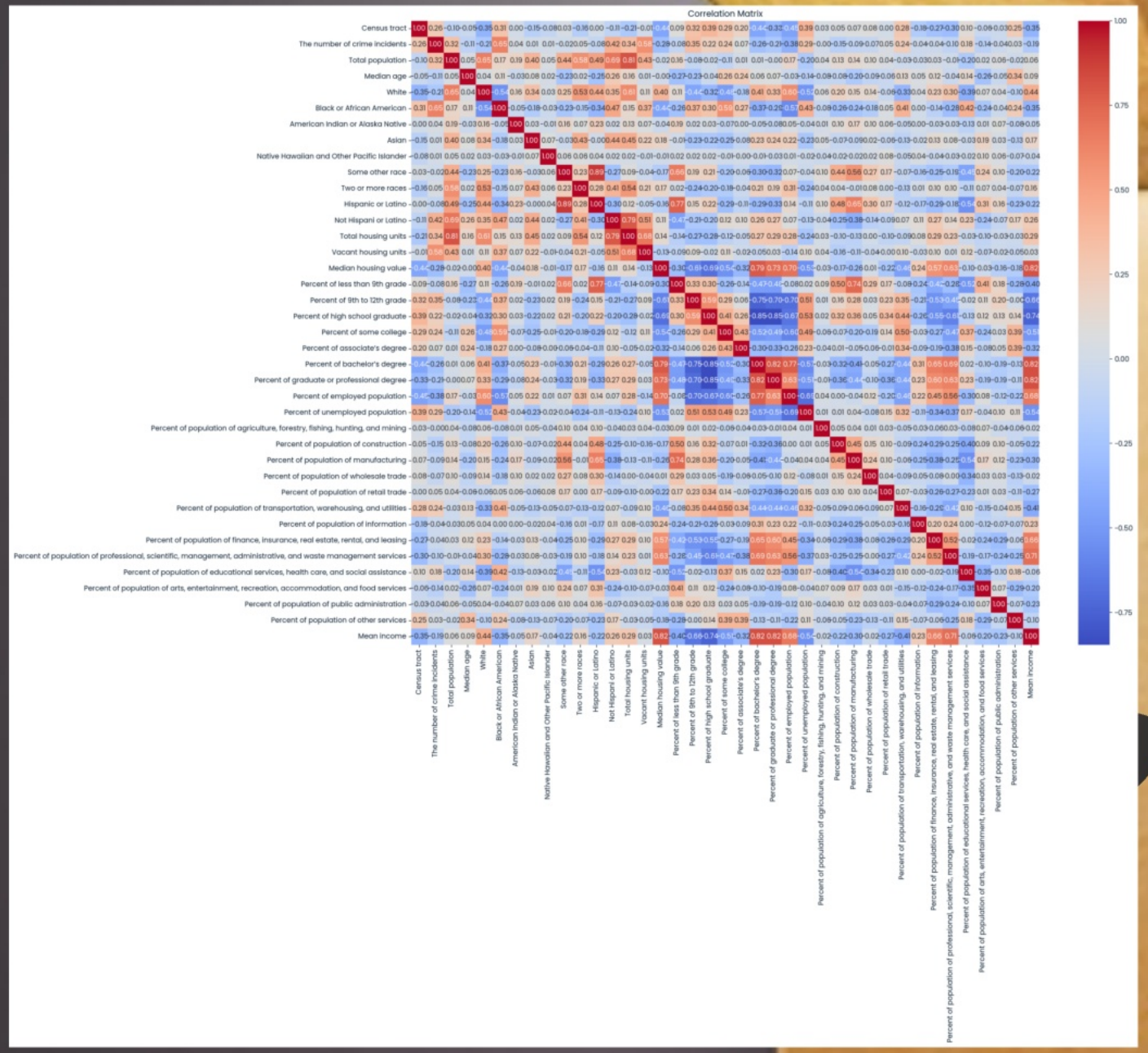
All the education related columns were plotted against the number of crime incidents.
On an overall less percent of education level were more prone to commit crime.

- Especially we can observe that the graduate_professional_degree's crime incidents points are observed to go lower as their percent of education level increases.
- If we focus on that particular education level we can observe the density of the points also lowers as the education level increases.
- Indicates that more percent of education level achieved can have positive impacts in reducing crime



CORRELATION MATRIX

The correlation of all variables where done and the highly correlated variables where removed based on a selected threshold. The matrix is shown on the right side.



SKEWNESS CHECK

- The skewness threshold was kept as <-2 or >3 and we had only some variables which were heavily skewed.
- Inorder to check if the skewness has an effect or not we scalled the variables and had a check by fitting the model and we did not find any changes. This is because of less variables being skewed.
- There were no outliers in the dataset as well

Variable Name	Skewness
American Indian or Alaska Native	4.174821
Asian	6.932746
Native Hawaiian and Other Pacific Islander	7.280884
Total housing units	3.429634
Vacant housing units	3.632187
Percent of population of agriculture, forestry, fishing, hunting, and mining	4.433094

SKEWNESS CHECK

- The skewness threshold was kept as <-2 or >3 and we had only some variables which were heavily skewed.
-
- Inorder to check if the skewness has an effect or not we scalled the variables and had a check by fitting the model and we did not find any changes. This is because of less variables being skewed.
-
- There were no outliers in the dataset as well.

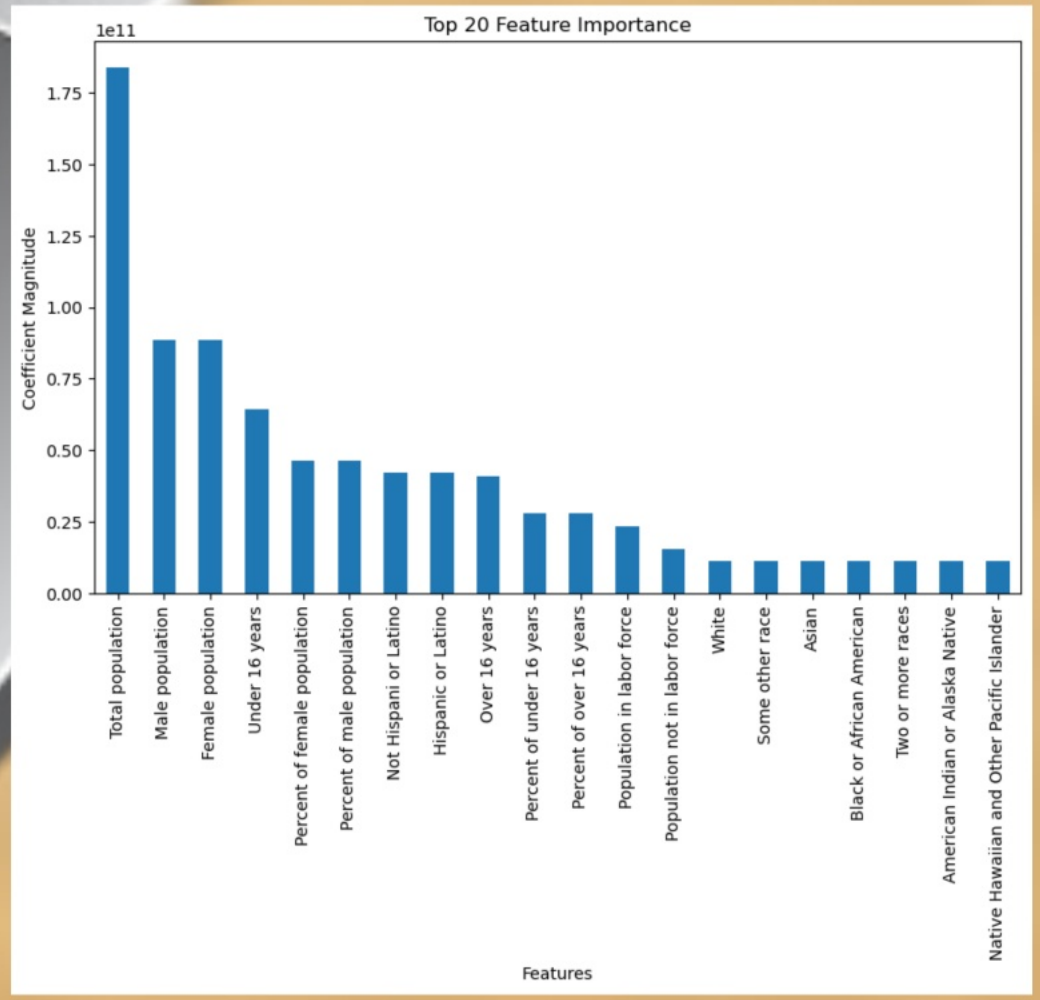
Variable Name	Skewness
American Indian or Alaska Native	4.174821
Asian	6.932746
Native Hawaiian and Other Pacific Islander	7.280884
Total housing units	3.429634
Vacant housing units	3.632187
Percent of population of agriculture, forestry, fishing, hunting, and mining	4.433094

RING CRIME DATA FOR

PREDICTIVE MODELING

MODEL 1 - MULTIPLE LINEAR REGRESSION

```
Mean Absolute Error: 139.49385630933546
Root Mean Squared Error: 323.6028419432187
Relative Absolute Error: 0.40528915056499515
Root Relative Squared Error: 0.9402042814039565
```



REGRESSION MODELING RESULTS

Models	RMSE	R-squared
Multiple Linear Regression	323.6	-0.58
MLR - Important Variables	220.5	0.28
Support Vector Machine	254.4	0.22
Support Vector Machine - Hypertuned	136.8	0.72
XGBoost Model	149.4	0.66
LASSO Model	144.3	0.71
XGBoost - Hypertuned	137.1	0.72
Random Forest Algorithm - Hypertuned	127.9	0.75

REGRESSION MODELING RESULTS

Models	RMSE	R-squared
Multiple Linear Regression	323.6	-0.58
MLR - Important Variables	220.5	0.28
Support Vector Machine	254.4	0.22
Support Vector Machine - Hypertuned	136.8	0.72
XGBoost Model	149.4	0.66
LASSO Model	144.3	0.71
XGBoost - Hypertuned	137.1	0.72
Random Forest Algorithm - Hypertuned	127.9	0.75

BAGGING & BOOSTING

Both are used to improve the accuracy of the prediction.

BAGGING

Each model is trained independently, and the final prediction is typically the average (for regression) or majority vote (for classification) of the predictions of the individual models. Bagging helps reduce variance and can improve the stability and accuracy of the model.

BOOSTING

Unlike bagging, boosting trains the models sequentially, where each subsequent model tries to correct the errors of the previous ones. Boosting algorithms assign weights to the observations, with misclassified observations receiving higher weights, so that subsequent models focus more on those observations.

BAGGING & BOOSTING

Both are used to improve the accuracy of the prediction.

BAGGING

Each model is trained independently, and the final prediction is typically the average (for regression) or majority vote (for classification) of the predictions of the individual models. Bagging helps reduce variance and can improve the stability and accuracy of the model.

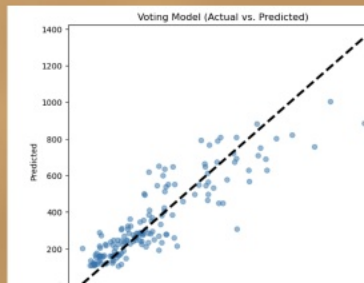
BOOSTING

Unlike bagging, boosting trains the models sequentially, where each subsequent model tries to correct the errors of the previous ones. Boosting algorithms assign weights to the observations, with misclassified observations receiving higher weights, so that subsequent models focus more on those observations.

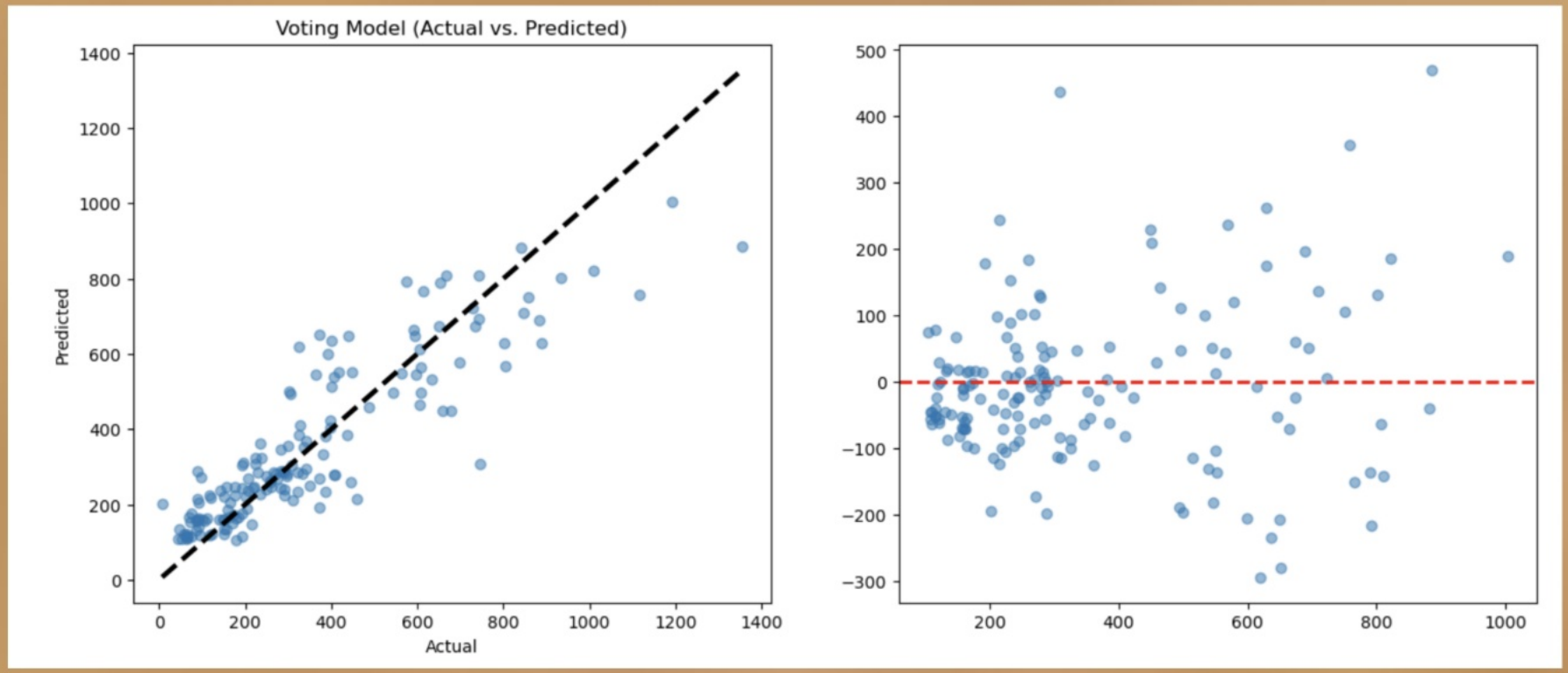
BEST REGRESSION MODELS

Models	RMSE	R-squared
Boosting Model - XGBoost & SVR	*132.6*	*0.73*
Bagging Model - SVR & RF	*119.5*	*0.78*
Boosting Model - SVR & RF	*131.5*	*0.76*

Bagging Model - Fit & Residuals plot



Bagging Model - Fit & Residuals plot



We have a lot of points cluttered arround the line and some predictions being away from the line thats why we have a RMSE value of arround 119 for the model. It might be pretty large but thats the best prediction we have got.

The residual plot shows the model is not overfitted or underfitted as we don't have all the dots on the same line but we have everything either near to the line where it is either overpredicted or underpredicted.

CLASSIFICATION MODELS

The main question here is Why Classification for a Regression problem ?

- Regression results seems not satisfactory thinking maybe the RMSE value is too high or we could have got a better r-squared value. The classification models where done to prove it right with a higher accuracy prediction.
- The major aim of the project is to suggest people whether the place is expected to be a higher crime occurence place or a low crime occurence place and this classification model with a very good accuracy proves us that as well. We get either 3 predictions which are High Crime Occurence, Low Crime Occurence or Averaging arround 400 crime occurences.
- The best 3 algorithm results have been posted.

	ACCCURACY
Random Forest Algorithm - Hypertunned	0.95
SVR - Classifier	0.60
K - Nearest Neighbors - Hypertunned	0.67

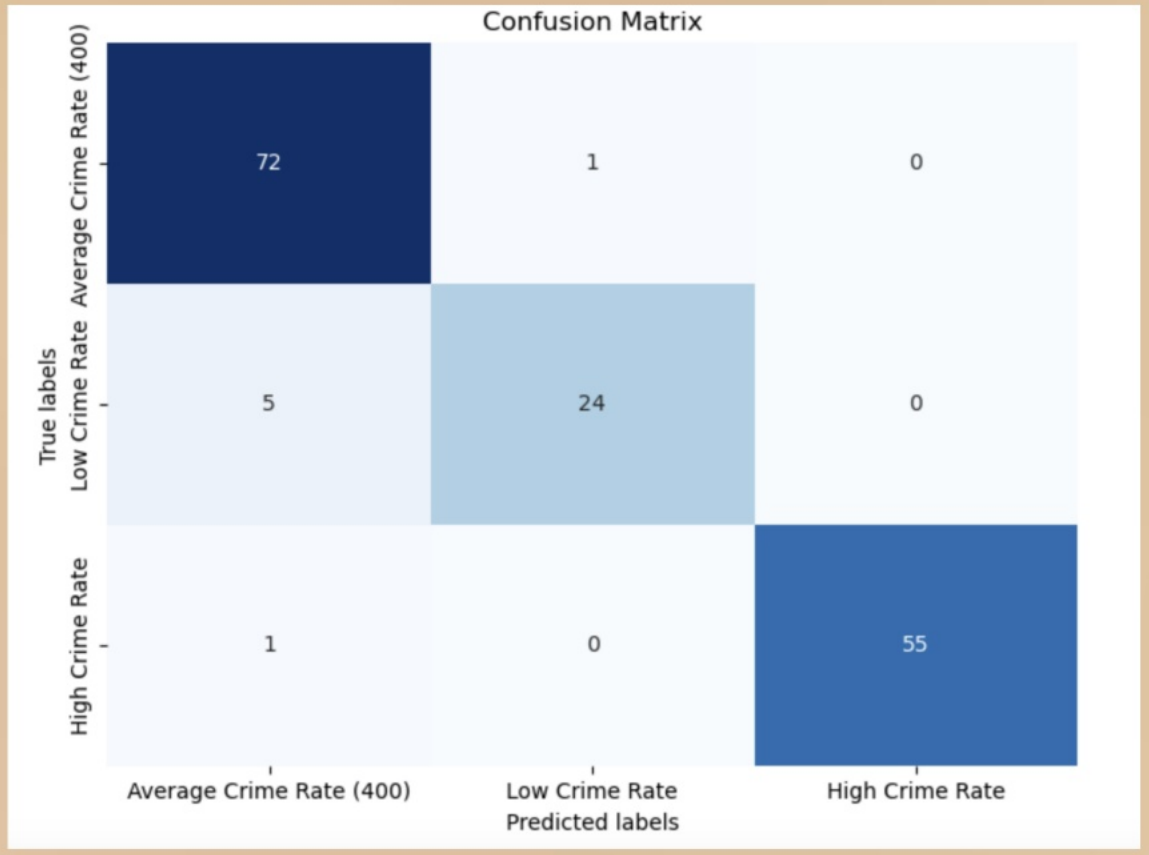
RF Confusion Matrix

- The RF Confusion Matrix was constructed where we had 72 correct predictions of Average Crime occurences, 24 correct predictions of Low Crime occurences and 55 correct predictions of High Crime

RF Confusion Matrix

- The RF Confusion Matrix was constructed where we had 72 correct predictions of Average Crime occurences, 24 correct predictions of Low Crime occurences and 55 correct predictions of High Crime occurences. Accuracy of prediction is 96%.

- Some incorrect predictions included, 5 predictions falling in the bracket of average crime rate where the true labels are low crime rate. This was the major minus which can be fixed in the future work.



MODEL DEPLOYMENT

- In order to deploy the hyperparameter tunned best model for regression and classification. We need to download the models so that it can be used by anyone at anytime in the future without rerunning the whole code.

- Hyperparameter models are computationally powerful and involves a lot of time to have it executed and completed.

REGRESSION MODEL DEPLOYMENT PKL: ***voting_model.pkl***
CLASSIFICATION MODEL DEPLOYMENT PKL: ***best_rf.pkl***

SAMPLE PREDICTIONS

REGRESSION PREDICTIONS

Actual Value	Predicted Value
207	237.314918
615	715.604217
110	164.536219
744	693.871582
150	133.352737
400	514.630534
659	450.535622
119	218.761097
407	538.854005
458	244.707115

CLASSIFICATION PREDICTIONS

Actual Value	Predicted Value
Average Crime Rate (400)	Average Crime Rate (400)
High Crime Rate	High Crime Rate
Low Crime Rate	Low Crime Rate
High Crime Rate	High Crime Rate
Low Crime Rate	Low Crime Rate
Average Crime Rate (400)	Average Crime Rate (400)
High Crime Rate	Average Crime Rate (400)
Low Crime Rate	Low Crime Rate

PROGRAMMING TECH USED

APPLICATIONS

- Jupyter Notebook
- R-Studio
- Tableau
- Prezi - Presentation

LIBRARIES

- Matplotlip
- Numpy
- Preprocessing Libraries
- Regression Algorithms
- Classification Algorithms
- Pipeline from sklearn
- Usage of Grid Search, Cross Validation
- Pickle for Model Deployment

- tidyverse
- ggplot2

FUTURE WORK

- Some future works can be done in many aspects. At first the number of records analysed were less which can be improved by foccusing more on data collection. We had just 800 records which played a vital role in the rmse. If more records where analysed we might have got a better accuracy.
- PCA can be done to limit the amount of variables to be analysed. This can be a better future work so that we can have less variables and also provide visualizations if less variables have been used in the model.
- Visual reports can be provided to improve the analytical process of the research. Visual reports on each census might help the people to have a look at the distribution of variables on each census. Tableau/Power BI or any of the visualization tools can be used on it.
- One important future work for classification models is we saw some wrong predictions in the braket of Low Crime/Average Crime which can be dealt with by trying different crime combinations to improve the accuracy.
- Include temploral analysis to understand how crime patterns evolve over time.
- Incoporate geospatial data and GIS techniques to analyze spatial patterns of crime and identify high-risk areas

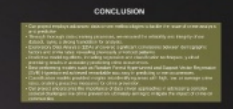
CONCLUSION

- Our project employs advanced data-driven methodologies to tackle the issue of crime analysis and prediction.
- Through thorough data cleaning processes, we ensured the reliability and integrity of our dataset, laying a strong foundation for analysis.
- Exploratory Data Analysis (EDA) uncovered significant correlations between demographic factors and crime rates, revealing previously unnoticed patterns.
- Predictive modeling efforts, including regression and classification techniques, yielded promising results in accurately predicting crime occurrences.
- Best-performing models such as Random Forest Hypertunned and Support Vector Regression (SVR) Hypertunned achieved remarkable accuracy in predicting crime occurrences.
- Classification models provided insights into identifying areas with high, low, or average crime rates, enabling proactive measures for crime prevention.
- Our project underscores the importance of data-driven approaches in addressing complex societal challenges like crime prevention, ultimately aiming to mitigate the impact of crime on communities.

Thank you!

EXPLORING CRIME DATA FOR ANALYSIS & PREDICTION

PREDICTIVE MODELING



What can we do?