# DS3021 Final Paper

Contributors: Vinith, Elaine, Ethan, Jolie, Chelsea

## Abstract:

Github link: here

In recent years, it has become increasingly important to study violent crimes and understand their patterns across different regions of the United States for various reasons, such as public safety and criminal justice reform. This study applies a multi-model machine learning and statistical approach to **investigate violent crime patterns and case solvability across the U.S grouped by region.** We used four models to answer our question: linear regression, random forest model predicting case solvability, clustering, and random forest models using data from solved cases in Virginia and the broader Southeastern. Using linear regression, we found that the Northeast region of the United States experiences a downward trend in homicide count while the other regions have no noticeable trend. Using the random forest models, we achieved a high accuracy in predicting case solvability and identifying suspect characteristics, particularly in sex and race. Using clustering, we found that rural states have much higher solve rates compared to urban states. These findings suggest that homicide trends and case solvability are highly distinct across the United States and heavily depend on region, demographics, and other factors. It is also important to note the biases in our data, which may have skewed our models. Overall, our findings highlight the distinct patterns in violent crimes across the United States.

## Introduction:

Using **linear regression**, we observed that while time and region together explained **80% of homicide count variation**, meaningful temporal trends were largely confined to the Northeast. This region had a moderate $R^2$ of 0.58, with a negative slope of approximately −35 homicides per year, reflecting a steady and **gradual decline** in cases over time. Conversely, the **southern states region showed weak and inconsistent trends** (non linear).

Our team also deployed a **Random Forest model predicting case solvability**; it achieved **73% accuracy**, with agency type and year emerging as key predictors. Additionally, for that model, hyperparameters were optimized using grid search with 5-fold cross-validation. Although, our model did have a tendency to incorrectly classify unsolved cases and classified them as solved ( **False Positives number was ~40,000**), which inherently limits the model's use in detecting cold cases.

Given this information, we understood that **region and external context**( weapon, relationship, and circumstance) may be helpful features to detect underlying patterns in state regions. After we performed **Clustering and PCA** , we found that **rural states** had the highest solve rates(**~90%**), often involving domestic homicides, and single-offender homicides, while **urban clusters** were striked by **complex**, unsolved cases involving strangers or multiple offenders ( anonymity was prevalent)- having a lowest solve rate of (**~71%**). **Arizona** stood out as an **outlier** with unique homicide characteristics- consistently.

Finally, we wanted to attempt to identify key suspect characteristics ( demographics) behind unsolved violent crime cases. We trained a series of **Random Forest models using data from solved cases in Virginia and the broader Southeastern.** The Predictive outcomes were best for offender sex and race, with accuracy scores above 90% for sex and approximately 87–89% for race. However, biases were evident — **Virginia** models skewed toward **White male offenders**, while the **Southeastern** model skewed toward **Black male offenders**, reflecting underlying data imbalances/ bias.

Our team also decided to perform a regional **subset** (Virginia vs. Southeastern States) which demonstrates stratified analysis- revealing both consistencies (high sex prediction accuracy) and disparities (varying racial model skew), exemplifying the importance of **localized model training** when investigating unsolved cases across multiple states/regions.

## Data Collection:

Our team started off the project by exploring several homicide-related datasets, each offering different strengths in scope, granularity, and context. However, as expected, we found the process of choosing a single dataset challenging, we wanted our analysis to be comprehensive while also having meaningful data that tells a story. We agreed that it was also important to consider multiple datasets—not only to compare and cross-validate findings but to provide context. For this reason, we initially considered both the UNODC's international homicide statistics and the U.S. crime report data.

The UN dataset included global homicide figures disaggregated by gender, region, and type of homicide (intimate partner/family-related), though many values were either missing or ambiguous (a lot of entries had marks with "-"). Despite this, we thought it would be insightful in understanding global trends across countries. In parallel, we also explored U.S. agency-level crime reporting data (including monthly reported homicides and clearance rates) which gave us the option to analyze regional differences and crime-solving capacity.

After considering the above options for characteristics like: the data structure, completeness, and narrative potential, we ultimately decided to use the dataset from **Murder Accountability Project**.This decision was

driven by the availability of variables —including victim and offender demographics, homicide context, case status(solved/unsolved) of each case—making it especially well-suited possibilities of predictive analyses.

As we anticipated, one major challenge when working with this dataset is the completeness and ambiguity of column headers. Additionally, there were a large number of null values: which we would have to address by discussing appropriate imputation strategies and removing features with little relevance.

Next Steps:

1. Data Cleaning: Impute missing values where necessary, convert data types according to the models used
2. Exploratory Data Analysis (EDA): Make initial visualization and use groupby function to relate and inform which features are most promising for modeling
3. Modeling Strategy: Consider which supervised and unsupervised models to pursue: maybe clustering agencies/ cases types, random forest to classify case solvability, regression to project homicide trends overtime.
4. Define research questions

## Data Cleaning:

The dataset we decided to work with to address our areas of interest/ uncover patterns was from The Murder Accountability Project. Below, we performed general data cleansing and feature selecting that we used for each model. Of course we did additional data cleaning for each model, but you could check that out on our GitHub page.

**Some general information about the dataset: There are 198,000 observations and 30 variables in our dataset.**

```
df= pd.read_csv('SHR65_23.csv')
df.head()
```

| | ID | CNTYFIPS | Ori | State | Agency | Agentype | Source | Solved | Year | Month | ... | OffRace | OffEthnic | Wea |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 197603001AK00101 | Anchorage, AK | AK00101 | Alaska | Anchorage | Municipal police | FBI | Yes | 1976 | March | ... | Black | Unknown or not reported | Handg p revc |
| 1 | 197604001AK00101 | Anchorage, AK | AK00101 | Alaska | Anchorage | Municipal police | FBI | Yes | 1976 | April | ... | White | Unknown or not reported | Handg p revc |
| 2 | 197606001AK00101 | Anchorage, AK | AK00101 | Alaska | Anchorage | Municipal police | FBI | Yes | 1976 | June | ... | Black | Unknown or not reported | Handg p revc |
| 3 | 197606002AK00101 | Anchorage, AK | AK00101 | Alaska | Anchorage | Municipal police | FBI | Yes | 1976 | June | ... | White | Unknown or not reported | Handg p revc |
| 4 | 197607001AK00101 | Anchorage, AK | AK00101 | Alaska | Anchorage | Municipal police | FBI | Yes | 1976 | July | ... | American Indian or Alaskan Native | Unknown or not reported | Kni cu instrur |

5 rows × 30 columns

*~This is the first 10 rows of our dataset, a glimpse of the features we'd be working with*

Portion of the State value counts (first few)

```
df['State'].value_counts()
```

```
State
California              129741
Texas                    86288
New York                 62695
Florida                  50854
Illinois                 39227
Michigan                 38641
Pennsylvania             34307
Georgia                  29698
North Carolina           29460
Ohio                     28230
Louisiana                27411
Maryland                 23742
Missouri                 22456
Virginia                 22454
Tennessee                22310
New Jersey               18750
Alabama                  18169
Arizona                  18068
South Carolina           17677
```

```
print(df.info())
print(df.describe())
print(df.shape)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 894636 entries, 0 to 894635
Data columns (total 30 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   ID            894636 non-null  object
 1   CNTYFIPS      894636 non-null  object
 2   Ori           894636 non-null  object
 3   State         894636 non-null  object
 4   Agency        894636 non-null  object
 5   Agentype      894636 non-null  object
 6   Source        894636 non-null  object
 7   Solved        894636 non-null  object
 8   Year          894636 non-null  int64
 9   Month         894636 non-null  object
 10  Incident      894636 non-null  int64
 11  ActionType    894636 non-null  object
 12  Homicide      894636 non-null  object
 13  Situation     894636 non-null  object
```

Image on the right shows the dimensions and the features' data type.

We then glanced at the statistical description of our data for the numerical variables.

```
None
                Year       Incident         VicAge         OffAge  \
count  894636.000000  894636.000000  894636.000000  894636.000000
mean     1998.868617      53.117632      47.313424     351.740167
std        14.134767     790.036562     117.434844     455.602403
min      1976.000000       0.000000       0.000000       0.000000
25%      1987.000000       1.000000      22.000000      24.000000
50%      1997.000000       2.000000      30.000000      38.000000
75%      2011.000000      11.000000      42.000000     999.000000
max      2023.000000   27113.000000     999.000000     999.000000

            VicCount       OffCount       FileDate
```

```
df.isnull().sum()
```

```
ID              0
CNTYFIPS        0
Ori             0
State           0
Agency          0
Agentype        0
Source          0
Solved          0
Year            0
Month           0
Incident        0
ActionType      0
Homicide        0
Situation       0
VicAge          0
VicSex          0
VicRace         0
```

By looking at the null values sum ( how much is missing from our data set)- separated by feature. We noticed that the features: Subcircum and FileDate have many missing data- so we can drop them.

This are the columns (features), we will be working with for our project:

```
df_new.columns
```

```
Index(['ID', 'CNTYFIPS', 'Ori', 'State', 'Agency', 'Agentype', 'Source',
       'Solved', 'Year', 'Month', 'Incident', 'ActionType', 'Homicide',
       'Situation', 'VicAge', 'VicSex', 'VicRace', 'VicEthnic', 'OffAge',
       'OffSex', 'OffRace', 'OffEthnic', 'Weapon', 'Relationship',
       'Circumstance', 'VicCount', 'OffCount', 'MSA'],
      dtype='object')
```

We split our dataset into 4 regions: Northeast, Midwest, South, and West. We mapped each state to one of the 4 regions. Then we grouped our dataset by the Region and Year and found the homicide count.

```
state_to_region = {
    'Maine': 'Northeast', 'New Hampshire': 'Northeast', 'Vermont': 'Northeast',
    'Massachusetts': 'Northeast', 'Rhode Island': 'Northeast', 'Connecticut': 'Northeast',
    'New York': 'Northeast', 'New Jersey': 'Northeast', 'Pennsylvania': 'Northeast',

    'Ohio': 'Midwest', 'Michigan': 'Midwest', 'Indiana': 'Midwest', 'Illinois': 'Midwest',
    'Wisconsin': 'Midwest', 'Minnesota': 'Midwest', 'Iowa': 'Midwest', 'Missouri': 'Midwest',
    'North Dakota': 'Midwest', 'South Dakota': 'Midwest', 'Nebraska': 'Midwest', 'Kansas': 'Midwest',

    'Delaware': 'South', 'Maryland': 'South', 'District of Columbia': 'South', 'Virginia': 'South',
    'West Virginia': 'South', 'North Carolina': 'South', 'South Carolina': 'South', 'Georgia': 'South',
    'Florida': 'South', 'Kentucky': 'South', 'Tennessee': 'South', 'Mississippi': 'South',
    'Alabama': 'South', 'Oklahoma': 'South', 'Texas': 'South', 'Arkansas': 'South', 'Louisiana': 'South',

    'Montana': 'West', 'Idaho': 'West', 'Wyoming': 'West', 'Colorado': 'West', 'New Mexico': 'West',
    'Arizona': 'West', 'Utah': 'West', 'Nevada': 'West', 'Washington': 'West', 'Oregon': 'West',
    'California': 'West', 'Alaska': 'West', 'Hawaii': 'West'
}
df['Region'] = df['State'].map(state_to_region)
```

## Methods:

With this data, we had a couple of questions that first came to mind.

1. What is the likelihood of a crime being solved?
2. How do crimes change over time for different parts of the US? (predicting crime counts)
3. Can we predict the demographic of the potential murderer in a cluster of unsolved cases?
4. **Are there trends between regions and urban/rural locations and homicide rates, i.e. suspicious clusters of murders?**
5. Are there differences in victim-offender relationships across different demographics?

In the end, we decided to focus on question 4, as we thought it would have the most interesting results. Since there are many approaches to determining trends between urban/rural locations and homicide rates, we chose to implement 4 different models.

- Multiple linear regression (supervised learning) to predict homicide rates using region and year as predictor variables.
- Random forest regression (supervised learning) to predict whether a case was solved or not using location, time and agency.
- Clustering (unsupervised learning) to determine clusters of solved/unsolved cases in different regions.
- Random forest regression (supervised learning) to predict unsolved cases using offender demographics.

To test whether these results would be "successful" or not, we decided we would use a testing dataset and apply our model to compare the accuracy to our training model. We would also use R^2, MSE, and RMSE as metrics for comparison across the regression models.

Some limitations we wanted to address include: working with a large dataset, which would lead to more computation and slower runtimes. Having 30 features could also lead to multicollinearity, where one or more predictor variables are strongly correlated with one another - leading to misleading coefficients. Since this dataset came from the Murder Accountability Project,, which is a nonprofit organization, its more likely that this data potentially has information that is biased/modified, and this could skew the results.

To potentially mitigate these limitations, we considered: isolating certain features and grouping similar states into regions when using them to train the models. To address multicollinearity, we could implement recursive feature elimination (RFE) to keep only the most significant features. Another way to address multicollinearity would be to perform PCA. For the multiple linear regression, instead of combining the cases from rural v urban areas, we could isolate the two and compare results between them. Lastly, we can't

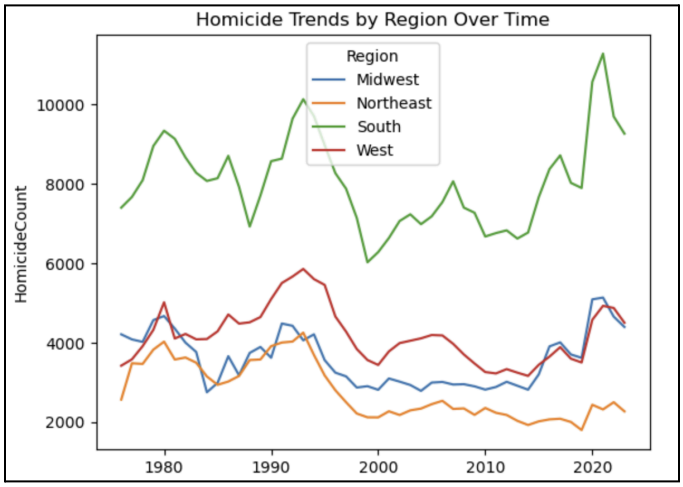
remove the inherent bias from the data, but we can acknowledge them in our conclusions and findings.

Lastly, we acknowledged that even if we couldn't answer our original question, we could definitely use our analysis/insight to leverage ideas for new, alternative methods.

## Results:

After considering the above processes and discussion, we decided to deploy 3 different machine learning models.

1. Linear Regression




| | Slope | Intercept | R_squared |
|---|---|---|---|
| Midwest | -3.104798 | 9830.298788 | 0.008350 |
| Northeast | -35.466910 | 73734.341937 | 0.579412 |
| South | 3.627895 | 943.334073 | -0.018787 |
| West | -8.719894 | 21649.037465 | 0.066501 |

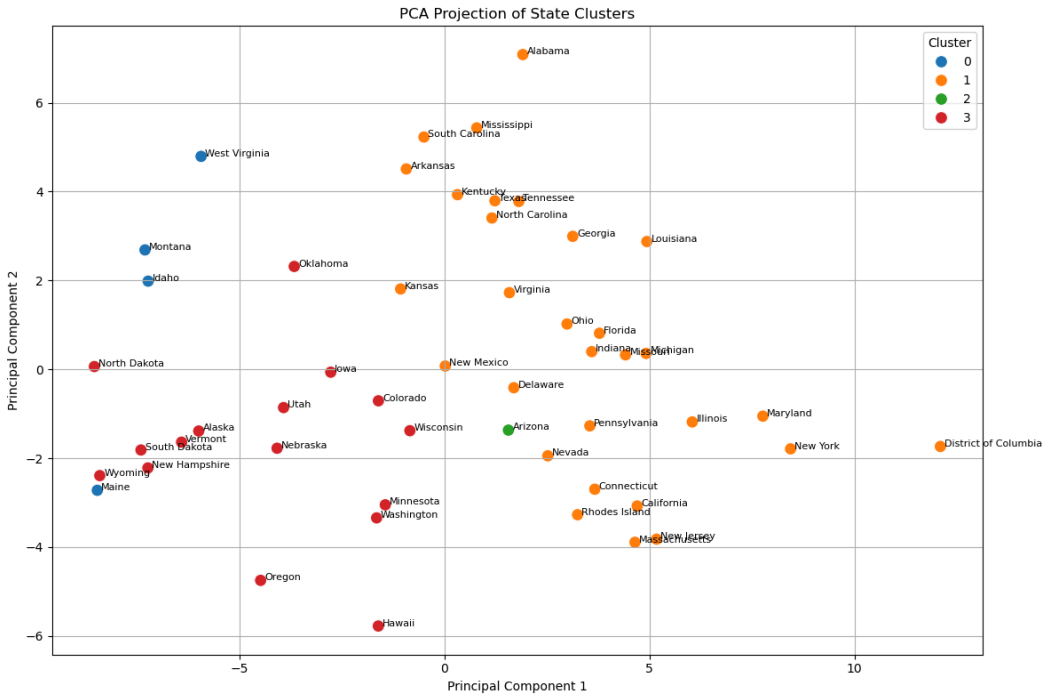
In the table above, we calculate the R-squared values of our linear regression models, which measures how much of the variation in homicide counts across time can be predicted or explained with our model. We got an output / R-squared value of 0.7998 while analyzing our entire dataset (all four regions combined), which means about 80% of the variations in homicide counts can be accounted for based on which region the data point is from (via dummy variables) and the year.

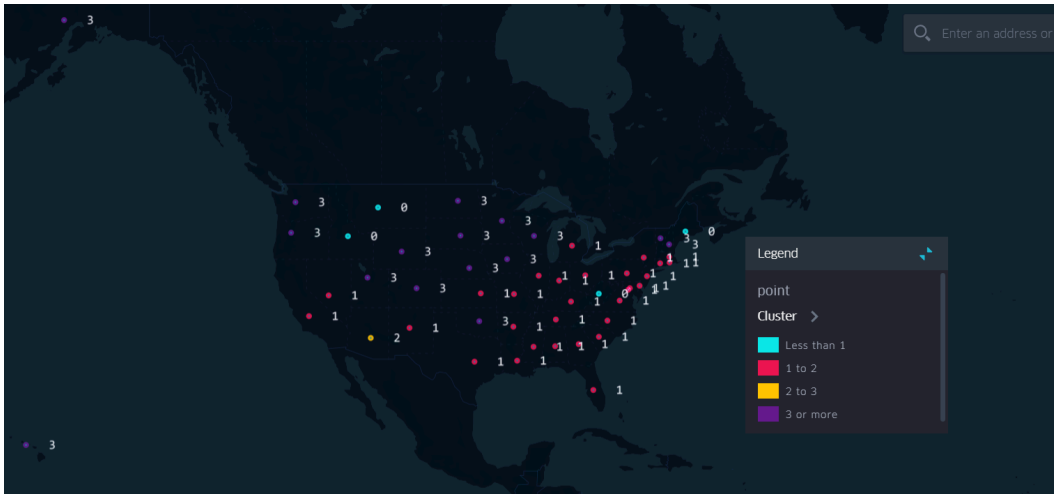
2. Random Forest model predicting case solvability

```
Test Accuracy: 0.7297795761423589
Classification report:              precision    recall  f1-score   support

           0       0.60      0.24      0.35     52565
           1       0.75      0.93      0.83    126363

    accuracy                           0.73    178928
   macro avg       0.67      0.59      0.59    178928
weighted avg       0.70      0.73      0.69    178928

Confusion matrix: [[ 12872  39693]
 [  8657 117706]]
Feature importances: [0.1552, 0.0382, 0.3447, 0.0463, 0.004, 0.2839, 0.1278]
```

A random forest regression model was trained using the State, Region, Agency, Agentype, Source, Year, and Month columns as predictors and the Solved column as the response . It had an R^2 score of 0.73.

3. Clustering



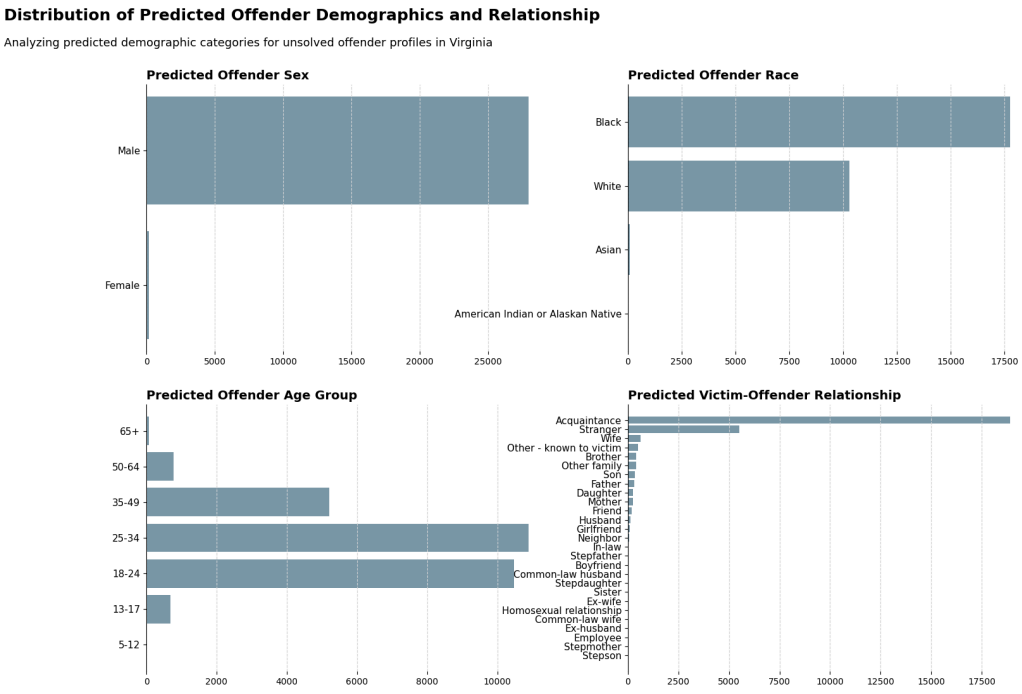



***Link to the MAP Data Visualization ( full screen + interactive tooltip):***

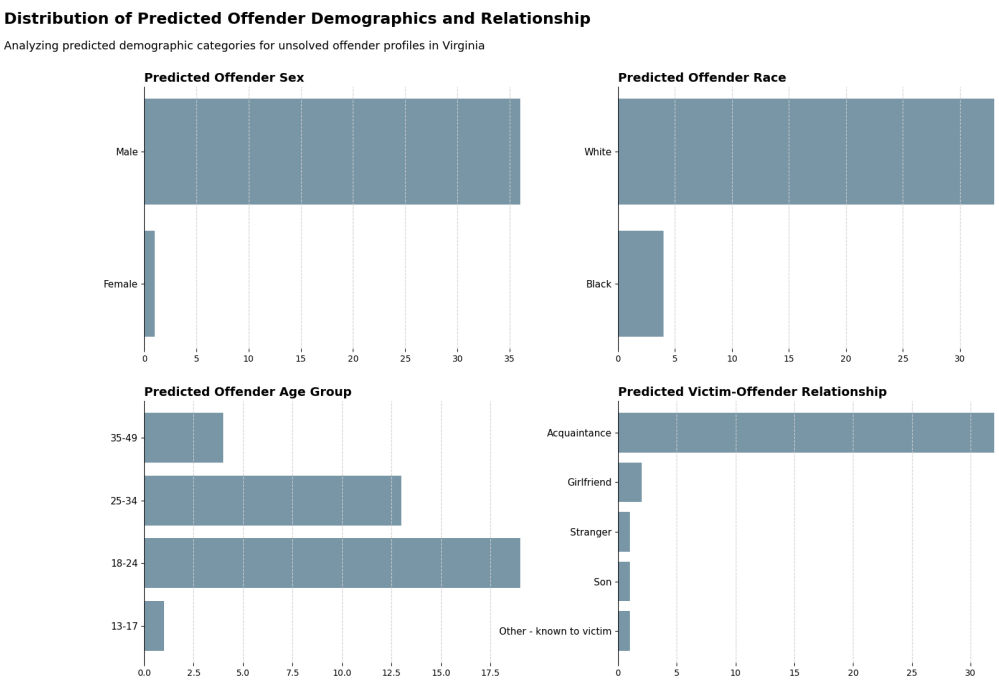
https://kepler.gl/demo/map?mapUrl=https://dl.dropboxusercontent.com/scl/fi/w4l05y4438za0zri19q6e/keplergl_d1y5z0i.json?rlkey=ouowt2qprldtrq64qvleiscog&dl=0

4. Random Forest models using data from solved cases in Virginia and the broader Southeastern.

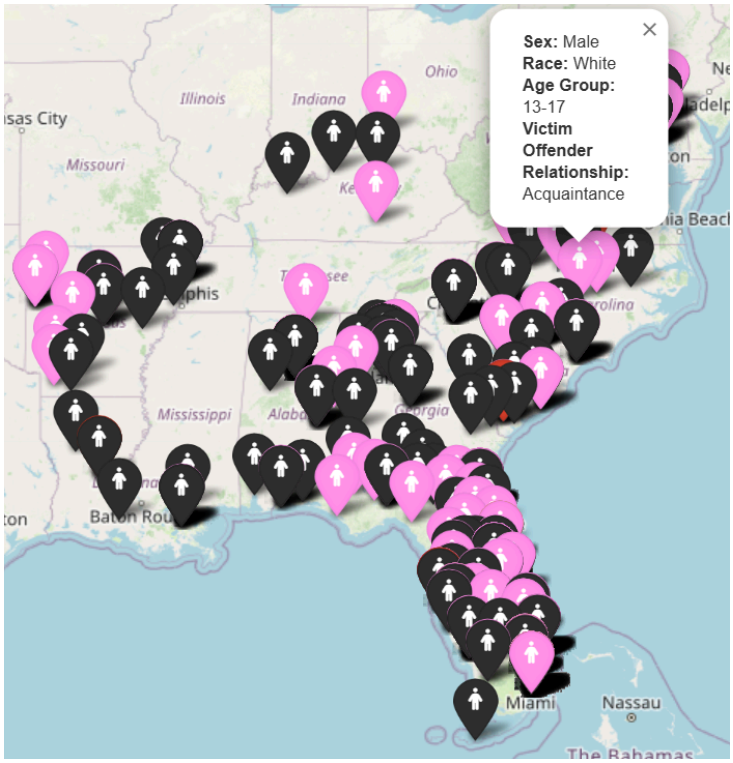
Southeastern States of the US: Georgia, Alabama, Virginia, Florida, North Carolina, etc.




State of Virginia: All counties included




Mapped out labels of the predicted results in the Southeastern region:



With limited data from the MAP dataset, as well as limited variables taken into consideration (Victim Demographics, Relationship, Crime Context, and etc), we're able to predict a hypothetical 'assumption' of the potential profiles for the offenders of the unsolved cold cases for both Virginia as well as the Southeastern region of the US.

## Conclusion:

To conclude, we wanted to summarize our results for each model.

In the linear regression method, the R-squared is pretty high, which tells us that region and time together is decently helpful in predicting homicide counts, but it doesn't tell us which part is doing the heavy lifting (region vs. year). So then we looked at each region on its own to see how well the year predicts homicide count for each region by fitting homicide count versus year. Our results showed that only the Northeast had a decent $R^2$ ( about 0.58) with a negative slope of about −35 per year, which tells us that the number of homicides in the Northeast have dropped quite steadily through time. However, in the other three regions, the R-squared values are basically zero, which tells us that for those regions, there are no clear linear trends for homicide count versus year. Lastly, we saw that most of the predictive accuracy for homicide counts (based on year and region) comes from region.

Looking at the random forest regression model, in general it correctly predicted 73% of all cases. Looking at the confusion matrix tells us that out of all the cases, a minority were incorrectly predicted, though the model is worse at predicting unsolved cases as solved cases. In the linear regression model, we weren't able to determine whether region or time was more important in predicting homicide counts. Looking at the feature importances in the random forest regression model, we can see that Agency and Year are the most important features, while Region and Source are less important in determining whether a case was solved or not. Overall, location, agency, and time are moderately useful in predicting whether a case was solved or not. To improve the model, we could add more columns such as case details like weapon used, victim information, offender information, etc, though that might also introduce multicollinearity and more noise into the predictions.

For the clustering method, our model ended up making 4 clusters. Cluster 0 (4 states) had the highest proportion of solved homicides (Solved_Yes ~ 90%), low urban crime types, and high rates of domestic relationships. These are mostly rural states that have solvable, domestic related homicides- low complexity of crimes. Cluster 1 had the lowest solved rate of 71%, it had a high rate of unknown offender cases, more complex urban homicides, and high population states. These states represent diverse, urban-driven homicide patterns where crimes are harder to solve. Cluster 2 had low solved rates, and high multiple victim cases and unknown offenders. Cluster 3 had an above average solved rate(85%), and high family or known-offender rates. These states form a moderate cluster— they may reflect balanced

investigative systems and a mix of urban/rural crime. Based on the above analysis, it looks like a low population, and higher social ties leads to more solvable domestic crimes. On the other hand, densely populated states have more unknown violence which leads to unsolved crimes, stranger homicides, and felony situations.

In our last method, the feature-specific accuracy for both experiments remained relatively similar: our RandomForest model yielded an approximately 90% accuracy for features such as race and sex. However the model yielded a relatively worse accuracy of about 43% for features like Age Group and Victim Offender Relationship. We concluded that because the data is from an objective, national criminal database, the model is likely more skewed toward statistical biases/inaccurate crime/murder profiles compared to it. This is why in our second model, by including data from all states in the Southeast region, we predicted that the majority of the potential homicide offenders to be black, male, 25-34 years old, who are acquaintances of the victim. To render this prediction effort more fair and justified, the currently available case-specific variables seem to be insufficient, and that more behavioral and socio-economic background data is needed to truly train a model that analyzes how murderers think, act, and behave towards their victims. These data are without a doubt going to be more difficult in terms of its acquisition, however, it will certainly be more fair and less prone to statistical biases.

While experimenting and implementing these Machine Learning models, we faced several complications and challenges. For example, as mentioned earlier- temporal trends were only meaningful when investigating Northeast ( with a moderate $R^2$ value and weak, non-linear trends in other regions like the South. Something we could try in future works is experimenting with non-linear models ( time-series models, spline regression) which may be helpful to capture regional temporal dynamics. Additionally, we also experienced a high false positive rate ( ~40,000 cases misclassified as solved) in our random forest model. We could possibly consider implementing class imbalance ( like weights or SMOTE). Not to mention, for the clustering model, something that we expected in our models was the possibility of Multicollinearity- which is why we deployed PCA with our clustering in hopes of the model to generalize better.

Call to Action:

1. States/ local governments should invest in localized crime analysis (our model shows regional variation in solvability)- Policymakers should work with law enforcement to develop prevention and investigative strategies by regions and surveillance in urban areas.
2. Invest in forensic tech - we already saw that higher unsolved rates mean that the evidence may be weak and forensic efforts may need more of investing
3. Deploy Specialized Task Forces in Outlier and High-Crime Clusters: for gang violence, trafficking, or serial offense.

4. Cross collaboration with CPS and Police- in state clusters that had elevated rates of arson, children homicides and stepfamily related cases
5. Additionally , for high solve rates ( domestic focused)- Leverage prior restraining orders or 911 call patterns to preempt escalation.

## References:

"Data & Docs." *Murder Accountability Project*, www.murderdata.org/p/data-docs.html?m=1. Accessed 9 May 2025.