# VIOLENT CRIMES IN TEXAS


A Presentation by: Adil Baksh, Ryan Pope, Griselda Rodriguez and Michael Sanchez

# PROJECT OVERVIEW

- For this project, we will analyze Census data alongside the National Incident-Based Reporting System to assess factors related to violent crime in Texas.
- Our goal is to predict the safety of specific cities on a per capita basis. We will employ the classification model Random Forest, and will categorize the data based on various features. Throughout the project, we will leverage Python, Pandas, Matplotlib, and Tableau for different analytical tasks.
- After analyzing the data, our goal changed to predicting the crime rate per 1000 people.

# DATA SOURCES

## US Census Data

ACS 5-year estimate data for all Texas cities reported to the US Census.

## National Incident-Based Reporting System

Number of crimes against a person for cities in Texas.

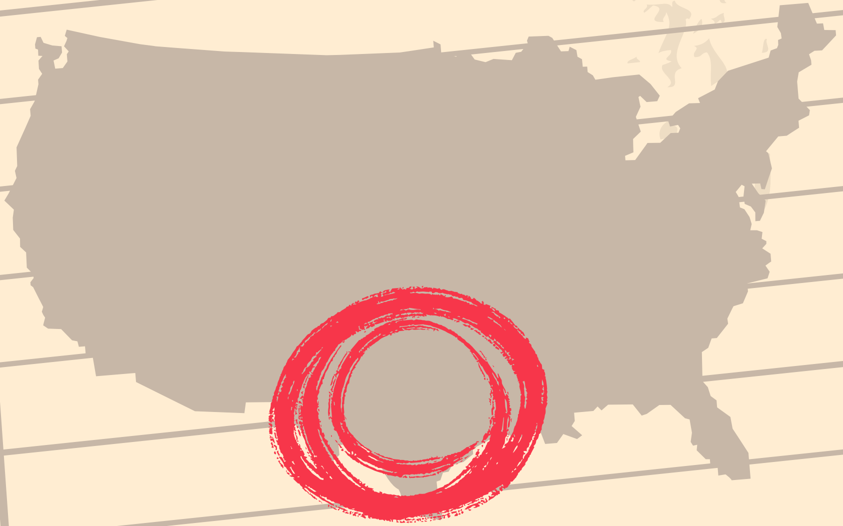
Data from 2022

# OVERVIEW OF DATA

- Census
  - Utilizes the 2022 ACS 5-year estimate data for all Texas cities reported to the US Census. Total count: 626 cities
- National Incident-Based Reporting System
  - Sources from the FBI Website
  - Reduces city count to 612 of reporting cities in Texas

# INTERACTIVE MAP

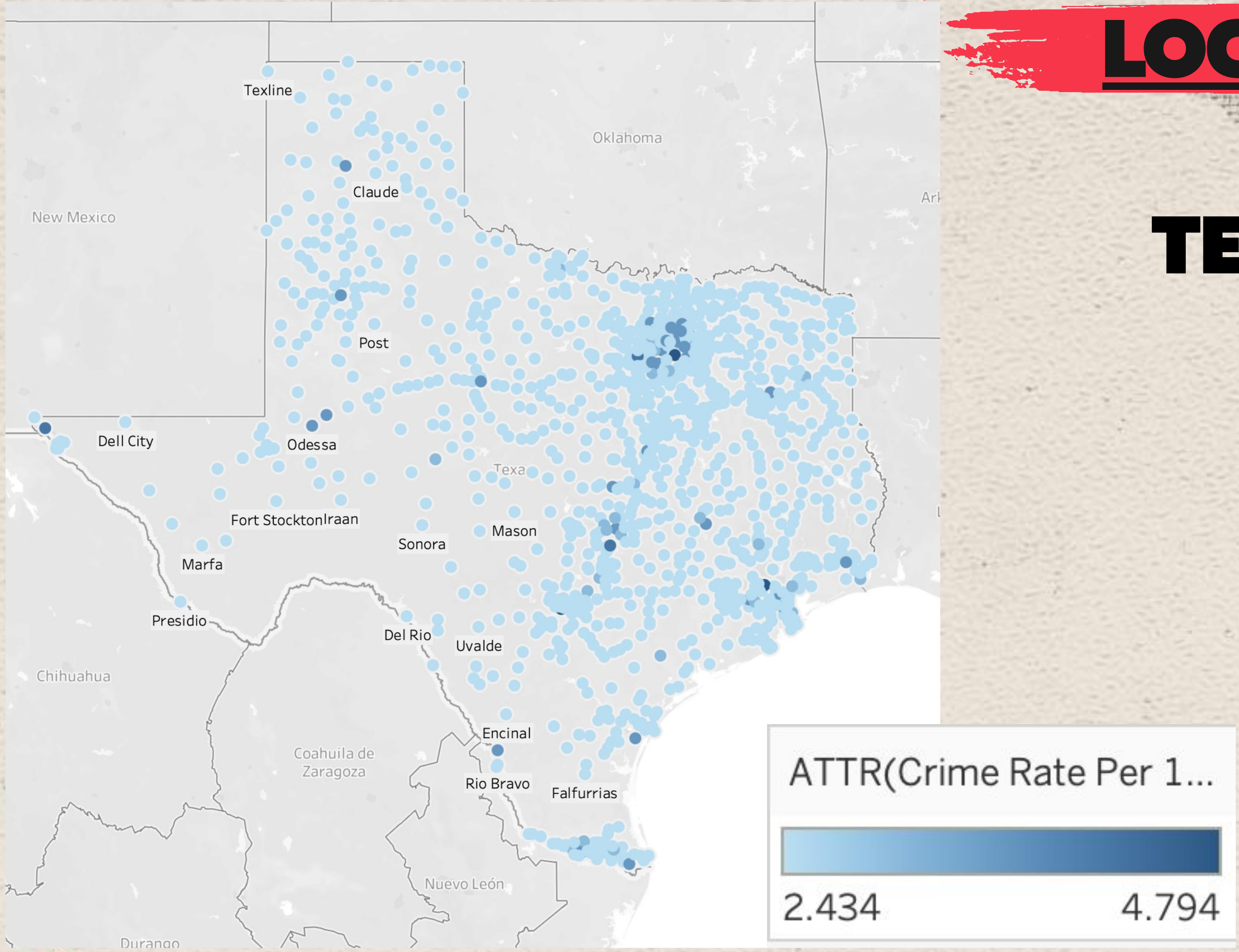
Main Focus: Texas



## TABLEAU

Used Tableau to create an interactive map.

# LOCATIONS OF REPORTED TEXAS CITIES



# MACHINE LEARNING MODEL
## OVERVIEW

# Code Overview

**Objective:** Predict Crime Rate based on a dataset of city-related features.

- **Process:**
    1. Load and preprocess data.
    2. Train and tune a RandomForestRegressor model.
    3. Evaluate the model.
    4. Analyze feature importance.
- **Tools Used:** Python, Pandas, Scikit-Learn, SciPy, and logging for tracking steps and errors.

# TYPE OF MODEL

- Model: RandomForest
- Target Variable: Crime Rate per 1000
  - Standard notation used by Researchers to determine the rate of crime.
- Changed focus from classifying “safe cities” to predicting the crime rate per 1000 people after initial analysis of data.

# TYPE OF MODEL

14 Features Used for Each City:

- Median Household Income
- Poverty Rates
  - Percentage of Males under the Poverty Line
  - Percentage of Females under the Poverty Line
  - Percentage of adults over 25 under the Poverty Line
  - Percentage of Poverty to City Population
- Education Rates
  - Percentage of adults with High School or Less Education
  - Percentage of adults with High School Degree or Equivalent
  - Percentage of adults with Some college and/or Associates
  - Percentage of adults with Bachelor's Degree or more
- Population
- Housing Data
  - Percent of Owner Occupied housing
  - Percent of Renter Occupied housing
  - Median Rent
- Total Offenses of Crime

# MACHINE MODEL LEARNING OPTIMIZATION

# What Changed?

- Removed Extraneous Column of Index (Left on the Original)
- Removed the Feature “Percentage of Owner Lived” as it was deemed irrelevant and had a stronger than expected pull to the data.
- Utilized a new model to help identify the best fit of important features using GridSearchCV

# FINAL MODEL RESULTS

# RESULTS

```
2024-10-24 20:57:06,497 - INFO - Mean Squared Error (MSE): 25.75740230317041
2024-10-24 20:57:06,497 - INFO - R-squared (R^2): 0.5670033494463131
2024-10-24 20:57:06,502 - INFO - Results saved to 'model_results.csv'
2024-10-24 20:57:06,502 - INFO - Analyzing feature importances
Mean Squared Error: 25.75740230317041
R^2 Score: 0.5670033494463131

Feature Importances:
                  Feature  Importance
9         Total\nOffenses    0.524622
1          Percent_Female    0.146334
7      Percent_Renter_Occ    0.080038
8             Population1    0.069288
2        Percent_over_25y    0.062498
3              Percent_HS    0.033471
0            Percent_Male    0.026887
6            Poverty_Rate    0.021068
5         Percent_BS_over    0.018051
4   Percent_Some_Col_Assoc    0.017744
```

| Actual | Predicted | City |
|---|---|---|
| 9.52813067150635 | 11.243545379293726 | Canton |
| 3.399192691735713 | 3.5223089171655637 | Hallsville |
| 2.930832356389215 | 2.052592749963237 | Bovina |
| 3.554502369668246 | 2.2574123536259743 | Wink |
| 1.7835909631391202 | 2.748996968551878 | Itasca |
| 12.094264117385503 | 11.76590962163798 | Elgin |
| 10.845986984815616 | 13.273375260642574 | San Augustine |
| 17.171476269973763 | 15.79376110486644 | Levelland |
| 4.800768122899664 | 5.499157117569636 | Weimar |
| 19.599075937104104 | 17.66937279662048 | Kilgore |
| 6.5048786589942456 | 4.33952254597418 | Highland Village |
| 8.637612877895563 | 4.3439122388990485 | Hudson Oaks |

# INTERPRETATION OF RESULTS

- Mean Squared Error: ~25
- $R^2$ Score: 0.58
- Top contributor to Results:
  - Total Offenses a City Holds
  - Percentage of Female Population under the Poverty Line

***Improvement: adding polynomial features to improve model R-Squared values

# FUTURE STUDIES

# RATIONALE AND POTENTIAL

- The purpose of the development of this machine learning model is to determine the crime rate of Texas Cities to potentially identify “safe” cities based on a set of features in order to:
  - Allocate resources to identify and alleviate crime in cities that may need while demographics change.
  - Analyze trends that may occur with growing populations and the changes in the economy.
- Future exploration to update and optimize the model would be to:
  - Include a larger set of information to include regions and potentially the entirety of the US cities.
  - Include more features outside of economic factors that may potentially lead to violent crime not considered in this model.

# WEBSITES USED

**US Census American Community Survey 2022 5-year Estimate**
https://data.census.gov/advanced. Accessed October, 2024

**FBI Crime Data Explorer 2022 for Texas**
https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/downloads. Accessed October, 2024

**Safewise. "100 Safest Cities in the US"**
https://www.safewise.com/safest-cities-america/. Accessed October, 2024