# Hate Crime Analysis and Forecasting

Pranjal Jain (pj2069)
Revati Trivedi (rst8739)
Yamini Narasimhan (yl9822)

Link for the most current version of the project proposal document:

Data Science project proposal

*What is the problem (including motivation and what is the specific outcome)?*

**Motivation**

Hate Crime is a crime motivated by victim's race, ethnicity, sexual orientation, color, national origin, gender identity, or disability. With the rising amount of Hate Crimes, it is crucial to analyze and understand the trends of Hate crimes to increase awareness and promote a more inclusive and healthy environment.

The goal of this project is to analyze historical data on hate crimes and forecast the number of hate crimes for coming years.

**Potential questions that we aim to answer with this project are**

- What is the trend of hate crime per population for different states?
- Are there certain events that have spiked the total hate crime or any specific hate crime?
- What are the season trends for certain types of hate crime?
- Is the crime targeted more towards individuals or groups?

*What kinds of data will you use? (describe the data fully including it's temporal and spatial dimensions, features and their types and scales (e.g. numerical or text, ordinal or nominal, etc.))*

**Source of dataset**

| Feature Name | Data Type | Description | Cleaning Procedure |
|---|---|---|---|
| incident_id: | Numerical, Nominal | Unique ID for Incident | Drop |
| data_year | DATE | Year of incident | |
| ori: | Text-Nominal | Code | Drop |
| pub_agency_name | Text-Nominal | Agency that reported | |
| pub_agency_unit | Text-Nominal | Agency Unit | Drop |
| agency_type_name | Text- Ordinal | Type of Agency reported | Drop |
| state_abbr | Text -Nominal | State Name Abbreviated | |
| state_name | Text -Nominal | State Name | Drop |
| division_name | Text -Nominal | Division area | |
| region_name | Text -Nominal | Region Name | |
| population_group_code | AlphaNum -Ordinal | Population of location-Coded | One hot Encoding and Standardize as using from multiple agencies |
| population_group_desc | Text -Ordinal | Population of location | Drop |
| incident_date | DATE | Exact Date of Incident | Standardize Date format |
| adult_victim_count | Numerical | Count of Adult Victims | |
| juvenile_victim_count | Numerical | Count of juvenile Victims | |

| | | | |
|---|---|---|---|
| total_offender_count | Numerical | Total Count of all Offenders | |
| adult_offender_count | Numerical | Count of Adult Offenders | |
| juvenile_offender_count | Numerical | Count of juvenile Offenders | |
| offender_race | Text -Nominal | Offender's Race | Drop: Irrelevant |
| offender_ethnicity | Text -Nominal | Offender's ethnicity | Drop: Irrelevant |
| victim_count | Numerical | Count of all Victims | |
| offense_name | Text -Nominal | Type of Offense | Standardize |
| total_individual_victims | Numerical | Count of all Victims | |
| location_name | Text - Ordinal | Type of Place incident Occured | |
| bias_desc | Text - Ordinal | Type of ethnicity Targeted | |
| victim_types | Text - Ordinal | Victim Type | |
| multiple_offense | Text - Ordinal | Single or Multiple offense | |
| multiple_bias | Text - Ordinal | Single or Multiple Bias | |

**Dimension**: 28 X 219,578

**Time range:** 1991- 2020

**About the data**

We are using the data provided on the FBI site - https://crime-data-explorer.app.cloud.gov/pages/explorer/crime/hate-crime.
This data is being collected per the Hate Crime Statistics Act passed on April 23, 1990. It also has extensive documentation defining each type of violation and their specific conditions. It also discusses the method of collection, data points collected and what is included.

The website https://www.fbi.gov/how-we-can-help-you/need-an-fbi-service-or-more-information/ucr/hate-crime is a great source of information for learning more about the dataset and about Hate Crime in general.

*What kind of model will you build? (What approach will you take for solving the problem and why not any other approaches, including how data will be cleaned, what specific algorithm(s) and any parameters used, and how you will evaluate your approach – describe a figure/table used to illustrate the evaluation)*

**Model and Evaluation approach**
**ARIMA ?**
**LSTM ?**

Given that we have historical data from 1991 to 2020, we plan to use the ARIMA model to forecast the number of Hate Crimes in the future. For parameters we can use the season ARIMA model for finding season trends of hate crime based on season order.

We are planning to split the dataset into training and test dataset yearwise. We will keep all the crimes from the year 2015-2020 as a test dataset. After training the model on all the crimes from the year 1991-2015, we will forecast for the next few years. We will compare this forecast with the test crime set to see if the pattern matches or not. We are planning on using the Mean squared error as an evaluation measure to see the error between actual test set and the predicted one.

Apart from the above mentioned approach, we are planning to use different approaches such as KNN where we form an unsupervised cluster for different types of crimes. Also, we are considering using deep learning models for the time series analysis to get better and more accurate forecasts.

*What assumptions are safe to make? (Explain clearly what the assumptions being made are and why that's okay, this could be in terms of features considered, potential confounding variables, variable types, etc.)*

**Assumption about the data**

All states/agencies across various states participate voluntarily in submission of the Hate Crime reports to the FBI. We assume that the states are reporting most of the hate crimes and there are no duplicate events considered more than once, this is due to unique ID tags for each crime.

We assume that most of the crimes reported have a specific motivation that leads it to be considered as a Hate Crime under how FBI crime explorer categorizes it as a hate crime.

We are dropping some features such as incident_id and population_group_desc due to irrelevance to predicting the time series of hate crime. We are also dropping some features such as offender race and ethnicity to avoid bias.