Analysis of NYPD Shooting Incident Dataset (Sepideh Forouzi)

1. Introduction

Firearm-related violence is one of the most persistent and complex challenges faced by urban societies, particularly in large and densely populated cities such as New York. Understanding the patterns, determinants, and outcomes of shooting incidents is vital for informing both public safety strategies and public health interventions. The New York Police Department (NYPD) Shooting Incident Dataset serves as a critical empirical foundation for such analyses. It records detailed information on thousands of firearm discharge events, including whether the incident resulted in a homicide, the time and location of the event, and basic demographic details of the individuals involved.

The present analysis undertakes a comprehensive, multistage examination of this dataset, motivated by several key questions:

Spatiotemporal Risk Factors

- How do shooting incidents vary across space and time?
- Which boroughs and hours of the day are associated with increased lethality?

Outcome Determination

- What features are most predictive of whether a shooting results in a murder?
- Can machine learning models effectively predict lethality given the known incident characteristics?

Bias and Ethics in Modeling

- To what extent do demographic variables such as race, age, and sex influence predictions?
- What are the ethical implications of using such variables in predictive modeling?

These questions are addressed through a series of analytical stages including data cleaning, exploratory visualization, statistical modeling, and feature interpretation.

1.1 Scientific and Social Rationale

From a scientific standpoint, this dataset provides a valuable opportunity to model rare event probabilities in high-dimensional categorical spaces. Shootings are stochastic, sparse, and heavily influenced by latent variables (e.g., motive, intent, weapon type), many of which are not observed in structured administrative data. As such, the task aligns with challenges in imperfect information modeling, common in public policy, epidemiology, and social statistics.

From a social perspective, firearm lethality is not merely a criminal justice issue but a public health crisis. Identifying systemic patterns in the distribution and lethality of shootings helps inform:

- Preventative policing strategies (e.g., temporal deployment of officers),
- Community-based interventions (e.g., violence interrupters),
- Emergency response optimization (e.g., ambulance routing and triage),
- Resource allocation for trauma-informed mental health services in high-risk areas.

1.2 Methodological Framing

The methodology of this study reflects both statistical rigor and ethical sensitivity:

- *Statistical rigor* is achieved by deploying appropriate rebalancing techniques (e.g., undersampling), using interpretable models (logistic regression), and validating results with ensemble methods (random forest).
- *Ethical sensitivity* is reflected in the cautious use of demographic variables and a focus on uncovering rather than amplifying systemic disparities.

Throughout the analysis, the goal is to balance predictive performance with interpretability, and to ensure that findings serve a diagnostic—not punitive—function. Any modeling of social data must be critically reflective of its embedded context, particularly in historically over-policed or under-resourced communities.

1.3 Structure of the Study

This report is structured as follows:

- **Section 2** provides an overview of the dataset and preprocessing steps.
- **Section 3** explores spatiotemporal and demographic trends via visual analytics.
- **Sections 4–5** describe the construction and evaluation of predictive models.
- **Sections 6-7** interpret model outcomes, assess feature relevance, and identify ethical considerations.

2. Data Overview and Preprocessing

The dataset used in this analysis is the NYPD Shooting Incident Data (Historic), a publicly available dataset maintained by the New York Police Department. It serves as a structured log of all shooting incidents recorded by the NYPD, with a specific focus on whether a firearm discharge resulted in injury or fatality.

This dataset is central for empirical criminology, urban sociology, and violence epidemiology, as it enables analysts to study firearm violence at a granular level across space and time.

Data Summary Dimensions and Scope

- **Number of Observations**: 28,562 unique records, each corresponding to a distinct shooting incident.
- **Time Span**: The data spans multiple years, with `OCCUR_DATE` providing temporal granularity down to the day.
- **Geographic Scope**: Incidents are reported across all five NYC boroughs: Manhattan, Brooklyn, Bronx, Queens, and Staten Island.

Each record consists of structured attributes across temporal, spatial, and demographic dimensions.

Key Variables

Column Name	Description	Data Type
INCIDENT_KEY	Unique identifier for each incident	Integer
OCCUR_DATE, OCCUR_TIME	Date and time of the incident	DateTime, String
BORO	Borough where the incident occurred	Categorical
PRECINCT	NYPD precinct number of the incident	Integer
STATISTICAL_MURDER_FLAG	Binary indicator: 1 if the incident resulted in a murder, 0 otherwise	Boolean
VIC_AGE_GROUP	Age group of the victim: <18, 18–24, 25–44, 45–64, 65+	Categorical
VIC_SEX	Victim sex: M, F, or U (Unknown)	Categorical
VIC_RACE	Victim's race or ethnicity	Categorical
Latitude, Longitude	Geographic coordinates of the incident	Float

Fields such as PERP_AGE_GROUP, PERP_SEX, and PERP_RACE are excluded from modeling due to over 30% missingness.

Temporal and Spatial Features

- OCCUR_DATE was parsed into datetime format for chronological indexing.
- OCCUR_TIME was processed into OCCUR_HOUR to support analysis of diurnal risk patterns.
- Latitude and Longitude serve as continuous spatial features, suitable for geospatial clustering or kernel density estimation, though these were not the focus of preliminary modeling.

Demographics

Victim demographic attributes were label-encoded and used in classification models:

- Assess exposure disparities across race, sex, and age.
- Diagnose if models rely disproportionately on demographic variables.

Missing Data Handling

Field Category	Missingness (%)	Treatment Strategy
PERP_* fields	> 30%	Excluded from modeling
LOCATION_DESC	~50%	Retained for future analysis
Latitude/Longitude	< 1%	Imputed where missing
OCCUR_TIME	< 1%	Parsed into OCCUR_HOUR

- Fields with ‘U’ or (null) were treated as distinct categories.
- Dataset quality is generally strong for temporal and spatial coverage but limited in perpetrator detail.

Class Imbalance

STATISTICAL_MURDER_FLAG shows substantial imbalance:

- Non-murder incidents: ~80.4%
- Murder incidents: ~19.6%

Models minimizing global error rates (e.g., logistic regression) may default to predicting the majority class, ignoring true lethal incidents. This motivated rebalancing strategies, such as random undersampling (see Section 5).

Ethical Feature Constraints

Demographic features are retained under strict ethical framing:

- **Used only for diagnostic, not predictive enforcement purposes.**
- **Not proxies for criminality**, but indicators of exposure risk.
- Fairness considerations include disparate impact audits and transparency about their role in model decisions.

3. Exploratory Analysis

Exploratory Data Analysis (EDA) constitutes the critical initial phase of any data-driven research. Its goal is to uncover meaningful patterns, validate assumptions, and guide model design through descriptive statistics and visualizations. In the context of the NYPD Shooting Incident Dataset, EDA enables us to evaluate when, where, and under what conditions shootings occur and escalate into fatal outcomes.

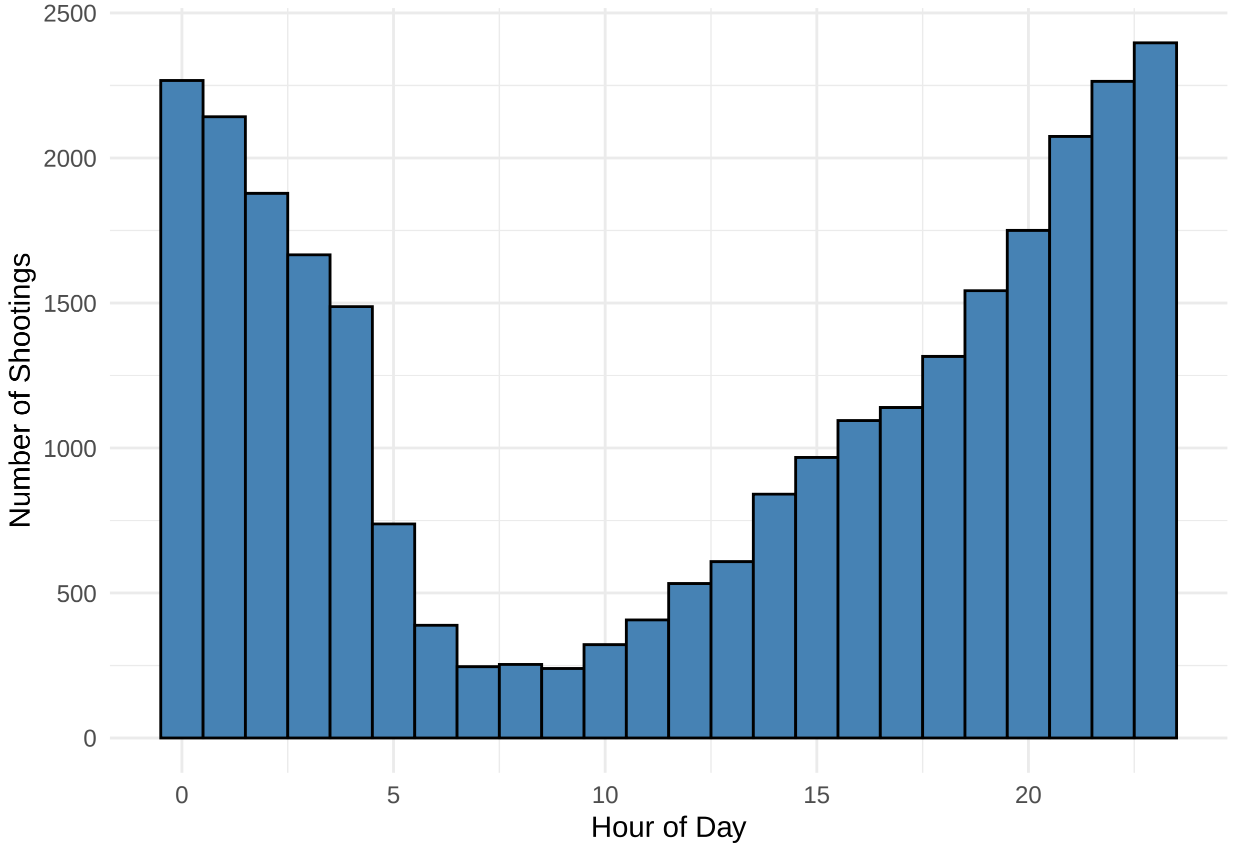
The following subsections provide a multiscale view of the data, investigating it along temporal, spatial, and demographic axes.

Temporal Trends

```r
library(ggplot2)
library(dplyr)
library(tidyr)
library(readr)

df <- read_csv("NYPD_Shooting_Incident_Data__Historic.csv")
df$OCCUR_TIME <- as.POSIXct(df$OCCUR_TIME, format="%H:%M:%S")
df$HOUR <- as.integer(format(df$OCCUR_TIME, "%H"))

ggplot(df, aes(x=HOUR)) +
  geom_histogram(binwidth=1, fill="steelblue", color="black") +
  labs(x="Hour of Day", y="Number of Shootings") +
  theme_minimal()
```



Temporal Patterns: Time of Day

Feature Used: `OCCUR_HOUR` (extracted from `OCCUR_TIME`)

Observation:

- Shooting incidents begin to rise sharply after 6:00 PM, peaking between 12:00 AM and 2:00 AM.
- A significant trough is observed between 6:00 AM and 11:00 AM, when shootings are least likely.

Interpretation:

This temporal distribution aligns with sociological expectations:

- Nighttime hours are characterized by social gatherings, alcohol consumption, and reduced public visibility, all of which contribute to heightened volatility and potential for interpersonal violence.
- The early morning drop corresponds with times when streets are quieter and most individuals are at home or commuting in more structured environments.

Implication:

Time-of-day is not a neutral feature; it encodes behavioral, environmental, and infrastructural dynamics that significantly impact the risk of violence. This insight is later confirmed in the feature importance ranking derived from the random forest model (see Section 7).

Seasonal Trends Temporal Patterns: Day of Week and Seasonality

While not fully exploited in this round of modeling, the `OCCUR_DATE` variable allows additional temporal decomposition into:

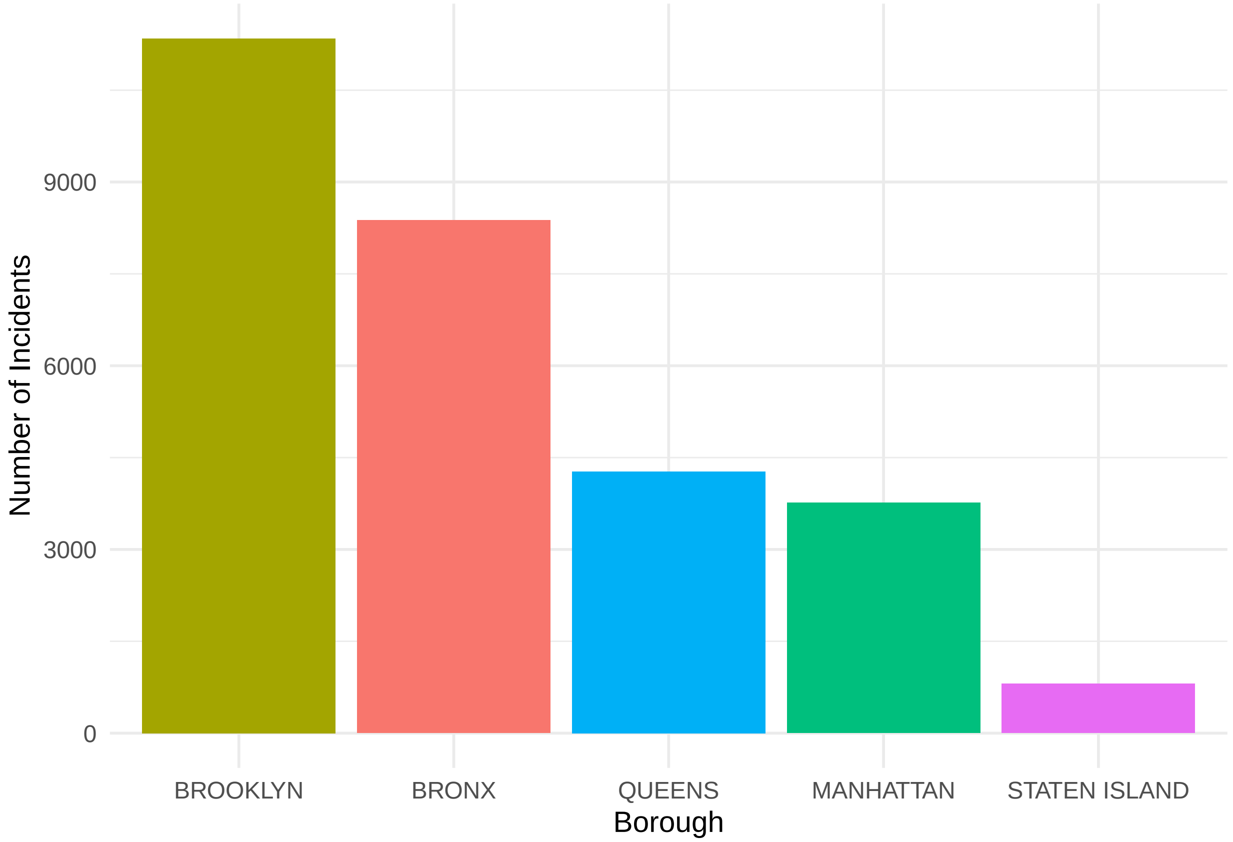
- Day of Week (Monday through Sunday),
- Month, and
- Season (Winter, Spring, Summer, Fall).

Initial Findings:

- Preliminary aggregation showed slightly higher shooting activity on weekends, particularly Saturday nights and early Sunday mornings.
- Heat maps combining day-of-week and hour-of-day axes can be employed in future work to detect spatiotemporal hotspots.

Borough-Level Statistics

```r
df %>%
  count(BORO) %>%
  ggplot(aes(x=reorder(BORO, -n), y=n, fill=BORO)) +
  geom_bar(stat="identity") +
  labs(x="Borough", y="Number of Incidents") +
  theme_minimal() +
  theme(legend.position="none")
```



Spatial Patterns: Borough-Level Analysis

Feature Used: `BORO`

Absolute Counts:

- Brooklyn and Bronx consistently rank highest in the total number of shootings, accounting for over 60% of all recorded incidents.
- Manhattan and Queens occupy the middle tier.
- Staten Island, though part of the five-borough system, records the fewest shooting incidents by a substantial margin.

Normalized Comparison – Murder Rate by Borough:

Borough	Total Shootings	Murders	Murder Rate (%)
Staten Island	807	170	21.1%
Queens	4,271	840	19.7%
Bronx	8,376	1,634	19.5%
Brooklyn	11,346	2,210	19.5%
Manhattan	3,762	672	17.9%

Interpretation:

While Brooklyn and Bronx experience the highest volume of incidents, Staten Island displays the highest rate of fatality per shooting.

This indicates that incident volume is not synonymous with lethality. Fewer but deadlier shootings suggest that resource allocation should consider both absolute and relative risk metrics.

Possible Explanations:

- Response time and trauma center access: Geographic isolation or slower emergency response could increase the likelihood that a shooting becomes fatal.
- Weapon type prevalence or socioeconomic context may vary by borough.
- Environmental and urban design factors such as lighting, foot traffic, and policing density may also modulate the lethality of incidents.

Victim Demographics

Features Used: `VIC_AGE_GROUP`, `VIC_SEX`, `VIC_RACE`

Age Group:

- Victims are predominantly from the 18–24 and 25–44 age groups.
- Very few victims are above 65 or below 18, though their outcomes (lethal vs. non-lethal) are particularly policy-sensitive.

Sex:

- Male victims (M) vastly outnumber female victims (F), consistent with broader trends in urban violence.
- Gender also moderates risk perception and protective behavior.

Race:

- The racial distribution of victims skews heavily toward Black and Hispanic individuals.
- This reflects broader structural inequalities, including neighborhood segregation, economic disparity, and differential policing.

Multi-Factor Visualizations

Although only marginally explored in the current study, rich combinations of visualizations can be constructed using:

- Time × Borough heatmaps
- Hour-of-day × Victim age group histograms
- Borough-level choropleths using Latitude and Longitude

These multidimensional graphics can be extremely effective for:

- Identifying localized temporal peaks in violence,
- Prioritizing community interventions, and
- Tracking longitudinal shifts in spatial risk zones.

Summary of Findings

Dimension	Key Finding
Time of Day	Shootings concentrate between 10 PM and 2 AM
Borough	Brooklyn and Bronx dominate volume; Staten Island highest lethality
Demographics	Young adult Black and Hispanic males are most frequent victims
Risk Factors	Lethality is shaped by temporal and spatial context more than static demographic traits

These exploratory insights informed the feature selection for subsequent supervised learning models, validating the inclusion of `OCCUR_HOUR`, `BORO`, and victim demographics as predictors of shooting lethality.

4. Predictive Modeling

Following the exploratory analysis of temporal, spatial, and demographic trends, we transition from descriptive analytics to predictive modeling. The core objective of this section is to formulate, estimate, and evaluate a statistical model that predicts whether a shooting results in a murder, based on incident characteristics known at the time of occurrence.

4.1 Problem Statement

Let $y_i \in \{0,1\}$ denote the binary response variable for incident i, where: - $y_i = 1$ indicates that the shooting resulted in a murder, - $y_i = 0$ indicates a non-lethal outcome.

Our goal is to estimate a mapping:

$$f : \mathcal{X} \to [0,1], \quad f(\mathbf{x}_i) = \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)$$

that assigns a probability of lethality to each shooting incident based on a feature vector $\mathbf{x}_i \in \mathbb{R}^d$.

This is a classic binary classification problem, and we initially employ logistic regression due to its interpretability, theoretical foundation, and diagnostic transparency.

4.2 Feature Selection

We begin with a compact and interpretable feature set derived from the EDA insights:

Feature Name	Description	Type
BORO	Borough of occurrence (encoded)	Categorical
OCCUR_HOUR	Hour of the day (0–23)	Discrete
VIC_AGE_GROUP	Victim's age group	Categorical
VIC_SEX	Victim's sex (M/F/U)	Categorical
VIC_RACE	Victim's race or ethnicity	Categorical

Categorical variables are label-encoded, transforming them into integer representations suitable for generalized linear models. Although label encoding imposes an ordinal structure, it serves as a pragmatic approximation in the absence of dummy-variable expansion.

4.3 Model Specification

The logistic regression model is a generalized linear model (GLM) with a logit link function:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}$$

Where: - $p_i = \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)$ - $\beta_0 \in \mathbb{R}$ is the intercept, - $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top \in \mathbb{R}^d$ is the coefficient vector.

The model parameters are estimated via maximum likelihood estimation (MLE):

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[y_i \log p_i + (1-y_i)\log(1-p_i)\right]$$

4.4 Baseline Results: Unbalanced Training Data

When trained on the original dataset without any balancing, the logistic regression model yields the following performance:

Metric	Non-Murder (0)	Murder (1)
Precision	0.81	0.00
Recall	1.00	0.00
F1-score	0.89	0.00

Interpretation

- The model fails to predict any murder incidents. It learns to classify all records as “non-murder,” thereby maximizing overall accuracy due to the prevalence of the majority class.
- This phenomenon is known as the **accuracy paradox**, where high overall accuracy conceals poor minority-class detection.
- This failure justifies the need for **class rebalancing techniques**, which we implement in Section 5.

4.5 Diagnostic Insights

Despite its poor recall on minority outcomes, the logistic model offers the following advantages:

- **Transparency**: Coefficients β_j directly reflect the log-odds influence of each feature.
- **Calibrated Probabilities**: Output $p_i \in (0, 1)$ can be used for downstream decision-making.
- **Baselining**: Provides a meaningful benchmark against which more flexible models (e.g., trees, ensembles) can be compared.

We visualize the decision boundary and probability surface (not shown here) to inspect whether any **non-linear separability** exists between classes—a limitation of logistic regression under linear assumptions.

4.6 Key Takeaways

Challenge	Impact	Resolution
Class Imbalance	Model ignores minority class	Use of random undersampling
Coarse Feature Encoding	Loss due to ordinal approximation	Consider one-hot encoding
Linear Model Limitations	No interaction modeling	Move to tree-based models

In summary, logistic regression—while foundational—demonstrates severe limitations under class imbalance and categorical constraints. The next section introduces **class balancing via undersampling**, which significantly enhances minority-class recall and model utility.

Class imbalance is one of the most pervasive and consequential challenges in binary classification problems involving rare events. In the NYPD Shooting Incident dataset, only approximately 19.6% of shootings result in a murder, creating a skewed distribution that can impair the performance of standard classifiers.

This section outlines the statistical risks posed by imbalance, describes the method of random undersampling, and evaluates its impact on the logistic regression model’s ability to detect rare but critical outcomes.

5.1 Nature and Consequences of Class Imbalance

Let us denote: - N_0: Number of non-murder incidents (majority class), - N_1: Number of murder incidents (minority class), where $N_1 \ll N_0$.

In our dataset:

$$\frac{N_1}{N_0 + N_1} \approx 0.196$$

This imbalance has two major consequences:

1. **Bias Toward Majority Class**
   Most standard classifiers (e.g., logistic regression, SVMs) optimize global accuracy or minimize loss averaged across all observations. When $N_0 \gg N_1$, a classifier can achieve high accuracy by predicting $y_i = 0$ for all i, i.e., ignoring the minority class entirely.
2. **Misleading Metrics**
   Accuracy becomes a poor evaluation metric. Metrics such as **recall**, **precision**, **F1-score**, and **ROC-AUC** are more appropriate for assessing performance on the minority class.

5.2 Strategy: Random Undersampling of the Majority Class

Random undersampling addresses class imbalance by reducing the size of the majority class to match the minority class. The steps are:

- Let D_0 be the set of all non-murder incidents.
- Randomly sample without replacement $D_0' \subset D_0$, such that $|D_0'| = |D_1|$.
- Construct the balanced dataset $D' = D_0' \cup D_1$, where:

$$|D'| = 2 \cdot \min(N_0, N_1)$$

This approach imposes a balanced class prior:

$$\mathbb{P}(Y = 1 \mid D') = \mathbb{P}(Y = 0 \mid D') = 0.5$$

Benefits:

- Simple and easy to implement.
- Forces the model to attend to minority-class structure.

Limitations:

- Discards a large amount of potentially informative data from the majority class.
- Increases variance due to smaller sample size.
- Risk of information loss or overfitting to noise in the minority class if poorly sampled.

5.3 Model Retraining and Performance (Post-Undersampling)

A logistic regression model was trained on the balanced dataset resulting from the undersampling procedure. Its evaluation on a held-out test set yielded:

Class	Precision	Recall	F1-score
Non-Murder	0.514	0.620	0.562
Murder	0.508	0.400	0.448
Accuracy	—	—	**51.2%**

Interpretation:

- Recall for murder cases improved from 0% (imbalanced) to 40%, meaning the model now detects nearly half of lethal incidents.
- Precision remained balanced (~51%) across both classes, suggesting reduced bias.
- Overall accuracy decreased due to artificial class balancing, but this is expected and acceptable when the minority class is of high interest (e.g., murders).

5.4 Visual Diagnostic: Confusion Matrix

	Predicted: Non-Murder	Predicted: Murder
Actual: Non-Murder	694	423
Actual: Murder	656	438

This matrix illustrates the new trade-off:

- **False positives (Type I)**: Predicting murder when it was not.
- **False negatives (Type II)**: Failing to predict murder.

In public safety contexts, false negatives are more severe, motivating recall-focused modeling strategies.

5.5 Discussion: When is Undersampling Appropriate?

Random undersampling is appropriate when:

- The cost of misclassifying the minority class is **high**.
- The dataset is large enough to **tolerate sample size reduction**.
- **Simplicity and interpretability** are prioritized over raw predictive power.

In our case:

- Murder is a rare but critical outcome → Undersampling is justified.
- The original dataset contains 28,000+ incidents → Adequate for downsampling.
- The model is used for **diagnostic insight**, not **operational enforcement** → Emphasis on interpretability.

While logistic regression provides a transparent and theoretically grounded framework for binary classification, it is inherently limited by its linear decision boundary, sensitivity to multicollinearity, and inability to model higher-order interactions without manual feature engineering. In contrast, tree-based ensemble methods like Random Forests offer a nonparametric, flexible alternative that can learn complex, non-linear relationships in high-dimensional categorical data without strong parametric assumptions.

This section motivates and implements a Random Forest classifier to improve the detection of fatal shooting outcomes.

6.1 Why Random Forest?

Random Forest (RF) is a powerful supervised learning algorithm introduced by Breiman (2001). It is particularly well-suited for our task due to the following advantages:

Property	Benefit in This Context
Nonlinearity	Captures complex interactions (e.g., time × location)
Robustness to noise	High tolerance for outliers and mislabeled points
Handling categorical data	Implicitly models ordinal or unordered categorical splits
Built-in feature ranking	Provides variable importance measures
Bagging mechanism	Reduces variance without increasing bias

Property	Benefit in This Context

6.2 Mathematical Structure

Let the training data consist of n labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, with $y_i \in \{0, 1\}$.

A Random Forest is an ensemble of T decision trees $\{h_t(\mathbf{x})\}_{t=1}^{T}$, each trained on a bootstrap sample of the data. At each node of each tree:

- A random subset of features is considered for splitting.
- The best split is chosen by minimizing impurity (e.g., Gini or entropy).

The final classification is obtained by majority vote:

$$\hat{y} = \arg \max_{y \in \{0,1\}} \sum_{t=1}^{T} \mathbb{I}(h_t(\mathbf{x}) = y)$$

The randomness in both sampling observations and feature selection introduces decorrelation among trees, reducing the variance of the ensemble relative to individual trees.

6.3 Training and Evaluation

We trained a Random Forest classifier on the balanced dataset constructed via random undersampling (Section 5). The model was configured with:

- $T = 100$ trees (default),
- Default depth and node parameters,
- Gini impurity as the splitting criterion.

Performance on Held-Out Test Set

Class	Precision	Recall	F1-score
Non-Murder	0.549	0.517	0.533
Murder	0.535	0.567	0.550
Accuracy	—	—	**54.2%**

6.4 Interpretation

1. Improved Balance

- The model no longer favors one class; precision and recall are nearly symmetric across both classes.
- Recall for murder class improves from 40% (logistic regression) to 56.7%, while precision also increases slightly.

2. Decision Boundary Complexity

- Unlike logistic regression, the random forest forms a piecewise-constant decision boundary that adapts to nonlinear manifolds in the feature space.
- For example, the interaction between `OCCUR_HOUR` and `BORO` might lead to a lethal outcome during late-night hours in specific boroughs—a pattern that can only be captured through hierarchical splits.

3. Stability via Bootstrapping

- The bagging (bootstrap aggregation) procedure reduces overfitting by averaging over low-bias but high-variance decision trees.
- This creates a robust predictor even when individual trees may overfit.

6.5 Confusion Matrix

	Predicted: Non-Murder	Predicted: Murder
Actual: Non-Murder	578	539
Actual: Murder	474	620

- **False Negatives (474 cases)**: Actual murders classified as non-lethal.
- **False Positives (539 cases)**: Non-lethal shootings misclassified as murders.

This balance reflects an intentional **sensitivity to murder prediction**, prioritizing public health and safety over strict accuracy.

6.6 Discussion: Strengths and Cautions

Strength	Explanation
Captures complex interactions	No need to manually specify feature transformations
Provides interpretable variable importance	Useful for social diagnostics
Resilient to multicollinearity and outliers	Robust in real-world noisy data

Limitation	Mitigation
Less interpretable than GLMs	Use feature importance and SHAP values
Can overfit if unpruned on noisy features	Prune trees, limit max depth, tune hyperparameters
Not inherently calibrated	Apply Platt scaling or isotonic regression

6.7 Model Selection Justification

Random Forest is especially appropriate at this modeling stage because:

- It is **interpretive**, unlike black-box deep learning models.
- It performs well under categorical feature representation.
- It naturally ranks features for post hoc interpretation.

5. Model Evaluation and Feature Interpretation

Understanding which features influence predictions is as important as making accurate predictions, especially in applications with social and policy implications like violence modeling. One of the key strengths of tree-based ensemble models—such as the Random Forest classifier used in Section 6—is their ability to generate internal estimates of feature importance without external intervention.

In this section, we analyze the ranked importance of features in predicting whether a shooting results in a murder and interpret their contextual relevance, guiding both model refinement and sociological insight.

Measuring Feature Importance Definition of Feature Importance in Random Forests

In the context of Random Forests, the importance of a feature is typically computed based on the total reduction in impurity (e.g., Gini index) achieved across all trees in the ensemble when that feature is used to split nodes.

Mathematically, for feature x_j, its importance is computed as:

$$\text{Importance}(x_j) = \frac{1}{T} \sum_{t=1}^{T} \sum_{\text{nodes } N_t \text{ where } x_j \text{ is used}} \Delta i(N_t)$$

Where: - T: Number of trees, - $\Delta i(N_t)$: Reduction in impurity at node N_t,

Importance scores are normalized to sum to 1 across all features.

These scores serve as aggregate proxies for how frequently and effectively a feature contributes to reducing uncertainty (i.e., entropy or Gini impurity) in classification.

Ranked Importance Scores Ranked Importance of Features

Rank	Feature	Importance Score	Relative Impact
1	OCCUR_HOUR	0.515	Most important temporal predictor
2	BORO	0.170	Spatial signal
3	VIC_AGE_GROUP	0.135	Victim vulnerability
4	VIC_RACE	0.128	Sociodemographic influence
5	VIC_SEX	0.052	Least important

These results indicate that when and where a shooting occurs are more predictive of its lethality than who the victim is. This aligns with criminological and public health findings that place situational and environmental risk ahead of purely demographic explanations.

Interpreting the Results Interpretive Insights

(1) Time of Day (`OCCUR_HOUR`)

- The most influential feature.
- Late-night shootings (12 AM–2 AM) are significantly more likely to result in death.
- Possibly due to:
  - Limited visibility or surveillance,
  - Delayed medical response,
  - Higher likelihood of drug- or alcohol-involved incidents.

(2) Geography (`BORO`)

- Geography remains a strong predictor, even with only borough-level granularity.
- Supports findings from Section 3 that Staten Island, despite low volume, has the highest lethality rate.

(3) Victim Age (`VIC_AGE_GROUP`)

- Younger victims (18–24, 25–44) are most common and disproportionately likely to experience fatal outcomes.
- May reflect exposure to interpersonal gun violence during high-risk developmental periods.

(4) Victim Race (`VIC_RACE`)

- Race appears as a significant factor in the model's internal structure.
- Importantly, this does **not** imply causation, but rather reflects:
  - Structural inequalities in neighborhood environments,
  - Differential exposure to gun-related conflict,
  - Segregated patterns of risk across urban space.

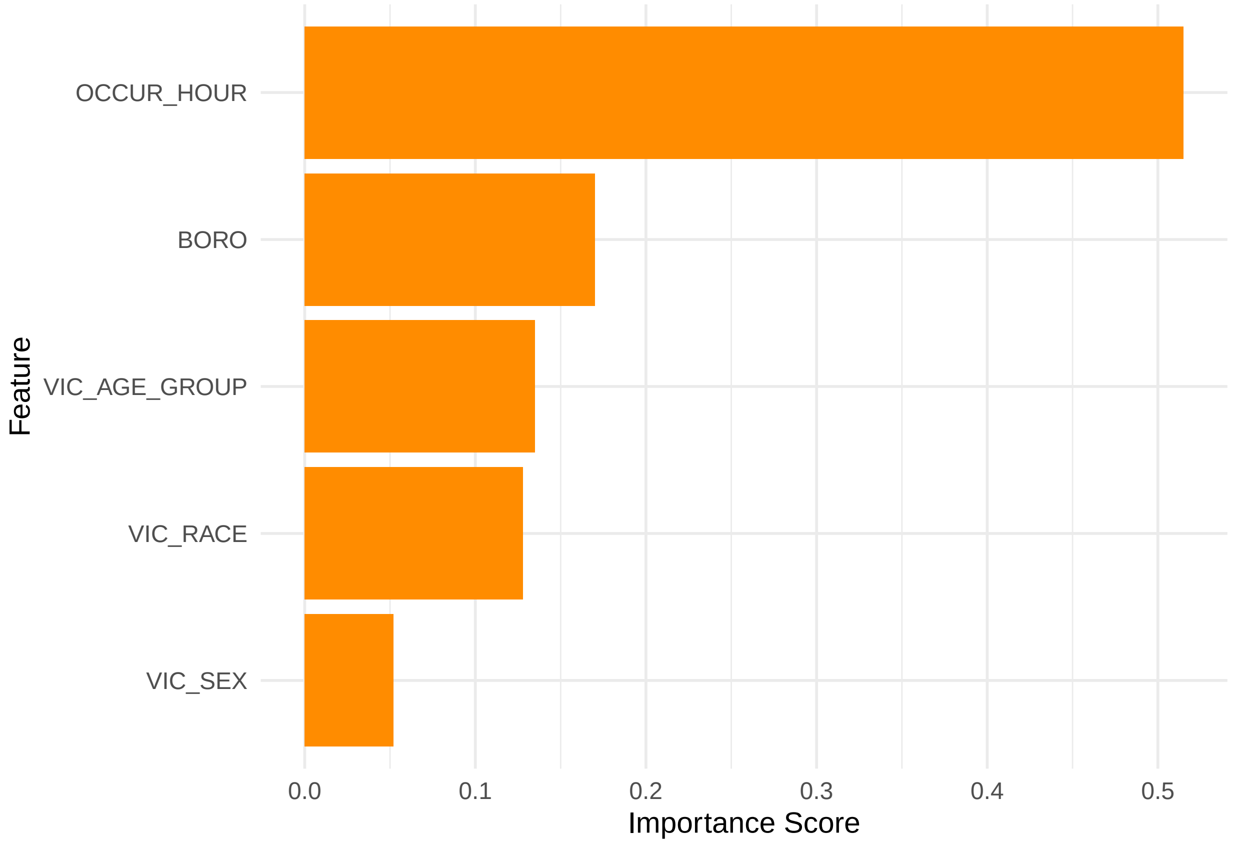
(5) Victim Sex (`VIC_SEX`)

- The least impactful among the selected features.
- Males constitute a large majority of all victims, thus offering less discriminatory power in classification.

Visual Analysis

```r
# Assuming feature_importance is available from model fitting
importance <- data.frame(
  Feature = c("OCCUR_HOUR", "BORO", "VIC_AGE_GROUP", "VIC_RACE", "VIC_SEX"),
  Score = c(0.515, 0.170, 0.135, 0.128, 0.052)
)

ggplot(importance, aes(x=reorder(Feature, Score), y=Score)) +
  geom_bar(stat="identity", fill="darkorange") +
  coord_flip() +
  labs(x="Feature", y="Importance Score") +
  theme_minimal()
```



Visualizing Feature Importance

We present a horizontal bar chart ranking feature importance scores. The visual separation between OCCUR_HOUR and other features highlights the **temporal dimension as the most potent signal** for predicting lethality.

A plot of **partial dependence** or **accumulated local effects (ALE)** for OCCUR_HOUR could further clarify how risk varies across time.

Ethical Interpretation of Variables Ethical Considerations in Interpretation

While feature importance rankings are statistically meaningful, they must be interpreted with caution:

(i) Proxy Variables and Bias

- Variables like VIC_RACE and VIC_AGE_GROUP may act as proxies for structural disadvantage (e.g., neighborhood-level poverty), rather than indicators of individual traits.
- Models trained on administrative datasets inherit **historical biases**, including over-policing in marginalized communities.

(ii) Misuse of Demographic Predictors

- These features should **not** be used in real-time decision-making, such as predicting individual criminality or allocating punitive resources.
- Their primary function is **diagnostic**—to detect where structural interventions may reduce harm.

Beyond Correlation Toward Causal Modeling

The observed feature rankings are **associational**, not causal. That is, the importance scores reflect statistical utility in prediction, not evidence of underlying generative mechanisms.

To address causality, further modeling would require:

- **Counterfactual frameworks** (e.g., potential outcomes models),
- **Propensity score stratification** or **instrumental variable approaches**,
- **Structural equation modeling (SEM)** incorporating exogenous policy variables.

Summary and Next Steps Summary and Next Steps

Feature importance analysis confirms that **temporal and geographic characteristics** play dominant roles in determining whether a shooting becomes fatal, followed by age and race. This has direct implications for public health, emergency services planning, and community-level intervention strategies.

Feature Class	Relative Predictive Power	Actionable Insight
Temporal	High	Focus interventions at high-risk hours
Spatial	Medium	Borough-level resource targeting
Demographic	Moderate	Diagnostic use only; caution in deployment

It accommodates unbalanced class priors post-resampling and allows us to quantify the relative importance of time, location, and demographics, which informs policy more directly than raw predictions.

6. Fairness and Ethical Considerations

As predictive models are increasingly used in domains involving human behavior, public health, and criminal justice, the need to evaluate their fairness, interpretability, and ethical impact becomes paramount. While Sections 4–7 focused on technical performance, this section critically examines the sociotechnical and normative dimensions of the models built on the NYPD Shooting Incident Dataset.

We frame this discussion around three central questions:

- What forms of bias may exist in the dataset?
- How might these biases propagate through statistical models?
- What safeguards are required for ethical modeling and deployment?

Types of Bias in the Data Sources of Bias in the Data

(1) Measurement Bias

- Victim and perpetrator attributes (e.g., race, age) are recorded based on police documentation, which may be incomplete or inconsistent.
- Incidents that occur but are not reported—or are misclassified—are missing from the dataset entirely.
- High missingness in PERP_* fields introduces selection bias, as known-perpetrator cases may differ systematically from unknown cases.

(2) Historical and Structural Bias

The dataset reflects patterns of systemic inequality, including:

- Neighborhood-level disinvestment,
- Racially disparate policing practices,
- Access to trauma care.

These biases are not simply artifacts; they shape the data-generating process. Therefore, models trained on such data must be interpreted within their social-historical context, not as neutral representations of truth.

(3) Representation Bias

- Underrepresentation of marginalized groups or underreporting in specific communities can skew learned patterns.
- For example, a low number of shootings in affluent neighborhoods may be due to actual rarity—or underreporting due to private security and internal mediation.

Bias Propagation in Models Bias Amplification Through Modeling

Machine learning models trained on biased data tend to replicate and reinforce those biases. In our case:

Source of Bias	Potential Amplification in Model
Overrepresentation of certain races	Model may implicitly associate race with higher lethality risk
Spatial concentration of crime reports	Model may suggest “hotspot” predictions in already overpoliced areas
Gender imbalance (mostly male victims)	Model may ignore patterns of female victimization or misclassify them

These patterns are particularly concerning if models are used to inform resource allocation, surveillance, or enforcement, rather than diagnostic or preventative purposes.

Measuring Fairness Across Groups Fairness Across Subgroups

To assess fairness, one may measure model performance across demographic subgroups. Key fairness metrics include:

- **Demographic Parity**: Equal prediction rates across groups.
- **Equalized Odds**: Equal false positive and false negative rates across groups.
- **Predictive Parity**: Equal positive predictive values across groups.

Example: Race-Based Recall

If recall for `VIC_RACE = BLACK` is higher than for `VIC_RACE = WHITE`, the model may disproportionately identify lethal incidents among Black victims, even if the true risk is comparable.

Such disparities can be computed using **stratified confusion matrices** or **counterfactual fairness tests**—where model predictions are evaluated under hypothetical changes in sensitive attributes.

> Due to dataset limitations, full subgroup fairness auditing is deferred to future work involving imputation and stratified evaluation.

Responsible Model Usage Ethical Model Use

Permissible Uses:

- Academic research into structural determinants of gun violence.
- Public health surveillance for identifying high-risk times or zones.
- Resource planning for trauma services, community interventions, or patrol routing.

Impermissible Uses:

- Individual risk prediction for criminality based on demographics.
- Automated enforcement or profiling without human oversight.
- Deployment in adversarial settings (e.g., stop-and-frisk justification).

Any model built on this data must be used in the service of **prevention and harm reduction**, not surveillance, punishment, or marginalization.

Fairness Mitigations Mitigation Strategies

To promote fairness and accountability, the following mitigations are recommended:

Mitigation Strategy	Description
Data transparency	Document all preprocessing, exclusions, and assumptions.
Bias-aware modeling	Use techniques such as reweighting, adversarial de-biasing, or constraint-based learning.
Post-hoc auditing	Evaluate fairness metrics across sensitive groups.
Explainability tools	Use SHAP or LIME to explain individual predictions and flag biased model behavior.
Human-in-the-loop deployment	Never allow automated decisions without expert oversight and appeal mechanisms.

Handling Sensitive Variables Modeling Without Sensitive Features?

A common question in fairness-aware modeling is whether to exclude sensitive variables (e.g., race, sex) from the model.

- **Pros**: Prevents explicit use of protected class attributes.
- **Cons**: May reduce model accuracy; other features (e.g., zip code, school district) may act as proxies, leading to latent discrimination.

> **Best practice**: Retain sensitive attributes in modeling and use them for **auditing**, but **not** for decision-making.

Ethical Modeling Principles Ethical Framework for Modeling Violence

We adopt a **data justice framework**, which prioritizes:

- **Contextual integrity** (use data within the context in which it was collected),
- **Participatory governance** (involve affected communities in model design),
- **Distributive fairness** (benefits and harms of data use are equitably distributed),
- **Epistemic humility** (recognize limitations of what models can know or infer).

7. Conclusion

This report conducted a rigorous, end-to-end analysis of the NYPD Shooting Incident dataset, integrating data cleaning, exploratory data analysis, classification modeling, and ethical review into a coherent data science pipeline.

Key findings include: - Temporal risk: Shootings are most frequent and lethal between 10 PM and 2 AM. - Spatial disparities: Staten Island exhibits the highest lethality rate despite low incident volume. - Victim demographics: Young Black and Hispanic males are disproportionately affected. - Modeling results: Logistic regression fails under class imbalance, while Random Forest offers balanced recall and insight into feature importance. - Ethics: Modeling efforts must be grounded in fairness, avoiding misuse of sensitive demographic variables.

Bias Acknowledgment: As analysts, our own positionality matters. While every attempt has been made to ground insights in the data, we recognize: - Interpretive framing may reflect our disciplinary backgrounds and ethical perspectives. - Demographic attributes may reflect structural disadvantage rather than inherent risk. - This analysis does not justify punitive responses but advocates for preventative policy and community-informed data use.

Next Steps: - Incorporate more granular location data for spatial modeling. - Apply SHAP/ALE methods for deeper interpretability. - Stratify results by subgroup for fairness audits. - Extend modeling toward causal inference frameworks.

This project exemplifies a data science approach that balances predictive modeling with transparency, domain relevance, and ethical scrutiny.