Analyzing the Risk of Civil Conflict in the United States: An Examination of Economic and Social Factors

Adam Humble[1], Kyle Montgomery[1], Nick Bartlett[1], and Caleb Kennedy[1]

[1]Arizona State University, Tempe, AZ 85281, USA

October 3, 2024

Abstract

The increasing political polarization in the U.S. along with economic disparities and widespread social unrest have raised concerns about the potential for civil conflict. This study explores the relationship between wealth inequality, economic instability, and the occurrence of civil unrest using various statistical and machine learning techniques. Economic and protest data was used in order to discover the relationship between these two variables.

The analysis incorporated data for inflation and unemployment in the U.S. compared to protest data for violent and non-violent events. While Mutual Information scores showed a moderate relationship between protest frequency and both the inflation rate alongside unemployment, Spearman's rank correlation showed weak relationships between these economic indicators and protest counts, with p-values indicating no statistically significant linear correlation.

Data analysis revealed significant wealth disparities, with the top 10% of earners holding an increasing share of wealth over time, while the bottom 50% saw their share diminish. Visualizations highlighted a growing disparity in median income across regions, contributing to heightened social tensions.

This research contributes to understanding the economic, social, and political conditions that drive social unrest and provides risk factors that may lead to a civil conflict in the United States. Through identifying indicators of instability, this research can help decision-makers address wealth inequality and address political polarization to mitigate the risk of civil unrest.

1 Introduction

Civil wars are incredibly destructive events with a broad impact on a society, and their frequency has increased over the past century. Between 1945 and 1999, approximately 127 of these conflicts took place across 73 nations [9]. This increase in civil wars has impacted much of the world, but The United States has avoided this level of conflict for nearly 150 years. With increasing political polarization, economic inequality, and the role social media plays in fueling unrest, there is a rising concern that the United States could be approaching another civil war. Understanding the causes of these events is essential to addressing this concern.

Political and economic tensions have consistently been found to be underlying factors that lead to civil war. Poverty was identified as a symptom that, if not addressed, could lead to aggression between a government and its citizens [9]. The likelihood of civil conflict is increased not only by poverty, but also by significant economic polarization between groups within a society [18]. Significant wealth disparity within a country can exacerbate social divisions, increasing the likelihood of unrest. Low income per capita is particularly associated with an increased probability of civil conflict, as individuals in economically deprived regions often face lower opportunity costs, making rebellion or conflict seem more appealing [6]. These findings reveal that a key topic to understand in regards to the escalation of societal tension into civil war is not only poverty, but wealth inequality.

1.1 Background and Motivation

The prediction of civil conflict has evolved significantly over the past few decades. We used methods to compare traditional models that are based on statistical analysis to modern approaches which use machine learning.

Early models for predicting civil conflict, such as those developed by Fearon and Laitin, and Collier and Hoeffler, relied heavily on regression analysis. These models sought to identify correlations between socio-economic and political variables—such as poverty, political instability, and resource dependence—and the onset of civil conflict [9, 6]. The models offered valuable insights, but struggled with real-world applicability. The in-sample success of traditional models did not consistently translate to accurate out-of-sample predictions. This discrepancy between statistical significance and predictive power emphasizes the limitations of traditional approaches [21]. Simpler models, which included fewer variables, sometimes outperformed more complex ones that incorporated numerous statistically significant factors. This was because models with too many variables struggled with generalization, reducing their accuracy in predicting civil unrest [21].

With machine learning, there is a more iterative framework for improving predictive models. Unlike traditional hypothesis-testing methods, machine learning models can continuously refine their performance based on new data inputs [14]. In the context of conflict forecasting, supervised learning techniques such as random forests and neural networks have demonstrated superior performance compared to simpler statistical models like logistic regression [5]. Machine learning models excel in their ability to handle vast datasets and uncover non-linear relationships that traditional models might overlook. Moreover, their emphasis on out-of-sample forecasting, which tests the model's ability to generalize to new and unforeseen data, is crucial for ensuring accuracy and avoiding overfitting [5]. Overfitting occurs when a model becomes too tailored to its training data, making it less effective in real-world applications. Processing large and diverse datasets such as social media activity, economic indicators, and political participation allow machine learning to have predictive capabilities [22]. These models can identify regions at high risk of conflict by analyzing complex interactions between variables.

The aim of this project was to examine socioeconomic causes of civil conflicts, drawing on both historical perspectives and contemporary research. We explored modern prediction techniques such as machine learning models which enhance the ability to predict the onset of civil conflict using modern data from the United States. By using historical insights and modern predictive analysis methods, we sought a better understanding of the factors that drive civil unrest and the tools available to anticipate it.

1.2 Methodologies in Literature

1.2.1 Economic Causes of Civil Conflict

Economic inequality, poverty, and competition for natural resources have long been recognized as critical drivers of civil conflict. Extensive literature highlights that significant wealth disparities within a nation can deepen social divisions, increasing the likelihood of unrest. Economic hardship, particularly in the form of low income per capita, reduces the opportunity cost of engaging in rebellion, making conflict more attractive for marginalized populations. As noted by previous research, economically deprived regions face higher probabilities of conflict, whereas wealthier nations generally provide more stability due to the high costs of rebellion outweighing any potential gains [6].

Additionally, population size and composition can exacerbate civil conflict risk. Larger populations may foster secessionist movements, as regional groups seek autonomy, while ethnic diversity—especially in the presence of moderate economic inequality—can enhance coordination among rebel groups. This is particularly evident in conflicts driven by economic disparity and competition for resources among different ethnic groups. The role of natural resources in civil conflict is complex: while the presence of valuable resources often incentivizes conflict, it can also strengthen governments' capacities to suppress rebellion when these resources bolster state power [6].

The dynamics of natural resource dependence follow a non-linear pattern, as evidenced by conflicts in resource-rich regions such as the Democratic Republic of Congo (DRC) and Nigeria. In these cases,

competition over resources like oil, gold, and minerals has fueled prolonged civil unrest, intensifying existing socio-economic inequalities [8]. These examples underscore the importance of governance structures and the equitable distribution of resources in mitigating conflict risk [19].

1.2.2 Political Instability and Governance

While economic factors lay the groundwork for civil conflict, political instability plays a decisive role in whether these tensions escalate into full-scale civil wars. Governments with weak institutional frameworks, particularly those that are neither fully democratic nor autocratic, often lack the capacity to manage political opposition effectively. Such regimes are more susceptible to civil conflict, as their inability to maintain governance structures creates opportunities for rebellion [13]. State capacity, including a government's ability to tax effectively and maintain bureaucratic quality, is a key determinant of a state's resilience against conflict. Weak state capacity, coupled with economic instability, heightens the likelihood of civil war [13].

In the modern era, political instability is further amplified by the role of social media, which facilitates protest coordination and the dissemination of information. Studies have demonstrated the predictive power of sentiment analysis tools in monitoring shifts in public mood, which can signal potential conflict. Social media platforms enable rapid organization among protesters, creating environments where civil unrest can escalate quickly, especially in politically fragmented states [3]. The Arab Spring uprisings serve as a prime example of how political grievances, combined with social media coordination, can lead to widespread protests and political instability. Social media played a crucial role in coordinating activists, encouraging activism in neighboring regions, and discouraged government repression as accounts of these activities would encourage more opposition [15].

1.2.3 Technological Causes: Social Media and Machine Learning

The integration of technology into conflict prediction has transformed how governments and researchers anticipate civil unrest. Social media, in particular, has become a double-edged sword—serving both as a platform for organizing movements and as a tool for predicting potential conflict. Platforms like Twitter and Facebook allow activists to coordinate rapidly, while also providing researchers with real-time data on public sentiment and political activity [1].

Predictive technologies, such as machine learning, have emerged as valuable tools in conflict forecasting. Machine learning models excel in analyzing vast datasets to detect patterns in economic disparity, resource dependence, and political participation. These models surpass traditional statistical approaches by offering higher out-of-sample performance, allowing them to generalize effectively to new and unforeseen conflict scenarios [5]. By incorporating real-time economic indicators and social media data, machine learning algorithms can identify regions at high risk for civil conflict, offering governments an opportunity to intervene proactively [4].

Machine learning also offers the potential to mitigate conflict before it escalates. By continuously monitoring social media feeds and applying sentiment analysis, governments can detect shifts in public mood and political participation that may indicate rising tensions. As technology continues to evolve, its role in both predicting and preventing civil conflict will become even more significant, making it an essential component of modern conflict prevention strategies [21].

1.3 Significance of the Study

This study analyzed the relationship between economic factors—specifically wealth inequality and economic instability—and the occurrence of civil unrest in the United States. Using statistical methods, along with machine learning techniques, we evaluated data on inflation rates, wealth inequality, and protest frequencies. Our findings indicate a moderate relationship between protest frequency and both inflation and wealth inequality. However, linear correlations were weak and not statistically significant. Additionally, wealth disparity visualizations showed an increasing concentration of wealth among the top earners and a diminishing

share for the bottom 50%. These results suggest that while economic factors contribute to social tensions, they may not directly predict civil unrest without considering other variables. This study contributes to the literature by providing a nuanced understanding of how economic indicators relate to civil conflict risk in the modern U.S. context.

2 Methods

2.1 Methods: Data Sources

The Geographic Wealth Inequality Database (GEOWEALTH-US) [20], distributed by the Inter-university Consortium for Political and Social Research (ICPSR), provides measures of wealth inequality based on surveys that were conducted by the U.S. Federal Reserve between the decades of 1960 and 2020. Multiple files are included in this database that separate the data based on varying geographical scales (e.g., regions, states, and metropolitan areas), but our research focused on Commuting Zones, the smallest division available, which were cross-referenced with Federal Information Processing Standards (FIPS) codes using a table from the IPUMS USA website [10]. Economic variables in the GEOWEALTH-US database include the share of wealth held by percentiles of the population, mean and median household wealth, mean income, home ownership, and the age and ethnicity of the respondents. This dataset allows for an analysis of wealth inequality trends in the United States over time.

We identified two datasets to aid in analyzing civil conflicts in The United States. To measure protest activity, we used the Crowd Counting Consortium Dataset (CCC), which monitors activities including marches, protests, riots, and demonstrations between 2017 and 2024 [7]. This data contains fields with geographical information such as the state, locality, and FIPS code for where each event took place. There are also fields describing the actors involved, various accounts on the size of the crowd, and binary fields which describe whether there were injuries, arrests or property damage.

The second dataset regarding civil conflict is the Nonviolent and Violent Campaigns and Outcomes (NAVCO) [11] dataset. This data provides insights into the prevalence and nature of various civil conflicts between 2007 and 2011. Fields from this dataset describe the date these events took place and categorizes the parties involved and the kinds of actions taken. These categorical variables allow for distinctions to be made between governmental and non-governmental actors as well as violent and non-violent events.

Additional economic data was required due to the GEOWEALTH-US database containing information by decade, while we aimed to analyze conflict on an annual or monthly basis. For inflation, we used the Consumer Price Index for All Urban Consumers (CPI-U) [16] dataset from the U.S. Bureau of Labor Statistics, which measures changes in the Consumer Price Index (CPI), a common measure of inflation, every month between January of 1999 and August of 2024 in the United States. To supplement CPI, we also used the Unemployment Rate in the United States [17] for the same months as CPI. For wealth inequality, we used the Selected Measures of Household Income Dispersion: 1967-2023 [12] data from the United States Census Bureau, which provides annual income percentiles of the US Population from the 10th to the 95th percentiles, along with ratios between the 90th to 10th, 90th to 50th, and 50th to 10th percentiles.

2.2 Methods: Data Preparation and Exploratory Data Analysis

We began our EDA process by cleaning and analyzing the dataset from the GEOWEALTH-US database [20] that contains wealth inequality data from 1960 - 2020 for commuting zones across The United States. This data contains 39 fields including the year, commuting zone, number of records observed in that zone during the year, and the average percentage of wealth held by percentiles of the population. There are 1,469 null values in this data, but they only occurred for the decades of 1960 and 1970 in the columns containing the percentage of respondents who owned their homes outright or were paying a mortgage in their commuting zone. We did not use these fields in our analysis, so no changes were required.

Our primary interest in this dataset was in the percentage of the total wealth owned by the percentiles of the population over time, so the first step of our exploration of the data was to visualize the distribution

of each of these fields. We then joined FIPS codes to this data using data from the IPUMS USA website [10] to aid in visualizing the median household wealth by commuting zone and decade using geopandas, along with the percentage of wealth held by the bottom 50th percentile of the population by commuting zone and decade. There was a significant skew in the distribution of both of these variables, so a natural log and log base 10 were applied for each of them respectively to provide a more insightful visualization on the map.

We used the CCC [7] and NAVCO [11] datasets for our statistical tests and machine learning models with the intent of predicting civil unrest and violent events using economic predictors.

The CCC data included many variables with null values, mostly due to incomplete reporting. The variables of interest for our analysis, however, contained no null values. This data also included FIPS code for each protest along with the state, year, and month they took place. There are four boolean variables that indicate whether there were any arrests, property damage, injuries to the crowd, or injuries to police for each protest. We created two new variables to use as predictors for the CCC data: a boolean variable indicating whether each protest was violent based on any of the other four variables being true, and a sum of the number of true values in those four variables as an attempt to measure levels of escalated violence for each protest.

The NAVCO dataset contains codes to categorize the actors and targets of each event, as well as a category for the kind of action that took place. To focus on civil violence against the US government, explanations for each category from the codebook were used to filter the dataset to only contain events that took place between non-government actors against the government. The codes in Table 1 were used to perform this filter. From there, we created a boolean target variable to indicate whether the event that took place was violent based on "verb" codes that were also described in the codebook.

Table 1: Actor Codes in the NAVCO Dataset [11]

Primary Role Codes	**Description**
COP	Police forces, officers.
GOV	Government: the executive, governing parties, coalitions partners.
JUD	Judiciary: judges, courts.
MIL	Military: troops, soldiers, all state-military personnel.
OPP	Political opposition: opposition parties, individuals.
REB	Rebels: armed and violent (non-state) groups and individuals.
LLY	Regime Loyalists: not otherwise specified.
ACT	Activists: primarily nonviolent non-state actors.
NON	Nonaligned third party.
SPY	State intelligence, secret service.
UAF	Armed forces that cannot be identified as MIL, COP, or REB.
UNS	Unidentified unarmed non-state actors.

To predict the target variables we created, we used CPI data from the U.S. Bureau of Labor Statistics as a measure of inflation and data from the United States Census Bureau that measured wealth inequality.

The CPI data has one field containing each year along with each month's CPI and an annual CPI. Each month column's header contained the three character abbreviation for the month (e.g. Jan, Feb, Mar). We converted these to their numerical equivalents to aid in joining datasets. We created separate annual and monthly datasets as the CCC data contains months and the NAVCO data does not. Once transposed, the months were then converted to integer data types along with the years. For both of these new datasets, we calculated a percentage change between each month or year and the one before. We created visualizations for this data including the distribution of monthly and annual percentage changes in CPI and the overall growth in the average CPI by year.

The wealth inequality data set contains no null values, but there were some duplicate years based on different practices that were used to measure income. We removed these duplicate rows by keeping the first row from each year. We also removed extra characters from some of the years where footnotes were added to

the original table. Once we cleaned the data and removed duplicate years, we proceeded with viewing some descriptive statistics for the variables in the dataset and creating visualizations to explore it further. The data contains the ratios between the annual income of some percentiles including the 90th/10th, 90th/50th, and 50th/10th. We visualized the distributions of these ratios and used a line chart to view the changes in each ratio over time. We then created the same visualizations for each of the three percentiles included in these ratios individually.

Finally, to prepare for statistical tests and machine learning, both the CCC and NAVCO datasets were enriched by joining the CPI and wealth inequality datasets. The CCC data was joined with the monthly CPI data using the year and month columns and the wealth inequality data using the year column. The NAVCO dataset was joined with both the annual CPI data and the wealth inequality data using the year column.

The CCC dataset captures protests and civil unrest in the United States from 2017 to 2024. With over 200,000 records, we began by counting occurrences of protests by state to observe trends in civil unrest.

2.3 Method: Mutual Information and Spearman's Correlation

Mutual Information was used to assess how much information is shared between protest frequency and economic indicators like CPI and unemployment rate. We also used Spearman's rank correlation to assess the strength and direction of the relationship between these variables. While Mutual Information identifies whether a dependency exists, regardless of its form, Spearman's rank correlation helps gauge the nature and strength of the relationship. We aimed to discover if these economic variables rate as good predictors to protest volume.

2.4 Method: Inferential Statistical Analysis

To investigate the relationship between economic factors and civil unrest in the United States, we conducted inferential statistical analyses using the Nonviolent and Violent Campaigns and Outcomes (NAVCO) dataset and the Crowd Counting Consortium (CCC) dataset. Our primary objective was to determine whether significant differences exist in economic indicators, such as the Consumer Price Index (CPI) and wealth inequality ratios, between violent and non-violent events.

For the NAVCO dataset, we first classified events as violent or non-violent based on specific action codes corresponding to violent activities, such as physical assaults or armed conflicts. We merged this event data with annual CPI data from the U.S. Bureau of Labor Statistics and wealth inequality data from the U.S. Census Bureau by aligning the dates of the events with the corresponding economic indicators. This alignment ensured that the economic conditions preceding or concurrent with each event were accurately represented.

We performed independent samples t-tests to compare the mean CPI values and wealth inequality ratios between violent and non-violent events. Prior to conducting the t-tests, we verified the assumptions of normality and homogeneity of variances using the Shapiro-Wilk test and Levene's test, respectively. These tests validated the use of t-tests for our data. Additionally, we conducted a one-way Analysis of Variance (ANOVA) to examine differences in economic indicators across multiple groups. Specifically, we selected three states—California, Georgia, and Idaho—with varying numbers of violent protests and compared the mean monthly percentage change in CPI during the months when violent protests occurred.

2.5 Method: Regression Modeling

To quantify the relationship between economic variables and the intensity of civil unrest, we developed regression models using the CCC dataset. We aimed to predict the level of violence in protests and the overall number of protests based on economic indicators.

We identified key economic predictors, including the CPI, monthly percentage change in CPI, and wealth inequality ratios (e.g., 90th/10th income percentile ratios). Protest-specific variables such as crowd size categories were also included to account for event characteristics. We addressed missing data by removing records

with incomplete information and standardized continuous variables using z-scores to ensure comparability. Categorical variables were encoded using one-hot encoding.

We developed linear regression models to predict two dependent variables: the level of violence in protests (using the violent_sum variable) and the number of protests per state per month. The data was divided into training and testing sets using an 80/20 split. We fitted the models using the ordinary least squares method and evaluated their performance using the R-squared value and Root Mean Squared Error (RMSE). Residual analysis was conducted to check for patterns that might suggest violations of regression assumptions.

2.6 Method: Random Forest Modeling

To improve prediction accuracy and capture non-linear relationships between economic indicators and civil unrest, we employed Random Forest models for both classification and regression tasks.

We selected relevant economic indicators and protest characteristics as features, including CPI values, wealth inequality ratios, and crowd size categories. In classification tasks predicting violent versus non-violent events, we addressed class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE), which synthetically generates samples of the minority class to balance the dataset.

For the Random Forest Classifier, we aimed to classify events as violent or non-violent based on economic factors. We optimized hyperparameters such as the number of trees (estimators), maximum depth of the trees (maxdepth), and minimum samples required to split an internal node (minsamplessplit) using grid search and cross-validation techniques. The model's performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices. Feature importance analysis was conducted to identify which variables contributed most to the predictions.

For regression tasks, the Random Forest Regressor was used to predict the number of protests and the level of violence as continuous variables. The model's performance was assessed using R-squared and RMSE metrics.

3 Results

3.1 Results: Data Exploration Analysis

3.1.1 Analysis of Conflict Datasets

The CCC dataset captures protests and civil unrest in the United States from 2017 to 2024.With over 200,000 records, we began by counting occurrences of protests by state to observe trends in civil unrest.

State	2017	2018	2019	2020	2021	2022	2023	2024 (Unfinished)
CA	1266	2602	1563	3726	4592	5635	4765	4558
DC	603	643	529	804	1111	1413	1579	1324
FL	415	962	430	1102	958	1114	1102	726
IL	254	677	333	1046	1271	1833	1367	769
NY	981	1849	1056	3097	3738	3776	3348	3518

Table 2: Number of Protests by State, Top 5 States (2017-2024)

These findings suggest that the economic triggers for civil unrest may vary by region, indicating that localized economic conditions may play a critical role in the intensity and frequency of protests. The variation in CPI changes across different states also highlights the need for more granular analyses of regional economic stressors.

3.1.2 Analysis of Wealth Inequality Using GEOWEALTH-US Database

One of our primary focuses for this project is wealth inequality. We began by exploring the distribution of wealth among the percentiles of the population provided in this dataset.

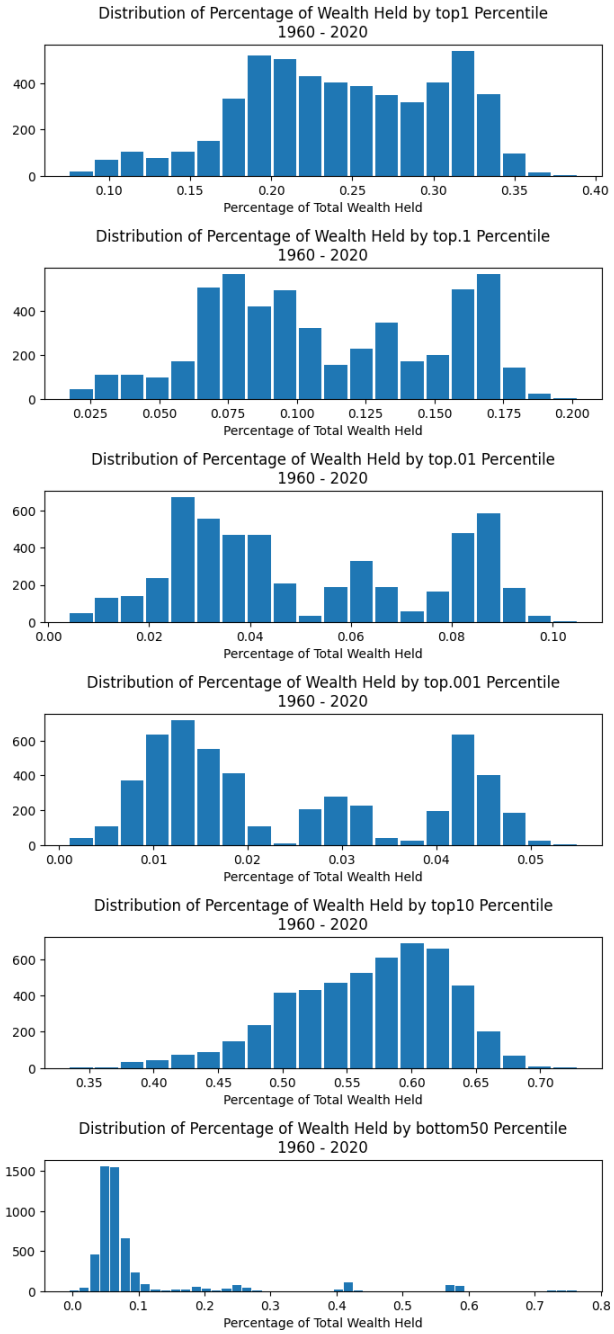

Figure 1: Distribution of Wealth by Percentiles

Figure 1 shows the distribution of variables in GEOWEALTH-US that represent the percentage of wealth held by various percentiles of the population in each Commuting Zone. The disparity between the percentage

of wealth held within the top percentiles of the population and the percentage held by the bottom 50th percentile of the population is immediately visible. The data shows that the bottom 50th percentile of the population owns less than 10% of the wealth in their Commuting Zone in 86.43% of the observations.

Visualizing the median income for different decades shows a decreasing median wealth across the country relative to the top earners, with lighter colors representing a greater median wealth. A natural log was applied to these values.

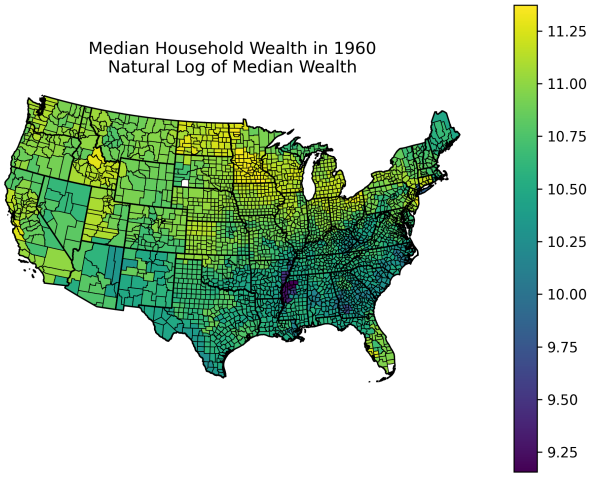

Figure 2: Median Wealth by Commuting Zone, 1960

Figure 2 shows a difference between median wealth across much of the south-eastern United States and the median wealth held in areas such as the West Coast, Midwest, and much of the Northeast.

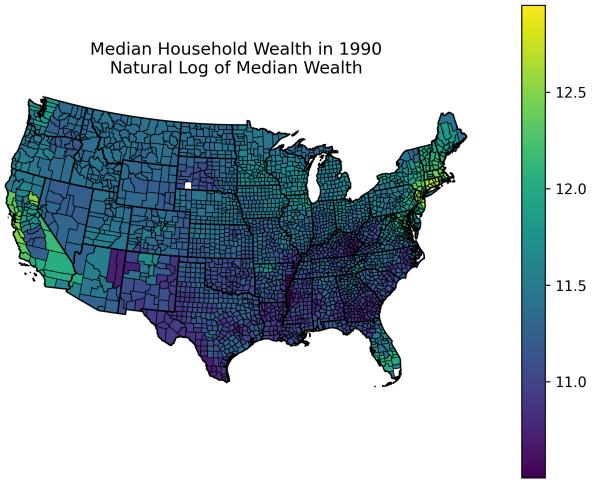

Figure 3: Median Wealth by Commuting Zone, 1990

Figure 3 has a much darker color across the country, with the exception of some areas such as Southern California, New York City and the surrounding area, and some other large cities.

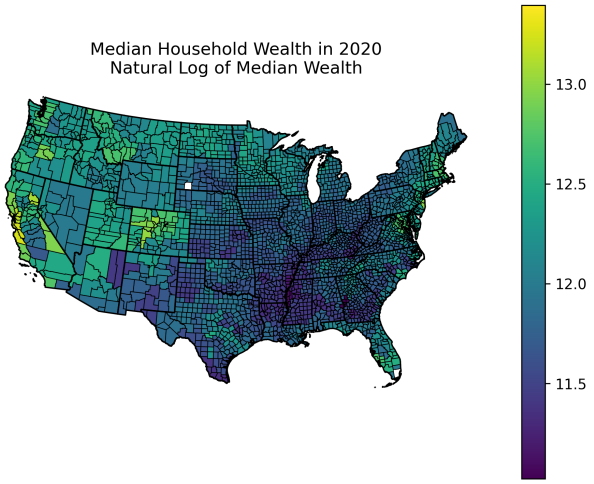

Figure 4: Median Wealth by Commuting Zone, 2020

The visualizations in figures 2, 3, and 4 demonstrate a growing disparity between median incomes in regions of the United States with larger populations and those with smaller populations. This trend is further demonstrated when the difference between the highest median income in each state is compared to the lowest median income.

The following figures 5, 6, and 7 visualize the percentage of wealth held by the bottom 50th percentile of the US population in 1960, 1990, and 2020. A log base 10 was applied to these values.

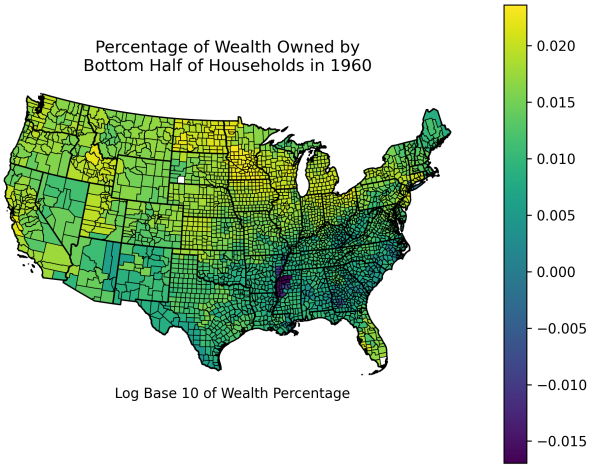

Figure 5: Percentage of Wealth Held by Bottom 50th Percentile, 1960

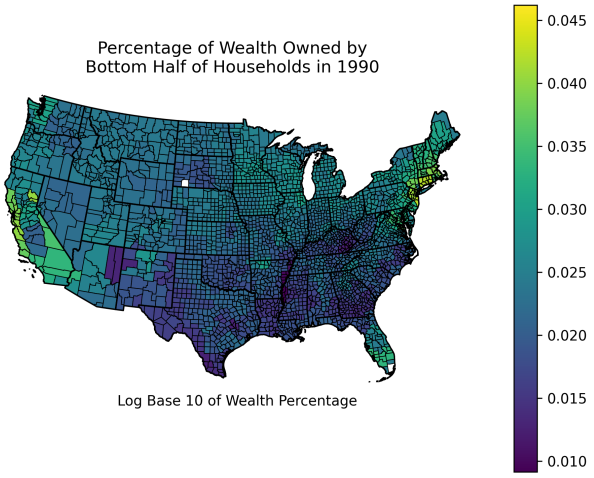

Figure 6: Percentage of Wealth Held by Bottom 50th Percentile, 1990

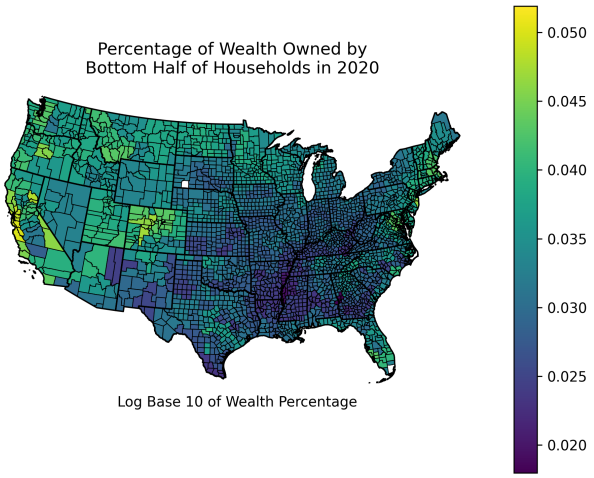

Figure 7: Percentage of Wealth Held by Bottom 50th Percentile, 2020

The difference shown between the wealth held by the bottom half of the US population between 1960 and 1990 is very similar to the difference in median wealth. Figures 5, 6, and 7 demonstrate the United States' transition from relative economic equality across the country to widespread inequality in 1990 and a slight improvement by 2020.

3.1.3 Analysis of CPI and US Census Bureau Wealth Inequality Data

The wealth inequality and CPI data used for statistical tests and machine learning also provided insights into how these situations have worsened over time. The wealth inequality data from the United States Census Bureau contains ratios between various percentiles of household income in the US by year. The distribution of these variables between 1967 and 2023 shows that the 90th and 50th percentiles of the population have a significantly higher household income, as illustrated in Figure 8.

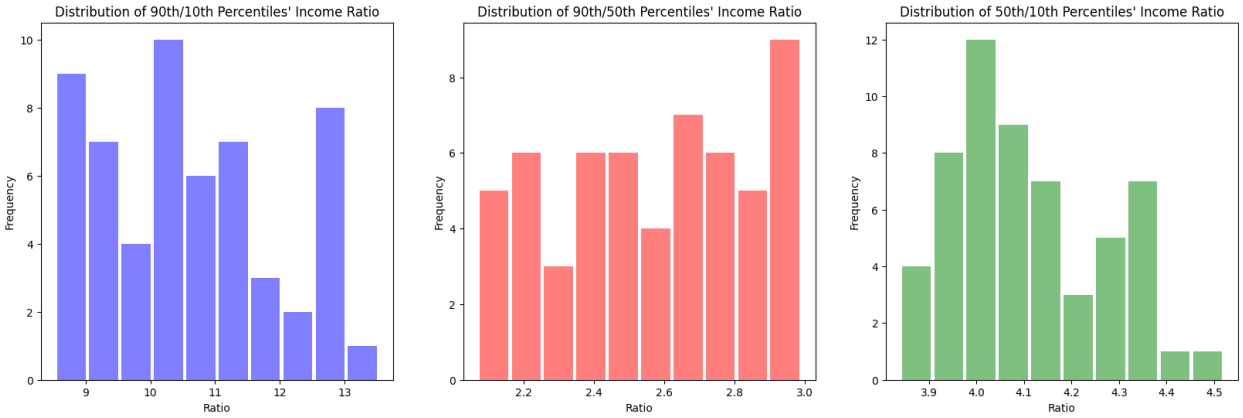

Figure 8: Distribution of Ratios of Household Income between Percentiles: 1967-2023

In Figure 9, there is a visible increase in the ratio between the 90th and 10th percentile over time.

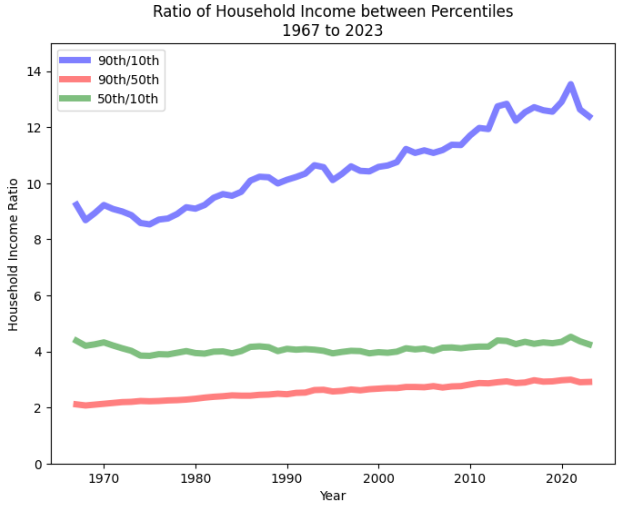

Figure 9: Changes in the Ratios between Percentiles of Household Income Over Time

We looked deeper into the wealth inequality data by visualizing the distribution of household income between these percentiles individually (Figure 10) and the changes in household income over time (Figure 11).

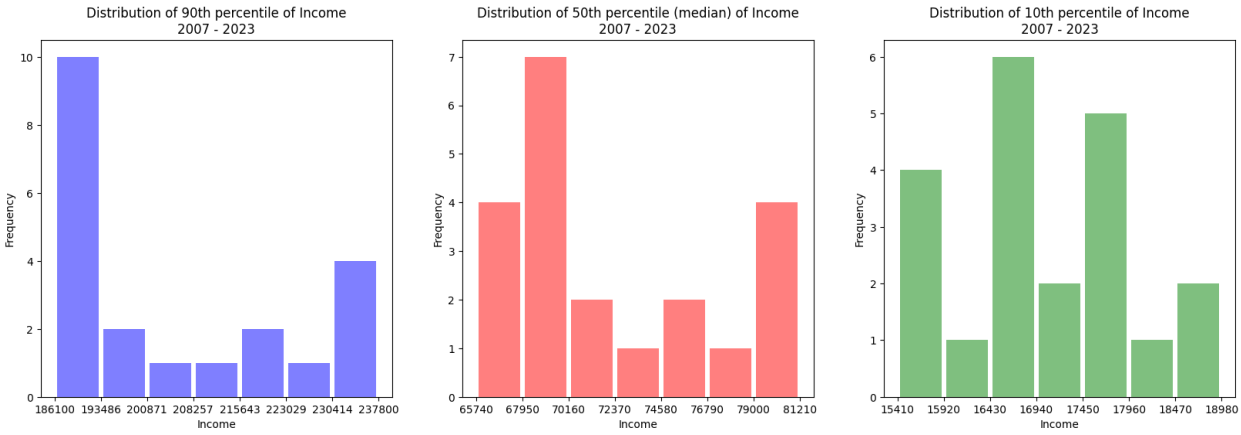

Figure 10: Distribution of the Percentiles of Household Income: 2007 - 2023

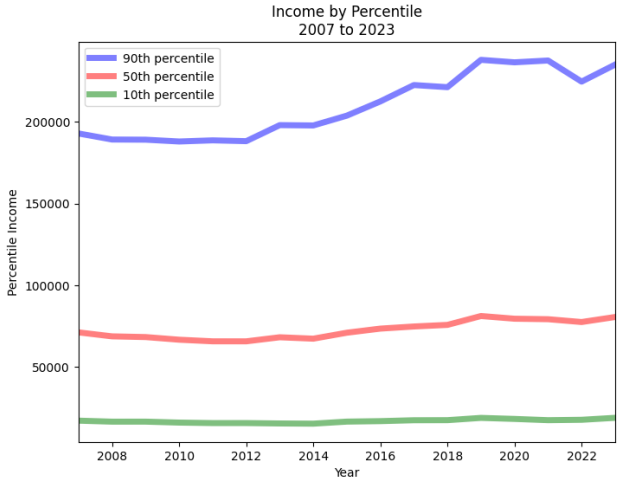

Figure 11: Changes in Percentiles of Household Income: 2007 - 2023

The majority of both the monthly and annual CPI percentages ranged between 0.7 and 4 percent as seen in their distributions below (Figure 12,13). The annual CPI over time is also visualized in Figure 13.

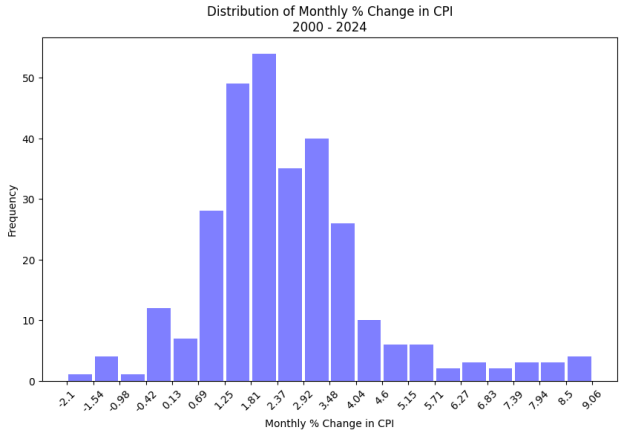

Figure 12: Distribution of Monthly Percentage Change in CPI: 2000 - 2024

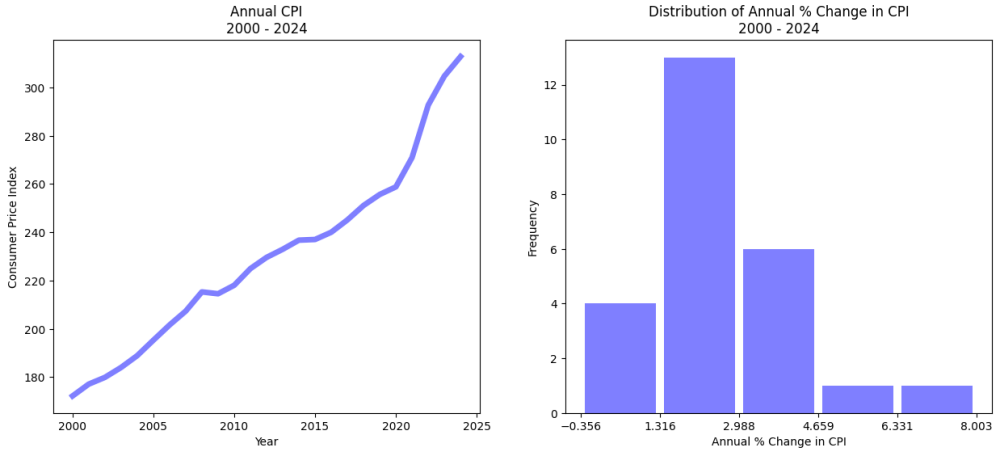

Figure 13: Change in CPI Over Time (Left) and Distribution of Annual Percentage Change in CPI (Right): 2000 - 2024

3.1.4 Analysis of Wealth Inequality and Protest Activity

One of the main focuses of this study was to analyze how growing wealth inequality might drive civil unrest. To explore this, we used the GEOWEALTH-US dataset to evaluate the wealth distribution by percentiles. Our findings highlighted the concentration of wealth in the top 1st and 10th percentiles, contrasted with the diminishing share of wealth held by the bottom 50th percentile over time.

In **Figure 14**, the data shows a steep increase in the share of wealth held by the top 10th percentile between 1960 and 2020, rising from around 35% to over 70% of the nation's wealth. This increase in wealth concentration aligns with growing income inequality, which has historically been a precursor to social unrest. Comparing the bottom 50th percentile reveals a contrasting trend. By 2020, the share of wealth held by the bottom 50% had fallen to below 10%, further indicating deepening inequality.

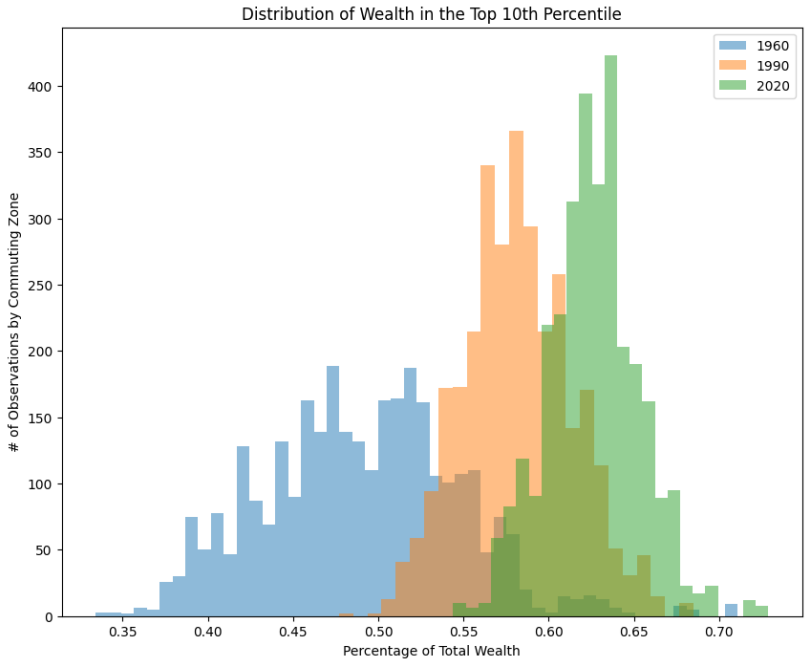

Figure 14: Distribution of Wealth: Top 10th Percentile

3.2 Results: Inferential Statistics Analysis

We conducted t-tests to further assess the disparity between violent and non-violent protest events in different years, particularly wealth inequality and inflation. We compared the CPI and wealth inequality metrics (90th/10th and 90th/50th ratios) for both violent and non-violent protest periods.

The results demonstrated statistically significant differences between violent and non-violent events. Specifically, violent events post-2010 were associated with higher CPI values and higher wealth inequality ratios, reinforcing the connection between economic distress and violent civil unrest.

Table 3: t-Test Results of CPI in Violent and Non-Violent Events

Year	**t-Statistic**	**p-Value**	**Statistically Significant**
(Pre-2010) CPI	-3.52	< 0.001	Yes
(Post-2010) CPI	6.13	< 0.001	Yes
(All Years) 90th/10th Wealth Ratio	4.39	< 0.001	Yes
(All Years) 90th/50th Wealth Ratio	4.37	< 0.001	Yes

As shown in Table 3, violent protests were more strongly associated with higher inflation and wealth inequality metrics, suggesting that these factors may play a critical role in escalating tensions during protest periods.

The results of t-tests on the NAVCO dataset showed significant differences between violent and non-violent events in relation to Consumer Price Index (CPI) [16]. In particular, Table 3 shows violent events before 2010 were associated with lower CPI values, while post-2010 violent events were linked to higher CPI values, suggesting a shift in economic distress patterns. The wealth inequality metrics, specifically the 90th/10th and 90th/50th wealth ratios, were significantly higher for violent events, indicating a potential link between extreme wealth disparity and violent resistance movements.

For our one-way ANOVA test, we aggregated the CCC dataset by state, filtered to only include violent protests, and then ordered it from most violent protests to least. We selected the first, twelfth, and twenty-

fifth states on this list (California, Georgia, and Idaho) to determine whether there was a statistically significant difference between these states regarding the monthly percentage change in CPI during the months these protests took place.

Our one way ANOVA test produced an F statistic of 22.04 with a p-value of 7.43e-10. These results would appear to indicate that there is a significant difference between the monthly change in CPI for each group. In order to determine a significant difference, a Tukey's Honestly Significant Difference (HSD) test was performed. This showed Idaho's monthly percentage change in CPI was statistically different from California and Georgia's, with the latter two having a similar variance between them. This can be seen in the boxplot in Figure 15.

Figure 15: Boxplot: Distribution of Monthly CPI % Changes from Periods with Violent Protests

3.3 Results: Regression Model

Our next approach to modeling the data involved regression analysis to explore the impact of wealth inequality and inflation on protest frequency. The datasets utilized in this analysis include the Crowd Counting Consortium Dataset (CCC) [7] for U.S. protest data, the GEOWEALTH-US [20] for wealth inequality data, and the Annual Inflation by GDP Deflator and Consumer Prices [2] for inflation rates. These datasets were merged and filtered for relevant time periods to create a comprehensive dataset for analyzing protest events.

Using linear regression, we tested whether increases in wealth inequality (measured by the wealth share of the bottom 50%) and inflation (CPI) were associated with higher frequencies of protest. The model found that both inflation and wealth inequality positively correlated with an increased number of protests. This finding aligns with the hypothesis that economic hardship, particularly rising inflation and greater wealth disparity, contributes to social unrest.

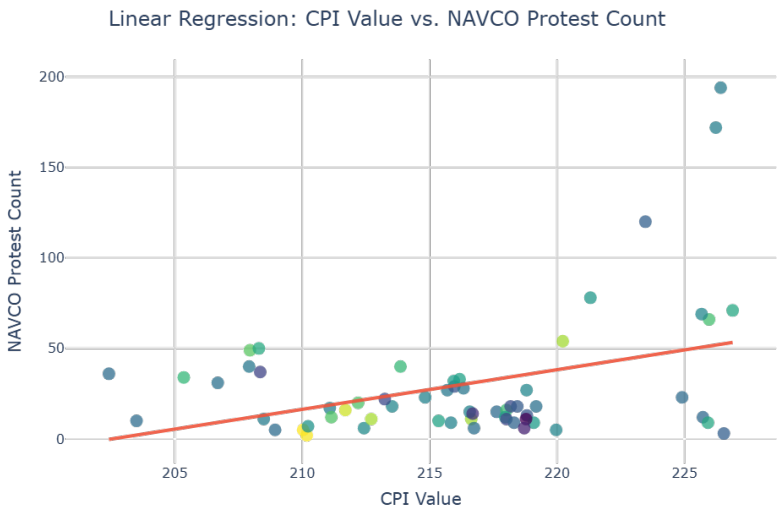

Figure 16: Linear Regression: CPI vs. Number of Protests

In Figure 16, the trend line illustrates that as the Consumer Price Index (CPI) increases, so does the frequency of protest events. This suggests a direct relationship between inflationary pressures and the mobilization of protesters, as inflation could exacerbate economic frustrations and create instability.

3.4 Results: Mutual Information and Spearman's Correlation

The Mutual Information (MI) scores between protest frequency and economic indicators CPI and unemployment were 0.3344 and 0.2698, respectively (Figure 17). These scores suggest a moderate or weak level of shared information, indicating that while there may be some dependency, it is not particularly strong.

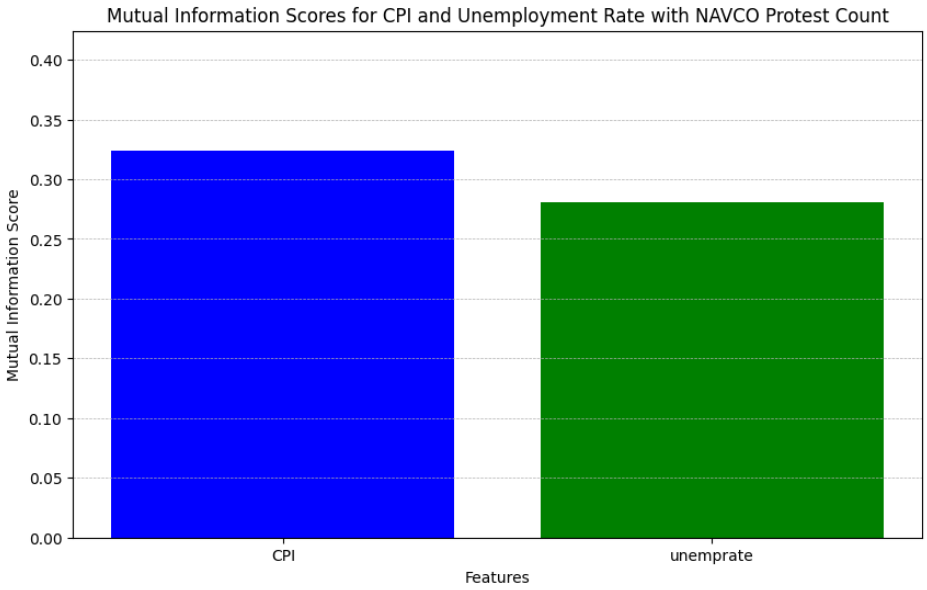

Figure 17: Mutual Information Scores: CPI and Unemployment

Spearman's rank correlation was also calculated to assess the strength and direction between protest

frequency and the same economic indicators. The results showed weak positive correlations, with a correlation coefficient of 0.1086 between NAVCO and CPI, and 0.0124 between NAVCO and unemployment. However, the p-values (0.4089 for CPI and 0.9253 for unemployment) indicate that these correlations are not statistically significant. This implies there is insufficient evidence to conclude that protest frequency is strongly associated with either CPI or unemployment.

3.5 Results: Random Forest Modeling Analysis

We also applied a Random Forest model to predict the likelihood of violent protest events based on key economic factors, including wealth inequality and inflation. This model allowed us to identify the most important variables contributing to violent protests and assess the predictive power of our dataset. The Random Forest model achieved an accuracy of 73%, with high precision in predicting violent events. The confusion matrix (Figure 18) shows that the model performed better at predicting violent events than non-violent ones, reflecting the imbalance in the dataset.

In Table 4, the model's precision for violent events is 0.79, with a recall of 0.81, indicating the model's strong performance in predicting violent protest events. However, the model struggled to accurately predict non-violent events, likely due to the skewed dataset favoring violent incidents.

Additionally, we evaluated the model's performance using the Receiver Operating Characteristic (ROC) curve (Figure 19). The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, providing a comprehensive measure of the model's ability to distinguish between classes. The area under the ROC curve (AUC) was approximately 0.83, indicating strong predictive performance. Notably, the curve begins to plateau at a TPR of 0.8 when the FPR is around 0.3, suggesting that beyond this point, increasing the threshold yields diminishing returns in improving the true positive rate without incurring a higher false positive rate. Feature importance analysis (Figure 20) revealed that CPI and wealth inequality ratios (90th/10th and 90th/50th) were the most significant predictors of violent protest events. This reinforces the connection between economic inequality and unrest, with higher levels of inequality contributing to an increased likelihood of violence during protests.

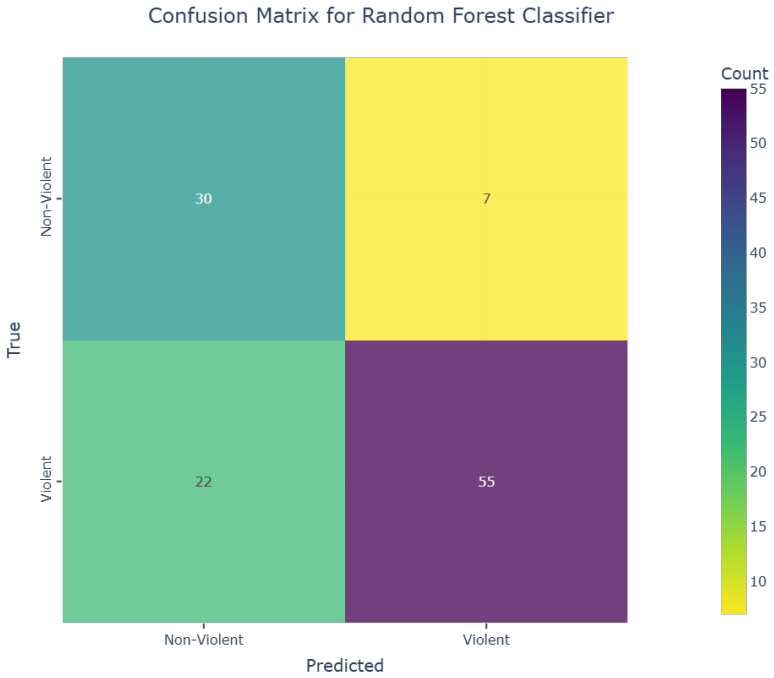

Figure 18: Confusion Matrix for Random Forest Classifier

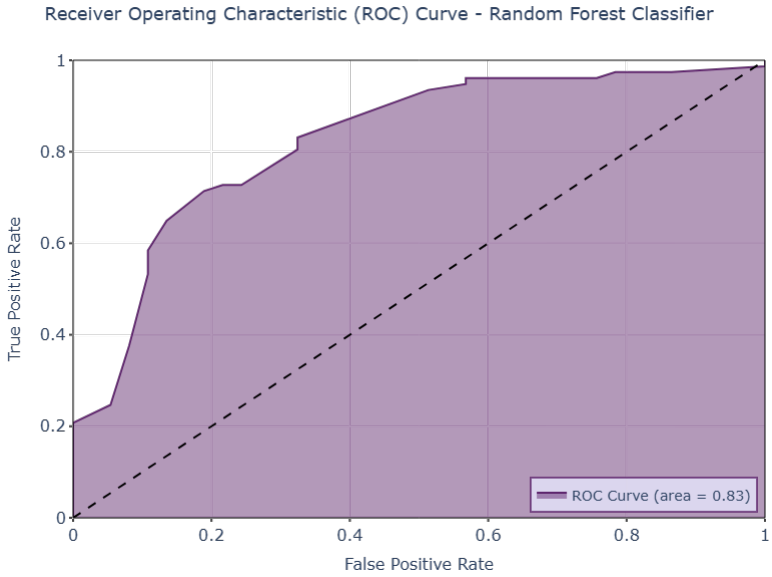

Figure 19: ROC Curve for Random Forest Classifier

Table 4: Random Forest Classification Report

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Violent	0.58	0.57	0.58	37
Violent	0.79	0.81	0.80	77

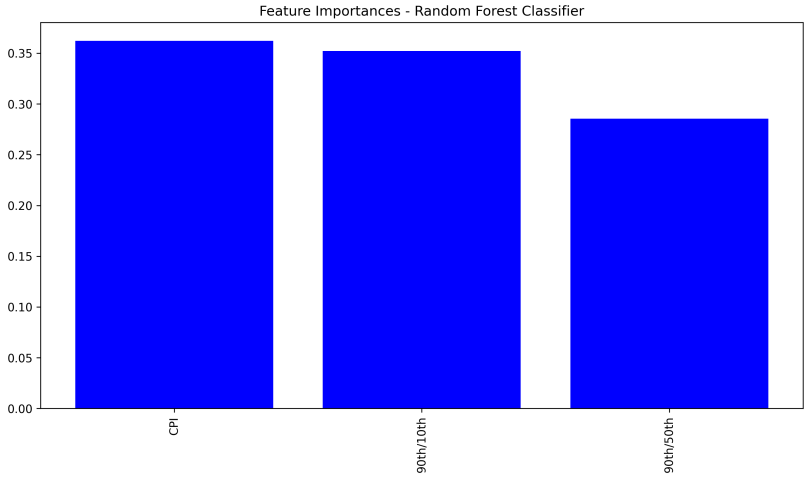

Figure 20: Feature Importance - Random Forest Classifier

4 Discussion

4.1 Summary of Findings

The analysis of wealth inequality and protest activity in the United States reveals significant insights into the relationship between economic disparities and social unrest. The GEOWEALTH-US dataset, in particular, demonstrates the growing concentration of wealth in the top 10% of earners, while the bottom 50% of the population has experienced a drastic decline in wealth share over time. This disparity correlates with an increase in protest activity, both in frequency and intensity. As wealth inequality deepened, the share of wealth held by the bottom 50% fell below 10% in most of the observations, a clear indicator of the widening economic gap. Our study also found a weak relationship between inflation, as measured by the Consumer Price Index (CPI), and protest activity. While higher CPI values seemed to have correlated with an increase in protest frequency, Mutual Information scores suggested a weak level of shared information. However, the results of the t-tests affirm that inflation is a critical factor contributing to civil unrest. Machine learning models, specifically Random Forest, were instrumental in predicting violent protest events. By incorporating economic factors such as CPI and wealth inequality ratios, these models achieved moderate accuracy, with wealth inequality emerging as a key predictor of violent unrest. However, the imbalance between violent and non-violent events in the dataset presented challenges, particularly in predicting non-violent protests, where the model's performance was less precise.

4.2 Interpretation of Results

4.2.1 Wealth Inequality and Social Unrest

The growing disparity in wealth distribution, particularly the increasing concentration of wealth in the top 10%, aligns with historical theories that link economic inequality to social unrest [9]. Our analysis highlights how wealth inequality is not only a persistent issue but one that has escalated over recent decades. The data shows that regions with higher concentrations of wealth in the hands of a few are also areas where protests are more frequent. This suggests that economic discontent, driven by the disparity between the wealthy and the lower-income population, is a significant driver of civil unrest. This finding supports existing literature that points to wealth inequality as a destabilizing factor in societies [18]. The visualizations of wealth distribution over time clearly depict how regions with higher median incomes, such as parts of California and New York, contrast sharply with the southeastern and rural areas, which exhibit much lower median wealth. The regional differences further illustrate that wealth inequality is not a uniform phenomenon but varies significantly across different parts of the country, likely contributing to varying levels of unrest.

4.2.2 Inflation and Protest Activity

The relationship between inflation and civil unrest has been well-documented, and much of our findings reinforce this connection. As inflation rises, particularly in economically distressed regions, it exacerbates the economic challenges faced by lower-income groups, leading to an increase in protest activity. The results of the t-tests on the NAVCO dataset further demonstrate that post-2010 violent events were significantly associated with higher CPI values, marking a shift in how economic pressures contribute to violence. This is indicative of the growing importance of inflation as a factor that incites unrest, especially in regions where economic hardship is more acute. The one-way ANOVA test comparing states with varying levels of protest activity (California, Georgia, and Idaho) provides additional evidence that inflation plays a critical role in fueling unrest. The significant differences in CPI changes between Idaho and the larger states suggest that inflationary pressures may have a more pronounced impact in states with smaller economies or less economic resilience, thus triggering more severe protests.

4.2.3 Predictive Power of Machine Learning Models

The application of machine learning models, particularly Random Forest, proved effective in predicting violent protest events. The model's high precision and recall in predicting violent events underscore the value of integrating economic factors, such as CPI and wealth inequality, into predictive models for civil unrest. The feature importance analysis highlights CPI and the wealth inequality ratios (90th/10th and 90th/50th) as the most significant predictors of violent protests. This reinforces the notion that economic inequality and inflation are not just correlated with unrest but are primary drivers of violent mobilizations.

However, the model's struggle to accurately predict non-violent events reveals a potential limitation. The skewed dataset, favoring violent incidents, likely influenced the model's performance. This imbalance suggests that while economic factors are critical in predicting violence, other social or political factors may be more influential in determining non-violent protest activity. Further research could explore the inclusion of additional variables, such as political polarization or social media activity, to improve the prediction of non-violent events.

4.3 Implications

The findings of this study have significant policy implications. Addressing wealth inequality is crucial to reducing tensions and preventing civil unrest. Policymakers should consider targeted measures such as progressive taxation, social welfare programs, and investments in education and job creation to mitigate economic disparities. These policies could help alleviate the economic frustrations leading to protests, especially in regions where the bottom 50% has experienced a sharp decline in wealth share. In addition, controlling inflation should be a priority for governments, especially in economically distressed regions. Stable inflation rates could help prevent the economic shocks that trigger unrest, as seen in the relationship between CPI and protest activity. Governments should focus on ensuring economic stability and protecting lower-income populations from the adverse effects of rising prices, which disproportionately affect those who are already economically vulnerable. The predictive power of machine learning models, such as Random Forest, offers valuable tools for monitoring and anticipating civil unrest. Governments and organizations could utilize these models to identify regions at high risk of protest activity, allowing for early intervention and the implementation of preventive measures. However, caution must be exercised to ensure that predictive models are used ethically and do not contribute to government overreach or the suppression of peaceful protests.

The use of machine learning models to predict civil unrest raises several ethical concerns. One of the primary risks is the potential breach of privacy when using sensitive data, such as protest records or social media activity. It is essential to ensure that these models comply with privacy regulations and that individuals' rights are protected. Moreover, there is a risk that predictive models could disproportionately target marginalized communities, leading to unfair treatment or increased surveillance in areas already facing economic or social challenges. Governments must be transparent about how these models are used and ensure that they do not contribute to the unjust suppression of peaceful protests or the over-policing of specific demographic groups. Addressing bias in machine learning models is also critical. The model's struggle to predict non-violent events suggests that there may be underlying biases in the data or the modeling process. Ensuring fairness and avoiding bias in predictive models is essential to prevent the perpetuation of existing inequalities or the unfair targeting of certain regions or communities.

4.4 Challenges and Limitations

One of the primary challenges encountered during this study was the issue of missing or incomplete data. Both the GEOWEALTH-US and CCC datasets contained gaps, which limited the scope of the analysis. For instance, the GEOWEALTH-US data was provided by decade, whereas the CCC data was organized monthly. This discrepancy in temporal granularity presented challenges in creating a cohesive dataset for analysis. Additionally, finding suitable conflict datasets that covered relevant time periods and contained comprehensive information on protest activity proved difficult. The NAVCO dataset, while useful, was limited in its coverage of non-violent events, further contributing to the imbalance in the dataset. These

limitations highlight the need for more comprehensive and standardized data collection on wealth inequality and civil unrest. Future research could benefit from more granular data that captures both violent and non-violent events in real time, as well as more detailed economic indicators at the local level.

5 Conclusions

Our study provided an in-depth analysis of the economic, political, and technological factors contributing to civil conflict, focusing on wealth inequality, resource dependence, and political instability. By integrating modern predictive tools, our research offers valuable insights into how these factors contribute to civil unrest. Our findings of a moderate relationship between economic indicators—particularly wealth concentration and poverty—and protest frequency align with previous studies that have identified economic disparity as a significant contributor to civil unrest [6, 18]. However, the weak linear correlations observed suggest that other variables, such as political polarization, may play a more significant role in the modern context. This echoes the work of researchers who emphasize the multifaceted nature of civil conflict causation [21]. Political instability, compounded by weak governance and amplified by social media, further escalates the risk of conflict. Our study contributes to the literature by providing contemporary empirical evidence from the United States, highlighting the need for integrated models that consider both economic and non-economic factors. By demonstrating the potential of predictive technologies to forecast and prevent civil conflict before it escalates, our research not only underscores the pressing need for reducing wealth inequality and strengthening governance structures but also illustrates how modern analytical tools can enhance conflict prediction. As the world continues to grapple with increasing polarization and economic disparity, these findings are crucial for policymakers, governments, and international organizations seeking to maintain peace and stability. By combining historical context with cutting-edge analytical methods, this study contributes to a more comprehensive understanding of civil conflict in the modern era and provides actionable solutions to mitigate its risks.

References

[1] S. Asur and B. A. Huberman. Predicting the future with social media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 1:492–499, 2010.

[2] T. W. Bank. Annual inflation by gdp deflator and consumer prices. *World Bank Open Data*, 2021.

[3] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *arXiv preprint arXiv:1010.3003*, 2010.

[4] C. Caragea, N. J. McNeese, A. R. Jaiswal, G. Traylor, H.-W. Kim, P. Mitra, D. Wu, A. H. Tapia, C. L. Giles, B. J. Jansen, et al. Classifying text messages for the haiti earthquake. In *ISCRAM*. Citeseer, 2011.

[5] M. Colaresi and Z. Mahmood. Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2):193–214, 2017.

[6] P. Collier and A. Hoeffler. On economic causes of civil war. *Oxford economic papers*, 50(4):563–573, 1998.

[7] C. C. Consortium. Crowd counting consortium dataset, 2024. Accessed September 22, 2024.

[8] Falola, Toyin, and A. Oyebade. *Hot Spot: Sub-Saharan Africa*. Greenwood, 2010.

[9] J. D. Fearon and D. D. Laitin. Ethnicity, insurgency, and civil war. *American political science review*, 97(1):75–90, 2003.

[10] I. C. for Data Integration. Ipums usa labor market area codes and commuting zones, 2024.

[11] A. C. for Democratic Governance and H. U. Innovation. Nonviolent and violent campaigns and outcomes data project (navco), 2021. The NAVCO dataset contains data on 627 mass mobilizations worldwide from 1900 to 2021, excluding maximalist campaigns.

[12] M. K. Gloria Guzman. Income in the united states: 2023, 2024.

[13] C. S. Hendrix. Measuring state capacity: Theoretical and empirical implications for the study of civil conflict. *Journal of peace research*, 47(3):273–285, 2010.

[14] W. Liu, B. Dai, A. Humayun, C. Tay, C. Yu, L. B. Smith, J. M. Rehg, and L. Song. Iterative machine teaching. In *International Conference on Machine Learning*, pages 2149–2158. PMLR, 2017.

[15] S. Mohsen. The role of new information and communication technologies and social media during the arab spring revolution. Master's Research Paper, University of Ottawa, 2010.

[16] U. B. of Labor Statistics. Consumer price index for all urban consumers (cpi-u), 2024. Accessed September 27, 2024.

[17] U. B. of Labor Statistics. (seas) unemployment rate, 2024. Accessed October 5, 2024.

[18] G. Østby. Polarization, horizontal inequalities and violent civil conflict. *Journal of Peace Research*, 45(2):143–162, 2008.

[19] J. Piombo. *Resources and Conflict in Sub-Saharan Africa*, page 17. Routledge, 1st edition, 2013.

[20] J. Suss, T. Kemeny, and D. S. Connor. Geowealth-us: Spatial wealth inequality data for the united states, 1960–2020. *Scientific Data*, 11(1):253, 2024.

[21] M. D. Ward, B. D. Greenhill, and K. M. Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of peace research*, 47(4):363–375, 2010.

[22] M. D. Ward, N. W. Metternich, C. L. Dorff, M. Gallop, F. M. Hollenbach, A. Schultz, and S. Weschle. Learning from the past and stepping into the future: Toward a new generation of conflict prediction. *International Studies Review*, 15(4):473–490, 2013.

APPENDIX A

Project code: https://github.com/AdamHumble/AsuDat490Fa24aCapstoneProject