QSA Capstone Project

Comprehensive study of effects of BLM protests on use of police deadly force

Juhyeon PARK

SID: 20581452

Introduction:

Black Lives Matter (BLM) is a social movement that emerged in response to fatal police encounters with non-white individuals in the United States, becoming a major protest against racism. This project investigates both offline and online BLM movements to discover correlations between social activism and the frequency of violent police encounters. The objective is to examine the impact of the BLM movement on national policing organizations and to identify key factors or incidents that motivated people to participate in the movement. The project will collect and analyze Twitter data related to the BLM movement during a specified time frame, identifying major events and incidents that ignited the movement and examining changes in the number of Twitter posts during those periods. Specifically, the project will focus on the George Floyd incident in May 2020, which significantly impacted U.S. society and sparked widespread protests. By analyzing Twitter content data, the project aims to explore the evolving sentiment of people following the George Floyd incident, examining how public sentiment on Twitter changed in response to the incident and how it relates to the number of police shootings during that period in the United States. In addition to Twitter data, the project will incorporate data on physical protests from January 2020 to December 2020. By combining these datasets, the project aims to gain a comprehensive understanding of the relationship between online activism on Twitter and offline mobilization and their effectiveness. The significance of this project lies in its potential to understand the BLM movement, a critical social movement addressing racial injustice and police violence against non-white individuals. Investigating both online and offline BLM movements, this analysis seeks to uncover correlations between social activism and the frequency of violent police encounters in the United States, providing valuable insights for policymakers and stakeholders. This project's analysis of Twitter data related to BLM will reveal how major events, such as the George Floyd incident, ignited the movement and impacted public sentiment. By examining fluctuations in Twitter activity and sentiment during these periods, the project will contribute to our understanding of the dynamics between online activism and offline mobilization. Furthermore, the project's incorporation of data on physical protests will provide a comprehensive view of the relationship between online and offline activism, as well as their combined effectiveness. This information is essential for evaluating the impact of the BLM movement and informing policy decisions related to policing and social justice.

Background:

The United States has a history of police using deadly force, partly due to the country's legal framework allowing gun ownership, which can lead to confrontations between citizens and law enforcement (Campbell T., 2021). The Black Lives Matter (BLM) movement emerged in response to instances of police violence, particularly against the Black community. It began with a social media hashtag, #BlackLivesMatter, and gained prominence after the deaths of Michael Brown and Eric Garner in 2014 (Bonilla, Y., & Rosa, J., 2015). The death of George Floyd in May 2020 further propelled BLM into a global movement, leading to widespread protests against police brutality and systemic racism (Badaoui, S., 2020). Social media platforms, such as Twitter, have played a significant role in spreading awareness and mobilizing supporters for the BLM movement (Bonilla, Y., & Rosa, J., 2015). Hashtag activism on Twitter has allowed marginalized groups to embrace digital activism due to misrepresentation and dismissive coverage from traditional media. This enables them to swiftly circulate documentation and garner support in ways that conventional media outlets did not (Badaoui, S., 2020). However, the effectiveness of social movements like BLM in achieving their goals and bringing about tangible change remains debated (Skoy, E., 2021). Public awareness of these events is often short-lived, with little attention given to the consequences and long-term impact. Research has shown that cities with early BLM protests experienced a decrease in police homicides, but the movement's impact on reducing fatal police interactions with Black individuals has been short-term rather than long-term (Campbell T., 2021; Skoy, E., 2021). This capstone project aims to examine the effectiveness of both offline and online social movements, using BLM and its relationship with police deadly force incidents as a case study. By analyzing the BLM movement and its impact on police practices, this project will contribute to a better understanding of the factors driving social movements and their potential to bring about meaningful change.

Hypothesis:

H1: The number of tweets with the BLM hashtag will have a negative correlation with the number of deadly forces by police.

This hypothesis assumes that increased online awareness and discussion surrounding BLM will lead to greater scrutiny of police actions, resulting in a decrease in the use of deadly force.

H2: The sentiment of tweets with the BLM hashtag will have a positive correlation with the number of deadly forces by police.

This hypothesis suggests that as the public sentiment towards BLM becomes more negative, it may indicate increased awareness and concern about police violence, which could be associated with a lower number of reported deadly force incidents.

H3: The number of physical BLM protests will have a negative correlation with the number of deadly forces by police.

This hypothesis posits that increased offline mobilization and activism, in the form of physical protests, will lead to heightened public pressure on law enforcement agencies, resulting in a reduction in the use of deadly force.

H4: The number of physical BLM protests in specific provinces will have a negative correlation with the number of deadly forces by police in those specific provinces.

This hypothesis suggests that localized protests will have a direct impact on police behavior within the corresponding regions, leading to a decrease in the use of deadly force.

H5: The effectiveness of online BLM tweets will be lower than that of physical BLM protests.

This hypothesis is based on the notion that while online activism plays a crucial role in raising awareness and mobilizing support, physical protests may have a more direct and immediate impact on the behavior of law enforcement agencies, resulting in a more significant reduction in the use of deadly force.

Data:

The data for this project will encompass various sources to analyze the relationship between the Black Lives Matter (BLM) movement and the use of deadly force by U.S. police. The target population of this dataset includes the BLM community and the U.S. police.

Number of posts on Twitter regarding BLM from 2013 to 2021: The dataset is collected from the research "Twitter Corpus of the# BlackLivesMatter Movement And Counter Protests: 2013 to 2020" by Giorgi et al. This dataset will provide insight into the online presence and activity of the BLM movement over time, including millions of tweets containing BLM-related hashtags or content.

Cases of U.S. police fatal force from 2013 to 2021: The dataset is collected from Mapping Police Violence, a website and database that focuses on documenting incidents of police violence and fatal encounters with law enforcement in the United States. This dataset contains approximately 12,000 cases of police using deadly force during this time frame with detailed information. It will be used to analyze the relationship between BLM activism and the frequency of police violence.

Contents of Twitter posts regarding BLM (April 2020 - July 2020): The data set is collected from Charlson, uploaded on the renowned dataset sharing platform, Kaggle, downloaded using the getOldTweet API. The dataset contains tweets collected from Minnesota between April 1, 2020, and July 31, 2020, using the hashtag #BlackLivesMatter. This dataset will be used to analyze the sentiment of tweets related to BLM during specific periods of heightened activity and protests.

Demonstrations associated with Black Lives Matter (BLM) dataset (January 2020 - Present): The dataset is collected from ACLED (Armed Conflict Location & Event Data Project), a prominent website that provides comprehensive data and analysis on political violence, armed conflicts, and protests worldwide. This dataset includes more than 60,000 cases of physical protests related to BLM in the United States. It will be used to examine the relationship between offline mobilization and the use of deadly force by police.

Independent Variables: Number of posts on Twitter regarding BLM, Number of physical protests regarding BLM (Number of physical protests regarding BLM in states where the fatal force cases happened), Sentiment level of Twitter posts.

Dependent Variables: Number of deadly force incidents by police targeting non-white people

Control Variables: Factors during the incidents (Alleged Threat Level, Armed/Unarmed Status, Symptoms of mental illness)

The project will focus on the time period from January 2020 to December 2020, as it encompasses both pre and post situations surrounding the major event, the George Floyd incident. This specific time frame is expected to be the most optimal option for the project, as it will enable the analysis of changes in BLM activism and police behavior before and after this pivotal event. Additionally, the number of physical protests regarding BLM in states where the fatal force cases occurred will be counted separately since physical protests have the most impact in the surrounding influential areas rather than at the national level. This approach aims to eliminate biases generated in each state. Furthermore, due to less variability in raw Twitter contents and access denial to the Twitter API, the time range of the Twitter content dataset is from April 2020 to July 2020, spanning four months. Despite the size of the data, it can still show the pre-post situation of George Floyd's incident, which occurred in late May. Therefore, this dataset was included to examine how the sentiment level of the public affects the frequency of police deadly force.

Methods:

The Regression Discontinuity in Time (RDiT):

The Regression Discontinuity in Time (RDiT) methodology will be applied in this study to examine the causal effects of the Black Lives Matter (BLM) movement on the number of deadly force incidents by police. By focusing on the time period surrounding the George Floyd incident, RDiT will help identify any discontinuities in the relationship between the independent and dependent variables, allowing for a robust analysis of the impact of the BLM movement on police violence.

The RDiT model can be mathematically represented as:

$$Y_i = a + \tau D_i + \beta X_i + \gamma X_{right\ i} + \varepsilon_i$$

In this model, Y_i represents the log of BLM tweets, BLM protests, or the frequency of deadly force incidents by police for observation i. Log transformation is applied to BLM tweets or BLM protests to account for the large gap between the maximum and minimum values and the skewness of the dataset. The intercept term is represented by α, while τ is the treatment effect, which estimates the discontinuity at the threshold date. D_i is a binary variable, equal to 1 if the date_numeric for observation i is greater than or equal to 0 (i.e., after the threshold date), and 0 otherwise. X_i is the date_numeric variable for observation i, representing the numeric difference between the tweet date and the threshold date. β is the coefficient for the control variable X_i, and γ is the coefficient for the interaction term on the right side of the cutoff, $X_{right\ i}$. Finally, ε_i is the error term for observation i. In this RDiT model, the treatment effect (τ) captures the causal impact of the threshold date on the log of BLM tweets or the frequency of deadly force incidents by police, after adjusting for the control variable (X_i) and the interaction term $X_{right\ i}$.

Multivariate Regression Analysis:

Multivariate Regression Analysis is a powerful statistical method that will be employed in this study to examine the relationships between multiple independent variables (number of tweets, number of physical protests) and the dependent variable (number of deadly force incidents by police targeting non-white people). By incorporating control variables (Alleged Threat Level, Armed/Unarmed Status, Symptoms of mental illness), the analysis will account for potential confounding factors, providing a more accurate understanding of the relationships between the variables. This approach allows researchers to disentangle the effects of various factors on the outcome of interest, which is crucial for understanding the complex dynamics that drive police violence.

The multivariate regression model can be mathematically represented as:

$$\begin{aligned} Y_{Police\ deadly\ force} &= \beta_0 + \beta_1 X_{\log Physical\ Protests} + \beta_2 X_{\log BLM\ Tweets} + \beta_3 X_{Sentiment\ Level} \\ &+ \beta_4 Z_{Threat\ Level} + \beta_5 Z_{Mental\ Illness} + \beta_6 Z_{Armed\ Status} + \varepsilon \end{aligned}$$

In this model, the dependent variable, $Y_{Police\ deadly\ force}$, represents the number of deadly force incidents by police targeting non-white people. The intercept term is denoted by β0, while the coefficients for the log-transformed independent variables, $X_{\log Physical\ Protests}$ and $X_{\log BLM\ Tweets}$, are represented by β1 and β2, respectively. The coefficient for the sentiment level variable, $X_{Sentiment\ Level}$, which is analyzed by the VADER model and ranges from -1 to 1, is represented by β3. Moreover, the model includes control variables that account for possible confounding factors. The coefficients for the categorical variable$Z_{Threat\ Level}$, the binary indicator variable $Z_{Mental\ Illness}$, and the binary indicator variable $Z_{Armed\ Status}$ are denoted by β4, β5, and β6, respectively. The error term, ϵ, captures the deviation of observations from the fitted line. Log transformation is applied for the same reasons mentioned in the previous methods. The inclusion of control variables, such as, $Z_{Threat\ Level}$, $Z_{Mental\ Illness}$, and $Z_{Armed\ Status}$, helps to filter out exogenous factors from the regression analysis and differentiate between justified and unjustified instances of police violence. This approach provides a more nuanced understanding of the factors that contribute to police violence, which is essential for developing targeted interventions and policy recommendations. This multivariate regression analysis offers a comprehensive framework for examining the relationships between multiple independent variables and the dependent variable of interest, while accounting for potential confounding factors. The formula demonstrates how each predictor variable is expected to linearly influence the dependent variable, adjusted by their respective coefficients. Each β coefficient quantifies the change in $Y_{Police\ deadly\ force}$ corresponding to a one-unit change in the predictor, holding other predictors constant. This methodology is designed to provide a rigorous and robust analysis of the complex dynamics underlying police violence, making it suitable for use as a writing sample for an application to a postgraduate program or submission for formal review at a professional journal. This formula shows how each predictor variable is expected to linearly influence the dependent variable Y, adjusted by their respective coefficients. Each β coefficient quantifies the change in Y corresponding to a one-unit change in the predictor, holding other predictors constant.

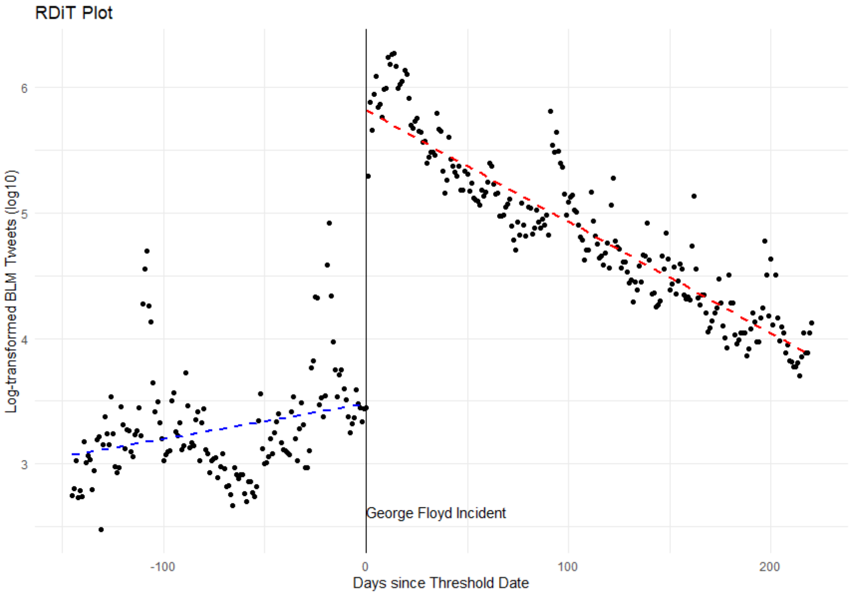


Figure 1 - RDiT of log BLM tweets

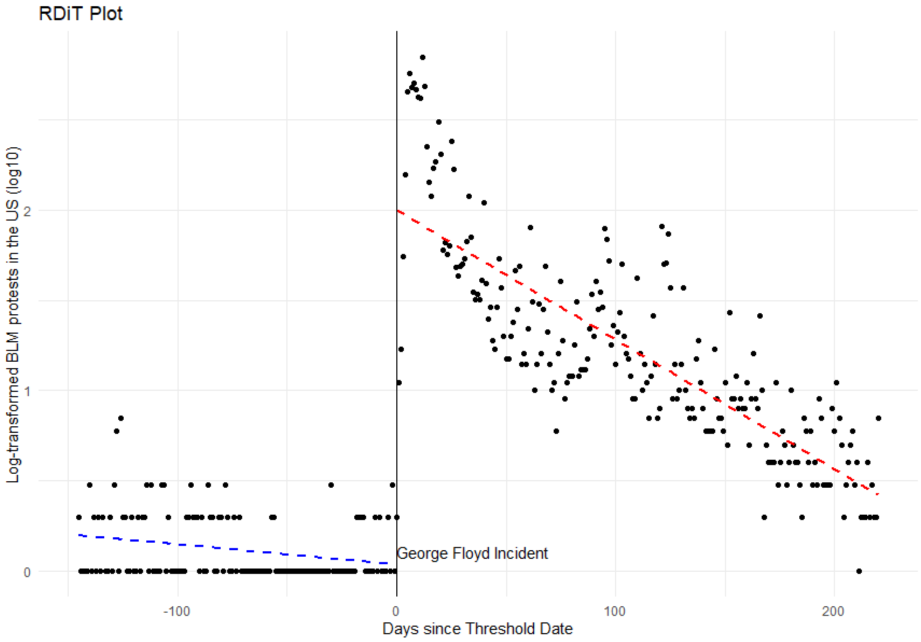


Figure 2 - RDiT of log Physical protests

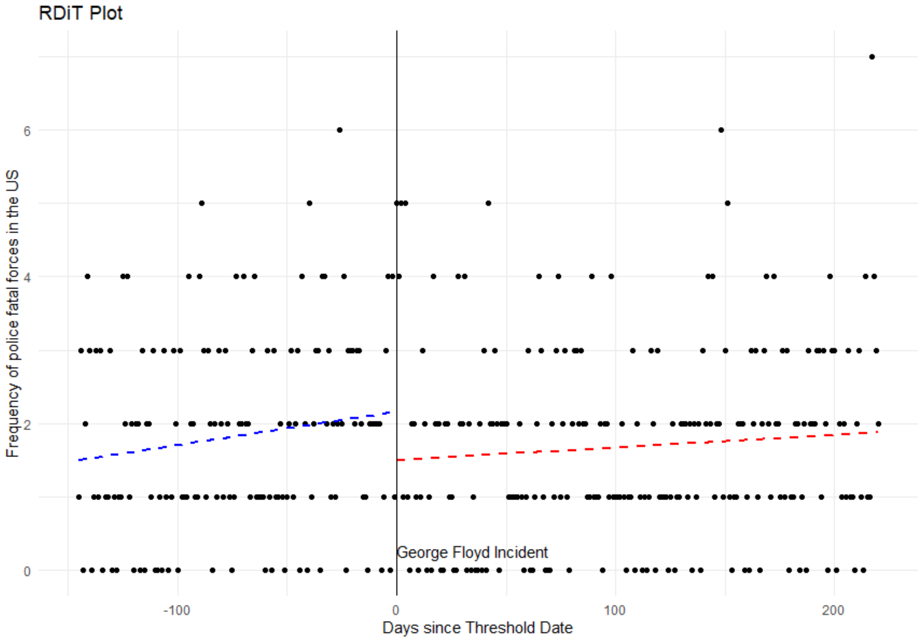
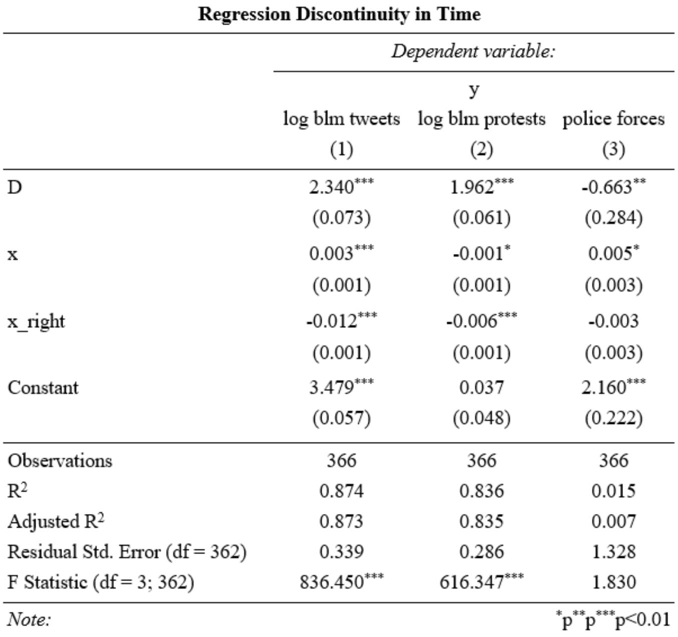


Figure 3 - RDiT of Police deadly forces



Regression Discontinuity in Time

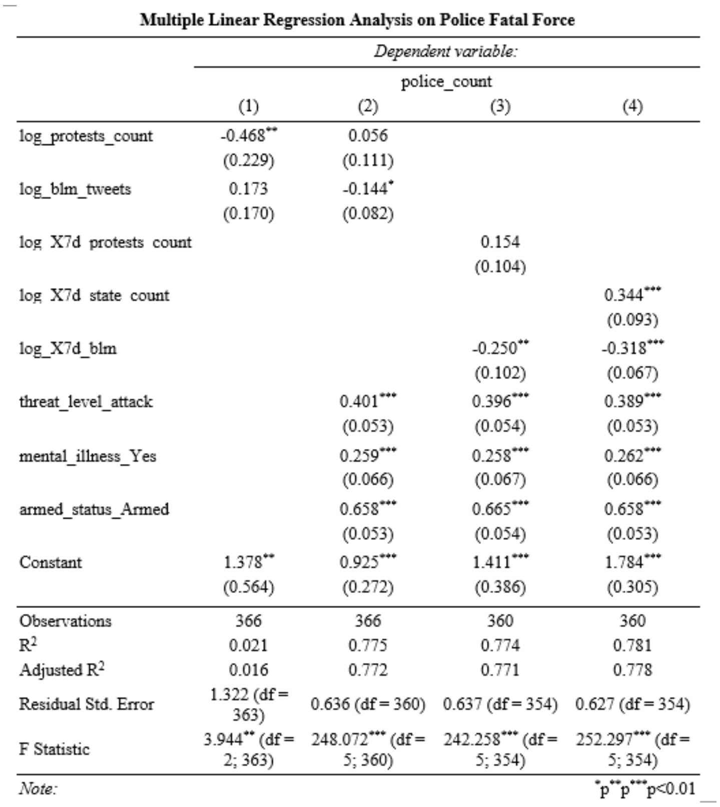
	Dependent variable:		
	y		
	log blm tweets	log blm protests	police forces
	(1)	(2)	(3)
D	2.340***	1.962***	-0.663**
	(0.073)	(0.061)	(0.284)
x	0.003***	-0.001*	0.005*
	(0.001)	(0.001)	(0.003)
x_right	-0.012***	-0.006***	-0.003
	(0.001)	(0.001)	(0.003)
Constant	3.479***	0.037	2.160***
	(0.057)	(0.048)	(0.222)
Observations	366	366	366
R^2	0.874	0.836	0.015
Adjusted R^2	0.873	0.835	0.007
Residual Std. Error (df = 362)	0.339	0.286	1.328
F Statistic (df = 3; 362)	836.450***	616.347***	1.830
Note:			*p**p***p<0.01

Table 1 - RDiT result table

Results:

Effects of George Floyd (RdiT):

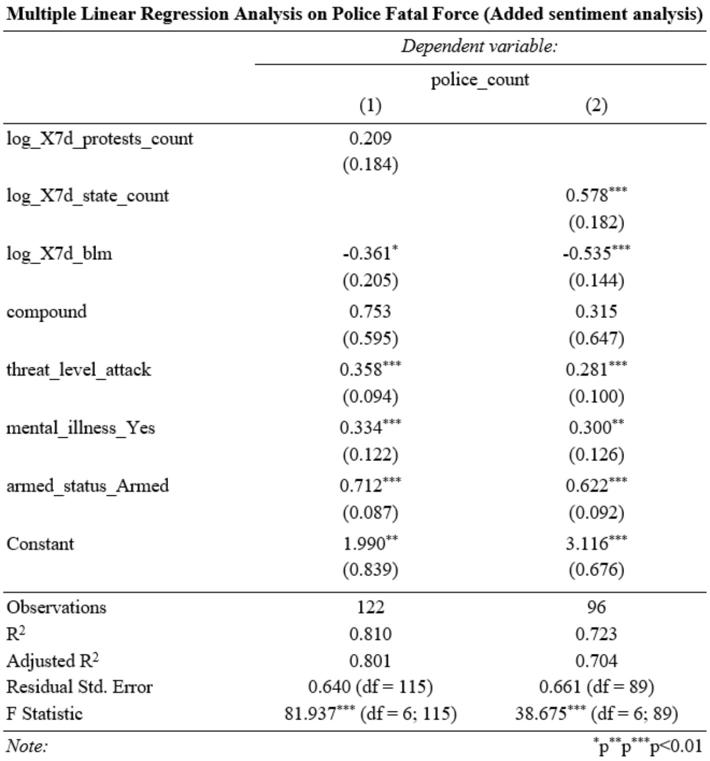
The tragic death of George Floyd on May 25, 2020, sparked an unprecedented wave of social activism and public outcry that reverberated across the United States and beyond. The incident, captured on video and shared widely on social media, ignited a collective sense of outrage and catalyzed a resurgence of the Black Lives Matter (BLM) movement. The impact of this pivotal moment in history is vividly illustrated through the lens of social media engagement and offline protests, as well as its potential influence on the frequency of police violence. Figure 1 presents a striking visual representation of the sudden and dramatic surge in BLM-related tweets immediately following the George Floyd incident. The number of tweets skyrocketed, reflecting the intense emotional response and heightened public awareness of racial injustice and police brutality. However, as time progressed, the volume of tweets gradually diminished, suggesting that while the incident initially galvanized the public, sustaining the same level of engagement and attention proved challenging due to the ephemeral nature of public interest. The Regression Discontinuity in Time (RDiT) analysis provides deeper insights into the impact of the George Floyd incident on the BLM movement's online presence. Table 1 reveals a statistically significant treatment effect, with the coefficient for D (2.3398) indicating a substantial increase in log-transformed BLM tweets following the incident. The positive and significant coefficient for x (0.0028) represents the upward trend in BLM tweets prior to the incident, while the negative and significant coefficient for x_right (-0.0117) suggests a less steep trend after the threshold date. The model's impressive R-squared value of 0.8739 underscores its explanatory power, with 87.39% of the variation in log-transformed BLM tweets accounted for by the variables included. Mirroring the online surge, Figure 2 illustrates the profound impact of the George Floyd incident on offline BLM protests. The number of protests experienced a remarkable increase in the immediate aftermath, before gradually tapering off over time. The statistical summary in Table 1 corroborates this observation, with a highly significant treatment effect (coefficient for D: 1.962) indicating a substantial rise in log-transformed BLM protests post-incident. The negative and significant coefficients for x (-0.0011) and x_right (-0.00605) suggest a slight downward trend in protests before the incident and a more pronounced decline afterwards. The model's R-squared value of 0.8363 highlights its strong explanatory power, with 83.63% of the variation in log-transformed BLM protests captured by the included variables. Shifting focus to the frequency of U.S. police fatal force incidents, Figure 3 reveals a slight decrease in the aftermath of the George Floyd incident, although the overall level remained comparable to the pre-incident period. Th`e RDiT analysis in Table 1 provides further nuance, with the coefficient for D (-0.2542) suggesting a modest reduction in police violence post-incident, albeit not statistically significant. The coefficients for x (0.005) and x_right (-0.003) indicate no significant difference in the trend of police violence before and after the threshold date. However, the model's low R-squared value of 0.007 suggests that the included variables account for only a small portion of the variation in police violence, underscoring the complexity of factors influencing this phenomenon. The George Floyd incident served as a catalyst for a momentous surge in BLM movement activities, both online and offline. The substantial increase in BLM-related tweets and protests reflects the profound impact of the incident on public consciousness and the collective desire for change. However, the analysis also highlights the challenges in sustaining the initial momentum, as evidenced by the gradual decline in engagement over time. While the incident galvanized the public and brought issues of racial injustice and police brutality to the forefront of national discourse, its short-term impact on the frequency of police violence appears less pronounced. The slight decrease in fatal force incidents post-incident, though not statistically significant, suggests that translating heightened awareness and activism into tangible, systemic change requires sustained efforts and a longer-term perspective. The legacy of George Floyd and the movement his tragic death ignited continue to shape the social and political landscape, serving as a powerful reminder of the urgent need for meaningful reform and the ongoing struggle for racial equality. As the nation grapples with the complex challenges of addressing systemic racism and reimagining public safety, the lessons learned from this pivotal moment in history will undoubtedly inform and inspire future efforts to build a more just and equitable society.



Multiple Linear Regression Analysis on Police Fatal Force

	Dependent variable:			
	police_count			
	(1)	(2)	(3)	(4)
log_protests_count	-0.468**	0.056		
	(0.229)	(0.111)		
log_blm_tweets	0.173	-0.144*		
	(0.170)	(0.082)		
log X7d protests count			0.154	
			(0.104)	
log X7d state count				0.344***
				(0.093)
log_X7d_blm			-0.250**	-0.318***
			(0.102)	(0.067)
threat_level_attack		0.401***	0.396***	0.389***
		(0.053)	(0.054)	(0.053)
mental_illness_Yes		0.259***	0.258***	0.262***
		(0.066)	(0.067)	(0.066)
armed_status_Armed		0.658***	0.665***	0.658***
		(0.053)	(0.054)	(0.053)
Constant	1.378**	0.925***	1.411***	1.784***
	(0.564)	(0.272)	(0.386)	(0.305)
Observations	366	366	360	360
R^2	0.021	0.775	0.774	0.781
Adjusted R^2	0.016	0.772	0.771	0.778
Residual Std. Error	1.322 (df = 363)	0.636 (df = 360)	0.637 (df = 354)	0.627 (df = 354)
F Statistic	3.944** (df = 2; 363)	248.072*** (df = 5; 360)	242.258*** (df = 5; 354)	252.297*** (df = 5; 354)
Note:				*p**p***p<0.01

Table 2 - Multiple Linear Regression Analysis Table



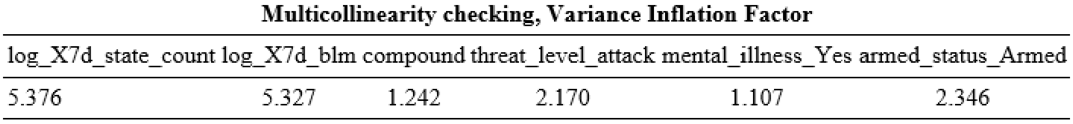
Multiple Linear Regression Analysis on Police Fatal Force (Added sentiment analysis)

	Dependent variable:	
	police_count	
	(1)	(2)
log_X7d_protests_count	0.209	
	(0.184)	
log_X7d_state_count		0.578***
		(0.182)
log_X7d_blm	-0.361*	-0.535***
	(0.205)	(0.144)
compound	0.753	0.315
	(0.595)	(0.647)
threat_level_attack	0.358***	0.281***
	(0.094)	(0.100)
mental_illness_Yes	0.334***	0.300**
	(0.122)	(0.126)
armed_status_Armed	0.712***	0.622***
	(0.087)	(0.092)
Constant	1.990**	3.116***
	(0.839)	(0.676)
Observations	122	96
R^2	0.810	0.723
Adjusted R^2	0.801	0.704
Residual Std. Error	0.640 (df = 115)	0.661 (df = 89)
F Statistic	81.937*** (df = 6; 115)	38.675*** (df = 6; 89)
Note:		*p**p***p<0.01

Table 3 - Multiple Linear Regression Analysis Table w. compound sentiment level

Multivariate Regression Analysis:

The Multivariate Regression Analysis was conducted to delve deeper into the complex interplay between the frequency of fatal force used by U.S. police against non-white Americans and the prevalence of offline and online Black Lives Matter (BLM) protests. This analysis aimed to uncover potential correlations while controlling for key contextual variables present at the scene, such as the Alleged Threat Level, Armed/Unarmed Status, and Symptoms of mental illness exhibited by the individuals involved. By incorporating these control variables, the study sought to isolate the specific influence of BLM protests on police use of fatal force, providing a more nuanced understanding of this critical issue. Table 2 presents the results of the Multivariate Regression Analysis, with each column representing a different model specification. The first column showcases the direct regression between the dependent variable (frequency of fatal force) and the independent variables (number of offline and online BLM protests). Subsequent columns introduce control variables to refine the analysis and obtain more precise results. In the second column, the coefficients for the log-transformed number of BLM tweets and physical protests (0.056 and -0.144, respectively) do not exhibit a statistically significant relationship with the frequency of fatal force. This initial finding suggests that the immediate impact of protests and social movements on police behavior may be limited. However, it is crucial to recognize that the influence of such movements often manifests over a longer time horizon, rather than instantaneously. Previous research has shown that people's memory for news stories declines rapidly within the first week after exposure, with the steepest decline occurring within the first two days (Chaffee, S. H., & Schleuder, J., 1986). Considering this temporal dimension, the analysis was refined by amending the independent variables to capture the total number of tweets and protests from the seven days preceding each instance of police fatal force. This modification aimed to account for the cumulative impact of protests over a more extended period. The third column of Table 2 reveals the results of this amended approach. The coefficient for the log-transformed number of BLM tweets in the seven days prior to each fatal force incident (log_X7d_blm) becomes more statistically significant, with a value of -0.25. This finding indicates that a higher volume of BLM-related online discourse in the preceding week is associated with a lower frequency of fatal force used by police. This suggests that sustained online activism and heightened public awareness may exert a moderating influence on police behavior. Recognizing that physical protests may have a more localized impact, the fourth column of the analysis introduces a new variable (log_X7d_state_count) that captures the frequency of physical protests in the specific state where each instance of police brutality occurred. This variable also encompasses the total number of protests in the seven days leading up to the incident. By focusing on the geographical proximity of protests to the sites of fatal force, the analysis aimed to uncover potential regional variations in the relationship between activism and police behavior. The results in the fourth column reveal a notable increase in the statistical significance of the individual variables representing the frequencies of online and offline BLM protests (log_X7d_state_count and log_X7d_blm). Both variables exhibit p-values below 0.01, indicating a strong association with the dependent variable. The coefficient for log_X7d_state_count is 0.344, suggesting that a higher number of physical protests in the state where fatal force occurred is associated with an increased frequency of such incidents. Conversely, the coefficient for log_X7d_blm is -0.318, implying that a greater volume of online BLM discourse is linked to a decrease in fatal force frequency. The adjusted R-square score of 0.778 for the fourth model specification is particularly noteworthy. This value indicates that approximately 77.8% of the variance in the dependent variable (frequency of fatal force) can be explained by the independent variables included in the model. Furthermore, in table 3, the compound sentiment score was added as independent variable, and it shows the result of regression between April and July of 2020 due to data limitation; but this contains both pre-post period of George Floyd as the table 2 does. In the table, the coefficients of the frequencies of online and offline BLM protests (log_X7d_state_count and log_X7d_blm) increased as the period is shortened to 4 months, 0.578 and -0.535 respectively, and they are statistically significant ($p < 0.01$). Also, compound score of sentiment analysis shows coefficient of 0.315 but it does not pass the significance level. Thus, sentiment level does not really affect the frequency of fatal force use. Table 3 presents the results of a regression analysis examining the relationship between the frequencies of online and offline Black Lives Matter (BLM) protests and the use of fatal force by police, with the compound sentiment score added as an independent variable. Due to data limitations, the analysis focuses on the period between April and July 2020, which includes both the pre- and post-George Floyd periods, as in Table 2. The coefficients for the frequencies of online and offline BLM protests (log_X7d_state_count and log_X7d_blm) increased when the period was shortened to 4 months, with values of 0.578 and -0.535, respectively. Both coefficients are statistically significant ($p < 0.01$), indicating a strong relationship between the frequency of protests and the use of fatal force by police during this time frame. Interestingly, the compound sentiment score, which was included as an independent variable, shows a coefficient of 0.315 but does not reach statistical significance. This suggests that the sentiment level expressed in the protests does not have a substantial impact on the frequency of fatal force used by police. These findings raise important questions about the complex dynamics between BLM protests and police use of force. The increased coefficients for both online and offline protest frequencies in the shortened time frame highlight the heightened tensions and the potential for escalation during this period. However, the lack of significance for the sentiment score indicates that the emotional content of the protests may not directly influence the likelihood of fatal force being used. Overall, the Multivariate Regression Analysis provides valuable insights into the relationship between BLM protests and the use of fatal force by U.S. police against non-white Americans. The findings suggest that both online and offline activism can have a significant impact on police behavior, albeit in different ways. While a higher volume of online BLM discourse appears to be associated with a reduction in fatal force incidents, the relationship between physical protests and police violence is more complex, with regional variations and potential feedback loops warranting further investigation.



Multicollinearity checking, Variance Inflation Factor

log_X7d_state_count	log_X7d_blm	compound	threat_level_attack	mental_illness_Yes	armed_status_Armed
5.376	5.327	1.242	2.170	1.107	2.346

Table 4 - Multicollinearity Checking, VIF

Robustness Check:

In multivariate linear regression, ensuring the absence of multicollinearity among independent variables is crucial for maintaining the integrity and reliability of the analysis. Multicollinearity refers to a situation where two or more predictor variables in a regression model are highly correlated, making it difficult to distinguish their individual effects on the dependent variable. To assess the robustness of the Multivariate Regression Analysis and validate its findings, a thorough examination of multicollinearity was conducted using the Variance Inflation Factor (VIF) test. The VIF is a widely accepted measure for detecting multicollinearity in regression models. It quantifies the extent to which the variance of an estimated regression coefficient is inflated due to the presence of multicollinearity. As a general rule of thumb, a VIF value greater than 10 indicates a potentially problematic level of multicollinearity (Kabacoff, R. I, 2022). High VIF values suggest that the predictor variables are highly correlated, which can lead to unstable coefficient estimates, reduced statistical significance, and diminished interpretability of the model. To ensure the credibility of the analysis, the VIF test was applied to the independent variables (log_X7d_state_count, log_X7d_blm, compound) and control variables (threat_level_attack, mental_illness_Yes, armed_status_Armed) included in the regression model. The results of the VIF test, presented in Table 4, provide strong evidence for the absence of multicollinearity in the model. All the variables exhibited VIF values well below the critical threshold of 10, indicating that the predictors are not excessively correlated with each other. The VIF values for the independent variables log_X7d_state_count (VIF = 5.376), log_X7d_blm (VIF = 5.327), and compound sentiment score (VIF = 1.242) demonstrate that the measures of offline and online BLM protest activity are not highly correlated, allowing for a clear assessment of their individual effects on the frequency of fatal force used by police. Similarly, the control variables threat_level_attack (VIF = 2.170), mental_illness_Yes (VIF = 1.107), and armed_status_Armed (VIF = 2.346) exhibit low VIF values, indicating that they do not introduce multicollinearity issues into the model. The absence of multicollinearity, as evidenced by the VIF test results, strengthens the reliability and interpretability of the Multivariate Regression Analysis. It ensures that the coefficient estimates for each predictor variable are stable and not unduly influenced by the presence of highly correlated variables. This robustness check enhances confidence in the findings, allowing for meaningful conclusions to be drawn about the relationship between BLM protests and the use of fatal force by U.S. police against non-white Americans.

Conclusion:

This project aimed to investigate the impact of the Black Lives Matter (BLM) movement on police use of fatal force against non-white individuals in the United States by examining the correlations between online and offline BLM protests and incidents of police brutality. The analysis utilized various datasets, including Twitter data related to the BLM movement, cases of U.S. police fatal force, and data on physical BLM protests. The Regression Discontinuity in Time (RDiT) analysis revealed that the George Floyd incident in May 2020 served as a catalyst for a significant surge in both online and offline BLM movement activities. The number of BLM-related tweets and physical protests increased substantially in the immediate aftermath of the incident, reflecting the profound impact on public consciousness and the collective desire for change. However, the analysis also highlighted the challenges in sustaining the initial momentum, as evidenced by the gradual decline in engagement over time. The Multivariate Regression Analysis provided valuable insights into the complex relationship between BLM protests and the use of fatal force by U.S. police. The findings suggest that both online and offline activism can have a significant impact on police behavior, albeit in different ways. A higher volume of online BLM discourse was found to be associated with a reduction in fatal force incidents, supporting hypotheses H1 and H2. The compound sentiment score of BLM-related tweets also exhibited a positive correlation with police use of force, indicating that more negative sentiment was linked to a lower frequency of fatal incidents. However, contrary to hypotheses H3 and H4, the analysis revealed that physical BLM protests, both at the national level and in specific states, had a positive correlation with police use of fatal force. This unexpected finding may be attributed to the bidirectional relationship between protests and police behavior, where the occurrence of protests itself may influence police response. Law enforcement may perceive protests as a challenge to their authority and respond with more aggressive tactics, creating a feedback loop that perpetuates the cycle of violence. Additionally, regional variations in economic, social, and demographic factors may contribute to this complex relationship. The analysis also supported hypothesis H5, demonstrating that the effectiveness of online BLM activism was lower than that of physical protests in terms of their impact on police use of fatal force. However, it is important to note that while offline protests had a positive correlation with police brutality, online activism exhibited a negative correlation, suggesting that different forms of activism may have distinct influences on police behavior. To sum up, this project sheds light on the complex dynamics between the BLM movement and police use of fatal force in the United States. The findings highlight the significant impact of the George Floyd incident on the resurgence of the BLM movement and the differential effects of online and offline activism on police behavior. While online discourse and negative sentiment were associated with a reduction in fatal incidents, physical protests exhibited a positive correlation with police brutality, emphasizing the need for further research to better understand the underlying mechanisms and regional variations.

References:

- Armed Conflict Location & Event Data Project. (2020). US Crisis Monitor 2020 Archive [Data set]. ACLED. https://acleddata.com/us-monitor-2020-archive/

- Bonilla, Y., & Rosa, J. (2015). #Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. American Ethnologist, 42(1), 4-17. https://doi.org/10.1111/amet.12112

- Badaoui, S. (2020). Black Lives Matter: A New Perspective from Twitter Data Mining (Policy Paper No. 20/28). Policy Center for the New South. https://www.policycenter.ma/sites/default/files/PP_20-28_Badaoui.pdf

- Campbell, T. (2023). Black Lives Matter's effect on police lethal use of force. Journal of Urban Economics, Article 103587. https://doi.org/10.1016/j.jue.2023.103587

- Chaffee, S. H., & Schleuder, J. (2006, March 17). Measurement and effects of attention to Media News. OUP Academic. https://doi.org/10.1111/j.1468-2958.1986.tb00096.x

- Giorgi, S., Guntuku, S. C., Himelein-Wachowiak, M., Kwarteng, A., Hwang, S., Rahman, M., & Curtis, B. (2021). Twitter Data of the #BlackLivesMatter Movement And Counter Protests: 2013 to 2020 (Version v2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4897616

- Giorgi, S., Guntuku, S. C., Rahman, M., Himelein-Wachowiak, M., Kwarteng, A., & Curtis, B. (2020). Twitter Corpus of the# BlackLivesMatter Movement And Counter Protests: 2013 to 2020. arXiv preprint arXiv:2009.00596.

- Kabacoff, R. I. (2022). R in Action (3rd Ed). Shelter Island, NY: Manning Publications.

- Mapping Police Violence. (2024). Mapping Police Violence. https://mappingpoliceviolence.us/

- Skoy, E. (2021). Black lives matter protests, fatal police interactions, and crime. Contemporary Economic Policy, 39(2), 280-291. https://doi.org/10.1111/coep.12508

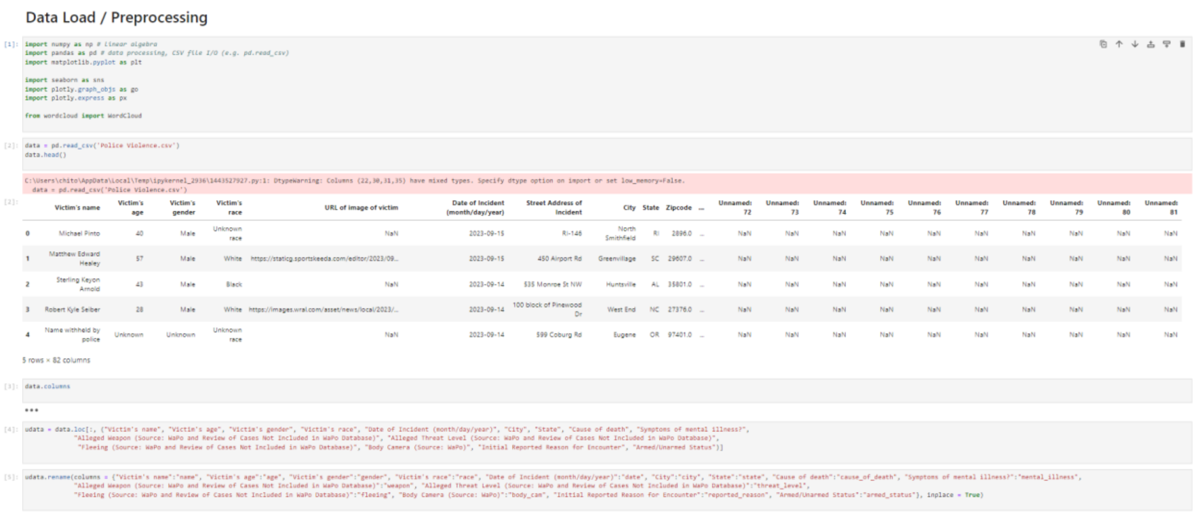
Appendices:

Journal:

Week	Meeting	Progress & Development
1	E-mail	To build framework of the project
2		To build framework of the project
3		To determine the direction
4	E-mail	Proposal submission
5		To write code for data-preprocessing
6		To write code for analysis & regression
7		To write code for analysis & regression
8		To write code for analysis & regression
9	E-mail	Mid-term Progress Report
10		To write my final paper
11		To write my final paper
12		Project Presentation
13	E-mail	To finalize my final paper
14		Submission of my final paper

Code:

Python data pre-processing:



Data Load / Preprocessing

[1]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt

import seaborn as sns
import plotly.graph_objs as go
import plotly.express as px

from wordcloud import WordCloud
```

[2]:
```
data = pd.read_csv('Police Violence.csv')
data.head()
```

```
C:\Users\chito\AppData\Local\Temp\ipykernel_2936\1443527927.py:1: DtypeWarning: Columns (22,30,31,35) have mixed types. Specify dtype option on import or set low_memory=False.
  data = pd.read_csv('Police Violence.csv')
```

[2]:

	Victim's name	Victim's age	Victim's gender	Victim's race	URL of image of victim	Date of Incident (month/day/year)	Street Address of Incident	City	State	Zipcode	...	Unnamed: 72	Unnamed: 73	Unnamed: 74	Unnamed: 75	Unnamed: 76	Unnamed: 77	Unnamed: 78	Unnamed: 79	Unnamed: 80	Unnamed: 81
0	Michael Pinto	40	Male	Unknown race	NaN	2023-09-15	RI-146	North Smithfield	RI	2896.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Matthew Edward Healey	57	Male	White	https://staticg.sportskeeda.com/editor/2023/09...	2023-09-15	450 Airport Rd	Greenvillage	SC	29607.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Sterling Keyon Arnold	43	Male	Black	NaN	2023-09-14	535 Monroe St NW	Huntsville	AL	35801.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Robert Kyle Seiber	28	Male	White	https://images.wral.com/asset/news/local/2023/...	2023-09-14	100 block of Pinewood Dr	West End	NC	27376.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Name withheld by police	Unknown	Unknown	Unknown race	NaN	2023-09-14	599 Coburg Rd	Eugene	OR	97401.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 82 columns

[3]:
```
data.columns
```

•••

[4]:
```
udata = data.loc[:, ("Victim's name", "Victim's age", "Victim's gender", "Victim's race", "Date of Incident (month/day/year)", "City", "State", "Cause of death", "Symptoms of mental illness?",
                 "Alleged Weapon (Source: WaPo and Review of Cases Not Included in WaPo Database)", "Alleged Threat Level (Source: WaPo and Review of Cases Not Included in WaPo Database)",
                  "Fleeing (Source: WaPo and Review of Cases Not Included in WaPo Database)", "Body Camera (Source: WaPo)", "Initial Reported Reason for Encounter", "Armed/Unarmed Status")]
```

[5]:
```
udata.rename(columns = {"Victim's name":"name", "Victim's age":"age", "Victim's gender":"gender", "Victim's race":"race", "Date of Incident (month/day/year)":"date", "City":"city", "State":"state", "Cause of death":"cause_of_death", "Symptoms of mental illness?":"mental_illness",
                 "Alleged Weapon (Source: WaPo and Review of Cases Not Included in WaPo Database)":"weapon", "Alleged Threat Level (Source: WaPo and Review of Cases Not Included in WaPo Database)":"threat_level",
                 "Fleeing (Source: WaPo and Review of Cases Not Included in WaPo Database)":"fleeing", "Body Camera (Source: WaPo)":"body_cam", "Initial Reported Reason for Encounter":"reported_reason", "Armed/Unarmed Status":"armed_status"}, inplace = True)
```

[6]:
```
udata['date'] = pd.to_datetime(udata['date'])
start_date = pd.to_datetime('2013-01-01')
end_date = pd.to_datetime('2021-12-31')
filtered_df = udata[(udata['date'] >= start_date) & (udata['date'] <= end_date)]
filtered_df['year'] = filtered_df['date'].dt.year
filtered_df['month'] = filtered_df['date'].dt.month
filtered_df['day'] = filtered_df['date'].dt.day
```

•••

[7]:
```
filtered_df
```

•••

[8]:
```
filtered_df['reported_reason'].unique()
```

•••

[9]:
```
filtered_df['armed_status'].unique()
```

•••

[10]:
```
na_columns = ['mental_illness', 'body_cam', 'city', 'fleeing']
filtered_df[na_columns] = filtered_df[na_columns].fillna('Unclear')

filtered_df['age'] = filtered_df['age'].fillna('Unknown')

filtered_df['threat_level'] = filtered_df['threat_level'].fillna('undetermined')
filtered_df['threat_level'] = filtered_df['threat_level'].replace(['other ', 'vehicle'], 'other')

filtered_df['reported_reason'] = filtered_df['reported_reason'].fillna('other')

filtered_df['race'] = filtered_df['race'].replace(['Black;Hispanic'], 'Black')
filtered_df['race'] = filtered_df['race'].replace(['Native American;Hispanic'], 'Hispanic')
filtered_df['race'] = filtered_df['race'].replace(['Native American', 'Pacific Islander', 'Unknown Race', 'Unknown race'], 'Other')

filtered_df['mental_illness'] = filtered_df['mental_illness'].replace(['No '], 'No')
filtered_df['mental_illness'] = filtered_df['mental_illness'].replace(['Drug or alcohol use'], 'Drug or Alcohol Use')

filtered_df['armed_status'] = filtered_df['armed_status'].replace(['Unclear', 'Unarmed/Did Not Have Actual Weapon'], 'Unarmed')
filtered_df['armed_status'] = filtered_df['armed_status'].replace(['Allegedly Armed', 'Allegedly armed'], 'Armed')
filtered_df['armed_status'] = filtered_df['armed_status'].replace(['Vehicle'], 'Vehicle')
```

[26]:
```
start_date = pd.to_datetime('2020-01-01')
end_date = pd.to_datetime('2020-12-31')
```

[27]:
```
blm_tweets_df = pd.read_csv('tweet_counts_per_day.csv')
blm_tweets_df['date_str'] = pd.to_datetime(blm_tweets_df['date_str'])
filtered_tweets_df = blm_tweets_df[(blm_tweets_df['date_str'] >= start_date) & (blm_tweets_df['date_str'] <= end_date)]
filtered_tweets_df = filtered_tweets_df[["date_str", "blm"]]
filtered_tweets_df
```

•••

[28]:
```
new_column_names = {
    'date_str': 'date',
    'blm': 'blm_tweets',
}
filtered_tweets_df = filtered_tweets_df.rename(columns=new_column_names)
```

[29]:
```
filtered_df_gf = filtered_df[(filtered_df['date'] >= start_date) & (filtered_df['date'] <= end_date)]
```

[30]:
```
filtered_df_gf['race'].unique()
```

•••

[31]:
```
filtered_race_df = filtered_df_gf[filtered_df_gf['race'] != 'White']
filtered_race_df
```

[32]:
```
filtered_race_df['state'].unique()
```

•••

[33]:
```
offline_df = pd.read_csv('protests_USA.csv')
```

[34]:
```
offline_df['ADMIN1'].unique()
```

•••

[35]:
```
state_dict = {
    'Colorado': 'CO','Washington': 'WA','Illinois': 'IL','Ohio': 'OH','Georgia': 'GA','Florida': 'FL','Iowa': 'IA','Kansas': 'KS','Massachusetts': 'MA','Maine': 'ME','New York': 'NY','Texas': 'TX','California': 'CA','Oregon': 'OR','South Dakota': 'SD','Wisconsin': 'WI','Indiana': 'IN','Michigan': 'MI','Louis
}

offline_df['ADMIN1'] = [state_dict[state] for state in offline_df['ADMIN1']]
```

[36]:
```
offline_df['ADMIN1'].unique()
```

•••

[37]:
```
offline_df['EVENT_DATE'] = pd.to_datetime(offline_df['EVENT_DATE'])
offline_df = offline_df[(offline_df['EVENT_DATE'] >= start_date) & (offline_df['EVENT_DATE'] <= end_date)]
offline_df['ASSOC_ACTOR_1'] = offline_df['ASSOC_ACTOR_1'].fillna("others")
filtered_offline_df = offline_df[offline_df['ASSOC_ACTOR_1'].str.contains('BLM', case=False)]
```

[38]:
```
filtered_offline_df = filtered_offline_df[['EVENT_DATE', 'ADMIN1']]
temp_column_names = {'EVENT_DATE': 'date',
                     'ADMIN1': 'state'}
filtered_offline_df = filtered_offline_df.rename(columns=temp_column_names)
```

[39]:
```
filtered_offline_df
```

•••

[40]:
```
# Group by 'date' and 'state', and count occurrences
filtered_offline_df = filtered_offline_df.groupby(['date', 'state']).size().reset_index(name='state_count')
```

[41]:
```
filtered_offline_df.sort_values('date')
```

[43]:
```
# Merge Data 1 and Data 2 on 'date' and 'state'
merged_data = filtered_race_df.merge(filtered_offline_df, on=['date', 'state'],how='left', indicator=False)
```

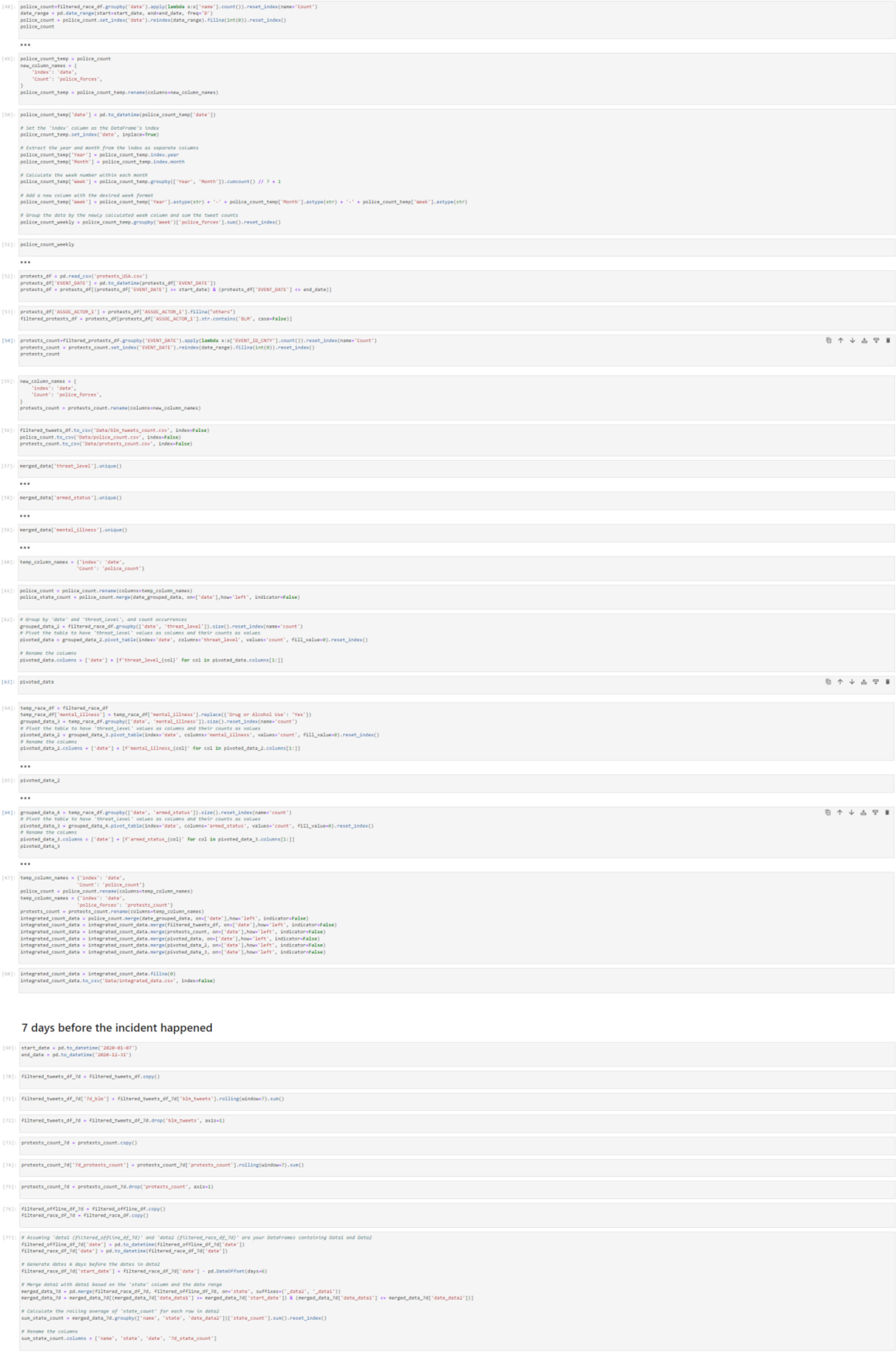
[44]:
```
merged_data
```

•••

[45]:
```
merged_data.to_csv('Data/just_checking.csv', index=False)
```

[46]:
```
merged_data['state_count'] = merged_data['state_count'].fillna(0)
```

[47]:
```
date_grouped_data = merged_data.groupby('date')['state_count'].sum().reset_index()
date_grouped_data
```



[48]:
```
police_count=filtered_race_df.groupby('date').apply(lambda x:x['name'].count()).reset_index(name='Count')
date_range = pd.date_range(start=start_date, end=end_date, freq='D')
police_count = police_count.set_index('date').reindex(date_range).fillna(int(0)).reset_index()
police_count
```

•••

[49]:
```
police_count_temp = police_count
new_column_names = {
    'index': 'date',
    'Count': 'police_forces',
}
police_count_temp = police_count_temp.rename(columns=new_column_names)
```

[50]:
```
police_count_temp['date'] = pd.to_datetime(police_count_temp['date'])

# Set the 'index' column as the DataFrame's index
police_count_temp.set_index('date', inplace=True)

# Extract the year and month from the index as separate columns
police_count_temp['Year'] = police_count_temp.index.year
police_count_temp['Month'] = police_count_temp.index.month

# Calculate the week number within each month
police_count_temp['Week'] = police_count_temp.groupby(['Year', 'Month']).cumcount() // 7 + 1

# Add a new column with the desired week format
police_count_temp['Week'] = police_count_temp['Year'].astype(str) + '-' + police_count_temp['Month'].astype(str) + '-' + police_count_temp['Week'].astype(str)

# Group the data by the newly calculated week column and sum the tweet counts
police_count_weekly = police_count_temp.groupby('Week')['police_forces'].sum().reset_index()
```

[51]:
```
police_count_weekly
```

•••

[52]:
```
protests_df = pd.read_csv('protests_USA.csv')
protests_df['EVENT_DATE'] = pd.to_datetime(protests_df['EVENT_DATE'])
protests_df = protests_df[(protests_df['EVENT_DATE'] >= start_date) & (protests_df['EVENT_DATE'] <= end_date)]
```

[53]:
```
protests_df['ASSOC_ACTOR_1'] = protests_df['ASSOC_ACTOR_1'].fillna("others")
filtered_protests_df = protests_df[protests_df['ASSOC_ACTOR_1'].str.contains('BLM', case=False)]
```

[54]:
```
protests_count=filtered_protests_df.groupby('EVENT_DATE').apply(lambda x:x['EVENT_ID_CNTY'].count()).reset_index(name='Count')
protests_count = protests_count.set_index('EVENT_DATE').reindex(date_range).fillna(int(0)).reset_index()
protests_count
```

[55]:
```
new_column_names = {
    'index': 'date',
    'Count': 'police_forces',
}
protests_count = protests_count.rename(columns=new_column_names)
```

[56]:
```
filtered_tweets_df.to_csv('Data/blm_tweets_count.csv', index=False)
police_count.to_csv('Data/police_count.csv', index=False)
protests_count.to_csv('Data/protests_count.csv', index=False)
```

[57]:
```
merged_data['threat_level'].unique()
```

•••

[58]:
```
merged_data['armed_status'].unique()
```

•••

[59]:
```
merged_data['mental_illness'].unique()
```

•••

[60]:
```
temp_column_names = {'index': 'date',
                     'Count': 'police_count'}
```

[61]:
```
police_count = police_count.rename(columns=temp_column_names)
police_state_count = police_count.merge(date_grouped_data, on=['date'],how='left', indicator=False)
```

[62]:
```
# Group by 'date' and 'threat_level', and count occurrences
grouped_data_2 = filtered_race_df.groupby(['date', 'threat_level']).size().reset_index(name='count')
# Pivot the table to have 'threat_level' values as columns and their counts as values
pivoted_data = grouped_data_2.pivot_table(index='date', columns='threat_level', values='count', fill_value=0).reset_index()

# Rename the columns
pivoted_data.columns = ['date'] + [f'threat_level_{col}' for col in pivoted_data.columns[1:]]
```

[63]:
```
pivoted_data
```

[64]:
```
temp_race_df = filtered_race_df
temp_race_df['mental_illness'] = temp_race_df['mental_illness'].replace({'Drug or Alcohol Use': 'Yes'})
grouped_data_3 = temp_race_df.groupby(['date', 'mental_illness']).size().reset_index(name='count')
# Pivot the table to have 'threat_level' values as columns and their counts as values
pivoted_data_2 = grouped_data_3.pivot_table(index='date', columns='mental_illness', values='count', fill_value=0).reset_index()
# Rename the columns
pivoted_data_2.columns = ['date'] + [f'mental_illness_{col}' for col in pivoted_data_2.columns[1:]]
```

•••

[65]:
```
pivoted_data_2
```

•••

[66]:
```
grouped_data_4 = temp_race_df.groupby(['date', 'armed_status']).size().reset_index(name='count')
# Pivot the table to have 'threat_level' values as columns and their counts as values
pivoted_data_3 = grouped_data_4.pivot_table(index='date', columns='armed_status', values='count', fill_value=0).reset_index()
# Rename the columns
pivoted_data_3.columns = ['date'] + [f'armed_status_{col}' for col in pivoted_data_3.columns[1:]]
pivoted_data_3
```

•••

[67]:
```
temp_column_names = {'index': 'date',
                     'Count': 'police_count'}
police_count = police_count.rename(columns=temp_column_names)
temp_column_names = {'index': 'date',
                     'police_forces': 'protests_count'}
protests_count = protests_count.rename(columns=temp_column_names)
integrated_count_data = police_count.merge(date_grouped_data, on=['date'],how='left', indicator=False)
integrated_count_data = integrated_count_data.merge(filtered_tweets_df, on=['date'],how='left', indicator=False)
integrated_count_data = integrated_count_data.merge(protests_count, on=['date'],how='left', indicator=False)
integrated_count_data = integrated_count_data.merge(pivoted_data, on=['date'],how='left', indicator=False)
integrated_count_data = integrated_count_data.merge(pivoted_data_2, on=['date'],how='left', indicator=False)
integrated_count_data = integrated_count_data.merge(pivoted_data_3, on=['date'],how='left', indicator=False)
```

[68]:
```
integrated_count_data = integrated_count_data.fillna(0)
integrated_count_data.to_csv('Data/integrated_data.csv', index=False)
```

7 days before the incident happened

[69]:
```
start_date = pd.to_datetime('2020-01-07')
end_date = pd.to_datetime('2020-12-31')
```

[70]:
```
filtered_tweets_df_7d = filtered_tweets_df.copy()
```

[71]:
```
filtered_tweets_df_7d['7d_blm'] = filtered_tweets_df_7d['blm_tweets'].rolling(window=7).sum()
```

[72]:
```
filtered_tweets_df_7d = filtered_tweets_df_7d.drop('blm_tweets', axis=1)
```

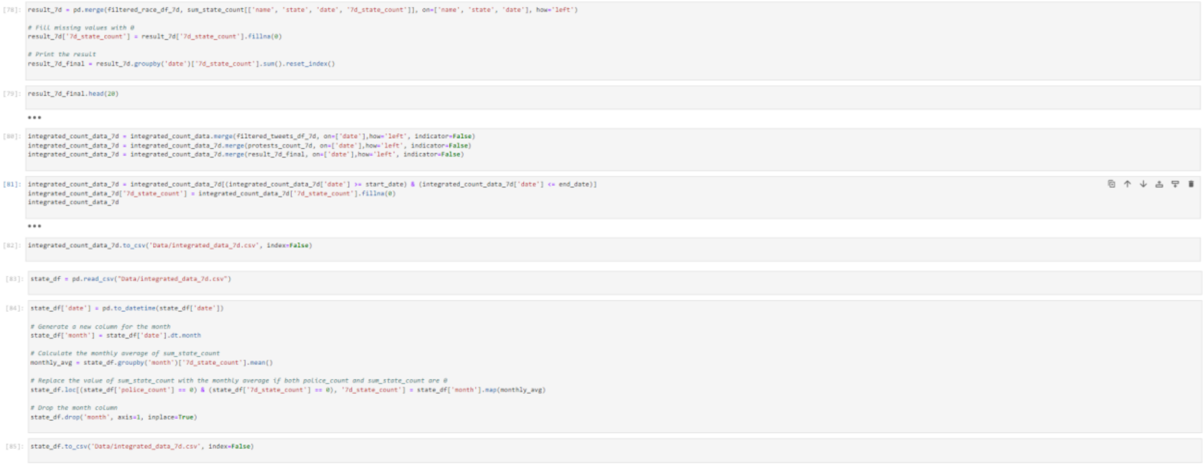
[73]:
```
protests_count_7d = protests_count.copy()
```

[74]:
```
protests_count_7d['7d_protests_count'] = protests_count_7d['protests_count'].rolling(window=7).sum()
```

[75]:
```
protests_count_7d = protests_count_7d.drop('protests_count', axis=1)
```

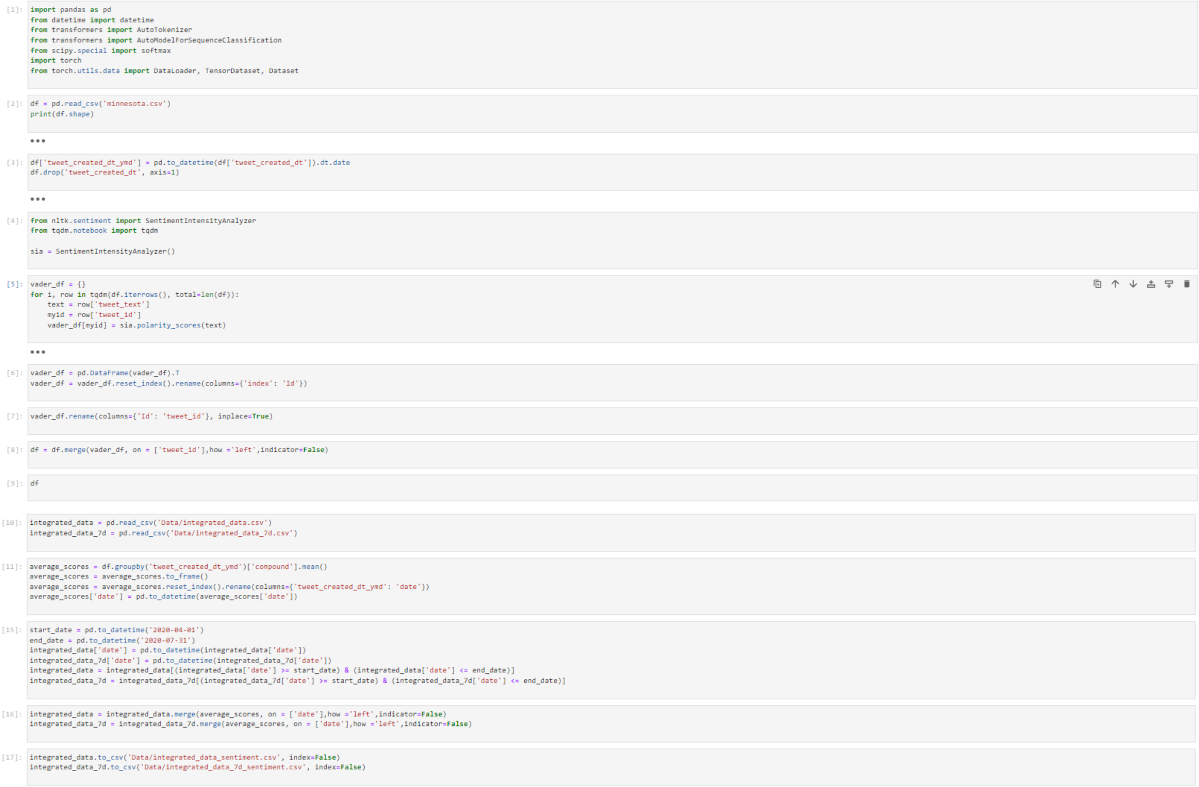
[76]:
```
filtered_offline_df_7d = filtered_offline_df.copy()
filtered_race_df_7d = filtered_race_df.copy()
```

[77]:
```
# Assuming 'data1 (filtered_offline_df_7d)' and 'data2 (filtered_race_df_7d)' are your DataFrames containing Data1 and Data2
filtered_offline_df_7d['date'] = pd.to_datetime(filtered_offline_df_7d['date'])
filtered_race_df_7d['date'] = pd.to_datetime(filtered_race_df_7d['date'])

# Generate dates 6 days before the dates in data2
filtered_race_df_7d['start_date'] = filtered_race_df_7d['date'] - pd.DateOffset(days=6)

# Merge data2 with data1 based on the 'state' column and the date range
merged_data_7d = pd.merge(filtered_race_df_7d, filtered_offline_df_7d, on='state', suffixes=('_data2', '_data1'))
merged_data_7d = merged_data_7d[(merged_data_7d['date_data1'] >= merged_data_7d['start_date']) & (merged_data_7d['date_data1'] <= merged_data_7d['date_data2'])]

# Calculate the rolling average of 'state_count' for each row in data2
sum_state_count = merged_data_7d.groupby(['name', 'state', 'date_data2'])['state_count'].sum().reset_index()

# Rename the columns
sum_state_count.columns = ['name', 'state', 'date', '7d_state_count']
```



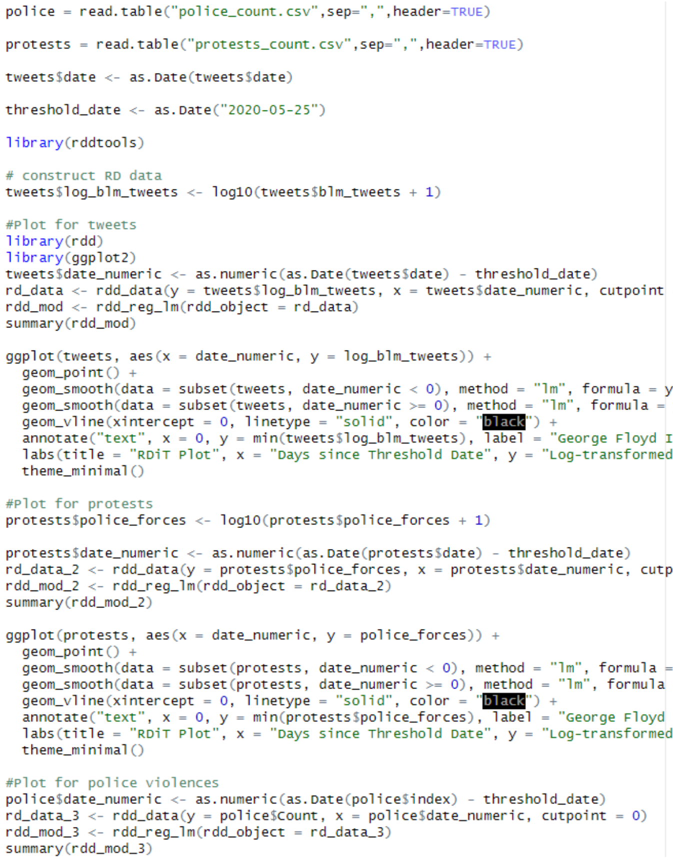
```
[78]: result_7d = pd.merge(filtered_race_df_7d, sum_state_count[['name', 'state', 'date', '7d_state_count']], on=['name', 'state', 'date'], how='left')

      # Fill missing values with 0
      result_7d['7d_state_count'] = result_7d['7d_state_count'].fillna(0)

      # Print the result
      result_7d_final = result_7d.groupby('date')['7d_state_count'].sum().reset_index()

[79]: result_7d_final.head(20)

      •••

[80]: integrated_count_data_7d = integrated_count_data.merge(filtered_tweets_df_7d, on=['date'],how='left', indicator=False)
      integrated_count_data_7d = integrated_count_data_7d.merge(protests_count_7d, on=['date'],how='left', indicator=False)
      integrated_count_data_7d = integrated_count_data_7d.merge(result_7d_final, on=['date'],how='left', indicator=False)

[81]: integrated_count_data_7d = integrated_count_data_7d[(integrated_count_data_7d['date'] >= start_date) & (integrated_count_data_7d['date'] <= end_date)]
      integrated_count_data_7d['7d_state_count'] = integrated_count_data_7d['7d_state_count'].fillna(0)
      integrated_count_data_7d

      •••

[82]: integrated_count_data_7d.to_csv('Data/integrated_data_7d.csv', index=False)

[83]: state_df = pd.read_csv("Data/integrated_data_7d.csv")

[84]: state_df['date'] = pd.to_datetime(state_df['date'])

      # Generate a new column for the month
      state_df['month'] = state_df['date'].dt.month

      # Calculate the monthly average of sum_state_count
      monthly_avg = state_df.groupby('month')['7d_state_count'].mean()

      # Replace the value of sum_state_count with the monthly average if both police_count and sum_state_count are 0
      state_df.loc[(state_df['police_count'] == 0) & (state_df['7d_state_count'] == 0), '7d_state_count'] = state_df['month'].map(monthly_avg)

      # Drop the month column
      state_df.drop('month', axis=1, inplace=True)

[85]: state_df.to_csv('Data/integrated_data_7d.csv', index=False)
```

Python sentiment analysis:

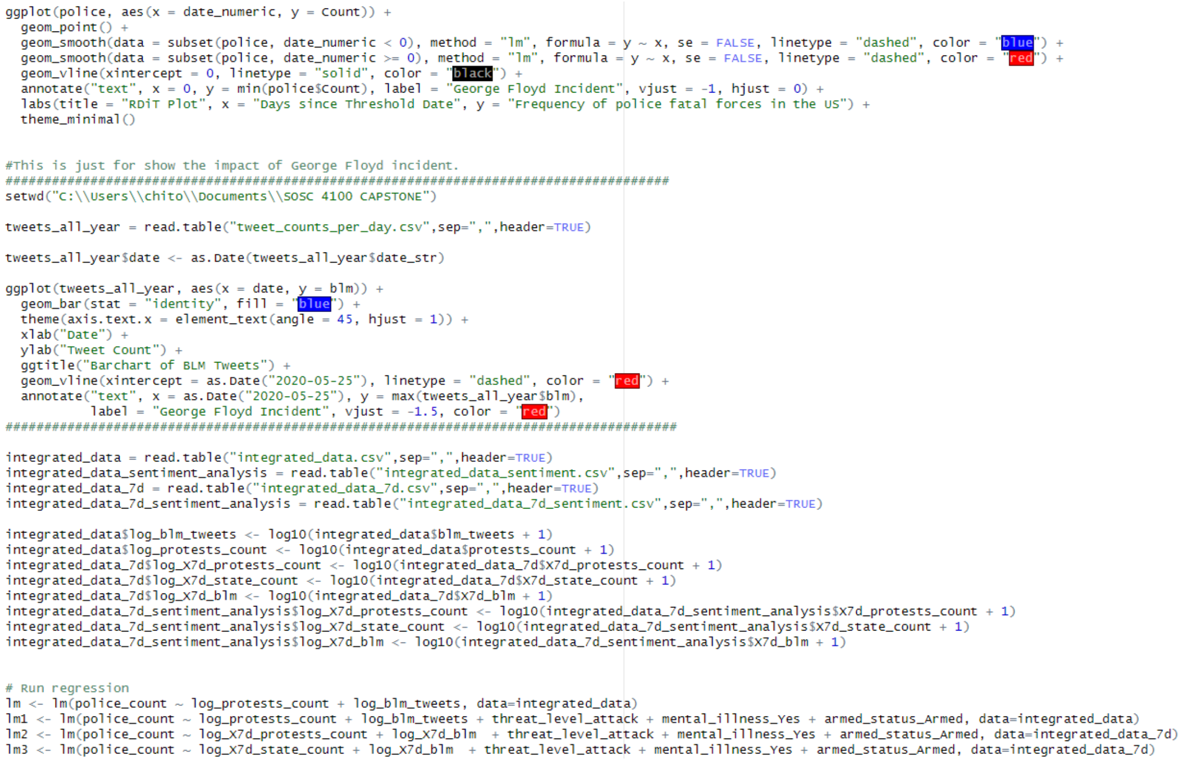


```
[1]: import pandas as pd
     from datetime import datetime
     from transformers import AutoTokenizer
     from transformers import AutoModelForSequenceClassification
     from scipy.special import softmax
     import torch
     from torch.utils.data import DataLoader, TensorDataset, Dataset

[2]: df = pd.read_csv('minnesota.csv')
     print(df.shape)

     •••

[3]: df['tweet_created_dt_ymd'] = pd.to_datetime(df['tweet_created_dt']).dt.date
     df.drop('tweet_created_dt', axis=1)

     •••

[4]: from nltk.sentiment import SentimentIntensityAnalyzer
     from tqdm.notebook import tqdm

     sia = SentimentIntensityAnalyzer()

[5]: vader_df = {}
     for i, row in tqdm(df.iterrows(), total=len(df)):
         text = row['tweet_text']
         myid = row['tweet_id']
         vader_df[myid] = sia.polarity_scores(text)

     •••

[6]: vader_df = pd.DataFrame(vader_df).T
     vader_df = vader_df.reset_index().rename(columns={'index': 'Id'})

[7]: vader_df.rename(columns={'Id': 'tweet_id'}, inplace=True)

[8]: df = df.merge(vader_df, on = ['tweet_id'],how ='left',indicator=False)

[9]: df

[10]: integrated_data = pd.read_csv('Data/integrated_data.csv')
      integrated_data_7d = pd.read_csv('Data/integrated_data_7d.csv')

[11]: average_scores = df.groupby('tweet_created_dt_ymd')['compound'].mean()
      average_scores = average_scores.to_frame()
      average_scores = average_scores.reset_index().rename(columns={'tweet_created_dt_ymd': 'date'})
      average_scores['date'] = pd.to_datetime(average_scores['date'])

[15]: start_date = pd.to_datetime('2020-04-01')
      end_date = pd.to_datetime('2020-07-31')
      integrated_data['date'] = pd.to_datetime(integrated_data['date'])
      integrated_data_7d['date'] = pd.to_datetime(integrated_data_7d['date'])
      integrated_data = integrated_data[(integrated_data['date'] >= start_date) & (integrated_data['date'] <= end_date)]
      integrated_data_7d = integrated_data_7d[(integrated_data_7d['date'] >= start_date) & (integrated_data_7d['date'] <= end_date)]

[16]: integrated_data = integrated_data.merge(average_scores, on = ['date'],how ='left',indicator=False)
      integrated_data_7d = integrated_data_7d.merge(average_scores, on = ['date'],how ='left',indicator=False)

[17]: integrated_data.to_csv('Data/integrated_data_sentiment.csv', index=False)
      integrated_data_7d.to_csv('Data/integrated_data_7d_sentiment.csv', index=False)
```

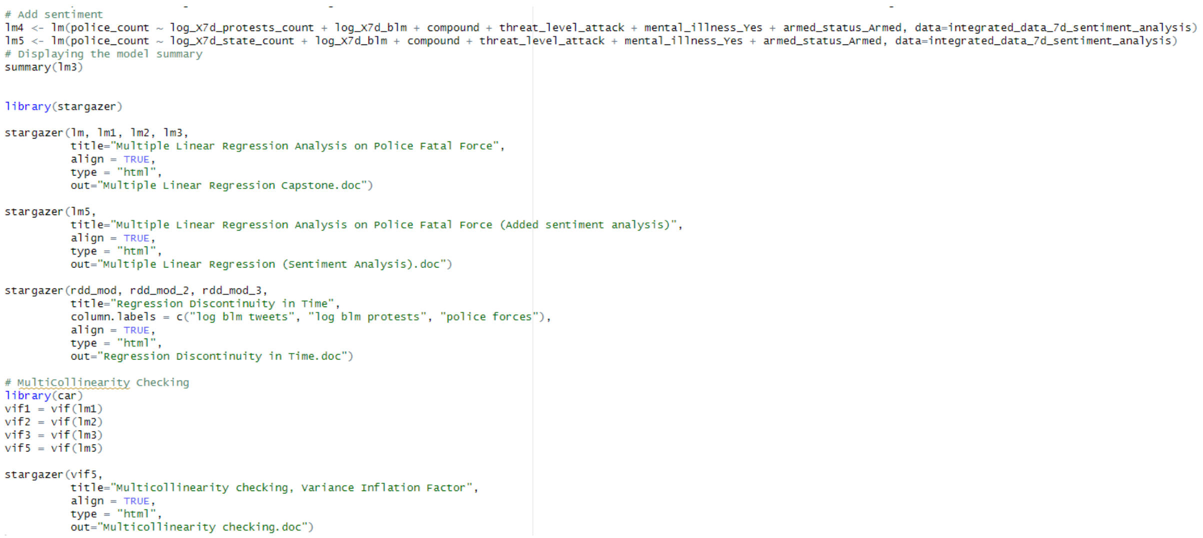
R Code:



```r
police = read.table("police_count.csv",sep=",",header=TRUE)

protests = read.table("protests_count.csv",sep=",",header=TRUE)

tweets$date <- as.Date(tweets$date)

threshold_date <- as.Date("2020-05-25")

library(rddtools)

# construct RD data
tweets$log_blm_tweets <- log10(tweets$blm_tweets + 1)

#Plot for tweets
library(rdd)
library(ggplot2)
tweets$date_numeric <- as.numeric(as.Date(tweets$date) - threshold_date)
rd_data <- rdd_data(y = tweets$log_blm_tweets, x = tweets$date_numeric, cutpoint = 0)
rdd_mod <- rdd_reg_lm(rdd_object = rd_data)
summary(rdd_mod)

ggplot(tweets, aes(x = date_numeric, y = log_blm_tweets)) +
  geom_point() +
  geom_smooth(data = subset(tweets, date_numeric < 0), method = "lm", formula = y ~ x, se = FALSE, linetype = "dashed", color = "blue") +
  geom_smooth(data = subset(tweets, date_numeric >= 0), method = "lm", formula = y ~ x, se = FALSE, linetype = "dashed", color = "red") +
  geom_vline(xintercept = 0, linetype = "solid", color = "black") +
  annotate("text", x = 0, y = min(tweets$log_blm_tweets), label = "George Floyd Incident", vjust = -1, hjust = 0) +
  labs(title = "RDiT Plot", x = "Days since Threshold Date", y = "Log-transformed BLM Tweets (log10)") +
  theme_minimal()

#Plot for protests
protests$police_forces <- log10(protests$police_forces + 1)

protests$date_numeric <- as.numeric(as.Date(protests$date) - threshold_date)
rd_data_2 <- rdd_data(y = protests$police_forces, x = protests$date_numeric, cutpoint = 0)
rdd_mod_2 <- rdd_reg_lm(rdd_object = rd_data_2)
summary(rdd_mod_2)

ggplot(protests, aes(x = date_numeric, y = police_forces)) +
  geom_point() +
  geom_smooth(data = subset(protests, date_numeric < 0), method = "lm", formula = y ~ x, se = FALSE, linetype = "dashed", color = "blue") +
  geom_smooth(data = subset(protests, date_numeric >= 0), method = "lm", formula = y ~ x, se = FALSE, linetype = "dashed", color = "red") +
  geom_vline(xintercept = 0, linetype = "solid", color = "black") +
  annotate("text", x = 0, y = min(protests$police_forces), label = "George Floyd Incident", vjust = -1, hjust = 0) +
  labs(title = "RDiT Plot", x = "Days since Threshold Date", y = "Log-transformed BLM protests in the US (log10)") +
  theme_minimal()

#Plot for police violences
police$date_numeric <- as.numeric(as.Date(police$index) - threshold_date)
rd_data_3 <- rdd_data(y = police$Count, x = police$date_numeric, cutpoint = 0)
rdd_mod_3 <- rdd_reg_lm(rdd_object = rd_data_3)
summary(rdd_mod_3)
```



```r
ggplot(police, aes(x = date_numeric, y = Count)) +
  geom_point() +
  geom_smooth(data = subset(police, date_numeric < 0), method = "lm", formula = y ~ x, se = FALSE, linetype = "dashed", color = "blue") +
  geom_smooth(data = subset(police, date_numeric >= 0), method = "lm", formula = y ~ x, se = FALSE, linetype = "dashed", color = "red") +
  geom_vline(xintercept = 0, linetype = "solid", color = "black") +
  annotate("text", x = 0, y = min(police$Count), label = "George Floyd Incident", vjust = -1, hjust = 0) +
  labs(title = "RDiT Plot", x = "Days since Threshold Date", y = "Frequency of police fatal forces in the US") +
  theme_minimal()


#This is just for show the impact of George Floyd incident.
######################################################################################
setwd("C:\\Users\\chito\\Documents\\SOSC 4100 CAPSTONE")

tweets_all_year = read.table("tweet_counts_per_day.csv",sep=",",header=TRUE)

tweets_all_year$date <- as.Date(tweets_all_year$date_str)

ggplot(tweets_all_year, aes(x = date, y = blm)) +
  geom_bar(stat = "identity", fill = "blue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Date") +
  ylab("Tweet Count") +
  ggtitle("Barchart of BLM Tweets") +
  geom_vline(xintercept = as.Date("2020-05-25"), linetype = "dashed", color = "red") +
  annotate("text", x = as.Date("2020-05-25"), y = max(tweets_all_year$blm),
           label = "George Floyd Incident", vjust = -1.5, color = "red")
#######################################################################################

integrated_data = read.table("integrated_data.csv",sep=",",header=TRUE)
integrated_data_sentiment_analysis = read.table("integrated_data_sentiment.csv",sep=",",header=TRUE)
integrated_data_7d = read.table("integrated_data_7d.csv",sep=",",header=TRUE)
integrated_data_7d_sentiment_analysis = read.table("integrated_data_7d_sentiment.csv",sep=",",header=TRUE)

integrated_data$log_blm_tweets <- log10(integrated_data$blm_tweets + 1)
integrated_data$log_protests_count <- log10(integrated_data$protests_count + 1)
integrated_data_7d$log_X7d_protests_count <- log10(integrated_data_7d$X7d_protests_count + 1)
integrated_data_7d$log_X7d_state_count <- log10(integrated_data_7d$X7d_state_count + 1)
integrated_data_7d$log_X7d_blm <- log10(integrated_data_7d$X7d_blm + 1)
integrated_data_7d_sentiment_analysis$log_X7d_protests_count <- log10(integrated_data_7d_sentiment_analysis$X7d_protests_count + 1)
integrated_data_7d_sentiment_analysis$log_X7d_state_count <- log10(integrated_data_7d_sentiment_analysis$X7d_state_count + 1)
integrated_data_7d_sentiment_analysis$log_X7d_blm <- log10(integrated_data_7d_sentiment_analysis$X7d_blm + 1)


# Run regression
lm <- lm(police_count ~ log_protests_count + log_blm_tweets, data=integrated_data)
lm1 <- lm(police_count ~ log_protests_count + log_blm_tweets + threat_level_attack + mental_illness_Yes + armed_status_Armed, data=integrated_data)
lm2 <- lm(police_count ~ log_X7d_protests_count + log_X7d_blm  + threat_level_attack + mental_illness_Yes + armed_status_Armed, data=integrated_data_7d)
lm3 <- lm(police_count ~ log_X7d_state_count + log_X7d_blm  + threat_level_attack + mental_illness_Yes + armed_status_Armed, data=integrated_data_7d)
```



```r
# Add sentiment
lm4 <- lm(police_count ~ log_X7d_protests_count + log_X7d_blm + compound + threat_level_attack + mental_illness_Yes + armed_status_Armed, data=integrated_data_7d_sentiment_analysis)
lm5 <- lm(police_count ~ log_X7d_state_count + log_X7d_blm + compound + threat_level_attack + mental_illness_Yes + armed_status_Armed, data=integrated_data_7d_sentiment_analysis)
# Displaying the model summary
summary(lm3)


library(stargazer)

stargazer(lm, lm1, lm2, lm3,
          title="Multiple Linear Regression Analysis on Police Fatal Force",
          align = TRUE,
          type = "html",
          out="Multiple Linear Regression Capstone.doc")

stargazer(lm5,
          title="Multiple Linear Regression Analysis on Police Fatal Force (Added sentiment analysis)",
          align = TRUE,
          type = "html",
          out="Multiple Linear Regression (Sentiment Analysis).doc")

stargazer(rdd_mod, rdd_mod_2, rdd_mod_3,
          title="Regression Discontinuity in Time",
          column.labels = c("log blm tweets", "log blm protests", "police forces"),
          align = TRUE,
          type = "html",
          out="Regression Discontinuity in Time.doc")

# MultiCollinearity Checking
library(car)
vif1 = vif(lm1)
vif2 = vif(lm2)
vif3 = vif(lm3)
vif5 = vif(lm5)

stargazer(vif5,
          title="Multicollinearity checking, Variance Inflation Factor",
          align = TRUE,
          type = "html",
          out="Multicollinearity checking.doc")
```