DATA 698

Capstone Project

# Association of Hateful Tweets and Hate Crimes in New York City

Prepared for:

Prof. Dr. Nasrin Khansari

City University of New York, School of Professional Studies

Prepared by:

Sie Siong Wong

Mario Pena

Joseph Shi

December 07, 2022

# Contents

# Abstract

Hate crime has become one of the major problems for communities within the New York City. Everyone is not born to hate someone because they are different in race, color, religion, gender, age, and so on. There are a lot of things we see and hear could influence our perception towards certain things. As we are all living in this digital age, large amounts of information are widely available to many people through social media. The objective of this study is to find out whether hateful tweets has any association with the hate crimes occurred in the NYC areas and do minority groups are mostly the target. Twitter's hate related data collection, sentiment analysis, and spatio-temporal analysis are the key players in this study. 3 years (2019-2021) of data are collected from Twitter and NYC Open Data source. Sentiment analysis is used to detect negative attitude tweets which we think is highly related to hate speech and consider them as hateful tweets for our further analysis. Spatio-temporal Bayesian modeling is widely used in the epidemiological studies to analyze local patterns over time by employing spatial and temporal random effects (Hu, 2019). Thus, we can use the same approach to study correlation between hateful tweets and hate crimes and any developing trend of hate crime risks over the 5 boroughs of New York City by observing the posterior distribution for the parameters in Bayesian Hierarchical spatio-temporal model. From the analysis results, we have found that hate crime victims are mostly the minority groups which races are black, other, and gender in female. Even though our analysis shows that there is no significant relationship between hateful tweets and hate crimes but we do see the developing trend of hate crime relative risk changes from 2019 to 2021 in Brooklyn, Queens, and Staten Island.

# Introduction

Many of us spend several hours per week in social media platforms such as Facebook, Twitter, and Instagram to name a few. Although there are many advantages that social media provides e.g. serving as the fifth estate of power, a place to express opinions and such, it has also become a platform for someone to spread hate, cyberbullying, harassments and so on. Messages of hatred and spread of violence could have a high potential for influencing and motivating someone to commit a crime. According to CNN (Studley & Tucker, 2022), hate crimes in New York City have increased 76% in the first quarter of 2022 compared to last year. Many social media companies such as Facebook, have started to control the contents being posted by their users. Because there is a big concern in our society that offensive speech can result in more crimes and eventually become a serious threat to social, political, and cultural stability, we see the need to find a solution.

Some may argue that the First Amendment allows individuals to express anything they want. Instead of limiting everyone's freedom of speech openly in social media platforms, there are possible alternatives to deal with this issue. For example, a platform can allow users to turn on or off a switch to hide possible sensitive contents. In order to do this, we need some kind of solution that is able to identify inappropriate speech, tag them, and warn users that their post may not be visible to everyone due to containing sensitive language.

We believe that offensive speech has a high potential for causing more violent crimes to happen in the city. To make our city a better place to live, we need to find a way not only to identify inappropriate speech in social media platforms and reduce the spread of such speech, but also when to deploy more law enforcement to patrol areas crowded with high numbers of offensive speech in the hope we interrupt crime prematurely. Our research hypothesis is that higher offensive tweets from an area have strong correlation to the number of crimes committed in that area and minority groups are mostly the target.

The challenging part for our research is to classify the tweets data into hate speech. Defining what is hate speech can be hard because it's constantly evolving and often dependent on context. Until today, artificial intelligence still struggles when it comes to identifying hate speech. But we have found that quite a few research scientists have tried different approaches to identifying offensive speech in social media text data. The more advanced techniques, deep learning methods such as gated recurrent unit (GRU), convolutional neural networks (CNN), gated convolutional recurrent - neural networks (GCR-NN), and combination of transfer learning and week supervision, are so far giving the best results compared to conventional classification approaches such as logistic regression, naive Bayes, decision tree, random forest, and gradient boosting.

There are many different types of hate speech that target different groups and individuals, and research has been done to differentiate and define them accordingly. Some approaches to the classification of hate speech from Twitter use the hashtag as a variable for predicting violence and hateful message from tweets. While other approaches for classification of hate speech on social media may also include the development of neural language models. Additionally, there are algorithms that have been created to detect hate speech in digital micro-environments, which facilitate and reduce analysis tasks undergone by law enforcement agencies and service providers. We have also come across research that found there may not be a correlation between social media messages and crime, and others that consider "weather" to be an important environmental factor that affects the occurrence of criminal incidents. There have been many different approaches tried to find a link between social media posts and crime, all with varying results, and what makes this research so interesting is that we may yet find different results that may contribute to this solution.

# Literature Review

The increase of using social media platforms such as Twitter and Facebook as a safe harbor to spread hate contents, and to cyberbully or harrass an individual or a minority group because of their identities such as race, religion, sexual orientation, etc., has promoted social disharmony in which may bring local crime rate to rise. (Williams et al., 2019) In this research article, a Weka tool was used to develop a machine learning classifier to annotate racism and religion and non-hateful tweets. Besides tweets and crime data, census data such as no qualifications, age, long-term unemployed, and black and minority ethnicity, were considered part of the study. Statistical models were built to study a temporal and spatial association. Random and fixed effects, Poisson regression models, and negative binomial models, which all results show consistent of positive association between hate speech in Twitter and offline racially and religiously discriminated crimes. Other research has also shown that online hate speech has a strong correlation with significant events such as terror attacks, political votes, etc (Hanes & Machin, 2013; Williams & Burnap, 2015). A recent example of this was the Christchurch extreme right terror attacker's post on 8chan. Although there is no direct causal link between online hate speech and offline hate crimes, hate speech is still part of the cumulative process for a hate formula from social status, political context and geographical perspective that bring harm to the physical world.

Nowadays, hate speech detection remains a big challenge for artificial intelligence. (Matsaki L., 2018) Because the definition of hate speech is constantly evolving and often hidden within context, AI can be fooled easily. AI can be fooled by inserting typos, adding extra alphabets, and removing spaces. An example of altered text to evade AI detection would be something like this: "MartiansAreDisgustingAndShouldBeKilled love." Humans can understand this message but the machine learning algorithms have trouble identifying it. Facebook CEO Mark Zuckerberg testified before congress in 2018 and mentioned he was optimistic that in 10 years AI would be able to identify hate speech more accurate. (Shead S., 2020) Only 24 percent of hate speech was able to be detected according to Facebook's chief technology officer, Mike Schroepfer. Billions of users consume the Facebook product and platform, a cutting-edge machine technology to accurately spot hate speech is a must-have, to protect its users from exposure to harmful content.

Many previous researchers have tried different machine learning approaches to best identify hate speech. (Lee et al., 2022) Stacked ensemble Gated Convolutional Recurrent- Neural Networks (GCR-NN), a deep learning approach was used to detect racist speech from a tweets dataset. Its 0.98 accuracy outperforms other conventional machine learning models such as Random Forest, K-Nearest Neighbors, SVM, etc. A large scale dataset of tweets related to racism from the world were collected, preprocessed to such detail as to remove stop words, annotated using the TextBlob into positive, negative and neutral score sentiments before performing Term Frequency - Inverse Document Frequency (TD-IDF), and Bag of Words (BoW) for machine learning models development. United States has the highest number of racist tweets, 50%, followed by United Kingdom, 31%. More than 53% of people age between 15-30 years old in the United States is exposed to online hate material (Hawdon et al., 2016). Both Logistic Regression and SVM performs better than other models using BoW features on average with a 0.97 accuracy score. For a fair comparison, few single deep learning models such as Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), were also implemented and optimized through hyper parameter tuning. (Crabb et al., 2019; Starosta A., 2019) Combination of Weak Supervision and Transfer Learning approach is useful for identifying which data lacks labeling. Weak supervision is a method for building a labeled training set from a large amount of unlabeled data. Transfer learning is a method where it reuses already existing pre-trained models for new tasks. It will use the training set created from weak supervision stage to build a new classifier model.

Research has found that there is a contradiction between freedom of speech and hate speech. Hate speech creates an environment that tests the limits of free speech, and it violates fundamental rights of a human being. On social networks, we can observe how other forms of violence such as terrorism, extremism and hate crimes, may each have their own space, victims and aggressors (Chetty & Alathur, 2018). Miró-Llinares and Rodriguez-Sala (2016) describe the existence of many different types of hate speech that target different groups and individuals. Research has been done to differentiate and define each type accordingly. It is important to keep in mind classification as it will help us group Tweets with similar speech together in one category. Some of the approaches to the classification of hate speech from Tweeter used by Miró-Llinares and Rodriguez-Sala (2016) take into account the hashtags as a variable for predicting violence and hate message from the tweet. While some of their other approaches for classification of hate speech on social media also include the development of neural language models. It is also important to underline that the authors warn the development of classification algorithms may not allow for a deeper understanding of the different nature of violent communications. The recommendation they make is to pursue a thorough study of the categorization of different expressions that are evident in these types of messages.

The study conducted by Miró-Llinares et al. (2018) focuses on designing an algorithm to detect hate speech in digital microenvironments. Its purpose is to facilitate and reduce analysis tasks undergone by law enforcement agencies and service providers. The algorithm used in this specific article uses machine learning classification techniques such as Random Forests. Furthermore, the study demonstrated that not all variables in the data relating to anonymity and visibility of users are applicable in distinguishing hate speech in tweets' content, and that tweet metadata proved to be more efficient in the classification process than account metadata. Compared to similar studies that have applied different classification approaches, the results obtained by this study slightly outperform the others. The Random Forest model applied reached a F1-score of 0.92, highlighting the accuracy of the model on the dataset. Previous attempts from other studies have obtained F-measures of 0.77, 0.90 and 0.76 according to the literature in this specific study.

We have also found research that is centered on the idea that social media metadata is a valuable input in the analysis of opinions and sentiment. From activism to detecting road traffic, access to these data is essential in today's comprehensive analyses. Using data from 18 Spanish-speaking Latin American countries, research found that similarly to mass media, social media suffers from a strong bias towards violent or sexual crimes. Simple models such as a linear regression were used in the research by Curiel et al. (2020) and found that countries with higher number of murders, murder rate and fear of crime are more likely to have crime-related tweets. However, it mainly represents the fears that people have of crime that may overemphasize certain types of crimes that are not as common as one would think. It was also found that there may not be a correlation between social media messages and crime, but a reflection of the level of fear of crime.

As explained by Chen et al. (2015), it is possible for weather to also be used as a feature for modeling crime around Tweeter data. Their work is also focused on predicting crime in order to maximize the allocation of scarce police resources. Their paper noted the limitations of past studies, suggesting that weather is a significant factor and it has been proved there is a correlation between weather and criminal activity. It is thought that certain environmental factors such as weather, should be considered as it may affect the occurrence of criminal incidents. Such data containing information about theft density, Twitter data, weather and geolocation points, have been modeled using logistic regression with recommendations for future analysis using support vector machine (Chen et al., 2015). More advanced techniques to predicting crime from Tweeter posts include Artificial Neural Networks approach that have achieved average accuracies of 0.903 on testing dataset and 0.933 on the training dataset as those demonstrated by Sandagiri et al. (2021).

To develop and train a hate detection model, we'll need a dataset with each record labeled with either hate or non-hate. This article (Alnazzawi, 2022) developed a corpus which has hate related context, hate crime types, and the motivation behind the hate crimes. This dataset is available in the Kaggle website and freely available to download. For our research, we could leverage this dataset in part of our annotation process to classify hate and non-hate tweets using the key words in the given hate related context. There were two steps in developing the dataset, corpus construction and annotation. The author used a Hashtag tracking tool "Hashtagify" to search for hate related hashtags which were later to be used in the TweetsScraper tool to collect tweets related to hateful contents. Nine years (2010-2019) of tweets data related to these hashtags were collected. These hashtags list contain hate crimes, racist, racism, Islamophobia, Islamophobic, sexism, disability, transgender, antisemitism, misogyny, and disabled. This hashtags list was found to be similar to what FBI used for the hate crime classification. For the annotation part, the author used the COGITO Tech (LLC) service to annotate each hate related tweets with 3 types of hate crime (physical assault, verbal abuse, incitement to hatred) and 5 types of motivation (racism, religion, disability, sexism, unknown) behind committing the crime. More than 60% of the 23,179 tweets, which hashtags' are hate crime related, are made up of physical assault and most of these physical assault crimes were motivated because of different races and ethnicities.

# Methodology

There are two sets of data we collect for this study, crime data and tweets data. The date range for these data is from January 2019 to December 2021. We may extend the data collection a year back if needed for our analysis.

## Data Collection

### Crime Data:

The crime data was collected from the NYC Open Data source (NYPD Complaint Data Historic) by using the available API. The read.socrata() function available from the "RSocrata" library was used to send SQL query command through the API to get NYC crime data. The return data from the query has about 1.31 million rows of crime data. Below are the variables and their values from the data collected.

- cmplnt_fr_dt : Date of incident occurred
- addr_pct_cd : Precinct of incident occurred
- ofns_desc : Offense description
- pd_desc : Offense description (more granular)
- boro_nm : Borough name
- susp_age_group : Suspect's age group
- susp_race : Suspect's race
- susp_sex : Suspect's sex
- latitude : Latitude coordinate
- longitude : Longitude coordinate
- vic_age_group : Victim's age group
- vic_race : Victim's race
- vic_sex : Victim's sex

### Tweets Data:

We applied for an academic research access developer account which has a higher tweets cap and got approved by Twitter for our research purpose to collect tweets data through API as well. The get_all_tweets() function available from "academictwitteR" library was used to query NYC tweets data. In the query, we use hashtags and keywords that are highly related to hate speech. We have collected about 1.2 million of tweets through API. From the data return, we select below variables, which will be used for model building.

- id : Unique identification for each tweet
- created_at : Date the tweets created
- text : Text written by user
- source : Source used by user to tweet
- possibly_sensitive : True/False content is sensitive
- type : Type of coordinates
- coordinates : Longitude and latitude coordinates
- retweet_count : Count of retweet
- reply_count : Count of reply to the tweet
- like_count : Count of like to the tweet
- quote_count : Count of retweet with comments
- longitude : Longitude coordinate
- latitude : Latitude coordinate
- borough : Borough name

## Data Analysis

We will use a sentiment analysis approach to annotate tweets into positive (score>0), negative (score<0), and neutral (score=0) sentiments, in order to exclude non-hate related tweets. Following this approach we consider tweets that have negative scores are more likely to be hateful speech. All hateful tweets will then be joined to the hate crimes reported data by using date and borough for spatial and temporal analysis. Both hate crime and hateful tweets count will be normalized using each place's population. A widely used regional statistics method, spatio-temporal Bayesian modeling, will be used to analyze spatio-temporal patterns, determine any developing trends, and test our hypothesis. Below is the architecture of the proposed methodology.

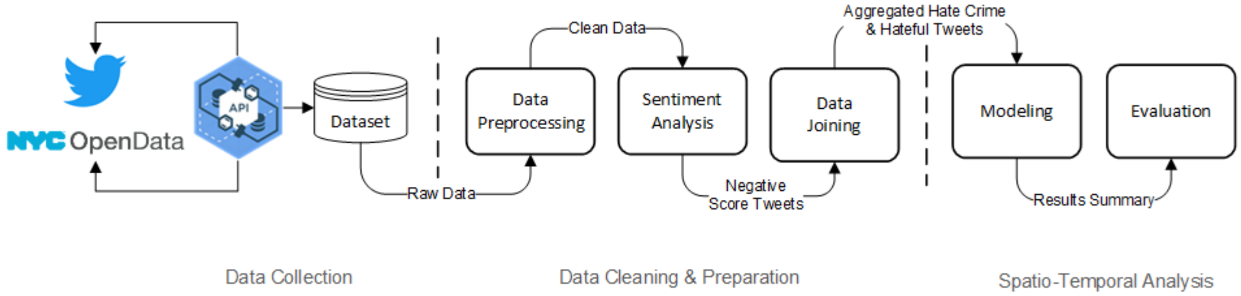

Figure 1: Architecture of the Proposed Methodology

## Model

In this study, we will develop two spatio-temporal Bayesian models. One obeys a Poisson distribution and another obeys Binomial distribution as below (Moraga, 2019; Hu et al., 2019).

$$Y_{ij} \sim Binomial(n_{ij}, p_{ij}) \tag{1}$$

where $Y_{ij}$ is the observed number of hate crime. $n_{ij}$ is the population of potential hate crimes occurred at i-th location in the j-th period of time. $p_{ij}$ is the hate crime rate occurred at i-th location in the j-th period of time. By taking the spatial and spatial-temporal correlation into account, model 1 above can be written as follows:

$$logit(p_{ij}) = \alpha + \beta X_i + u_i + s_i + \gamma t_j + \sigma_i t_j \tag{2}$$

where $logit(p_{ij})$ is the logit connectivity function for the hate crime rate. The spatio-temporal Bayesian model based on Poisson distribution is as below.

$$Y_{ij} \sim Poisson(E_{ij}\theta_{ij}) \tag{3}$$

where $E_{ij}$ is the expected number of hate crime occurred at i-th location in the j-th period of time. $\theta_{ij}$ is the relative risk of i-th location in the j-th period of time. Model 3 above can be written as follows:

$$log(\theta_{ij}) = \alpha + \beta X_i + u_i + s_i + \gamma t_j + \sigma_i t_j \tag{4}$$

The terms which are used in both model 2 and 4 are explained here. $\alpha$ is the intercept, $\beta$ is the coefficient of covariate X at i-th location. $\beta X_i$ is the fixed effect of the model. $u_i + s_i$ is the area random effect. $\gamma t_j$ represents the purely temporal term, and its coefficient, $\gamma$. $\sigma_i t_j$ represents the space and time interaction term.

In order to obtain the expected count of hate crime, $E_{ij}$ in the Poisson model, SpatialEpi package (Kim et al, 2018) will be used. Popular metrics such as DIC (*Deviance Information Criterion*), CPO (*Conditional Predictive Ordinate*), and WAIC (*Watanabe–Akaike Information Criterion*), will be used as a performance measurement to compare and measure how good a model is. All these metrics are particularly useful in Bayesian model selection problems and are available in the result summary generated from the INLA package (Integrated Nested Laplace Approximation) package (Rue et al, 2009).

## Scope and Limitations

Our scope is to analyze the tweet's data created in the five boroughs of New York City: Manhattan, Brooklyn, Queens, Bronx, and Staten Island. Some tweets may have videos embedded and that would be out of our capability to analyze the video's content. Socio-economic factors such as unemployment rate, income, education, etc., which may correlate with the hate crimes, are not considered in this study.

32,224 tweets out of the 1.2 million tweets have geographic coordinates available; thus, we may not be able to fully perform spatial analysis at the zip code level. Another limitation to our study is that assuming negative attitude tweets are hateful tweets is unlikely to be 100% reliable compared to human moderation. Same thing happens if we apply a trained binary classification model for labeling hundreds of thousands of negative attitude tweets into hateful or non-hateful tweets. With that being said, some of the hateful tweets could simply be expression of unhappiness, disagreement, sadness, disappointment, etc.

# Results

The crime data described as harassment and assault are considered as hate crimes, and tweets data which are negative sentiments are selected for analysis in this study. 920 thousand out of 1.2 million of tweets are identified as negative attitude tweets. Many of these negative tweets contain key terms commonly used in hateful messages. Figure 2 shows the top 100 most frequent mentioned words.

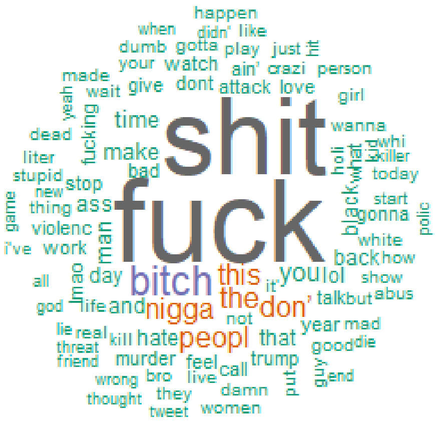

Figure 2: Word clouds of top 100 most often mentioned words in the negative attitude tweets

From the observed hate crime rate data, we found that the following minority groups; black, other races and female, are facing higher chances of being harassed and assaulted in most of the boroughs. Figure 3 and 4 below show these trends, and particularly the borough of Bronx has the highest hate crime rate in both race and sex compared to the other boroughs.

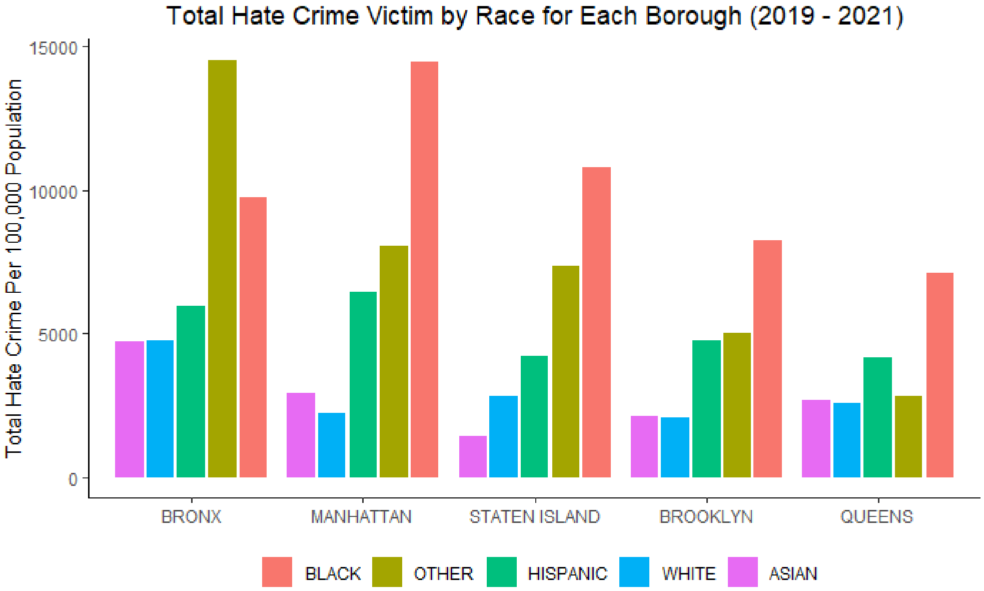

Figure 3: Total hate crime by race for each borough between 2019 and 2021

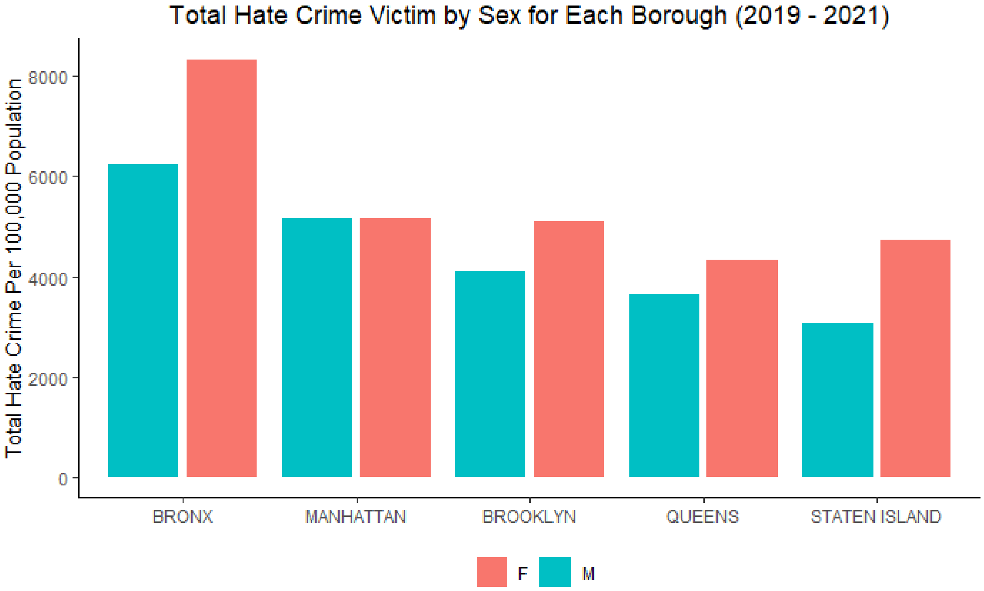

Figure 4: Total hate crime by sex for each borough between 2019 and 2021

The observed hate crime rate data and detected hateful tweets from year 2019 to 2021 are analyzed. In figure 5, we can see the trend of hate crimes is pretty stable over the 3 year period with a slightly downward trend in 2020, and then increase again in 2021. In contrast, the hateful tweets drop sharply, nearly half compared to previous periods after July 2020.

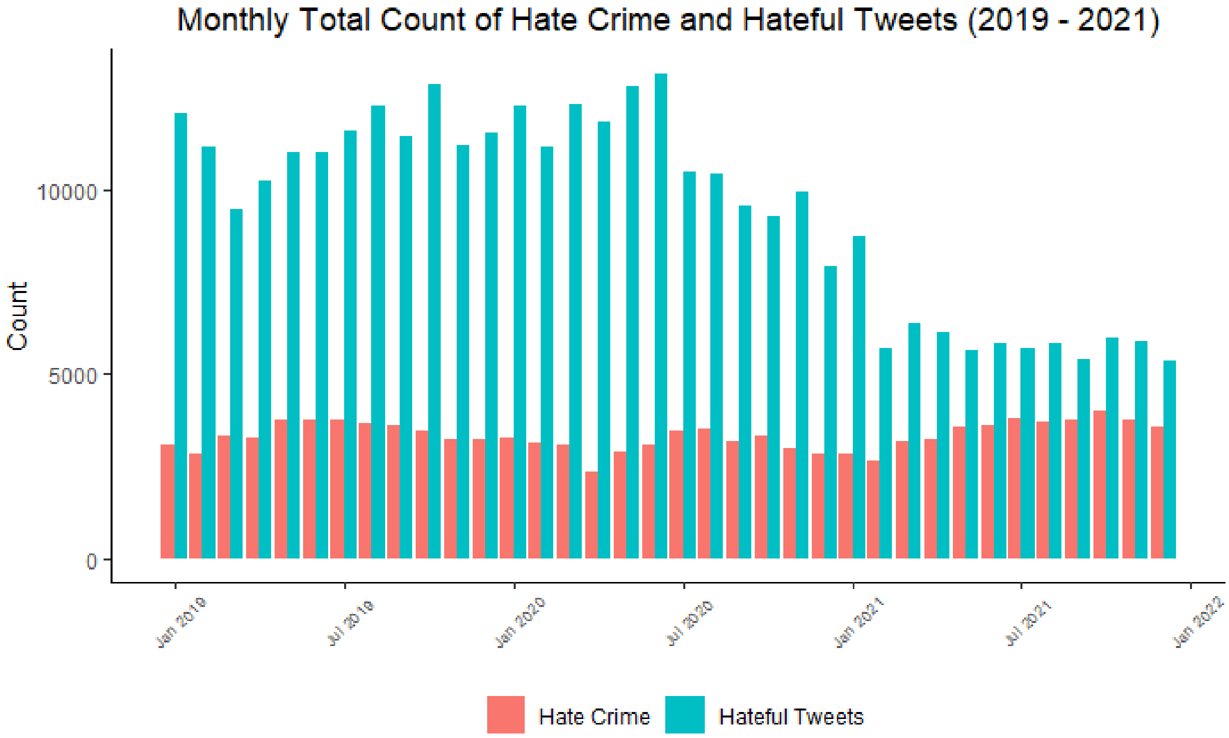

Figure 5: Monthly count of crime and negative attitude tweets between 2019 and 2021

Now, if we put hate crimes and hateful tweets count together and compare them for each borough, as shown in figure 6 and 7, Manhattan has the highest number of hate crimes and hateful tweets followed by Bronx, Brooklyn, Queens, and Staten Island.

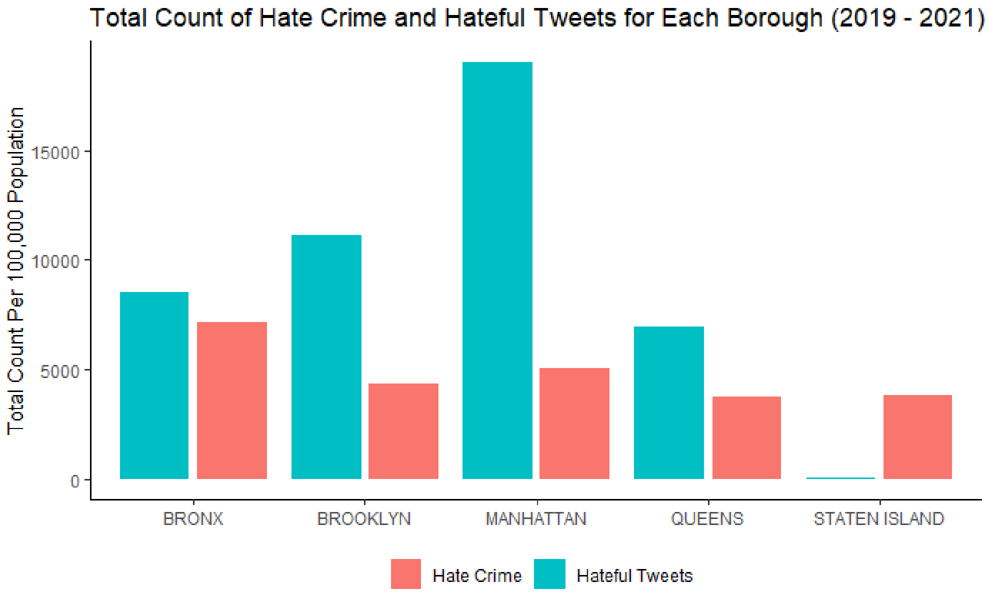

Figure 6: Total hate crime and hateful tweets for each borough between 2019 and 2021

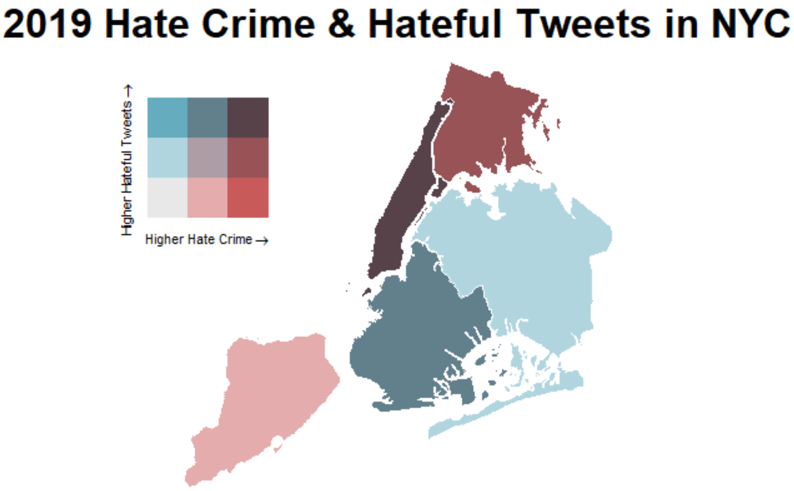

Figure 7: Bivariate chloropleth map for total hate crime and hateful tweets in 2019

Bayesian inference approximation method, INLA (Integrated Nested Laplace Approximation) package (Rue et al, 2009), is used to test the hypothesis of whether hateful tweets covariate has a statistically significant correlation to the hate crimes, and developing trend of relative risk of hate crimes occurred in the 5 boroughs.

Two types of Bayesian models are developed in this study, one is based on Poisson distribution and the other is based on Binomial distribution. Both models are considered to be mixed effects model, which means each model has both fixed and random effects. For model performance evaluation, DIC, WAIC, CPO metrics are used. A smaller value or closer to zero of these 3 metrics indicates a better model fit. Table 1 shows Poisson model has a better performance than Binomial model.

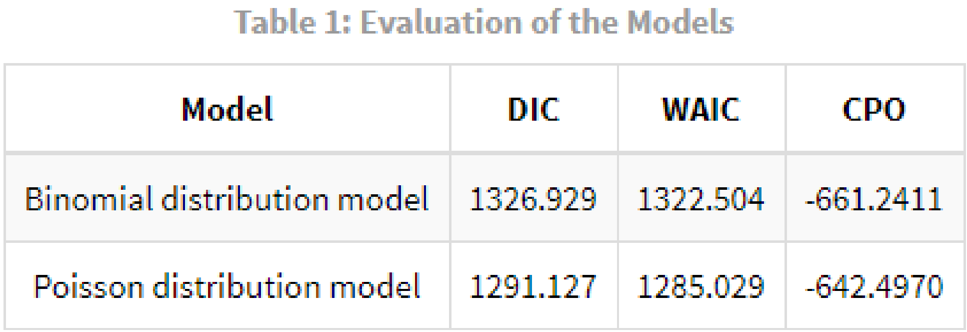

Table 1: Evaluation of the Models

| Model | DIC | WAIC | CPO |
|---|---|---|---|
| Binomial distribution model | 1326.929 | 1322.504 | -661.2411 |
| Poisson distribution model | 1291.127 | 1285.029 | -642.4970 |

Figure 8: Performance evaluation for Poisson and Binomial distribution models using DIC, WAIC, and CPO metrics

From the analysis results as shown in figure 9, we can observe that the hateful tweets covariate and Time ID variables coefficient at 95% confidence interval contains zero for both models. This means we are uncertain whether both of these variables have any effect on contributing to hate crimes and implies they could have a positive or negative effect on our target variable. Tables 2 and 3 show the coefficient at 95% confidence interval for Poisson models.

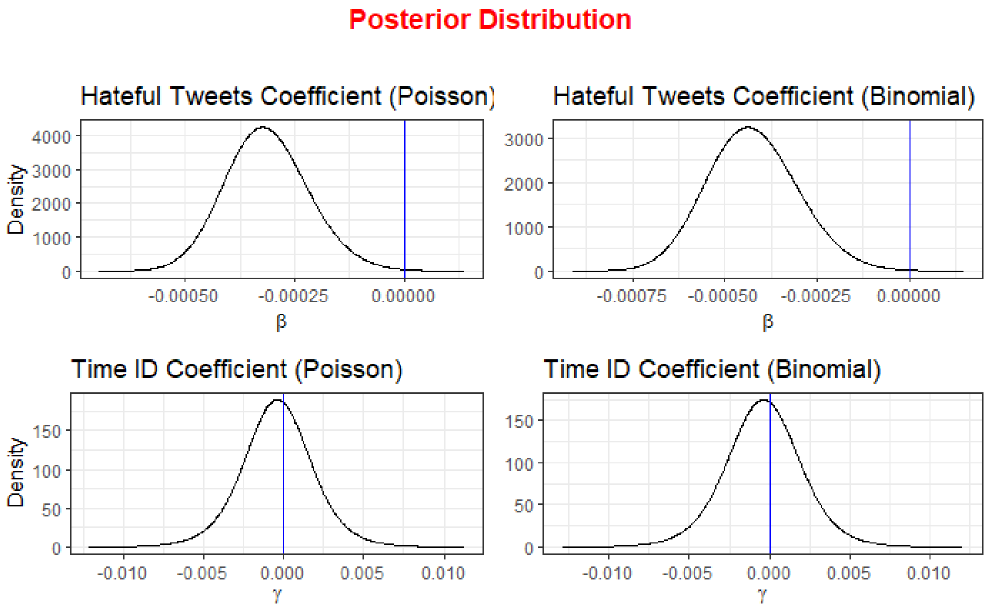

Figure 9: Posterior distribution of hateful tweets covariate and time id variables coefficient at 95% confidence intervals

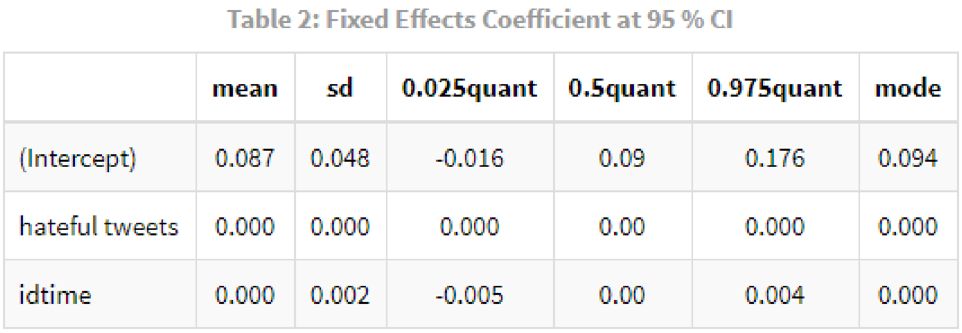

Table 2: Fixed Effects Coefficient at 95 % CI

| | mean | sd | 0.025quant | 0.5quant | 0.975quant | mode |
|---|---|---|---|---|---|---|
| (Intercept) | 0.087 | 0.048 | -0.016 | 0.09 | 0.176 | 0.094 |
| hateful tweets | 0.000 | 0.000 | 0.000 | 0.00 | 0.000 | 0.000 |
| idtime | 0.000 | 0.002 | -0.005 | 0.00 | 0.004 | 0.000 |

Figure 10: Fixed Effects Coefficient at 95% confidence intervals for Poisson model

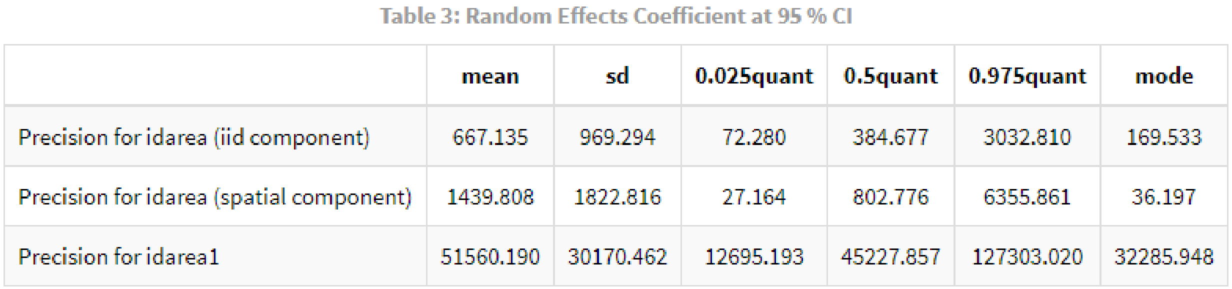

Table 3: Random Effects Coefficient at 95 % CI

| | mean | sd | 0.025quant | 0.5quant | 0.975quant | mode |
|---|---|---|---|---|---|---|
| Precision for idarea (iid component) | 667.135 | 969.294 | 72.280 | 384.677 | 3032.810 | 169.533 |
| Precision for idarea (spatial component) | 1439.808 | 1822.816 | 27.164 | 802.776 | 6355.861 | 36.197 |
| Precision for idarea1 | 51560.190 | 30170.462 | 12695.193 | 45227.857 | 127303.020 | 32285.948 |

Figure 11: Random Effects Coefficient at 95% confidence intervals for Poisson model

In addition to verifying the fixed variables' statistical significant, average posteriors of relative risks in each year for the 5 boroughs are also computed and developing trends are showed in figure 12. From the figure, we can see that the relative risk for Manhattan and Bronx has not changed over the 3 years period. Because the relative risk of Bronx is much greater than 1, hate crimes are more likely to occur in this area than its neighborhoods. The relative risk of Queens and Staten Island have become lower from 2019 to 2021 but in Brooklyn it has increased. This shows the developing trend of risk increase of hate crime occurring in Brooklyn areas over the 3 years period.

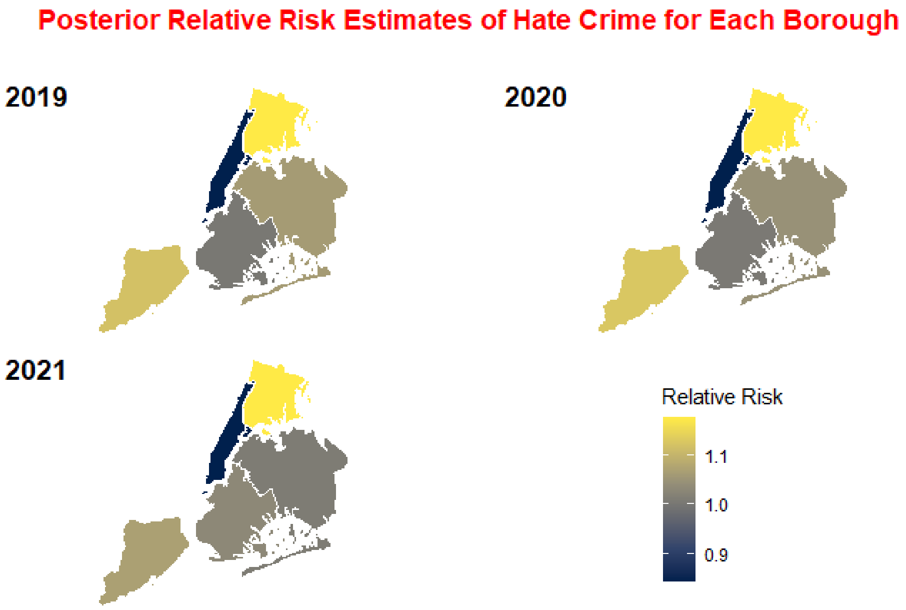
Posterior Relative Risk Estimates of Hate Crime for Each Borough
2019
2020
2021
Relative Risk
1.1
1.0
0.9

Figure 12: Average posteriors of relative risks in each year for the 5 boroughs

# Discussion

First of all, the Syuzhet package (Kim H., 2022) has done a great job in identifying the negative attitude tweets from the raw data we collected from Twitter in this study. Most of the top 100 terms from the negative tweets dataset are very offensive and we assume that many of these tweets are hate speech. The trend of total count of negative attitude tweets from 2019 and 2021, as shown in figure 4, is interesting to discuss here. Surprisingly, the total count of hateful tweets is starting to drop sharply after July 2020 compared to previous months since January 2019, which are pretty stable. The main factor that probably contributed to this is the fact that people were working from home. It could have been less stressful without commuting to work and in turn expressing more positive attitude toward their daily social media interaction.

The analysis results in above have verified that our hypothesis of hateful tweets covariate has no relationship with hate crimes occurred since its coefficient is statistically insignificant. This matches with the result found in these (Williams et al., 2019; Curiel et al., 2020) studies where there was no association found between online hate speech and hate crimes. But then the observed crime data have shown that hate crimes did happen mostly among minority groups, especially black, other races, and females.

In the formula we built to run in the INLA package, we put hateful tweets as covariate because we think hateful tweets can explain the hate crime risks, but the results don't show the association. There are of course other factors we have not taken into account for the model and we are modeling those factors using random effects. For instance, the spatial random effect as it's possible that two areas that are close to each other have similar risks.

A popular spatial model, BYM (Besag et al., 1991), is used in the model. This model will account for data that is spatially correlated, for example, neighboring areas could have a more similar trend than those areas that are not sharing a border with one another or are farther away. These patterns can be observed in figure 8. The Poisson model performs better than the Binomial model in this study because there is no fixed sample size being collected, but number of crimes happening each month is random. Because total sample size is random, an important characteristic of the Poisson distribution, Poisson model performs better in our case study (Howell D., 2003).

Although the fixed variables, except the intercept, have a coefficient that is not statistically significant, the spatial random effect components do play a big role in the model because we are observing that areas that are close to each other have similar risks as those seen in Brooklyn and Queens. The inverse of relative risk between Bronx and Manhattan could be due to the difference in socio-economic status among the residents in these two areas. If analysis can be done at the zip code level of NYC or at the county level for the New York state, we may be able to observe more examples of areas that share a border having similar risks.

# Conclusion

Nowadays, social media plays a critical role in influencing people's perception on everything we see, hear, and read in our daily lives. The goal of this study is to find if hateful tweets existing in Twitter have any association with the hate crimes in the New York City areas and whether minority groups are mostly affected. We found that hateful tweets do not have a statistically significant correlation with hate crimes, but based on the observed NYPD crime data we do see minority groups such as black, other races, and females being mostly the victims of hate crimes. The study definitely has room for improvements and for other considerations. Instead of focusing in the New York City areas, we can expand the scope to the whole state of New York and conduct analysis at the county level as most tweets don't share their exact location. This would make the spatial random effect more effective as it has many measured levels. Random effects variance estimates will be unstable if the grouping variable has too few levels. For future research, we can incorporate an unsupervised learning technique such as topic modeling using LDA (Latent Dirichlet Allocation) to further filter out non-hateful tweets. However, this requires a computer that has a much bigger RAM and better GPU than regular laptop or desktop to perform matrix factorization on large datasets, 920k rows of text in our case. Additionally, we could also leverage the same methodology used in this study to do hate crime prediction for each location.

# References

1. Kshirsagar V. (2019, December 24). *Detecting Hate tweets — Twitter Sentiment Analysis.* Towards Data Science. Retrieved from https://towardsdatascience.com/detecting-hate-tweets-twitter-sentiment-analysis-780d8a82d4f6.

2. Crabb et al. (2019, May 28). *Classifying Hate Speech: an overview.* Towards Data Science. Retrieved from https://towardsdatascience.com/classifying-hate-speech-an-overview-d307356b9eba.

3. Pang, G. (2022, March 4). *Deep Learning for Hate Speech Detection: A Large-scale Empirical Evaluation.* Towards Data Science. Retrieved from https://towardsdatascience.com/deep-learning-for-hate-speech-detection-a-large-scale-empirical-evaluation-92831ded6bb6.

4. Kim, H. (2022). *Sentiment Analysis: Limits and Progress of the Syuzhet Package and Its Lexicons.* Texas A&M University. Retrieved from http://www.digitalhumanities.org/dhq/vol/16/2/000612/000612.html.

5. Williams et al. (2019, July 23). *Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime.* The British Journal of Criminology. Retrieved from https://academic.oup.com/bjc/article/60/1/93/5537169.

6. Lee et al. (2022, January). *Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model.* ResearchGate. Retrieved from https://www.researchgate.net/publication/357916429_Racism_Detection_by_Analyzing_Differential_Opinions_Through_Sentiment_Analysis_of_Tweets_Using_Stacked_Ensemble_GCR-NN_Model.

7. Matsaki L. (2018, September 26). *To Break a Hate-Speech Detection Algorithm, Try 'Love'.* Wired. Retrieved from https://www.wired.com/story/break-hate-speech-algorithm-try-love/.

8. Rizwan et al. (2020, January). *Hate-Speech and Offensive Language Detection in Roman Urdu.* ACL Anthology. Retrieved from https://aclanthology.org/2020.emnlp-main.197.pdf.

9. Miró-Llinares et al. (2018, November 15). *Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments.* Springer Link. Retrieved from https://link.springer.com/article/10.1186/s40163-018-0089-1.

10. Curiel et al. (2020, April 02). *Crime and its fear in social media.* Nature. Retrieved from https://www.nature.com/articles/s41599-020-0430-7#Sec8.

11. Sandagiri et al. (2021, January 19). *Detecting Crime Related Twitter Posts using Artificial Neural Networks based Approach.* IEEE Xplore. Retrieved from https://ieeexplore.ieee.org/document/9325485.

12. Kumar, A. (2022, March 20). *Hate Speech Detection Using Machine Learning.* Vitalflux. Retrieved from https://vitalflux.com/hate-speech-detection-using-machine-learning/#:~:text=The%20techniques%20for%20dete

13. Alnazzawi, N. (2022, May 24). *Using Twitter to Detect Hate Crimes and Their Motivations: The HateMotiv Corpus.* MDPI. Retrieved from https://www.mdpi.com/2306-5729/7/6/69.

14. Kocon et al. (2021, June 03). *Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach.* ScienceDirect. Retrieved from https://www.sciencedirect.com/science/article/pii/S0306457321001333.

15. Shead, S. (2020, November 19). *Facebook claims A.I. now detects 94.7% of the hate speech that gets removed from its platform.* CNBC. Retrieved from https://www.cnbc.com/2020/11/19/facebook-says-ai-detects-94point7percent-of-hate-speech-removed-from-platform.html#:~:text=Facebook%20announced%20Thursday%20that%20artificial,and%20just%2024%25%20in%202017.

16. Williams, M. & Burnap, P. (2015, June 25 ). *Cyberhate on Social Media in the Aftermath of Woolwich: A Case Study in Computational Criminology and Big Data.* British Journal of Criminology. Retrieved from https://academic.oup.com/bjc/article/56/2/211/2462519

17. Hanes, E. & Machin, S. (2013, September). *Hate Crime in the Wake of Terror Attacks: Evidence from 7/7 and 9/11.* Journal of Contemporary Criminal Justice. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1038.9462&rep=rep1&type=pdf.

18. Hawdon et al. (2016, July 13). *Exposure to Online Hate in Four Nations: A Cross-National Consideration.* ResearchGate. Retrieved from https://www.researchgate.net/publication/303480309_Exposure_to_Online_Hate_in_Four_Nations_A_Cross-National_Consideration.

19. Starosta A. (2019, February 15). *CBuilding NLP Classifiers Cheaply With Transfer Learning and Weak Supervision. Medium.* Retrieved from https://medium.com/sculpt/a-technique-for-building-nlp-classifiers-efficiently-with-transfer-learning-and-weak-supervision-a8e2f21ca9c8.

20. Chetty N. & Alathur S. (2018, May 8). *Hate speech review in the context of online social networks.* Retrieved from https://www.sciencedirect.com/science/article/pii/S1359178917301064.

21. Miró-Llinares F. & Rodriguez-Sala J.J. (2016, July). *Cyber Hate Speech on Twitter: Analyzing Disruptive Events from Social Media to Build a Violent Communication and Hate Speech taxonomy.* ResearchGate. Retrieved from https://www.researchgate.net/publication/308487177_Cyber_hate_speech_on_twitter_Analyzing_disruptive_events_from_social_media_to_build_a_violent_communication_and_hate_speech_taxonomy.

22. Chen et al. (2015, June 8). *Crime prediction using Twitter sentiment and weather.* IEEE Xplore. Retrieved from https://ieeexplore.ieee.org/abstract/document/7117012/authors#authors.

23. Hashtagify. Search And Find The Best Twitter Hashtags. Available online: https://hashtagify.me/ (accessed on 15 March 2022).

24. Training Data for AI, ML with Human Empowered Automation. Cogit. Available online: https://www.cogitotech.com/about-us (accessed on 15 March 2022).

25. The HateMotiv Corpus. Kaggle. Available online: https://www.kaggle.com/datasets/nohaalnazzawi/the-hatemotiv-corpus (accessed on 3 October 2022).

26. NYPD Complaint Data Historic. NYC Open Data. Available online: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i (accessed on 17 October 2022).

27. 2020 Census Data. NYC Planning. Available online: https://www.nyc.gov/site/planning/planning-level/nyc-population/2020-census.page#2020-census-results (accessed on 29 October 2022).

28. Moraga P. (2019, November 25). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny.* Chapman & Hall/CRC Biostatistics Series. Retrieved from https://www.paulamoraga.com/book-geospatial/sec-arealdataexamplest.html#model-2.

29. NYCdata. Baruch College. Available online: https://www.baruch.cuny.edu/nycdata/population-geography/age_distribution.htm (accessed on 24 November 2022),

30. Hu et al. (2018, October 31). *Urban crime prediction based on spatio-temporal Bayesian model.* National Library of Medicine. Retrieved from https://ncbi.nlm.nih.gov/pmc/articles/PMC6209226/.

31. Rue et al. (2009, April 6). *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.* Journal of the Royal Statistical Society. Retrieved from https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2008.00700.x.

32. Besag et al. (1991). *Bayesian image restoration, with two applications in spatial statistics.* Springer. Retrieved from https://link.springer.com/article/10.1007/BF00116466.

33. Kim et al. (2018). *SpatialEpi: Methods and Data for Spatial Epidemiology.* Github. Retrieved from https://github.com/rudeboybert/SpatialEpi.

34. Howell D. (2003, January 3). *The Poisson and Binomial Distributions.* University of Vermont. Retrieved from https://www.uvm.edu/~statdhtx/StatPages/More_Stuff/PoissonBinomial/PoissonBinom.html.

35. Studley L. & Tucker E. (2022, April 16). *Hate crimes in New York City are up 76% this year compared to the same period in 2021.* CNN. Retrieved from https://www.cnn.com/2022/04/16/us/hate-crimes-rise-in-new-york-city.

# Appendices

## Appendix A1. Twitter Data Collection

```
# To run below codes for collecting tweets, first you'll need to create a .Renviron file
# by running set_bearer(), then store their bearer token in one line in the file like this
# TWITTER_BEARER = 'bearer token', close the file, and restart R session.

# Run this code to use API credentials saved in the .Renviron file connects to the Twitter
# to get tweets data.
get_bearer()

# Function to get all tweets from year 2020 to 2022 by borough
get_tweets <- function(borough) {
  get_tweets <- get_all_tweets(
    query = c("#hate crime", "#racist", "#racism", "#Islamophobia", "#Islamophobic",
              "#sexism", "#disability", "#transgender", "#antisemitism", "#misogyny",
              "#disabled", "slurs", "violent", "murder", "xenophobic", "nigga",
              "threat", "punching", "harrass", "assault", "black babies", "islam",
              "nazi", "genocide", "fuck", "shit", "discrimination", "bitch",
              "verbal abuse", "black babies", "violence", "white genocide", "graffiti",
              "explosion", "enslaved", "attack", "bomb", "harrassment",
              "killing minorities", "monsters", "black man", "killer", "shot dead",
              "hijab", "insult", "discriminate", "intimidation", "sexual assault",
              "black folks", "white supremacist", "homophobic", "burning a LGBT",
              "hatred", "muslim", "stabbing", "harrasing", "smashed", "punched",
              "choked", "knocking", "rape", "burned to death", "bastard", "stripped",
              "interracial", "slavery", "bomb", "torture", "abuse", "fucking",
              "slaughter", "brutally", "bully", "go back to your"),
    n = 1000000,
    start_tweets = "2019-01-01T10:00:00Z",
    end_tweets = "2021-12-31T10:00:00Z",
    country = "US",
    place = borough,
    lang = "en",
    page_n = 500,
    point_radius = c(-73.989121, 40.702944, 18)

  )

  return(get_tweets)
}

# Function to select variables, transform date and coordinates,
# extract list-like columns, longitude, and latitude into new columns
extract_cols <- function(tweets, borough) {
  all_tweets <- cbind(tweets %>%
                        select(id,
                               created_at,
                               text,
                               source,
                               possibly_sensitive),
                      as.data.frame(tweets$geo$coordinates),
```

```r
                      as.data.frame(tweets$public_metrics))


  all_tweets <- all_tweets %>%
    mutate(coordinates = as.character(coordinates))

  all_tweets <- all_tweets %>%
    mutate(created_at = as.Date(created_at))

  all_tweets <- all_tweets %>%
    mutate(longitude = parse_number(sub("[^-\\d]+", "", coordinates)))

  all_tweets <- all_tweets %>%
    mutate(latitude = parse_number(sub("[^, \\d]+", "", coordinates)))

  all_tweets$borough <- borough

  return(all_tweets)
}

# Get historical tweets for each borough
m_twt_raw <- get_tweets("Manhattan")
q_twt_raw <- get_tweets("Queens")
k_twt_raw <- get_tweets("Brooklyn")
x_twt_raw <- get_tweets("Bronx")
s_twt_raw <- get_tweets("Staten Island")

# Clean tweets data for each borough
m_twt_clean <- extract_cols(m_twt_raw, "MANHATTAN")
q_twt_clean <- extract_cols(q_twt_raw, "QUEENS")
k_twt_clean <- extract_cols(k_twt_raw, "BROOKLYN")
x_twt_clean <- extract_cols(x_twt_raw, "BRONX")
s_twt_clean <- extract_cols(s_twt_raw, "STATEN ISLAND")

# Read data from Github.
# As it takes time run above codes to collect NYC tweets data and also quota limitation
# per month to get # tweets data from Twitter, we uploaded the each borough data we
# collect from running above codes to a Github repo for convenient to read into
# R at any time.

# Github path
github <- "https://raw.githubusercontent.com/SieSiongWong/DATA-698/master/Data/"
m_twt <- read.csv(paste0(github,"manhattan_2019_2021.csv"), header=TRUE, sep=",")
q_twt <- read.csv(paste0(github,"queens_2019_2021.csv"), header=TRUE, sep=",")
k_twt <- read.csv(paste0(github,"brooklyn_2019_2021.csv"), header=TRUE, sep=",")
x_twt <- read.csv(paste0(github,"bronx_2019_2021.csv"), header=TRUE, sep=",")
s_twt <- read.csv(paste0(github,"staten_island_2019_2021.csv"), header=TRUE, sep=",")

# Combine borough data sets
tweets_data <- rbind(m_twt,q_twt,k_twt,x_twt,s_twt)
```

## Appendix A2. Crime Data Collection

```
# Collect crime data from year 2019 to 2021.
crime_data <-
  read.socrata("https://data.cityofnewyork.us/resource/qgea-i56i.json?$\
               select=cmplnt_fr_dt,addr_pct_cd,ofns_desc,pd_desc,\
               susp_age_group,boro_nm,susp_race,susp_sex,latitude,\
               longitude,vic_age_group,vic_race,\
               vic_sex&$where=cmplnt_fr_dt between \’2019\’ and \’2022\’")
```

## Appendix B. Data Preprocessing

```r
# Drop row which borough is NA
crime_data <- crime_data %>% drop_na(boro_nm)

# Filter hate related offensive corresponding description
hate_crime <- filter(crime_data, grepl(’ASSAULT|HARRASSMENT’, ofns_desc))

# Convert to date type
hate_crime$cmplnt_fr_dt <- as.Date(hate_crime$cmplnt_fr_dt)

# Count crime by year, month and borough
hate_crime_count_boro <- hate_crime %>%
  mutate_at(vars(cmplnt_fr_dt), funs(year, month)) %>%
  group_by(year, month, boro_nm) %>%
  summarise(total = n())

# Count total crime by year, month and borough
total_crime_count_boro <- crime_data %>%
  mutate(as.Date(cmplnt_fr_dt)) %>%
  mutate_at(vars(cmplnt_fr_dt), funs(year, month)) %>%
  group_by(year, month, boro_nm) %>%
  summarise(total_crime = n())

# Combine total crime and hate crime into a single dataframe
crime_count_boro <- merge(hate_crime_count_boro, total_crime_count_boro,
                          by = c("year", "month", "boro_nm"))
```

```r
# Remove meaningless characters and symbols in tweets text
tweets_data$text <- gsub("&amp","", tweets_data$text)
tweets_data$text <- gsub("<[^>]+>","", tweets_data$text)
tweets_data$text <- gsub("#\\w+","", tweets_data$text)
tweets_data$text <- gsub("@\\w+","", tweets_data$text)
tweets_data$text <- gsub("[[:punct:]]","", tweets_data$text)
tweets_data$text <- gsub("http\\w+","", tweets_data$text)
tweets_data$text <- gsub("[ \t]{2,}"," ", tweets_data$text)

# Get negative emotion score tweets
tweets_vector <- as.vector(tweets_data$text)
emotion_score <- get_sentiment(tweets_vector)
tweets_negative <- cbind(tweets_data, emotion_score) %>% filter(emotion_score < 0)

# Count tweets by year, month and borough
tweets_count_boro <- tweets_negative %>%
  mutate_at(vars(created_at), funs(year, month)) %>%
  rename(boro_nm = borough) %>%
  group_by(year, month, boro_nm) %>%
  summarise(total = n())
```

## Appendix C. Data Exploration

```r
# Tokenize the text and see frequency of words greater than
top_words <- tweets_negative %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  count(word, sort=TRUE) %>%
  filter(n>15000)

# Visualization of top words within the complete tweets data.
theme_set(theme_classic())
options(scipen = 999)

ggplot(top_words, aes(x=reorder(word,n), y=n))+
  geom_bar(stat="identity", width = 0.5, fill="tomato2")+
  xlab("Terms") + ylab("Count") + coord_flip() +
  theme(axis.text.x = element_text(angle=65, vjust=0.6, size=7))

# Count tweets which have coordinates by borough
tweets_negative %>%
  group_by(borough) %>%
  summarise(total = n(), geo_avail =  sum(coordinates!="NULL"))

# Select text and id column
tweets_corpus <- tweets_negative %>%
  select(X, text) %>%
  rename(doc_id = X) %>%
  mutate(doc_id = as.character(doc_id))

# Create a corpus
tweets_corpus <- VCorpus(DataframeSource(tweets_corpus))

# Remove all punctuation from the corpus
tweets_corpus  <- tm_map(tweets_corpus, removePunctuation)

# Remove all English stopwords from the corpus.
tweets_corpus <- tm_map(tweets_corpus, removeWords, stopwords("en"))
tweets_corpus <- tm_map(tweets_corpus, removeWords, stopwords("SMART"))

# Remove all number from the corpus.
tweets_corpus <- tm_map(tweets_corpus, removeNumbers)

# Strip extra white spaces in the corpus.
tweets_corpus <- tm_map(tweets_corpus, stripWhitespace)

# Stem words in the corpus.
tweets_corpus <- tm_map(tweets_corpus, stemDocument)

# Visualize word cloud
wordcloud(tweets_corpus,
          max.words = 100,
          random.order = FALSE,
          rot.per = 0.15,
          min.freq = 5,
          colors = brewer.pal(8, "Dark2"))

# 5 boroughs race population
boro_race_pop <- data.frame(boro_nm = c(’BROOKLYN’,
                                        ’MANHATTAN’,
                                        ’QUEENS’,
                                        ’BRONX’,
                                        ’STATEN ISLAND’),
                            HISPANIC = c(516426, 402640, 667861, 806463, 96960),
                            WHITE = c(968427, 793294, 549358, 130796, 277981),
                            BLACK = c(729696, 199592, 381375, 419393, 46835),
                            ASIAN = c(370776, 219624, 656583, 67766, 58753),
                            OTHER = c(37579+113170, 62989+16112, 66175+84112, 19866+28370,
                                      3900+11318)) %>%
  gather(vic_race,total_pop,HISPANIC:OTHER, factor_key=TRUE)

# 5 boroughs sex population
boro_sex_pop <- data.frame(boro_nm = c(’BROOKLYN’,
                                       ’MANHATTAN’,
                                       ’QUEENS’,
                                       ’BRONX’,
                                       ’STATEN ISLAND’),
                            F = c(1346912, 857428, 1159829, 748852, 244706),
                            M = c(1212991, 771278, 1094029, 669355, 231437)) %>%
  gather(vic_sex,total_pop,F:M, factor_key=TRUE)

# Visualize total count of crimes and hateful tweets
tweets_count_boro$yr_mth <- as.yearmon(with(tweets_count_boro,
                                            sprintf("%d-%02d", year, month)))
crime_count_boro$yr_mth <- as.yearmon(with(crime_count_boro,
                                           sprintf("%d-%02d", year, month)))
tweets_count_boro$name <- "Hateful Tweets"
crime_count_boro$name <- "Hate Crime"
rbind(tweets_count_boro, crime_count_boro) %>%
  ggplot(aes(x=yr_mth, y=total, fill=name))+
  geom_bar(position="dodge", stat="identity")+
  xlab("Date") + ylab("Count") +
  theme(axis.text.x = element_text(angle=45, vjust=0.6, size=7)) +
  theme(legend.title=element_blank(),
        legend.position="bottom",
        axis.title.x=element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  ggtitle("Monthly Total Count of Hate Crime and Hateful Tweets (2019 - 2021)")

# Visualize total count of victim race for each borough between year 2019 and 2021
hate_crime %>%
  group_by(boro_nm, vic_race) %>%
  summarise(total=n()) %>%
  spread(vic_race, total) %>%
  mutate(HISPANIC = ‘BLACK HISPANIC‘ + ‘WHITE HISPANIC‘,
         WHITE = WHITE,
         BLACK = BLACK,
         ASIAN = ‘ASIAN / PACIFIC ISLANDER‘,
         OTHER = ‘AMERICAN INDIAN/ALASKAN NATIVE‘+UNKNOWN) %>%
  select(HISPANIC, WHITE, BLACK, ASIAN, OTHER) %>%
  gather(vic_race,total, HISPANIC:OTHER, factor_key=TRUE) %>%
  left_join(., boro_race_pop, by=c(’boro_nm’, ’vic_race’)) %>%
  mutate(total=  total/total_pop*100000) %>%
  ggplot(aes(x = reorder(boro_nm, -total), y = total, fill = reorder(vic_race, -total))) +
  geom_col(position = position_dodge2(reverse = TRUE)) +
  theme(legend.title=element_blank(),
        legend.position="bottom",
        axis.title.x=element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  ggtitle("Total Hate Crime Victim by Race for Each Borough (2019 - 2021)") +
  ylab("Total Hate Crime Per 100,000 Population")

# Visualize total count of victim sex for each borough between year 2019 and 2021
hate_crime %>%
  group_by(boro_nm, vic_sex) %>%
  filter(!any(is.na(vic_sex))) %>%
  summarise(total=n()) %>%
  filter(vic_sex==’F’ | vic_sex==’M’) %>%
  left_join(., boro_sex_pop, by=c(’boro_nm’, ’vic_sex’)) %>%
  mutate(total=  total/total_pop*100000) %>%
  ggplot(aes(x = reorder(boro_nm, -total), y = total, fill = reorder(vic_sex, -total))) +
  geom_col(position = position_dodge2(reverse = TRUE)) +
  theme(legend.title=element_blank(),
        legend.position="bottom",
        axis.title.x=element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  ggtitle("Total Hate Crime Victim by Sex for Each Borough (2019 - 2021)") +
  ylab("Total Hate Crime Per 100,000 Population")

# Visualize total count of victim age group for each borough between year 2019 and 2021
hate_crime %>% group_by(boro_nm, vic_age_group) %>% summarise(total=n()) %>%
  filter(vic_age_group %in% c(’<18’,’18-24’,’25-44’,’45-64’,’65+’,’unknown’)) %>%
  ggplot(aes(x = boro_nm, y = total, fill = reorder(vic_age_group, -total))) +
  geom_col(position = position_dodge2(reverse = TRUE)) +
  theme(legend.title=element_blank(),
        legend.position="bottom",
        axis.title.x=element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  ggtitle("Total Hate Crime by Victim Age Group for Each Borough (2019 - 2021)") +
  ylab("Count")
```

## Appendix D. Data Preparation

```
# 5 boroughs population
boro_pop <- data.frame(boro_nm = c(’BROOKLYN’,
                                   ’MANHATTAN’,
                                   ’QUEENS’,
                                   ’BRONX’,
                                   ’STATEN ISLAND’),
                       population = c(2736074,1694251,2405464,1472654,495747))

# Combine hate tweets and crime data and then normalize by population
crime_tweets_combine_mth <- merge(crime_count_boro,
                                  tweets_count_boro,
                                  by = c("yr_mth", "year", "month", "boro_nm")) %>%
  select(yr_mth, year, month, boro_nm, total_crime, total.x, total.y) %>%
  rename(total_hate_crime = total.x, total_hate_tweets = total.y) %>%
  left_join(., boro_pop, by=’boro_nm’) %>%
  mutate(total_crime = total_crime/population*100000,
         total_hate_crime = total_hate_crime/population*100000,
         total_hate_tweets = total_hate_tweets/population*100000) %>%
  arrange(yr_mth)

# Normalized hate tweets and crime data group by year for chloropleth map
crime_tweets_combine_yr <- crime_tweets_combine_mth %>%
  group_by(year, boro_nm) %>%
  summarise(total_crime = sum(total_crime),
            total_hate_crime = sum(total_hate_crime),
            total_hate_tweets = sum(total_hate_tweets))

# Visualize total count by borough from 2019 to 2021
crime_tweets_combine_boro <- crime_tweets_combine_yr %>%
  group_by(boro_nm) %>%
  summarise(total_crime = sum(total_crime),
            total_hate_crime = sum(total_hate_crime),
            total_hate_tweets = sum(total_hate_tweets))

# Convert to long form to plot
crime_tweets_combine_boro %>% gather(condition,
                                     total,
                                     total_crime,
                                     total_hate_crime,
                                     total_hate_tweets,
                                     factor_key=TRUE) %>%
  filter(condition == ’total_hate_crime’ | condition == ’total_hate_tweets’) %>%
  mutate(condition = ifelse(condition=="total_hate_crime",
                            "Hate Crime",
                            "Hateful Tweets")) %>%
  ggplot(aes(x = boro_nm, y = total, fill = condition)) +
  geom_col(position = position_dodge2(reverse = TRUE)) +
  theme(legend.title=element_blank(),
        legend.position="bottom",
        axis.title.x=element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  ggtitle("Total Count of Hate Crime and Hateful Tweets for Each Borough (2019 - 2021)") +
  ylab("Total Count Per 100,000 Population")


# Bivariate chloropleth map for year 2019, 2020, and 2021
# Note: borough name and geometry columns must be in the first and second column in order

boro_boudaries <- nyc_boundaries(geography = "borough", add_acs_data = T) %>%
  st_transform(2263) %>%
  select(borough_name, geometry)

# Function to plot bivariate chloropleth map for hate crime vs hate tweets
plot_chloro_map_tw <- function(df, year) {

data <- df %>%
  ungroup() %>%
  filter(year==year) %>%
  rename(borough_name = boro_nm) %>%
  mutate(borough_name = str_to_title(borough_name)) %>%
  select(borough_name, total_hate_crime, total_hate_tweets) %>%
  left_join(boro_boudaries, ., by = c("borough_name")) %>%
  bi_class(x = total_hate_crime, y = total_hate_tweets, style = "quantile", dim = 3)

map <- ggplot() +
  geom_sf(data = data,
          mapping = aes(fill = bi_class),
          color = "white",
          size = 0.1,
          show.legend = FALSE) +
  bi_scale_fill(pal = "GrPink", dim = 3) +
  labs(
    subtitle = paste0(year, " Hate Crime & Hateful Tweets in NYC")) +
  bi_theme()  +
  theme(plot.subtitle=element_text(size=18,
                                   hjust=0.5,
                                   color="black"))

legend <- bi_legend(pal = "GrPink",
                    dim = 3,
                    xlab = "Higher Hate Crime",
                    ylab = "Higher Hateful Tweets",
                    size = 8)

final_plot <- ggdraw() +
  draw_plot(map, 0, 0, 1, 1) +
  draw_plot(legend, 0.1, 0.5, 0.35, 0.35)

return(final_plot)

}

# Plot all 3 years chloropleth maps
plot_chloro_map_tw(crime_tweets_combine_yr, 2019)
plot_chloro_map_tw(crime_tweets_combine_yr, 2020)
plot_chloro_map_tw(crime_tweets_combine_yr, 2021)
```

## Appendix E. Build Models

```
# Create a neighborhood matrix to define the spatial random effect
nb <- poly2nb(boro_boudaries)
nb2INLA("map.adj", nb)

# Read the matrix map file, store to a g object for later use in specifying the spatial
# structure in the model
g <- inla.read.graph(filename = "map.adj")

# Create the index vector for the boroughs and date used for specifying the random effects
# Index vector of area is required for the two random effects
d <- crime_tweets_combine_mth
d$idarea <- as.numeric(as.factor(d$boro_nm))
d$yr_mth <- as.yearmon(with(d, sprintf("%d-%02d", year, month)))
d$idarea1 <- d$idarea

# Copy total hate crime to a new variable, Y and required no decimal by INLA algorithm
d$Y <- round(d$total_hate_crime, 0)

# Calculate expected crime cases
# Number of strata is equal to 1 for our study because our collected data
# is not stratified random sampling
d <- d[order(d$boro_nm,d$yr_mth), ]
E <- expected(population = d$total_crime, cases=d$total_hate_crime, n.strata=1)
d$E <- E

# Reorder date in ascending order and then assign incremental time id for each unique date
d <- d[order(d$yr_mth), ]
d$idtime <- cumsum(!duplicated(d$yr_mth))

# Create formula to run in INLA function
# INLA package requires R version 4.1 or above to run
formula <- Y ~ total_hate_tweets + f(idarea, model = "bym", graph = g, scale.model=TRUE) +
  f(idarea1, idtime, model = "iid") + idtime

# Poisson distribution model
res_p <- inla(formula,
            family = "poisson",
            data = d,
            E = E,
            control.predictor = list(compute = TRUE),
            control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE),
            verbose = FALSE)

# Binomial distribution model
res_b <- inla(formula,
            family = "binomial", Ntrials=total_crime,
            data = d,
            control.family=list(link=’logit’),
            control.predictor = list(link=1,compute = TRUE),
            control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE),
            verbose = FALSE)
```

## Appendix F. Model Evaluation

```r
# Get summary for poisson and binomial models
summary(res_p)
summary(res_b)

# Plot the posterior distribution of the total hate tweets coefficient
# for poisson distribution model
options(scipen = 999)
marginal_tw_p <- inla.smarginal(res_p$marginals.fixed$total_hate_tweets)
marginal_tw_p <- data.frame(marginal_tw_p)
marginal_tw_p_plot <- ggplot(marginal_tw_p, aes(x = x, y = y)) +
  geom_line() +
  labs(x = expression(beta), y = "Density") +
  geom_vline(xintercept = 0, col = "blue") +
  theme_bw()  +
  ggtitle("Hateful Tweets Coefficient (Poisson)")

marginal_tm_p <- inla.smarginal(res_p$marginals.fixed$idtime)
marginal_tm_p <- data.frame(marginal_tm_p)
marginal_tm_p_plot <-ggplot(marginal_tm_p, aes(x = x, y = y)) +
  geom_line() +
  labs(x = expression(gamma), y = "Density") +
  geom_vline(xintercept = 0, col = "blue") +
  theme_bw() +
  ggtitle("Time ID Coefficient (Poisson)")

# Plot the posterior distribution of the total hate tweets coefficient
# for binomial distribution model
marginal_tw_b <- inla.smarginal(res_b$marginals.fixed$total_hate_tweets)
marginal_tw_b <- data.frame(marginal_tw_b)
marginal_tw_b_plot <- ggplot(marginal_tw_b, aes(x = x, y = y)) +
  geom_line() +
  labs(x = expression(beta), y = "Density") +
  geom_vline(xintercept = 0, col = "blue") +
  theme_bw() +
  ggtitle("Hateful Tweets Coefficient (Binomial)")

marginal_tm_b <- inla.smarginal(res_b$marginals.fixed$idtime)
marginal_tm_b <- data.frame(marginal_tm_b)
marginal_tm_b_plot <- ggplot(marginal_tm_b, aes(x = x, y = y)) +
  geom_line() +
  labs(x = expression(gamma), y = "Density") +
  geom_vline(xintercept = 0, col = "blue") +
  theme_bw() +
  ggtitle("Time ID Coefficient (Binomial)")

# Arrange all plots into a single figure
annotate_figure(ggarrange(marginal_tw_p_plot,
                          marginal_tw_b_plot + rremove("y.title"),
                          marginal_tm_p_plot,
                          marginal_tm_b_plot + rremove("y.title"),
                           ncol = 2, nrow = 2) ,
          top = text_grob("Posterior Distribution \n",
                          color = "red",
                          face = "bold",
                          size = 14))

# Get DIC, WIC value for both models into a table
kable(data.frame(Model = c(’Binomial distribution model’, ’Poisson distribution model’),
                 DIC = c(res_b$dic$dic, res_p$dic$dic),
                 WAIC = c(res_b$waic$waic, res_p$waic$waic),
                 CPO = c(sum(log(res_b$cpo$cpo)), sum(log(res_p$cpo$cpo)))),
      format="html",
      align = "c",
      table.attr = "style = \"color: black;\"",
      caption = "<center><b>Table 1: Evaluation of the Models<b></center>") %>%
  kable_styling(bootstrap_options = c("bordered", "striped"), full_width = F) %>%
  row_spec(row = 0, bold = TRUE)

# Get fixed effects coefficient CI into a table
fixed_tbl <- data.frame(summary(res_p)[3])
colnames(fixed_tbl) = gsub("fixed.", "", colnames(fixed_tbl))
rownames(fixed_tbl)[rownames(fixed_tbl) == "total_hate_tweets"] <- "hateful tweets"
fixed_tbl %>%
  select(-kld) %>%
  kable(format="html",
        align = "c",
        table.attr = "style = \"color: black;\"",
        caption = "<center><b>Table 2: \
        Fixed Effects Coefficient at 95 % CI <b></center>")  %>%
  kable_styling(bootstrap_options = c("bordered", "striped"), full_width = F) %>%
  row_spec(row = 0, bold = TRUE)

# Get random effects variables coefficient CI into a table
hyperpar_tbl <- data.frame(summary(res_p)[4])
colnames(hyperpar_tbl) = gsub("hyperpar.", "", colnames(hyperpar_tbl))
hyperpar_tbl %>% kable(format="html",
      align = "c",
      table.attr = "style = \"color: black;\"",
      caption = "<center><b>Table 3: \
      Random Effects Coefficient at 95 % CI <b></center>")  %>%
  kable_styling(bootstrap_options = c("bordered", "striped"), full_width = F) %>%
  row_spec(row = 0, bold = TRUE)

# Merge relative risk value to the dataset
d$RR_P <- res_p$summary.fitted.values[, "mean"]
d$RR_B <- res_b$summary.fitted.values[, "mean"]

# Group by year and borough and calculate average of relative risk for each year
d2 <- d %>%
  group_by(year, boro_nm) %>%
  summarise(total_crime = sum(total_crime),
            total_hate_crime = sum(total_hate_crime),
            total_hate_tweets = sum(total_hate_tweets),
            avg_rr_p = mean(RR_P),
```

```r
            avg_rr_b = mean(RR_B))

# Plot chloropleth map for relative risk data from poisson distribution model
map1 <- d2 %>%
  ungroup() %>%
  filter(year==2019) %>%
  rename(borough_name = boro_nm) %>%
  mutate(borough_name = str_to_title(borough_name)) %>%
  left_join(boro_boudaries, ., by = c("borough_name")) %>%
  ggplot() +
    geom_sf(aes(fill = avg_rr_p),
            color = "white",
            lwd = 0.2) +
    scale_fill_viridis_c(
      name = "Relative Risk",
      option = "cividis") +
  theme_void() +
  theme(plot.title.position = ’plot’,
        plot.title = element_text(hjust = 0.5)) +
  theme(panel.grid = element_line(color = "transparent"))

map2 <- d2 %>%
  ungroup() %>%
  filter(year==2020) %>%
  rename(borough_name = boro_nm) %>%
  mutate(borough_name = str_to_title(borough_name)) %>%
  left_join(boro_boudaries, ., by = c("borough_name")) %>%
  ggplot() +
    geom_sf(aes(fill = avg_rr_p),
            color = "white",
            lwd = 0.2) +
    scale_fill_viridis_c(
      name = "Relative Risk",
      option = "cividis") +
  theme_void() +
  theme(plot.title.position = ’plot’,
        plot.title = element_text(hjust = 0.5)) +
  theme(panel.grid = element_line(color = "transparent"))


map3 <- d2 %>%
  ungroup() %>%
  filter(year==2021) %>%
  rename(borough_name = boro_nm) %>%
  mutate(borough_name = str_to_title(borough_name)) %>%
  left_join(boro_boudaries, ., by = c("borough_name")) %>%
  ggplot() +
    geom_sf(aes(fill = avg_rr_p),
            color = "white",
            lwd = 0.2) +
    scale_fill_viridis_c(
      name = "Relative Risk",
      option = "cividis") +
  theme_void() +
  theme(plot.title.position = ’plot’,
        plot.title = element_text(hjust = 0.5)) +
  theme(panel.grid = element_line(color = "transparent"))

plot_legend <- get_legend(map3)

annotate_figure(ggarrange(map1 + theme(legend.position = "none"),
                          map2 + theme(legend.position = "none"),
                          map3 + theme(legend.position = "none"),
                          plot_legend,
                          ncol = 2, nrow = 2,
                          labels = c("2019", "2020", "2021")) ,
          top = text_grob("Posterior Relative Risk Estimates of Hate Crime for Each Borough \n",
                          color = "red",
                          face = "bold",
                          size = 14))
```