Investigating the Influences and Dynamics: A Trend Perspective on COVID-19 Booster Uptake in the United States.

Paper Choice: Option 3: COVID Campus Reopening Decision

Jiatai Zhang

Jiayi Zhang

Course: CS169-Statistical Bioinformatics in R

Tufts university

Abstract

This research focused on the trends in COVID-19 vaccine booster uptake in the United States. It replicated and built upon the model from the paper ”Modeling for COVID-19 College Reopening Decisions: Cornell, a Case Study,” which utilized epidemiological models to predict pandemic progression, aiding in safer reopening decisions, as exemplified by Cornell University. The study ”Investigating the Influences and Dynamics: A Trend Perspective on COVID-19 Booster Uptake in the the United States,” delved into the impact of various parameters on booster vaccine doses, examining the relationship between long-term infection prevention and people's willingness to receive cumulative vaccine doses from 2023 to 2026. It employed logistic regression and ARIMA models to analyze statistical significance and explored trends related to age, race, and geography in vaccine uptake, as well as people's intentions. The model's accuracy was approximately 0.8, effectively reflecting the variety in trends.

Introduction :

During the COVID-19 pandemic, the issue of resuming operations at U.S. colleges and universities garnered significant attention. The research paper, "Modeling for COVID-19 College Reopening Decisions: Cornell, a Case Study," addressed the use of epidemiological models to predict the progression of the COVID-19 pandemic in order to make safer college reopening decisions. It demonstrated this framework's application at Cornell University and assesses the efficacy of various intervention measures. According to the paper, the university intended to control on-campus breakouts by increasing vaccination rates and increasing monitoring frequency. Small differences in behavioral and biological parameters can cause huge differences in predicted case counts [1]. However, the uncertainty in parameters could also impact outcomes, potentially resulting in an increase in actual infection rates.

While the paper considered parameter uncertainty, it appeared that the parameters it took into account were not comprehensive in light of today's data. For instance, the status of booster vaccine doses, the continued availability of online classes for the fall of 2021, and the presence of international students opting for gap years all influence the predictive data presented in the paper. Our research incorporated a wide range of parameters to predict infection rates and compared them. This research further investigated people's willingness to get a fourth dose of the vaccine in the current environment, where the pandemic was under control but still present.

Literature Review:

Early 2020 signified a particularly tumultuous period during the COVID-19 pandemic, with approximately 86,600 cases of SARS-CoV-2 recorded within the initial months of that year, according to the Centers for Disease Control and Prevention (CDC)'s report [2]. Several studies indicated that patients with chronic and degenerative diseases were the vulnerable group, facing life-threatening symptoms from SARS-CoV-2. Under the circumstances concerning the unknown risks of the SARS-CoV-2 virus and the lack of comprehensive patient data, it was imperative to prevent the virus from spreading within the population effectively. On December 11, 2020, the Food and Drug Administration approved the Pfizer-BioNTech COVID-19 vaccine in the U.S., alongside the Moderna, Janssen (Johnson and Johnson), and Novavax bivalent vaccines. Presently, 68.5% of the world's population had received at least one dose of an approved vaccine [3]. However, one of the significant characteristics of the virus was its adaptability to the environment. The spike protein, a surface protein that allows the virus to penetrate host cells, is constantly changing. Mutations that altered the conformation of the spike protein could enable the virus to escape detection by immune systems [4]. Vaccines had continuously evolved to respond to various virus variants, such as Alpha, Beta, Gamma, Delta, and Omicron, which appeared in 2020 and 2021. To prevent infections from SARS-CoV-2 virus variants, vaccines against variants were gradually being recommended. However, several studies indicated that the average vaccine validity rate was decreasing which was from 94% for Alpha and Beta, to 76% for Delta, and down to 34% for Omicron [5]. This decreasing effectiveness raised concerns about the vaccine's utility and the public's willingness to regularly receive vaccinations.

Michael Jordan, University Infection Control Director, announced in a Jan. 5 email to the Tufts community that the bivalent COVID-19 booster vaccine was no longer mandatory for all university personnel and students. Patrick Collins, Executive Director of Media Relations at Tufts, elaborated, "It became increasingly clear over the fall semester that, after nearly three years of the pandemic, we needed to try a new strategy to achieve this goal." [6]. He emphasized that the continued mandate for the bivalent booster did not achieve the desired effects, prompting a re-assessment of its practicality in enforcement. Even though the bivalent booster was no longer required at Tufts, the university continued to track students' uploaded vaccination files, because Tufts University agreed that a high rate of vaccination was the best way to protect Tufts communities, which was consistent with Cornell University's case study results. This non-mandatory approach toward booster vaccinations can be attributed to potential risks, underlined by the relatively recent emergence of the virus and a lack of long-term efficiency data for the vaccine, as well as challenges in tracking recipient feedback.

Shira Doron, a Tufts Medical Center physician and epidemiologist, and Monica Gandhi, a professor of medicine at the University of California – San Francisco, mentioned that a determining whether all population groups should receive booster shots requires clear knowledge [7].They noted that age was significant in risk assessment following a booster. Supporting this, research from the Journal of Medical Ethics indicated that within a 6-month period, 31,207 to 42,836 young adults (aged 18–29) who received a third mRNA vaccine could expect between 1,430 and 4,626 cases to experience side effects, with 18.5 serious cases from mRNA vaccine requiring hospitalization and a potential diagnosis of myopericarditis [8].Early vaccination was utilized to minimize transmissions, whereas, two years later, most individuals had already been infected with COVID. With vaccines proving effective against serious illness for the higher-risk population but not offering high protection against future infection, the discussion regarding the importance of boosters for young people, especially university students, becomes crucial.

Furthermore, the research, "Simulating COVID-19 Classroom Transmission on a University Campus," studied SARS-CoV-2 transmission in classrooms, taking into

account situations with vaccinated and unvaccinated individuals [9]. Comparing with findings from Cornell University, which showed little to no interaction leading to SARS-CoV-2 transmission between employees and students under the hypothesis that they were fully vaccinated, the research simulation model first confirmed that increases in the parameters of mask usage and vaccination rate led to a significant reduction in transmission risk on campus [1]. However, when altering the mask usage parameter (with or without a mask), transmission variance was 64% within a 95% confidence interval. Considering the highest mask usage parameters, the Delta variants displayed a critical vaccination threshold of 93%. Under the circumstance of decreasing efficacy of variant vaccines, mask usage should be considered as a more sensitive parameter for transmission on campus.

Vaccine concerns could be defined as the collective patterns of attitudes and beliefs toward a vaccine among individuals who refused vaccination [10]. This paper pointed out the diversity in attitudes towards the COVID-19 vaccine among different racial groups in the United States. The paper ”Modeling for COVID-19 college reopening decisions: Cornell, a case study” lacked references to this particular data. Studying the attitudes of different racial groups toward vaccines could have persisted a better understanding to address issues of vaccination acceptance and health disparities, aiming to safeguard a greater segment of society from the threat of infectious diseases. According to this study, White and Asian Americans were more ready to get vaccinated, while African Americans and Hispanic Americans might have been more hesitant. This hesitation could have been attributed to socioeconomic considerations, historical and cultural origins, and social influences. Subsequent studies were suggested to look more closely at this part of the data to discuss attitudes and changes in attitudes toward vaccines by race.

The research investigated parameters, including age, race, and gender of populations, as well as social attitudes toward vaccines. It was conducted to determine which parameter was most significant in vaccine acceptance for SARS-CoV-2 variants from 2020 to 2023. The paper ’Modeling for COVID-19 College Reopening Decisions: Cornell, a Case Study’ lacked references to this diversity. Understanding these

attitudes was crucial in addressing vaccination acceptance and health disparities, with further studies recommended to delve deeper into these attitudes and their changes over time. The results of the current research could be used to explore whether booster vaccines should have been mandated for the university population as well as the general public in New York. Data were collected from the CDC and other public, up-to-date GitHub platforms. Machine learning models, including logistic regression and ARSRM time series, were utilized to analyze the results and determine the appropriate weights to be assigned to the parameters, observing the trend of vaccine acceptance over time periods.

Data Description: Cornell Simulation

The study "Modeling for COVID-19 College Reopening Decisions: Cornell, A Case Study" had examined cases from the year 2020 in New York. The datasets "covid_county_census_data" and "county_logalpha" both contained common columns such as date, county_name, province_state, death_count, and population. These datasets were merged using the "county_name" column and then filtered to only include "New York" records. As the research had primarily focused on the exploration period of the pandemic, there was a noticeable spike in death counts in March 2020, spanning the observation period from January 2020 to April 2020 (Figure1).

On January 21st, 2020, the CDC confirmed the first COVID case in the United States. Following this, universities and communities initiated emergency policies to devise solutions. On March 17th, 2020, the University of Minnesota was the first to begin testing hydroxychloroquine [11]. This effort was later joined by other universities, including Cornell University. The spike in death counts, attributable to diagnosed SARS-CoV-2 cases, was taken into account statistically. This made the urgency for a vaccine to prevent further infections even more paramount.

The simulation of individual Disease Progression had depicted the disease's progression through various stages for infected individuals across five age groups (Table1). The simulation had distinguished between asymptomatic and symptomatic individuals, with the symptomatic group self-reporting daily. When examining the distribution between age and severity level of asymptomatic individuals, $P\text{-}Sev\text{-}Age$ had represented the total probability within each age group, segmented into severity levels using CDC numbers for hospitalization and ICU rates in the nominal planning

scenario. This scenario had a higher density in the 18-44 age group (Figure2). The probability density function (PDF) across ages showed a higher occurrence of asymptomatic cases within this age group. The cumulative density function also indicated a gradual increase, reaching close to 1 in 55-59 and above groups (Figure3).

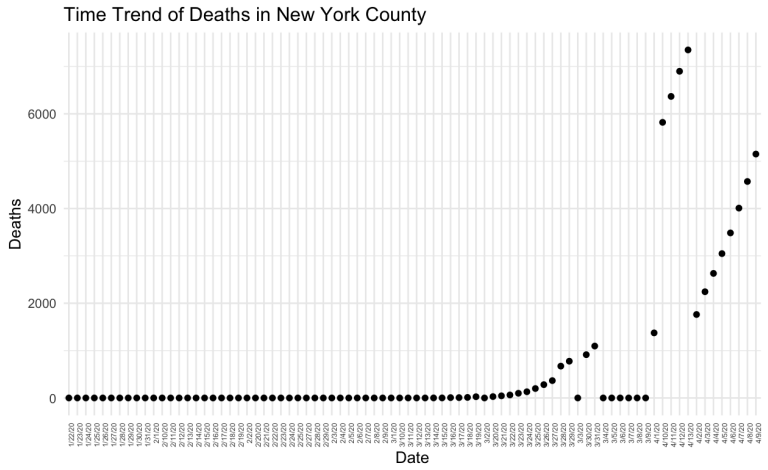

Figure 1: The time trend of deaths in New York County.

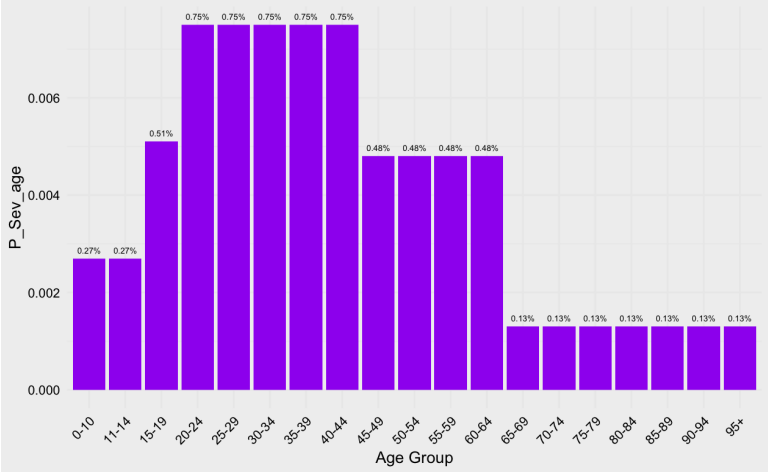

Figure 2: Distribution of total probability across severity levels for age groups (before re-grouping into five categories).

The dataset ”case-hosp-death” recorded cases for the first five days of March. To predict the infection trend, the cumsum() function was used to compute the cumulative sum of the CASE_COUNT for each day, tracking the total number of cases over those five days. The exponential trend of deaths shown in Figure 1 suggested that the rate of infection could vary during outbreak and surge periods. Computing t-square provided a quadratic regression that fit the data more accurately. The changes in cases over time

were displayed in Figure 4.

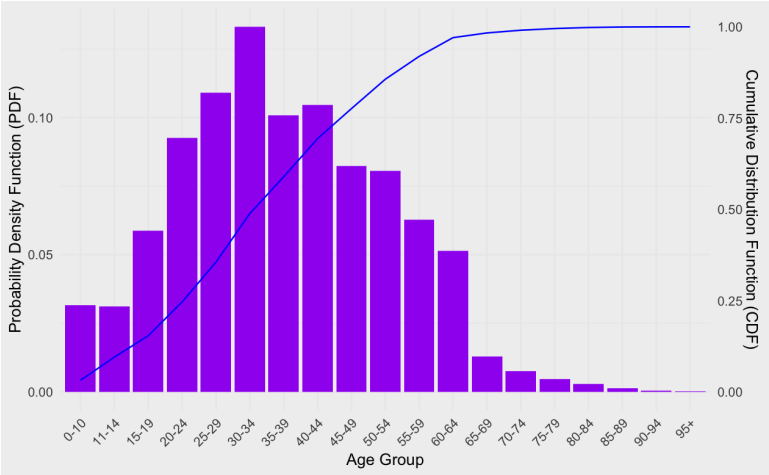

Figure 3: Distribution of total probability across severity levels for age groups (before re-grouping into five categories).

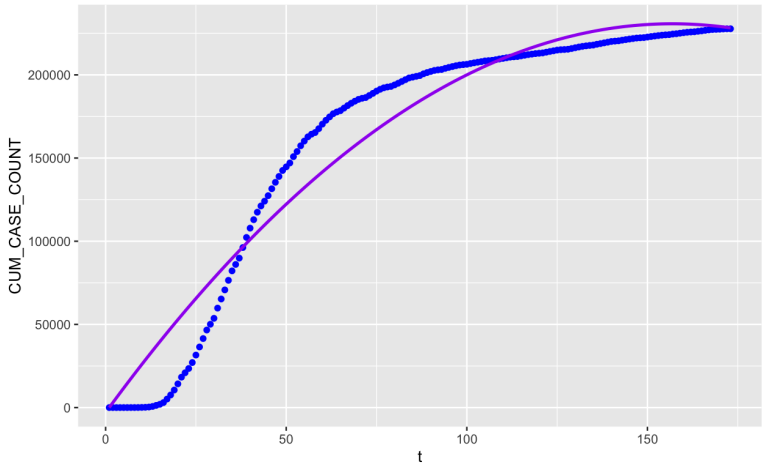

Figure 4: Poisson regression of cumulative case changes over linear time (t) and quadratic time (t^2). The actual cumulative cases are represented by blue dots, depicting the real trend, while the predicted increase in cases, modeled by t^2, is illustrated in purple.

As the predicted cumulative cases had increased, the rate of change had decreased (Figure 5). However, the actual case count had first increased and then decreased, resulting in a negative t-square value. The downward U-shaped distribution had indicated that there was a surge when t = 50. Figure 6 had demonstrated the case prediction with the unit of a week. It had aimed to predict the cases for the next day by using the case trend from the previous 7 days.

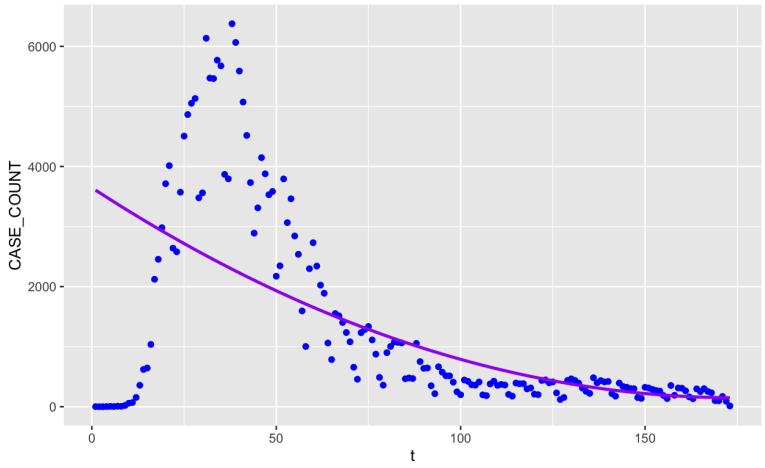

Figure 5: Poisson regression of case changes by day over linear time (t) and quadratic time (t^2). The actual cases, represented by blue dots, depict the real trend, while t^2, shown in purple, illustrates the predicted change in cases.

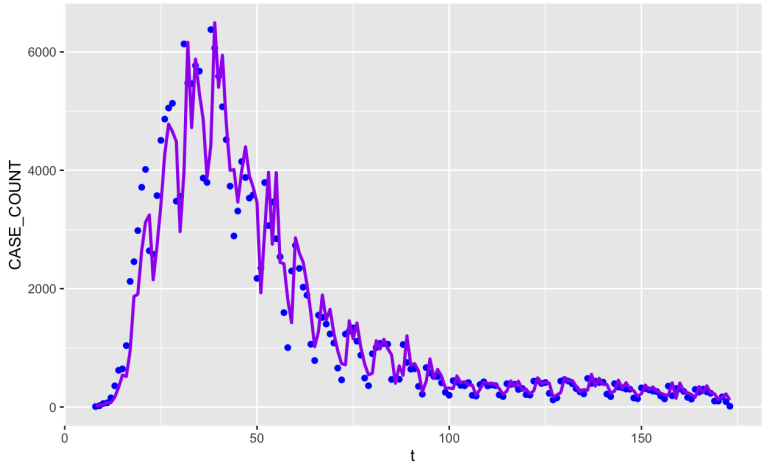

Figure 6: Poisson regression to predict the week-long case change. The actual cases, represented by blue dots, depict the real trend, while the quadratic time (t^2), shown in purple, illustrates the predicted change in cases.

In this regard, the analysis of Cornell University's COVID-19 study had underscored the effectiveness of vaccines for the university population. The study had traced the rise in deaths and case counts in New York county, demonstrating the importance of preventive measures. The data from the beginning of March had shown the rapid and, at times, unpredictable spread of the virus, emphasizing the potential effectiveness of the vaccine. From the data on symptomatic and asymptomatic individuals, different age groups had exhibited varying vulnerabilities to SARS-CoV-2 infections, necessitating further investigation into the efficacy of vaccines over the subsequent two years.

	Age_Group <chr>	Total_Population <int>	Asymptomatic_Count <dbl>	Asymptomatic_Proportion <dbl>	P-Sev-Age <int>
1	0-17	255277978	8080350.0	0.0056356828	0.27
2	18-44	599140623	61516750.8	0.0429051824	0.75
3	45-64	414878433	29683550.9	0.0207029492	0.48
4	65-74	112823290	1228650.5	0.0008569288	0.13
5	75+	51663368	173738.2	0.0001211746	0.13
Overall	NA	1433783692	100683040.4	0.0702219177	NA
Highest Asymptomatic Group	NA	*NA*	*NA*	0.0429051824	NA

Figure 7: Table 1. Summary of Asymptomatic COVID-19 Cases and Severity Proportions by Age Group.

Data Description: SARS-CoV-2 Vaccine

The New Jersey Immunization Information System (NJIIS) tracked the cumulative vaccine rate in New York State. Complete primary vaccination was defined for residents who had received either the two-dose series of the Moderna, Pfizer, or Novavax vaccine, or the single-dose series of the Johnson & Johnson vaccine. This status was previously referred to as 'fully vaccinated.' The cumulative count was based on this fully vaccinated status. On September 22nd, 2021, the FDA amended the emergency use authorization for the Pfizer-BioNTech COVID-19 Vaccine to allow the use of a single booster dose, to be administered at least six months after completing the Pfizer-BioNTech COVID-19 vaccine primary series [12]. Trends of booster doses by age groups from August 2021 to August 2022, around the time period of the booster, were displayed in Figure 8, It indicated that residents around 45-55 and 65-74 had higher vaccine rates than other age groups. This was shown in purple and pink, suggesting that people belonging to these age groups might have had a higher willingness to accept the booster.

The booster distribution by race indicates that White and Asian populations had the highest rates. Conversely, American Indian and Alaska Native populations might be less inclined to receive the booster after full vaccination, as shown in Figure 9 .

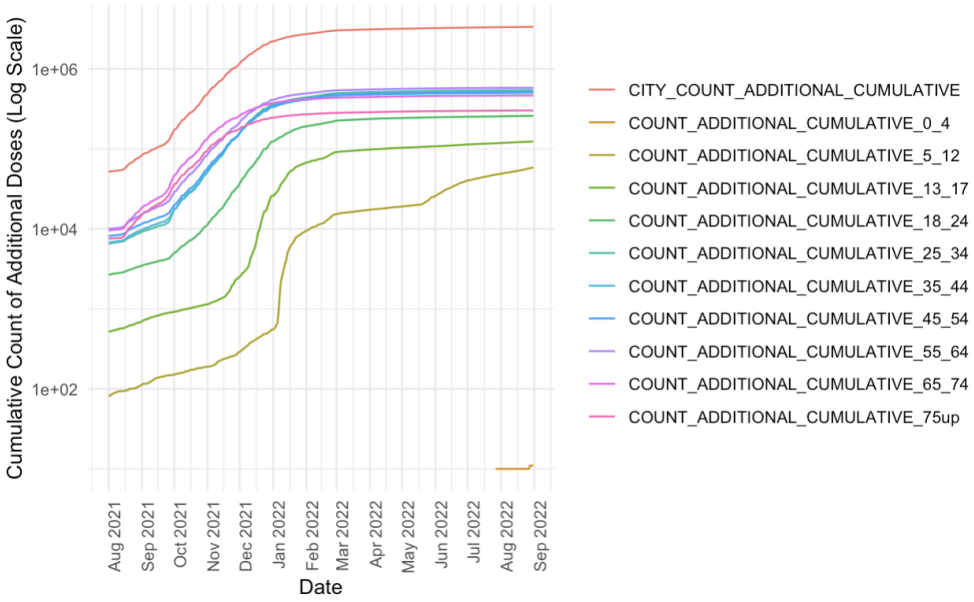

Figure 8: Trends of Booster Doses by Age Group from August 2021 to August 2022. The y-axis represents the COUNT_ADDITIONAL_CUMULATIVE of booster doses, taken on a log scale, indicating the number of people who received a monovalent booster

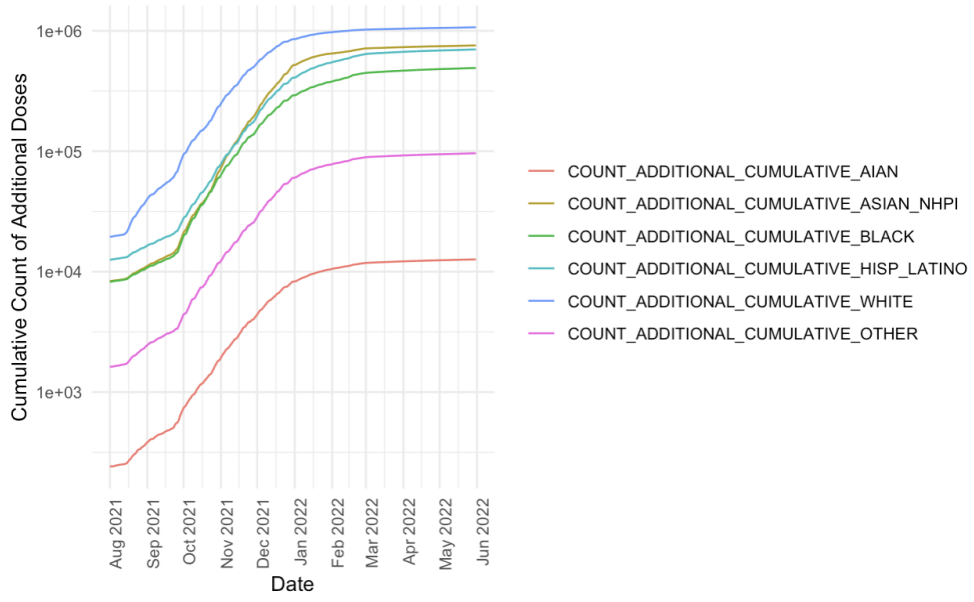

Figure 9: Trends of Booster Dose by Race Group from Aug 2021 to Aug 2022. The y-axis represents the COUNT_ADDITIONAL_CUMULATIVE of booster doses, taken on a log scale, indicating the number of people who received a monovalent booster.

Building on the insights gained from the Cornell University COVID-19 study and the analysis of New York's vaccination trends, our investigation now extends to the vaccination landscape in Connecticut, focusing specifically on the vaccination rates across different racial and ethnic groups. By delving into the New Jersey Immunization Information System (NJIIS), we aim to enhance the accuracy of our project and provide a comprehensive understanding of the vaccination dynamics. Understanding that vaccine acceptance can vary among diverse populations, our exploration into Connecticut's vaccination data will shed light on potential disparities and nuances within different racial and ethnic communities.

The Connecticut COVID-19 vaccination data used in this study focuses on ethnically diverse and age-specific populations. It details the number and percentage of people who were vaccinated, received two doses of the vaccine, and were fully vaccinated, with special attention to ethnic breakdowns. Race was broken down into five main categories: 'Hispanic', 'New Hampshire American Indian or Alaska Native', 'New Hampshire Asian or Pacific Islander', 'New Hampshire Black', and 'White' to allow for a more in-depth analysis of how different racial groups fare in terms of vaccination. The most recent data available as of August 2, 2023 was extracted to better support our study.

Through visualization using ggplot(), and disregarding the age distribution, each ethnic group generally had a high rate of vaccination when receiving the first dose, but experienced a significant drop when receiving the second dose. Notably, the NH American Indian or Alaska Native groups had vaccination rates even lower than 25%, highlighting the challenges this group faced in completing the full vaccination course. Whites were slightly more likely than Asians to receive a second dose of vaccine.

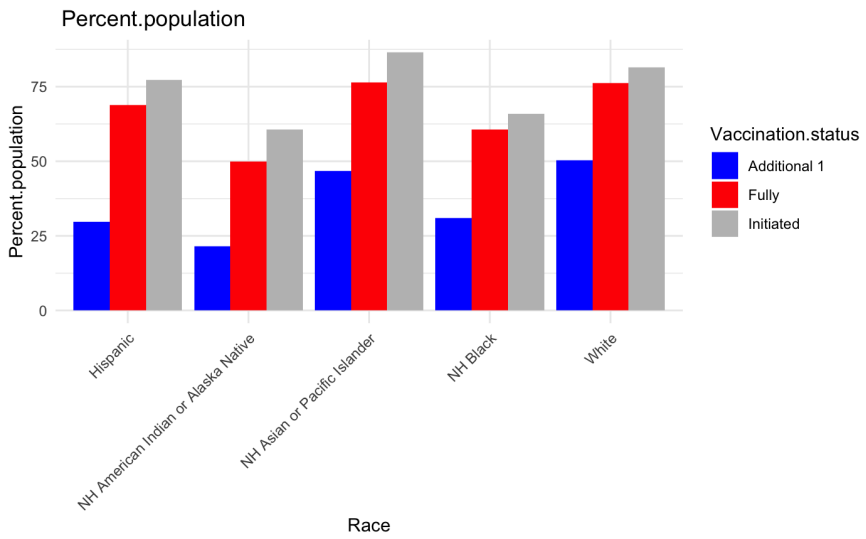

Figure 10: The vaccination status among different racial/ethnic groups in all age

However, when considering the 15-24 age group, Asians had a higher probability of a second dose of vaccination in this age group compared to Whites. It was important to note that estimates for multiple races were considered unreliable and therefore this component was omitted from the analysis.

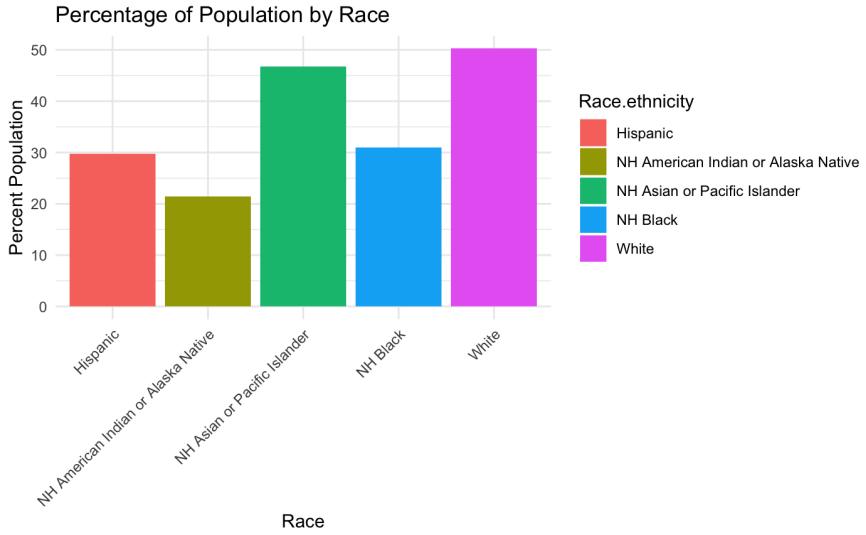

Figure 11: Displaying the percentage of population by race among individuals aged 15-24

Method Overview

The paper 'Modeling for COVID-19 College Reopening Decisions: Cornell, a Case Study' employed a variety of methods to investigate the transmission of COVID-19 and infection control strategies on university campuses. The study utilized simulation modeling to investigate the spread of COVID-19 within a university campus and predicted infection transmission under various interventions based on different parameters and assumptions. To address parameter uncertainty, Bayesian analysis was used to sample the parameter space, assess the uncertainty range of parameters, and generate multiple parameter configurations to understand the spread under different parameter settings. Additionally, the paper retrospectively reviewed data from past academic years to evaluate the consistency between past decisions and model predictions with actual outcomes, thus assessing the quality of the model and decisions. Ultimately, these methods provided support for university decision-making in the context of the COVID-19 pandemic, including whether to reopen campuses, the design of regular testing strategies, and other strategies such as isolation and tracing.

To investigate the research on booster acceptance, logistic regression was implemented to observe the additional SARS-CoV-2 vaccine uptake among populations up to the most recent update on February 2023, based on the statistical significance of race and groups. The ARIMA model focused on changing quantitative subjects to predict the average change, 2023 to 2026. Similarly, it resulted in identifying the groups with the highest positive moving average, referring to a higher willingness for booster shots in the future.

Model Logistic Regression

The Logistic Regression model is a statistical method employed for predicting the probability of an event occurrence based on one or more predictor variables. In binary classification scenarios, where the outcome variable has two possible outcomes, Logistic Regression estimates the probability that a given instance belongs to a specific category.

The mathematical expression of the Logistic Regression model is as follows:
Let Y be the binary outcome variable (0 or 1), and $X_1, X_2, \ldots, X_n$ be the predictor variables. The model assumes a linear relationship between the predictor variables and the log-odds of the event:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Here, p represents the probability of the event occurring, and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients associated with each predictor variable. The logistic function is then applied to transform the linear combination into the probability scale:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

[13]

The Logistic Regression model provides insights into the relationship between the predictors and the likelihood of the event, making it a valuable tool for binary classification tasks. Model evaluation involves assessing accuracy, precision, recall, and other relevant metrics to gauge its performance on training and test datasets. Additionally, the significance of coefficients and statistical tests can reveal the impact of each predictor on the odds of the event.

Model ARIMA

An Autoregressive Integrated Moving Average (ARIMA) Prediction Model applied statistical analysis using an integrated moving average. It predicted future trends based on past data performance.

There are three sections associated with the model. Autoregression, represented by p, refers to the number of lag observations or values in the model. Integration, represented by d, is referred to as the degree of difference. Moving average, represented by q, calculates the dependency between an observation and a residual error applied to lagged observations.

A process $\{X_t\}$ followed an Integrated ARIMA model, denoted by ARIMA(p, d, q).

If $\nabla^d X_t = (1 - B)^d X_t$ is ARIMA(p, q), we write the model as

$$\phi(B)(1 - B)^d X_t = \theta(B) Z_t,$$

$$Z_t \sim WN(0, \sigma^2).$$

[14]

The integration parameter d is a non-negative integer.

Empirical Analysis: Logistic Regression

This model aimed to delve into COVID-19 vaccination data through logistic regression modeling, focusing particularly on the influence of demographic factors on vaccination status. Utilizing the dataset named "COVID_19_Vaccinations_by_Race_Ethnicity_and_Age_-_ARCHIVED.csv," which included various variables such as ethnicity, age group, and vaccination status.

The initial step involved creating a focused subset where "Vaccination.status" was limited to "Additional 1" or "Fully," streamlining the dataset for vaccination-centric analysis. Subsequent preprocessing steps ensured data accuracy, particularly converting the character-type vaccination status to a binary variable and cleaning the ethnicity variable to ensure model accuracy.

A univariate logistic regression unveiled initial trends in the relationship between "Vaccination.status" and "Race.ethnicity." This initial exploration laid the groundwork for a deeper understanding of vaccination dynamics. However, since this was a one-dimensional logistic regression, the next study aims to more fully explore other factors that may influence vaccination status.

In pursuit of a more nuanced comprehension, the dataset underwent a split into training and testing sets using the caTools package. The multivariate logistic regression included key variables such as "Race.ethnicity," "Count," "Percent.population," and "Population." This broader approach facilitated a comprehensive analysis of how

various factors collectively influence vaccination outcomes.

Model evaluation was a crucial step to validate predictive performance. Accuracy metrics, confusion matrices, and the pseudo-R square value provided a holistic assessment, gauging the model's effectiveness in distinguishing between positive and negative vaccination statuses. In summary, the logistic regression model demonstrated promising accuracy, with an accuracy of approximately 81.61% on the train-data and 82.29% on the test-data. The pseudo-R square value of 0.8418 indicates that the model is effective in explaining the variability in vaccination status. This comprehensive evaluation supports the model's robustness in predicting and understanding the factors influencing vaccination behavior.

Integrating the comprehensive analysis and applying the logistic regression model yielded probabilities for each sample belonging to the 'Vaccination.status' category. Organizing these probabilities alongside the actual observed values in a data frame allowed us to observe variations in the inclination towards continuing booster vaccinations among different demographic groups. Through these predicted probabilities, distinct patterns emerged, revealing a higher willingness among the white population to continue receiving booster shots, followed by the Hispanic community. It reflected the attitudes and tendencies of various demographic groups towards vaccine booster uptake.

Empirical Analysis: ARIMA Time Series

The model focused on the data-set ”trends_by_age” and ”trends_by_race”. The preprocessing steps involved converting the ”date” column to DATE format, then filtering the data frame to exclude data outside the date range of August 1, 2021, to August 31, 2022. The melt() function from the reshape2() package was used to transform the data-set from a wide to a long format. Finally, the ”COUNT_ADDITIONAL_CUMULATIVE” column was converted from a string to a numerical data type.

The ARIMA model utilized the `auto.arima()` function in R, which applied a variation of the Hyndman-Khandakar algorithm. The initial results were obtained from the default model processed by `auto.arima()` in the `forecast` library, with cross-validation performed using `tsCV()`. The prediction period extended up to the end of 2025, with each result forecast one step ahead using $h = 1$, assuming 12 observations per time unit. The forecasts were stored in a list, indexed by age groups. Error metrics computed for each cross-validation fold were also stored in a list corresponding to each forecast. The results were interpreted through visualization, with two lines representing historical and forecast data, as shown in Figure 11.

The autocorrelation function (ACF) explained an exponentially decaying pattern in the ARIMA forecasting, as illustrated in Figure 12, which displayed the model’s performance. The lag, extending up to 24, indicated that the correlation of the series with its own past values was calculated for 24 time periods back, and the autocorrelation coefficient ranged between -0.1 and 0.2, reflecting the strength and

direction of the linear relationship between the time series and its lagged values. The blue dashed line represented the significance level; for instance, the bar at lag 6 was slightly above this line indicated moderate statistical significance. Overall, most bars were within the blue dashed line, suggesting that there was no significant autocorrelation at any of the lags. The residuals appeared to be white noise, indicating that the model had captured the observations well, and could be further corroborated with the given data.

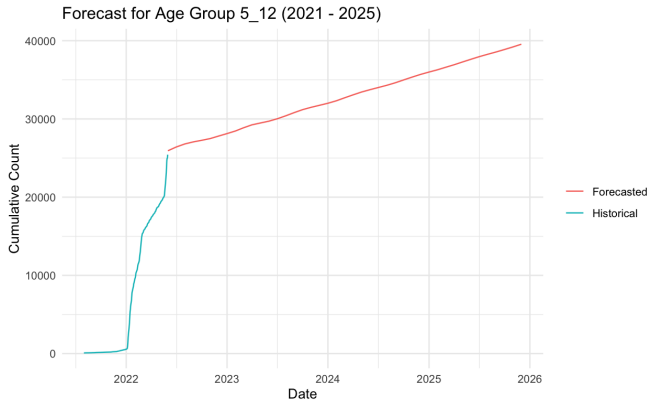

Figure 12: Comparison of Historical and Forecast Cumulative Counts of Booster Trends Among Age Group 5 to 12.

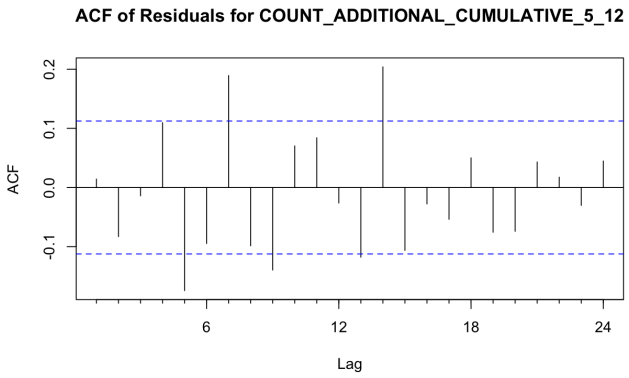

Figure 13: ACF of Residual for Forecast Cumulative Counts of Booster Trends Among Age Group 5 to 12

The ARIMA forecasting predicted the trend in SARS-CoV-2 booster vaccinations, which could indirectly indicate people's willingness to receive future boosters. The rate of change in these forecasts, observed through trends in the data, served as an indicator of changes in people's preferences. A higher increase in the cumulative count suggested an increased number of people willing to receive the COVID vaccine. Even

though for some age groups the base cumulative count was not the highest, a higher increasing rate could indicate a higher future cumulative count in those groups. Figure 13 illustrated the slope of forecasting trends in booster uptake among different age groups, with a steeper trend line possibly indicating a higher acceptance rate. It was important to note that these initial results, derived from an ARIMA model, suggested that the age groups 5-12 and 25-34 showed a higher tendency towards receiving future COVID boosters. Figure 14 displayed the changing rate of covid vaccine booster accpetance for racial groups based on the similar method as that of the age groups, indicating that White, Hispanic, Latin, and Asian, as well as Native Hawaiian and Pacific Islander groups, could have a higher willingness and booster acceptance in coming 3 years.

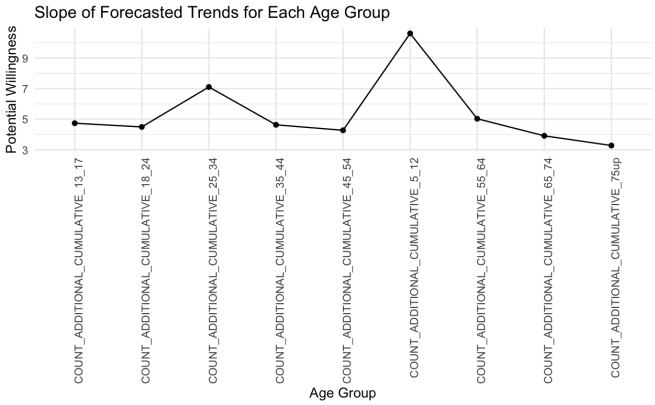

Figure 14: The Changing Rate for Age Groups, Indicating the Increasing Rate of Cumulative Count Up to the End of 2025.

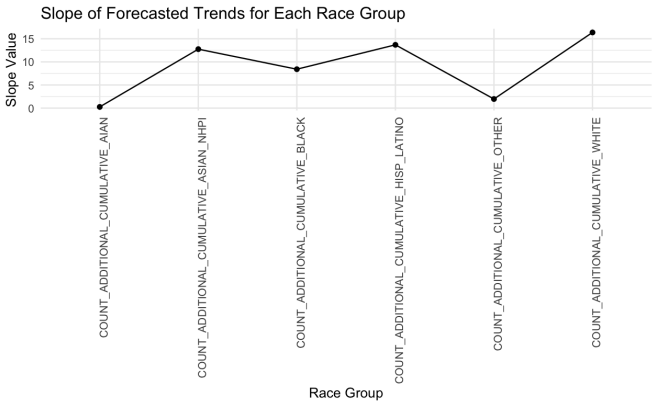

Figure 15: The Changing Rate for Race Groups, Indicating the Increasing Rate of Cumulative Count Up to the End of 2025.

Error Analysis

Even though the ACF reflected that the performance of the ARIMA model had no particular shortcomings in fitting the New Jersey Immunization Information System dataset, it was a challenge to process sentiment analysis based solely on the quantitative subject. The increase in data provided evidence that people did this in an objective sense, but it did not directly prove that all people had positive intentions about it. The outcome of an action could be influenced by many factors, such as the requirements of the community, school, or workplace. For example, in families where the elderly were the vulnerable group, but due to their own illnesses and weak immune systems to cope with vaccine complications, the younger members of the family would indirectly block the spread of the virus to the vulnerable group in the family by vaccinating themselves so that they did not bring the virus into the family. In such cases, the predictions of the age group's willingness to take action were inaccurate. Similarly, for race parameters, the higher increase rate of cumulative count observed in Asian groups could have been caused by satisfying customs policies, which was one of the reasons. The prediction of the population group's willingness to receive a vaccine booster was also inaccurate, and this was a problem that could not be avoided through data processing and model manipulation.

Sentiment analysis was used to analyze digital text to determine whether the emotional tone was positive, negative, or neutral. The text data was provided by Shuo Zhang, a Tufts Lecturer researching deep learning in natural language processing. The data was collected from Reddit comments on posts related to the COVID-19 pandemic from 2022 to 2023, associating each with a sentiment label: 0 for negative and 1 for positive [15]. The data was filtered by age group 18 to 34, all race groups, COVID vaccine, and

geographical location patterns, and preprocessed by converting to lowercase, removing stop words, punctuations, and tokenization. The cleaned data was displayed in Figure 16.

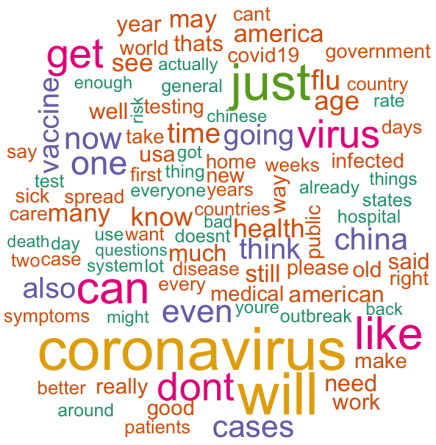

Figure 16: Overview of Reddit comments on COVID-19 vaccine.

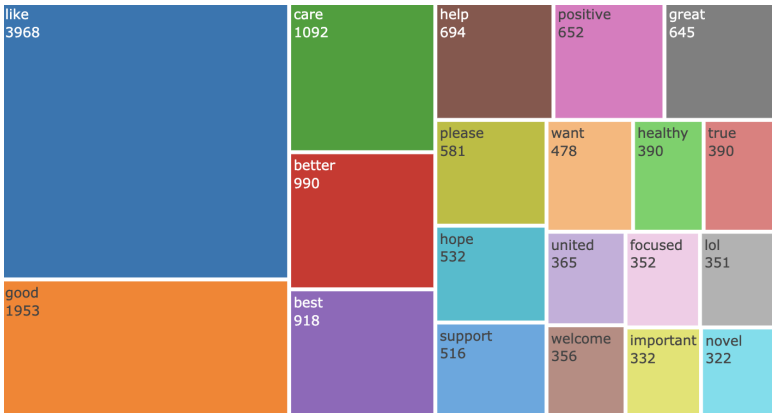

Figure 17: Top 20 positive words base on sentiment score

The sentiment score was assigned using `get_sentiments("afinn")` from the `dplyr` library. The top 20 selected words with positive tones and the top 20 with negative tones were shown in Figures 17 and 18. The overall emotional tone was positive calculated with total sentiments score. The results of the sentiment analysis were consistent with the logistic and ARIMA model predictions under the age group 15 to 24, and 25 to 34.

Figure 18: Top 20 negative words base on sentiment score

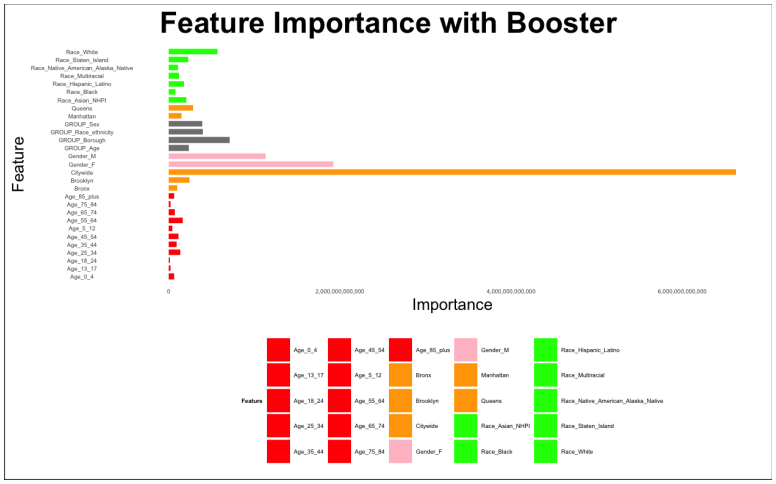

Figure 19: Feature importance in predicting COVID-19 vaccine booster acceptance on Sep.2023.

Additionally, the reliability of the findings depends on the sample size, necessitating that the data be comprehensive and diverse in terms of both quality and variety. This research considered only two parameters, a decision that might be seen as somewhat naive. The random forest model was capable of identifying the importance of each feature within the entire set of parameters present in the dataset. Modifying parameters, either through inclusion or exclusion, could potentially alter the overall predictive outcomes. This is particularly true if the new parameters demonstrate significant statistical significance. The outcomes derived from the random forest feature selection process are illustrated in Figure 19. The geographical feature was identified as the predominant factor, emerging as a critical element for future improvements in prediction models regarding the acceptance of COVID-19 vaccine boosters.

Conclusion

In conclusion, through an in-depth examination of Cornell University's decision-making during the COVID-19 pandemic and a comprehensive analysis of vaccination trends in New York and Connecticut, this study has yielded several key findings. Utilizing Logistic Regression and ARIMA time series models, we delved into the changing patterns within the data, with a particular focus on vaccination rates across different age and ethnic groups, ultimately forecasting future trends in booster shot administration. The Logistic Regression results indicate a higher likelihood of future COVID booster acceptance among the White population within certain age groups. Meanwhile, the ARIMA model emphasizes a potentially higher tendency for the 5-12 and 25-34 age groups to receive COVID booster shots in the future.

These findings underscore the importance of considering a multitude of factors when making decisions during future pandemics. Vaccine decisions are influenced not only by scientific factors but also by societal, emotional, and geographical dimensions. Through deep exploration and comprehensive analysis of the data, we provide profound insights for future decision-making, highlighting the importance of considering the characteristics and trends of different population groups in this process.

Data and Code Availability

Please visit our CS169-Final-Project GitHub Repository for more details.

References

[1] Peter I. Frazier, J. Massey Cashore, Ning Duan. *Modeling for COVID-19 college reopening decisions: Cornell, a case study*. National Academy of Sciences, 2022;2. Available at: https://doi.org/10.1073/pnas.2112532119.

[2] CDC. COVID Data Tracker. (accessed Feb. 14, 2022 & Oct 4, 2023). Available at: https://covid.cdc.gov/covid-data-tracker/#datatracker-home.

[3] Olivia C. Coiado, Aashka Shah. *COVID-19 vaccine and booster hesitation around the world: A literature review*. Front, Volume 9 - 2022 - Sec. Infectious Diseases: Pathogenesis and Therapy. Available at: https://doi.org/10.3389/fmed.2022.1054557.

[4] Rahul Shekhar, Ishan Garg. *COVID-19 Vaccine Booster: To Boost or Not to Boost*. Infectious Disease Reports, 2021; 13(4):924-929. Available at: https://doi.org/10.3390/idr13040084.

[5] Hannah Ritchie, Edouard Mathieu. *Coronavirus Pandemic (COVID-19)*. Our World in Data, Global Change Data Lab, 2020. Available at: https://ourworldindata.org/coronavirus.

[6] Katie Spiropoulos. *Tufts ends bivalent COVID-19 booster and flu vaccination requirements, The Tufts Daily.*. Available at: https://www.tuftsdaily.com/article/2023/01/tufts-ends-bivalent-covid-19-booster-and-flu-vaccination-requirements.

[7] CW. *Study says COVID vaccine mandates for university students likely causing net harm*. Available at: https://commonwealthmagazine.org/study-says-covid-vaccine-mandates-for-university-students-likely-causing-n

[8] Rahul Shekhar, Ishan Garg. *COVID-19 Vaccine Booster: To Boost or Not to Boost*. Infectious Disease Reports, 2021; 13(4):924-929. Available at: https://doi.org/10.3390/idr13040084.

[9] Arvin Hekmati, Mitul Luhar. *Simulating COVID-19 classroom transmission on a university campus, National Academy of Sciences, 2023; 120(16)*. Available at: https://doi.org/10.1073/pnas.2116165.

[10] Jen Hao Chen, Cheng Shi Shiu. *Race, ethnicity and COVID-19 vaccine concerns: A latent class analysis of data during early phase of vaccination*. ScienceDirect, Population Health, Volume 18, June 2022, 101073. Available at: https://doi.org/10.1016/j.ssmph.2022.101073.

[11] AJMC. *A Timeline of COVID-19 Developments in 2020.* (accessed Oct.18,2023 & published Jan. 2021). Available at: https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020.

[12] U.S Department of Health and Human Services. *SECRETARIAL DIRECTIVE ON ELIGIBILITY TO RECEIVE PARTICULAR COVID-19 VACCINE BOOSTERS, Sep. 25, 2021.*. Available at: https://www.hhs.gov/coronavirus/covid-19-vaccines/index.html.

[13] Donna Slonim. *Fa23-CS-0169-01-Statistical Bioinformatics,Logistic regression, Nov.06, 2023* Available at: https://www.cs.tufts.edu/cs/169.

[14] Rob J Hyndman, George Athanasopoulos. *Forecasting: Principles and Practice (2nd ed). Monash University, Australia, 2018, sec. 8.4, Moving Average Models*. Available at: https://otexts.com/fpp2/MA.html#MA.

[15] Shuo Zhang. *NLP-course-fall2023*. Available at: https://github.com/Tufts-University/NLP-course-fall2023.git