Flu Shot Learning: Predicting Vaccine Status

Gabi Rivera

University of San Diego

Master of Science, Applied Data Science

ADS501-01 Foundations of Data Science

Section 01

17 Oct 2022

Table of Contents

Business Understanding

Background

The United States National Center for Immunization and Respiratory Diseases (NCIRD) together with the National Center for Health Statistics (NCHS) has conducted the National 2009 H1N1 Flu Survey (NHFS) designed to monitor the vaccination rate upon the availability of the H1N1 vaccine to the public. The two departments of the Centers for Disease Control and Prevention (CDC) have found a new purpose for the survey. A competition was opened to explore the randomized dataset of NHFS in creating a model that will predict the likelihood of individuals acquiring vaccines comparable to COVID-19 virus vaccine demand (DrivenData). The H1N1 survey was primarily sponsored by NCIRD with the dataset provided to the public courtesy of the NCHS as stated on the competition webpage. Detailed organizational charts of each CDC department are shown in figures 1 and 2 to identify key business collaborators. The result of the predicted model will be handed to the US NCIRD for future use and purpose at their discretion.

Figure 1

National Center for Immunization and Respiratory Diseases (NCIRD) Organizational Chart

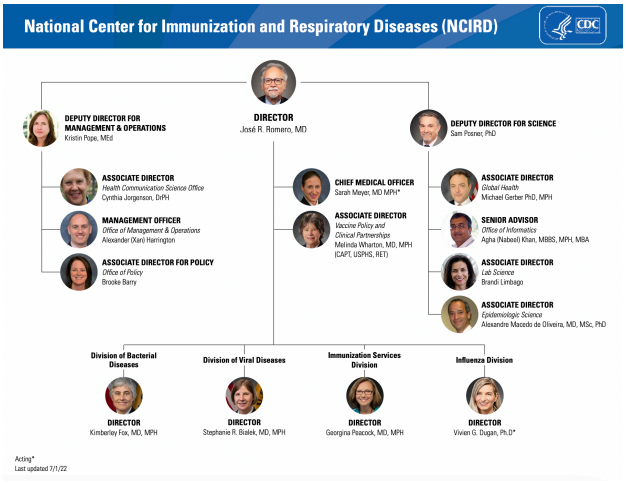
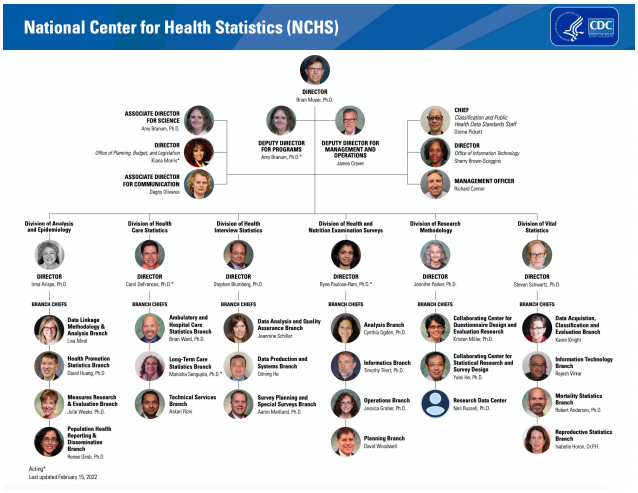


Figure 2

National Center for Health Statistics (NCHS) Organizational Chart



At the launch of the competition, COVID-19 virus was on the rise, and the need to develop a predictive tool that will provide efficient forecasting of the vaccinated population is highly practical. The challenge at the time is that COVID-19 vaccines are still in development. In response, a substitute using the National 2009 H1N1 Flu Survey dataset is employed to address the challenge of projecting the likelihood of an individual acquiring the H1N1 vaccine and/or seasonal flu vaccine based on associated background, personal opinions, and health behaviors (DrivenData). The H1N1 influenza virus vaccine survey is an ideal synthetic dataset given that its clinical spectra closely resemble that of COVID-19 as they are both respiratory illnesses (de Costa et al., 2020). Although, it is important to note that the epidemiology impact of each pandemic is incomparable (de Costa et al., 2020). To justify the scientific differences between the two viruses, the model will heavily rely on the collected behavioral tendency to generate the predictive outcome instead of the scientific makeup of the vaccines. Ultimately, the goal is simply to determine whether the associated personal profile can predict the tendency of an individual to get vaccinated against respiratory pandemics.

Business objectives and success criteria

The National Center for Immunization and Respiratory Diseases is tasked to address the prevention of disease, disability, and death by respiratory pathogens through the aid of immunization. It is imperative for this branch of the CDC to thoroughly understand the scope and reach of vaccination efforts to better address the pandemic and fulfill both their domestic and global responsibility. In addition, the results and application of the generated model are to be used in determining the flow of government resources allotted to current pandemic reliefs. With this the following objectives and success criteria are listed below:

Table 1

Business Objective and Success Criteria

Business Objective	Success Criteria
Decrease COVID-19 reinfection within 6 months.	Expected decrease by 20% threshold and lowest total at 70%
Decrease transmission rate.	Expected decrease by 20% threshold and lowest total at 70%
Decrease vaccine hesitancy and misinformation.	Expected decrease by 10% threshold and lowest total at 60%.
Predict the likelihood of vaccination.	Expected increase of vaccination rate between 20-30%.
Prevent uneven access to vaccination.	Expected community vaccination at 90%.

Inventory of resources

The raw CSV dataset was provided by the National Center for Health Statistics and was obtained directly from the DrivenData competition webpage at about 4.6MB in file size for both test and training materials (see Table 5). Additional CSV files were included for submission formatting and training labels at 374KB and 256KB, respectively. Features of the dataset are described in detail on the DrivenData webpage under the Problem Submission tab (see Table 6).

There are 38 features including the respondent_id with total records of 26,706 for the training set and 26,707 for the test dataset.

To detail the survey approach, the random-digit-dialing technique was used to phone households from late 2009 to early 2010 with questionnaires regarding their vaccination statuses on both H1N1 and seasonal flu vaccines. In conjunction with the query regarding vaccination status, additional questions pertaining to the individual's "social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission" were also collected (DrivenData).

For software, Python and R were used as the programming languages in performing data preparations, visualization, and modeling through Jupyter and RStudio platforms. Datasets were also imported in Microsoft Excel for initial data collection and simple data table creation. Regarding systems, MacBook Pro Apple M1 MacOS Monterey 8GB 251GB Storage was utilized to execute all the programming necessities and produce valuable data relayed to the sponsor.

Lastly, domain experts for business and scientific resources will be led by NCIRD's Immunization Services and Influenza teams. The two divisions will provide supplementary information regarding evolving business needs and the scientific approach of the project. Information regarding the NHFS dataset will be supported by the NCHS Division of Vital Statistics to bridge the gap in any dataset follow-up and patient consent. The acquisition of the COVID-19 survey will be headed by the NCHS Division of Health Interview statistics together with the business operation team to fulfill the Health Insurance Portability and Accountability Act of 1996 (HIPPA) compliance and patient authorization requirements. The dataset for both

N1H1 and COVID-19 will be handed over to the data administrator, IT department, and data science team for dataset keeping, governance, as well as model application.

Requirements, assumptions, constraints, and RESOLVED Strategy

The project is designed to provide useful statistical predictions about the affected population's vaccination status. Ultimately, the goal will be to deploy conscientious action items based on carefully compiled predictions on the general public that will negatively influence the spike of a blossoming pandemic. Because of the nature of how a pandemic spread in real-time, it is favorable to set a comprehensible timeline that will account for a robust predictive model that can be released without scarifying the overall accuracy score of 70% cut-off. With this, an allowance of 3.5 months will be dedicated to completing deployment tasks once a working model is available within 1 month of the beginning kick-off using the H1N1 dataset available to the public and provided by NCHS. That is a projection of 1 month dedicated to electronic random sampling data collection, 1 month for dataset compiling and cleaning, 1 month for analysis of validation to reporting, 2 weeks for the review board, and monthly maintenance together with repeatability testing running side-by-side with the deployment of action items will be performed till the threshold of 20% - 30% vaccination rate is achieved. Running along the dedicated timeline for data collection is the review of the data security, privacy, and legal restriction propositions. The data administration team headed by the IT department will initiate data governance and handling practices 1 month ahead of the model completion and will emphasize favorable terms toward the participants.

The main assumption from the generated model is that the H1N1 vaccine survey will perform similarly to the COVID-19 vaccine survey. However, the impact and scale of the H1N1 virus are incomparable having 284,000 deaths compared to the COVID-19 virus's 2 million

death worldwide in the first 12 months (Seladi-Schulman, 2021). Because of this, variables used to represent opinion and behavior (i.e., hand washing, vaccine effectiveness, health risk, etc) might not be truly comparable and representative due to the differences in responders' perceived severity. However, the model will be allowed to have a 20-30% error to account for this gap as well as adjusted to prevent overfitting with the H1N1 survey dataset. With the completion of the model, the reportable requires comprehensive documentation of the model's analysis results with appropriate statistical scores. The final review with the board of NCIRD will comprise a compiled summary of training, validation, and deployment results in conjunction with the COVID-19 vaccine and boosters.

Gaps or shortcomings in the timeline and execution of the model from both internal and external (i.e., vaccine shortage, availability, distribution to qualified margins from phase 1a to phase 2, unpredicted consequences of new strains, etc.) reasons will be broadcasted to NCIRD for transparency and quick resolution. Any budget constraints that are tied to delays will be handled between the company and the sponsor for discussion and appropriation. Due to legal constraints concerning privacy and security issues pertaining to participants' traceability, the unique identification number will be removed from the analysis. No foreseeable consequence against the model is projected from the removal of ID numbers as the categorical variables are sufficient to yield the desired likelihood status. Lastly, action items administered to decrease the pandemic reach are restricted to not-for-profit educational routes and incentives.

To fully evaluate the business ethic of the vaccine status model, the RESOLVEDD strategy was utilized. The breakdown of the approach is as follows:

1. Review: Pandemics are a common reoccurrence throughout human history. It is an almost guaranteed phenomenon but its emergence and the scale of its impact are difficult

to predict. This is the case for the COVID-19 virus. In December 2019, the virus emerged as a highly contagious pathogen that garnered alarming morbidity and mortality rate. By March 2020, the World Health Organization (WHO) declared COVID-19 a pandemic after reaching 114 countries and accounting for more than 4,291 deaths with 118,000 cases (Centers for Disease Control and Prevention, 2022). The availability of FDA-released mRNA vaccines did not materialize until about 1 year after the virus's emergence as part of an emergency use authorization (EUA) effort (Centers for Disease Control and Prevention, 2022).

2. Estimate: The challenge with emerging pandemics in the general population is the rate of contagiousness and the mortality caused by the pathogen as well as its high mutation rate that impedes the success rate of disease eradication through vaccine availability. In the case of COVID-19, the mutation rate of the virus is immoderate which competes with the ability to release and produce new vaccines. This leads to difficulty in containing the infection rate and achieving herd immunity as variability in the public's attitude toward the vaccine adds to the complication. Vaccine hesitancy is a valid skepticism but oftentimes has dire consequences during pandemics when time is of the essence. In addition, another challenge posed by the acquired dataset as sensitive patient information regarding vaccination status requires HIPAA and patient authorization consent. Ethically, handling vaccine status information tied to a patient's identification number demands proper data governance, privacy, security, and exploitation protection.
3. Solution: The approach is to increase the vaccination rate to control the spread of the pandemic and the mortality counts associated. Exploration of ways to encourage the population to acquire available vaccines during a pandemic can help boost the

vaccination count. With this, a predictive model on vaccination status through surveyed information about the opinions, behavior, and personal profile is highly desirable with its propensity to reveal solutions against vaccine hesitancy in the general population. To address sensitive patient information, unique identification will be masked or removed altogether as mentioned previously. The model will be allowed to stand alone without strings associated with a unique patient ID.

4. Outcome: The ideal outcome is to increase the vaccination rate which has the aptitude to attain faster herd immunity. The advantage of herd immunity is that once a certain threshold of the population is immune to the disease, the spread of the pathogen becomes unlikely and so the whole population becomes protected. More importantly, the vulnerable portion of the population will also receive protection without getting the vaccination. For patient data protection, the ideal outcome is that masking patient ID removes apparent bias in the model generation and analysis throughout the process.
5. Likely Impact: Addressing vaccine hesitancy through the help of predictive models has the propensity to influence the general public's perception regarding mandated medical approaches. Just like personalized advertising, the aim of using the model is to cause a positive impact on the population's vaccine status. With the removal of patient ID from the model, the desired likely impact is to lessen or remove any perceptible bias that may be inadvertently perpetuated in the model.
6. Values upheld and violated: The intent of generating a vaccine status model is meant for the greater good of the population. The goal is to help control the spread of the pandemic and lessen its damage by increasing the vaccination rate. The downside to this is the intent of purposely influencing the public's opinion and behavior toward vaccines. There

is also the fear of incorporating subtle but inevitable biases in the model that have the consequence to misrepresent disadvantaged groups.

7. Evaluate: In the chance, the vaccine status model is utilized properly and contributed to a min of 20% increase in vaccination rate, the likely impact is to lower the damages and human suffering associated with the pandemic. However, newer evaluations and criticism of the model's use will go on long after the pandemic. Rebuttal on propagandist use of the model's output and its deliberate intention to sway people's free-will will be challenged especially when unpredicted outcomes arise. It is not guaranteed that the outcome will be true to its original purpose and with that, some unforeseeable consequences can occur.
8. Decide: The decision to proceed with the creation and use of a predictive model outweighs the current pending consequences. That is because at present the need to save human lives is greater than the future misappropriation of the model.
9. Defend: The community needs to be flexible in exploring novel avenues to help address the need to increase the vaccination rate during a pandemic because human lives are at risk. It is a tangible and immediate dilemma so exceptions to rules should be considered. Once the pandemic is over, limitations and restrictions can be imposed on the use of the technology. It can be reevaluated as necessary however during unpredictable times it is imperative to contribute toward the success of the many.

Risks and contingencies

Assessment of expected risks associated from developing a predictive model along with proposed contingencies are listed below:

Table 2

Risk and Contingency Assessment

Type	Identified Risk	Contingency
Organizational	Expert staff available might limit the projected milestones in completing the model.	Communication of the projected milestones and securing schedule priorities will be deployed to meet deadlines.
Financial	Government funding is based on the success criteria and reliability of reviewed results through competition.	Expert department heads will scrutinize results together with a combined real-time assessment of competition to lock funding wins.
Technical	Complexity of the predictive model might be too incomprehensible with 36 variables.	A technical approach to trimming down variables based on statistical significance will be employed.
Data and source	Application to COVID-19 virus vaccine might be incomparable using H1N1 and Seasonal vaccines dataset.	A validation dataset will be included to avoid overfitting the predictive model against the H1N1 and Seasonal vaccine datasets.
	Health Insurance Portability and Accountability Act of 1996 (HIPPA) compliance	Appropriate staff training regarding sensitive patient information and handling will be provided. Patient identification will be masked or removed as needed.

Terminology

Important terminologies were put together for ease of business communication.

Table 3

Technical Terminology

Area	Terminology	Definition
Business	B2G	Business to government transactions.
	Benchmarking	evaluation of business standards through market competitors comparison.
	Corporate Social Responsibility (CSR)	Self-regulation practice that aims to consider environmental and social impact throughout business decisions.
	Ethical Investment	Investments on companies with environmental and moral credentials.

	Ethical Trade	Umbrella term for business practices promoting social and/or environmental responsible trading.
	Gantt Chart	Visualization of WBS
	Work Breakdown Structure (WBS)	Breakdown of deliverables from major to individual tasks.
Data Mining	Machine learning	This is a field that explores the use of algorithms that can learn from the data and use that knowledge to make predictions on data they have not seen before.
	Nominal	Categorical values from an unordered set.
	Ordinal	Categorical values from an ordered set.
	Predictor variable	A predictor variable is a variable that is being used to measure some other variable or outcome. In an experiment, predictor variables are often independent variables, which are manipulated by the researcher rather than just measured.
	Response variable	An outcome or response variable is in most cases the dependent variable, which is observed and measured by changing the independent variable(s).
	Softmax Regression	Multinomial logistic regression is used to predict discrete values.
	Supervised learning	This is a branch of machine learning that includes problems where a model could be built using the data and true labels or values.

Note. The majority of terminology was obtained from "A hands-on introduction to Data Science" by Chirag Shah.

Data mining goals and success criteria

In response to the business objectives, data mining goals are listed in Table 4 to address the technical criteria of the project model.

Table 4

Data Mining Objective and Success Criteria

Data Mining Objective	**Success Criteria**
Compare vaccination rate base on variables (i.e., background, personal opinions, health behaviors, etc.).	Visualization and K-S statistical score of 70% precision and accuracy.
Compare personal opinion vs health behaviors.	Visualization and K-S statistical score of 70% precision and accuracy.

Determine the public's attitude against vaccination base on the given variables (i.e., background, personal opinions, health behaviors, etc.).	Provide positive and negative scores or visualization based on an individual's social, economic, and demographic backgrounds. Include ROC curves for statistical significance.
Determine the relationship of each variable (i.e., background, personal opinions, health behaviors, etc.) with an individual's tendency to get the vaccine.	Provide top categorical variables that are highly correlated to vaccination tendencies with meaningful statistical scores and visualization. Include ROC curves for model performance.

Project plan/ Order of tasks

As a summary of weekly sponsor meetings (see Appendix A), business and legal documentation mark the start of the project where all necessary data security, privacy, and patient health consent will be laid out and acquired ahead of time before phase II of the data mining. Phase I of data mining will begin a week before business and legal documentation wrap up. This phase entails weekly deliverables and communications during the model creation projected to complete within a month. Phase II of data mining comprises model deployment using the acquired COVID-19 surveyed dataset with a completion time of 3 months. A review board will be held as a final assessment of the predictive model's success criteria. Maintenance and repeatability testing of the model will be determined till a 20% - 30% vaccination rate is retained.

Data Understanding

Initial data collection report

The predictive model for vaccine status will be created using features from the dataset that address the individual's background profile together with personal opinions on health behaviors as listed in Table 6. Unique features that show the most meaningful categorical contribution backed with a statistical relevance score will be selected for the final model after exploration and analysis. The training set will initially be used to create the predictive model and

later applied to the test dataset to assess robustness. As shown in Table 5, the primary datasets are the Test and Training Features. The Training Labels and Submission Format was acquired for additional information and reference. With this, two variables were added to the Training dataset to incorporate whether respondents have been vaccinated against H1N1 and Seasonal flu. Vaccination status is the main feature to focus on throughout the project. Not all of the 36 features will be included in the final model. Redundant features under the same category such as behavior and opinion will be omitted once the correlation pattern is explored. Lastly, missing data of each categorical variable will be addressed for removal or imputation during pre-processing.

Table 5

Raw File Names and Format

File Names	Format	File Size
Flu_Shot_Learning_Predict_H1N1_and_Seasonal_Flu_Vaccines_-_Test_Features	CSV	4.6MB
Flu_Shot_Learning_Predict_H1N1_and_Seasonal_Flu_Vaccines_-_Training_Features	CSV	4.6MB
Flu_Shot_Learning_Predict_H1N1_and_Seasonal_Flu_Vaccines_-_Training_Labels	CSV	256KB
Flu_Shot_Learning_Predict_H1N1_and_Seasonal_Flu_Vaccines_-_Submission_Format	CSV	374 KB

Note. Only Test and Training Features will be utilized for data delivery.

Table 6

Dataset Features Description

Features	Description	Variables
respondent_id	Unique and random identifiers	
h1n1_vaccine	Whether respondent received H1N1 flu vaccine	0 = No; 1 = Yes
seasonal_vaccine	Whether respondent received seasonal flu vaccine	0 = No; 1 = Yes
h1n1_concern	Level of concern about the H1N1 flu.	0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned
h1n1_knowledge	Level of knowledge about H1N1 flu.	0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge
behavioral_antiviral_meds	Has taken antiviral medications. (binary)	

behavioral_avoidance	Has avoided close contact with others with flu-like symptoms. (binary)	
behavioral_face_mask	Has bought a face mask. (binary)	
behavioral_wash_hands	Has frequently washed hands or used hand sanitizer. (binary)	
behavioral_large_gatherings	Has reduced time at large gatherings. (binary)	
behavioral_outside_home	Has reduced contact with people outside of own household. (binary)	
behavioral_touch_face	Has avoided touching eyes, nose, or mouth. (binary)	
doctor_recc_h1n1	H1N1 flu vaccine was recommended by doctor. (binary)	
doctor_recc_seasonal	Seasonal flu vaccine was recommended by doctor. (binary)	
chronic_med_condition	Has any of the following chronic medical conditions: asthma or another lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. (binary)	
child_under_6_months	Has regular close contact with a child under the age of six months. (binary)	
health_worker	Is a healthcare worker. (binary)	
health_insurance	Has health insurance. (binary)	
opinion_h1n1_vacc_effective	Respondent's opinion about H1N1 vaccine effectiveness.	1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective
opinion_h1n1_risk	Respondent's opinion about risk of getting sick with H1N1 flu without vaccine.	1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high
opinion_h1n1_sick_from_vacc	Respondent's worry of getting sick from taking H1N1 vaccine.	1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried
opinion_seas_vacc_effective	Respondent's opinion about seasonal flu vaccine effectiveness.	1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective
opinion_seas_risk	Respondent's opinion about risk of getting sick with seasonal flu without vaccine.	1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high
opinion_seas_sick_from_vacc	Respondent's worry of getting sick from taking seasonal flu vaccine.	1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried
age_group	Age group of respondents.	
education	Self-reported education level.	
race	Race of respondent.	
sex	Sex of respondent.	
income_poverty	Household annual income of respondent with respect to 2008 Census poverty thresholds.	

marital_status	Marital status of respondent.
rent_or_own	Housing situation of respondent.
employment_status	Employment status of respondent.
hhs_geo_region	Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
census_msa	Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
household_adults	Number of other adults in household, top-coded to 3.
household_children	Number of children in household, top-coded to 3.
employment_industry	Type of industry respondent is employed in. Values are represented as short random character strings.
employment_occupation	Type of occupation of respondent. Values are represented as short random character strings.

Data description report

Figure 1 shows a brief summary of the training dataset. As mentioned, both the training and the test dataset share the same 38 variables with 26,706 total training records and 26,707 total test records. This shows that the training and test datasets are about 50:50 or split in half as material sources. The need to split the data to potentially perform validation was assessed throughout to avoid over-fitting the model and produce a robust data evaluation. Going back to the summary, the initial variables comprise 3 data types which are 64-bit Integer, Object, and 64-bit Float with 1, 12, and 23 variable counts under each type respectively. Based on Table 6, some of the variables, especially those identified as Array and Boolean, were adjusted to their respective categories for suitable data exploration. Depending on the need of the predictive model process, features were transformed into the best representative type to satisfy plot and library requirements.

In addition, there is a considerable amount of missing data with a total of 1014.9K (Shown in Figure 2). The max missing data per column is at 50% or 13,470 records while the

max missing data per row is at 58%. Removing all the missing data altogether will comprise ~76% of data loss. As a resolution, missing data were subjected to Multiple Imputation by Chained Equation (MICE) in R using the random forest methodology. Five iterations were calculated and the best-suited iteration used was assessed through the generated comparative regression plot and by comparing it against the original data distribution. A completed cleaned training dataset was transferred back to Jupyter notebook for data exploration.

Data exploration report

Data distribution was explored as a preliminary evaluation of the training dataset features. The general shape and spread of key predictors are shown in Figures 3-4 for integer and object data types to help guide the next steps in the data pre-processing analysis. In addition, a correlation matrix was calculated in Figure 5 using Theil's U statistics to quantify the relationship between the categorical features. Understanding the relationship between the variables is essential to identify possible predictors of interest as well as to eliminate unnecessary predictors that don't add relevant information to the desired model of predicting an individual's likelihood to seek vaccination. Overall, the yield across the matrix displays low associations, especially along the main features of interest. For vaccination status, the highest score for H1N1 is 39% when paired with opinions on seasonal risk while 38% is the score of seasonal versus opinions on the risk of H1N1.

To further explore the relationship between the categorical variables, plots were created to display combinations of features involving behaviors, demographic, opinions, population, and socio-economics that best represent each subject. Figure 6 shows an increasing trend as responders believe in the effectiveness of the seasonal flu vaccine along with the H1N1 flu vaccine. Health insurance policyholders slightly lead the frequencies across the five options. In

Figure 7, frequent hand washing and opinions on the risk of getting H1N1 without a vaccine were paired along with race. Frequent hand washer responders have a slightly higher frequency along with higher scores on believing that no vaccine exposes them to an increased risk of H1N1 infection. Race predictor also seems to have a slightly increasing pattern with Hispanic on the lead for negative and positive outcomes. From here, it seems that positive opinions and behavior toward flu prevention and vaccination have some association that can be utilized in the final model.

Figure 8 looks at the area's population density and the frequency of respondents who received the H1N1 flu vaccine along with income distinctions. There is no pronounced disadvantage attached to the denser areas in relation to H1N1 vaccination access. However, income disparity favors respondents earning $75,000 and above annually in acquiring H1N1 vaccination compared to lower income earners or those below the poverty line. With this, income appears to be a better predictor to use overpopulation density in determining people's likelihood of getting vaccinated. Lastly, additional plots were created to see the relationships between other predictors that identify which of the 36 dependent variables apply best to the final model. Ultimately, the goal is to create a multinomial logistic regression that determines people's vaccination status based on their background, health opinions, and health behaviors.

Data quality report

Overall, the quality of the datasets is in a good shape. The flat files were saved as CSV and have consistent attributes throughout. The only caveat is that the training dataset had an enormous amount of labeled and unlabeled missing data as well as incorrect datatype descriptions. The challenge was to transform the datatypes and missing data in order to utilize R studio's MICE library package sensitive to types of categorical variables and blank fields. So, 15

features were initially labeled as float64 and were re-labeled as logical to reflect their binary nature which is false or true instead of 0 or 1, respectively. All 10 ordinal variables were converted as factors from integer64 datatype to capture the in-variable scaling categories as well as the NA values. Before the 12 object variables were also transformed into factors, the blanks were filled in with NA to represent the package's acceptable null value. The resulting data was then saved in a CSV format and imported back into Jupyter for data exploration.

References

Centers for Disease Control and Prevention. (2022, August 16). *CDC Museum Covid-19 Timeline. Centers for Disease Control and Prevention*. Retrieved October 3, 2022, from https://www.cdc.gov/museum/timeline/covid19.html

Centers for Disease Control and Prevention. (2022, January 18). *Similarities and differences between flu and covid-19*. Centers for Disease Control and Prevention. Retrieved September 11, 2022, from https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm

Centers for Disease Control and Prevention. (2020, July 30). *Immunization and respiratory diseases (NCIRD) overview*. Centers for Disease Control and Prevention. Retrieved September 11, 2022, from https://www.cdc.gov/ncird/overview/index.html

da Costa, V. G., Saivish, M. V., Santos, D. E., de Lima Silva, R. F., & Moreli, M. L. (2020). *Comparative epidemiology between the 2009 H1N1 influenza and COVID-19 pandemics*. Journal of Infection and Public Health, 13(12), 1797–1804. https://doi.org/10.1016/j.jiph.2020.09.023

DrivenData. (n.d.). *Flu shot learning: Predict H1N1 and seasonal flu vaccines*. DrivenData. Retrieved September 11, 2022, from https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/

Seladi-Schulman, J. (2021, April 16). *H1N1 influenza vs. COVID-19 comparison: Similarities & Differences*. Healthline. Retrieved October 2, 2022, from https://www.healthline.com/health/h1n1-vs-covid-19#quick-comparison-table

Shah, C. (2020). *A hands-on introduction to Data Science*. Cambridge University Press.

Supplemental Data

Figure 1

Initial Summary Description of Flu Shot Training Dataset

```
fs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26707 entries, 0 to 26706
Data columns (total 36 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   respondent_id                26707 non-null  int64  
 1   h1n1_concern                 26615 non-null  float64
 2   h1n1_knowledge               26591 non-null  float64
 3   behavioral_antiviral_meds    26636 non-null  float64
 4   behavioral_avoidance         26499 non-null  float64
 5   behavioral_face_mask         26688 non-null  float64
 6   behavioral_wash_hands        26665 non-null  float64
 7   behavioral_large_gatherings  26620 non-null  float64
 8   behavioral_outside_home      26625 non-null  float64
 9   behavioral_touch_face        26579 non-null  float64
 10  doctor_recc_h1n1             24547 non-null  float64
 11  doctor_recc_seasonal         24547 non-null  float64
 12  chronic_med_condition        25736 non-null  float64
 13  child_under_6_months         25887 non-null  float64
 14  health_worker                25903 non-null  float64
 15  health_insurance             14433 non-null  float64
 16  opinion_h1n1_vacc_effective  26316 non-null  float64
 17  opinion_h1n1_risk            26319 non-null  float64
 18  opinion_h1n1_sick_from_vacc  26312 non-null  float64
 19  opinion_seas_vacc_effective  26245 non-null  float64
 20  opinion_seas_risk            26193 non-null  float64
 21  opinion_seas_sick_from_vacc  26170 non-null  float64
 22  age_group                    26707 non-null  object 
 23  education                    25300 non-null  object 
 24  race                         26707 non-null  object 
 25  sex                          26707 non-null  object 
 26  income_poverty               22284 non-null  object 
 27  marital_status               25299 non-null  object 
 28  rent_or_own                  24665 non-null  object 
 29  employment_status            25244 non-null  object 
 30  hhs_geo_region               26707 non-null  object 
 31  census_msa                   26707 non-null  object 
 32  household_adults             26458 non-null  float64
 33  household_children           26458 non-null  float64
 34  employment_industry          13377 non-null  object 
 35  employment_occupation        13237 non-null  object 
dtypes: float64(23), int64(1), object(12)
memory usage: 7.3+ MB
```

Figure 2

Missing Value Plot of Flu Shot Training Dataset

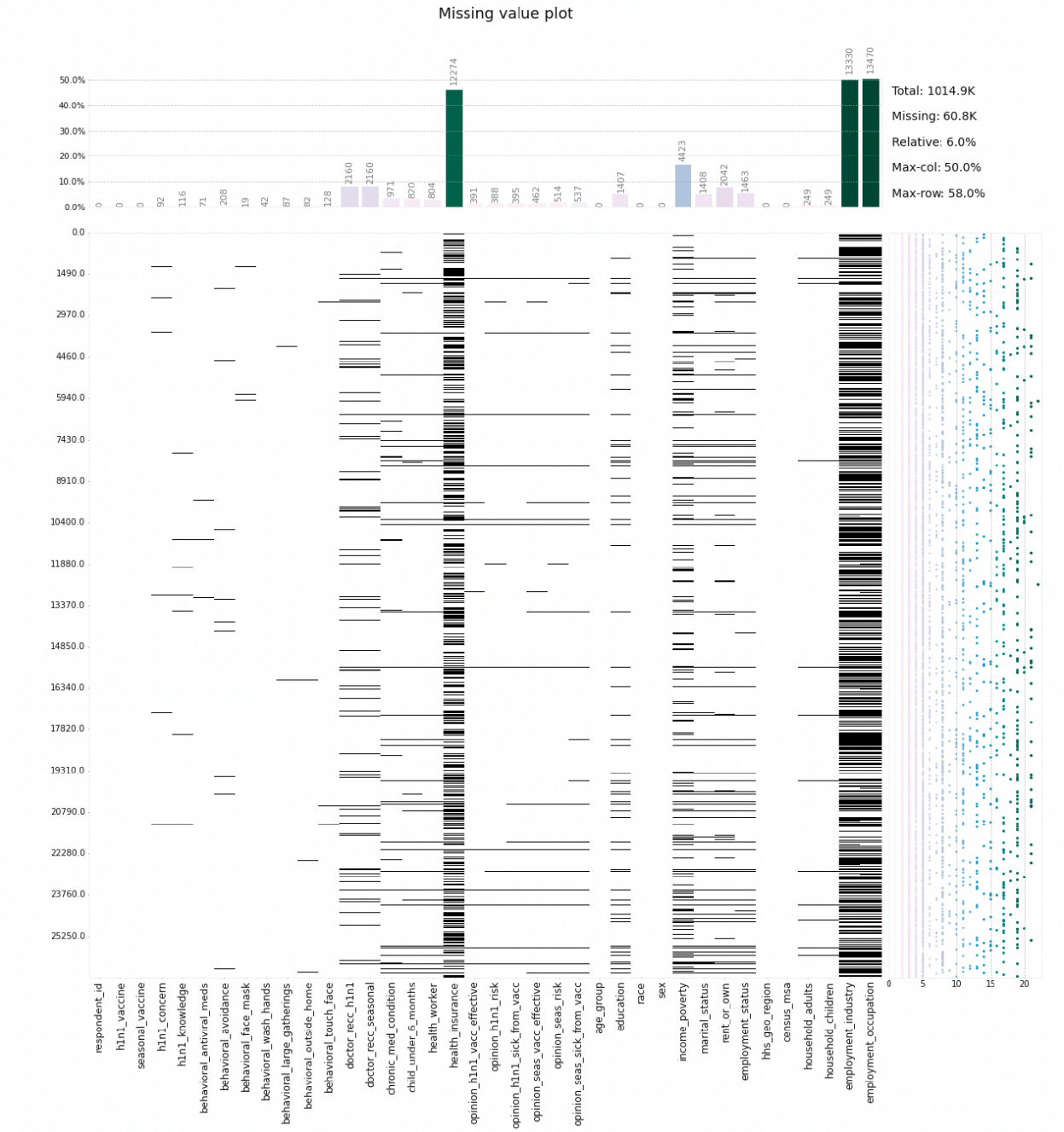
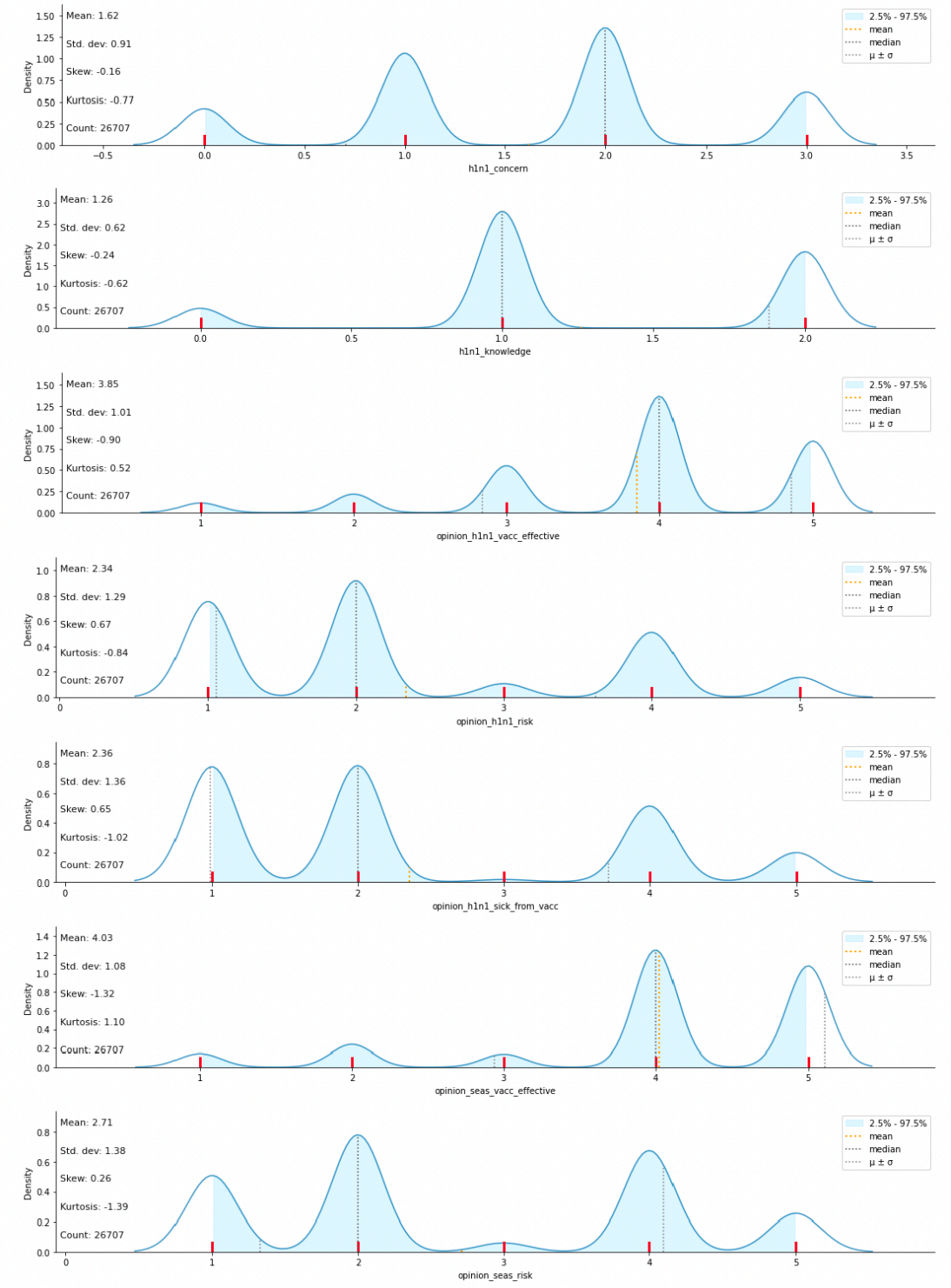


Figure 3

Distribution Plots of Training Dataset with Integer Labeled Features



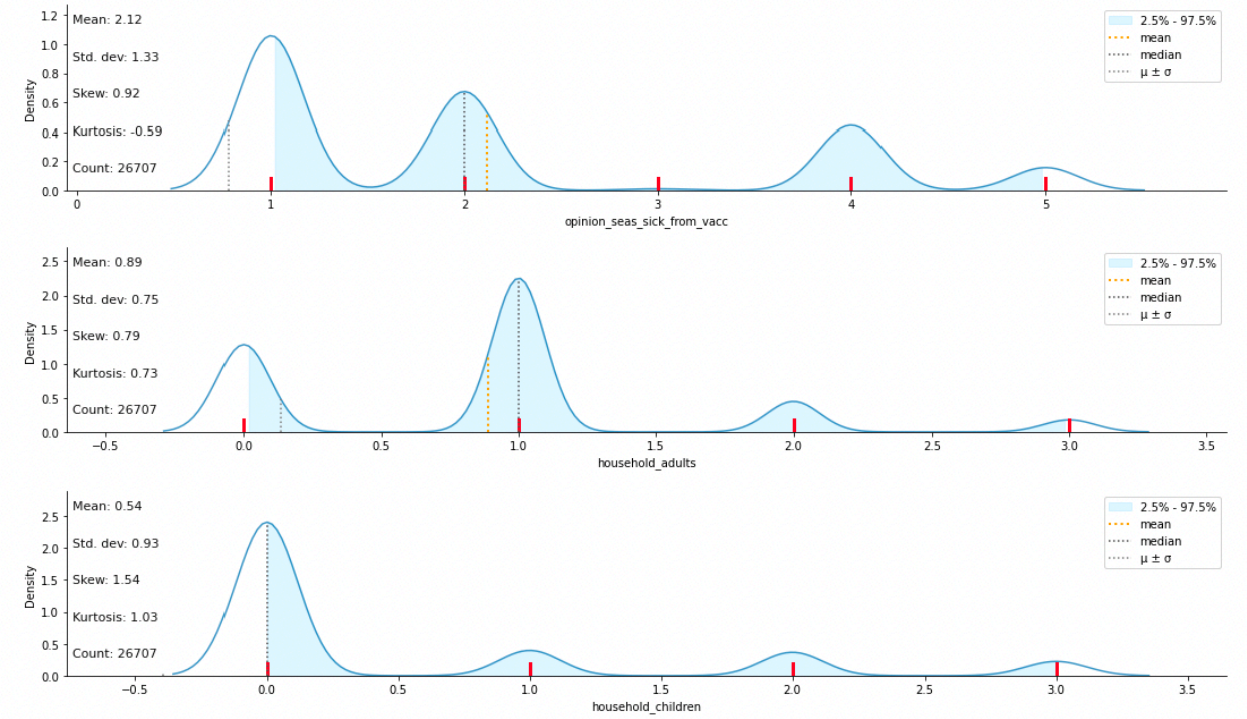

Mean: 2.12
Std. dev: 1.33
Skew: 0.92
Kurtosis: -0.59
Count: 26707
Density
opinion_seas_sick_from_vacc
2.5% - 97.5%
mean
median
μ ± σ
Mean: 0.89
Std. dev: 0.75
Skew: 0.79
Kurtosis: 0.73
Count: 26707
Density
household_adults
2.5% - 97.5%
mean
median
μ ± σ
Mean: 0.54
Std. dev: 0.93
Skew: 1.54
Kurtosis: 1.03
Count: 26707
Density
household_children
2.5% - 97.5%
mean
median
μ ± σ

Figure 4

Categorical Plots of Training Dataset with Object Labeled Features

Figure 5

Categorical Correlation Matrix using Thiel's U Statistics

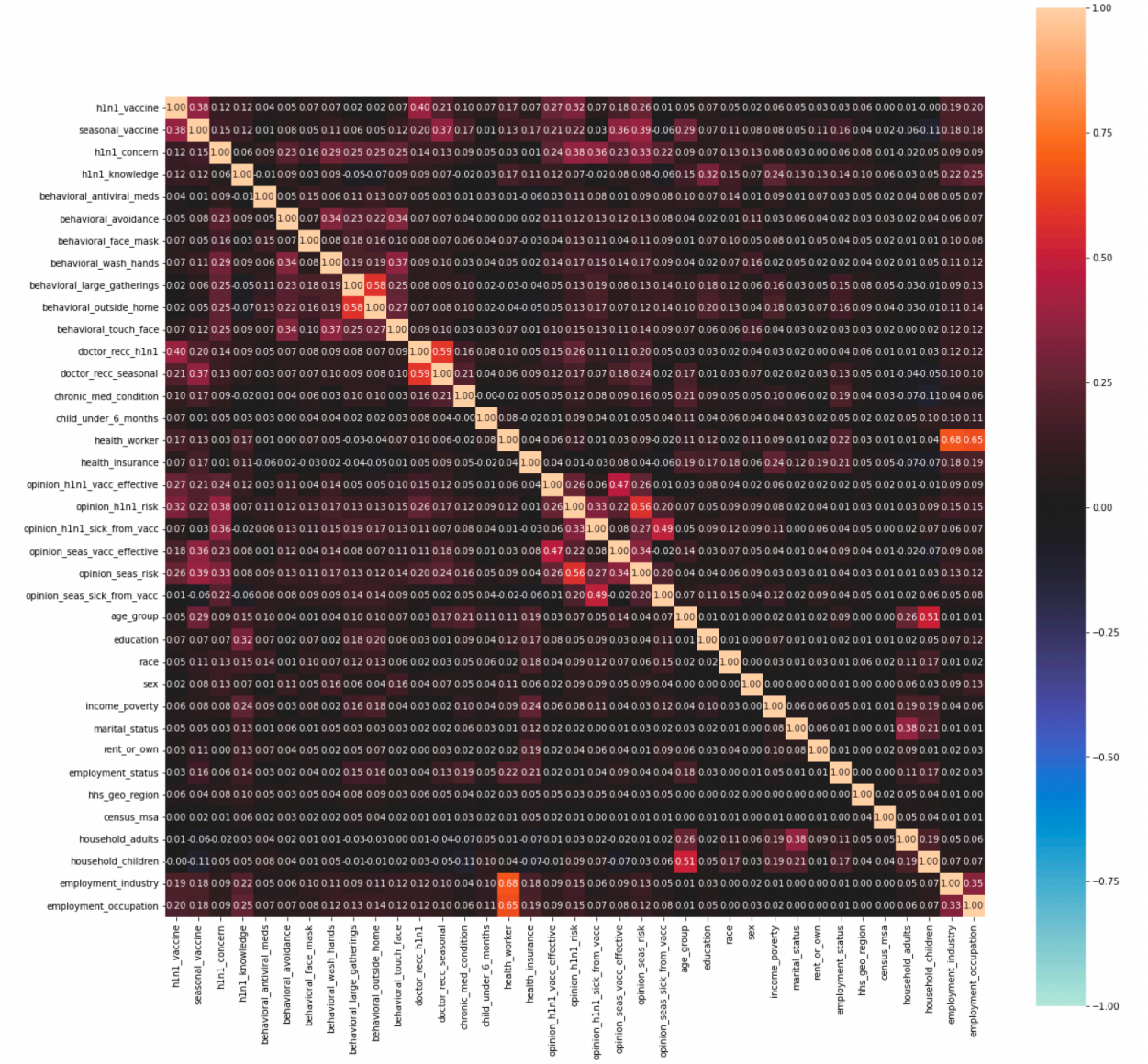


Figure 6

Opinions on the effectiveness of H1N1 vs Seasonal Vaccine with having health insurance

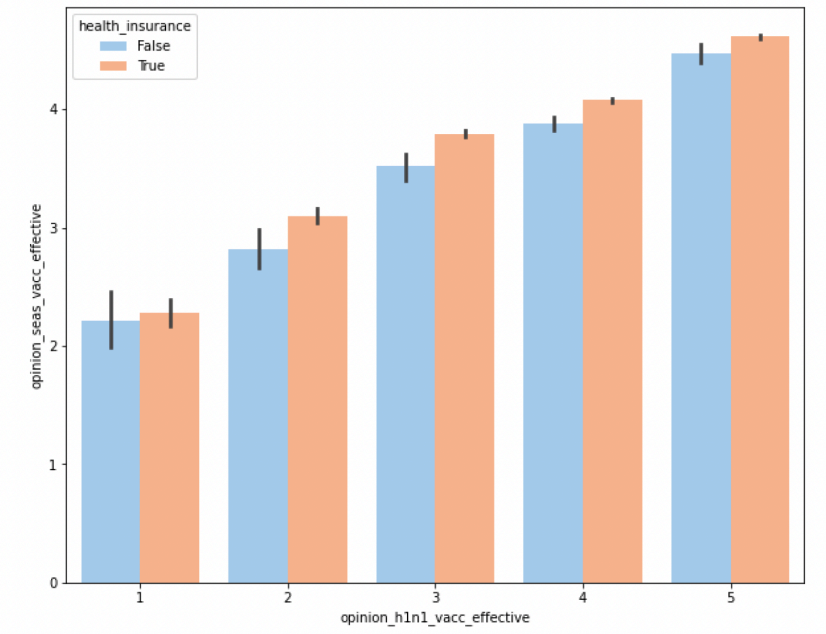


Figure 7

Hand washing vs opinions on the risk of H1N1 by race

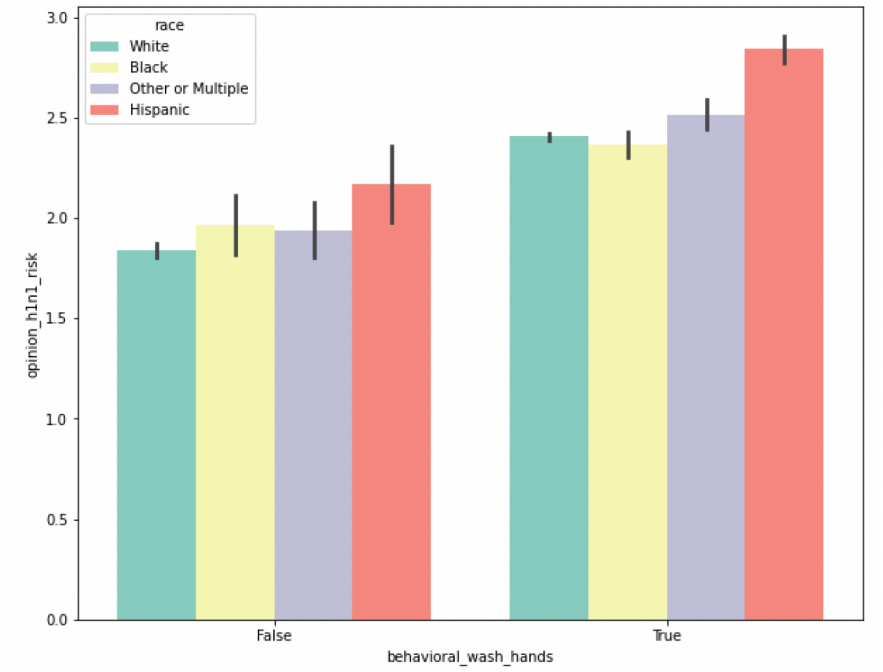
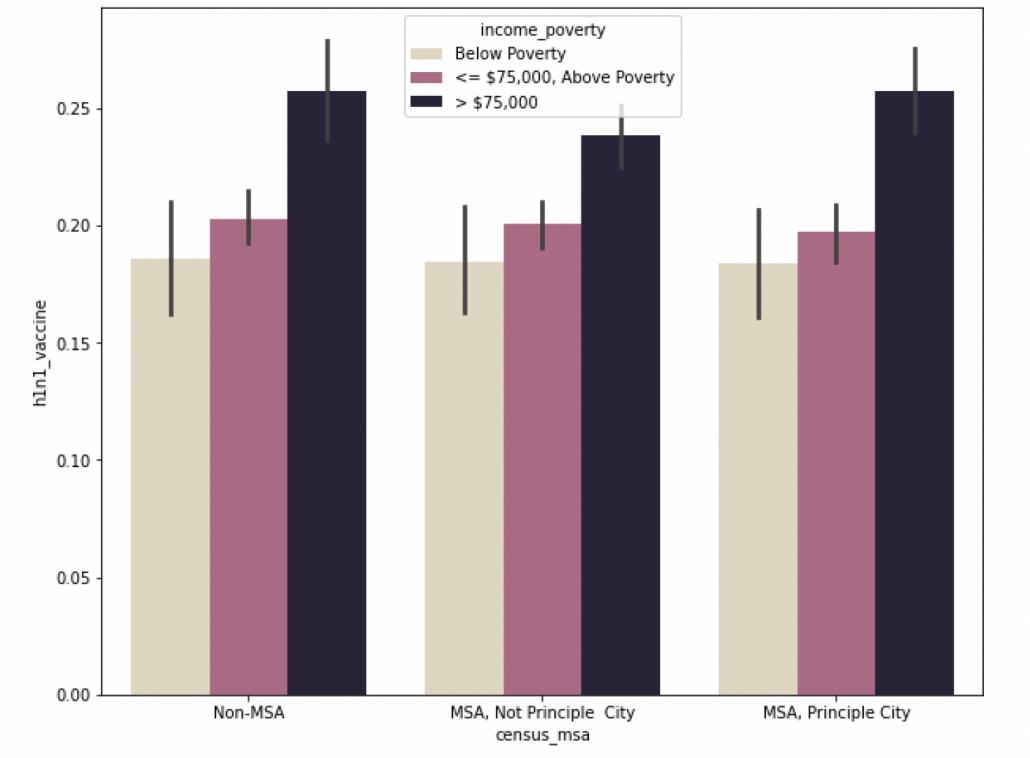
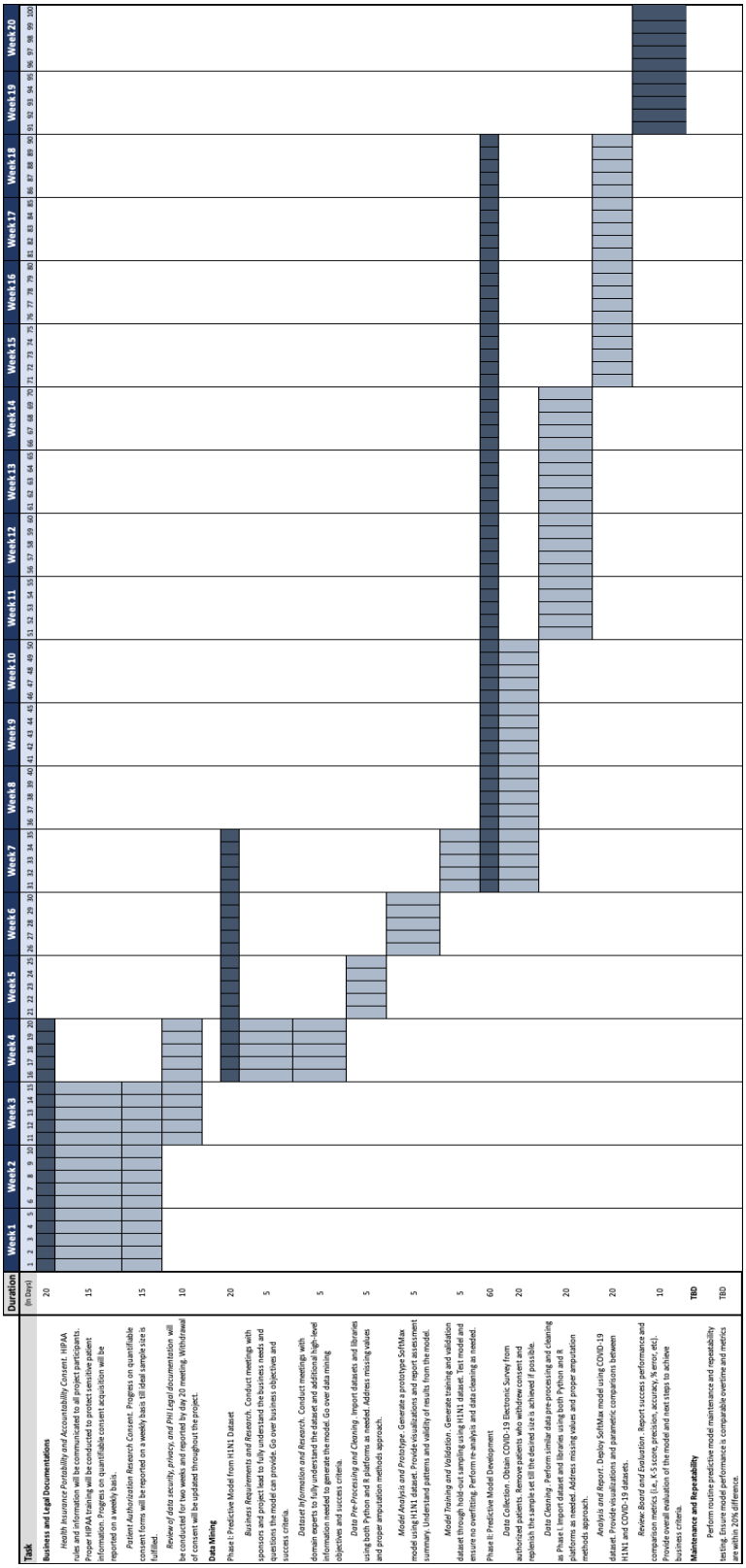


Figure 8

Population density vs H1N1 vaccination by poverty thresholds



Appendix A


Task
Duration
Week1
Week2
Week3
Week4
Week5
Week6
Week7
Week8
Week9
Week10
Week11
Week12
Week13
Week14
Week15
Week16
Week17
Week18
Week19
Week20
Business and Legal Documentations
Health Insurance Portability and Accountability Consent. HIPAA rules and information will be communicated to all project participants. Proper HIPAA training will be conducted to protect sensitive patient information. Progress on quantifiable consent acquisition will be reported on a weekly basis.
Patient Authorization Research Consent. Progress on quantifiable consent forms will be reported on a weekly basis till ideal sample size is fulfilled.
Review of data security, privacy, and PHI Legal documentation will be conducted for two weeks and reported by day 20 meeting. Withdrawal of consent will be updated throughout the project.
Data Mining
Phase I: Predictive Model from H1N1 Dataset
Business Requirements and Research. Conduct meetings with sponsors and project lead to fully understand the business needs and objectives. The model can provide. Go over business objectives and success criteria.
Dataset Information and Research. Conduct meetings with domain experts to fully understand the dataset and additional high-level information needed to generate the model. Go over data mining objectives and success criteria.
Data Pre-Processing and Cleaning. Import datasets and libraries using both Python and R platforms as needed. Address missing values and proper amputation methods approach.
Model Analysis and Prototype. Generate a prototype SoftMax model using H1N1 dataset. Provide visualizations and report assessment summary, Understand patterns and validity of results from the model.
Model Training and Validation. Generate training and validation dataset through hold-out sampling using H1N1 dataset. Test model and ensure no overfitting. Perform re-analysis and data cleaning as needed.
Phase II: Predictive Model Development
Data Collection. Obtain COVID-19 Electronic Survey from authorized patients. Remove patients who withdrew consent and replenish the sample set till the desired size is achieved if possible.
Data Cleaning . Perform similar data pre-processing and cleaning as Phase I. Import dataset and libraries using both Python and R platforms as needed. Address missing values and proper amputation methods approach.
Analysis and Report. Deploy SoftMax model using COVID-19 dataset. Provide visualizations and parametric comparisons between H1N1 and COVID-19 datasets.
Review Board and Evaluation. Report success performance and comparisons results (i.e. Confusion matrix, accuracy, Kstats, etc). Provide overall evaluation of the model and next steps to achieve business criteria.
Maintenance and Repeatability
Perform routine predictive model maintenance and repeatability testing. Ensure model performance is comparable overtime and metrics are within 10% differences
20
15
15
10
20
5
5
5
5
5
60
20
20
20
10
TBD
TBD