Rotman

Executive Summary

The unprecedented COVID-19 pandemic has impacted public health and society, making it crucial for our team of analysts at the Public Health Agency Canada to thoroughly examine the COVID data collected. By analyzing the 2022 COVID data, we can gain valuable insights into the virus's spread, identify trends, and inform effective strategies for future public health initiatives. The goal of this analysis is to predict whether an individual has received a booster or third vaccine.

The dataset provided to the team consisted of 512 columns and was quite disorganized, presenting significant challenges for analysis. Through data processing and cleaning, the dataset was trimmed to 61 columns by replacing blank values with the mode and applying binary and integer ranking. This cleaned dataset will now facilitate effective analysis.

The exploratory data analysis allowed us to dive-deeper into the 61 clean columns. We categorized the predictors into seven key groups: “personal information,” “emotional factors,” “perceptions of government measures,” “attitudes towards Coronavirus,” “preventive actions against COVID-19,” “social distancing practices,” and “views on vaccines.”. These 7 key groups encompass 16 columns that will contribute to the k-nearest-neighbors (kNN) model.

The k-nearest neighbors (kNN) model identifies records that are closest to a selected data point, allowing for comparisons based on their proximity in the feature space. The target variable selected for this analysis is to predict whether an individual has received a booster or third vaccine (*vac_boost_1).* Model tuning using k-fold cross validation yielded the following model parameters; k=22. The 22 nearest neighbors using the Minkowski distance will be used to predict the vaccine target. The model yielded the following results: **Accuracy** (74.34%), **Overall Error Rate** (25.66%), **Sensitivity** (84.03%), **Specificity** (62.24%).

Section 1: Data Preparation

Data preparation is the most important step in data analysis, as the quality of insights depends directly on how well the data is prepared. A thorough data preparation process ensures that the analysis is accurate and meaningful. The provided COVID dataset was highly disorganized, containing numerous NULL columns that hindered initial analysis. Cleaning this data was essential to remove empty fields, handle missing values, and ensure the dataset was usable.

Starting with 512 columns:

The dataset contained 512 columns, this needed to be reduced to a reasonable amount to begin an efficient and effective exploratory data analysis (EDA). 139 columns contained more than 50% NULL values and were removed. NULL values in this dataset were in 3 different forms; ‘ ‘, ‘__NA__’, and actual NaN values. While a 50% threshold is typically aggressive, it was justified in this case due to the dataset’s excessive number of columns, allowing for a more focused and effective analysis. Additionally, 3 more columns (qweek, RecordNo, endtime) were removed, as they served only as unique identifiers from the survey and do not contribute meaningful information for the analysis. Furthermore, 2 more columns (future_1, future_2) were removed due to ambiguous metadata, which made their interpretation unclear. Understanding each column’s meaning is essential for evaluating how it influences the target variable, so columns lacking clarity are better excluded. Removing these ensures the dataset is focused on variables that meaningfully contribute to the analysis.

78 remaining columns:

With the dataset now reduced to 78 columns, in-depth analysis is possible. This set allows us to thoroughly examine each column, identifying any duplicates or unusual anomalies that might impact the quality of insights. There were no duplicates in the dataset, reassuring that each record is unique.

D1_health columns contain information about pre-existing medical conditions. There are 13 columns for 13 various conditions, 1 column for prefer not to say, and 1 column for none of the above. Looking into the data more shows that there are 3028 records that are NULL for all 15 columns. This was not removed in earlier steps since it was below the 50% threshold. Replacing these NULL values will be tough since there is no information for any of them. There are some records with none of the above, so it is not safe to assume the blanks are none of the above. Based on this, all 15 columns were removed.

I2_health column has the number of people outside household contact within 6-feet. There are 2007 NULL values, and this will be difficult to impute. Additionally, had_covid had many ‘ ‘ and NaN values, which is very difficult to impute since it’s a binary variable. These columns contain useful information but should not be prioritized over the other 60+ cleaner columns.

61 remaining columns:

61 columns remain and need to be cleaned for EDA. Using the provided metadata, integer mappings were created for all columns that contained strings for survey responses (1-Agree should be 1). Mode imputation was used for any blank/NaN values. After data preparation, the dataset has been refined to 61 columns, all of which are integer-based (either binary or ranking values) and free of NULL values. This clean, consistent dataset is now primed for in-depth exploration to identify the top features that will be most effective for the kNN model.

Section 2: Exploratory Data Analysis

After preparing a cleaned dataset, we explored different target variables to identify those that could provide valuable and interesting insights. Our group was particularly interested in examining whether the target population chose to receive a booster shot or third dose of the vaccine. This dose was not mandated but was primarily taken by individuals who believed in the vaccine's positive effects.

Target variable:

The target variable, *vac_boost_1*, is a binary categorical variable capturing whether an individual received a COVID-19 booster or third dose. Among the 6,430 records, approximately 56.7% represent individuals who received the booster or third dose, while 43.3% did not, resulting in a relatively balanced sample across both categories.

Our analysis aims to understand how various factors influence people's decisions regarding this vaccination step. We categorized the predictors into seven key groups: “personal information”, “emotion factor”, “view towards government measures”, “view towards Coronavirus”, “preventing actions against COVID-19”, “social distancing”, and “view towards vaccines”. This structure enables a comprehensive exploration of the circumstances and attitudes shaping vaccination decisions.

Personal Information:

We examined various personal information, such as weight, age, employment status, region, and more. However, the most interesting and useful patterns emerged primarily from the variables related to age and household size.

The average age of our study population is approximately 50 years, with a standard deviation of 18 years (Appendix - Ref #1). We decided to categorize the ages into four groups: Young Adults (18-29), Adults (30-49), Middle-Aged Adults (50-64), and Older Adults (65-90). We then created a graph to visualize the distribution of individuals who received booster shots within each age group. The graph reveals a clear pattern: as age increases, the percentage of people who took booster shots also rises. (Appendix - Ref #2) .

Looking at the household_size population, we observe household sizes ranging from 1 to 8 members. A clear trend emerges: larger households tend to have a lower percentage of individuals receiving the booster shot, while smaller households (especially those with one or two members) show higher booster uptake. This result might seem surprising at first, but it could be explained by the idea that individuals in larger households may feel less at risk, as they likely have a built-in support network or share the belief that they can rely on one another for care (Appendix - Ref #3).

Emotion Factor:

Throughout our analysis, we were also interested in examining whether an individual's level of happiness with their current life status might have influenced their decision to receive a booster shot. Although we are not suggesting that an individual’s happiness directly causes changes in their beliefs about vaccination, we thought it would be insightful to explore this potential relationship.

In the dataset, we used a variable called **'cantril_ladder,'** which ranges from 0 (feeling very bad) to 10 (feeling very good). A clear pattern emerged: as the level of happiness increases, the rate of booster uptake also rises. However, an unexpected anomaly was observed at the highest level (10), where booster uptake decreased for reasons that are unclear. (Appendix - Ref #4)

View towards Government Measures:

Another factor we were interested in exploring is whether individuals believe that the government's measures for handling COVID-19 were effective, which is represented by the variable **WCREX_1** in the dataset. We believe this variable is significant, as it likely influences an individual's decision to receive a booster shot—people who perceive government measures as effective may be more inclined to get vaccinated.

Based on the results, we observed a clear pattern: there is a lower percentage of people receiving booster shots among those who do not trust the government's COVID-19 protocols. (Appendix - Ref #5)

View Towards Coronavirus:

There are several important variables in the dataset related to individuals' beliefs about COVID-19 that we aimed to explore (scaling from 1(Disagree) - 7 (Agree)). Specifically:

- **r1_1**: "Coronavirus is very dangerous for me."
- **r1_2**: "It is likely that I will get coronavirus in the future."
- **r1_3**: "Getting a vaccine will protect me against coronavirus."
- **r1_9**: "Getting a vaccine will protect others against coronavirus."

By conducting a correlation matrix on these variables, we identified key relationships. Notably, the **strongest correlation** is between **r1_8** and **r1_9**, with a value of **0.69**. This strong correlation suggests that individuals who believe vaccination protects others are also likely to believe it offers personal protection. (Appendix - Ref #6)

After visualizing the data through graphs, we identified the following trends (Appendix - Ref #7)::

1. **Booster Uptake vs. COVID-19 Danger Perception (r1_1):**
   - ○ Individuals who perceive COVID-19 as very dangerous are more likely to receive a booster shot.
2. **Booster Uptake vs. Likelihood of Infection (r1_2):**
   - ○ The belief that one is likely to contract COVID-19 in the future has little to no influence on the decision to get a booster.
3. **Booster Uptake vs. Vaccine Protects Me (r1_8):**
   - ○ People who strongly believe that the vaccine protects them are significantly more likely to get a booster shot.
4. **Booster Uptake vs. Vaccine Protects Others (r1_9):**

- The belief that the vaccine protects others also has a strong positive impact on booster uptake.

Preventing Actions Against COVID-19:

We hypothesize that an individual's preventive behaviors against COVID-19 may provide insights into their views on receiving a booster. To explore this, we examined key factors related to COVID-19 prevention actions.

The variable ***i9_health*** assesses willingness to self-isolate after feeling unwell. Among individuals willing to self-isolate, over 60% received a third COVID-19 vaccine dose, while this figure dropped to 46% among those unwilling to self-isolate. This suggests that individuals with self-isolation awareness may be more inclined to receive a booster dose (see Appendix, Ref #8). A cross-tab analysis between *i9_health* and the target variable yielded a p-value close to 0, reinforcing *i9_health* as a significant predictor. Additionally, we considered a potential correlation between *i9_health* and *household_size*, hypothesizing that individuals in larger households may be more inclined to self-isolate to protect family members. However, the observed correlation is low (approximately 0.1), indicating minimal overlap. Therefore, it is suitable to retain both variables in the dataset without redundancy concerns.

Another set of predictors examines external actions taken to prevent COVID-19, such as wearing face masks, using hand sanitizer, and avoiding crowded areas. This set initially included 20 predictors, which were grouped by us into four categories based on action similarities: *face_mask*, *dangerous_contact* (avoiding high-risk crowded places), *contact* (avoiding small social gatherings), and *sanitation* (hand washing, sanitizer use). Since over 95% of respondents reported occasional mask use and sanitation practices, we excluded these variables from further analysis and focused on *contact* and *dangerous_contact*.

Cross-tab analysis of these two variables with the target variable yielded near-zero p-values, indicating potential significance. Given their relevance to social distancing, we assessed them alongside other social distancing factors to check for correlation. Although most correlations were weak, *dangerous_contact* and *contact* showed a moderate correlation of 0.45. To reduce redundancy, we opted to retain ***dangerous_contact*** only (Appendix, Ref #9). Specifically, among individuals who avoided high-risk places, 58% received a booster, compared to only 31.4% of those who did not (Appendix, Ref #10).

Social Distancing:

We identified several social distancing factors that may help predict attitudes toward boosters. We believe that individuals who practice social distancing may be more risk-averse and proactive in protecting themselves, making them more likely to opt for a booster. Notably, these social distancing variables show weak correlations with one another (Appendix, Ref #9), suggesting that they can be included together in the analysis without significant redundancy.

The first set of variables, ***soc1_1*** and ***soc1_2***, captures an individual's participation in social gatherings with more than six people, both outdoors and indoors. Both variables demonstrate trends in relation to the target. Individuals who avoid large gatherings, whether outdoors or indoors, have a higher likelihood (approximately 60%) of receiving a booster. In contrast, those who frequently participate in large gatherings are less likely to have received a booster, with only around 45% doing so (see Appendix, Ref #11 and Ref #12).

Another ordinal variable, ***Vent_3***, indicates how frequently individuals avoided social gatherings in the past seven days, with a scale from 1 (Always) to 5 (Not at all). This variable also shows a significant relationship with the target. Individuals who more frequently avoided social gatherings had a higher likelihood of receiving a booster (see Appendix, Ref #13).

View Towards Vaccines:

Individuals' views on vaccines may provide insights into their willingness to receive a third dose. To explore this, we examined several vaccine-related factors in relation to the target variable.

The first set of predictors captures attitudes towards vaccine side effects and effectiveness. ***vac2_2*** reflects concerns about potential side effects, showing that individuals with fewer concerns about side effects are more likely to receive a third dose (Appendix, Ref #14). ***vac7*** measures trust in vaccines, revealing that individuals with high trust in vaccines are significantly more likely to take boosters, with around 80% of those expressing strong trust opting for a booster (Appendix, Ref #15). Since *vac2_2* and *vac7* have a correlation of only 0.35, it is feasible to include both predictors in the prediction without redundancy.

The second set of predictors examines opinions on whether COVID-19 vaccinations should be mandatory for certain high-risk groups. Specifically, *vac_man_1* captures views on mandatory vaccination for healthcare workers, while *vac_man_3* reflects views on frontline emergency workers. These two groups face heightened exposure to the virus, and support for mandatory vaccination for them may indicate strong beliefs in vaccine effectiveness. Cross-tab analysis showed that both variables have near-zero p-values with the target, indicating potential significance. However, they also have a strong correlation of 0.88 with each other, so we chose to retain ***vac_man_3***, as it has a slightly higher correlation with the target. Individuals who believe that frontline emergency workers should be vaccinated are more likely to receive a booster, with around 70% opting for it, compared to just 30% among those who do not hold this belief (Appendix, Ref #16).

Summary of EDA:

Our group focused on understanding factors influencing whether the target population opted to receive a booster shot or third dose of the COVID-19 vaccine. After thoroughly examining the relationships between various predictors and the target variable, we selected 16 predictors to model *vac_boost_1* in the following section. A summary table of our selected variables across each category is provided below:

Category	Selected Variables
personal information	age, household_size
emotion factor	cantril_ladder, PHQ4_3
view toward government measures	WCRex1
view toward Coronavirus	r1_1, r1_2, r1_9
preventing actions against COVID-19	i9_health, dangerous_contact
social distancing	Soc1_1, Soc1_2, Vent_3
view toward vaccines	vac2_2, vac7, vac_man_3

Section 3: K-NN Model Tuning

The K-Nearest Neighbour algorithm is used to fit the dataset to classify whether a person has had a booster or a third dose of a Coronavirus (COVID-19) vaccine. To classify a new record, the method finds “similar” records in the training data. These “neighbours” are used to derive a classification for the new record by voting.

We firstly converted the categorical variables into dummy/indicator variables from the preprocessed dataset from previous steps. The dataset is then partitioned into training data (80%) and validation data (20%). Next, 10-fold cross-validation is applied onto the training dataset, splitting the dataset into 10 groups, and the model would be trained and tested 10 separate times so each data subgroup would be the test dataset once. Each time when the model ran on the dataset, an accuracy score was calculated. Using cross-validation, the mean score of the model is about 74.49%. This is a representation of how the model will perform on hidden data.

Distance-based algorithms, like kNN, can be easily impacted by feature magnitudes. Features with larger magnitudes could negatively impact the distance calculation, resulting in biased results. Thus, feature scaling is crucial and needs to be applied before fitting the model. By looking at the dataset, the magnitudes for each column are pretty close to each other except for age. We decided to train two models, one on the scaled dataset, one on the original dataset to compare their accuracy scores.

To find the optimal value of parameter K, Grid Search cross-validation is used to tune the parameter to get the best accuracy result. We specified a range of k values for testing from 1 to 30 to train the model multiple times. The new model then used grid search by taking in a new kNN classifier to find the optimal k value for neighbours. Based on Appendix - Ref #17, it shows accuracy scores for different k values for two models. For the model with scaled data, the highest accuracy score is about 74.34% with a k value of 22, whereas the best accuracy score for the model with unscaled data is about 74.65% with a k value of 7, which is slightly higher. Then we rerun the algorithm using the chosen k, and got accuracy scores of 74.34% and 73.56% for models with scaled data and with original data respectively for validation dataset. From the testing results, we would choose k=22 with scaling the dataset as it produces a good score for testing data which means the model is not overfitted with the training data while keeping the predictor information at the same time.

Section 4: Model Testing

The confusion matrix (Appendix - Ref #18) provides a breakdown of the model's performance, classifying the data into four categories:

- True Positive (TP): People who received the booster correctly predicted as having received it (600).
- True Negative (TN): People who did not receive the booster correctly predicted not to have received it (356).
- False Positive (FP): People who did not receive the booster incorrectly predicted as having received it (216).
- False Negative (FN): People who received the booster were incorrectly predicted to need it (144).

Accuracy(74.34%) is the ratio of correctly predicted observations to the total observations:

The **Overall Error Rate**(25.66%) represents the percentage of incorrect predictions:

Sensitivity (84.03%) is the ability of the model to identify all actual positives correctly:

Specificity(62.24%) is the ability of the model to identify all actual negatives correctly:

The model correctly identifies 62.24% of individuals who have not received a booster. This indicates a lower ability to detect true negatives than true positives. Low specificity may mean the model incorrectly categorizes individuals as having received a booster when they have yet to, leading to unnecessary follow-up.

The **False Positive Rate**(37.76%) is the proportion of actual negatives that were incorrectly classified as positive:

37.76% of people who have yet to receive the booster were wrongly predicted as having received it. This could be a concern, as it might lead to incorrect assumptions about vaccine coverage.

The **False Negative Rate**(8.86% is the proportion of actual positives that were incorrectly classified as unfavourable:

15.97% of people who received the booster were predicted as not having received it. This is important because it means the model needs to include vaccinated people, which could affect public health campaigns.

Overall Interpretation:

The **high sensitivity** suggests that the model effectively identifies those who have received the booster. This is important for ensuring that the number of vaccinated individuals is accurately tracked. However, the **specificity** is relatively lower, meaning the model struggles to identify those who have yet to receive the booster correctly. This could lead to incorrect assumptions about how many people still need a booster. The **false positive and false negative rates also indicate potential misclassification issues**. In public health, **false negatives** (missing people who need a booster) can be especially problematic.

Section 5: Conclusion

This project aimed to gain insights from Canada's 2022 COVID-19 Behavior Data, which captures a range of COVID-19-related behaviors, including symptoms, testing, isolation, social distancing, and vaccination. Through a systematic process of data preparation, exploratory data analysis, and K-Nearest Neighbors (K-NN) model training and testing, we developed an effective classification model to predict the likelihood of receiving a COVID-19 booster or third dose.

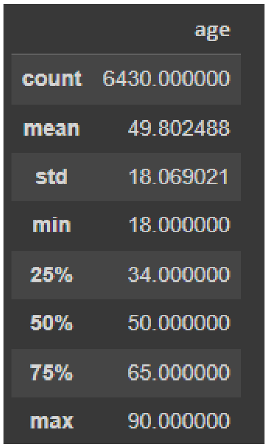
Our data preparation phase focused on transforming the initial dataset of 512 columns into a streamlined set of 61 columns, all of which are integer-based (either binary or ranking values) and free of NULL values. Exploratory data analysis allowed us to identify *vac_boost_1* as our target variable, a binary indicator reflecting whether an individual received a booster. We were particularly interested in this non-mandated dose, often chosen by individuals who believed in the vaccine's benefits. Predictors were organized into seven categories: "Personal Information," "Emotional Factors," "Views on Government Measures," "Perspectives on COVID-19," "Preventive Actions," "Social Distancing," and "Attitudes Towards Vaccines." From these, we selected 16 relevant predictors for our model.

After selecting our target variable and 16 key predictors, we applied K-Nearest Neighbors (K-NN) classification to build a predictive model, using an appropriate train-test split. An optimal k-value of 22 was selected through grid search cross-validation to maximize accuracy. With this k-value and scaled data, our model achieved an accuracy of 74.34% on the testing set.

Model evaluation using a confusion matrix provided additional performance insights. High sensitivity suggests that the model effectively identifies individuals who received the booster, which is crucial for accurately tracking vaccinated individuals. However, the model's lower specificity indicates difficulty in correctly identifying those who did not receive the booster, which could lead to misestimates of the remaining unvaccinated population. False positive and false negative rates further reveal potential misclassification issues; in a public health context, false negatives (missing individuals needing a booster) are particularly concerning. To improve model performance, incorporating additional predictors or exploring alternative classification methods could enhance accuracy and better balance sensitivity and specificity.

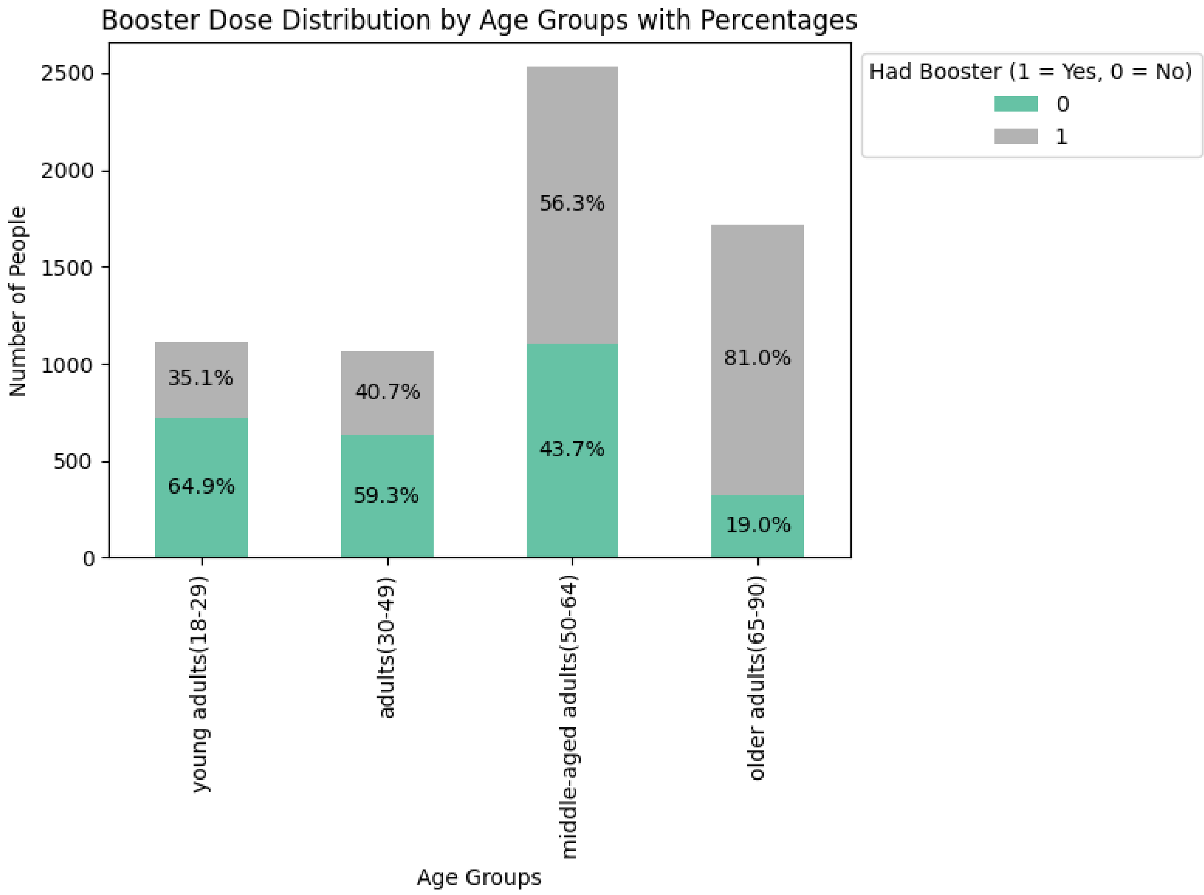
Appendix

Reference 1: age

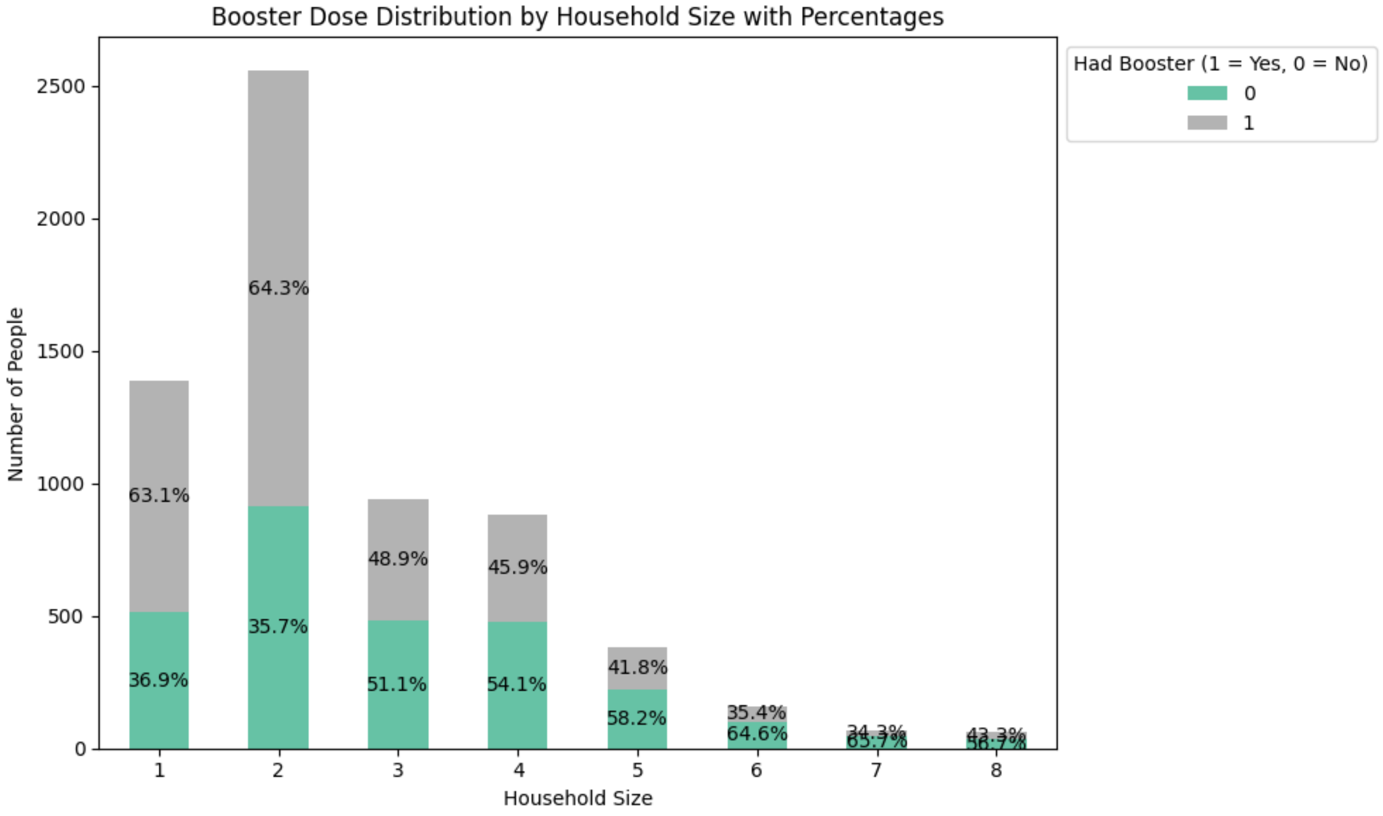


	age
count	6430.000000
mean	49.802488
std	18.069021
min	18.000000
25%	34.000000
50%	50.000000
75%	65.000000
max	90.000000

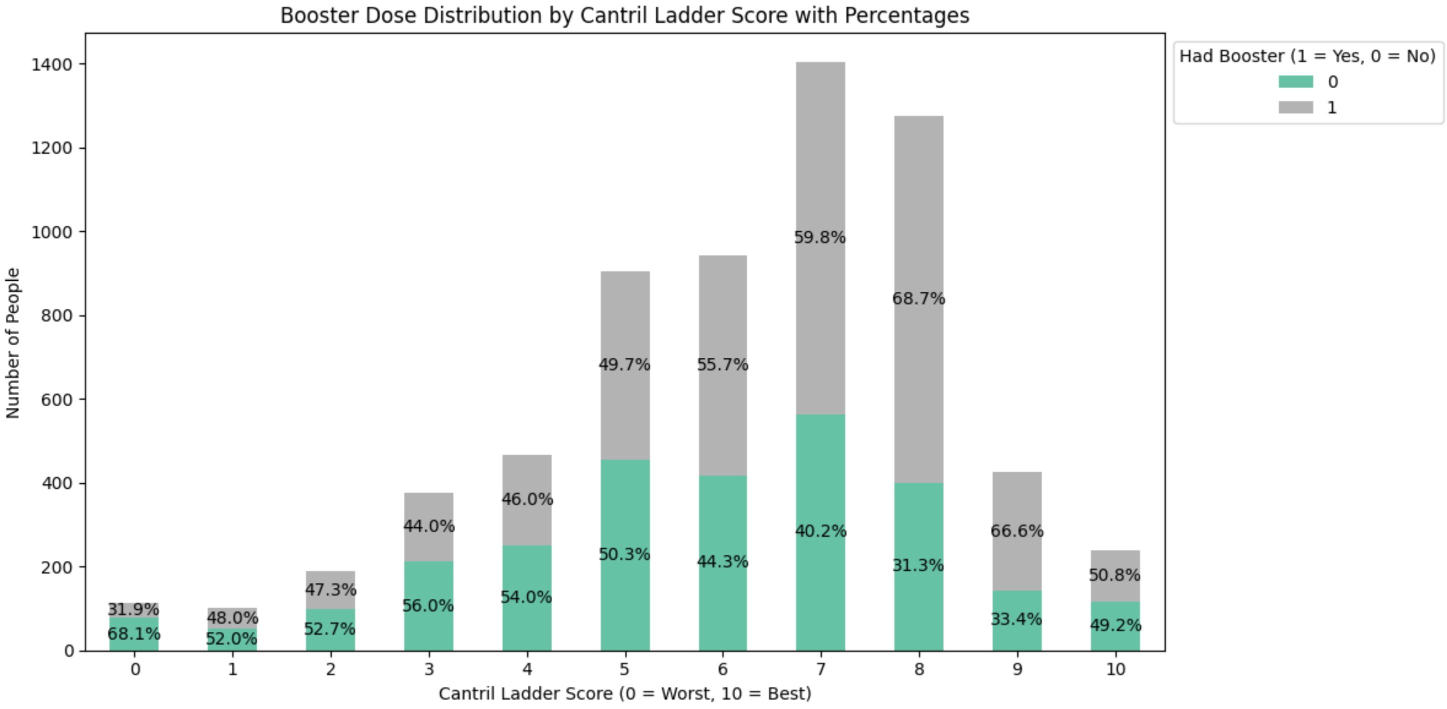
Reference 2: Booster Dose Distribution by Age Groups



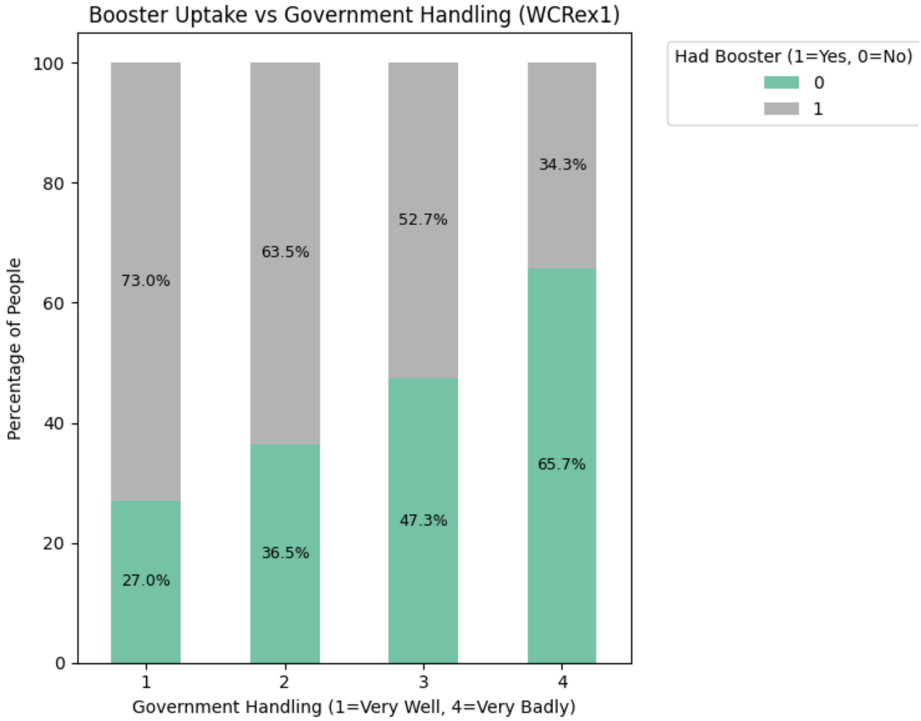
Reference 3: Booster Dose Distribution by Household Size



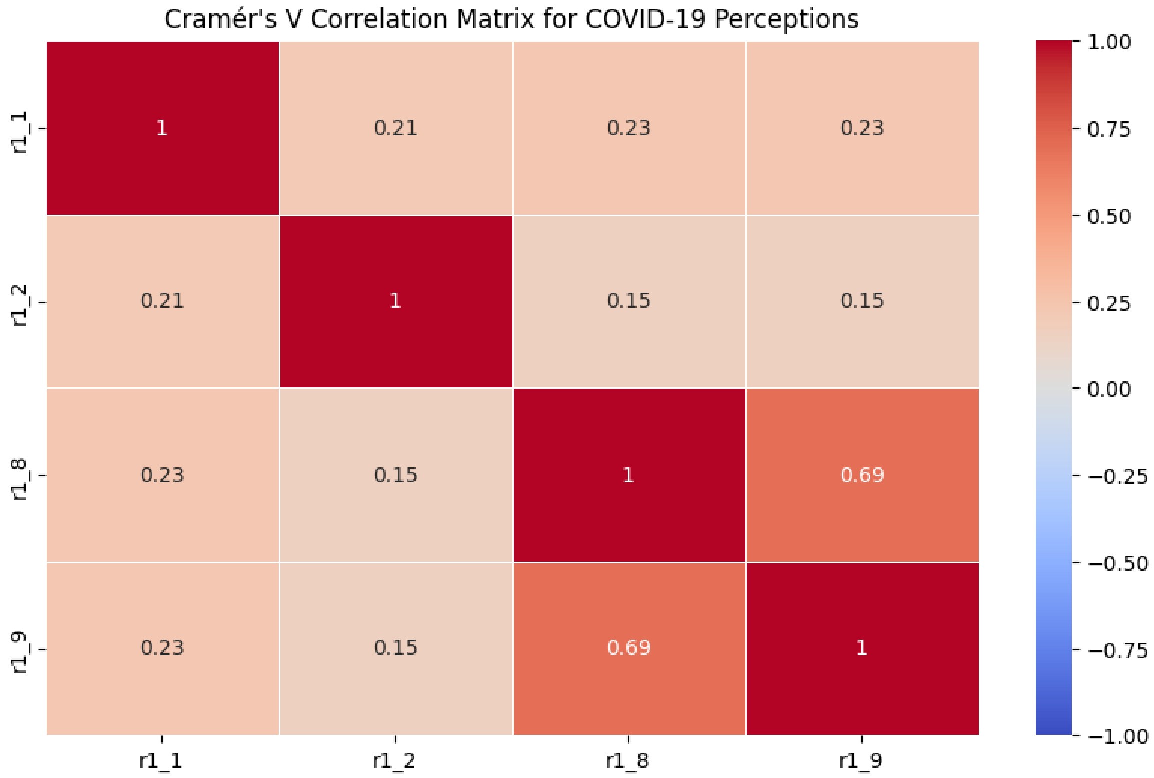
Reference 4: Booster Dose Distribution by Cantril Ladder Score



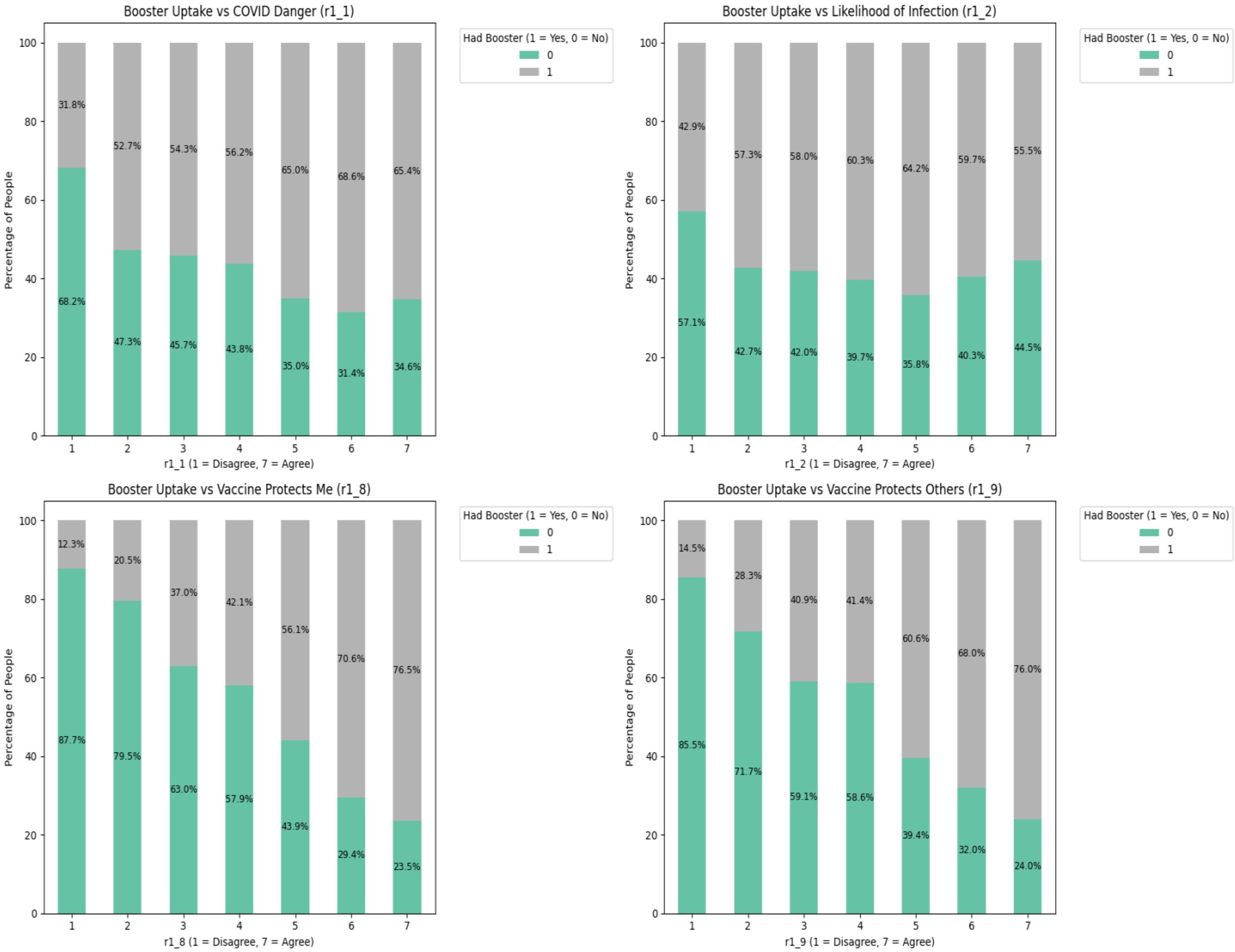
Reference 5: Booster Uptake vs Government Handling (WCRex1)



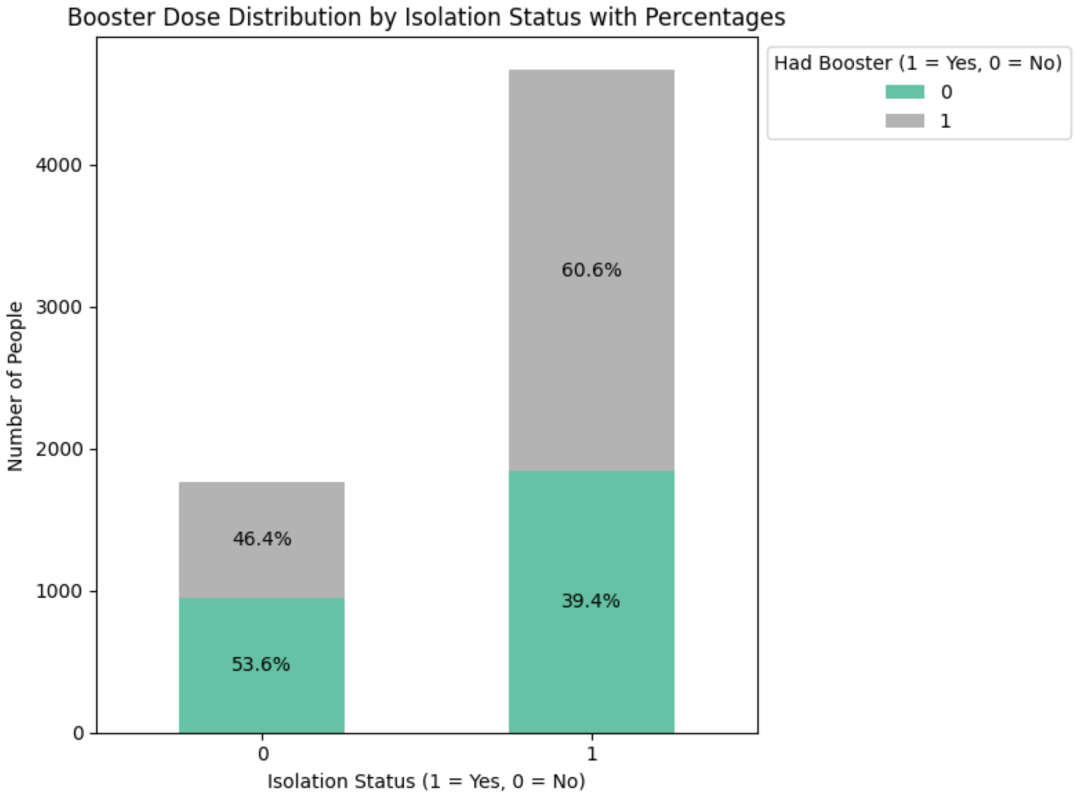
Reference 6: Correlation Matrix for COVID-19 Perceptions



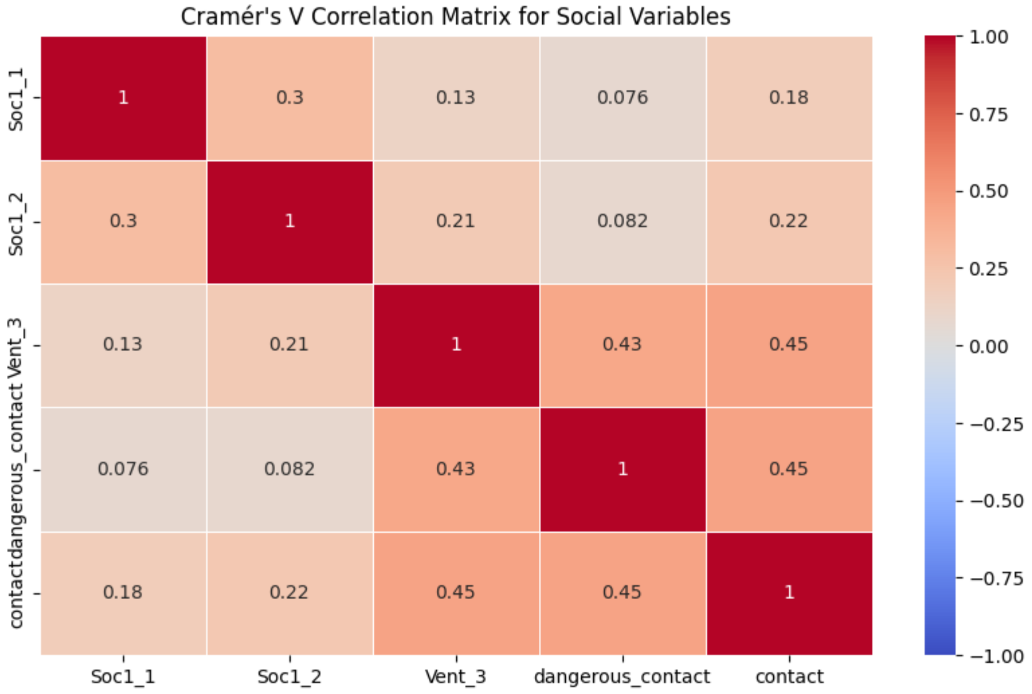
Reference 7: Booster Uptake vs. Factors about Views towards Coronavirus



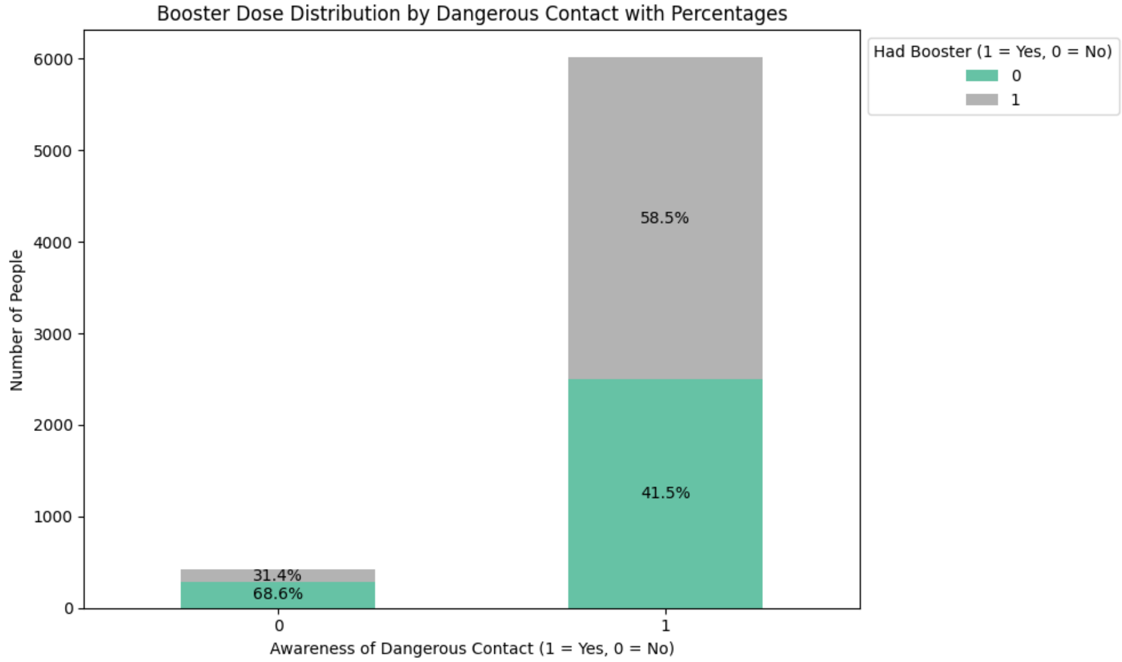
Reference 8: Booster Dose Distribution by i9_health



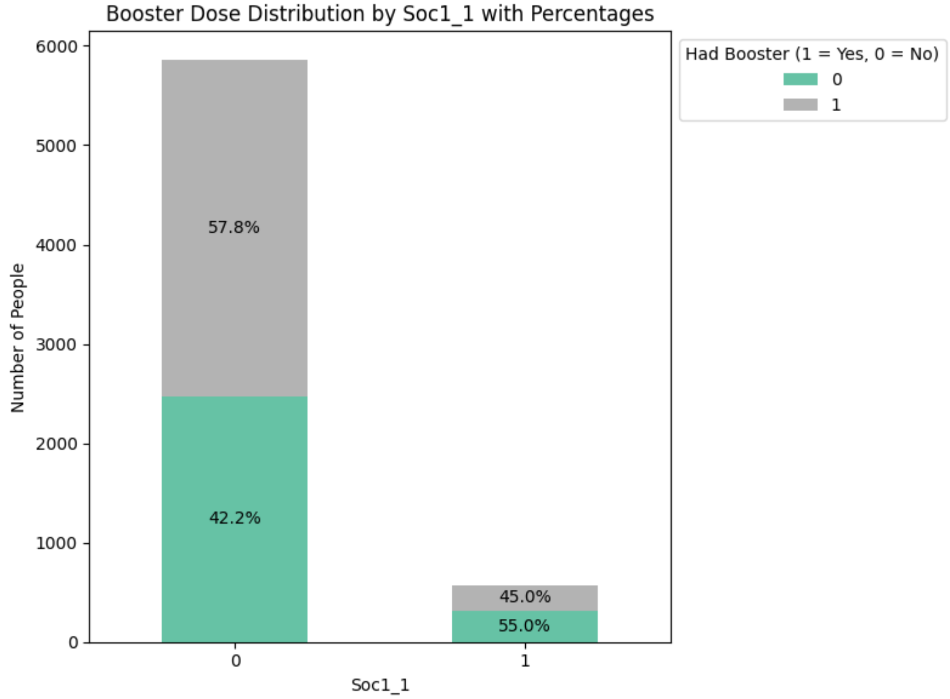
Reference 9: Correlation Matrix for Social-related Factors



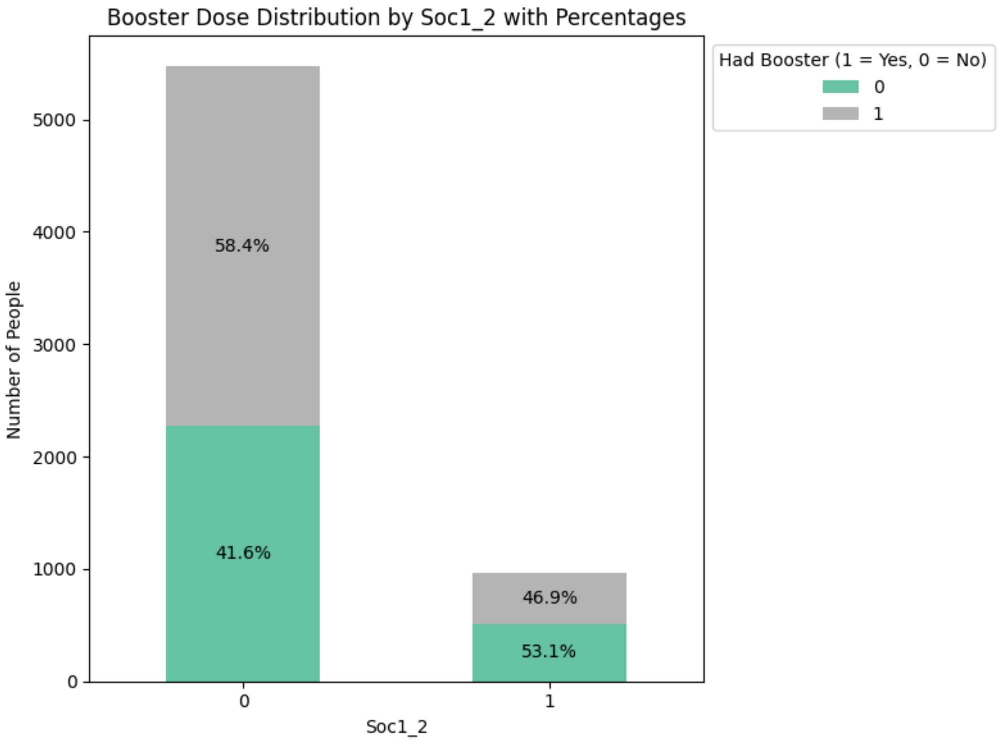
Reference 10: Booster Dose Distribution by dangerous_contact



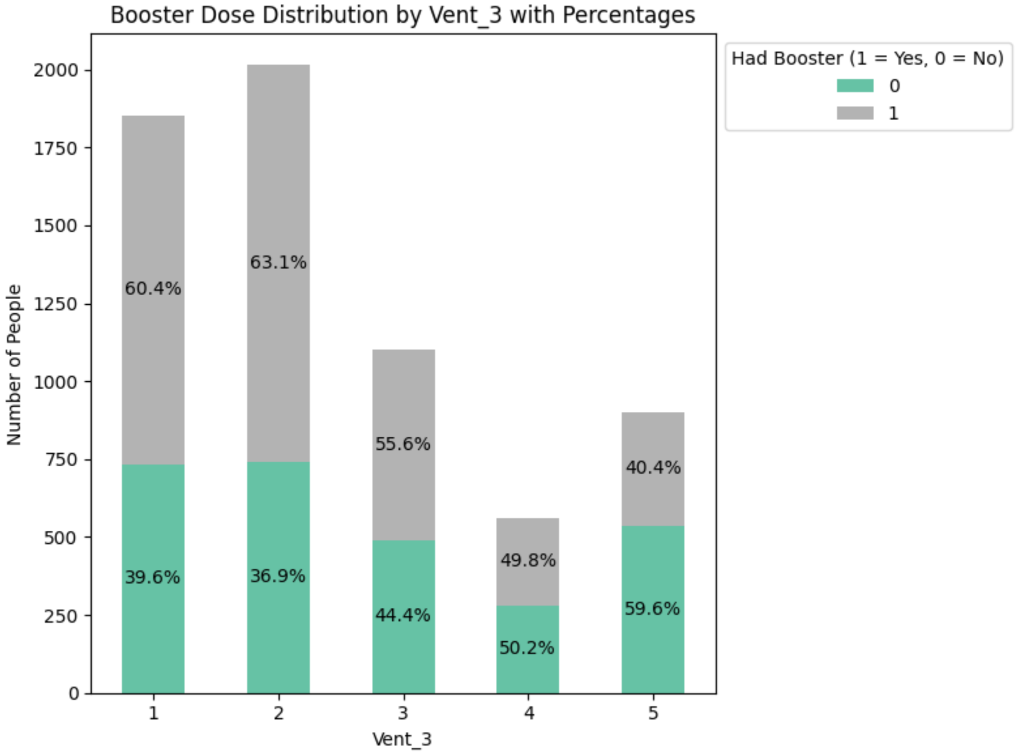
Reference 11: Booster Dose Distribution by Soc1_1



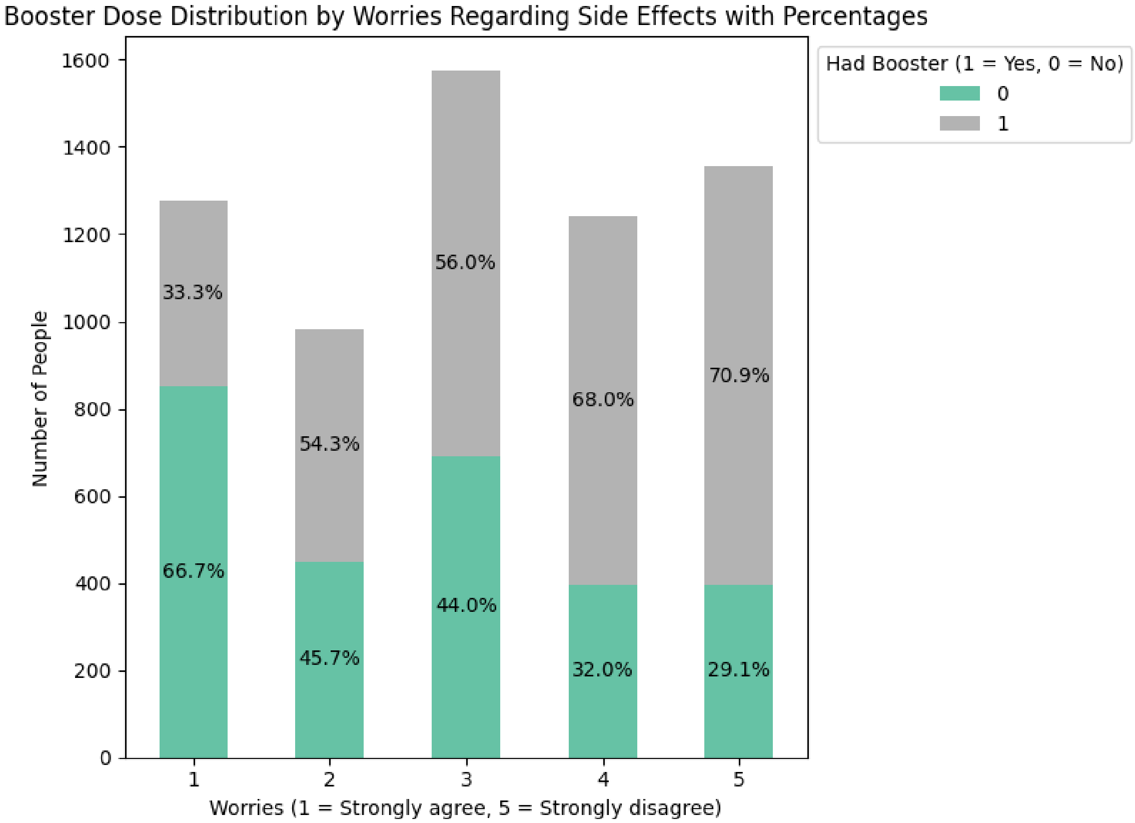
Reference 12: Booster Dose Distribution by Soc1_2



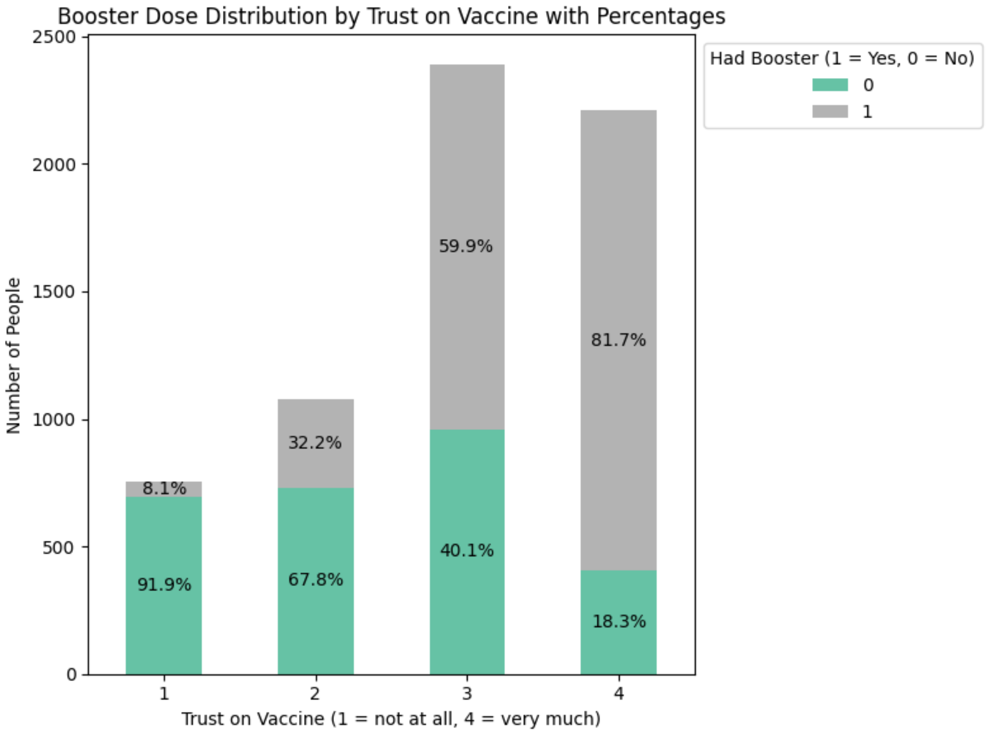
Reference 13: Booster Dose Distribution by Vent_3



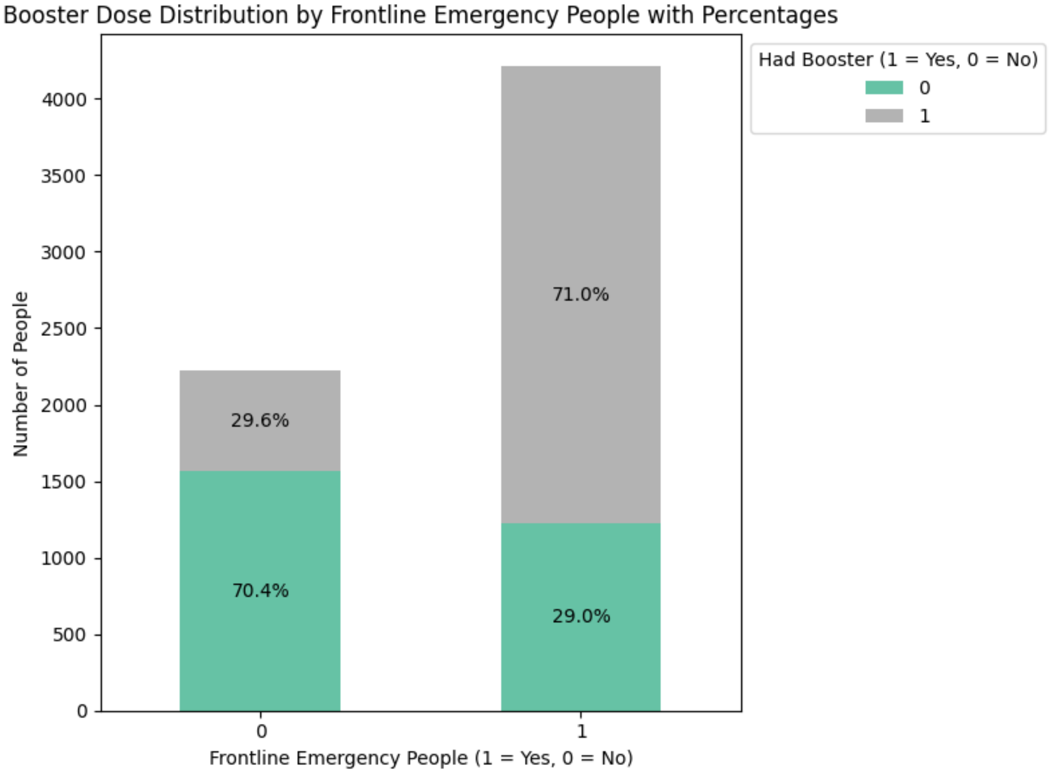
Reference 14: Booster Dose Distribution by Worries Regarding Side Effects



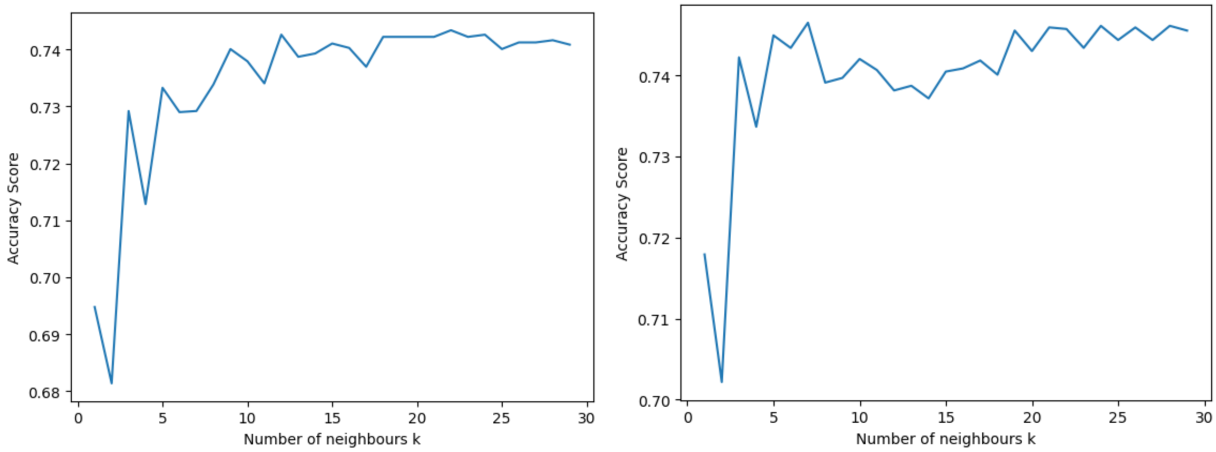
Reference 15: Booster Dose Distribution by Trust on Vaccines



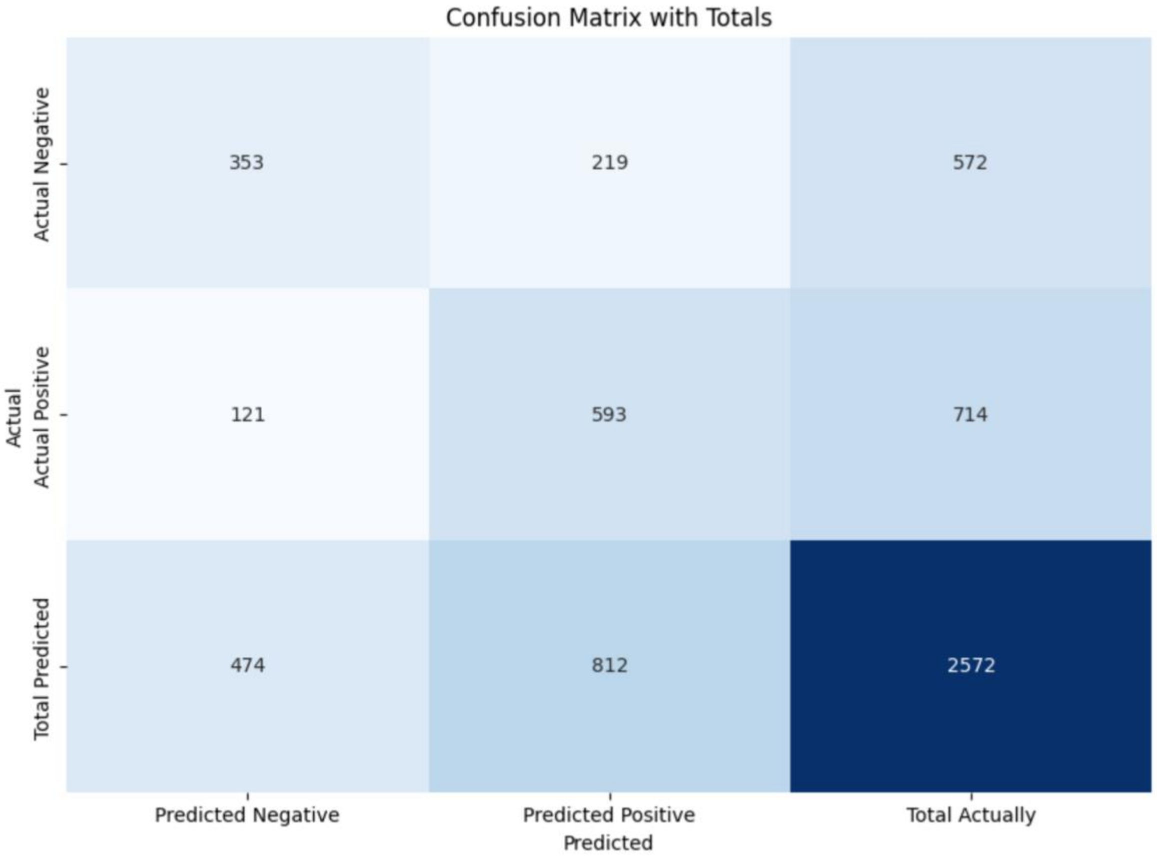
Reference 16: Booster Dose Distribution by vac_man_3



Reference 17: Accuracy score between model with scaled data (left) versus model with unscaled data (right)



Reference 18: Confusion Matrix



Confusion Matrix with Totals

	Predicted Negative	Predicted Positive	Total Actually
Actual Negative	353	219	572
Actual Positive	121	593	714
Total Predicted	474	812	2572

Actual
Predicted