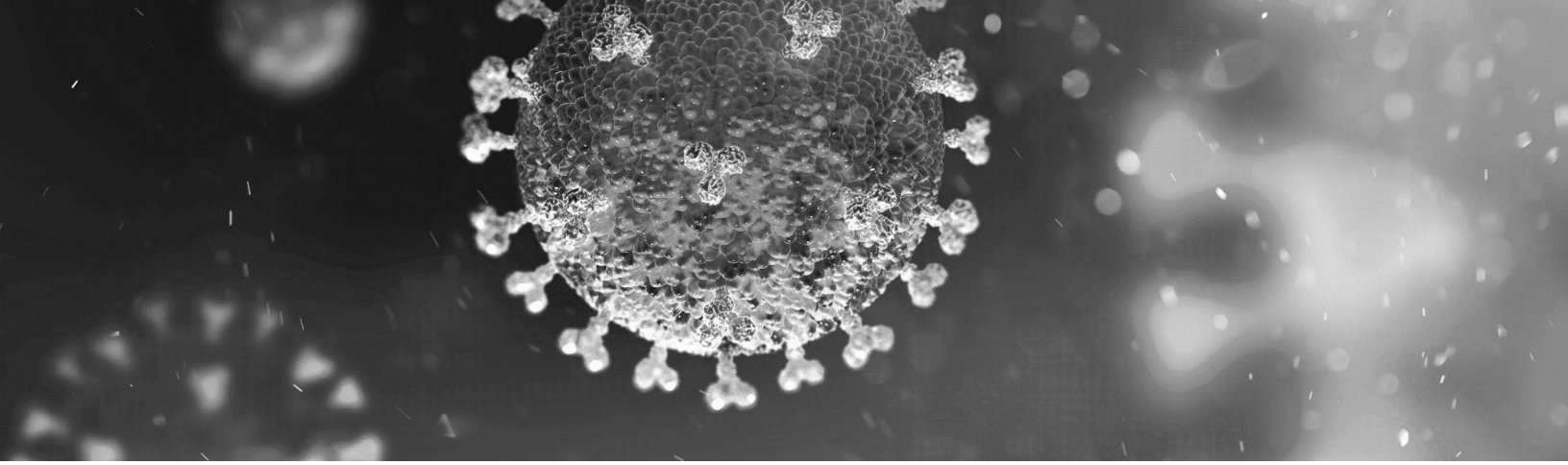


GROUP 1: VACCINE INTENT & IMPACT OF THE COVID-19 PANDEMIC

By: Kaitlyn Bleiweiss, James Butler, Janelle Gillis, Brett Vaccaro & Brian Weiss

Table of Contents

Executive Summary	2
Introduction	3
Data Description	4
Data Preparation	5
Descriptive Data Analysis	6
Predictive Data Analysis	24
Business Recommendations	28
Appendix	30
References	35

Executive Summary

This analysis seeks to explore pandemic and vaccine related findings from publicly available data captured from the COVID-19 Household Pulse Survey conducted by the United States Census Bureau from January 1 to July 5, 2021.

GOAL:

Successfully explore and describe the effects of the COVID-19 pandemic on United States residents and identify a model that can accurately predict vaccine intent.

OBJECTIVES:

- What are the effects of COVID-19 on the population? Can a descriptive explanation be created that defines the different effects on areas such as childcare, education, employment, energy use, food security, health, housing, and household spending?
- Select and build model(s) that will accurately predict vaccine intent.

ANALYTICAL METHODS:

- Descriptive Analysis: Chi-Square Tests of Independence, T-Tests, Frequency Analysis, and K-Modes Clustering
- Predictive Analysis: Multivariate Imputation, Decision Trees, Random Forests, and Extreme Gradient Boosting
- Most Effective Model: Extreme Gradient Boosting

CONCLUSION:

Through in-depth descriptive and predictive analysis, the effects of COVID-19 were defined in part using seven clusters or personas. In addition, an Extreme Gradient Boosting model allowed for the effective prediction of vaccine intent in this population. Some key findings included abundant evidence that food insecurity has increased during the pandemic. This may be related to unexpected job losses, income loss, or other unexpected life events during this time. In addition, it was alarmingly discovered that those with higher levels of anxiety were disproportionately reporting delaying needed medical care, which implies that those who potentially needed medical care the most delayed obtaining it during this pandemic. In conclusion, it is suggested that PwC consider using the findings of this report to aid in the building of effective and personalized communication, client, and employee strategies. One example could be a vaccination education project that benefits from personalizing

communication to audience segments that align with clusters three and four, as well as those identified in the vaccine intent predictive model.

Introduction

In 2020, the COVID-19 virus shut down life as we knew it and the full effects are still developing. The resulting global pandemic has severely disrupted our social and economic environment. Our best efforts to curb the spread of this infectious disease are centered around two main strategies - physical distancing to minimize new infections and the rapid development of a vaccine to inoculate the population.

As a result, it is increasingly important to use existing research and available data to examine the pandemic effects on society. Some potential avenues to investigate are how COVID-19 has impacted key economic and social areas, such as childcare, education, employment, energy use, food security, health, housing, and household spending. Understanding these implications of the pandemic can help inform future intervention, government and private spending, and business trends as well as aiding in the recovery of the general populous after experiencing a devastating world-wide crisis.

Throughout 2021, there has been a mass effort to administer COVID-19 vaccines, but it has become evident that some people in the United States are still hesitant about receiving the COVID-19 vaccine. This trend drives the need to investigate what factors contribute to vaccine hesitation to help define future business strategy and, on a humanitarian level, effectively inoculate enough of the world to provide herd immunity and put an end to this pandemic.

For this data analysis project, PricewaterhouseCoopers (PwC) has specifically requested that COVID-19 data be explored, including the dataset from the Household Pulse Survey conducted by the United States Census Bureau. There are two main problems to address:

- *What are the effects of COVID-19 on the population? Can the data descriptively explain the different effects on areas such as childcare, education, employment, energy use, food security, health, housing, and household spending?*
- *What are the factors that predict vaccine intent? Select and build model(s) that will accurately predict vaccine intent.*

This analysis seeks to explore findings from publicly available data captured from the COVID-19 Household Pulse Survey conducted by the United States Census Bureau from January 1 to July 5, 2021, which includes data from both Phase 3 and 3.1. This survey was sent to residents of the United States of America during weekly increments to assess changing

perceptions, evolving effects of the pandemic, and in later phases of the survey whether vaccinations were received. The goal is to help inform a direction and potential solution to the identified business problems, including understanding the effects on COVID-19 in the United States and the factors that will help to predict vaccine intent. PwC has expressed that understanding the impacts of COVID-19 is critical to making strategic decisions for themselves and on behalf of their clients. COVID-19 has had wide-ranging impacts on a variety of sectors in business and in individual lives. It is important to understand these effects, describe them in tangible ways, and expand the possibility of making changes based on empirical findings.

In addition to the business needs, it is in the best interest of organizations and the population to increase vaccine acceptance. The more people who get vaccinated, the better equipped society will be to overcome COVID-19 and end this pandemic. At this point, it is of the utmost importance to understand why some individuals may be hesitant or resistant to receiving the vaccine. If key characteristics, personas, or predictors of vaccine intent are identified, it may be possible to create ways to mitigate fears of the vaccine or address other concerns specific to each population.

Data Description

Before analysis could begin, a detailed data review was required. Data understanding and pre-processing are one of the most difficult but imperative stages of data analysis. As noted, this project will rely on the publicly available COVID-19 Household Pulse Survey data. The United States Census Bureau, in collaboration with other federal agencies, began administering the Household Pulse Survey in April of 2020 (Fields et al., 2020). Since the initial roll-out, there have been multiple phases of data collection starting with Phase 1 and currently ongoing in Phase 3.2 with each phase utilizing variations of the original survey instrument.

The project scope focuses on data beginning in January 2021 and ending in the first week of July 2021 encompassing the last six weeks of Phase 3 and the first six weeks of Phase 3.1. The Phase 3 survey instrument had 204 variables while Phase 3.1 had 239 variables. There were 459,235 and 425,460 observations from the six weeks of Phase 3 and six weeks of Phase 3.1 data, respectively, totaling 884,695 observations. Table 1 provides a week-by-week summation of the data utilized for analysis. No differences were found between variables in datasets of the same phase.

Table 1: Dimensionality of Phase 3 and Phase 3.1 Household Pulse Survey Datasets

Phase	Data File	Date Range	# Variables	# Observations
3	Week 22	1/6/21-1/18/21	204	68,348

3	Week 23	1/20/21-2/1/21	204	80,567
3	Week 24	2/3/21-2/15/21	204	77,122
3	Week 25	2/17/21-3/1/21	204	77,788
3	Week 26	3/3/21-3/15/21	204	78,306
3	Week 27	3/17/21-3/29/21	204	77,104
3.1	Week 28	4/14/21-4/26/21	239	68,913
3.1	Week 29	4/28/21-5/10/21	239	78,467
3.1	Week 30	5/12/21-5/24/21	239	72,897
3.1	Week 31	5/26/21-6/7/21	239	70,854
3.1	Week 32	6/9/21-6/21/21	239	68,067
3.1	Week 33	6/23/21-7/5/21	239	66,262
Total				**884,695**

Data Preparation

The data preparation began with studying the data dictionary of the Household Pulse Survey to identify which variables, representing survey questions, should be merged and which should be dropped. Variables from each phase were compared by variable name, question text, answer choices, and range. This allowed for the identification of instances when the variable names did not match and pointed toward the necessary steps to merge common variables between Phase 3 and 3.1, if possible.

Missing data was found to be significant in this dataset. The Household Pulse Survey data had two types of NA values, those where the questions were seen but an answer was not selected (-99) and those where the response was missing or not reported (-88). (-99) values were changed to zero for “select all that apply” variables lacking a “not applicable” answer choice. In essence, the lack of response in this specific case was treated as a genuine answer to the survey question. For all other cases, (-88) and (-99) values were unified as “NA” values. Quite a few variables contained more than 70% missing values. With many follow-up questions to specific responses of other survey questions, several survey participants had fewer than 50% responses.

In order to achieve the two goals outlined by PwC, the formation of two different datasets was required. A “descriptive” dataset would encompass a broader range of variables while a “predictive” dataset would be more limited in dimensionality due to some variables and observations being irrelevant or biasing the models. Once the arduous process of cleansing and pre-processing the data was completed, the descriptive analysis and predictive model

selection and creation was ready to begin. The differences in the preparation of these two datasets are described in greater detail below.

DESCRIPTIVE DATASET PREPARATION

To begin, an unfiltered larger dataset, which consisted of 136 variables and 884,695 observations, was created to address the goal of describing how United States residents have been impacted by the COVID-19 pandemic. This dataset was utilized to explore preliminary demographics and some initial descriptive findings. Once those datapoints were captured, further dataset preparation was conducted as appropriate for descriptive analysis. As previously mentioned, missing data proved to be problematic in this dataset. Many variables were calculated to have missing data as high as 99%. This can create a scenario where the observations that have an answer may become difficult to interpret in a meaningful way. An additional complication included drawing comparisons and conclusions on participant data that is complete and comparing it to participants that had significant missing data. This scenario proved to be problematic. As a result, a decision was made to eliminate variables that were

70% or higher for missing values, as well as observations that included missing data. This reduced the size of the descriptive dataset down to 46 variables and 601,429 observations.

PREDICTIVE DATASET PREPARATION

The target variable to be predicted was called “GETVACC” in the original Household Pulse Survey dataset and may be referred to as “intent” or the target variable hereafter. The target variable asked participants about their intention to get vaccinated with “definitely get”, “probably get”, “probably not get” and “definitely not get” vaccinated as answer choices. Phase 3.1 differed from Phase 3 in that it had a fifth “unsure” answer choice. Participants from Phase 3.1 who selected “unsure” were removed prior to merging. The intent variable was then reduced from four classes to two by merging “definitely get” with “probably get” and “definitely not get” with “probably not get”. All follow-up questions to the intent variable were removed as to not bias the predictive models. Participants who indicated they already received the vaccine were also removed. Filtering for only the unvaccinated participants made some variables irrelevant and were subsequently removed.

Most conditional variables, or follow-up questions, posed concerns over significant missingness. Moreover, it was desired to create a predictive model from survey responses that every participant could see. Consequently, all conditional variables were removed. After reducing the dimensionality for the predictive dataset, ten variables were discovered to need reworking to effectively merge Phase 3 and 3.1. Five variables simply needed to be renamed

to be identical. The remaining five, which included the target variable, needed values to be adjusted such that their representation would remain the same once merged. When completed, the Phase 3 and 3.1 datasets were merged, and the final dimensionality of the predictive dataset was 74 variables and 161,848 observations.

The next task was to deal with missing values. Rows with greater than 50% missing values or where the value of the target variable was missing were removed. Down sampling was used to correct class imbalance in the target variable since further dimensionality reduction was preferred to lessen the computational time of training the classification models. NA values were then imputed using ten rounds of multivariate imputation by chained equations (MICE) with logistic regression for all two-factor variables and predictive mean matching for all other variables.

Descriptive Data Analysis

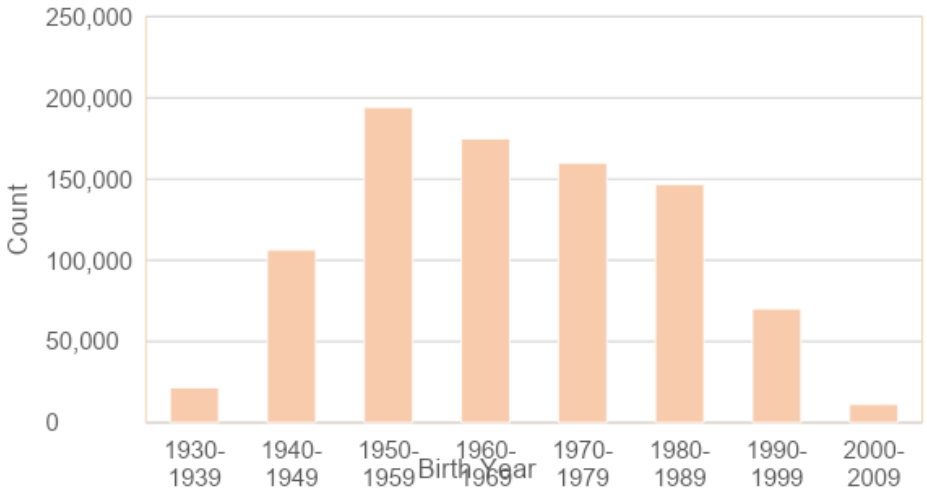
After initial descriptive analysis was conducted on the full dataset of 884,695 responses, the more streamlined dataset of 601,429 observations was created for deeper analysis techniques. Chi-square tests of independence were then used to examine the relationship between the categorical variables and t-tests were used when comparing groups of numeric variables. Due to the large sample size, nearly everything was proven to have a statistically significant relationship. As a result, those relationships that had a more notable effect size became the focus of this analysis. Specifically, for chi-square tests the effect size used was Cramer's V and for t-tests Cohen's d was utilized. It is important to note that interpretation of Cramer's V has been subject to debate in scientific literature, with some researchers indicating anything below a 0.1 is considered non-substantive (Crewson 2015, Doring 2018) while others have different metrics (Akoglu 2018). A threshold of 0.1 was used in this analysis to determine which relationships were meaningful. Cramer's Vs range from 0 to 1, with larger values indicating stronger relationships. In terms of Cohen's d, 0.2 indicates a small effect size, 0.5 indicates a moderate effect size, and 0.8 indicates a large effect size. In addition, all chi-square and t-test results reported in subsequent sections are significant at $p<.001$. Cramer's V and Cohen's d will be reported to indicate the effect size.

INITIAL FINDINGS

First, descriptive analyses were conducted on the original larger descriptive dataset to investigate variables of interest including demographics. Most respondents were female and Caucasian. It is important to note that while a special focus was placed on analyzing participants of a minority race, no distinct differences in trends compared to the overall dataset

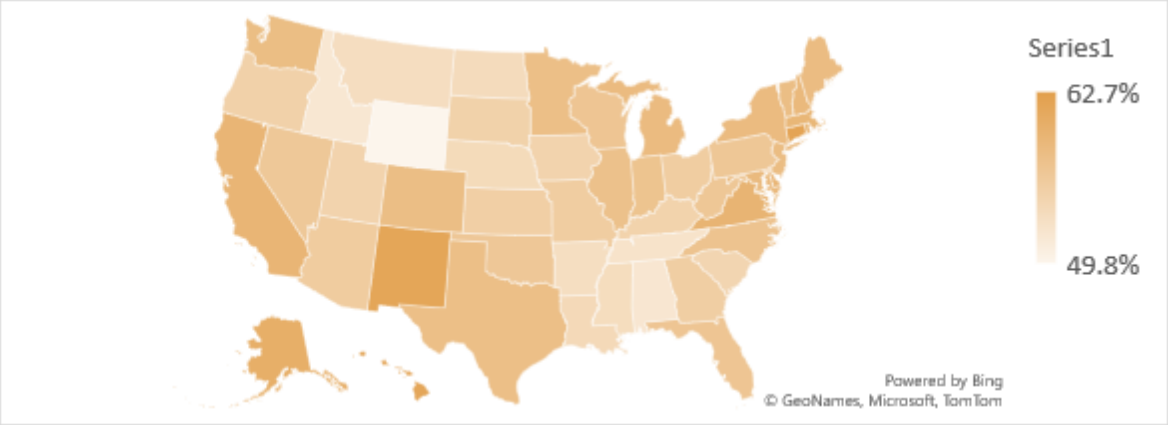
were found. In terms of education,13.6% of respondents had a high school degree or lower, 60.8% had completed at least some college, and 25.6% had graduate degrees. 18.4% were never married, 58.3% married, and 22.5% widowed, separated, or divorced. Households contained an average of 2.72 people. As noted in Figure 1, the frequencies by birth year in a decade range were reported. While the median birth year for this dataset was found to be 1966, the largest group of participants had a birth year between 1950-1959. The smallest age bracket, not surprisingly, was found to be the very youngest, those born between 2000-2009.

Figure 1: Birth Year Frequency – Descriptive Dataset



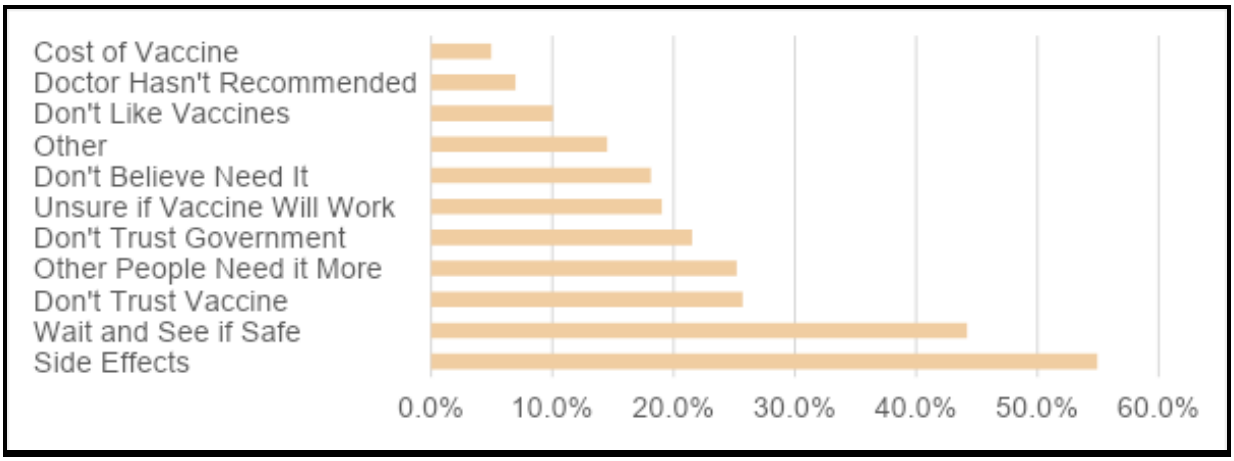
Next, analyses regarding differences between vaccinated and unvaccinated respondents were conducted. An overview of the vaccination rate by state was provided in the Figure 2 heat map. To accomplish this calculation, the number of survey participants by state who indicated they were vaccinated was divided by the number of responses from that state. Calculating the vaccination rate by state this way ensured that there was an accounting for states with different response rates. As shown in Figure 2, all vaccinations rates ranged from 49.8% to 62.7%. Those with the highest vaccination rates were found to be from: District of Columbia (62.7%), Connecticut (62.0%), New Mexico (61.6%), Hawaii (61.6%), and Alaska (60.4%). Conversely, those with the lowest vaccination rates were from: Arkansas (53.3%), Tennessee (52.5%), Alabama (52.0%), Idaho (51.9%), and Wyoming (49.8%).

Figure 2: Vaccination Rates by State



For individuals who had not received the vaccine and indicated that they were “probably going to get”, “probably not going to get”, or “definitely not going to get the vaccine”, they provided information on why they were hesitant. Figure 3 provides an overview of the most common reasons for the participant’s decision. 46.5% cited concern for side effects, 44.2% waiting to see how it works for other people, 25.7% reported not trusting the vaccine, 21.5% did not trust the government, and 25.2% believed others needed it more.

Figure 3: Reasons for Vaccine Hesitancy



These findings align with published research about COVID-19 vaccine hesitancy where it was documented that nearly 40% of adults were somewhat hesitant about getting the COVID-19 vaccine in February 2021 (Ruiz & Bell, 2021). Vaccine rejection rates were relatively fixed at 8.2% from January to March 2021 (Tram et al., 2021). Adults having a previous COVID-19 diagnosis or were unsure if they have had COVID-19 were more hesitant (Nguyen et al., 2021). Young adults were also found to be the age group with the highest COVID-19 infection rate, while also being found to be one of the highest age groups for vaccine hesitancy (Adams et al., 2021; Baack et al., 2021). This research was substantiated in this particular dataset, where 10.5% of survey respondents reported having been diagnosed with COVID-19

by a medical professional. A significant t-test indicated that those who had COVID-19 were younger than those who did not have COVID-19, with a small effect size (*d* = .26). In the independent research, those who expressed vaccine hesitancy gave mostly deliberative reasons, while the main reasons given by those who rejected the vaccine are listed below (Nguyen et al., 2021):

- Distrust of COVID-19 vaccines (47.9%)
- Concern about possible side effects (46.5%)
- Distrust in the government (40.1%)

18.1% of hesitant respondents believed they did not need the vaccine. Table 2 illustrates the rationale behind those decisions, with feeling they were not part of a high-risk group, followed by not believing COVID-19 to be a serious illness, or already having had COVID-19 proving the most common reasons.

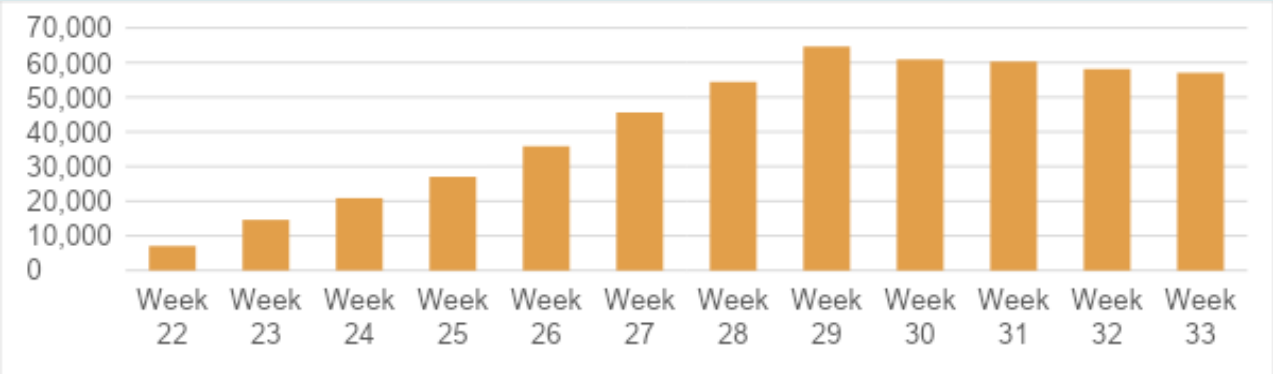
Table 2: Reasons for Not Believing a COVID-19 Vaccine is Needed

Reason for Not Believing	% Selected
Not Member of High-Risk Group	53.0%
Do Not Believe COVID-19 is a Serious Illness	37.9%
Already Had COVID-19	26.8%
Other	21.1%
Plan to Take Other Precautions	18.3%
Do Not Think Vaccines are Beneficial	17.7%

Next, the frequency of vaccination and hesitancy by data collection week was calculated, as shown in Figure 4. Not surprisingly, the vaccination rates were lower in the beginning weeks of the 2021 surveys but continued to increase in frequency as the vaccine became more widely available. There was a steady increase each week from Week 22 to 29

(January through the end of April 2021) and then the frequency of the vaccine plateaus from May to the beginning of July. It is important to remember that vaccines were not readily available at the beginning of this data collection. Beginning in April 2021 (Week 28), the vaccine became more widely available.

Figure 4: Vaccination Counts by Data Collection Week



In Figure 5, the vaccination hesitancy was calculated by the same data collection weeks that were utilized in Figure 4. As shown, the not hesitant group of participants reduces dramatically in April (Week 28), which was when the vaccine became available to all United States residents aged 18 and older. This makes sense since the not hesitant audience would have started to transform into the vaccinated group. Conversely, the hesitant audience continued to grow, and now represents a much higher proportion after the vaccine became readily available. Lastly, Figure 6 showcases vaccination status by employment type. 55.5% of survey participants noted that they were employed in the private business sector, with 60.1% indicating they weren't vaccinated compared to 52% that were vaccinated. 16.9% represented the government sector, with 18.3% of these employees being vaccinated and 15.1% that weren't.

Figure 5: Vaccination Hesitancy Rates by Data Collection Week

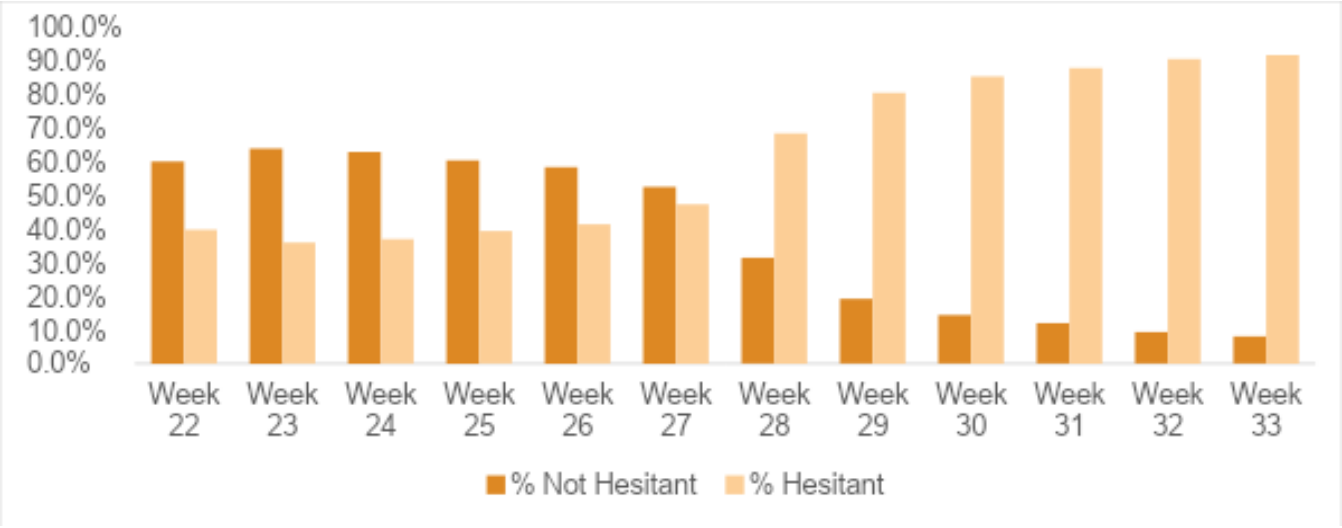
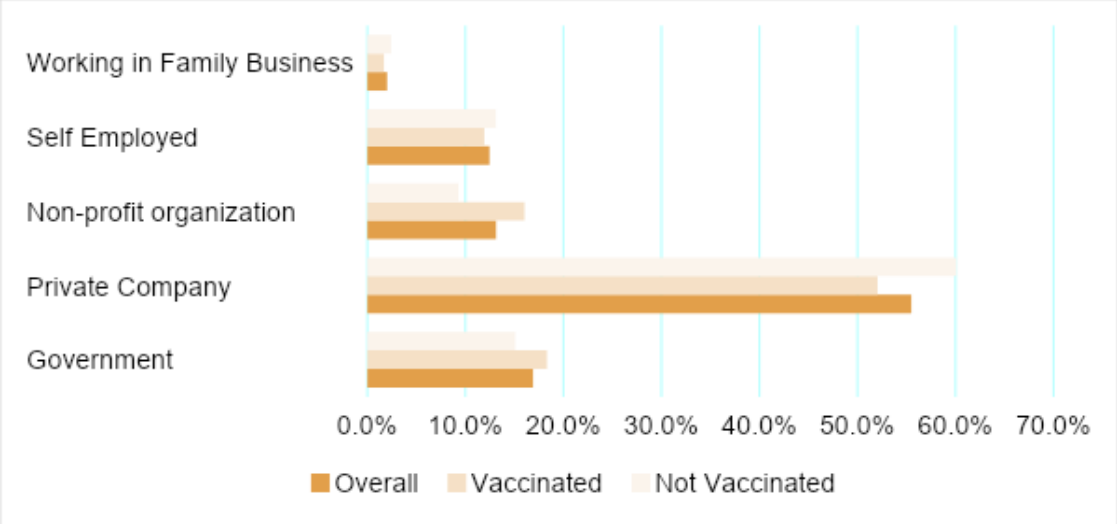
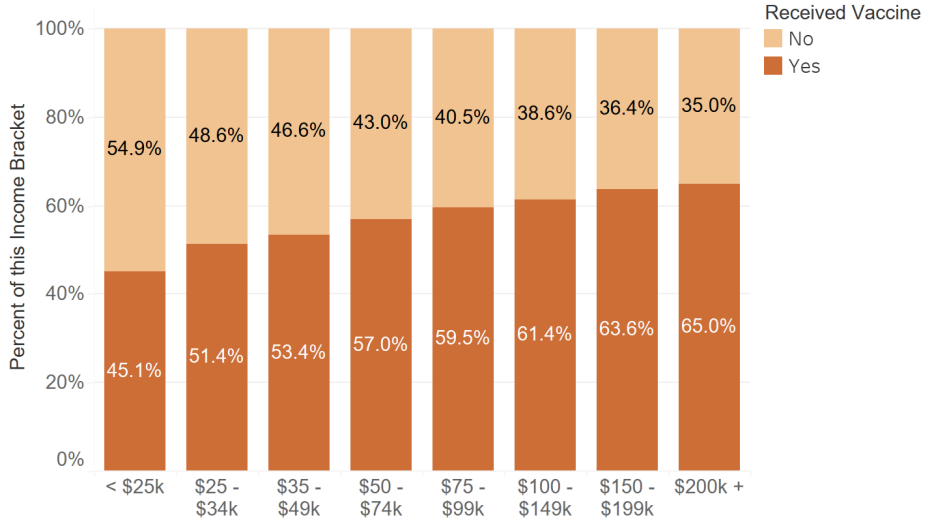


Figure 6: Employment Type by Vaccination Status



Next, the relationship was examined between income and vaccination status. In previous literature, it was noted that those who were not vaccinated were more likely to have lower incomes (Beleche et al., 2021; Hamel et al., 2021; Nguyen et al., 2021; Ruiz & Bell, 2021; Tram et al., 2021). This was substantiated by this dataset, as well. A chi-square test of independence was conducted to examine the potential relationship between vaccination and income. As mentioned previously, all reported chi-square tests of this dataset are significant at the $p<.001$ level. If these were truly independent of each other, it is expected that the ratio of vaccinated to unvaccinated would be the same for each income bracket. As shown in Figure 7, the unvaccinated are overrepresented in the lower income levels. Conversely, the vaccinated are overrepresented in the high-income levels. This supports that there is significant relationship between income and vaccination status. It is important to note, however, that income brackets vary in width, so it's not a strictly linear relationship. This relationship was found to have a small effect size of $V = .12$.

Figure 7: Vaccination Status by Income Level



In addition to the demographical information described above, PwC identified six key social and economic categories to be explored for COVID-19 impact that included childcare, education, employment, energy use, food security, health, housing, and household spending. As part of the descriptive analysis, certain trends have been identified. It is important to note that it was difficult to assess childcare trends with this dataset as 99% of survey respondents opted to not participate in child-related questions. In addition, energy use was not a topic that was included in this dataset outside of internet usage. Therefore, these two socio-economic categories will not be discussed in this report. The subsequent analyses were all conducted on the smaller descriptive dataset with 601,429 respondents and 46 variables.

MENTAL HEALTH

Mental health-related areas were continuously proven to be important during the descriptive analysis. Intuitively, this makes sense as the pandemic transformed every aspect of individuals lives. In fact, 11.1% of respondents received counseling or therapy from a mental health professional such as a psychiatrist, psychologist, psychiatric nurse, or clinical social worker in the four weeks preceding survey participation. Overall, this is a bit lower than trends in the United States prior to the pandemic. According to the CDC, 19.2% of adults received some sort of mental health care in 2019 (Terlizzi & Zablotsky, 2020). This includes 15.8% who were treated with prescription medication for mental health ailments (Terlizzi & Zablotsky, 2020). This discrepancy in trend could potentially be explained by the fact that 10.2% of survey participants identified that counseling or therapy from a mental health professional was needed but did not get it for some reason.

Across all respondents, 46.1% indicated that they had been bothered by having little interest or pleasure in doing things in the past seven days, 46.5% reported not being able to stop or control their worrying, and 55.9% reported feeling nervous, anxious, or on edge. There was a significant relationship between gender and anxiety, with 53.0% of men reporting having no anxiety compared to 38.1% of women. This relationship had a small effect size (V = .15). Not surprisingly, as shown in Table 3, participants who reported more feelings of anxiety, received mental health services at a greater rate (V = .23). A similar relationship was found for those who experienced a lack of interest (V = .19) and worry (V = .21).

Table 3: Mental Health Level

Level	Anxious % Receiving Mental Health Services	Interest % Receiving Mental Health Services	Worry % Receiving Mental Health Services
Not at all	4.0%	6.2%	5.6%
Several days	12.3%	13.8%	13.7%

More than half the days	18.7%	19.2%	20.5%
Nearly every day	24.2%	24.2%	24.9%

There was a moderate relationship identified between age and receiving mental health care services (*d* = .51). The average birth year for someone receiving mental health services was 1973.6 compared to 1965.9 for those who didn't receive mental health services. The same trend was evidenced with those who put off receiving mental health services (*d* = .59), as they also tended to be younger (average birth year 1974.7) than those who didn't put off mental health services (average birth year 1965.8).

23.1% took prescription medication to help with emotions, concentration, behavior, and/or mental health, with women reporting that they were taking prescription medication to help with mental health more often than men (women 28.2% and men 15.7%) at a small effect size (*V* = .15). Those who take prescriptions to help with their mental health were more likely to report having food insecurity/insufficiency, 33.2% compared to 21.2% who don't (*V* = .12).

Income was found to have varying degrees of impact on mental health statuses. Table 4 showcases the percentage of participants with specific mental health related behaviors broken down by the represented income levels. Some interesting findings include participants from lower incomes were significantly more likely to put off receiving mental health services when they needed it compared to those with higher income (*V* = .11). In addition, lower income levels report more use of prescription medication to help with mental health ailments (*V* = .11). Compounding this effect is the reality that medical institutions were overwhelmed with COVID-19 patients and often could not provide the services needed.

Table 4: Mental Health by Income

Income	% Receiving Mental Health Services	% Use of Prescription Medications Related to Mental Health	% Experiencing Anxiety
Less than $25,000	17.5%	33.2%	70.2%
$25,000 - $34,999	13.7%	27.9%	63.7%
$35,000 - $49,999	12.3%	25.8%	60.5%
$50,000 - $74,999	10.7%	24.0%	57.0%
$75,000 - $99,999	9.5%	22.3%	54.0%
$100,000 - $149,999	8.1%	20.4%	51.7%
$150,000 - $199,999	7.0%	18.6%	49.5%
$200,000 and above	5.5%	16.4%	46.2%

During analysis, it was discovered that those that had increased feelings of anxiety were disproportionately reporting the delay of medical care at 26.4%-45.7% depending upon the level of anxiety. This relationship was found to have moderate effect size, indicative of its importance (*V* = .30). Of those with no anxiety reported, only 10.8% reported delaying medical care. This is important as the impacts of COVID-19 are realized. The previous two combined indicate that those who may need medical help (mental health or other medical care) the most are delaying or not getting the help they need at much higher rates than those who are not experiencing mental health concerns.

Not surprisingly, there was a relationship of small to moderate effect size (*V* = .22) between those who expected employment related income loss and higher levels of anxiety (78.9%) compared to those who did not expect income loss (52.5%). People reporting higher levels of difficulty paying for normal expenses also reported increased levels of anxiety at a moderate effect size (*V* = .25). Further investigation went into how individuals were paying for their typical expenses. 59.9% of those who borrowed money from friends or family to meet their spending needs indicated higher levels of anxiety than those who did not borrow money from friends (23.2%). This relationship had a small to moderate effect size (*V* = .23). Additionally, those who needed to use money from savings or by selling assets to meet their spending needs had increased anxiety (30.5%) compared to 21.5% of those that did not use money from savings and assets (*V* = .21).

While education levels did not have a meaningful relationship with anxiety levels (*V* = .05), vaccination status did. In fact, 19.4% of non-vaccinated respondents reported the highest level of anxiety, while only 11.5% of vaccinated respondents reported that same level. Those that have received the COVID-19 vaccine report lower levels of anxiety than those who hadn't received the vaccine with a small effect size (*V* = .15).

General Health

In addition to mental health treatment, the analysis revealed that 23.4% delayed getting medical care because of the pandemic, as reported within 4 weeks of study participation. As shown in Table 5, people with lower incomes were more likely to not seek medical care than those with higher incomes, with an effect size of *V* = .11. This is similar to those with greater food insecurity who noted they did not get needed medical care at a higher rate, with the relationship having a moderate effect size of *V* = .26. Conversely, of those with no food insecurity, 11.9% reported not getting needed medical care.

Table 5: General Health Trends by Income

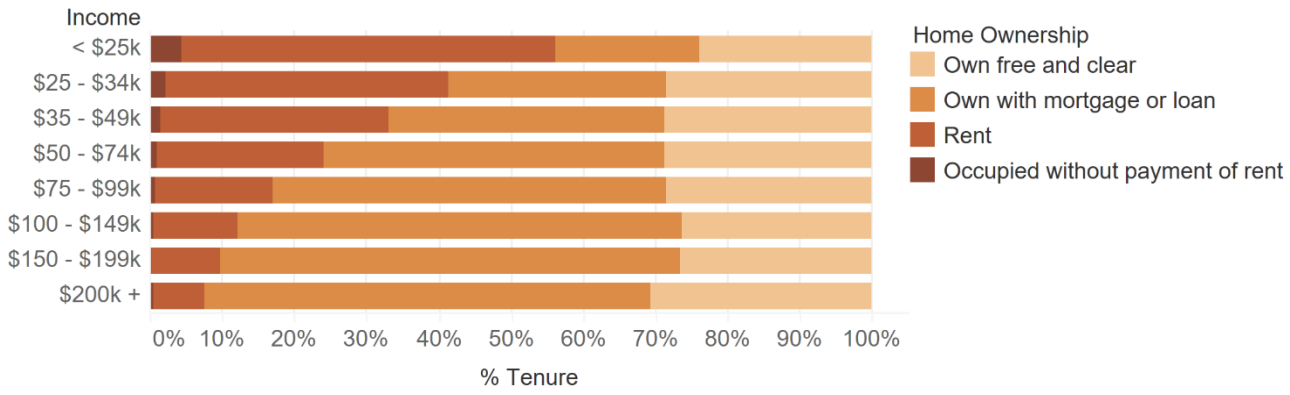
Income	% Report Not Getting Medical Care

Less than $25,000	26.5%
$25,000 - $34,999	21.9%
$35,000 - $49,999	20.3%
$50,000 - $74,999	17.8%
$75,000 - $99,999	16.2%
$100,000 - $149,999	14.4%
$150,000 - $199,999	13.1%
$200,000 and above	11.9%

HOUSING

Trends in housing weren't as revealing as the other socio-economic categories. It is worth noting that 27.4% own their house free and clear, 49.3% own with a mortgage, 22.1% rent, and 1.2% occupy without payment. A significant relationship was found between income and housing type. This had a small to moderate effect size (V = .21). Interestingly, there isn't a huge difference between income levels in those that own their house with no mortgage. The lowest income level owns without a mortgage at 23.6%, while the remaining income brackets own without a mortgage at a rate of 25.9% to 30.2%. A possible explanation is that younger people with high incomes may own their home without a mortgage and older individuals without a large income (retired/social security) own their home with no mortgage because they have since paid it off. Bigger differences appear for those who own a house with a mortgage or rent. For example, the lowest income level owns housing with a mortgage at 20% and the highest income level owns with a mortgage at 62.2%. The inverse is true for renting. 52.0% of the lowest income level rent their housing, while only 7.3% of the highest income level rent.

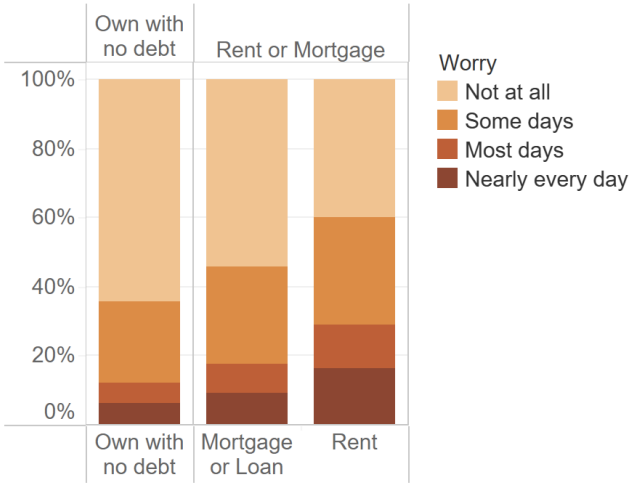
Figure 8: Relationship between Income and Housing



A significant relationship between housing and worry was also found (V = .11). Specifically, those who rent or have a mortgage experience more frequent worry than those who own their home with no mortgage. While those who occupy without payment of rent

experience the most frequent worry at the greatest proportion. Figure 9 illustrates this relationship and suggests that those who have less stable housing experience greater worry.

Figure 9: Relationship Between Housing and Worry



FOOD INSECURITY

Food security was another aspect of the COVID-19 pandemic that is often discussed, as consumers saw empty shelves in food stores, hoarding behavior, and supply chain disruptions affect their access to goods. In related published research, it was documented that unemployment hit an historic high of 14.8% in April 2020 due to the COVID-19 pandemic. As a result, food insecurity increased significantly to 38% in March 2020, as compared to 11% in 2018 (Niles et al., 2021).

In this dataset, 7% reported receiving Supplemental Nutrition Assistance Program (SNAP) or food stamp benefits. Those who receive SNAP benefits report experiencing higher levels of difficulty paying for normal expenses, with a moderate effect size (V = .29). For example, 82.9% of SNAP recipients report difficulty paying for expenses compared to just 38.5% of non-SNAP recipients (V = .29). Those who received SNAP were younger than those who did not receive SNAP, as indicated by the t-test's moderate effect size (d = .30). People

receiving SNAP benefits more frequently reported worry (37.8%) over the last seven days than those that did not receive SNAP (17.4%), with a small effect size (V = .15). Those who rent or occupy without payment of rent (or mortgage or owning) are more likely to be SNAP recipients (V = .21). Specifically, 56.2% of SNAP recipients rent or occupy without rent, while only 20.8% of non-SNAP recipients do.

In Table 6, the differences between SNAP recipients and non-SNAP recipients have been outlined for other food and expense related issues. After chi-square tests were conducted, it was noted that SNAP recipients report greater rates of receiving free food and/or groceries, (V = .22), experiencing food insufficiency, such as not enough food, or not the types of food they want to eat (V = .23), and having difficulty paying for normal expenses, such as rent, food, car payments, medical expenses, etc. (V = .29).

Table 6: SNAP Status Characteristics

SNAP	% Receive Free Groceries or Food	% Experiencing Food Insufficiency/Insecurity	% Having Difficulty Paying for Normal Expenses
SNAP Recipient	24.0%	55.9%	82.9%
Not a SNAP Recipient	4.2%	21.6%	38.5%

Overall, 5.5% of respondents reported receiving free groceries or meals in the last seven days. As outlined in Table 7, there is a relationship between education level and proportion receiving free groceries or food. This was supported through a chi-square test with a small effect size (V = .11).

Table 7: Received Free Food by Education Level

Education Level	% Receiving Free Food
Less than high school	20.8%
Some high school	17.6%
High school graduate or equivalent (i.e., GED)	8.8%
Some college, but degree not received or in progress	7.2%
Associate's degree (for example AA, AS)	6.9%
Bachelor's degree (for example BA, BS, AB)	4.3%
Graduate degree (for example master's, professional, or doctorate	3.3%

As shown in Table 8, there is a relationship between income and food insufficiency/ insecurity (*V* = .22). With those with lower incomes, reporting more frequently that they have enough to eat but not the types they wanted or not having enough to eat. Those who reported receiving free food or groceries were also more likely to use their regular income less (60.9% compared to 82.4%) for normal expenses than those who did not receive free food (*V* = .13). In fact, 21.1% reported borrowing money from friends and family, while only 5.6% of people who did not receive free food reported the same (*V* = .14). In addition, 43.5% of free food recipients used their stimulus payment for regular expenses compared to only 22.2% of those who did not receive free food (*V* = .11). Not surprisingly, there is also a relationship between those who expected to have employment related income loss and the receipt of free groceries or food (*V* = .11).

Sadly, 5.5% reported not having enough food to eat during the seven days preceding the survey, with an additional 18.5% reporting they had enough to eat, but not the types of food they wanted or liked. The USDA reported in 2020, that 3.9% of households had low food security, which was defined as having reduced food intake and disrupted eating patterns (Coleman et al., 2021). This dataset revealed a slightly higher rate for January through July 2021. In addition, the USDA reported that 10.5% of households experienced some form of food insecurity, including those who had to rely on basic foods and lack of variety in their diet (Coleman et al., 2021). This was substantiated in this analysis as there were 18.5% who reported not having the types of food they want to eat. While this is a distinct increase from the USDA's research, it is important to note that the two cannot be compared directly since the survey questions are worded differently.

Table 8: Food Insecurity by Income

Income	% Enough Food/Preferred Kind	% Enough Food/Not Preferred Kind	% Sometimes Not Enough to Eat	% Often Not Enough to Eat
Less than $25,000	45.0%	33.8%	15.8%	5.4%
$25,000 - $34,999	55.7%	31.3%	10.4%	2.5%
$35,000 - $49,999	64.5%	26.9%	7.1%	1.5%
$50,000 - $74,999	74.0%	21.5%	3.8%	0.7%
$75,000 - $99,999	81.1%	16.4%	2.1%	0.4%
$100,000 - $149,999	87.1%	11.7%	1.0%	0.2%

$150,000 - $199,999	91.1%	8.3%	0.5%	0.1%
$200,000 and above	94.5%	5.1%	0.2%	0.2%

According to chi-square tests performed, those who owned their home (with or without a mortgage) are less likely (3.2%) to report food insufficiency than those (13.2%) who rent or occupy without rent (V = .14).

Regarding vaccination status, an interesting but disproportionate relationship was detected (V = .16). At the time of survey, 61.8% of vaccinated participants had enough food compared to 38.2% of those unvaccinated.

Lastly, the relationship between some level of food insecurity and marital status was explored and substantiated (V = .11). Those with the marital status of separated (47.9%) had the highest levels of food insecurity followed by divorced (32.9%), single (31.4%), widowed (26.2%), and married (18.5%).

EMPLOYMENT

Regarding employment, 58.4% indicated they had worked in the last seven days, 13.24% expected to have employment income loss in the upcoming four weeks from the date of the survey, 31.8% received social security benefits, and 43.5% reported difficulty paying for usual household expenses in the last seven days. This aligns with published research that discussed the 26 million Americans that filed for unemployment during the pandemic (one in six US workers, roughly 16.7%) (Lund et al., 2020). An additional 57 million US jobs were vulnerable to reduced hours, pay cuts, unpaid leave, and layoffs (Lund et al., 2020). Food service, leisure, and hospitality jobs were impacted first followed by retail, business services, and manufacturing (Lund et al., 2020).

Similarly, 41.6% of survey respondents indicated experiencing some level of difficulty paying for normal household expenses such as food, rent/mortgage, bills, etc. The United States government, in hopes of assisting citizens during the pandemic economic shutdown, released stimulus payments to a certain percentage of the population. Specifically, 31.2% of respondents reported receiving a COVID-related stimulus payment, 26.1% of them reported spending most of it, 28.5% reported saving it, and 45.4% reported using it to pay off debt. 32% received Social Security benefits, Supplemental Security Income, or Medicare benefits. A chi-square test revealed a slight relationship between those with lower incomes who utilized the stimulus payments to pay off debts at higher rates (V = .14).

12.7% of survey participants believed they would experience employment related income loss over the next 4 weeks. Chi-square testing revealed a relationship between expected employment income loss and respondent reported income with a small effect size (V = .16). It

is suggested that those with high paying jobs may work in fields that were less impacted by the pandemic, while those with low paying jobs may have worked in places that were more pandemic vulnerable, such as food service or retail.

40.5% of respondents reported not working for pay or profit during the last seven days. Chi-square tests revealed a relationship (V = .24), between income and working. Those with lower incomes reported more often not having any work for pay or profit during the last seven days. This is substantiated by published independent research that revealed that 86% of jobs that were pandemic vulnerable paid less than $40k per year (Lund et al., 2020). Not surprisingly, there is also a relationship between working in the past seven days and experiencing difficulty paying expenses (V = .11). Interestingly, the biggest differences were in the extremes. 61.3% of those who had worked in the past seven days reported no difficulty paying for normal expenses and 54.3% of those who have not worked report no difficulty paying for normal expenses. For those that report a little difficulty or that it is somewhat difficult, there was not much difference in proportion of those who worked or didn't work. There were differences again when looking at those who found it very difficult to deal with normal expenses but worked (5.9%) compared to 11.4% of those who did not work. It is important to note that some participants who have not worked in the last seven days were more likely to report receiving Social Security benefits (62.7% of people who have not worked compared to 11.2% of those that worked). This relationship is among the strongest effect sizes reported in this report at V = .54.

Marital status and working in the last seven days were also found to be related (V = .18). For example, those who reported they were widowed worked the least in the last seven days (27.1%). Conversely, those that were single reported working the most (70.0%). Divorced and separated were similar (55.6% and 56.8% respectively) and 60.6% of married people reported working in the past seven days.

Related independent published research mentioned working adults without a bachelor's degree were twice as likely to be in pandemic vulnerable positions, which represents 58% of the US workforce but 82% of pandemic vulnerable jobs (Lund et al., 2020). Table 9 illustrates similar findings in those with higher education reporting they worked more often in the past seven days at an effect size of V = .14.

Additional sources of income for general spending needs were identified, including regular income (81.3%), credit cards and loans (25.9%), money from savings or assets (21.3%), borrowing from friends/family (6.4%), unemployment benefits (6.4%), stimulus (23.4%), savings from deferred/forgiven payments (3.4%), and SNAP benefits (4.4%). An

interesting relationship was determined through chi-square testing that supports men reported making more money than women (*V* = .12).

Finally, with a Chi-square Cramer's V effect size of 0.29, a relationship was revealed between those with private health insurance and those who had worked more often during the past seven days. 66.8% of those with private health insurance worked in the last 7 days, compared to 34.1% of those without private health insurance.

Table 9: Education Level and Percentage of Work

Education Level	% Worked in Last 7 Days
Less than high school	39.3%
Some high school	39.0%
High school graduate or equivalent	47.5%
Some college, but degree not received or is in progress	52.3%
Some college, but degree not received or is in progress	57.1%
Bachelor's degree	64.1%
Graduate degree	66.2%

CLUSTER ANALYSIS

As part of our descriptive analysis and to aid in the identification of key survey participant audience segments, cluster analysis was performed. Before this could be attempted, however, the dataset needed to be filtered again as its dimensionality was too large for processing. The first step taken was to remove dependent variables, which showed an unnatural correlation. In addition, the missing values were evaluated once again. Any rows that contained missing values were removed to ensure that only survey participants that fully filled out the survey were evaluated. In addition, any remaining variables with a high level of missing values were also removed. Lastly, a random sample dataset was created to improve run time efficiency of the clustering model. This left a sample dataset with a dimensionality of 46 variables and 51,121 observations. This sample size allowed multiple clustering models to be run, with K-modes clustering proving most successful with this primarily categorical dataset.

An elbow curve or scree plot graph, as shown in Figure 10, was created to identify that the optimum number of clusters was seven. Table 10 provides the size of each of the seven clusters, which range in size from 3,769 observations to 11,769.

Figure 10: Elbow or Scree Plot for Cluster Analysis

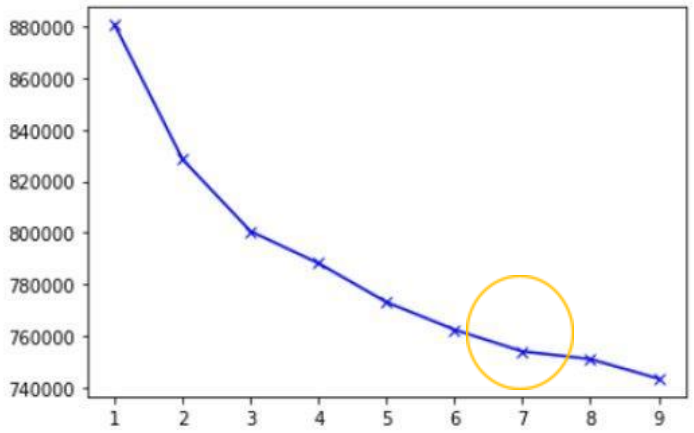


Table 10: Cluster Size

Cluster	Size
1	7,045
2	11,769
3	8,591
4	3,834
5	8,058
6	3,769
7	8,055

While 46 variables were reviewed using k-mode clustering, the most distinguishable differences were identified between the seven clusters using the modes (or mean if the variable was numeric) of each of the variables. As a result of this descriptive analysis, seven personas were created, as shown below, to showcase each cluster in a meaningful way. Personas are easy-to-use tools that can aid a business, such as PwC, in identifying key communication, sales, marketing or relationship strategies based on key behaviors and demographics identified. As shown below, out of the seven clusters only one cluster features males as the predominant gender. Mental health topics and income levels also seem to be particularly important to the cluster development.

CLUSTER ONE

- Female
- Born in the 1940s
- From the West
- Some College Education
- Has Public Health Insurance
- Income: <$75k
- Single Family Home: No Mortgage
- Receives Social Security
- Divorced
- No Paid Work Last 7 Days

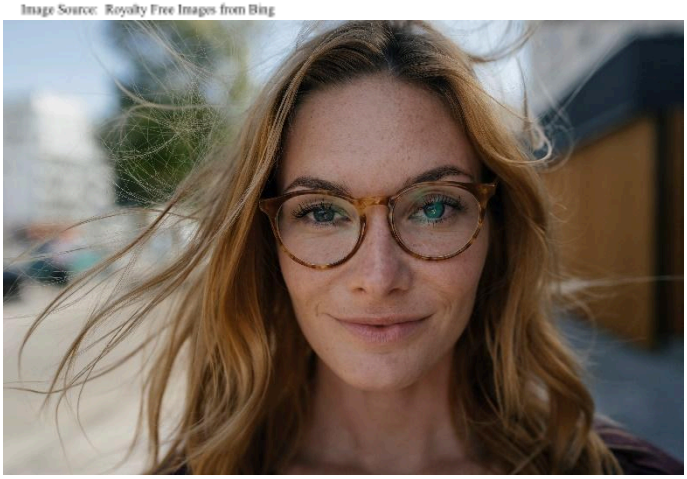
- Vaccinated (COVID)

Image Source: Royalty Free Images from Bing

CLUSTER TWO

- Female
- Born in the 1960s/1970s
- From the Midwest
- Graduate and Bachelor's Degrees
- Income: $100k-$150k
- No Difficulty Paying Expenses
- Has Private Insurance
- Single Family Home: Mortgage
- Married
- Worked for Income in Past 7 Days
- Vaccinated (COVID)

Image Source: Royalty Free Images from Bing



CLUSTER THREE

- Female
- Born in the 1970s
- From the South
- Bachelor's Degree
- Credit Cards/Funds Used in Pandemic
- Income: $50k-$75k
- Single Family Home: Mortgage
- Moderate Worry, No Interest & Anxiety
- Little Difficulty Paying Bills in Pandemic
- Private Insurance
- Married
- Worked in Last 7 days
- Non-Vaccinated (COVID)

CLUSTER FOUR

- Female
- Born in the 1970s
- From the West
- Some College but No Degree
- No Regular Income Like Before Pandemic
- Income: $25k-$35k
- Renting
- Delayed Getting Medical Care Due to COVID
- Often Little Interest in Activities
- Worried and Anxious Nearly Every Day
- Enough Food but Not Always the Kind Liked
- Very Difficult Paying Bills
- Never Married
- No Work in 7 Days and Using Savings for Income or Funds
- Non-Vaccinated (COVID)

Image Source: Royalty Free Images from Bing

CLUSTER FIVE

Image Source: Royalty Free Images from Bing

- Male
- Born in the 1960s
- From the South
- Graduate Degree
- Income: $150K - $200k
- Single Family Home: Mortgage
- Married
- Private Insurance
- Worked in the Past 7 days
- Not Experiencing Worry, Anxiety, or Lack of Interest
- No Difficulty Paying Bills
- No Food Insecurities
- Vaccinated (COVID)

CLUSTER SIX

- Female
- Born in the 1970s
- From the West
- Graduate Degree
- Income: $100k-$150k
- Single Family Home: Mortgage
- Delayed Getting Medical Care Due to Pandemic, Even When Sick
- Several days with Worry, Anxiety, and Little Interest in Activities
- No Food Insecurity
- No Difficulty Paying Expenses
- Married
- Private Insurance
- Worked in the Past 7 Days
- Vaccinated (COVID)

Image Source: Royalty Free Images from Bing

- Hasn't Worked in the Past 7 Days
- Vaccinated (COVID)

CLUSTER SEVEN

- Female
- Born in the 1940s/1950s
- From the South
- Bachelor's Degree
- Using Public Health Insurance
- Income: $75k-$100k
- Single Family Home: No Mortgage
- Receives Social Security
- Married
- No Worry, Anxiety, or Lack of Interest
- Not Experiencing Food Insecurity

Predictive Data Analysis

The Household Pulse Survey data presents many challenges in wrangling and cleaning the data. It's comprised mostly of categorical variables and a handful of numeric variables. The dimensionality is very large in both columns (variables) and rows (observations) making it difficult to assess redundant or irrelevant variables. There is a considerable amount of missing data throughout. Tree-based supervised learning algorithms are among the best equipped to

deal with these obstacles since they are flexible with variable types and robust against redundant or irrelevant variables and outliers. For this reason, decision trees, random forest, and extreme gradient boosting (all tree-based models) were chosen to predict vaccine intent.

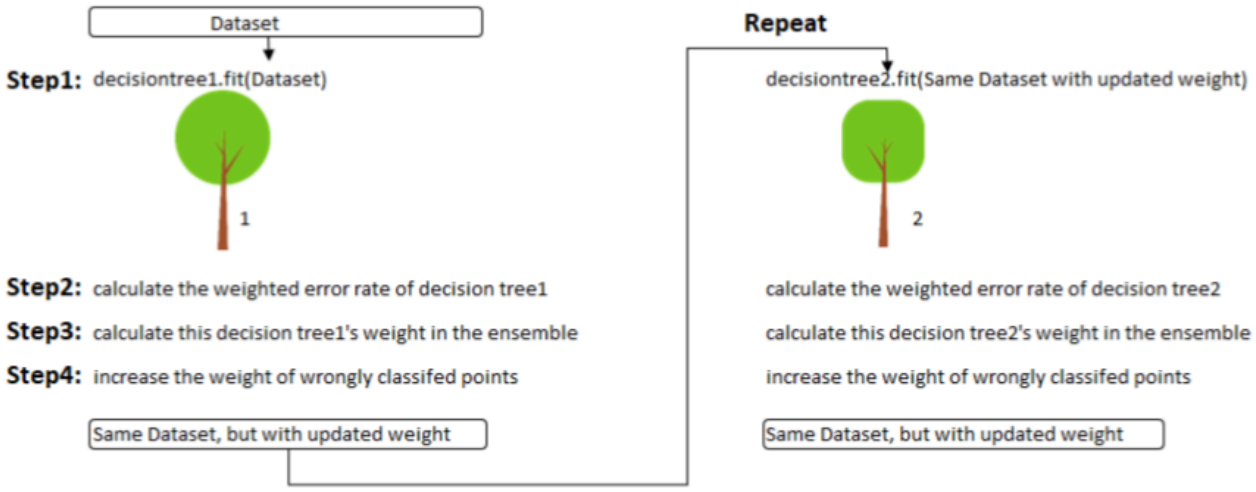
For all models discussed the positive class in the target variable, vaccine intent, are those who leaned toward not getting the vaccine (i.e., those who responded “Definitely not get the vaccine” or “Probably not get the vaccine” in the survey).

The predictive dataset was partitioned such that 80% of the observations would be used to train each of the predictive models and tested on the remaining 20%. The decision tree model used a 10-fold cross-validation with three repeats to minimize potential model bias. Due to the higher computational demands of the random forest and extreme gradient boosting models, the cross validation was reduced to 5-fold with three repeats.

Decision trees are the simplest tree-based supervised learning algorithm. They break down the dataset into smaller subsets by splitting participants at “nodes” by their survey responses. The variable utilized to determine the first split is the one that provides the most homogeneous output. The model continues to further subset the participants many times over in a similar manner making a complex tree-like structure when visualized. Random forests expand on this by modeling several iterations of trees where nodes are split with only a randomized subset of the predictor variables. Since the decision tree and random forest models were not the best for predicting vaccine intent, the results are discussed in the Appendix.

Extreme gradient boosting also expands on decision trees by modeling several iterations of trees in series with prediction errors of the previous decision tree used to train the next tree. Figure 11 provides an intuitive visualization of how boosting works. Extreme gradient boosting narrowly outperformed the other two models and is discussed in greater detail.

Figure 11: Visualization of How Boosting Works



EXTREME GRADIENT BOOSTING ANALYSIS

The best classification model to predict vaccine intent was extreme gradient boosting. To set up for extreme gradient boosting, the categorical variables of the dataset were binarized and the resulting dataset converted to training and testing matrices. There are seven tunable parameters for extreme gradient boosting. A short description of each parameter is discussed in the Appendix.

Given the large number of tunable parameters, a broad grid search across all parameters to find the best tuned model would be computationally inefficient and infeasible. Instead, several iterations of training were conducted using a narrow grid search. The best performing parameters were used to inform the grid search of the next training iteration. In all training iterations gamma was kept at the default value of 0, and subsample was kept at the default value of 1. The other tunable parameter values chosen by the final training iteration were nrounds = 130, max_depth = 8, min_child_weight = 6, eta = 0.09, and colsample_bytree = 0.9.

Table 11 shows the performance metrics of the best tuned boosting model. Like random forest, boosting is also an ensemble method, so only the testing results for the best model should be assessed. No overfitting was observed when comparing the accuracy and kappa between training and testing performance. The extreme gradient boosting model was slightly better at avoiding incorrect predictions in the positive class (those who responded “Definitely not get the vaccine” or “Probably not get the vaccine” in the survey) than the negative class. The performance metrics also suggest the model produces a fairly balanced prediction between the two classes and supports this being a good predictive model.

Table 11. Performance Metrics of the Best Tuned Extreme Gradient Boosting Model

	Training	Testing
Accuracy	0.7158	0.7150
Kappa	0.4315	0.4299
Sensitivity / Recall	-	0.6961
Specificity	-	0.7339
Positive Predictive Value / Precision	-	0.7234
Negative Predictive Value	-	0.7071

F-score	-	0.7095

But what do each of these metrics represent? Accuracy is the ratio of those correctly predicted from all the survey participants. Kappa is not as intuitive, but simply put, it's a metric that expands on percent agreement to account for chance (higher values instill greater modeling confidence). These two metrics are class independent. For this reason, they will not be sensitive to balancing predictions for the positive or negative class.

Sensitivity (recall), specificity, positive predictive value (precision), and negative predictive value are class specific statistical measures. Sensitivity and specificity pertain to the ratio of correct predictions to the *actual* positive and negative class, respectively, while positive predictive value and negative predictive value pertain to the ratio of correct predictions to the *predicted* positive and negative class, respectively. The F-score is a combination of recall and precision meant to identify imbalances in predictive class.

Since class imbalance in the target variable was corrected by undersampling, accuracy and kappa should be sufficient and concise metrics to compare model performance so long as there are no glaring imbalances in the class specific statistical measures. For all three models no significant class prediction imbalances were found. Since the predictive value of one class was not prioritized by PwC at the outset of the project, it can be assumed the predictions in both classes are of equal importance. If that assumption is correct, it is recommended that extreme gradient boosting is utilized for predicting vaccine intent. Both the accuracy and kappa are highest for the extreme gradient boosting model. Thus, it will perform the best overall predictions.

VARIABLE IMPORTANCE

Variable importance was assessed using the decision tree and random forest models as well as the Boruta algorithm which uses a wrapper method based on the random forest algorithm. Importance determined by the decision tree and random forest models are included in the Appendix. Table 12 shows the top 10 variables based on their importance as determined by the Boruta algorithm.

Table 12. Variable Importance Using the Boruta Algorithm (Top 10)

Rank	Variable Name	Variable Description	Importance
1	WEEK	Week of the survey	125.4
2	FEWTRANS	Fewer public transit or rideshare trips due to pandemic	71.4
3	TBIRTH_YEAR	Participant birth year	63.5
4	EEDUC	Educational attainment	61.6

5	INCOME	Household income level	47.6
6	EXPNS_DIF	Difficulty with expenses	46.0
7	ANXIOUS	Frequency of anxiety over the previous 7 days	42.7
8	THHLD_NUMKID	Total number of people under 18-years-old in household	35.0
9	CURFOODSUF	Household food sufficiency for last 7 days	33.5
10	INTEREST	Frequency of having little interest in things over the previous 7 days	32.7

Looking at variable importance can help inform which factors most influence the predictability of the survey participants' intent to vaccinate. Understanding the mechanisms behind the predictive model lead to understanding what drives an individual toward a vaccine-hesitant mindset and will aid PwC in forming actionable insights. The Boruta algorithm found all 73 predictor variables to be important with the most important being the week of the survey, whether participants used less public transit or rideshares due to pandemic concerns, participant birth year and education level, and household income level.

Business Recommendations

Recalling the goals of this data analysis project, PwC has specifically requested that two main challenges be addressed:

- *What are the effects of COVID-19 on the population? Can the data descriptively explain the different effects on areas such as childcare, education, employment, energy use, food security, health, housing, and household spending?*
- *What are the factors that predict vaccine intent? Select and build model(s) that will accurately predict vaccine intent.*

While both challenges have been addressed in the descriptive and predictive analysis sections, it is also suggested that PwC consider using the seven personas that were created using the k-modes clustering to potentially build effective and personalized strategies dependent on the characteristics of each group. For example, a vaccination education project could benefit from personalizing communication to audience segments that align with clusters three and four, as well as those identified in the vaccine intent predictive model. In the descriptive analysis, economic insecurity and challenges were highlighted as significantly affecting many individuals across various demographics. By identifying employees that were most impacted by the pandemic and creating support strategies, such as loan forgiveness, tuition remission, or economic stipends, PwC can create outreach to those segments.

Members of cluster four, for example, could benefit from an effective employee retention and recruitment strategy.

Mental health challenges were revealed to be of particular importance in identifying the effects of the COVID-19 pandemic. PwC could utilize this information for quite a few potential consultant projects, including partnering with key mental health provider networks. Together, public service campaigns could be created to encourage appropriate audience segments, such as clusters three, four, and six, to seek treatment. A similar strategy could be utilized internally to employees most likely to be affected. In addition to mental health, food insecurity was perhaps the most impactful to certain segments of the United States during the pandemic. Utilizing the information provided in this analysis could present an opportunity to PwC to work in a government contractor capacity with local, regional, state, and federal government agencies to assist in public health initiatives such as addressing growing food insecurity challenges.

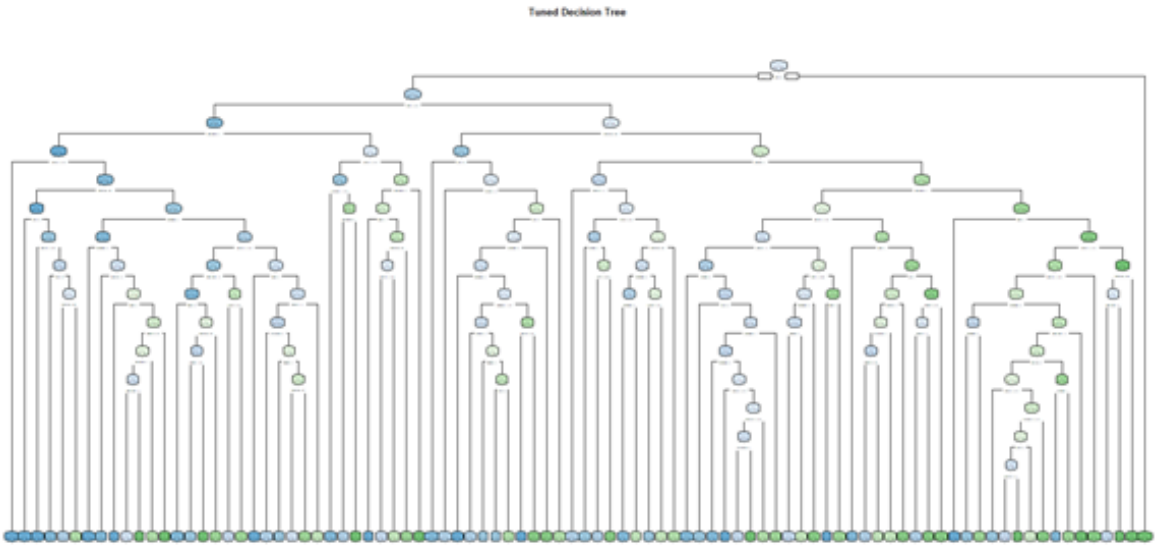
In terms of future analyses with public health data of this dimensionality and replicating findings from this report, PwC may want to consider cloud-based computing options for running subsequent models. In addition, it is recommended that PwC utilize an extreme gradient boosting model for predicting vaccine intent.

Appendix

DECISION TREE ANALYSIS

The first classification model to predict vaccine intent was a decision tree. This model broke down the dataset into smaller subsets at nodes which split participants by a particular survey response using the variable that provides the most homogeneous output. The dataset is allowed to further subset into a complex tree-like form with many branches and nodes until reaching a terminal node called a leaf. A visual representation of the final decision tree model shown in Figure A helps to make more intuitive sense of the model's classification method.

Figure A. Visualization of the Best Tuned Decision Tree Model



The decision tree has only one tunable parameter called a complexity parameter. Decision trees tend to overfit the data when the algorithm is unchecked resulting in poor prediction outcomes, and the complexity parameter reduces this overfitting by forcing the algorithm to make a simpler tree. The model was set to automatically select the optimal complexity parameter based on a fit that results in the highest accuracy. Since no overfitting was observed when comparing the training and testing performances, the final model parameter of 0.0002 was found to be optimal.

Table A shows the performance metrics of the training and testing sets from the best tuned decision tree model. The metrics of each set are comparable, suggesting there is no overfitting and further tuning of the model is unnecessary. The decision tree model was slightly better at avoiding incorrect predictions in the positive class (those who responded "Definitely

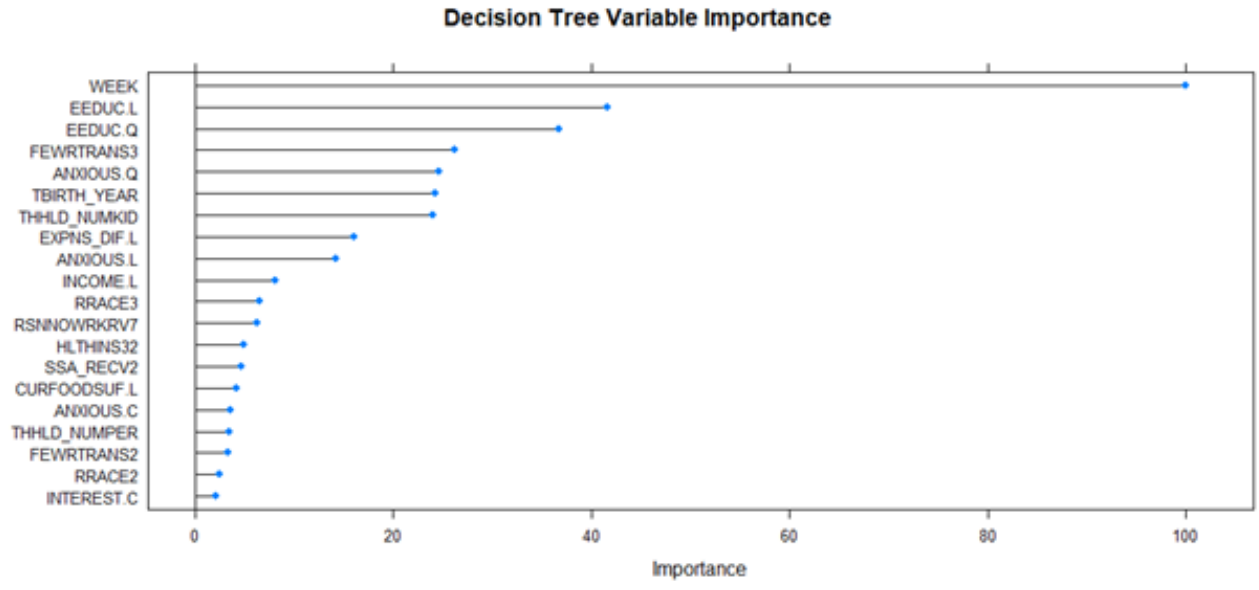
not get" or "Probably not get" in the survey) than the negative class. The performance metrics also suggest the model produces a balanced prediction between the two classes. This result should be expected given the two classes are equal in prevalence and further supports this being an acceptable predictive model.

Table A. Performance Metrics of the Best Tuned Decision Tree Model

	Training	Testing
Accuracy	0.6999	0.6909
Kappa	0.3998	0.3818
Sensitivity / Recall	0.6512	0.6429
Specificity	0.7486	0.7389
Positive Predictive Value / Precision	0.7215	0.7112
Negative Predictive Value	0.6822	0.6742
F-score	0.6845	0.6753

Figure B show the top 20 variables based on their importance as determined by the decision tree model.

Figure B. Decision Tree Variable Importance (Top 20)



RANDOM FOREST ANALYSIS

The next classification model to predict vaccine intent was random forest. This model runs several iterations of a decision tree and selects the best fit tree that produces the highest accuracy. Each decision tree is different in that at each node a randomized subset of all possible variables are chosen to split participants into smaller subsets. Like in decision tree classification, only the variable from the randomized subset that results in the most homogenous output is chosen. An advantage of random forests is that they tend to be fairly robust against overfitting.

Random forests have two tunable parameters. The first is the number of iterations (or trees) to make before stopping and selecting the best performing one. As the number of decision tree iterations increases the best fit model's accuracy also tends to increase. The computational demands of more iterations beyond the default of 500 trees tends to not be worth the very small increases in model performance. For this reason, the number of trees parameter was kept at the default setting of 500.

The second tunable parameter is the sample size of random variables from which the model can pick. The optimal value for this parameter typically lies somewhere close to the square root of the number of predictor variables in the dataset (in this case $\sqrt{73} = 8.5$). A grid search was used to make several models varying the parameter from eight to fourteen by even numbers. The best resulting model found the optimal value for the random variable sample size was twelve.

Table B shows the performance metrics of the best tuned random forest model. For ensemble methods only the training results for the best model should be assessed. Like in the decision tree model there appears to be no overfitting when comparing the accuracy and kappa between training and testing performance. However, unlike the decision tree model, the random forest model was slightly better at avoiding incorrect predictions in the negative class (those who responded "Definitely get" or "Probably get" in the survey) rather than the positive class. The performance metrics also suggest the model produces a fairly balanced prediction between the two classes and supports this being a good predictive model.

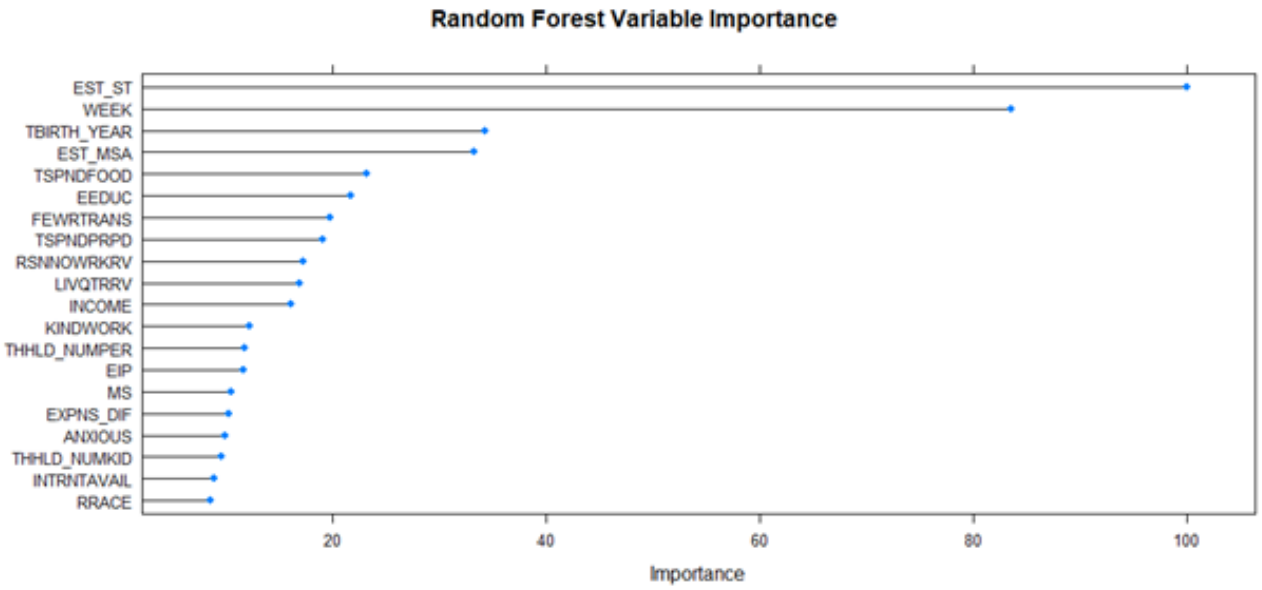
Table B. Performance Metrics of the Best Tuned Random Forest Model

	Training	Testing
Accuracy	0.7053	0.7043
Kappa	0.4105	0.4087

Sensitivity / Recall	-	0.7537
Specificity	-	0.6550
Positive Predictive Value / Precision	-	0.6860
Negative Predictive Value	-	0.7267
F-score	-	0.7182

Figure C show the top 20 variables based on their importance as determined by the random forest model.

Figure C. Random Forest Variable Importance (Top 20)



EXTREME GRADIENT BOOSTING TUNING PARAMETERS

1) *nrounds* – This is the number of boosting rounds, or decision trees, to make before stopping. More rounds of boosting will usually increase model performance up to a point where the benefit eventually levels off. The default value is 100.
2) *max_depth* – This is the maximum number of nodes allowed from root to the farthest leaf. More depth allows for a more complicated tree. This may increase performance, but too much complexity results in overfitting. The default value is six.

3) *min_child_weight* – This is the minimum weight required to create a new node (or split). A smaller value will allow for a small subset of samples in child nodes. Like max_depth, this allows for more complexity, but at the risk of overfitting.
4) *gamma* – This sets a minimum amount of error reduction required for a node to split if not already limited by the max_depth or min_child_weight parameters. Higher values will reduce tree complexity to prevent overfitting. The default value is zero.
5) *eta* – This is the learning rate for the boosting rounds to achieve optimal prediction error reduction. In other words, it determines the amount of correction applied between boosting rounds. A lower learning rate is computationally slower and needs more boosting rounds to reach the optimum model. However, the benefit of a slower learning rate is that it can be robust to overfitting and improve model performance. The default value is 0.3.
6) *colsample_bytree* – This sets the ratio of variables sampled for each tree. Like with random forest, random sampling a smaller subset of variables for each tree can help to avoid overfitting. The default value is one which uses all variables.
7) *subsample* – This sets the ratio of observations sampled for each tree. Lower values can decrease computational demand and prevent overfitting. However, a very low ratio will result in underfitting. The default value is one which uses all observations.

COMPARING MODEL PERFORMANCE

Table C shows a comparison of the testing performance from each of the three classification models used on the Household Pulse Survey data to predict vaccine intent. The best value for each performance metric is highlighted. For all models the positive class in the target variable are those who leaned toward not getting the vaccine (i.e., those who responded "Definitely not get the vaccine" or "Probably not get the vaccine" in the survey).

Table C. Comparing Decision Tree, Random Forest, and Extreme Gradient Boosting

	Decision Tree	Random Forest	Extreme Gradient Boosting
Accuracy	0.6909	0.7043	0.7150
Kappa	0.3818	0.4087	0.4299
Sensitivity / Recall	0.6429	0.7537	0.6961
Specificity	0.7389	0.6550	0.7339
Positive Predictive Value / Precision	0.7112	0.6860	0.7234

Negative Predictive Value	0.6742	0.7267	0.7071
F-score	0.6753	0.7182	0.7095

References

Adams, S.H, Schaub, J.P., Nagata, J.M., Park, M.J., Brindis, C.D., & Irwin Jr., C.E. (2021). Young adult perspectives on COVID-19 vaccinations. *Journal of Adolescent Health, 69*(3), 511-514. https://doi.org/10.1016/j.jadohealth.2021.06.003

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91-93. https://doi.org/10.1016/j.tjem.2018.08.001

Baack, B.N., Abad, N., Yankey, D., Kahn, K.E., Razzaghi, H., Brookmeyer, K., Kolis, J., Wilhelm, E., Nguyen, K.H., & Singleton, J.A. (2021). *COVID-19 vaccination coverage and intent among adults aged 18-39 years - United States, March-May 2021*. Morbidity and Mortality Weekly Report, Centers for Disease Control and Prevention, *70*(25), 928-933. http://dx.doi.org/10.15585/mmwr.mm7025e2

Beleche, T., Ruhter, J., Kolbe, A., Marus, J., Bush, L., & Sommers, B. (2021). *COVID-19 vaccine hesitancy: Demographic factors, geographic patterns, and changes over time* (Issue Brief). Assistant Secretary for Planning and Evaluation. https://aspe.hhs.gov/sites/default/files/private/pdf/265341/aspe-ib-vaccine-hesitancy.pdf

Coleman-Jensen, A., Rabbitt, M.P., Gregory, C.A., & Singh, A. (2021). *Household food security in the United States in 2020 (ERR-298).* U.S. Department of Agriculture, Economic Research Service. https://www.ers.usda.gov/webdocs/publications/102076/err-298.pdf?v=8240.1

Crewson, P. (2015). *Applied statistics*. Acastat Software. https://www.acastat.com/statbook/chisqassoc.htm

Doring, M. (2018). Effect sizes: Why significance alone is not enough. *Data Science Blog.* https://www.datascienceblog.net/post/statistical_test/effect_size/

Dror, A.A., Eisenbach, N., Taiber, S., Morozov, N.G., Mizrachi, M., Zigron, A., Srouji, S., & Sela, E. (2020). Vaccine hesitancy: The next challenge in the fight against COVID-19. *European Journal of Epidemiology, 35*, 775-779. https://doi.org/10.1007/s10654-020-00671-y

Fields, J.F., Hunter-Childs, J., Tersine, A. Sisson, J., Parker, E., Velkoff, V., Logan, C., & Shin, H. (2020). Design and operation of the 2020 Household Pulse Survey, *U.S. Census Bureau.* Forthcoming. Retrieved from https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/2020_HPS_Background.pdf

Internet Consultants LLC. (n.d.). PWC clients (PWC 2021 client list). The Big 4 Accounting Firms. Retrieved October 2, 2021, from https://big4accountingfirms.com/pwc-audit-clients-list/.

Lund, S., Ellingrud K., Hancock, B. & Manyika, J. (2020). COVID-19 and Jobs: Monitoring the US impact on people and places. *McKinsey Global Institute*. https://www.mckinsey.com/~/media/McKinsey/Industries/Public%20and%20Social%20Sector/Our%20Insights/COVID%2019%20and%20jobs%20Monitoring%20the%20US%20impact%20on%20people%20and%20places/COVID-19-and-jobs-Monitoring-the-US-impact-on-people-and-places.pdf

Nguyen, K.H., Nguyen, K., Corlin, L., Allen, J.D., & Chung, M. (2021). Changes in COVID-19 vaccination receipt and intention to vaccinate by socioeconomic characteristics and geographic area, United States, January 6 – March 29, 2021. *Annals of Medicine, 53*(1), 1419-1428. https://doi.org/10.1080/07853890.2021.1957998

Nguyen, K.H., Srivastav, A., Razzaghi, H., Williams, W., Lindley, M.C., Jorgensen, C., Abad, N., & Singleton, J.A. (2021). COVID-19 vaccination intent, perceptions, and reasons for not vaccinating among groups prioritized for early vaccination – United States, September and December 2020. *American Journal of Transplantation, 21*(4), 1650-1656. https://doi.org/10.1111/ajt.16560

Niles, M.T., Beavers, A.W., Clay, L.A., Dougan, M.M., Pignotti, G.A., Rogus, S., Savoie-Roskos, M.R., Schattman, R.E., Zack, R.M., Acciai, F., Allegro, D., Belarmino, E.H., Bertmann, F., Biehl, E., Birk, N., Bishop-Royse, J., Bozlak, C., Bradley, B., Brenton, B.P., ...Yerxa, K. (2021). *A multi-site analysis of the prevalence of food security in the United States, before and during the COVID-19 pandemic.* MedRxiv. https://www.medrxiv.org/content/10.1101/2021.07.23.21260280v1

Ruiz, J.B. & Bell, R.A. (2021). Predictors of intention to vaccinate against COVID-19: Results from a nationwide survey. *Vaccine, 39*(7), 1080-1086. https://doi.org/10.1016/j.vaccine.2021.01.010

Terlizzi, E.P. & Zablotsky, B. (2020, September). *Mental health treatment among adults: United States, 2019.* (NCHS Data Brief, No. 380). National Center for Health Statistics. https://www.cdc.gov/nchs/data/databriefs/db380-H.pdf

Tram, K. H., Saeed, S., Bradley, C., Fox, B., Eshun-Wilson, I., Mody, A., & Geng, E. (2021). Deliberation, dissent, and distrust: Understanding distinct drivers of COVID-19 vaccine hesitancy in the United States. *Clinical Infectious Diseases*. Advance online publication. https://doi.org/10.1093/cid/ciab633