Analysis of IPUMS NHIS Survey Data

Jay Bendre, John Dinh, Grant Gambetta and Ignat Kulinka

November 2021

1 Introduction

In the United States, there have been over 47 million cases of COVID since the pandemic started in early March of 2020. Based on socioeconomic factors, certain groups of people have been affected by the pandemic more than others [1]. In order to take a closer look at the disparities between sub-populations and the impact COVID has had on different groups of people, it is important to analyze socioeconomic factors including, but not limited to, household income, race, education, and access to healthcare and how these constituents relate to COVID infection and testing rates. Another point of analysis will be determining what are the driving factors behind high rates of COVID infections. Uncovering similarities and differences between high risk groups will reveal which demographic needs more government aid during future pandemics. In addition, the spread of an infectious disease like COVID-19 is limited through keeping community exposure as low as possible. This means that people need to be frequently tested for COVID and those who test positive need to properly quarantine in order to limit their exposure to other people. Therefore, determining which factors influence people's ability to get tested for COVID can help ensure that everyone gets tested for COVID when necessary and ultimately keep the spread of COVID at a minimum.

2 Dataset Overview and Questions of Interest

The data set to be analyzed in this project is a subset of variables from the National Health Interview Series (NHIS) for 2020. The NHIS is a individual level interview survey from households that were randomly sampled based on their locations. As a whole, NHIS is a part of the Integrated Public Use Microdata Series (IPUMS) [2], which is a database for individual level census data and health surveys. For the first quarter of 2020, the survey data was collected via in-person interviews, and for subsequent quarters, was collected via phone call. The subset of variables we selected for this study were primarily socioeconomic related and were used to find relationships with COVID related variables. The response variables that we modeled using these socioeconomic predictors were COVID test results, diagnoses, and the ability to get tested for COVID. The goal from using these variables was to determine if there are statistically significant relationships between sub-populations and its proneness to COVID. With this in mind, the questions of interest for our project are:

1. How does socioeconomic status affect COVID-19 infection status?
2. What are the driving factors behind infections?
3. What factors impact someone's ability to get tested for COVID?

3 Exploratory Data Analysis

A crucial preliminary step in any statistical analysis is performing a thorough examination of the data set prior to application of any techniques or methods. While potentially time consuming, exploratory data analysis can help spot any potential issues, as well as expose various trends and relationships that are not readily apparent. As such, we began by performing several sanity and routine diagnostic checks of the data. Namely, the initial data set was noted to include 40 variables and 37,358 rows. Next, the data type

(e.g. numeric or categorical) for each variable was recorded and adjusted for analysis in R. In addition, distributions of each variable were visually inspected. Figure 1 demonstrates summary distributions for two of the COVID-related response variables relied on in this project. At this stage of the analysis, miscellaneous columns such as Serial Number and NHIS identification were dropped, leaving us with 29 data-rich columns for subsequent examination.

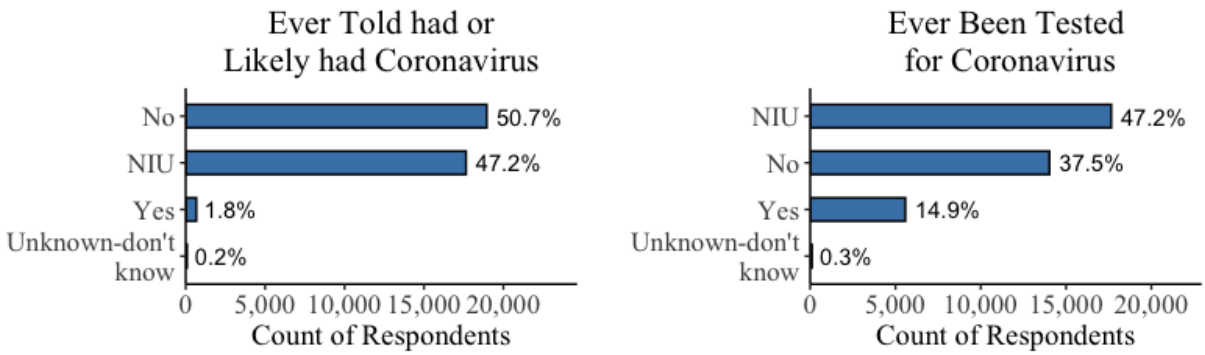

Figure 1: Responses for CVDDIAG and CVDTEST, respectively.

Next, the full data set was reviewed for any missing or incomplete rows. While there were no missing values recorded in the data set, due to the nature of the data source, most of the columns contained entries coded as Not In Universe ("NIU"). These entries indicated that a respondent did not answer a particular survey question and as a result, NIU values were treated as missing values and were subsequently dropped from the data. Note that, in addition to a separate documentation, each variable was *haven labelled* (i.e. provided with meta data or a description of the variable and it's levels). Table 1 below shows a sample of a few variables that we analyzed in this project. Table 7 provides a more complete summary of all the variables in the dataset, including variable name, description, type, and the number of NIU responses grouped by general variable theme.

Table 1: Brief description of some of the variables used in this study.

Variable Name	Description
AGE	Age of the respondent
CVDTEST	Ever been tested for coronavirus
EMPSTAT	Employment status in past 1 to 2 weeks
FAMTOTINC	Total family income for the year 2020
HINOTCOVE	Respondent Health insurance coverage status

3.1 Feature Engineering

Numerous variables in the dataset had multiple levels that conveyed the same type of information for the questions at interest. For example, the demographic variable SEX in the original data set had 5 factors. Three of these 5 factors were a variation of "unknown", and were therefore combined into a single factor. As a result, the new variable SEX had 3 levels: male, female, and unknown. In the demographic variables (Table 8), this procedure was done for sexual orientation (SEXORIEN), race (RACEA), legal marriage status (MARSTAT), current marital status (MARST), veteran status (ARMFEV), and family size (FAMSIZE).

In the education variables (Table 8), this procedure was done to educational attainment (EDUC), spouse education level (SPOUSEDUC), employment status of cohabiting partner (PARTNEREMP), school attending status (SCHOOLNOW), and employment status within the last 2 weeks (EMPSTAT). In the hours

worked variable (HOURSWRK), all values of unknown were replaced with 0.

In the healthcare section (Table 8), the variable that indicated whether or not individuals had a usual place for medical care, USUALPL, was mutated in such a way that respondents that had at least 1 usual place were merged and respondents that answered with a variation of unknown were also merged. Health insurance coverage status (HINOTCOVE) and employer health care offering (EMPHI) were also altered such that all unknowns were combined into one factor.

Finally, the COVID variables were adjusted in a similar manner. The unknown values in test result (CVDTESTRSLT), test opportunity (CVDTEST), and diagnosis (CVDDIAG) were all merged into one factor in their respective variables.

3.2 Data Visualization

Prior to fitting any model, further exploratory analysis of the relationship between the response and predictor variables was conducted. One of the first comparisons conducted was Figure 2, which juxtaposed COVID status of respondents according to a test (CVDTESTRSLT) to whether or not they had health insurance coverage (HINOTCOVE). The percentages in the histogram show break out of respondent's COVID infection status by health insurance coverage. At first glance, we can see that the majority of the respondents have tested negative and have health insurance coverage. In addition to the above observation, close examination of respondents within health insurance groups produced a stark discrepancy between those who tested positive and negative. That is, out of the respondents who had health coverage, only 8.6% (440 out of 5099) tested positive while for those without health insurance, 18.6% tested positive (59 out of 318). [INSERT TEST MAYBE]. The discrepancy in this relationship between infection status and health insurance coverage led to continued examination of both variables in the modeling stage of the project. Overall, this analysis indicates that those lacking viable health insurance were at a greater risk for COVID infection. Additional research into this trend indicates that this finding is corroborated and generalized to many countries world-wide such as China, Israel, and Belgium [3].

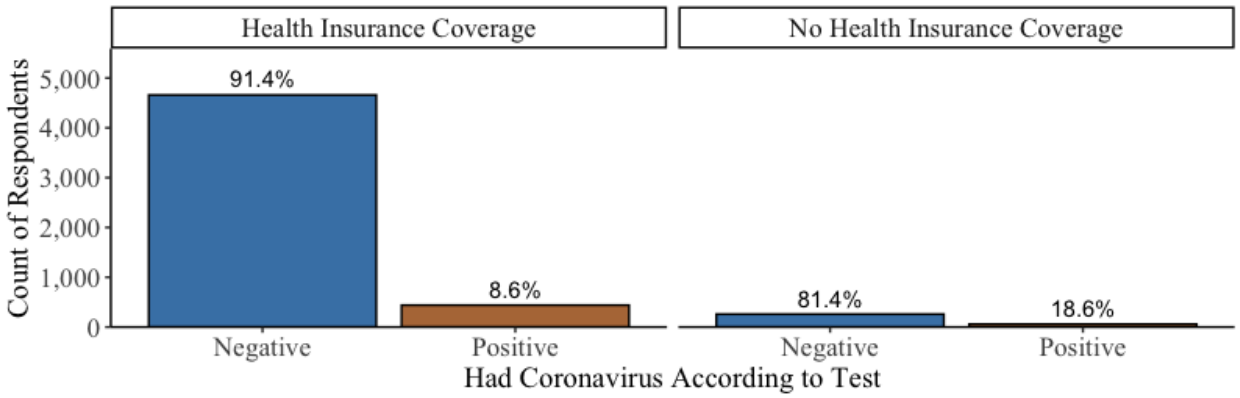

Figure 2: Relationship COVID test status and health insurance coverage.

In order to better understand potential trends in the data, further analysis of the CVDTESTRSLT variable was performed. As such, Figure 3 shows another set of histograms created to examine the relationship between availability of paid sick leave at a respondent's job and infection status according to a COVID test. Note that the percentages above the bars of the histogram indicate the proportion of respondents within the class based on the availability of paid sick leave. The data indicates that while there were fewer respondents who did not have employer sponsored sick leave, a larger portion of them tested positive. While this difference only amounts to approximately 2.4%, it is indicative of a potential at-risk group. In other words, the data indicates that providing sick leave to employees can lead to a positive impact on driving COVID infection rates down.

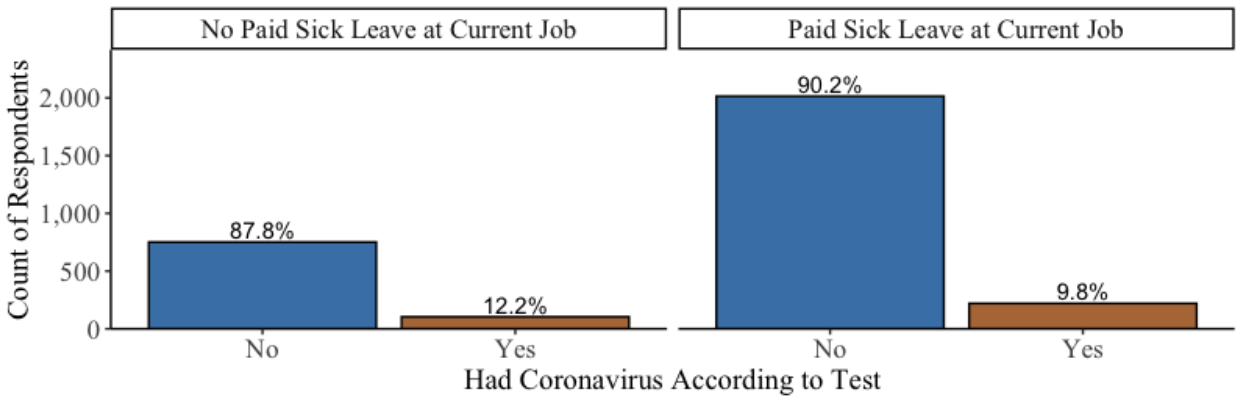

Figure 3: Relationship between COVID testing and employer insurance offering.

Another relationship we analyzed was the impact of whether or not the respondent's employer offered health insurance (EMPHI) versus the respondent getting tested for COVID (Figure 4). Interestingly enough, regardless of the employer health insurance offering, the majority of the respondents did not get tested for COVID. In fact, in both groups, roughly twice as many individuals have not gotten tested compared to those that have gotten tested. A closer comparison between the percentages of respondents who got tested for COVID shows that respondents whose employers offered healthcare insurance were more likely like to get a COVID test.

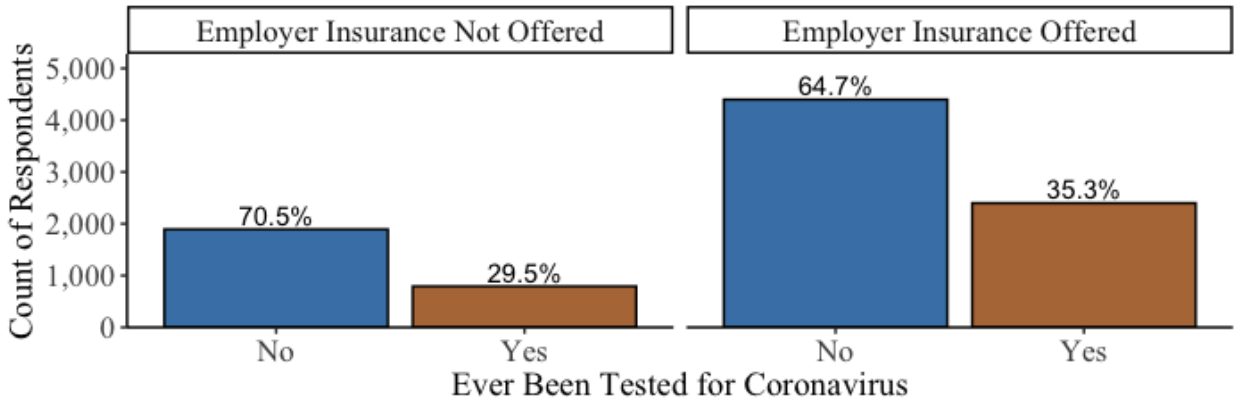

Figure 4: Relationship between COVID testing and employer insurance offering.

4 Model Fitting

4.1 Question 1

To address question one, "How does socioeconomic status affect COVID-19 infection status?" we fit a logistic regression to model the response variable CVDDIAG. Recall that CVDDIAG represented whether or not an individual was told they had or likely had COVID-19. As mentioned in the feature engineering section, the unknown values in this variable were merged into one factor. After re-leveling the variable, the response of interest is 0, which represents an individual never being told they likely had COVID-19, and 1, which represents an individual being told they likely had COVID-19. Prior to fitting a model on the data, we

had to address the class imbalance that was present in CVVDIAG. Roughly 85% of the labels in CVVDIAG were 0 (did not have COVID) and a class imbalance this large could have caused the model to be biased. Therefore, we used stratified sampling (an oversampling technique) to even out the class distribution in CVVDIAG. After oversampling the data, we had a more balanced distribution in CVDDIAG that consisted of roughly 75% zeros and 25% ones (Figure 5).

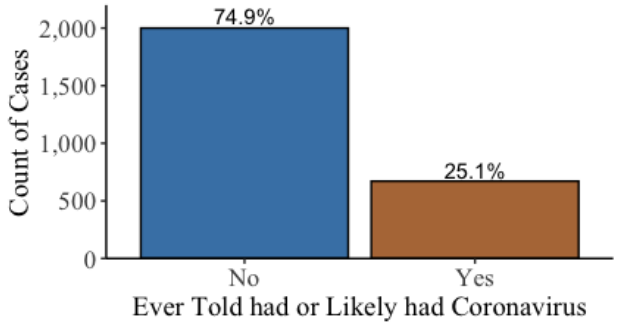

Figure 5: CVVDIAG Distribution after Stratified Sampling

After oversampling the data, a 70/30 split was performed to partition the data into a training and test set and a logistic regression model was fit using age (AGE), sex (SEX), race (RACEA), marriage status (MARST), sexual orientation (SEXORIEN), education level (EDUC), sample adult flag (ASTATFLG), sample child flag (CSTATFLG), and region of residence (REGION) as predictor variables.

After fitting the full model, we ran a forward selection procedure to find the optimal model. The forward selection procedure was performed using the Akaike information criterion (AIC) with the model mentioned before as the full model. After the forward stepwise procedure was performed, the optimal model based on AIC was as follows:

```
CVDDIAG ~ AGE + SEX + RACEA + ASTATFLG
```

The reduced model had an accuracy of 73% on the test data and the model summary can be viewed in Table 2. Refer to Table 4 to view the confusion matrix for this model.

4.2 Question 2

The focal point of question 2 was to determine the driving factors behind COVID infection rates. Our response variable of interest is the same variable of interest in question 1, CVDDIAG. A random forest model was fit on the same train/test split mentioned in Section 4.1 and the model consisted of 500 decision trees because we had a smaller size of available predictors. The size of the subset sample split at each step is 5. By design, a random forest measures the randomness of the variables and assigns importance based on the gini index of the variable (Figure 6). A few of these important predictors, outside of COVID test result (CVDTESTRSLT) and COVID testing ability (CVDTEST), were age (AGE), family income (FAMOTINC), and education level (EDUC). The random forest model had an accuracy of 91.5% on the test data.

4.3 Question 3

Lastly, to address question 3, we decided to fit a logistic regression model to analyze the relationship that several variables had on someone's ability to get tested for COVID-19. As a reminder, question 3 was "What types of working conditions impact someone's ability to get tested for COVID?" The response variable of interest for this question was *CVDTEST* where 0 represented someone who never got tested for COVID and 1 represented someone who did get tested for COVID.

We began by performing the stratified sampling procedure to oversample the data with the goal of evening out the class distribution in the CVDTEST variable. Using a 70/30 split, we fitted the model on 70% of the entire dataset and held out the remaining 30% for testing. The full model consisted of the following predictor variables: sex (SEX), age (AGE), family size (FAMSIZE), employment status (EMPSTAT), number of hours worked per week (HOURSWRK), paid sick leave at current job (PAIDSICK), workplace offered health insurance (EMPHI), usually work 35 or more hours per week (EMPFT), has usual place for medical care (USUALPL), health insurance coverage status (HINOTCOVE), total combined family income since 2007 (INCFAM07ON), and family total income (FAMTOTINC).

The full model had an accuracy of 58% on the test dataset. We decided to run another forward selection procedure using the AIC criterion and the optimal model was as follows:

```
CVDTEST ~ EMPSTAT + HINOTCOVE + SEX + EMPHI + AGE
```

The reduced model had an accuracy of 61% on the test data, which was 3% higher than the full model. Refer to Table 6 to view the confusion matrix for the reduced model and Table 3 for the model summary.

5 Results and Discussion

5.1 How does socioeconomic status affect COVID-19 infection status?

In order to determine which socioeconomic factors actually affect the diagnosis of COVID, we decided to fit a logistic regression model that would be able to predict the diagnosis of the patient based on factors like their race, age, gender etc. In order to determine which variables were statistically significant, we had to employ a forward selection method, wherein we started with the null model and tried to reach the full model.

An important caveat to note is the variable ASTATFLG, which is just an indicator of having an adult in the household to take the survey, does not have much meaning in this analysis, although it was marked as statistically significant. The response variable of interest we are looking at is CVDDIAG, which measured if an individual had been told they had or likely had COVID. Table 2 gives us a summary of the statistically significant predictors after the forward selection procedure. The estimated regression coefficient for age is -0.017, and it is statistically significant at the 0.001 level. This implies that for every additional year in age the patient is, the odds decrease by 1.7%. It suggests that the *log* odds of them getting diagnosed with COVID decreases as their age increases, however, the log odds increase if they are an adult as opposed when they are minors. From the relation of both these variables it's indicative that adults with ages in the range of $[20, 25]$ are more likely to get COVID than people belonging to age ranges higher than these. The estimated regression coefficient for race (RACEA) is 0.001, and is statistically significant at the 0.01 level. This implies that although race (RACEA) is a significant predictor, the minuscule value of the regression coefficient does not change the log-odds that much compared to other variables. Finally, the estimated coefficient for sex (SEX) is 0.213 and is statistically significant at the 0.05. level. This implies that males are 123.7% more likely to contract COVID than other genders.

5.2 What are the driving factors behind infections?

To determine which factors are most significant, not only limited to the socioeconomic variables, but also including variables describing their access to health coverage, work conditions and current COVID-19 status, we performed a random forest classifier on all the variables in the dataset. The random forest classifier helps to combat the class bias that exists within the dataset by assigning levels of importance to variables and using them accordingly in the predictions. Table 5 shows the the confusion matrix obtained for the random forest classifier. From the table we get an accuracy of 91.52% which is significantly better than all the logistic regression models employed in this study. Figure 6 represents a graph that ranks the importance of all the variables according to the model, based on the mean decreasing gini index from all the forests generated.

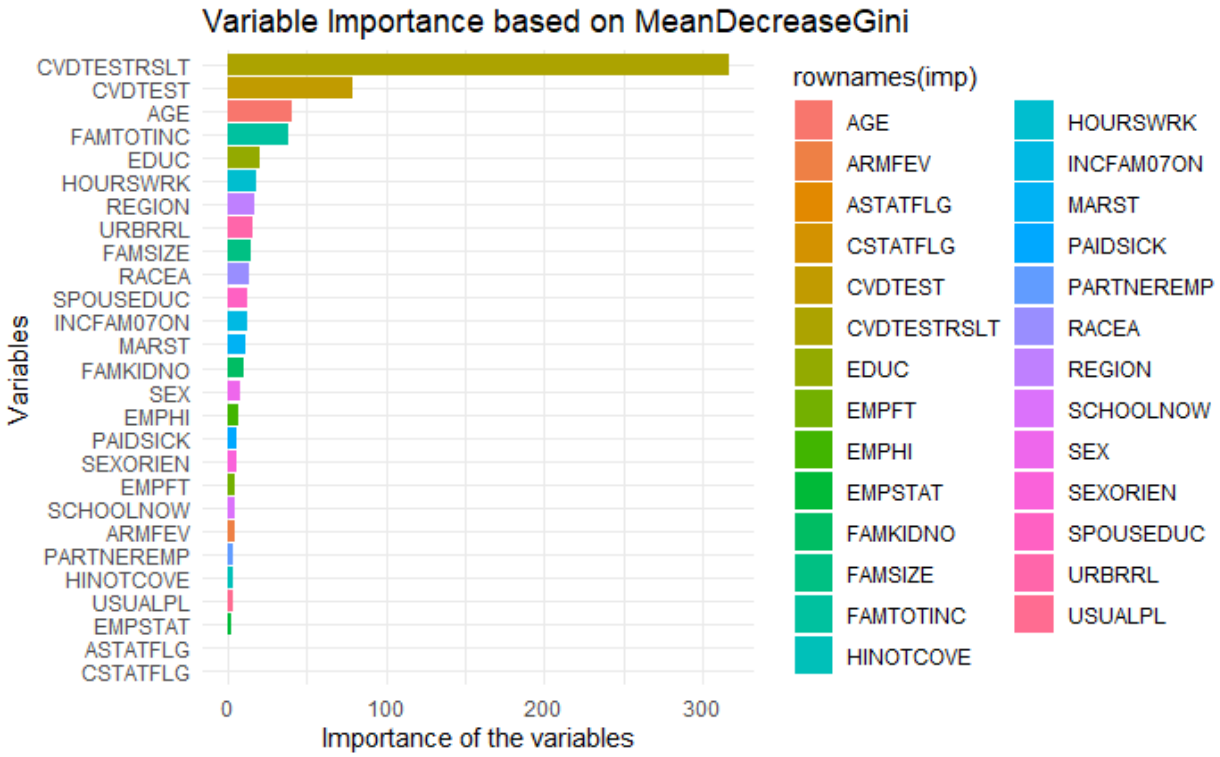

Figure 6: Variable importance for the Random Forest based on mean decrease gini.

It is obvious that the variables CVDTESTRSLT and CVDTEST are highly important in determining the diagnosis of the patient since they are directly correlated to the results (disregarding the case for false positives in the testings). Other important variables when it comes to diagnosing COVID infections are age (AGE), total family income (FAMTOTINC), and education (EDUC) because they have higher mean decrease gini. This means that older people tend to be more vulnerable with regard to being diagnosed with COVID as opposed to younger people. Similarly, people with lower family income in turn may not have the financial capabilities or means to get tested, which gives them a higher chance of being diagnosed with COVID as opposed to their affluent counterparts. People with a higher level of education almost guarantees that they are more aware about the widespread impact of the pandemic and therefore are more likely to get tested than people with a lower level of education.

5.3 What factors impact someone's ability to get tested for COVID?

To understand the factors that impact someone's ability to get tested for COVID-19, we developed a logistic regression model using a variety of predictors such as age (AGE), employment status (EMPSTAT), has a place for medical care (USUALPL), and health insurance coverage status (HINOTCOVE). The model achieved an accuracy of 61% on the test set and the confusion matrix for the model can be found in Table 6 in the appendix. Table 3 shows all the variables that were statistically significant in the model after the forward selection procedure. From looking at the confusion matrix (Table 6), the model had a difficult time classifying people who got tested for COVID (label 1). This is because the model predicted 1,381 out of 1,451 records in the test set as label 0 (did not get tested for COVID) even though the class distribution was not severely imbalanced.

Referring to Table 3, we can observe the estimated coefficients that were statistically significant in the logistic regression model. Regarding employment status (EMPSTAT), the estimated coefficients for employed people (EMPSTAT100) and not employed people (EMPSTAT200) were 1.346 and 1.142, respectively. Both of these coefficients were statistically significant at the 0.001 level. It is worth nothing that the coefficient for EMPSTAT999 was statistically significant as well, however, EMPSTAT999 represents people who had

an unknown employment status. The coefficients for EMPSTAT100 and EMPSTAT200 imply that people were more likely to get tested for COVID if they were employed because the coefficient for EMPSTAT100 was slightly larger than the coefficient for EMPSTAT200. It seems that people who were not employed (EMPSTAT200) got tested as well, however maybe not as frequently as those who were employed. For whether someone had health insurance coverage, HINOTCOVE2 was statistically significant at the 0.001 level. The coefficient for HINOTCOVE2 was -0.474 which implies people who had health insurance coverage were not as likely to get tested for COVID. Also, the coefficient for age (AGE) was -0.004 and was statistically significant at the 0.05 significance level. Since this coefficient was slightly negative but very close to 0, it is difficult to conclude whether there was a clear relationship between age and getting tested for COVID. Regarding whether someone had a usual place for medical care (USUALPL), the estimated coefficients for people who had a usual place (USUALPL2) and people who had more than one usual place (USUALPL3) were 0.440 and 0.771, respectively. The coefficient for USUALPL2 was significant at the 0.01 level and the coefficient for USUALPL3 was significant at the 0.05 level. The coefficient for people who did not have a usual place for medical care (USUALPL1) was not statistically significant. From these coefficients, we can conclude that people with more than one place for medical care tend to get tested more for COVID than those who only have one place or do not have any place for usual medical care.

6 Conclusion

There are a multitude of socioeconomic factors that affect an individual's susceptibility to COVID and access to the proper care. An initial point of analysis was to analyze which socioeconomic factors are most related to COVID infection status. From this analysis, we found that people in the range of 20 to 25 years old were most likely to get COVID. Males were 123.7% found more likely to contract COVID than other genders. In addition, we found that race did not change the log odds of someone getting COVID, compared to the other variables.

Another point of analysis was to examine what were was the driving factors behind infection rates. After fitting a random forest and and ranking the importance of variables, it was determined that age, total family income, and education level were the most important variables that influenced COVID diagnosis. Younger people are less likely to be diagnosed with COVID compared to their older counterparts. Also, families with lower income are not able to get tested as often and therefore have a higher chance of being diagnosed with COVID. Additionally, people with lower education levels are more likely to be diagnosed with COVID based on lack of public health knowledge.

Lastly, being tested for COVID is vital when it comes to lowering infection rates and there are many factors that can influence whether someone gets tested. We were able to conclude that people who were employed tend to get tested for COVID more than the people who were unemployed. Also, we determined that people who had health insurance coverage were less likely to get tested for COVID. Age did not end up having a clear effect on COVID testing ability. Lastly, we found that people who had more than one usual place for medical care got tested more than people who had only one usual place for medical care or no usual place for medical care.

7 Function

The function below allows one to assess model sensitivity based on inputted data by comparing classical and bootstrap standard error estimates for all predictors. At the core, the function fits a logistic regression model as specified in Question 3 above but allows one to specify the data set, number of bootstrap replicates, and the random seed used in the bootstrap procedure. As a result, the function outputs a summary table that compares classical and bootstrap standard error estimates for all predictors. In addition, a plot containing the empirical bootstrap distribution of the standard errors is returned for each variable.

This function is helpful for the analyses detailed above because it allows one to explore an important issue of model sensitivity and asses how the standard errors of the predictors change as new data is used to fit the model. In addition, understanding the distributions of standard errors of the predictors in the model is crucial to developing reliable models since standard errors are used to estimate predictor significance as well as confidence and prediction intervals.

```r
# A. Write replicate function
cust.boot.fn <- function(x, indices){
        x <- x[indices, ]
        # fit logistic regression
        m_log_tmp <- glm(CVDTEST ~ EMPSTAT + HINOTCOVE + AGE + USUALPL + FAMTOTINC,
        data = x, family = binomial)

        # obtain and return the estimates
        return(summary(m_log_tmp)$coefficients[, c(1)])
}

# B. Run boot
run_boot <- function(dt, repl_num, seed){

        # 1. Set seed and run boot
        set.seed(as.numeric(seed))
 boot_est <- boot(data = dt, statistic = cust.boot.fn, R = repl_num)

 # 2. Fit model on all data
 m_log <- glm("CVDTEST ~ EMPSTAT + HINOTCOVE + AGE + USUALPL + FAMTOTINC",


 # 3. Create summary comparison
 tbl_out <- data.frame(og_est = summary(m_log)$coefficients[, c(2)],

         mutate(diff = og_est - boot_est) %>%
         kable(col.names = c("Original Model", "Boostrap", "Difference"), digits = 2)

 # 4. Plot
 p <- ggplot(melt(data.frame(boot_est$t)) %>%
 mutate(variable=as.numeric(str_remove(variable, "X"))-1), aes(x=value)) +
        geom_histogram(bins = 30) +
        facet_wrap(variable~., scales = "free_x",
        labeller = label_bquote(beta [.(variable)])) +
        theme_classic2() +
        ylab("Frequency")+
        xlab("Value")

 # 5. Return results
 return(list(tbl_out, p))
}
```

8 Appendix

8.1 Tables and Figures

Table 2: Summary of the Logistic Regression model for Question 1.

Dependent Variable: CVDDIAG	Logistic Regression Model
AGE	-0.017^{***}
	(0.003)
SEX	0.213^{*}
	(0.111)
RACEA	0.001^{**}
	(0.0003)
ASTATFLG	1.423^{***}
	(0.255)
Constant	-2.076^{***}
	(0.283)
Log Likelihood	$-1{,}012.959$
Akaike Inf. Crit.	2,035.918

Table 3: Summary of Logistic Regression model for question 3.

	Dependent variable:
	CVDTEST
EMPSTAT100	1.346^{***}
	(0.166)
EMPSTAT200	1.142^{***}
	(0.185)
EMPSTAT999	1.473^{***}
	(0.279)
HINOTCOVE2	-0.474^{***}
	(0.153)
AGE	-0.004^{**}
	(0.002)
USUALPL2	0.440^{***}
	(0.150)
USUALPL3	0.771^{**}
	(0.321)
Constant	-1.826^{***}
	(0.212)
Observations	3,500
Log Likelihood	−2,299.774
Akaike Inf. Crit.	4,621.548
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Confusion Matrix for Logistic Regression in question 1.

		True Label		
		Not Infected	Infected	Total
Prediction	Not Infected	586	214	800
	Infected	2	0	2
	Total	588	214	802

Table 5: Confusion Matrix for Random Forest in question 2

		True Label		
		Not Infected	Infected	Total
Prediction	Not Infected	579	59	638
	Infected	9	155	164
	Total	588	214	802

Table 6: Confusion Matrix for Logistic Regression in question 3.

		True Label		
		Never Tested	Tested	Total
Prediction	Never Tested	852	529	1381
	Tested	35	35	70
	Total	887	564	1451

Table 7: Summary of the Variables Analyzed

Group	Variable Name	Description	Converted Data Type	NIU Count
Demographic	AGE	Age	numeric	0
	SEX	Sex	factor	0
	SEXORIEN	Sexual orientation	factor	5,790
	RACEA	Main Racial Background (Pre-1997 Revised OMB Standards), self-reported or interviewer reported	factor	0
	MARSTAT	Legal marital status	factor	5,790
	MARST	Current marital status	factor	5,790
	FAMSIZE	Number of persons in family	numeric	0
	FAMKIDNO	Number of family members under 18 (fam record)	numeric	0
	ARMFEV	Ever served in U.S. Armed Forces, Reserves, or National Guard	factor	5,790
Education	EDUC	Educational attainment	factor	5,790
	SPOUSEDUC	Education level of sample adult's spouse	factor	23,385
	SCHOOLNOW	Currently attending school	factor	5,790
	PARTNEREMP	Employment status of cohabiting or unmarried partner	factor	35,523
	EMPSTAT	Employment status in past 1 to 2 weeks	factor	5,790
	HOURSWRK	Total hours worked last week or usually	factor	19,978
	EMPFT	Usually work 35+ hours per week	factor	19,954
	INCFAM07ON	Total combined family income (2007+)	factor	0
	FAMTOTINC	Total family income, last year (top coded)	numeric	0
Healthcare	USUALPL	Has usual place for medical care	factor	0
	PAIDSICK	Paid sick leave at current job	factor	19,978
	EMPHI	Workplace offered health insurance	factor	19,978
	HINOTCOVE	Health Insurance coverage status	factor	0
COVID	CVDDIAG	Ever told had or likely had coronavirus	factor	17,649
	CVDTEST	Ever been tested for coronavirus	factor	17,649
	CVDTESTRSLT	Had coronavirus, according to test	factor	31,779

Table 8: Contributions Tables

Name	Email	Contributions
Jay Bendre	jdbendre@ucdavis.edu	Data Cleanup, Model Fitting and Diagnostics, Report Editing, Results and Discussion, Conclusions.
John Dinh	jndinh@ucdavis.edu	Report Drafting and Editing, Introduction, Dataset Overview, Feature Engineering, Results and Discussion, Conclusions
Grant Gambetta	gkgambetta@ucdavis.edu	Data Cleanup, Model Fitting and Diagnostics, Report Editing, Results and Discussion, Conclusions.
Ignat Kulinka	ikulinka@ucdavis.edu	Report Drafting, EDA Section, Data Visualizations, RMD editor, Function, Results and Discussion, Conclusions.

8.2 Code

```
# Importing all the required libraries
library(tidyverse)
library(ggplot2)
library(plotly)
library(ipumsr)
library(haven)
library(splitstackshape)
library(caret)
library(randomForest)
library(extrafont)
library(ggpubr)
library(stargazer)
library(randomForestExplainer)
library(extrafont)
library(ggpubr)
library(MASS)

# fix for dplyr select
select <- dplyr::select


# A. Loading the data and creating a dataframe
df_ddi <- read_ipums_ddi("Dataset/nhis_00001.xml")
df <- as.data.frame(read_ipums_micro(df_ddi, verbose = FALSE))
# df %>% head(5)

# B. Preliminary check on the data
#  1. Rows/cols
# c(nrow(df), ncol(df))

#  2. Review classes
# sapply(df, class)
# Note: some of the variables have attached data definitions "haven labeled"

# C. Check for missing values
# sapply(df, function(x) sum(is.na(x)))
```

```r
# D. Dropping variables that are not that important or provide no inisghts in the data
droppable_cols <- c("YEAR","SERIAL","STRATA","PSU",
"NHISPID","HHX","SAMPWEIGHT","LONGWEIGHT",
"PARTWEIGHT", "PERNUM")
df <- df %>% select(-all_of(droppable_cols))
# colnames(df)

# E. Function to plot variables
#  1. Make a function to plot a single distribution
plot_var <- function(col, title){
        plt_dt <- df %>% mutate(var_factor2 := as_factor({{col}})) %>%
                group_by(var_factor2) %>%
                summarize(cnt = n()) %>%
                mutate(prc = cnt/sum(cnt))

        p <- ggplot(plt_dt, aes(x=reorder(var_factor2, cnt, sum), y=cnt)) +
                geom_col(fill = "steelblue", color = "black", width = .5) +
                scale_y_continuous(name = "Count of Respondents",
                labels = scales::comma_format(), expand = expansion(mult = c(0, .3))) +
                scale_x_discrete(name = "", labels = function(x) str_wrap(x, width = 15)) +
                labs(title = paste0(title)) +
                coord_flip() +
                geom_text(label = scales::percent(plt_dt$prc, accuracy = 0.1), hjust = -0.15) +
                theme_classic() +
                theme(plot.title = element_text(size = 16, hjust = 0.5),
                axis.text = element_text(size = 13),
                axis.title = element_text(size = 14),
                aspect.ratio = 1/2,
                text = element_text(family = "Times New Roman"))

        return(p)
}

#  2. Plot COVID-related response variables as an example
covid_vars_plt <- ggarrange(plot_var(CVDDIAG, "Ever Told had or\nLikely had Coronavirus"),



# F. Summarize data variables
#  1. Quick info on the variables
smry <- data.frame(var_name = names(df),
# initial_data_type = sapply(df, function(x) class(x)[3]),
desc = sapply(df, function(x) attributes(x)$label),
final_data_type = ifelse(sapply(df, function(x) class(x)[3]) == "integer", "factor", "numeric"),
num_niu = sapply(df, function(x) sum(as_factor(x)=="NIU")), row.names = NULL)

#  2. Stargazer for the report
# stargazer(smry, summary = FALSE)

# G. Handling SEX AND SEXORIEN variable
#  1. Combining all 'unknown categories' into one
unk <- c(7,8,9)
df$SEX[df$SEX %in% unk] <- 9
```

```r
#  2. Combining Unknown SEXORIEN into "Something else" category
unk <- c(5,7,8)
df$SEXORIEN[df$SEXORIEN %in% unk] <- 4

# H. Handling RACEA Variable
#  1. Combining all 'unknown categories' into one
unk <- c(580, 900, 970, 980, 990)
df$RACEA[df$RACEA %in% unk] <- 900

# I. Handling MARSTAT & MARST Variable
#  1. Review/compare distributions
# df %>%
#         group_by(MARSTAT=as_factor(MARSTAT), MARST=as_factor(MARST)) %>%
#         summarise(n())

#  2. Keep current marital status for this analysis since both are similar
df <- df %>%
        select(-MARSTAT)

#  3. Combine NIU and Unknown into one
unk <- c(0,99)
df$MARST[df$MARST %in% unk] <- 99

#  4. Combine all married labels into one
marr <- c(10,11,12,13)
df$MARST[df$MARST %in% marr] <- 10

# J. Handling FAMSIZE
#  1. Combine Unknowns into one
unk <- c(98,99)
df$FAMSIZE[df$FAMSIZE %in% unk] <- 98

# K. Handling PARTNEREMP
#  1. Combine various levels of unknown response
unk <- c(7,8,9)
df$PARTNEREMP[df$PARTNEREMP %in% unk] <- 9

# L. Handling ARMFEV
#  1. Combine all levels of unknown
unk <- c(97,98,99)
df$ARMFEV[df$ARMFEV %in% unk] <-  99

# M. Handling EDUC, SPOUSEDUC, SCHOOLNOW, by combining all unknowns
unk <- c(996,997,998,999)
df$EDUC[df$EDUC %in% unk] <- 996

unk <- c(97,98,99)
df$SPOUSEDUC[df$SPOUSEDUC %in% unk] <- 99

unk <- c(7,8,9)
df$SCHOOLNOW[df$SCHOOLNOW %in% unk] <- 9

# N. Handling Employment Status
#  1. Working responses
```

```
work <- c(110,111,112)
df$EMPSTAT[df$EMPSTAT %in% work] <- 110

#  2. With job
w_job <- c(120,121,122)
df$EMPSTAT[df$EMPSTAT %in% w_job] <- 120

#  3. Unemployed
unemployed <- c(200,210,211:217)
df$EMPSTAT[df$EMPSTAT %in% unemployed] <- 200

#  4. Unknown
unk <- c(997:999)
df$EMPSTAT[df$EMPSTAT %in% unk] <- 999

# O. Handling HOURSWRK by replacing number of hours unknown into 0
unk <- c(0,97:99)
df$HOURSWRK[df$HOURSWRK %in% unk] <- 0

# P. Handling PAIDSICK
#  1. Combined unknowns into one
df <- df %>%
        mutate(PAIDSICK = replace(PAIDSICK, PAIDSICK > 4, 9))

# Q. Mutating EMPHI and EMPFT
df <- df %>%
        mutate(EMPHI = replace(EMPHI, EMPHI > 4, 9))

df <- df %>%
        mutate(EMPFT = replace(EMPFT, EMPFT > 7, 7))

# R. Mutating USUALPL
df <- df %>%
        mutate(USUALPL = replace(USUALPL, USUALPL == 3,2))

df <- df %>%
        mutate(USUALPL = replace(USUALPL, USUALPL >= 7,9))

# S. Mutating HINOTCOVE
#  1. Combine all unknowns into one
df <- df %>%
        mutate(HINOTCOVE = replace(HINOTCOVE,HINOTCOVE >4,9))

# T. Mutating CVDTEST
#  1. Combine all unknowns into one
df <- df %>%
        mutate(CVDTEST = replace(CVDTEST,CVDTEST >4,9))

# U. Mutating CVDDIAG
#  1. Combine all unknowns into one
df <- df %>%
        mutate(CVDDIAG = replace(CVDDIAG,CVDDIAG >4,9))

# V. Mutating CVDTESTRSLTS
```

```r
#  1. Combine all unknowns into one
df <- df %>%
        mutate(CVDTESTRSLT = replace(CVDTESTRSLT,CVDTESTRSLT >4,9))

# W. Review relation ship between health insurance coverage and coronavirus infection
#  1. Summarize data and filter to complete data
plt_dt <- df %>%
        mutate(HINOTCOVE=as_factor(HINOTCOVE),CVDTESTRSLT=as_factor(CVDTESTRSLT)) %>%
        filter(HINOTCOVE %in% c("Yes, has no coverage", "No, has coverage")
        & CVDTESTRSLT %in% c("Yes", "No")) %>%
        group_by(HINOTCOVE, CVDTESTRSLT) %>%
        summarize(cnt = n()) %>%
        ungroup() %>%
        group_by(HINOTCOVE) %>%
        mutate(prcnt = cnt/sum(cnt)) %>%
        arrange(HINOTCOVE)

#   HINOTCOVE            CVDTESTRSLT   cnt  prcnt (total)
# 1 No, has coverage     No           4659 0.860
# 2 No, has coverage     Yes           440 0.0812
# 3 Yes, has no coverage No            259 0.0478
# 4 Yes, has no coverage Yes            59 0.0109

#  2. Plot stacked barchart for all four groups
ggplot(plt_dt, aes(x=CVDTESTRSLT, y=cnt, fill=CVDTESTRSLT)) +
        geom_bar(position = "dodge", stat="identity", color="black") +
        facet_wrap(~HINOTCOVE, labeller = labeller(HINOTCOVE =
        c("No, has coverage" = "Health Insurance Coverage",
        "Yes, has no coverage" = "No Health Insurance Coverage"))) +
        scale_x_discrete("Had Coronavirus According to Test", labels=c("Negative", "Positive")) +
        scale_y_continuous(name = "Count of Respondents",
        expand = expansion(mult = c(0, .2))) +
        scale_fill_manual(values=c("#4682B4", "#B47846")) +
        geom_text(label = scales::percent(plt_dt$prcnt, accuracy = 0.1),
        position = position_dodge(width=0.9), vjust=-0.5) +
        theme_classic() +
        theme(legend.position="none",
        text = element_text(family = "Times New Roman"),
        aspect.ratio = 1/2,
        axis.text = element_text(size = 13),
        axis.title = element_text(size = 14),
        strip.text = element_text(size = 13))

# X. Seeing results between getting sick leave and covid result tests
#  1. Summarize and filter data
plt_dt2 <- df %>%
        group_by(PAIDSICK = as_factor(PAIDSICK), CVDTESTRSLT = as_factor(CVDTESTRSLT)) %>%
        summarize(count = n()) %>%
        filter(PAIDSICK %in% c("Yes", "No") &
        CVDTESTRSLT %in% c("Yes", "No")) %>%
        group_by(PAIDSICK) %>%
        mutate(prcnt = count/sum(count)) %>%
        arrange(PAIDSICK)
```

```r
ggplot(plt_dt2, aes(fill=CVDTESTRSLT, y=count, x=CVDTESTRSLT)) +
        geom_bar(position="dodge", stat="identity", color = "black") +
        facet_wrap(~PAIDSICK, labeller = labeller(PAIDSICK =
        c("No" = "No Paid Sick Leave at Current Job",
        "Yes" =  "Paid Sick Leave at Current Job"))) +
        scale_y_continuous(name = "Count of Respondents",
        labels = scales::comma_format(),
        expand = expansion(mult = c(0, .2))) +
        geom_text(label = scales::percent(plt_dt2$prcnt,
        accuracy = 0.1), position =
        position_dodge(width=0.9), vjust=-0.25) +
        xlab("Had Coronavirus According to Test") +
        scale_x_discrete(labels = c("No","Yes")) +
        scale_fill_manual(values = c("#4682B4","#B47846")) +
        theme_classic() +
        theme(legend.position="none",
        text = element_text(family = "Times New Roman"),
        aspect.ratio = 1/2,
        axis.text = element_text(size = 13),
        axis.title = element_text(size = 14),
        strip.text = element_text(size = 13))

# Y. Seeing the number of people who got tested
# given they have health insurance coverage from the company
sample_df <- df %>%
        group_by(CVDTEST = as.factor(CVDTEST), EMPHI = as.factor(EMPHI)) %>%
        summarize(count = n()) %>%
        filter(CVDTEST %in% c(1,2) &
        EMPHI %in% c(1, 2)) %>%
        group_by(EMPHI) %>%
        mutate(prcnt = count/sum(count)) %>%
        arrange(EMPHI)

ggplot(sample_df, aes(fill=CVDTEST, y=count, x=CVDTEST)) +
        geom_bar(position="dodge", stat="identity", color = "black") +
        facet_wrap(~EMPHI, labeller =
        labeller(EMPHI = c("0" = "NIU","1" =  "Employer Insurance Not Offered",
        "2" = "Employer Insurance Offered", "9" = "Unknown"))) +
        scale_y_continuous(name = "Count of Respondents",
        labels = scales::comma_format(), expand =
        expansion(mult = c(0, .2))) +
        geom_text(label = scales::percent(sample_df$prcnt, accuracy = 0.1),
        position = position_dodge(width=0.9), vjust=-0.25) +
        xlab("Ever Been Tested for Coronavirus") +
        scale_x_discrete(labels = c("No","Yes")) +
        scale_fill_manual(values = c("#4682B4","#B47846")) +
        theme_classic() +
        theme(legend.position="none",
        text = element_text(family = "Times New Roman"),
        aspect.ratio = 1/2,
        axis.text = element_text(size = 13),
        axis.title = element_text(size = 14),
        strip.text = element_text(size = 13))
```

```r
# Z. Does having a usual place for medical care affect the testing
plt_dt3 <- df %>%
        group_by(USUALPL = as_factor(USUALPL), CVDTEST = as_factor(CVDTEST)) %>%
        summarize(count = n()) %>%
        filter(USUALPL %in% c("Yes, has a usual place or Yes",
        "There is no place or No") &
        CVDTEST %in% c("Yes", "No")) %>%
        group_by(USUALPL) %>%
        mutate(prcnt = count/sum(count)) %>%
        arrange(USUALPL)

ggplot(plt_dt3, aes(x=CVDTEST, y=count, fill=CVDTEST)) +
        geom_bar(position="dodge", stat="identity", color = "black") +
        facet_wrap(~USUALPL, labeller =
        labeller(USUALPL = c("There is no place or No" =
        "No Usual Place Medical Care","Yes, has a usual place or Yes" =
        "Has Usual Place for Medical Care"))) +
        scale_y_continuous(name = "Count of Respondents",
        labels = scales::comma_format(),
        expand = expansion(mult = c(0, .2))) +
        geom_text(label = scales::percent(plt_dt3$prcnt,
        accuracy = 0.1), position =
        position_dodge(width=0.9), vjust=-0.25) +
        xlab("Ever Been Tested for Coronavirus") +
        scale_x_discrete(labels = c("No","Yes")) +
        scale_fill_manual(values = c("#4682B4","#B47846")) +
        theme_classic() +
        theme(legend.position="none",
        text = element_text(family = "Times New Roman"),
        aspect.ratio = 1/2,
        axis.text = element_text(size = 13),
        axis.title = element_text(size = 14),
        strip.text = element_text(size = 13))


# AA. Create a stratified sample
#  1. Review CVDDIAG distribution
df %>%
        group_by(CVDDIAG_CD=CVDDIAG,
             CVDDIAG=as_factor(CVDDIAG)) %>%
        summarize(Count = n())

#  2. Exclude NIU and Unknown values
sample_df <- df %>%
        filter(!(CVDDIAG %in% c(0,9)))

#  3. Create a stratified sample of the data
set.seed(1234)
strat <- stratified(sample_df,group = 'CVDDIAG',size = 2000)

#  4. Refactor the sample to zero's and one's
strat <- strat %>%
        mutate(CVDDIAG = as.factor(CVDDIAG-1))
```

```r
#  5. Review the distribution
strat %>%
        group_by(CVDDIAG) %>%
        summarize(Count = n()) %>%
        mutate(Percent_Total = Count/sum(Count))

#  6. Visualising it
strat_plt <- strat %>%
        group_by(CVDDIAG) %>%
        summarize(cnt = n()) %>%
        mutate(prc = cnt/sum(cnt))

ggplot(strat_plt, aes(x=CVDDIAG, y=cnt)) +
        geom_col(color = "black", fill=c("#4682B4","#B47846")) +
        scale_y_continuous(name = "Count of Cases",
        labels = scales::comma_format(),
        expand = expansion(mult = c(0, .1))) +
        theme_classic() +
        geom_text(label = scales::percent(strat_plt$prc, accuracy = 0.1),
                        position = position_dodge(width=0.9), vjust=-0.25) +
        scale_x_discrete("Ever Told had or Likely had Coronavirus",
        labels = c("No","Yes")) +
        theme(legend.position="none",
        text = element_text(family = "Times New Roman"),
        aspect.ratio = 1/2,
        axis.text = element_text(size = 13),
        axis.title = element_text(size = 14),
        strip.text = element_text(size = 13))
# AB. Split data into training and testing
set.seed(101)
idx <- sample.int(n = nrow(strat),


s.train <- strat[idx,]
s.test <- strat[-idx,]

# AC. Selecting personal information variables
#  1. Create a vector of column names
per_info_vars <- c("AGE", "SEX", "MARST", "RACEA",



#  2. Subset training and testing datasets
data_to_study.train <- s.train %>%
        select(all_of(per_info_vars))

data_to_study.test         <- s.test %>%
        select(all_of(per_info_vars))

#  3. Use stepAIC to find the optimal combination of variables
model <- glm(CVDDIAG ~ ., data_to_study.train, family = 'binomial') %>%
        stepAIC(trace = FALSE, k=2)
```

```
#  4. Review model summary
summary(model)

#  5. Review the confusion matrix
lr_probs <- predict(model, newdata = data_to_study.test, type = 'response')
lr_predicted <- ifelse(lr_probs < 0.5, 0, 1)
confusionMatrix(factor(lr_predicted),


library(stargazer)
conf.mat <- as.data.frame.matrix(cf$table)
stargazer(conf.mat, title = "Confusion Matrix", summary = FALSE)
```


```
# A. Using logistic regression to see what
# affects the diagnosis and selecting the best model using stepAIC

library(randomForestExplainer)

# Creating samples

x.train <- s.train %>% dplyr::select(-CVDDIAG)
y.train <- s.train %>% dplyr::select(CVDDIAG)

x.test <- s.test %>% dplyr::select(-CVDDIAG)
y.test <- s.test %>% dplyr::select(CVDDIAG)

model.rf <- randomForest::randomForest(x = x.train,
y = as.factor(y.train$CVDDIAG), ntree = 500,
importance = T, proximity = T)
print(model.rf)
print(importance(model.rf,2))

plot(model.rf)

predictions <- predict(model.rf, newdata = x.test)
cf.rf <- confusionMatrix(predictions,droplevels(as.factor(y.test$CVDDIAG)))


frame <- measure_importance(model.rf)
frame
plot_importance_ggpairs(frame)


plot_importance_rankings(frame)

imp <- as.data.frame(model.rf$importance[,3:4])

imp <- imp %>% arrange(desc(MeanDecreaseGini))

ggplot(imp, aes(x = reorder(rownames(imp), MeanDecreaseGini),
y = MeanDecreaseGini, fill = rownames(imp))) +
geom_bar(stat = "identity") + ggtitle("Variable Importance based on MeanDecreaseGini") +
```

```
theme_minimal() + coord_flip() +
xlab("Variables") + ylab("Importance of the variables")


df_ddi <- read_ipums_ddi("Dataset/nhis_00001.xml")
df <- as.data.frame(read_ipums_micro(df_ddi, verbose = FALSE))
cvd_test_df <- df %>% dplyr::filter(CVDTEST == 1 | CVDTEST == 2)
# select only yes or no for covid test variable
cvd_test_df$CVDTEST  <- ifelse(cvd_test_df$CVDTEST == 1, 0, 1)
# relevel covid test variable: # 0 = no test, 1 = test
barplot(table(cvd_test_df$CVDTEST))
cvd_test_df <- cvd_test_df %>% dplyr::select(NHISHID, SEX, AGE,
FAMSIZE, EMPSTAT, HOURSWRK, PAIDSICK,
EMPHI, EMPFT, USUALPL, HINOTCOVE,
INCFAMO7ON, FAMTOTINC, CVDTEST)
# re factor variables, 7 = unknown
cvd_test_df$PAIDSICK[cvd_test_df$PAIDSICK == 8] = 7
cvd_test_df$PAIDSICK[cvd_test_df$PAIDSICK == 9] = 7
cvd_test_df$EMPHI[cvd_test_df$EMPHI == 8] = 7
cvd_test_df$EMPHI[cvd_test_df$EMPHI == 9] = 7
cvd_test_df$EMPFT[cvd_test_df$EMPFT == 8] = 7
cvd_test_df$EMPFT[cvd_test_df$EMPFT == 9] = 7
cvd_test_df$USUALPL[cvd_test_df$USUALPL == 8] = 7
cvd_test_df$USUALPL[cvd_test_df$USUALPL == 9] = 7
cvd_test_df$HOURSWRK <- as.double(cvd_test_df$HOURSWRK)
cvd_test_df <- cvd_test_df %>% mutate_if(is.integer, as.factor)
cvd_test_df
# oversample the data
set.seed(123)
vec <- c(3000, 2000)
names(vec) <- c(0, 1)
strat_sample <- stratified(cvd_test_df, group = 'CVDTEST', size = vec)
strat_sample

table(strat_sample$CVDTEST)/length(strat_sample$CVDTEST)
barplot(table(strat_sample$CVDTEST))
set.seed(123)
lr_train <- strat_sample %>% dplyr::sample_frac(0.70)
lr_test <- dplyr::anti_join(strat_sample, lr_train, by = "NHISHID")
lr_train <- lr_train %>% dplyr::select(-c("NHISHID"))
lr_test <- lr_test %>% dplyr::select(-c("NHISHID"))

lr2 <- glm(data = lr_train, CVDTEST ~ ., family = 'binomial')
summary(lr2)
null_model <- glm(data = lr_train, CVDTEST ~ 1, family = 'binomial')
full_model <- glm(data = lr_train, CVDTEST ~ ., family = 'binomial')
step <- stepAIC(null_model,
scope = list(lower = ~1, upper = full_model),
direction = 'both', k = 2, trace = 0)
step$anova
# final model
lr2 <- glm(data = lr_train, CVDTEST ~ EMPSTAT + HINOTCOVE + AGE + USUALPL + FAMTOTINC,
family = 'binomial')
summary(lr2)
```

```
lr2_probs <- predict(lr2, newdata = lr_test, type = 'response')
lr2_predicted <- ifelse(lr2_probs < 0.5, 0, 1)

confusionMatrix(factor(lr2_predicted, levels=min(lr_test$CVDTEST):max(lr_test$CVDTEST)),
factor(lr_test$CVDTEST, levels=min(lr_test$CVDTEST):max(lr_test$CVDTEST)))
```

References

[1] JA Patel, FBH Nielsen, AA Badiani, S Assi, VA Unadkat, B Patel, R Ravindrane, and H Wardle. Poverty, inequality and covid-19: the forgotten vulnerable. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7221360/, Jun 2020.

[2] Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. Ipums usa: Version 10.0. https://doi.org/10.18128/D010.V10.0.

[3] Efrat Shadmi, Yingyao Chen, Inês Dourado, Inbal Faran-Perach, John Furler, Peter Hangoma, Piya Hanvoravongchai, Claudia Obando, Varduhi Petrosyan, Krishna D Rao, and et al. Health equity and covid-19: global perspectives. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7316580/, Jun 2020.