Predicting COVID-19 Vaccination Status in America

Calvin Kapral

Dataset:

The dataset for this project originates from a Pew Research Center survey conducted in September 2022, which focused on U.S. public opinions about scientists, religion, and COVID-19 policies. This nationally representative survey, part of the American Trends Panel (ATP), includes responses from over 10,000 adults randomly selected from across the United States. The survey aims to explore public trust in elected officials and scientists, analyze behaviors and opinions related to COVID-19 vaccination uptake, and examine demographic and psychological factors influencing vaccine hesitancy.

Key variables in the dataset include whether individuals were fully vaccinated against COVID-19 (COVID_vaccinated), levels of trust in elected officials (CONF_a, aggregated into two groups), and confidence in scientists (Conf_SCI, derived from confidence in medical scientists (CONF_f) and other scientists (CONF_g)). Additional demographic variables such as age, race, education, and income are also included, offering insights into how various factors contribute to vaccine uptake and hesitancy. The primary goal of this study is to identify risk factors that pose obstacles to COVID-19 vaccination uptake. The analysis focuses on understanding the relationship between trust in elected officials and scientists, as well as the impact of demographic characteristics on vaccine hesitancy. This project adheres to a strict confidentiality agreement: *"I understood and agreed to that the data used in this report must be kept confidential and must not be shared or distributed outside this class. The statistical analysis, the result and the report were agreed to be only used in this class and will not be posted or published without the authorization of Prof. Jimin Ding."*

Problem 1: Association Between COVID Vaccination Uptake and Trust in Elected Officials:

Null and Alternative Hypotheses:

1. **Null Hypothesis (H_0):** There is no association between COVID vaccination uptake and trust in elected officials.
2. **Alternative Hypothesis (H_1):** There is an association between COVID vaccination uptake and trust in elected officials.

The data was summarized in the following contingency table:

```
##
##     High Trust Low Trust
##   N       1785       442
##   Y       5759      2430
```

Statistical Test

We used **Pearson's Chi-Square Test with Yates' continuity correction** to assess the association between trust in elected officials and COVID vaccination uptake. This test is

appropriate because it evaluates the independence between two categorical variables (trust and vaccination status).

Results

- **Chi-Square Statistic (X^2):** 84.168
- **Degrees of Freedom (df):** 1
- **P-value:** < 2.2e-16

The expected values for each category were calculated as follows:

```
##
##     High Trust Low Trust
##   N    1612.95  614.0499
##   Y    5931.05 2257.9501
```

Since the **p-value** is significantly less than 0.05, we **reject the null hypothesis**. There is a statistically significant association between trust in elected officials and COVID vaccination uptake.

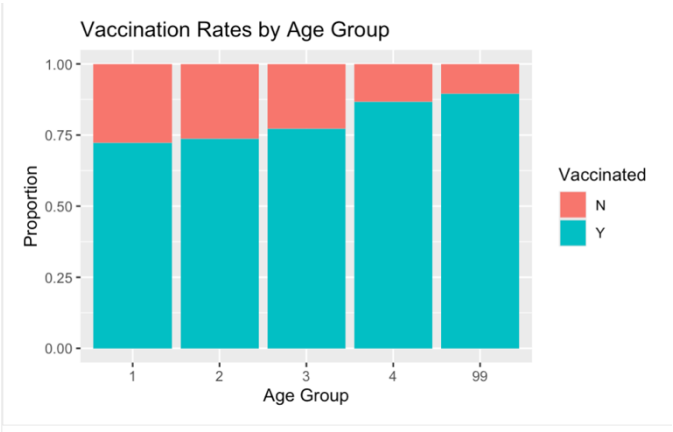
Individuals with **High Trust** in elected officials are significantly more likely to be vaccinated (5759) compared to those with **Low Trust** (2430). Similarly, among those not vaccinated, individuals with High Trust (1785) outnumber those with Low Trust (442). This result underscores the importance of trust in elected officials as a predictor of vaccination behavior. Also, since our raw values are higher in high trust for yes, this is evidence that there is strong correlation in this regard.

Problem 2: Stratified Analysis of Vaccination Rates

1. Vaccination Rates by Age Group

The analysis stratifies vaccination rates by age groups (F_AGECAT), as shown in the bar plot **"Vaccination Rates by Age Group"**:

- Younger age groups (1 and 2) have lower vaccination proportions compared to older groups (3 and 4).
- Older individuals (Group 4) exhibit the highest vaccination rates, while younger individuals (Group 1) show the lowest.



Key Insight: Age has a significant impact on vaccination rates, with older individuals being more likely to be vaccinated. This may reflect factors such as higher perceived risk or targeted vaccination campaigns for older adults.

2. Vaccination Rates by Confidence in Science

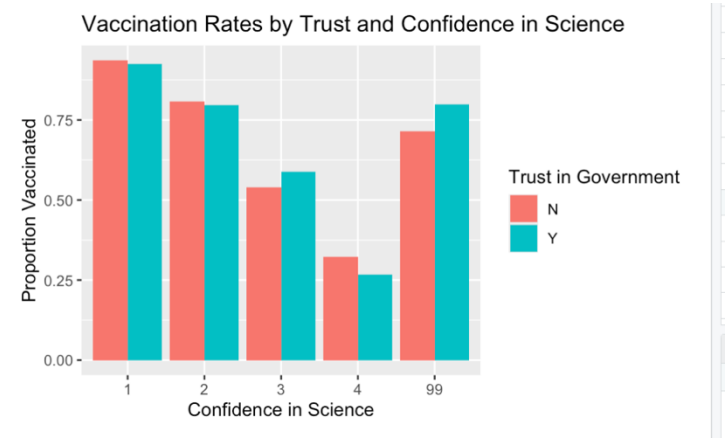
The stratified analysis of vaccination rates by confidence in science (Conf_SCI), shown in the bar plot **"Vaccination Rates by Confidence in Science"**, reveals:

- High confidence levels (Categories 1 and 2) are associated with high vaccination rates.
- Low confidence levels (Categories 3 and 4) show a substantial drop in vaccination uptake.
- Unknown confidence (Category 99) also demonstrates relatively high vaccination rates.

Key Insight: Confidence in science is a critical predictor of vaccination behavior. Individuals with higher confidence are more likely to be vaccinated, while low confidence correlates with vaccine hesitancy.

3. Vaccination Rates by Trust and Confidence in Science

The bar plot **"Vaccination Rates by Trust and Confidence in Science"** examines the interplay between trust in government and confidence in science:



- When confidence in science is high (Categories 1 and 2), vaccination rates are high regardless of trust in government.
- At lower confidence levels (Categories 3 and 4), trust in government plays a larger role in influencing vaccination behavior, with individuals with high trust showing higher vaccination rates compared to those with low trust.
- For unknown confidence levels (Category 99), trust in government appears to influence vaccination rates more significantly.

4. Why Vaccination Rates Vary Less by Trust When Conditioned on Confidence in Science

- High confidence in science appears to mitigate the effect of trust in government on vaccination behavior. When individuals trust scientific institutions, they are more likely to vaccinate, reducing reliance on trust in government officials.
- Conversely, when confidence in science is low, trust in government becomes a more significant factor, as individuals may look to government figures for guidance in the absence of scientific trust.
- **Age**: Older individuals are more likely to be vaccinated, possibly due to perceived risk or public health targeting.

- **Confidence in Science**: Higher confidence in science correlates with higher vaccination rates and reduces reliance on government trust.
- **Trust in Government**: Trust in government plays a larger role when confidence in science is low, highlighting its compensatory effect in vaccine decision-making.

These results emphasize the importance of strengthening public confidence in science and tailoring vaccination campaigns to address age-related disparities and trust dynamics.

Problem 3: Analysis of Selected Variables Education and Political Affiliation

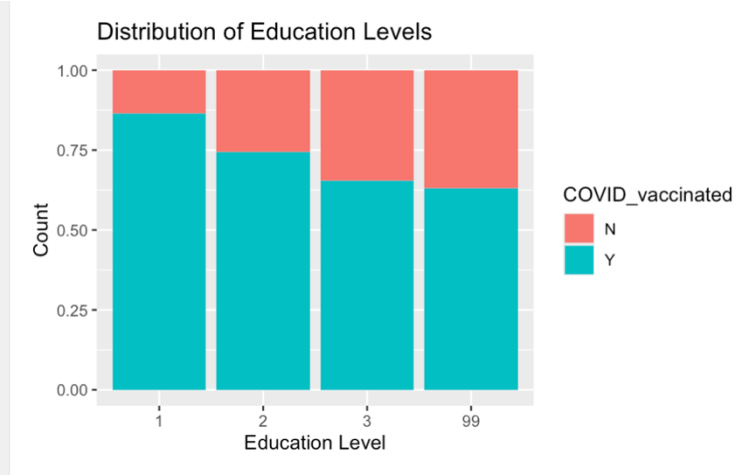
Variable Justification

- **Education Levels (F_EDUCCAT)**: Education is a key determinant in shaping health literacy and vaccine acceptance. By analyzing this variable, we can assess how education influences vaccination rates.
- **Political Affiliation (F_PARTY_FINAL)**: Political beliefs can significantly impact attitudes toward public health measures, including vaccination. This variable provides insights into how party affiliation correlates with vaccine uptake.

1. Distribution of Education Levels

The visualization of education levels and vaccination rates shows:

- **Education Level 1 (College Graduates)**: The highest vaccination rates are observed among individuals with college degrees.
- **Education Level 2 (Some College)**: Vaccination rates are slightly lower compared to college graduates but remain relatively high.
- **Education Level 3 (High School Graduate or Less)**: The lowest vaccination rates are observed in this group, indicating a potential gap in health literacy or access to vaccines.
- **Category 99 (Unknown)**: Vaccination rates are similar to lower education levels.



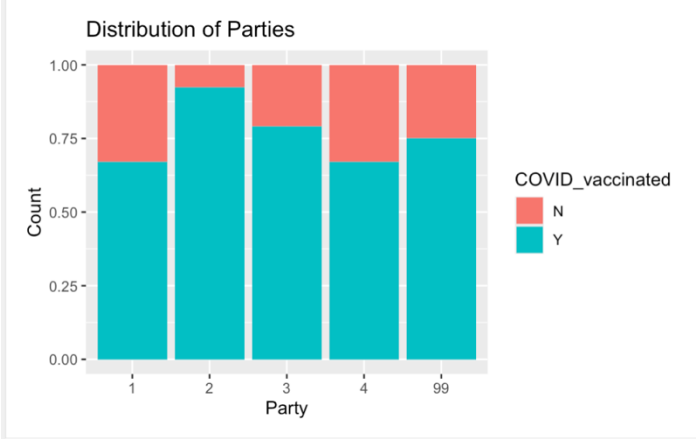
Key Insight: Higher education levels correlate with increased vaccination uptake, reflecting the role of education in shaping attitudes and behaviors toward vaccines.

2. Distribution of Political Affiliation

The visualization of political affiliation and vaccination rates shows:

- **Party 1 (Republicans)**: Lower vaccination rates are observed in this group compared to other affiliations.

- **Party 2 (Democrats)**: The highest vaccination rates are seen among Democrats, suggesting greater acceptance or promotion of public health measures.
- **Party 3 (Independents)**: Vaccination rates are moderate, between Republicans and Democrats.
- **Party 4 (Others)**: Rates are comparable to Independents but slightly lower.
- **Category 99 (Unknown)**: Vaccination rates show a mix, reflecting uncertain political alignment.



Key Insight: Political affiliation significantly influences vaccination behavior, with Democrats showing the highest acceptance and Republicans the lowest. This trend highlights the impact of political ideology on health decisions.

Conclusion

- Education and political affiliation are strong predictors of vaccination rates. College graduates and Democrats exhibit the highest uptake, while high school graduates or less and Republicans show lower vaccination rates.
- Public health campaigns should focus on addressing gaps in vaccine acceptance, particularly among lower-educated populations and groups with political hesitancy.

These findings emphasize the importance of tailoring vaccine outreach programs to demographic and political contexts to increase overall vaccination rates.

Problem 4: Cleaning the Dataset

1. Replacing Missing Values (98 and 99)

The values 98 and 99 in the dataset represent missing data. These values were replaced with NA to standardize the missing data representation and facilitate proper handling during analysis. This approach ensures that the dataset complies with R's missing value conventions.

2. Removing Unnecessary Columns

The following columns were removed because they are irrelevant to the analysis or redundant:

- QKEY: A unique identifier that is unnecessary for analysis.
- INTERVIEW_START and INTERVIEW_END: Metadata columns that do not contribute to the analysis.

- DEVICE_TYPE: Information about how the survey was conducted, not directly relevant to vaccination analysis.
- FORM: Refers to the survey form and does not impact analysis.
- WEIGHT_*: Survey weights that are unnecessary for the current analysis.

Problem 5: Building a Binomial Model to Identify Risk Factors for Non-Vaccination

Model Selection

We used a **binomial logistic regression model** to identify the risk factors for non-vaccination. Logistic regression is suitable for binary outcomes (e.g., vaccinated or not vaccinated) and allows us to estimate the odds of non-vaccination based on multiple categorical predictors.

Model Setup

In logistic regression, choosing a reference group is crucial because all comparisons in the model are made relative to that group. By setting non-vaccinated (0) as the reference group for the dependent variable (COVID_vaccinated), we gain insights into how different predictors (e.g., age, education, race, etc.) affect the odds of being vaccinated relative to being non-vaccinated. So, for all the predictor variables, we set the reference group as the subgroup within the variable thought most likely to be un-vaccinated. It is also important to turn the categorical variables into factors that can be grouped which I do using the as.factor function in R.

1. **Response Variable:**
    - COVID_vaccinated: Binary outcome where 1 indicates vaccinated and 0 indicates not vaccinated and 0 is the reference group.
2. **Predictor Variables:**
    - F_AGECAT: Age group, with older adults (level 4) as the reference group.
    - F_EDUCCAT: Education level, with high school graduates or less (Level 3) as the reference group.
    - F_PARTY_FINAL: Political affiliation, with Republicans (level 1) as the reference group.
    - F_ATTEND: Religious attendance, with "never attends" (level 6) as the reference group.

- F_INC_TIER2: Income tier, with the lowest income (level 1) as the reference group.
- F_RACETHNMOD: Race, with black (level 2) as the reference group.

Model Findings

Key Findings

In the output of the code, we get the log-odds ratio for the sub-group of the variable compared to the reference group.

```
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.53997    0.14407  -3.748 0.000178 ***
## data$F_AGECAT1     -1.29764    0.11185 -11.601  < 2e-16 ***
## data$F_AGECAT2     -1.19100    0.07805 -15.259  < 2e-16 ***
## data$F_AGECAT3     -0.86491    0.07767 -11.135  < 2e-16 ***
## data$F_EDUCCAT1     1.04399    0.07702  13.555  < 2e-16 ***
## data$F_EDUCCAT2     0.39872    0.07131   5.591 2.26e-08 ***
```

From the snippet of code, we can see that age levels have negative log-odds ratios indicating that they are evidence of lower vaccination rates among young people compared to the reference group of older people, whereas higher educated people shower a higher log odds ratio of being vaccinated.

1. **Significant Predictors:**
    - **Age**: Younger individuals (Categories 1, 2, and 3) are significantly less likely to be vaccinated compared to older adults (Category 4).
    - **Education**: College graduates (Category 1) and individuals with some college education (Category 2) are more likely to be vaccinated compared to high school graduates or less.
    - **Political Affiliation**: Democrats (Category 2) and Independents (Category 3) are significantly more likely to be vaccinated compared to Republicans.
    - **Religious Attendance**: Individuals who rarely attend religious services (Categories 2-6) are more likely to be vaccinated compared to those who never attend.
    - **Income**: Higher income levels (Categories 2 and 3) are associated with increased vaccination likelihood.
    - **Race**: Asian individuals (Category 5) are significantly more likely to be vaccinated compared to White individuals.

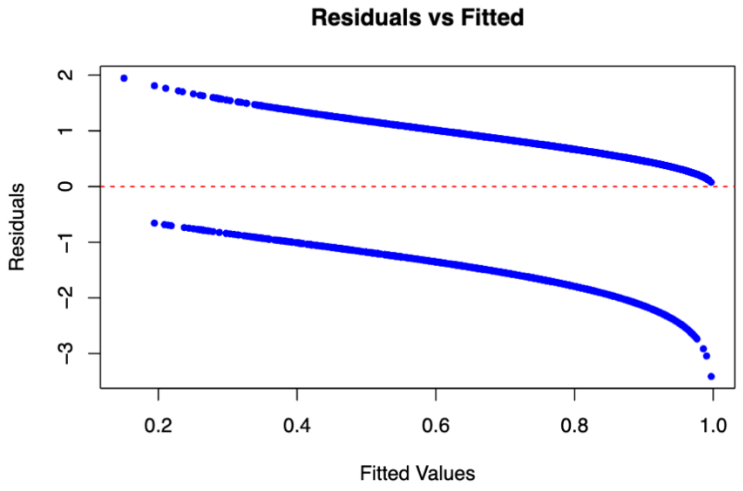
2. **Non-Significant Predictors:**
    - Category 4 for race did not show significant differences in vaccination likelihood.

Logistic regression estimates the relationship between predictor variables and a binary outcome (e.g., vaccinated vs. non-vaccinated) using **log-odds** as the scale. To make the results more interpretable, the log-odds coefficients are converted to **odds ratios (OR)**, which quantify the likelihood of an event occurring (e.g., vaccination) for one group compared to a reference group. In this case, an OR of 1 means no difference among the groups, whereas an OR above one means that there is an increased odds outcome compared to the reference group. We further show this by creating confidence intervals of odds-ratios, where we can we sinc the age variables are less than one, they are lower odds of being vaccinated at lower ages, whereas education is greater than one, so they have higher odds ratios of being vaccinated at higher education levels.

```
## (Intercept)          0.4395807 0.7733033
## data$F_AGECAT1       0.2194520 0.3402646
## data$F_AGECAT2       0.2605735 0.3538619
## data$F_AGECAT3       0.3613774 0.4900280
## data$F_EDUCCAT1      2.4428714 3.3039562
## data$F_EDUCCAT2      1.2954970 1.7133927
```

Model Diagnostics

The **Residuals vs Fitted Plot** is a key diagnostic tool used to evaluate the goodness-of-fit of a regression model, including logistic regression models. It helps assess whether the model's assumptions are satisfied and whether there are patterns in the residuals that indicate potential issues with the model.



Residuals: These are the differences between the observed outcomes and the model's predicted probabilities. For logistic regression, the residuals can be measured in various ways (e.g., Pearson residuals or deviance residuals).

Fitted Values: These are the predicted probabilities of the outcome (e.g., the predicted probability of being vaccinated).

The plot displays **residuals** (y-axis) against **fitted values** (x-axis).

In the provided plot:

- **Mild curvature in residuals**:
    - There is a slight curve or systematic pattern in the residuals. This suggests potential **non-linearity** in the relationship between predictors and the log-odds of the outcome. The model may not fully capture some complexities in the data.
- **Consistent spread but slight skewness**:
    - The residuals are mostly evenly distributed around zero, indicating a reasonably good fit. However, at higher fitted values (closer to 1), the residuals spread slightly wider and become more negative, suggesting **skewness or systematic bias** in the model.
- **Influential points**:
    - Some data points with large residuals are present. These might represent **outliers** or **cases poorly explained** by the model.

Overall, despite the fact there are some areas for improvement, the residuals vs fitted plots suggests that the logistic regression model fits the data reasonably well.

Next, I checked for overdispersion, and found that the parameter was 0.99 in the model which is less than the threshold at 1.5 so we do not need to account for this by using a different model like the quasi-linear model.

The **Receiver Operating Characteristic (ROC) curve** is a diagnostic tool used to evaluate the performance of a logistic regression model in predicting a binary outcome. In this case, it assesses the model's ability to differentiate between vaccinated and non-vaccinated individuals.

 Shape of the Curve:

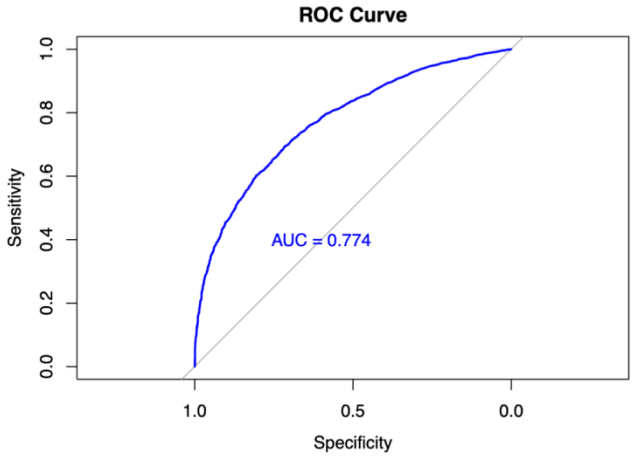
- The ROC curve plots **sensitivity** (true positive rate) on the y-axis against **1 - specificity** (false positive rate) on the x-axis for different thresholds of the predicted probabilities. A perfect model would have an ROC curve that hugs the top-left corner, indicating high sensitivity and specificity.

 AUC (Area Under the Curve):

- The AUC quantifies the overall performance of the model: AUC = 0.5: The model performs no better than random guessing. AUC = 1: The model perfectly separates the classes. AUC between 0.7 and 0.9: Indicates moderate discriminative power.

AUC Value:

The AUC is **0.774**, which indicates that the model has **moderate discriminative power**. This means the model is reasonably good at distinguishing between vaccinated and non-vaccinated individuals but leaves room for improvement.



Conclusion:

Our analysis identified several key factors contributing to vaccine hesitancy. Younger individuals are significantly less likely to be vaccinated compared to older age groups, highlighting a generational divide. Education also plays a crucial role—individuals with higher education levels, such as college graduates, are more likely to get vaccinated than those with less education. Political affiliation strongly influences vaccination status, with Democrats being far more likely to be vaccinated than Republicans. Additionally, those who rarely or never attend religious services tend to have higher vaccination rates compared to regular attendees. Finally, racial and ethnic background is a factor, with Asian and Hispanic individuals showing higher vaccination rates compared to White non-Hispanic groups. These findings can help tailor public health strategies to address specific concerns within these populations and improve vaccination outreach.

Appendix

```
library(dplyr)
library(tidyverse)
load("/Users/calvinkapral/Downloads/ATP (7).RDA")
library(ggplot2)
library(tidyr)

library(reshape2)
library(haven)
library(pROC)
```

Problem 1:

```
data_problem_1 <- ATP %>% mutate(CONF_GOV_grouped = ifelse(CONF_GOV %in% c(1, 2), "Low Trust", "High Trust"))
contingency_table <- table(data_problem_1$COVID_vaccinated, data_problem_1$CONF_GOV_grouped)

contingency_table

chi_sqr <- chisq.test(contingency_table)
chi_sqr
chi_sqr$expected
```

Problem 2:

```
filtered_data <- data_problem_1 %>% select(COVID_vaccinated, F_AGECAT, Conf_SCI, trust) %>% drop_na()

age_vacc_rates <- filtered_data %>% group_by(F_AGECAT, COVID_vaccinated) %>% summarise(count = n()) %>% mutate(prop = count / sum(count))

conf_sci_vacc_rates <- filtered_data %>% group_by(Conf_SCI, COVID_vaccinated) %>% summarise(count = n()) %>% mutate(prop = count / sum(count))

ggplot(age_vacc_rates, aes(x = as.factor(F_AGECAT), y = prop, fill = COVID_vaccinated)) + geom_bar(stat = "identity", position = "stack") + labs(title = "Vaccination Rates by Age Group", x = "Age Group", y = "Proportion", fill = "Vaccinated")

ggplot(conf_sci_vacc_rates, aes(x = as.factor(Conf_SCI), y = prop, fill = COVID_vaccinated)) + geom_bar(stat = "identity", position = "stack") + labs(title = "Vaccination Rates by Confidence in Science", x = "Confidence in Science", y = "Proportion", fill = "Vaccinated")
```

```
trust_confidence_vacc <- filtered_data %>% group_by(trust, Conf_SCI, COVID_vaccinated) %>% summarise(count = n()) %>% mutate(prop = count / sum(count))

grouped_bar_data <- trust_confidence_vacc %>% filter(COVID_vaccinated == "Y")

ggplot(grouped_bar_data, aes(x = as.factor(Conf_SCI), y = prop, fill = trust)) + geom_bar(stat = "identity", position = "dodge") + labs(title = "Vaccination Rates by Trust and Confidence in Science", x = "Confidence in Science", y = "Proportion Vaccinated", fill = "Trust in Government")
```

Problem 3:

```
filtered_data_2 <- data_problem_1 %>% select(COVID_vaccinated, F_ATTEND, F_EDUCCAT, F_PARTY_FINAL, F_RACETHNMOD) %>% drop_na()

education_rates <- filtered_data_2 %>% group_by(F_EDUCCAT, COVID_vaccinated) %>% summarise(count = n()) %>% mutate(prop = count / sum(count))

#1 = college grad, 2 = some college, 3 = hs grad or less

religious_attendance_rates <- filtered_data_2 %>% group_by(F_ATTEND, COVID_vaccinated) %>% summarise(count = n()) %>% mutate(prop = count / sum(count))

#6 = never attends --> #1 = more than once a week
party_vaccination <- filtered_data_2 %>% group_by(F_PARTY_FINAL, COVID_vaccinated) %>% summarise(count = n()) %>% mutate(prop = count / sum(count))

#1 = republican, #2 = democrat, #3 = independent, #4 = other

race_vaccination <- filtered_data_2 %>% group_by(F_RACETHNMOD, COVID_vaccinated) %>% summarise(count = n()) %>% mutate(prop = count / sum(count))

#1 = white, #2 = black, #3 = hispanic, #4 = other, #5 = asian

ggplot(education_rates, aes(x = as.factor(F_EDUCCAT), y = prop, fill = COVID_vaccinated)) + geom_bar(stat = "identity", position = "stack") + labs(title = "Distribution of Education Levels", x = "Education Level", y = "Count")

ggplot(religious_attendance_rates, aes(x = as.factor(F_ATTEND), y = prop, fill = COVID_vaccinated)) + geom_bar(stat = "identity", position = "stack") + labs(title = "Distribution of Religion Attendance", x = "Religion", y = "Count")
```

```
ggplot(party_vaccination, aes(x = as.factor(F_PARTY_FINAL), y = prop, fill =
COVID_vaccinated)) + geom_bar(stat = "identity", position = "stack") + labs(title =
"Distribution of Parties", x = "Party", y = "Count")

ggplot(race_vaccination, aes(x = as.factor(F_RACETHNMOD), y = prop, fill =
COVID_vaccinated)) + geom_bar(stat = "identity", position = "stack") + labs(title =
"Distribution of Diff Races", x = "Party", y = "Count")
```

Problem 4:

```
#First, get rid of 98 and 99 values

data <- ATP #establish data as the call
data[data == 98 | data == 99] <- NA

data$QKEY <- NULL
data$INTERVIEW_START <- NULL
data$INTERVIEW_END <- NULL
data$DEVICE_TYPE <- NULL
data$FORM <- NULL
data$WEIGHT_W114 <- NULL
data$WEIGHT_W64_W66_W83_W114 <- NULL
data$WEIGHT_W84_W114 <- NULL
```

Problem 5:

```
data$COVID_vaccinated <- ifelse(data$COVID_vaccinated == "Y", 1,
                        ifelse(data$COVID_vaccinated == "N", 0, NA))

data <- data %>%
  mutate(across(c(COVID_vaccinated,F_ATTEND
, F_AGECAT, F_GENDER, F_EDUCCAT, F_INC_TIER2, , F_PARTY_FINAL            ,
F_RACETHNMOD), as.factor))

data$COVID_vaccinated <- relevel(data$COVID_vaccinated, ref = "0") #No as reference for
vaccinated
data$F_AGECAT <- relevel(data$F_AGECAT, ref = "4")        # Older as reference
data$F_EDUCCAT <- relevel(data$F_EDUCCAT, ref = "3")      # dint go to college
data$F_PARTY_FINAL           <- relevel(data$F_PARTY_FINAL     , ref = "1")# "republican" as
reference for party
data$F_ATTEND <- relevel(data$F_ATTEND, ref = "1") # reference is never attends
data$F_INC_TIER2 <- relevel(data$F_INC_TIER2, ref = "1")#reference is low
data$F_RACETHNMOD <- relevel(data$F_RACETHNMOD, ref = "2")# reference is white
```

```
data_model <- glm(formula = data$COVID_vaccinated ~ data$F_AGECAT +
data$F_EDUCCAT + data$F_PARTY_FINAL + data$F_ATTEND+ data$F_INC_TIER2 +
data$F_RACETHNMOD,
   family = binomial(link = "logit"), data = data)


summary(data_model)

pchisq(deviance(data_model), df.residual(data_model), lower = FALSE)
exp(confint(data_model))
plot(fitted(data_model), residuals(data_model),
    xlab = "Fitted Values",
    ylab = "Residuals",
    main = "Residuals vs Fitted",
    pch = 20, col = "blue")
abline(h = 0, col = "red", lty = 2)

pearson_residuals <- residuals(data_model, type = "pearson")
overdispersion <- sum(pearson_residuals^2) / df.residual(data_model)
print(overdispersion)


data$predicted <- predict(data_model,newdata = data, type = "response")


# Actual values (binary outcome)
data$actual <- data$COVID_vaccinated


roc_curve <- roc(data$COVID_vaccinated, data$predicted)
auc_value <- auc(roc_curve)
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)

text(0.6, 0.4, paste("AUC =", round(auc_value, 3)), col = "blue")
print(paste("AUC:", round(auc_value, 3)))
```