Robertson and Cao Project

Part 1:Teenage HPV Vaccination Coverage and Socioeconomic Factors

Kelly Cao and Rachel Robertson

2024-04-18

# 1. Introduction

## 1.1 Background

Human Papillomavirus (HPV) is a double-stranded DNA virus that infects the cells of the skin and mucosal linings [1]. HPV is the most common sexually transmitted infection (STI) in the United States, where over 90% of sexually active males and 80% of sexually active females will likely become infected within their lifetime [2, 3]. Most cases of HPV are asymptomatic and cleared over time, but certain infections tend to become persistent and can cause genital warts. Specific types of HPV are known to be “high-risk” or increase the likelihood of developing cancer in persistently infected cells [4]. The HPV group contains over 200 related viruses, but only 12 types are considered “high-risk” due to the presence of oncogenes. These include HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59, though HPV-16 and HPV-18 are associated with the highest number of HPV-associated cancer cases. In the U.S., there are 37,300 cases of HPV-associated cancer cases every year, including cases of cervical, anal, oropharyngeal (throat and mouth), vulvar, vaginal, and penile cancers [5,6]. In an effort to prevent HPV-associated cancers worldwide, two HPV vaccines were developed and made available in 2006. These included a bivalent and quadrivalent vaccine that contained viral-like particles (VLPs) of the viral capsid protein (L1) from “high-risk” HPV types [7]. In recent years, the vaccine used in the U.S., Gardasil 9, has been expanded to contain HPV types 6, 11, 16, 18, 31, 33, 45, 52, and 58 [4]. Currently, the United States Centers for Disease Control and Prevention (CDC) recommends a two-dose Gardasil 9 regime to all teenagers, with a primary dose at age 11 to 12 and a second dose within 6 to 12 months of the first [8]. In 2022, the CDC estimated that 76.0% of teenagers aged 13 to 17 had received at least 1 dose of the HPV vaccine while only 62.6% of teenagers were up-to-date with their vaccination regime [9].

## 1.2 Data

The data we chose to use is the National Immunization Survey (NIS) of 2022. It consists survey data that was collected to monitor vaccination coverage for teenagers age 13-17 in the United States of America. The surveys were conducted by the National Center for Immunization and Respiratory Diseases of the Center (NCIRD) for Disease Control and Prevention (CDC). The survey itself consists of two parts: (1) the household telephone (random digit dialing) survey answered by a parent or guardian and (2) a mailed survey for the vaccination provider, called “Immunization History Questionnaire.” The original data set is a fixed width file (FWF). In the data set, there are 43,738 rows and 672 columns, with each row representing an individual and each column representing the answer to a question. The questions pertain to the teenager’s immunization history, demographics, and additional household-reported health information. We have truncated the data set to include 26-selected columns for the ease of observation. We have also filtered the data set to include one year, 2022, which is the most recent year of data that is published. The provided survey data will require cleaning, as there are signs of human errors present.

## 1.3 Research Question

Data from the 2022 National Immunization Survey for teens will be used to identify which socioeconomic and geographic factors are associated with HPV vaccination completion for teenagers aged 13 to 17 in the U.S. Specifically, we want to find the most important socioeconomic and geographic factors for a teenager being up-to-date for their HPV vaccinations in the U.S. The outcome of this study will be measured using the overall up-to-date HPV vaccination completion indicator, which includes all age groups in any step of the vaccination process. The variable includes teenagers aged 13 to 17 who are up-to-date and have received at least one HPV shot.

## 1.4 Purpose

Measuring HPV vaccine completion with “up-to-date” variables may be more robust than vaccination rate based on those who have completed the regime (2 or more shots), because not all teens included in the survey are at the age in which they would’ve completed the full regime. This means that younger teens who are up to date with 1+ shot will not be excluded from the analysis. We aim to measure determinants of socioeconomic status that are engrained into the survey questions. These factors include: family income, poverty status, income-to-poverty ratio, insurance status, insurance breaks, maternal education, living arrangement, and geographic mobility status. Additional demographic factors that are associated with healthcare access and may be examined include: race, ethnicity, language, facility in which the vaccine was administered, and whether the teen had completed a wellness exam between the ages of 11-12. Geographic distribution of the teens will be assessed by true state of residence. We would like to examine the data for correlations between HPV vaccination completion status and the aforementioned socioeconomic determinants, demographics, and geographic location. This serves as valuable information to determine differences in healthcare access and vaccination coverage for teenagers living in the U.S. It is generally known that those of lower income and without insurance have limited healthcare access in the U.S. [10], but it will be interesting to see if this pattern is reflected at the state level. If so, policy should be examined as it may be related to healthcare access disparities. Recent studies have examined HPV vaccination completion according to social determinants, but they have not examined geographic distribution and include data prior to 2019 [11]. More recent years display differences in global HPV vaccination trends due to the COVID19 pandemic [12]. Our study will fill in geographical gaps and provide updated vaccination trends based on socioeconomic factors.

# 2. Methods

## 2.1 Loading Data

The data was loaded from the original DAT file, entitled ‘NISTEENPUF22.DAT’ by specifying the column positions within the data file. The columns were then read and made into a data frame. Each of the factor variables was defined and the corresponding levels were assigned labels. The structure and summary statistics of all of the variables were examined. The loaded data is saved as a rds and a csv in the ‘processed-data’ folder, as to not alter the original raw data file. The definition of each factor variable in included in the README file within the ‘raw-data’ folder.

## 2.2 Cleaning Data

The structure of the processed data was examined. The data was examined for missing values. No NAs were present, however missing values were present as labeled factor levels. The missing factor levels were printed, identified, and replaced with NA values. A string of 27,000 NAs were present in the outcome variable, indicating vaccination completion. All rows corresponding to the NAs for the missing outcome value were removed. Approximately 160 additional rows with NA values for facility or insurance status variables were removed. Over 16,000 rows of data remained. The year column was eliminated, as all of the data is from 2022. We saved the cleaned data as an rds and csv in the ‘processed-data’ folder. ## Exploratory Analysis Exploratory analysis was used to determine the distribution of the predictors, including the socioeconomic variables that were chosen for analysis. The distribution of single predictors including income-to-poverty ratio, income, state, race, ethnicity, vaccination status, and facility were examined using histograms and scatter plots. The following predictors were compared visually and fitted to a linear regression to examine preexisting associations: vaccination status and ethnicity, vaccination status and poverty status, income poverty ratio and insurance status, income poverty ratio and ethnicity. ## Analysis ### Generalized Logistic Regression A generalized logistic regression was performed with bi-directional step-wise selection for predictor reduction. Variables that were redundant of a major variable were excluded from analysis. The variables initially included were age, sex, state, race/ethnicity, insurance status, income, insurance breaks, poverty status, education, language, geographic mobility, facility, and housing status. The model was evaluated with a residual plot and correlation matrix to reveal strong multicollinearity. This is the basis for the subsequent analysis methods, as they are more resistant to collinearity. ### Principal Component Analysis Principal component analysis (PCA) was performed to do dimensionality reduction. PCA would ideally transform the highly correlated predictors into uncorrelated components. The h2o package was used to provide a machine-learning platform in RStudio, to allow machine-learning methods ot be implemented. Due to the data being mainly categorical variables, the PCA proved to be unsuccessful. A LASSO Ridge Regression was performed with the h2o package, to mitigate the multicollinearity issue. ### Random Forest Model A random forest model was used, as it is resistant to multicollinearity. The tidymodels package was used to produce a random forest model with the classification method. The model was trained and fit from training data, which was an 80% split of the original data. After evaluation, the model parameters for minimal node size and number of selected predictors were tuned for using a grid search with cross-validation resamples. After tuning, the best model was extracted based on the ROC AUC metric and was fit to the training data. The tuned model was used to make predictions and evaluated with performance metrics, a confusion matrix, and a comparison to a single-predictor random forest model. Finally, the tuned random forest model was re-evaluated by being fit to the testing data. ### LASSO Regression

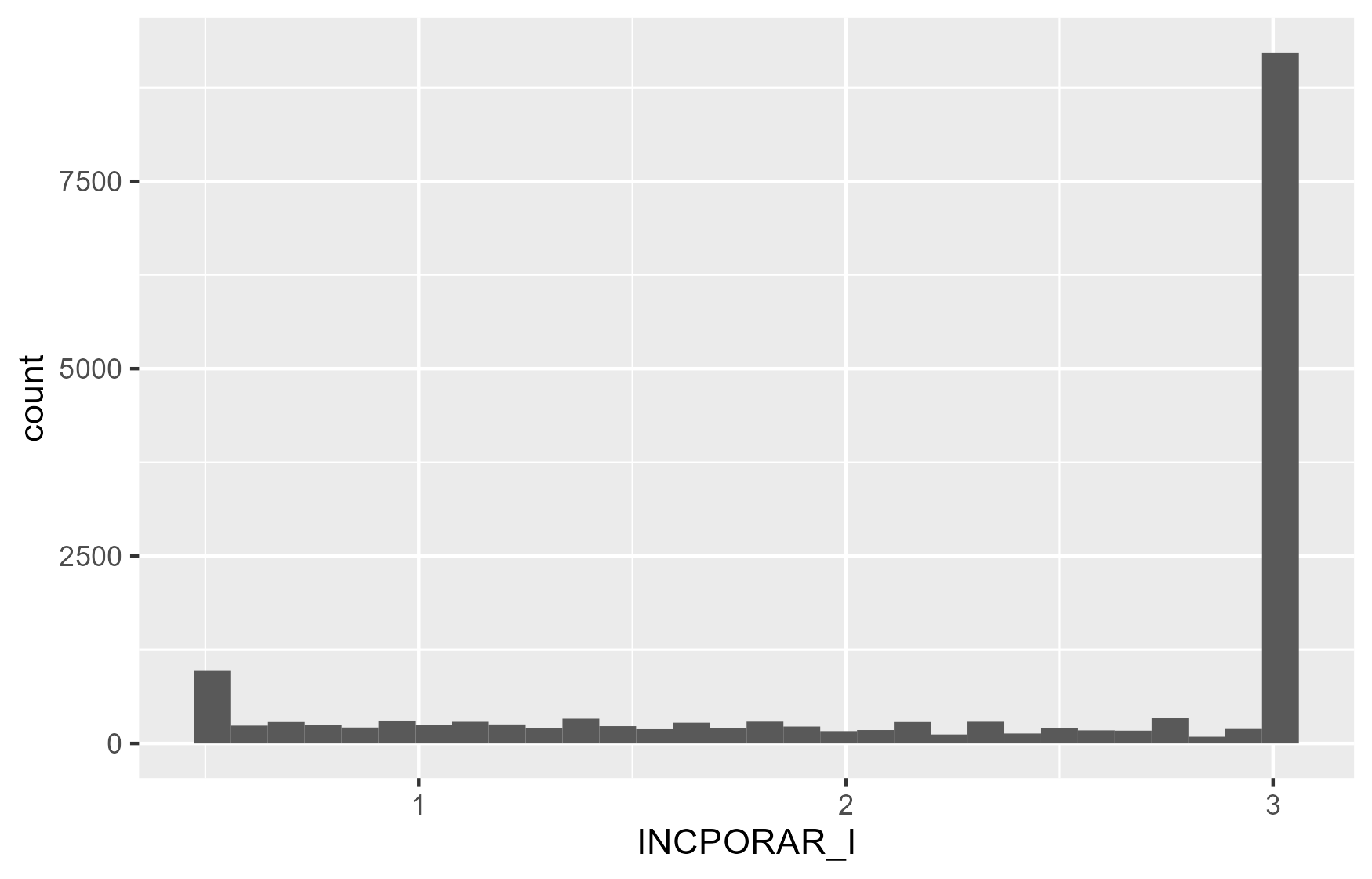
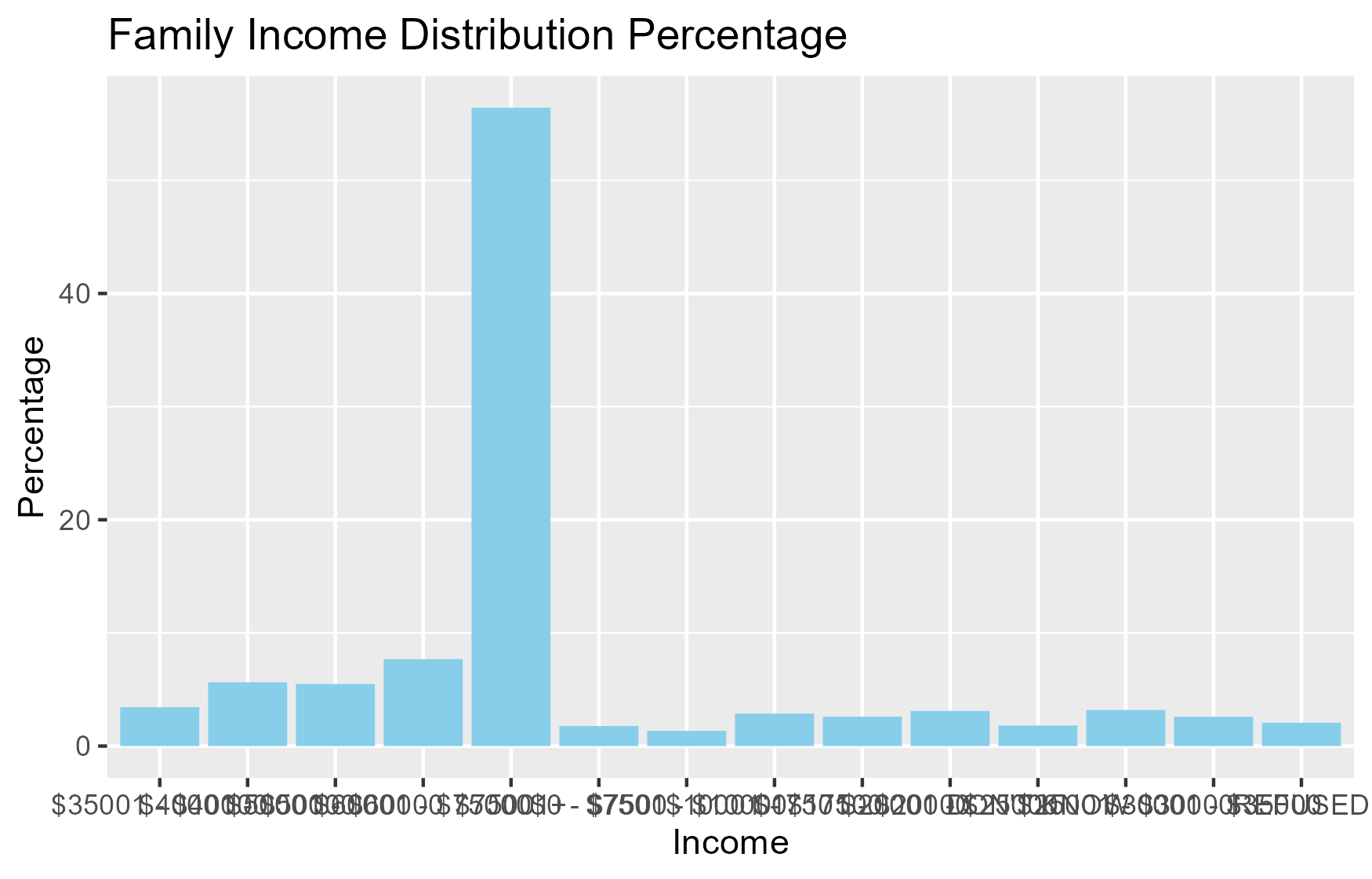
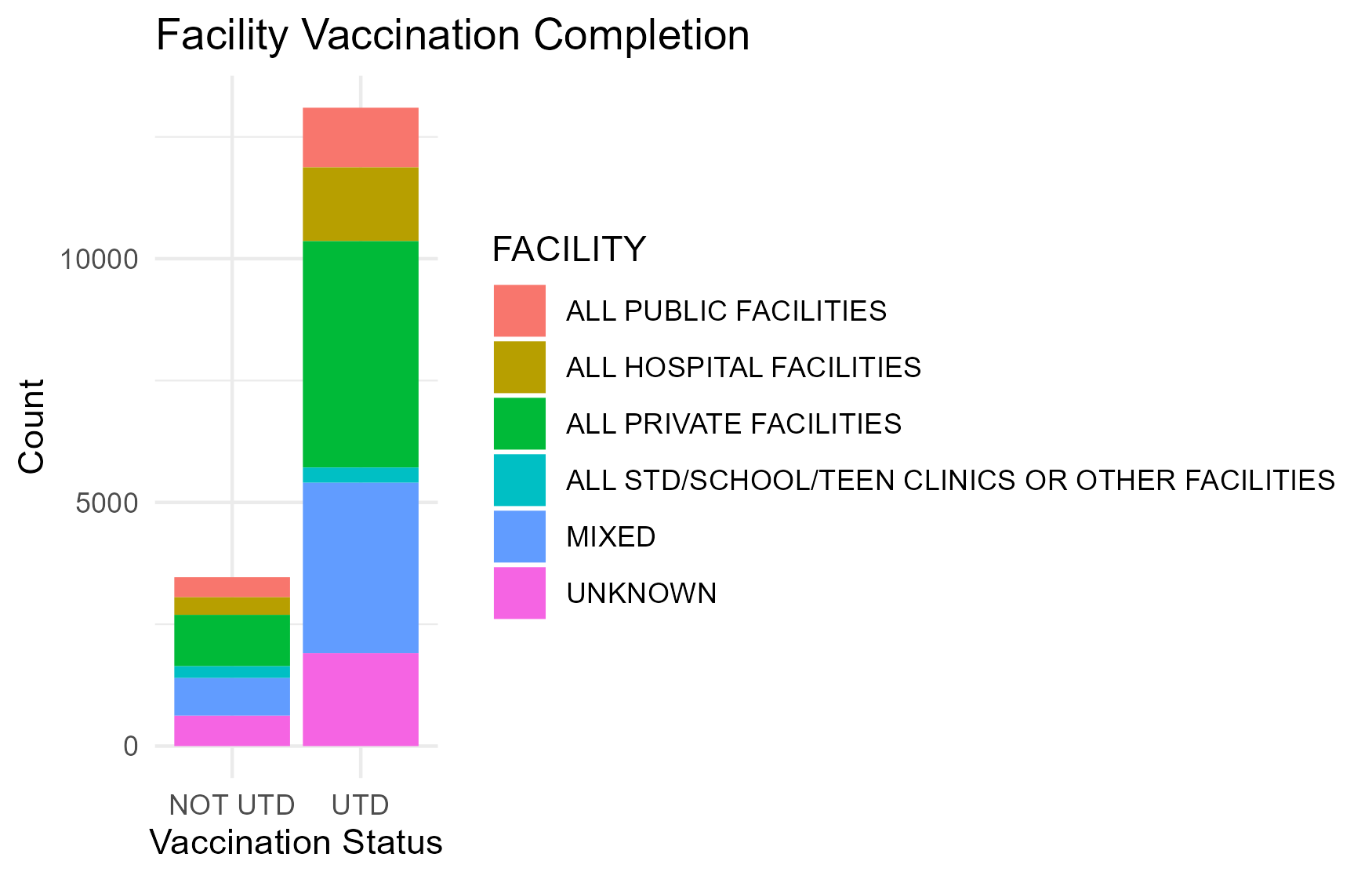
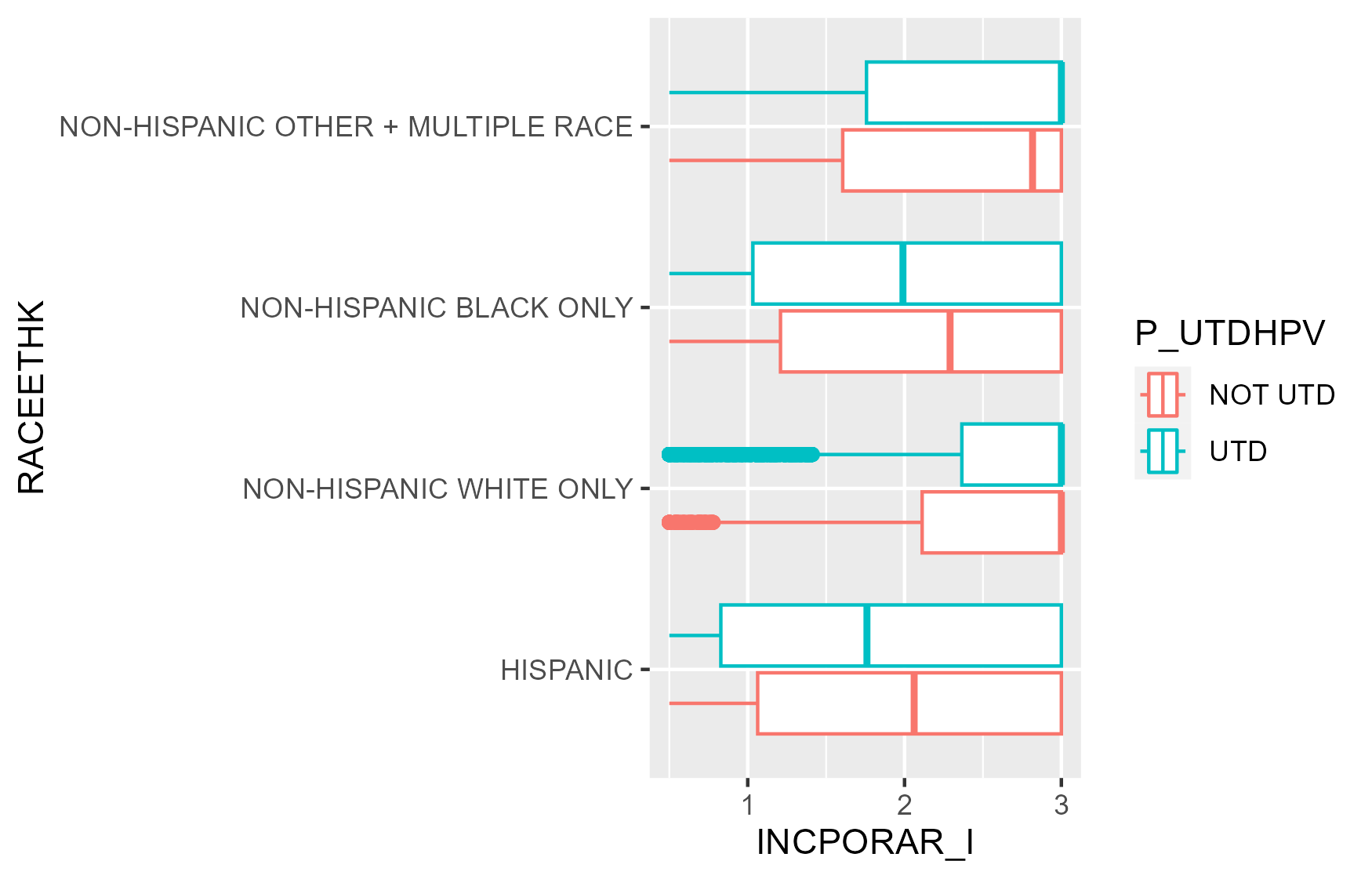
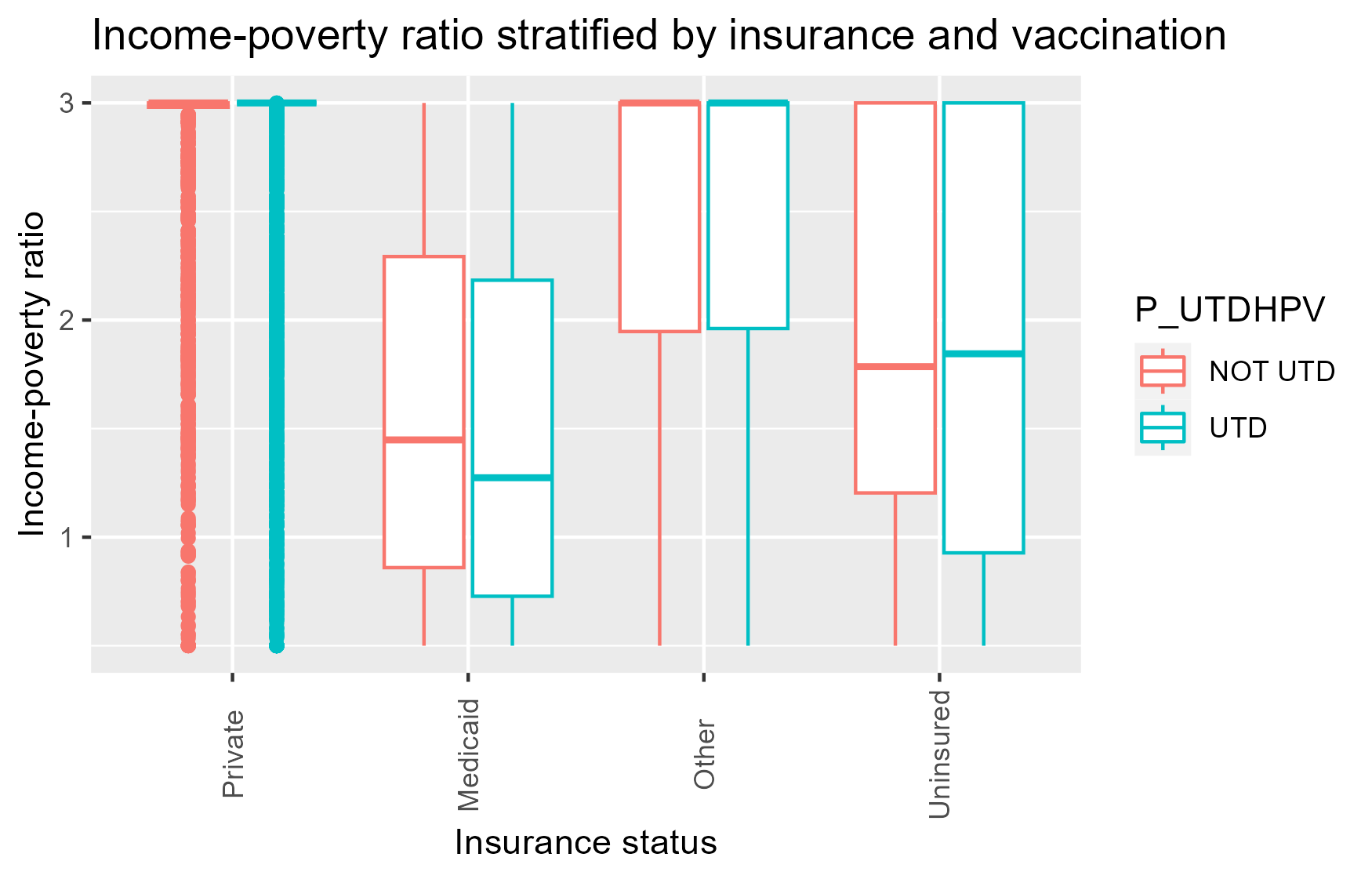
# 3. Results

## 3.1 Exploratory Data Analysis

Exploratory analysis is used to determine the distribution of our predictors, including the socioeconomic variables that we chose for analysis. Nearly all of the demographic factors have a similar distribution between those who have completed their HPV vaccination regime versus those who have not completed it.

Table 1.Socioeconomic characteristics of U.S. teenagers who have completed or are up-to-date with HPV vaccination series.

| **Characteristic** | **NOT UTD**, N = 3,465 | **UTD**, N = 13,099 | **Missing Data**, N = 0 |
| --- | --- | --- | --- |
| Age (years) | NA | NA | NA |
| 13 | 845 (24%) | 2,448 (19%) | 0 (NA%) |
| 14 | 753 (22%) | 2,757 (21%) | 0 (NA%) |
| 15 | 657 (19%) | 2,667 (20%) | 0 (NA%) |
| 16 | 603 (17%) | 2,707 (21%) | 0 (NA%) |
| 17 | 607 (18%) | 2,520 (19%) | 0 (NA%) |
| Sex | NA | NA | NA |
| MALE | 1,945 (56%) | 6,741 (51%) | 0 (NA%) |
| FEMALE | 1,520 (44%) | 6,358 (49%) | 0 (NA%) |
| Insurance Status | NA | NA | NA |
| PRIVATE INSURANCE ONLY | 2,020 (58%) | 7,682 (59%) | 0 (NA%) |
| ANY MEDICAID | 1,025 (30%) | 4,205 (32%) | 0 (NA%) |
| OTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE) | 305 (8.8%) | 968 (7.4%) | 0 (NA%) |
| UNINSURED | 115 (3.3%) | 244 (1.9%) | 0 (NA%) |
| Poverty Status | NA | NA | NA |
| ABOVE POVERTY > $75K | 1,930 (56%) | 7,415 (57%) | 0 (NA%) |
| ABOVE POVERTY <= $75K | 1,063 (31%) | 3,523 (27%) | 0 (NA%) |
| BELOW POVERTY | 370 (11%) | 1,814 (14%) | 0 (NA%) |
| UNKNOWN | 102 (2.9%) | 347 (2.6%) | 0 (NA%) |
| Race and Ethnicity | NA | NA | NA |
| HISPANIC | 575 (17%) | 2,728 (21%) | 0 (NA%) |
| NON-HISPANIC WHITE ONLY | 2,224 (64%) | 7,514 (57%) | 0 (NA%) |
| NON-HISPANIC BLACK ONLY | 298 (8.6%) | 1,218 (9.3%) | 0 (NA%) |
| NON-HISPANIC OTHER + MULTIPLE RACE | 368 (11%) | 1,639 (13%) | 0 (NA%) |
| Maternal Education-level | NA | NA | NA |
| LESS THAN 12 YEARS | 165 (4.8%) | 782 (6.0%) | 0 (NA%) |
| 12 YEARS | 508 (15%) | 1,943 (15%) | 0 (NA%) |
| MORE THAN 12 YEARS, NON-COLLEGE GRAD | 1,047 (30%) | 3,095 (24%) | 0 (NA%) |
| COLLEGE GRADUATE | 1,745 (50%) | 7,279 (56%) | 0 (NA%) |
| Survey Language | NA | NA | NA |
| ENGLISH | 3,297 (95%) | 11,993 (92%) | 0 (NA%) |
| SPANISH | 148 (4.3%) | 1,019 (7.8%) | 0 (NA%) |
| OTHER | 20 (0.6%) | 87 (0.7%) | 0 (NA%) |
| Housing Status | NA | NA | NA |
| OWNED OR BEING BOUGHT | 2,622 (76%) | 9,810 (75%) | 0 (NA%) |
| RENTED | 737 (21%) | 2,951 (23%) | 0 (NA%) |
| OTHER ARRANGMENT | 98 (2.8%) | 312 (2.4%) | 0 (NA%) |
| DON’T KNOW | 4 (0.1%) | 8 (<0.1%) | 0 (NA%) |
| REFUSED | 4 (0.1%) | 18 (0.1%) | 0 (NA%) |

Income-poverty-ratio is a numeric value that indicates how close to poverty a family is living. It is calculated by income over the poverty income threshold and ranges from values 1 (closest to poverty) and 3 (furthest from poverty). The distribution of the income poverty ratio indicates a strong left skew. The highest frequency of observations was gathered from households with an income poverty ratio, meaning most survey participants live well above the poverty threshold. The family income is right skewed with the majority of the observations coming from $75000+ income level. This indicates that a high frequency of participants are above the poverty income threshold, but also in a modest income class.   The state variable shows a more uniform distribution, suggesting that the observations were gathered fairly across the US states and territories. States and territories with a significantly larger number of observations include Texas, West Virginia, Pennsyvania, Puerto Rico, New York, and Illinois. Although it is noted a few territories provide 0 observations. The race and ethnicity variables show a distribution that is similar to the U.S. census data. The census data states that the racial composition of the U.S. is 58.9% white, 13.6% African American, and 19.1% Hispanic or Latino, while our survey population is composed of 58.% white, 9.15% African American, and 19.9% Hispanic or Latino participants. The majority of observations came from households with private insurance only The Insurance status variable shows that the majority of the observations came from households with private insurance only (58.6%), followed by Medicaid (31.6%), other insurance (7.7%), and uninsured (2.2%). This is mirrored in the facility variable, as survey location, where a majority of its observations show private facilities (34.4%), followed by mixed facilities (25.8%), hospitals (11.3%), and all public facilities (9.9%). 15.2% of the facility types were unknown. Upon stratification of facility type by HPV vaccination completion status, it was found that the majority of participants who were up-to-date were surveyed at private or mixed facilities. They likely received their HPV vaccination there as well.  A total of 79% of participants had up-to-date status for receiving 1 or more shots of HPV vaccination. U.S. states and territories had an HPV vaccination completion rate ranging from 69% to 95%, fluctuating around ~80%. There are different counts of each race/ethnicity for each factor level, so the percentage of those with up-to-date vaccine status out of each race/ethnicity level was adjusted. The Hispanic ethnicity was found to have the highest vaccine completion while the non-Hispanic white ethnicity had the lowest. It was also found that for Hispanic and Non-Hispanic black ethnicities, the income poverty ratio was higher for those who were not up-to-date in HPV vaccination status, compared to those who were. The opposite was true for mixed ethnic groups, where vaccine completion corresponded with a higher mean income-poverty ratio. When examining the overall HPV vaccination status stratified by income poverty ratio, the income poverty ratio did not differ between those who are up to date and those who are not. This may be accounted for by other trends such as that shown with ethnicity, above.  Subsequently, income poverty ratio and insurance status were examined, and stratified by HPV vaccination status. It is apparent that those with private and other insurance status had the highest income-poverty ratio, while those with Medicaid or uninsured were closer to the poverty income threshold. For those with Medicaid, those who were up-to-date in HPV vaccinations had a slightly lower income-poverty ratio than those who were not up to date.  ## Analysis We begin with various socioeconomic predictors that interest us. Some of these predictors are a subset of the other, or may be used to describe an overall category.

The variables used were chosen based on their interpretability and applicability to our research question “Which socioeconomic and geographic factors impact HPV vaccine completion rates among teenagers?” . There are seven potential outcome variables for HPV vaccine completion:

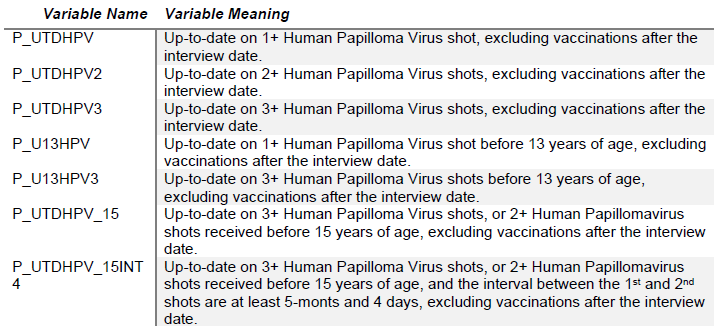


Table 3. Human Papillomavirus vaccine status variables and meanings.

The first option, P\_UTDHPV was the chosen outcome as it is the most general. It has the benefit of being able to use Age as a predictor variable and also not biasing the model towards older ages who are required to receive more than one shot to be considered up-to-date. Additionally, in April of 2022, the SAGE work group of the WHO officially recommended a single-dose HPV vaccine regime, because vaccine efficacy was found to be significantly high with a single dose. This recommendation was given with the goal of increasing vaccine coverage in areas with lower vaccine accessibility and availability (13). In September of 2023, PAHO followed suit in recommending a single-dose HPV vaccine regime for the Americas (14).

The predictor variables we chose to examine are listed below:

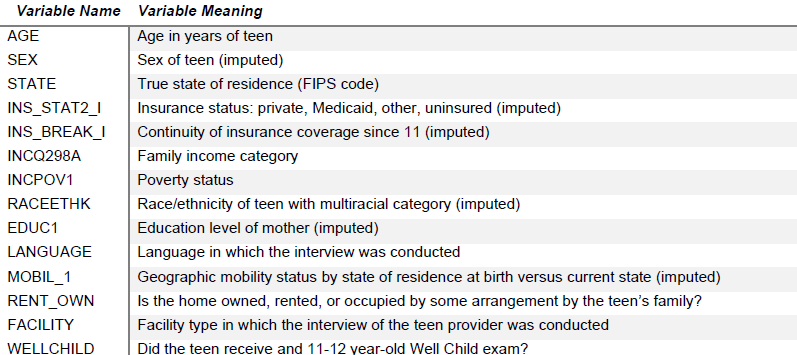


Table 4. Potential predictor variable names and meanings.

INCOPAR or income-poverty ratio was excluded because this is reflected by the income and poverty status variables, separately. Additionally, this is a numeric variable while the others are categorical. RACE\_K or race was excluded because race is included within the RACEETHK variable. RACEETHK is ideal because it includes both race and ethnicity.

To determine which predictor variables will be best suited for a general linear model, we employed a stepwise comparison is performed and then assessed with AIC and WAIC. The stepwise model is set up with P\_UTDPHPV as the outcome and the remaining chosen variables as the predictors. Each predictors were added and dropped in different variations, and the resulting AIC were recorded for each of the different models. From this, we were able to determine that the model with the lowest AIC (and thus the greatest fit) was the model with all of the predictors that we previously picked out.

Moving from that, we plotted the residuals from the logistic regression model to check for non-linearity and discovered that the predictors potentially had a high indication of collinearity. to further confirm this, a VIF analysis was attempted but was unable to proceed as the multicollinearity was too great for too many variables. As a result, we decided to perform a correlation matrix, which performed as expected and verified our concerns.

INCLUDE EITHER REGRESSION PLOT OR STEPWISE FINAL RESULT IMAGE

In order to reduce the multicollinearity issue, we initially attempted to use PCA. To employ the unsupervised learning model, we use the h2o package. The PCA resulted in one principle component with a high standard deviation and high proportion of variance, showing that the Pc1 captures a large portion (97.9%) of variability in the data.

INCLUDE IMAGE OF PCA RESULTS

The result of having only one principle component further supports the fact that many of the predictor variables are highly collinear. One PC shows that a PCA must reduce the data object to a single dimension. Having a single principal component could be problematic for my analysis, particularly if it leads to the loss of crucial information or structure in my data due to collinearity among the original features.

Using the same package, I used the LASSO logistical regression model and performed a co-validation test on it.

MSE: 0.1558815  
RMSE: 0.3948183  
LogLoss: 0.4850906  
Mean Per-Class Error: 0.4906615  
AUC: 0.6596287  
AUCPR: 0.8704217  
Gini: 0.3192575  
R^2: 0.05771195  
Residual Deviance: 16070.08  
AIC: 16266.08

Random Forest

LASSO Regression # Discussion

# 4. Conclusion

# 5. References

1. https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/vaccine-standardization/human-papillomavirus#:~:text=Human%20papillomavirus%20(HPV)%20is%20a,approximately%208%2Dkb%20in%20length.
2. https://www.womenshealth.gov/a-z-topics/human-papillomavirus#:~:text=Genital%20HPV%20is%20the%20most,79%20million%20Americans%20have%20HPV.&text=It%20is%20so%20common%20that,some%20point%20in%20their%20lifetime.
3. https://www.health.state.mn.us/data/mcrs/data/qfhpv.html#:~:text=Human%20Papillomavirus%20(HPV)%20is%20a,with%20HPV%20in%20their%20lifetime.&text=Around%2050%20percent%20of%20HPV,HPV%2C%20which%20can%20cause%20cancer.
4. HPV and Cancer. In: National Cancer Institute [Internet]. 1 Mar 2019 [cited 12 Mar 2024]. Available: https://www.cancer.gov/about-cancer/causes-prevention/risk/infectious-agents/hpv-and-cancer
5. https://www.cancer.gov/about-cancer/causes-prevention/risk/infectious-agents/hpv-and-cancer
6. https://www.cdc.gov/vaccines/vpd/hpv/hcp/vaccines.html#:~:text=Vaccine%20Composition,45%2C%2052%2C%20and%2058.
7. https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/vaccine-standardization/human-papillomavirus#:~:text=Human%20papillomavirus%20(HPV)%20is%20a,approximately%208%2Dkb%20in%20length.
8. https://www.cdc.gov/vaccines/vpd/hpv/hcp/recommendations.html#:~:text=Two%20doses%20of%20HPV%20vaccine,third%20dose%20of%20HPV%20vaccine.
9. https://www.cdc.gov/mmwr/volumes/72/wr/mm7234a3.htm#:~:text=In%202022%2C%2089.9%25%20of%20adolescents,HPV%20vaccination%20(HPV%20UTD).
10. Sciences, N. A. of, Engineering, & Medicine, and. (2018, March 1). Factors that affect health-care utilization. Health-Care Utilization as a Proxy in Disability Determination. https://www.ncbi.nlm.nih.gov/books/NBK500097/
11. Mansfield, L. N., Chung, R. J., Silva, S. G., Merwin, E. I., & Gonzalez-Guarda, R. M. (2022). Social determinants of human papillomavirus vaccine series completion among U.S. adolescents: A mixed-methods study. SSM - population health, 18, 101082. https://doi.org/10.1016/j.ssmph.2022.101082
12. Casey, R. M., Akaba, H., Hyde, T. B, et al. (2024). Covid-19 pandemic and equity of global human papillomavirus vaccination: descriptive study of World Health Organization-Unicef vaccination coverage estimates. BMJ Medicine, 3, 000726. doi: 10.1136/bmjmed-2023-000726
13. https://www.who.int/news/item/20-12-2022-WHO-updates-recommendations-on-HPV-vaccination-schedule
14. https://www.paho.org/en/news/5-9-2023-paho-technical-advisory-group-recommends-countries-americas-use-single-dose-hpv