Robertson and Cao Project

Part 1:Teenage HPV Vaccination Coverage and Socioeconomic Factors

Kelly Cao and Rachel Robertson

2024-02-22

# 1. Introduction

## 1.1 Data Set

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.

## 1.2 Research Question

We want to use the 2022 National Immunization Survey for teens to identify which socioeconomic and geographic factors are associated with HPV vaccination completion for teenagers in the U.S. Specifically, we are asking the question, What is the likelihood of a teenager being up-to-date for their HPV vaccinations in the U.S., based on socioeconomic factors and geographic distribution. The outcome of this study will be measured using the columns that pertain to HPV vaccination completion. These variables are provider-collected, and they state whether or not the teen is (1) up-to-date with 1+ HPV shot, (2) up-to-date with 2+ HPV shots, or (3) up-to-date with 3+ HPV shots (excluding all vaccinations post-survey). 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.

## 1.3 Proposed Analysis

For the analysis portion, we have decided to do a correlation analysis between HPV vaccination completion and each of the socioeconomic factors. When performing this, we may choose to stratify by age or gender, if we find these to be confounders. The correlation analysis cannot be performed using a multivariate model because there may be collinearity present between many of the determinants for socioeconomic status. This will lead to spurious results. Instead, the correlation between vaccination status and each factor is examined separately.   
Next, we may stratify this by state. Comparing these correlation values between states and regions may lead us to find patterns of healthcare accessibility in the U.S. It is generally known that those of lower income and without insurance have limited healthcare access in the U.S.(1), 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 (2). More recent years display differences in global HPV vaccination trends due to the COVID19 pandemic (3). Our study will fill in geographical gaps and provide updated vaccination trends based on socioeconomic factors.

## 1.4 Methods

### 1.4.1 Data Cleaning and Processing

#### 1.4.1.1 Loading the 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. Next, each of the factor variables were 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. This does not alter the original raw data file. The definition of each factor variable in included in the README file within the ‘raw-data’ folder.

Next, we explored the structure of the processed data and searched for any missing values. However, there were no NAs present. There are, however, missing values as labelled factor levels which must be accounted for. We identified the missing value labels by printing the factor levels. We then replaced the all of the values labelled missing with NA values. Next, we found a string of over 27000 NAs all of the variables for vaccination completion. These rows corresponded for all variables related to vaccine completion. Since this is the response variable of our scientific question, we eliminated these rows. We also eliminated rows with small numbers missing values, which were within the columns for facility, Wellchild exams, and insurance breaks. Lastly, we eliminatethe year column, as all of the data comes from the year 2022. We saved the cleaned data as an rds and csv in the ‘processed-data’ folder.

### 1.4.2 Exploratory Analysis

First we will load the packages needed for the exploratory analysis.

#load needed packages. make sure they are installed.  
library(here) #for data loading/saving  
library(dplyr)  
library(skimr)  
library(ggplot2)  
library(knitr)

Next, we will load the data called ‘cleandata1’ from the ‘processed-data’ folder.

#Path to data. Note the use of the here() package and not absolute paths  
data\_location <- here::here("data","processed-data","cleandata1.rds")  
#load data  
mydata <- readRDS(data\_location)

The following data set has been cleaned. Any observations with NAs has been removed.

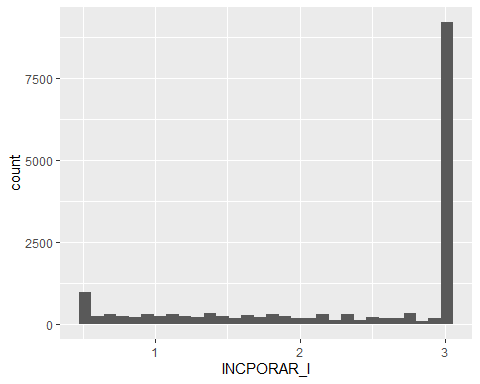
summary\_df = skimr::skim(mydata)  
print(summary\_df)

── Data Summary ────────────────────────  
 Values  
Name mydata  
Number of rows 16564   
Number of columns 23   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Column type frequency:   
 factor 20   
 numeric 3   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Group variables None   
  
── Variable type: factor ───────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate ordered n\_unique  
 1 EDUC1 0 1 FALSE 4  
 2 INCPOV1 0 1 FALSE 4  
 3 INCQ298A 0 1 FALSE 14  
 4 LANGUAGE 0 1 FALSE 3  
 5 MOBIL\_1 0 1 FALSE 2  
 6 RACEETHK 0 1 FALSE 4  
 7 RACE\_K 0 1 FALSE 3  
 8 RENT\_OWN 0 1 FALSE 5  
 9 SEX 0 1 FALSE 2  
10 STATE 0 1 FALSE 52  
11 FACILITY 0 1 FALSE 6  
12 P\_U13HPV 0 1 FALSE 2  
13 P\_U13HPV3 0 1 FALSE 2  
14 P\_UTDHPV 0 1 FALSE 2  
15 P\_UTDHPV\_15 0 1 FALSE 2  
16 P\_UTDHPV\_15INT 0 1 FALSE 2  
17 P\_UTDHPV2 0 1 FALSE 2  
18 P\_UTDHPV3 0 1 FALSE 2  
19 INS\_STAT2\_I 0 1 FALSE 4  
20 INS\_BREAK\_I 0 1 FALSE 4  
 top\_counts   
 1 COL: 9024, MOR: 4142, 12 : 2451, LES: 947   
 2 ABO: 9345, ABO: 4586, BEL: 2184, UNK: 449   
 3 $75: 9345, $60: 1273, $40: 934, $50: 908   
 4 ENG: 15290, SPA: 1167, OTH: 107   
 5 DID: 12849, MOV: 3715, DON: 0, MIS: 0   
 6 NON: 9738, HIS: 3303, NON: 2007, NON: 1516  
 7 WHI: 12378, OTH: 2379, BLA: 1807   
 8 OWN: 12432, REN: 3688, OTH: 410, REF: 22   
 9 MAL: 8686, FEM: 7878, DON: 0, MIS: 0   
10 TEX: 784, PEN: 750, PUE: 668, ILL: 623   
11 ALL: 5700, MIX: 4271, UNK: 2528, ALL: 1874  
12 UTD: 10874, NOT: 5690, Mis: 0   
13 NOT: 15838, UTD: 726, Mis: 0   
14 UTD: 13099, NOT: 3465, Mis: 0   
15 UTD: 10838, NOT: 5726, Mis: 0   
16 UTD: 10747, NOT: 5817, Mis: 0   
17 UTD: 10966, NOT: 5598, Mis: 0   
18 NOT: 15510, UTD: 1054, Mis: 0   
19 PRI: 9702, ANY: 5230, OTH: 1273, UNI: 359   
20 CUR: 15500, CUR: 705, CUR: 254, CUR: 105   
  
── Variable type: numeric ──────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate mean sd p0 p25 p50  
1 SEQNUMT 0 1 21670. 12635. 1 10652. 21614.  
2 AGE 0 1 15.0 1.40 13 14 15   
3 INCPORAR\_I 0 1 2.35 0.886 0.5 1.62 3   
 p75 p100 hist   
1 32628. 43737 ▇▇▇▇▇  
2 16 17 ▇▇▇▇▇  
3 3 3 ▂▁▁▁▇

# save to file  
summarytable\_file = here("results","tables", "summarytable.rds")  
saveRDS(summary\_df, file = summarytable\_file)

The distribution of the numerical income to poverty ratio can be shown below. It displays a value from 1-3. A ratio displaying less than 1 depicts an income less than the poverty level. Any income ratio >1 indicates an income greater than the poverty level. Most of the observations were gathered from households with income ratios of 3, as shown by the skewed plot below.

p1 <- mydata %>% ggplot(aes(x=INCPORAR\_I)) + geom\_histogram()   
plot(p1)



figure\_file = here("results", "figures", "income.png")  
ggsave(filename = figure\_file, plot=p1)

The state variable shows a more uniformed distribution, suggesting that the observations were gathered fairly across the US states and territory. Although it is noted a few territories provides 0 observations.

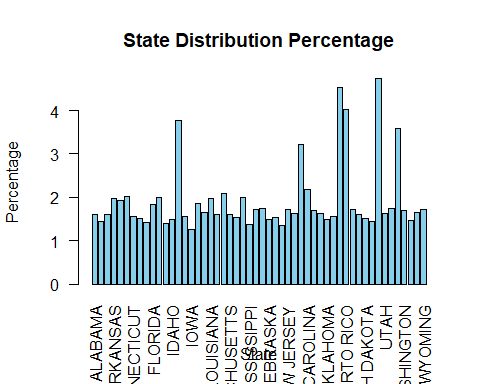
summary(mydata$STATE)

ALABAMA DELAWARE DISTRICT OF COLUMBIA   
 266 250 235   
 FLORIDA GEORGIA HAWAII   
 305 330 233   
 IDAHO ILLINOIS INDIANA   
 247 623 256   
 IOWA ALASKA KANSAS   
 207 238 306   
 KENTUCKY LOUISIANA MAINE   
 275 328 267   
 MARYLAND MASSACHUSETTS MICHIGAN   
 345 267 254   
 MINNESOTA MISSISSIPPI MISSOURI   
 330 228 284   
 MONTANA NEBRASKA NEVADA   
 289 246 255   
 NEW HAMPSHIRE NEW JERSEY NEW MEXICO   
 224 285 270   
 NEW YORK NORTH CAROLINA NORTH DAKOTA   
 534 360 280   
 OHIO ARIZONA OKLAHOMA   
 269 267 245   
 OREGON PENNSYLVANIA RHODE ISLAND   
 258 750 284   
 SOUTH CAROLINA SOUTH DAKOTA TENNESSEE   
 264 251 239   
 TEXAS UTAH ARKANSAS   
 784 268 328   
 VERMONT VIRGINIA WASHINGTON   
 290 592 279   
 WEST VIRGINIA WISCONSIN WYOMING   
 243 272 285   
 CALIFORNIA GUAM PUERTO RICO   
 320 0 668   
 U.S. VIRGIN ISLANDS COLORADO CONNECTICUT   
 0 335 256

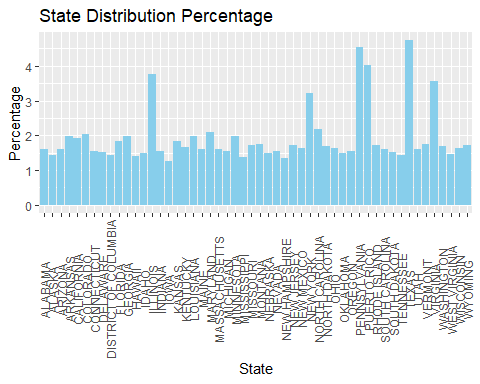
state\_distribution <- table(mydata$STATE)  
# Convert STATE to character  
mydata$STATE <- as.character(mydata$STATE)  
  
# Calculate distribution percentage  
state\_distribution\_percentage <- prop.table(table(mydata$STATE))\*100  
  
print(state\_distribution\_percentage)

ALABAMA ALASKA ARIZONA   
 1.605892 1.436851 1.611929   
 ARKANSAS CALIFORNIA COLORADO   
 1.980198 1.931901 2.022458   
 CONNECTICUT DELAWARE DISTRICT OF COLUMBIA   
 1.545520 1.509297 1.418739   
 FLORIDA GEORGIA HAWAII   
 1.841343 1.992272 1.406665   
 IDAHO ILLINOIS INDIANA   
 1.491186 3.761169 1.545520   
 IOWA KANSAS KENTUCKY   
 1.249698 1.847380 1.660227   
 LOUISIANA MAINE MARYLAND   
 1.980198 1.611929 2.082830   
 MASSACHUSETTS MICHIGAN MINNESOTA   
 1.611929 1.533446 1.992272   
 MISSISSIPPI MISSOURI MONTANA   
 1.376479 1.714562 1.744748   
 NEBRASKA NEVADA NEW HAMPSHIRE   
 1.485149 1.539483 1.352330   
 NEW JERSEY NEW MEXICO NEW YORK   
 1.720599 1.630041 3.223859   
 NORTH CAROLINA NORTH DAKOTA OHIO   
 2.173388 1.690413 1.624004   
 OKLAHOMA OREGON PENNSYLVANIA   
 1.479111 1.557595 4.527892   
 PUERTO RICO RHODE ISLAND SOUTH CAROLINA   
 4.032842 1.714562 1.593818   
 SOUTH DAKOTA TENNESSEE TEXAS   
 1.515334 1.442888 4.733156   
 UTAH VERMONT VIRGINIA   
 1.617967 1.750785 3.574016   
 WASHINGTON WEST VIRGINIA WISCONSIN   
 1.684376 1.467037 1.642115   
 WYOMING   
 1.720599

p2 <- barplot(state\_distribution\_percentage,   
 main = "State Distribution Percentage",   
 xlab = "State",   
 ylab = "Percentage",  
 col = "skyblue", # Set bar color  
 las = 2)



#Recreating the graph in ggplot  
  
#adjusting the percent disitrbution data into a data frame  
data\_df <- as.data.frame(as.table(state\_distribution\_percentage))  
  
#plot with ggplot  
p3 <- ggplot(data\_df, aes(x = Var1, y = Freq)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "State Distribution Percentage",  
 x = "State",  
 y = "Percentage") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
  
#Viewing the plot to confirm it is done properly  
print(p3)



figure\_file = here("results", "figures", "state.distribution.png")  
ggsave(filename = figure\_file, plot=p3)

The family income variable provides a more normal distribution with the majority of the observation coming from $75000+ income level.

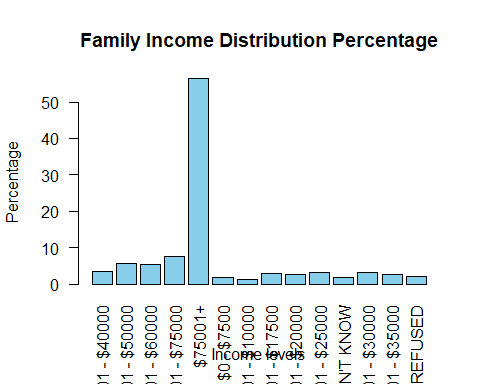
summary(mydata$INCQ298A)

$35001 - $40000 $40001 - $50000 $50001 - $60000 $60001 - $75000 $75001+   
 569 934 908 1273 9345   
 $0 - $7500 $7501 - $10000 $10001 - $17500 $17501 - $20000 $20001 - $25000   
 293 223 475 431 514   
 DON'T KNOW $25001 - $30000 $30001 - $35000 REFUSED   
 301 528 429 341

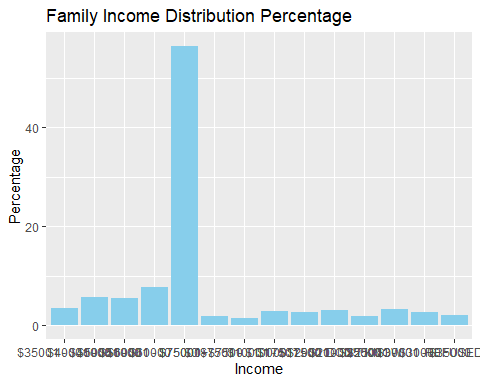
Family\_income\_distribution <- table(mydata$INCQ298A)  
  
# Calculate distribution percentage  
Family\_Income\_distribution <- prop.table(table(mydata$INCQ298A))\*100  
  
print(Family\_Income\_distribution)

$35001 - $40000 $40001 - $50000 $50001 - $60000 $60001 - $75000 $75001+   
 3.435161 5.638735 5.481768 7.685342 56.417532   
 $0 - $7500 $7501 - $10000 $10001 - $17500 $17501 - $20000 $20001 - $25000   
 1.768896 1.346293 2.867665 2.602028 3.103115   
 DON'T KNOW $25001 - $30000 $30001 - $35000 REFUSED   
 1.817194 3.187636 2.589954 2.058681

p4 <- barplot(Family\_Income\_distribution,   
 main = "Family Income Distribution Percentage",   
 xlab = "Income levels",   
 ylab = "Percentage",  
 col = "skyblue", # Set bar color  
 las = 2) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))



#Recreating the graph in ggplot  
  
#adjusting the percent disitrbution data into a data frame  
family\_income\_df <- as.data.frame(as.table(Family\_Income\_distribution))  
  
#plot with ggplot  
p5 <- ggplot(family\_income\_df, aes(x = Var1, y = Freq)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Family Income Distribution Percentage",  
 x = "Income",  
 y = "Percentage")  
  
#Viewing the plot to confirm it is done properly  
print(p5)



figure\_file = here("results", "figures", "family.income.distribution.png")  
ggsave(filename = figure\_file, plot=p5)

The race and ethnicity variables show a distribution that is similar to the U.S. census data, which concludes that the racial composition of the U.S. is 58.9% white, 13.6% African American, and 19.1% Hispanic or Latino.

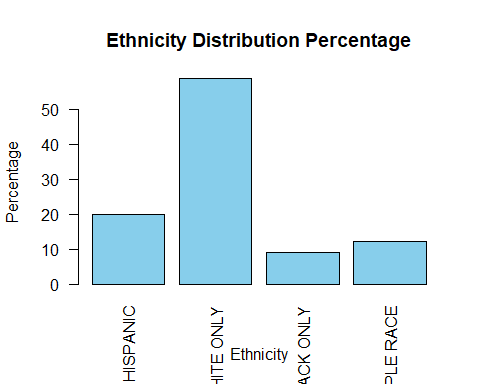
summary(mydata$RACEETHK)

HISPANIC NON-HISPANIC WHITE ONLY   
 3303 9738   
 NON-HISPANIC BLACK ONLY NON-HISPANIC OTHER + MULTIPLE RACE   
 1516 2007

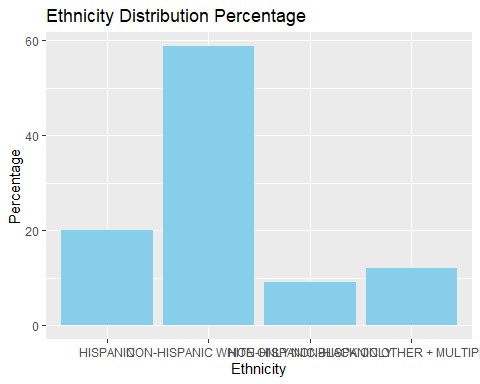
ethnicity\_distribution <- table(mydata$RACEETHK)  
  
# Calculate distribution percentage  
Ethnicity\_distribution <- prop.table(table(mydata$RACEETHK))\*100  
  
print(Ethnicity\_distribution)

HISPANIC NON-HISPANIC WHITE ONLY   
 19.940836 58.790147   
 NON-HISPANIC BLACK ONLY NON-HISPANIC OTHER + MULTIPLE RACE   
 9.152379 12.116638

p6 <- barplot(Ethnicity\_distribution,   
 main = "Ethnicity Distribution Percentage",   
 xlab = "Ethnicity",   
 ylab = "Percentage",  
 col = "skyblue", # Set bar color  
 las = 2) +  
theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))



#Recreating the graph in ggplot  
  
#adjusting the percent disitrbution data into a data frame  
ethnicity\_df <- as.data.frame(as.table(Ethnicity\_distribution))  
  
#plot with ggplot  
p7 <- ggplot(ethnicity\_df, aes(x = Var1, y = Freq)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Ethnicity Distribution Percentage",  
 x = "Ethnicity",  
 y = "Percentage")  
  
#Viewing the plot to confirm it is done properly  
print(p7)



figure\_file = here("results", "figures", "ethnicity.distribution.png")  
ggsave(filename = figure\_file, plot=p7)

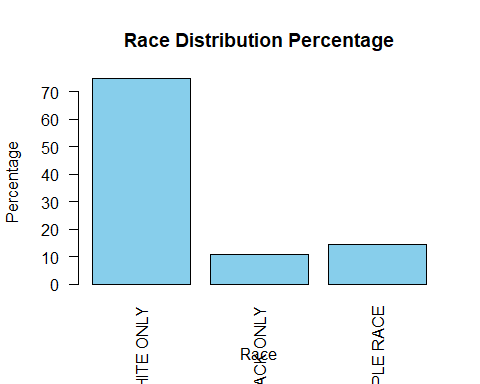
summary(mydata$RACE\_K)

WHITE ONLY BLACK ONLY OTHER + MULTIPLE RACE   
 12378 1807 2379

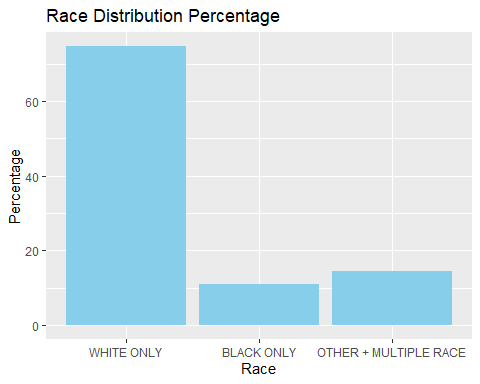
race\_distribution <- table(mydata$RACE\_K)  
  
# Calculate distribution percentage  
Race\_distribution <- prop.table(table(mydata$RACE\_K))\*100  
  
print(Race\_distribution)

WHITE ONLY BLACK ONLY OTHER + MULTIPLE RACE   
 74.72833 10.90920 14.36247

p8 <- barplot(Race\_distribution,   
 main = "Race Distribution Percentage",   
 xlab = "Race",   
 ylab = "Percentage",  
 col = "skyblue", # Set bar color  
 las = 2)



#Recreating the graph in ggplot  
  
#adjusting the percent disitrbution data into a data frame  
race\_df <- as.data.frame(as.table(Race\_distribution))  
  
#plot with ggplot  
p9 <- ggplot(race\_df, aes(x = Var1, y = Freq)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Race Distribution Percentage",  
 x = "Race",  
 y = "Percentage")  
  
#Viewing the plot to confirm it is done properly  
print(p9)



figure\_file = here("results", "figures", "race.distribution.png")  
ggsave(filename = figure\_file, plot=p9)

The Insurance status variable shows that the majority of the observations came from households with private insurance only. This is followed by any medicaid and then other insurance. This is mirrored in the facility variable, where a majority of its observations show private facilities being the location.

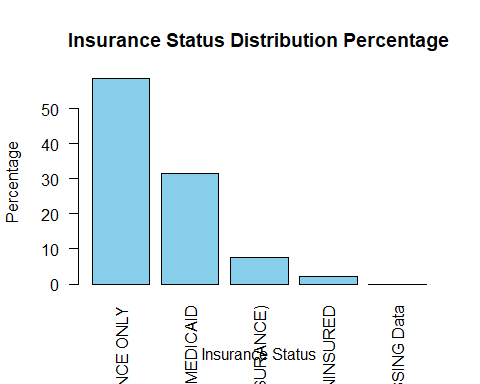
summary(mydata$INS\_STAT2\_I)

PRIVATE INSURANCE ONLY   
 9702   
 ANY MEDICAID   
 5230   
OTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE)   
 1273   
 UNINSURED   
 359   
 MISSING Data   
 0

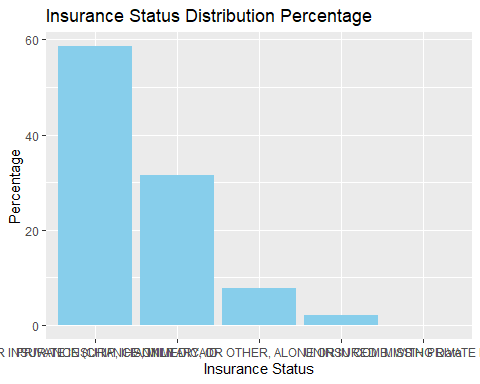
insurance\_status\_distribution <- table(mydata$INS\_STAT2\_I)  
  
# Calculate distribution percentage  
Insurance\_distribution <- prop.table(table(mydata$INS\_STAT2\_I))\*100  
  
print(Insurance\_distribution)

PRIVATE INSURANCE ONLY   
 58.572809   
 ANY MEDICAID   
 31.574499   
OTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE)   
 7.685342   
 UNINSURED   
 2.167351   
 MISSING Data   
 0.000000

p10 <- barplot(Insurance\_distribution,   
 main = "Insurance Status Distribution Percentage",   
 xlab = "Insurance Status",   
 ylab = "Percentage",  
 col = "skyblue", # Set bar color  
 las = 2)



#Recreating the graph in ggplot  
  
#adjusting the percent disitrbution data into a data frame  
insurance\_df <- as.data.frame(as.table(Insurance\_distribution))  
  
#plot with ggplot  
p11 <- ggplot(insurance\_df, aes(x = Var1, y = Freq)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Insurance Status Distribution Percentage",  
 x = "Insurance Status",  
 y = "Percentage")  
  
#Viewing the plot to confirm it is done properly  
print(p11)



figure\_file = here("results", "figures", "insurance.status.distribution.png")  
ggsave(filename = figure\_file, plot=p11)

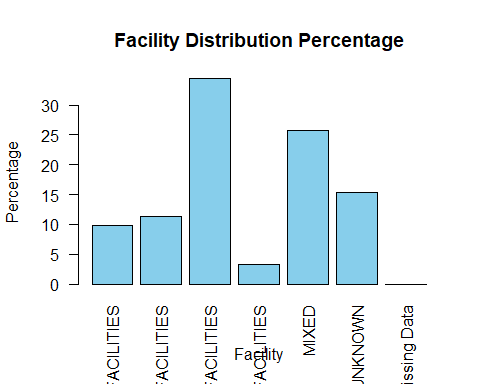
summary(mydata$FACILITY)

ALL PUBLIC FACILITIES   
 1635   
 ALL HOSPITAL FACILITIES   
 1874   
 ALL PRIVATE FACILITIES   
 5700   
ALL STD/SCHOOL/TEEN CLINICS OR OTHER FACILITIES   
 556   
 MIXED   
 4271   
 UNKNOWN   
 2528   
 Missing Data   
 0

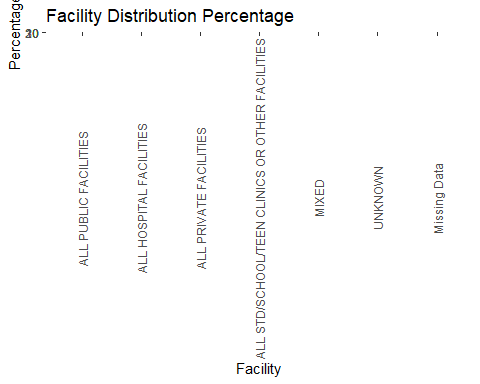
Facility\_distribution <- table(mydata$FACILITY)  
  
# Calculate distribution percentage  
Facility\_distribution <- prop.table(table(mydata$FACILITY))\*100  
  
print(Facility\_distribution)

ALL PUBLIC FACILITIES   
 9.870804   
 ALL HOSPITAL FACILITIES   
 11.313692   
 ALL PRIVATE FACILITIES   
 34.411978   
ALL STD/SCHOOL/TEEN CLINICS OR OTHER FACILITIES   
 3.356677   
 MIXED   
 25.784835   
 UNKNOWN   
 15.262014   
 Missing Data   
 0.000000

p12 <- barplot(Facility\_distribution,   
 main = "Facility Distribution Percentage",   
 xlab = "Facility",   
 ylab = "Percentage",  
 col = "skyblue", # Set bar color  
 las = 2) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))



#Recreating the graph in ggplot  
  
#adjusting the percent disitrbution data into a data frame  
facility\_df <- as.data.frame(as.table(Facility\_distribution))  
  
#plot with ggplot  
p13 <- ggplot(facility\_df, aes(x = Var1, y = Freq)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Facility Distribution Percentage",  
 x = "Facility",  
 y = "Percentage") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
  
#Viewing the plot to confirm it is done properly  
print(p13)



figure\_file = here("results", "figures", "facility.distribution.png")  
ggsave(filename = figure\_file, plot=p13)

The vaccination variable indicates nearly 80% of the observations had an Up-to-date status for the 1+ shot HPV vaccination.

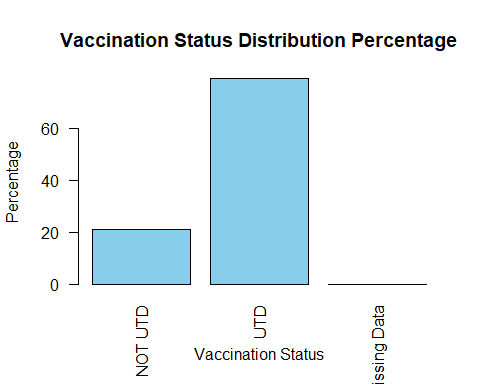
summary(mydata$P\_UTDHPV)

NOT UTD UTD Missing Data   
 3465 13099 0

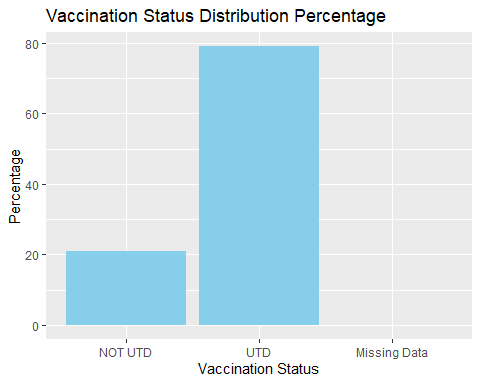
Vaccination\_status\_distribution <- table(mydata$P\_UTDHPV)  
  
# Calculate distribution percentage  
Vaccination\_status\_distribution <- prop.table(table(mydata$P\_UTDHPV))\*100  
  
print(Vaccination\_status\_distribution)

NOT UTD UTD Missing Data   
 20.91886 79.08114 0.00000

p14 <- barplot(Vaccination\_status\_distribution,   
 main = "Vaccination Status Distribution Percentage",   
 xlab = "Vaccination Status",   
 ylab = "Percentage",  
 col = "skyblue", # Set bar color  
 las = 2)



#Recreating the graph in ggplot  
  
#adjusting the percent disitrbution data into a data frame  
Vaccine\_df <- as.data.frame(as.table(Vaccination\_status\_distribution))  
  
#plot with ggplot  
p15 <- ggplot(Vaccine\_df, aes(x = Var1, y = Freq)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "Vaccination Status Distribution Percentage",  
 x = "Vaccination Status",  
 y = "Percentage")  
  
#Viewing the plot to confirm it is done properly  
print(p15)



figure\_file = here("results", "figures", "vaccination.status.distribution.png")  
ggsave(filename = figure\_file, plot=p15)

We examine the percentage of up-to-date vaccination status by state.

#Aggregate vaccine status stratified by state  
vaccine\_state\_percent <- mydata %>%  
 group\_by(STATE) %>%  
 summarise(percentage\_UTD = sum(P\_UTDHPV == "UTD") / n() \* 100)  
  
#Create table to display only up-to-date status in percents for each state  
table1 <- mydata %>%  
 filter(P\_UTDHPV == "UTD") %>%  
 count(STATE) %>%  
 rename(UTD\_count = n) %>%  
 left\_join(vaccine\_state\_percent, by = "STATE") %>%  
 select(STATE, percentage\_UTD) %>%  
 arrange(STATE)  
  
#Print the table  
kable(table1, caption = "Vaccine Distribution Percentage per State (Only 'UTD' Status)")

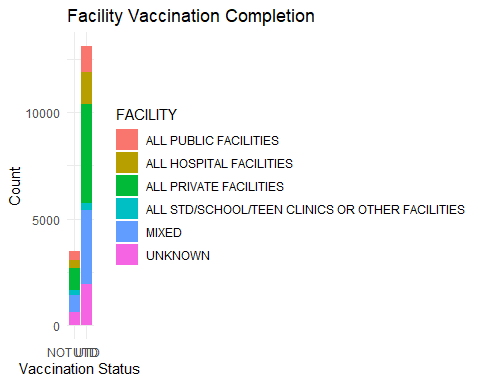
Vaccine Distribution Percentage per State (Only ‘UTD’ Status)

| STATE | percentage\_UTD |
| --- | --- |
| ALABAMA | 73.68421 |
| ALASKA | 75.21008 |
| ARIZONA | 76.02996 |
| ARKANSAS | 73.47561 |
| CALIFORNIA | 74.37500 |
| COLORADO | 81.49254 |
| CONNECTICUT | 79.29688 |
| DELAWARE | 84.80000 |
| DISTRICT OF COLUMBIA | 86.38298 |
| FLORIDA | 74.42623 |
| GEORGIA | 71.51515 |
| HAWAII | 87.98283 |
| IDAHO | 77.73279 |
| ILLINOIS | 82.82504 |
| INDIANA | 77.34375 |
| IOWA | 87.92271 |
| KANSAS | 73.85621 |
| KENTUCKY | 68.72727 |
| LOUISIANA | 79.87805 |
| MAINE | 78.65169 |
| MARYLAND | 85.21739 |
| MASSACHUSETTS | 86.51685 |
| MICHIGAN | 76.37795 |
| MINNESOTA | 85.45455 |
| MISSISSIPPI | 56.14035 |
| MISSOURI | 74.29577 |
| MONTANA | 82.35294 |
| NEBRASKA | 83.73984 |
| NEVADA | 73.72549 |
| NEW HAMPSHIRE | 85.71429 |
| NEW JERSEY | 72.98246 |
| NEW MEXICO | 81.85185 |
| NEW YORK | 81.46067 |
| NORTH CAROLINA | 75.27778 |
| NORTH DAKOTA | 85.35714 |
| OHIO | 78.81041 |
| OKLAHOMA | 70.61224 |
| OREGON | 81.39535 |
| PENNSYLVANIA | 83.06667 |
| PUERTO RICO | 86.52695 |
| RHODE ISLAND | 94.71831 |
| SOUTH CAROLINA | 75.37879 |
| SOUTH DAKOTA | 81.27490 |
| TENNESSEE | 71.12971 |
| TEXAS | 73.59694 |
| UTAH | 76.86567 |
| VERMONT | 87.93103 |
| VIRGINIA | 80.06757 |
| WASHINGTON | 81.00358 |
| WEST VIRGINIA | 70.78189 |
| WISCONSIN | 82.35294 |
| WYOMING | 69.47368 |

#Write the table to save to the results  
write.table(table1, file = here("results","tables", "state-vaccine-completion.txt"), sep = "\t", row.names = FALSE)

We examine the facility that submitted the survey, which is stratified by the vaccine completion status of the survey subject.

#Stratifying the HPV vaccine completion by facility  
##Make a stacked bar plot to compare which facilities reported UTD patients  
p18 <- ggplot(mydata, aes(x = P\_UTDHPV, fill = FACILITY)) +  
 geom\_bar(stat = "count") + # Use geom\_bar() with the count of each facility  
 labs(x = "Vaccination Status", y = "Count", title = "Facility Vaccination Completion") +  
 theme\_minimal()  
 print(p18)



#Save the bar plot  
figure\_file <- here("results", "figures", "facility-vaccination.png")  
ggsave(filename = figure\_file, plot = p18, width = 8, height = 6, dpi = 300)

We examine the percentage of vaccination completion by race and ethnicity. Because there are different counts of each race/ethnicity for each factor level, the percentage of those with up-to-date vaccine status out of each race/ethnicity level was determined. This adjusts for the different counts in each racial category. The Hispanic ethnicity was found to have the highest vaccine completion while non-hispanic white has the lowest.

#Aggregate total count per each race-ethnicity factor level  
total\_count\_by\_race <- mydata %>%  
 group\_by(RACEETHK) %>%  
 summarise(total\_count = n())  
  
#Aggregate count of those with UTD status for each race-enthnicity factor level  
UTD\_by\_race <- mydata %>%  
 filter(P\_UTDHPV == "UTD") %>%  
 group\_by(RACEETHK) %>%  
 summarise(UTD\_count = n())  
  
#Merge the total count and UTD count data frames  
vaccine\_percent\_by\_race <- total\_count\_by\_race %>%  
 left\_join(UTD\_by\_race, by = "RACEETHK") %>%  
 mutate(percentage\_UTD = UTD\_count / total\_count \* 100)  
  
# Print the table  
kable(vaccine\_percent\_by\_race, caption = "Percentage of UTD Vaccination Status by RACEETHK")

Percentage of UTD Vaccination Status by RACEETHK

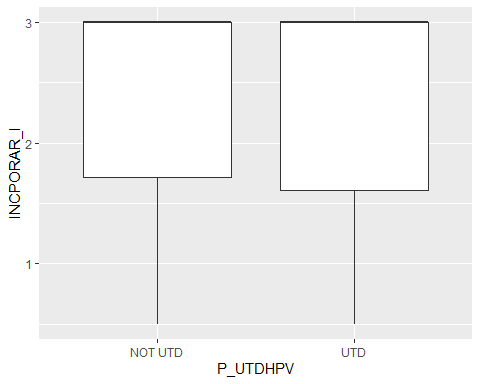
| RACEETHK | total\_count | UTD\_count | percentage\_UTD |
| --- | --- | --- | --- |
| HISPANIC | 3303 | 2728 | 82.59158 |
| NON-HISPANIC WHITE ONLY | 9738 | 7514 | 77.16163 |
| NON-HISPANIC BLACK ONLY | 1516 | 1218 | 80.34301 |
| NON-HISPANIC OTHER + MULTIPLE RACE | 2007 | 1639 | 81.66418 |

# Save the table as a text file  
write.table(vaccine\_percent\_by\_race, file = here("results", "tables", "vaccine\_percent\_by\_race.txt"), sep = "\t", row.names = FALSE)

### 1.4.3 Model and Plot Fitting

A boxplot is created with income-poverty ratio to vaccination status. The box plot indicates little difference between the two status. A statistical model is fitted to the same variables. A p-value of 0.1358 suggests there is little significance between the income-poverty variable and the vaccination status.

#Stratifying the Income-poverty levels to the vaccination status in a boxplot figure  
p16 <- mydata %>% ggplot(aes(x=P\_UTDHPV, y=INCPORAR\_I)) + geom\_boxplot()  
plot(p16)



figure\_file = here("results","figures", "income-vaccination.png")  
ggsave(filename = figure\_file, plot=p16)   
  
# Fitting Income-poverty levels and Vaccination status to a statistical model  
model <- lm(INCPORAR\_I ~ P\_UTDHPV, data = mydata)  
summary(model)

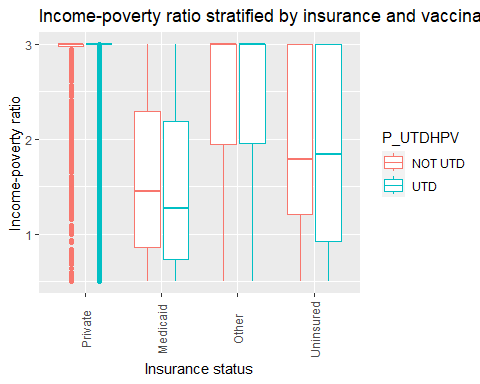
Call:  
lm(formula = INCPORAR\_I ~ P\_UTDHPV, data = mydata)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1.8680 -0.7304 0.6320 0.6572 0.6572   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.36800 0.01505 157.381 <2e-16 \*\*\*  
P\_UTDHPVUTD -0.02524 0.01692 -1.492 0.136   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.8857 on 16562 degrees of freedom  
Multiple R-squared: 0.0001343, Adjusted R-squared: 7.394e-05   
F-statistic: 2.225 on 1 and 16562 DF, p-value: 0.1358

A boxplot is created with income-poverty ratio to insurance status. The box plot indicates there are distinct difference between the two status. A statistical model is fitted to the same variables. The p-value of < 2.26e-16 suggests there is statistical significance between the income-poverty variable and the insurance status.

#Stratifying the Income-poverty levels to the insurance status in a boxplot figure  
original\_labels <- levels(mydata$INS\_STAT2\_I)  
print(original\_labels)

[1] "PRIVATE INSURANCE ONLY"   
[2] "ANY MEDICAID"   
[3] "OTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE)"  
[4] "UNINSURED"   
[5] "MISSING Data"

p17 <- mydata %>%   
 ggplot(aes(x=INS\_STAT2\_I, y=INCPORAR\_I, color = P\_UTDHPV)) +   
 geom\_boxplot() +  
 labs(x = "Insurance status", y = "Income-poverty ratio", title = "Income-poverty ratio stratified by insurance and vaccination") +  
 scale\_x\_discrete(labels = c("PRIVATE INSURANCE ONLY" = "Private", "ANY MEDICAID" = "Medicaid", "OTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE)" = "Other", "UNINSURED" = "Uninsured")) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5)) # Rotate x-axis text by 90 degrees  
plot(p17)



figure\_file = here("results","figures", "insurance-income-stratified.png")  
ggsave(filename = figure\_file, plot=p17)   
  
# Fitting Income-poverty levels and insurance status to a statistical model  
model <- lm(INCPORAR\_I ~ INS\_STAT2\_I, data = mydata)  
summary(model)

Call:  
lm(formula = INCPORAR\_I ~ INS\_STAT2\_I, data = mydata)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.2951 -0.3791 0.2049 0.2049 1.4747   
  
Coefficients:  
 Estimate  
(Intercept) 2.795130  
INS\_STAT2\_IANY MEDICAID -1.269800  
INS\_STAT2\_IOTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE) -0.343549  
INS\_STAT2\_IUNINSURED -0.911345  
 Std. Error  
(Intercept) 0.006797  
INS\_STAT2\_IANY MEDICAID 0.011486  
INS\_STAT2\_IOTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE) 0.019959  
INS\_STAT2\_IUNINSURED 0.035985  
 t value  
(Intercept) 411.21  
INS\_STAT2\_IANY MEDICAID -110.56  
INS\_STAT2\_IOTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE) -17.21  
INS\_STAT2\_IUNINSURED -25.33  
 Pr(>|t|)  
(Intercept) <2e-16  
INS\_STAT2\_IANY MEDICAID <2e-16  
INS\_STAT2\_IOTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE) <2e-16  
INS\_STAT2\_IUNINSURED <2e-16  
   
(Intercept) \*\*\*  
INS\_STAT2\_IANY MEDICAID \*\*\*  
INS\_STAT2\_IOTHER INSURANCE (CHIP, IHS, MILITARY, OR OTHER, ALONE OR IN COMB. WITH PRIVATE INSURANCE) \*\*\*  
INS\_STAT2\_IUNINSURED \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.6695 on 16560 degrees of freedom  
Multiple R-squared: 0.4287, Adjusted R-squared: 0.4286   
F-statistic: 4142 on 3 and 16560 DF, p-value: < 2.2e-16

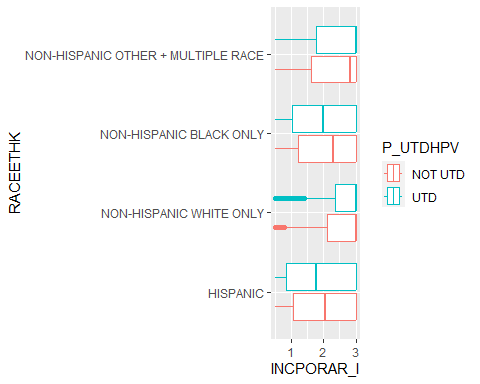
Income-poverty ratio was fit to vaccine status and insurance status in a multi-linear model.

# Fitting Income-poverty levels and insurance status to a statistical model  
model <- lm(INCPORAR\_I ~ P\_UTDHPV, INS\_STAT2\_I, data = mydata)  
summary(model)

Call:  
lm(formula = INCPORAR\_I ~ P\_UTDHPV, data = mydata, subset = INS\_STAT2\_I)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.691e-12 -3.300e-15 0.000e+00 0.000e+00 2.216e-11   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.292e+00 2.434e-15 9.419e+14 <2e-16 \*\*\*  
P\_UTDHPVUTD 7.076e-01 2.942e-15 2.405e+14 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.76e-13 on 16562 degrees of freedom  
Multiple R-squared: 1, Adjusted R-squared: 1   
F-statistic: 5.783e+28 on 1 and 16562 DF, p-value: < 2.2e-16

A box-plot was created for income-poverty level versus race/ethnicity and this was stratified by vaccine completion (UTD or NOT UTD). A lower income-poverty ratio means the group is closer to the poverty line. For the non-Hispanic black and Hispanic racial and ethnic groups, vaccine completion corresponded with a lower mean income-poverty ratio. For mixed ethnic groups, vaccine completion corresponded with a higher mean income-poverty ratio.

#Stratifying the HPV vaccine completion by both Race-Ethnicity and (choose either insurance or income-poverty status)  
p19 <- mydata %>%   
 ggplot(aes(x=INCPORAR\_I, y=RACEETHK, color = P\_UTDHPV)) + geom\_boxplot()  
plot(p19)



figure\_file = here("results","figures", "raceeth-income-stratified.png")  
ggsave(filename = figure\_file, plot=p19)   
  
# Fitting Income-poverty levels and insurance status to a statistical model  
model <- lm(INCPORAR\_I ~ P\_UTDHPV, RACEETHK, data = mydata)  
summary(model)

Call:  
lm(formula = INCPORAR\_I ~ P\_UTDHPV, data = mydata, subset = RACEETHK)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.411e-12 -5.000e-15 0.000e+00 0.000e+00 4.160e-11   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.292e+00 3.294e-15 6.959e+14 <2e-16 \*\*\*  
P\_UTDHPVUTD 7.076e-01 5.131e-15 1.379e+14 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 3.251e-13 on 16562 degrees of freedom  
Multiple R-squared: 1, Adjusted R-squared: 1   
F-statistic: 1.901e+28 on 1 and 16562 DF, p-value: < 2.2e-16

## 1.5 References

1. 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/
2. 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
3. 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