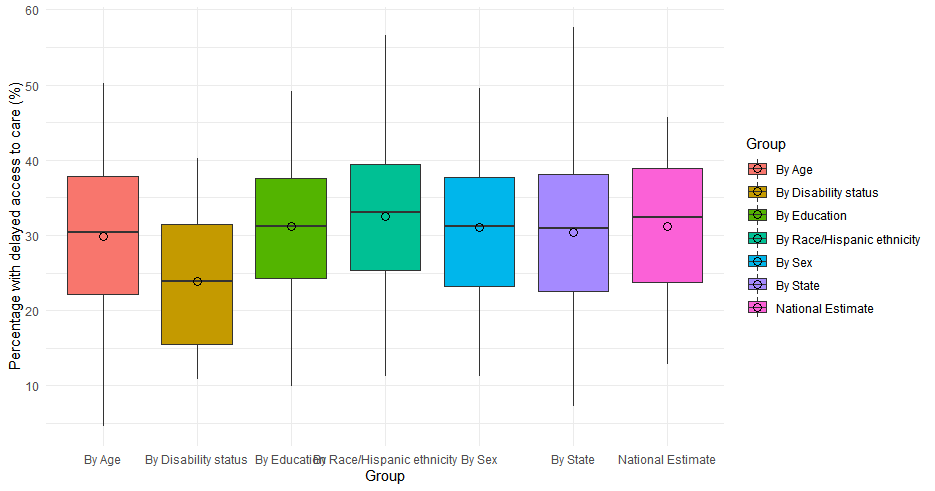
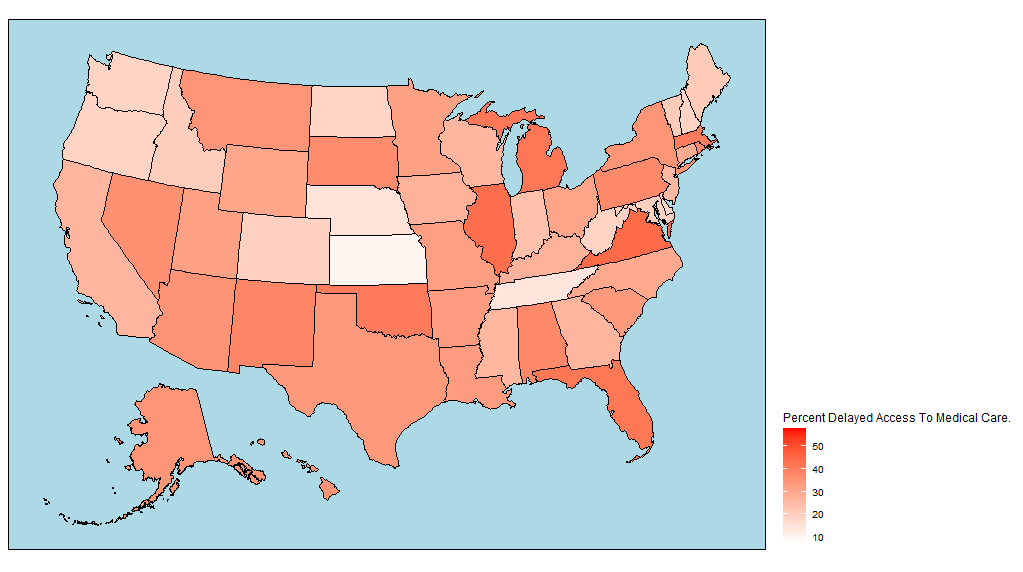
Interactions of Indicators of Reduced Access to Covid-19 Medical Care

Early on in the pandemic, as part of the government’s response to Covid-19, the Census Bureau launched its Household Pulse Survey nationwide. Designed as a 20-minute online survey, the data covers a wide variety of questions and topics. This paper focuses specifically on one response variable—Have you been delayed in getting medical care because of the pandemic?—and several related identifiers that might have a relationship with the response variable. These categorical variables include Age groups, Education level, race, sex, disability status and state. The aim was to explore the question, “What were some of the most important factors that prevented someone from getting medical care during the pandemic?”

Initially, results came from a post-collection government-processed dataset that analyzed the estimated percentage of U.S. residents with their access to medical care delayed because of the pandemic. Responses were categorized by the above identifiers and collected over 33 separate “Weeks” from April 23rd, 2020 until July 5th, 2021. Note that designated “Weeks” often included longer or shorter cycles. The collection criteria included a combination of both an inability and self-guided choice not to receive care; the data does not account for whether the preventative factor was perceived or directly experienced. Data was adjusted for nonresponse cases. Below are summary boxplots by group and a five-number summary that takes into account every data point.

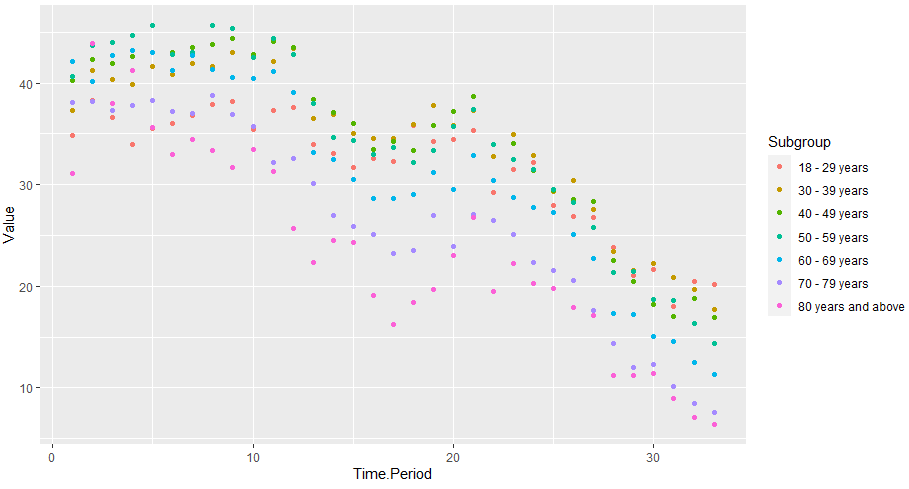
| National Estimate | | | | | | | All Data | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Min | Q1 | Med | Q3 | Max | Mean | SD | Min | Q1 | Med | Q3 | Max | Mean | SD |
| 14.6 | 27.7 | 33.1 | 40.3 | 41.5 | 31.1 | 9.69 | 4.6 | 22.8 | 31.2 | 38.1 | 57.7 | 30.9 | 9.80 |



In the national estimate (the “average” category), with IQR\*1.5 = 18.9, no value falls far enough out of range to be considered an outlier. The same is true when looking at all categories together. This means that during no week was access suddenly very difficult or easy and that the data is distributed approximately symmetrically. The boxplots confirm this visually; no line was more than 1.5 times the size of its respective box. They also show that the groups have similar means—this is to be expected given that they all are supposed to represent samples of the entire population. Note that disability status was not added as a variable until weeks into the survey, so I did not use it.

Below are three of the several preliminary plots made to explore the data further by separating into groups and subgroups.

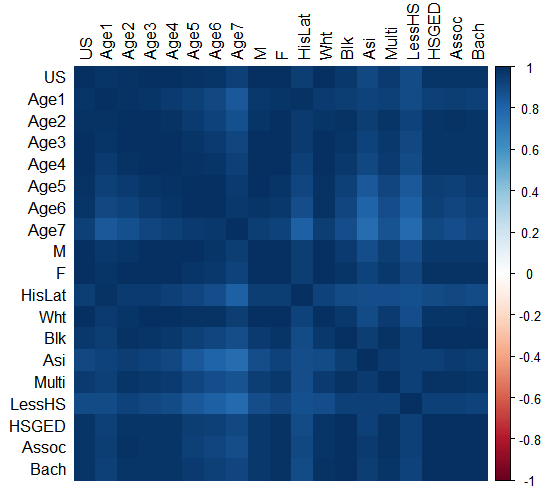
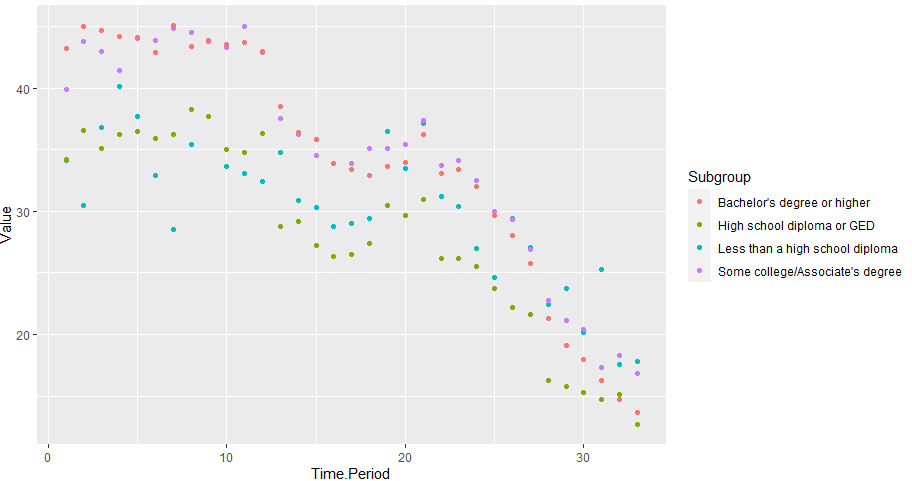
There was clear variation among each category’s subgroups; people with less than a college education generally had less delayed medical care, as did older people and those living in, say, Kansas or Tennessee. I had presumed older people would be prioritized, and it is possible that more educated people tend to live in more densely-packed areas (or maybe they were more cautious).

To confirm statistically that there were actually differences between the subgroups, I ran a chi-squared test for homogeneity. Each of my groups/samples was one of the different subgroups within a group (for example, the decade-long age groups among all data categorized by age) and the categorical variable was the number of people who did/didn’t experience delayed access (based on a different version of the same government data with numbers instead of percentages). H0 was that the proportions are all the same, and Ha was that they are different. The data was randomly collected and every count was much greater than 5. I did not run too many tests because of how many weeks and subgroups exist, but the ones I did consistently output p-values of 1, like the one given below analyzing by age subgroup for Week 33:

Pearson's Chi-squared test

data: oldyoung

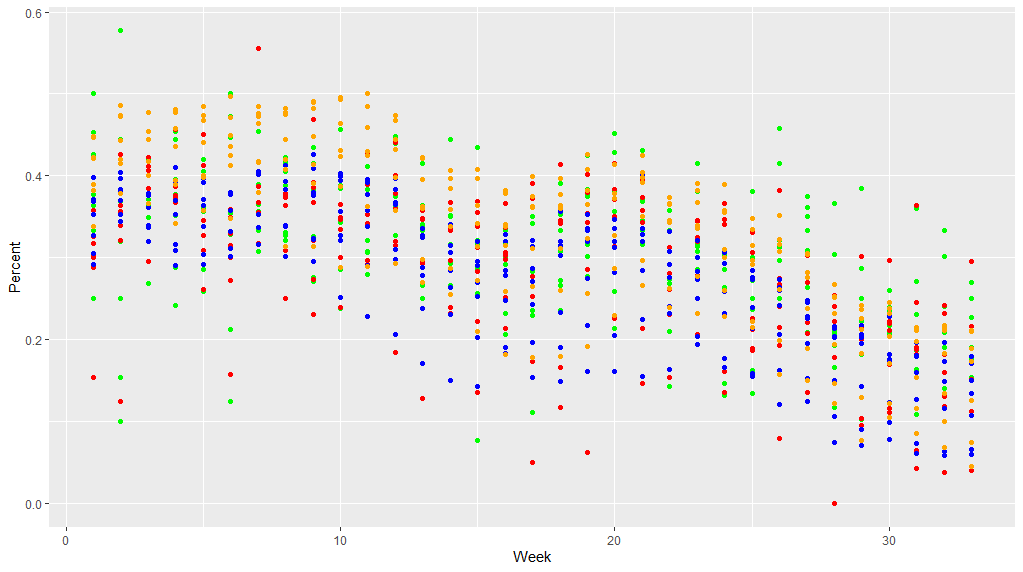
X-squared = 1.4623, df = 18, p-value = 1

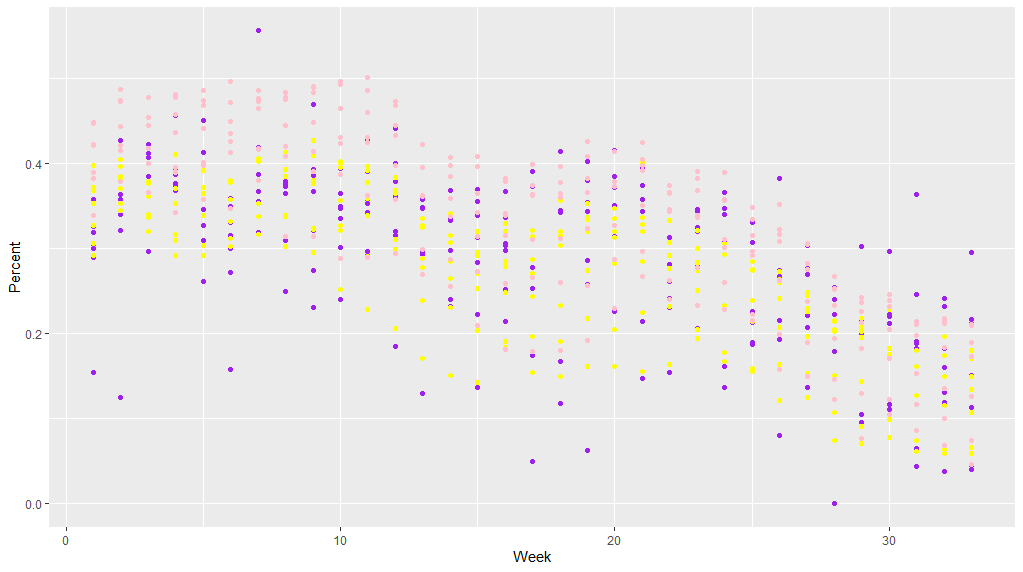
The test reports that the subgroups are maybe not as separate as they seem. I failed to reject the null hypothesis at the α = 0.05 significance level and had to conclude that the subgroups could have come from the same population/there was not strong evidence to suggest a difference in access. It is also very possible this is some sort of computing error, but R did not output the warning about unreliability that it sometimes does when returning a statistical test.

Regardless, every graph I tested had that similar overall downward trend with a peak somewhere around the 22nd week (which corresponds to early January 2021, when there was a significant spike in cases). I confirmed this with a correlation matrix, demonstrated below. Regardless of the magnitude differences in access, no subgroup stood out as peaking or dropping dramatically at different times. Possible outlier groups were the very oldest people, those with less than a high school education, and certain Asian and Hispanic/Latino groups. I could not easily pick out a “most important” variable.

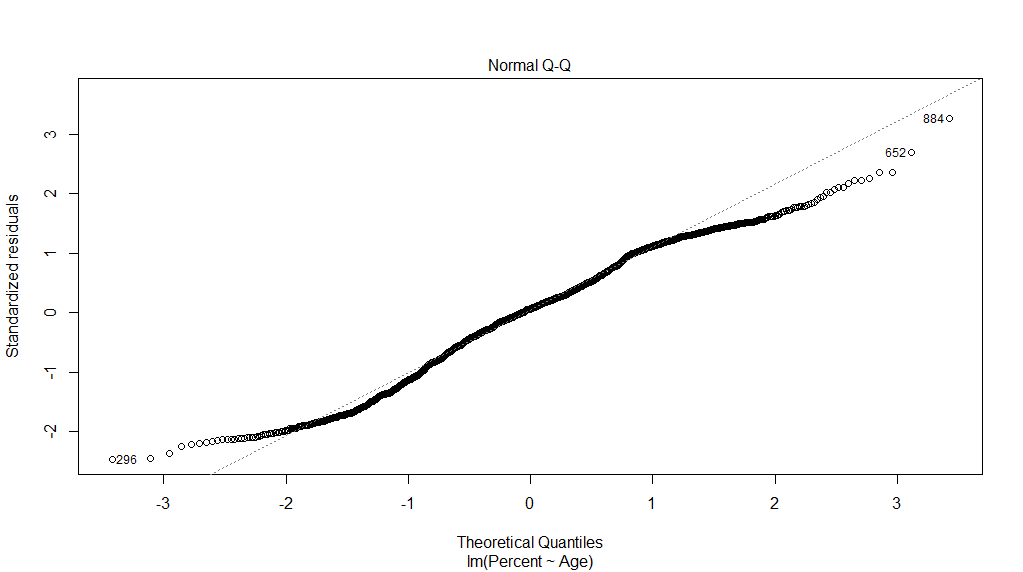
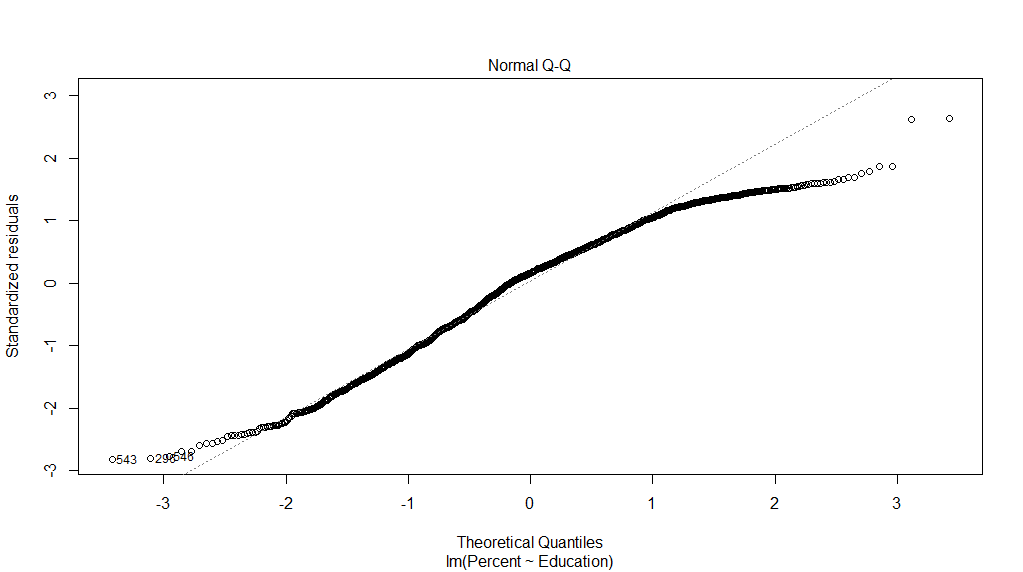
To further explore the data and compare groups by more than one factor, I needed to find a way to attach percent access data points to labels with more than one identifier; the dataset I initially found only contained statistics by single subgroups. To do so, I had to access the individual response data so that I could find combinations of groups and estimate access that way. I selected age group and education level as my new points of focus, and I wanted to see if certain combinations of the two might predict an even lower or higher likelihood of access to care than either variable would predict alone.

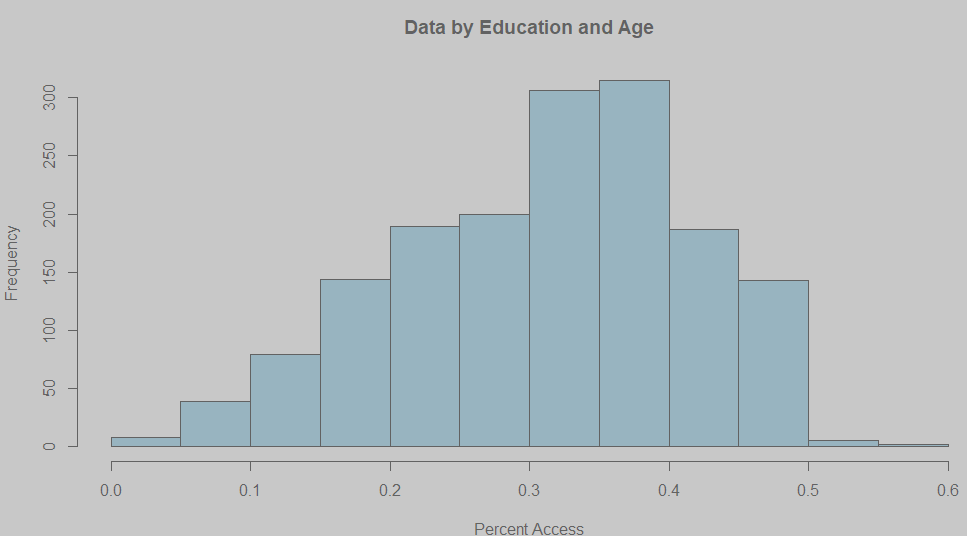
To help make processing slightly easier, I combined some of the seven education levels in the individual response data to match the four levels in the processed data. I also combined several age groups: in total, there were 16 groups that accounted for every combination of age and education level that I set. I made a separate script, denoted in the code pages as the Microdata Processor Script, that would filter out responses for each of the 16 groups and calculate the percentage that had delayed access to medical treatment. Note that unlike the government-processed data, I did not weight for nonresponse, which may have had a slight impact (most weights in the government data only shifted the percentages 1-2%). However, this was essentially the same dataset.

Below I tried to graph both age and education at the same time; each color corresponds to a certain level of education and you can see the vertical distributions during each week for the same color that show the variation by age. Unfortunately R would not let me overlay more than four groups at once. Note that the patterns are (very roughly) similar but obviously there are magnitude differences. Also note that the green (Less than high school education) and red (Some high school education) and, to an extent, purple (Associate’s degree earned) plots have very inconsistent trends.

The next step was to test if the possibility of these standouts in age and education was possibly due to some sort of intersectional effect that could come as the result of a combination of certain age groups and education levels. I determined, based on the statistical tests outlined in the textbook, that the way to do this was through a two-way ANOVA test.

I performed the ANOVA with R’s integrated aov() function. The H0 here is that changing age and education level do not affect access to treatment. An additional H0 states that all γij values (which quantify the level of interaction/the change in the response variable that is more than the sum of the explanatory variables themselves) will be equal to 0. I performed this test at the α = 0.05 significance level. A quick histogram showed the population data was approximately normally distributed, and the data is randomly collected and independent. Normal probability plots were taken as roughly linear, though it could be argued the Education plot perhaps tailed off too much at the end.





Initially, while the first two H0 were of course very strongly rejected (p-values on the order of 10E-16), there was no observed interaction effect. Looking back at the correlation plot, I decided to further split up the ages and education levels so that groups like people 80+ years old or those with less than a high school education would be separated out. Running the aov() test again, I obtained the following output.

|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| --- | --- | --- | --- | --- | --- |
| Age | 6 | 2.491 | 0.4152 | 47.171 | < 2e-16 \*\*\* |
| Education | 6 | 0.828 | 0.1380 | 15.678 | < 2e-16 \*\*\* |
| Age:Education | 36 | 0.566 | 0.0157 | 1.787 | 0.00295 \*\* |
| Residuals | 1568 | 13.801 | 0.0088 |  |  |

This time I was able to reject all three null hypotheses because of p-values < 0.05; while I was not quite able to obtain information about which variable had the most significant effect, I did show that access to medical care was more complicated than might possibly be assumed. Simply looking at a person’s education and age identifiers alone would *not* necessarily be a good predictor of how their total likelihood of access to quick and on-time medical treatment during the pandemic would be. Instead, we can conclude that there is some sort of interaction effect on experiencing delayed access to medical care that comes from being in certain age/education groups and is not simply the addition of the two independent effects.

###################################################################################

Code Part 1

###################################################################################

## setwd("C:/Users/Thirty-Eight/Dropbox/PC/Documents/R/HCBPS-Final-Project/Datasets/PewScience")

## PewScience = read.spss('PewScience.sav')

## head(PewScience)

## setwd("C:/Users/Thirty-Eight/Dropbox/PC/Documents/R/HCBPS-Final-Project/Datasets/Pew2")

## pewMedia = read.spss('ATP W85.sav')

## head(pewMedia)

library(dplyr)

library(ggplot2)

library(stringr)

#library(ggcorrplot)

library(corrplot)

library(ggpubr)

library(tidyverse)

library(broom)

library(AICcmodavg)

############################ Load Data, Exploratory Plots ############################################

setwd("C:/Users/Thirty-Eight/Dropbox/PC/Documents/R/HCBPS-Final-Project/Datasets/GovCovidIndicators")

govCovid = read.csv('Indicators\_of\_Reduced\_Access\_to\_Care\_Due\_to\_the\_Coronavirus\_Pandemic\_During\_Last\_4\_Weeks.csv')

head(govCovid)

govCovid$Time.Period.Start.Date

hist(govCovid$Time.Period, col = 'SKy Blue', main = 'Missed Covid Treatment Last 4 Weeks')

scatter.smooth(govCovid$Time.Period, y = govCovid$Value, xlab = "Days Since 5/5/2020", ylab = "Value", main = "Did Not Get or Delayed Covid Treatment")

table(govCovid$Subgroup)

barplot(eduCovid$Value, col = 'Blue')

plot(eduCovid$Value)

eduData = govCovid %>% filter(Group == "By Education")

head(eduData)

bachData = govCovid %>% filter(Subgroup == "Bachelor's degree or higher", Indicator == "Delayed Medical Care, Last 4 Weeks")

collegeData = govCovid %>% filter(Subgroup == "Some college/Associate's degree")

hsgedData = govCovid %>% filter(Subgroup == "High school diploma or GED")

lhsData = govCovid %>% filter(Subgroup == "Less than a high school diploma")

#arrange(bachData, by\_group = TRUE)

scatter.smooth(bachData$Time.Period, y = bachData$Value, xlab = "Days Since 5/5/2020", ylab = "Value", main = "Did Not Get or Delayed Covid Treatment")

scatter.smooth(collegeData$Time.Period, y = collegeData$Value, xlab = "Days Since 5/5/2020", ylab = "Value", main = "Did Not Get or Delayed Covid Treatment")

scatter.smooth(hsgedData$Time.Period, y = hsgedData$Value, xlab = "Days Since 5/5/2020", ylab = "Value", main = "Did Not Get or Delayed Covid Treatment")

scatter.smooth(lhsData$Time.Period, y = lhsData$Value, xlab = "Days Since 5/5/2020", ylab = "Value", main = "Did Not Get or Delayed Covid Treatment")

############################ Separate By Subgroups ############################################

filterer <- function() {

inType <- readline(prompt="Delayed Medical Care (1) OR Did Not Get Needed Care (2) OR Delayed or Did Not Get Care (3)? ")

if (inType == "1"){

indicator = "Delayed Medical Care, Last 4 Weeks"

}

else if (inType == "2"){

indicator = "Did Not Get Needed Care, Last 4 Weeks"

}

else if (inType == "3"){

indicator = "Delayed or Did Not Get Care, Last 4 Weeks"

}

else{

break

}

subgroup <- readline(prompt="What is the group? ")

return(govCovid %>% filter(Group == subgroup, Indicator == indicator))

}

scatterer <- function(){

customData = filterer()

customData$numYear = as.Date(as.character(customData$Time.Period.Start.Date))

sp<-ggplot(customData, aes(x=numYear, y=Value, color=Subgroup)) + geom\_point() + scale\_x\_continuous(breaks = c(1,5), limits = customData$Time.Period.Start.Date)

sp

}

scatterer()

nationalEstimate = govCovid %>% filter(Group == "National Estimate", Indicator == "Delayed Medical Care, Last 4 Weeks")

ggplot(data = govCovid, aes(fill = Group)) +

geom\_boxplot(mapping = aes(x = Group, y = Value))+

stat\_summary(mapping = aes(x = Group, y = Value), fun = mean, geom = "point", shape = 1, size = 3)+

scale\_y\_continuous("Percentage with delayed access to care (%)")+

theme\_minimal()

fivenum(govCovid$Value)

################################## By State ##############################

library(usmap)

stat = govCovid %>% filter(Group == "By State")

stat

names(stat)[names(stat) == 'State'] <- 'state'

plot\_usmap(data = stat, values = "Value", color = "black") +

scale\_fill\_continuous(low = "white", high = "red", name = "Percent Delayed Access To Medical Care.", label = scales::comma) +

theme(legend.position = "right", panel.background = element\_rect(color = "black", fill = "lightblue"))

################################## By State ##############################

prop.test(x, n, p = NULL, alternative = "two.sided",

correct = TRUE)

#########################Correlation Matrix ############################################

head(cormat)

setwd("C:/Users/Thirty-Eight/Dropbox/PC/Documents/R/HCBPS-Final-Project/Datasets/GovCovidIndicators")

cormat <- read.csv("CorrelatableData.csv", header = TRUE)

vars <- cormat[,c('18 - 29 years', '30 - 39 years', '40 - 49 years', '50 - 59 years', '60 - 69 years', '70 - 79 years', '80 years and above', 'Male',

'Female', 'Hispanic or Latino', 'Non-Hispanic White, single race', 'Non-Hispanic Black, single race', 'Non-Hispanic Asian, single race',

'Non-Hispanic, other races and multiple races', 'Less than a high school diploma', 'High school diploma or GED', "Some college/Associate's degree",

"Bachelor's degree or higher")]

corrplot.mixed(cor(mtcars)) #, order = "hclust", tl.col = "black"

corrplot.mixed(cor(mtcars), lower = 'square', upper = 'number', order = 'hclust')

corrplot(cor(cormat), method = 'color',tl.col = "black")

############################# Homogeneity Chi Squared ####################

setwd("C:/Users/Thirty-Eight/Dropbox/PC/Documents/")

xsq <- read.csv("chisquarabledata.csv", header = TRUE)

#Male/Female

malefemale <- rbind(xsq[9,], cormat[10,])

chisq.test(malefemale)

#Old/Young

oldyoung <- rbind(cormat[1,], cormat[7,])

chisq.test(oldyoung)

############################ Loading Custom Data By Age and Education ############################################

setwd("C:/Users/Thirty-Eight/Dropbox/PC/Documents/R/HCBPS-Final-Project/Datasets/HouseholdPulseMicrodata")

newgovCovid = read.csv('superAgeEduDataset')

head(newgovCovid, n = 20)

eduData1 = newgovCovid %>% filter(Education == "Less than high school")

eduData2 = newgovCovid %>% filter(Education == "Some high school")

eduData3 = newgovCovid %>% filter(Education == "High school graduate or equivalent (for example GED)")

eduData4 = newgovCovid %>% filter(Education == "Some college, but degree not received or is in progress")

eduData5 = newgovCovid %>% filter(Education == "Associate’s degree (for example AA, AS)")

eduData6 = newgovCovid %>% filter(Education == "Bachelor's degree (for example BA, BS, AB)")

eduData7 = newgovCovid %>% filter(Education == "Graduate degree (for example master's, professional, doctorate)")

ggplot() +

geom\_point(data=eduData1, aes(x=Week, y=Percent), color='green') +

geom\_point(data=eduData2, aes(x=Week, y=Percent), color='red') +

geom\_point(data=eduData3, aes(x=Week, y=Percent), color='blue') +

geom\_point(data=eduData4, aes(x=Week, y=Percent), color='orange')

ggplot() +

geom\_point(data=eduData2, aes(x=Week, y=Percent), color='purple') +

geom\_point(data=eduData3, aes(x=Week, y=Percent), color='yellow') +

geom\_point(data=eduData4, aes(x=Week, y=Percent), color='pink')

############################# Two-Way ANOVA ####################

hist(newgovCovid$Percent, col = 'SKy Blue', main = 'Data by Education and Age', xlab = 'Percent Access')

#Normal

ages <- subset (newgovCovid, select = -Education)

educations <- subset (newgovCovid, select = -Age)

plot(lm(Percent~Education,data= educations))

plot(lm(Percent~Age,data= ages))

plot(lm(Percent, data = newgovCovid))

#NPP Linear

var.test(ages, educations, alternative = "two.sided")

eduAgeInteraction <- aov(Percent ~ Age\*Education, data = newgovCovid)

summary(eduAgeInteraction)

###################################################################################

Code Part 2-Microdata Processing Script

###################################################################################

library(foreign)

library(dplyr)

library(ggplot2)

library(stringr)

library(data.table)

#################### HCBPS Final | Microdata Processor Script ###########

customDataSet <- data.frame(Age = character(), Education = character(), Percent = numeric(), Week = numeric())

customDataSet

processor <- function(week\_num) {

setwd(sprintf("C:/Users/Thirty-Eight/Dropbox/PC/Documents/R/HCBPS-Final-Project/Datasets/HouseholdPulseMicrodata/HPS\_Week%s\_PUF\_CSV/", week\_num))

####### Change pulse2020 <-> pulse2021 ###########

govCovid = read.csv(sprintf("pulse2020\_puf\_%s.csv", week\_num))

head(govCovid)

# Age 2021-Answer Education

# 18-29 1992 2003 1) Less than high school

# 30-39 1982 1991 2) High school or GED

# 40-59 1962 1981 3) Some college/associate’s degree

# 60+ 1933 1961 4) Bachelor’s degree or higher

#-99 Ignored

#-88 Missing/Was not given questtion

govCovid = govCovid %>% filter(DELAY != -88 & DELAY != -99)

tabler <- function(x){

return(

x %>%

####### Change years 2020 -> 10002, 2021 -> 21113 ###########

group\_by(Age = cut(TBIRTH\_YEAR, breaks = c(1931, 1960, 1980, 1990, 2002)), Education =cut(EEDUC, breaks = c(0.9, 2, 3.9, 5, 5.9,7))) %>%

summarise(Percent = (sum(EEDUC[DELAY == 1])/(sum(EEDUC[DELAY <= 2]))), Week = week\_num)

)

}

tab = tabler(govCovid)

tab$Age <- as.character(tab$Age)

tab$Education <- as.character(tab$Education)

tab[tab == "(1.93e+03,1.96e+03]"] <- as.character("60+")

tab[tab == "(1.96e+03,1.98e+03]"] <- as.character("40-59")

tab[tab == "(1.98e+03,1.99e+03]"] <- as.character("30-39")

tab[tab == "(1.99e+03,2e+03]"] <- as.character("18-29")

tab[tab == "(0.9,2]"] <- as.character("Less Than/Incomplete HS")

tab[tab == "(2,3.9]"] <- as.character("High School or GED")

tab[tab == "(3.9,5]"] <- as.character("Some College/Assoc Degree")

tab[tab == "(5.9,7]"] <- as.character("Bachelor’s Degree or Higher")

tab

}

#####1-21 is 2020, 22-33 is 2021########

for (x in 1:21) {

i <- sprintf("%02d", x)

setwd(sprintf("C:/Users/Thirty-Eight/Dropbox/PC/Documents/R/HCBPS-Final-Project/Datasets/HouseholdPulseMicrodata/HPS\_Week%s\_PUF\_CSV/", i))

newdata <- processor(i)

customDataSet <- rbind(customDataSet, newdata)

}

write.csv(customDataSet,"C:/Users/Thirty-Eight/Dropbox/PC/Documents/R/HCBPS-Final-Project/Datasets/HouseholdPulseMicrodata/ageEduDataset", row.names = FALSE)

Data (In Order of Usage)

<https://data.cdc.gov/NCHS/Indicators-of-Reduced-Access-to-Care-Due-to-the-Co/xb3p-q62w>

<https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html>

<https://www.census.gov/programs-surveys/household-pulse-survey/data.html>

Nonresponse Weighting (Present in first dataset from the CB, not present in my custom-filtered datasets).

https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/2020\_HPS\_NR\_Bias\_Report-final.pdf