Final Project Milestone 2

Holly Figueroa

5/22/2021

Importing, Cleaning, Slicing, and Dicing the Data

A central dataset to be used to ascertain relationships that may exist between variables and one's willingness to vaccinate is relatively clean to start. Each variable was offered in a separate table. So each must file must be read, cleaned, and then combined. Each census table contains extra information in the first rows that will have to be skipped. Columns were selected for use and renamed.

```
library(dplyr)
# WILLINGNESS CHANGES
# Open and clean willingness to vaccinate dataframe
orig_vaccine_df <- read.csv('final_project/vaccine_will.csv', skip=1)
head(orig_vaccine_df)
```

```
##   Week          Area Total.Individual.Population.age.18. Measure.Universe
## 1   27 United States                           249170916        130998203
## 2   27       Alabama                             3717378          2007289
## 3   27        Alaska                              524925           203863
## 4   27       Arizona                             5597268          2972441
## 5   27      Arkansas                             2246527          1191733
## 6   27    California                            29939021         15594638
##     Number Margin.of.Error.... Percent Percent.Margin.of.Error....
## 1 62520073             1226564    47.7                         0.9
## 2   723497              129853    36.0                         5.9
## 3    48167               11786    23.6                         4.9
## 4  1208476              143519    40.7                         4.5
## 5   446572               76040    37.5                         6.1
## 6  9250244              500362    59.3                         3.3
```

```
# rename columns and view
orig_vaccine_df <-orig_vaccine_df%>%
  rename(week = Week,
         state = Area,
         state_adult_pop = Total.Individual.Population.age.18.,
         willing_sample = Measure.Universe,
         total_willing = Number,
         pc_MoE_willing = Percent.Margin.of.Error....,
         pc_willing = Percent)
head(orig_vaccine_df)
```

```
##   week         state state_adult_pop willing_sample total_willing
## 1   27 United States       249170916      130998203      62520073
## 2   27       Alabama         3717378        2007289        723497
## 3   27        Alaska          524925         203863         48167
## 4   27       Arizona         5597268        2972441       1208476
## 5   27      Arkansas         2246527        1191733        446572
## 6   27    California        29939021       15594638       9250244
##   Margin.of.Error.... pc_willing pc_MoE_willing
## 1             1226564       47.7            0.9
## 2              129853       36.0            5.9
## 3               11786       23.6            4.9
## 4              143519       40.7            4.5
## 5               76040       37.5            6.1
## 6              500362       59.3            3.3
```

```
# Select columns to keep and combine with others later.
vaccine_willing_percents <- orig_vaccine_df%>%
  select(state, pc_willing)
head(vaccine_willing_percents)
```

```
##           state pc_willing
## 1 United States       47.7
## 2       Alabama       36.0
## 3        Alaska       23.6
## 4       Arizona       40.7
## 5      Arkansas       37.5
## 6    California       59.3
```

```
# EXPECTED INCOME LOSS CHANGES
# read expected loss of income due to Covid data file
orig_exp_income_loss_df <- read.csv('final_project/exp_income_loss.csv', skip = 1)
head(orig_exp_income_loss_df)
```

```
##   Week          Area Total.Individual.Population.age.18. Measure.Universe
## 1   12 United States                           249170916        247851443
## 2   12       Alabama                             3717378          3672268
## 3   12        Alaska                              524925           522814
## 4   12       Arizona                             5597268          5545526
## 5   12      Arkansas                             2246527          2236004
## 6   12    California                            29939021         29848918
##     Number Margin.of.Error.... Percent Percent.Margin.of.Error....
## 1 87332680             1450190    35.2                         0.6
## 2  1034020              131027    28.2                         3.6
## 3   167304               22462    32.0                         4.3
## 4  2090442              173641    37.7                         3.1
## 5   623541               67904    27.9                         3.1
## 6 13807861              606119    46.3                         2.0
```

```
colnames(orig_exp_income_loss_df)
```

```
## [1] "Week"                                "Area"
## [3] "Total.Individual.Population.age.18." "Measure.Universe"
## [5] "Number"                              "Margin.of.Error...."
## [7] "Percent"                             "Percent.Margin.of.Error...."
```

```
# create percentages only data frames to combine later and rename columns
exp_income_loss_percents <- orig_exp_income_loss_df%>%
  select(2,7)%>%
  rename(state = Area, pc_exp_income_loss = Percent)
head(exp_income_loss_percents)
```

```
##           state pc_exp_income_loss
## 1 United States               35.2
## 2       Alabama               28.2
## 3        Alaska               32.0
## 4       Arizona               37.7
## 5      Arkansas               27.9
## 6    California               46.3
```

```
# INCOME LOST CHANGES
# read data on people with income lost due to Covid data file
orig_income_lost_df <-read.csv('final_project/income_lost.csv', skip = 1)
head(orig_income_lost_df)
```

```
##   Week          Area Total.Individual.Population.age.18. Measure.Universe
## 1   12 United States                           249170916        247855856
## 2   12       Alabama                             3717378          3686297
## 3   12        Alaska                              524925           522612
## 4   12       Arizona                             5597268          5551517
## 5   12      Arkansas                             2246527          2239763
## 6   12    California                            29939021         29862562
##      Number Margin.of.Error.... Percent Percent.Margin.of.Error....
## 1 126554411             1457948    51.1                         0.6
## 2   1689166              149103    45.8                         4.1
## 3    256356               20925    49.1                         4.0
## 4   2841364              193646    51.2                         3.4
## 5    980085               94012    43.8                         4.2
## 6  17489568              529057    58.6                         1.8
```

```
colnames(orig_income_lost_df)
```

```
## [1] "Week"                                "Area"
## [3] "Total.Individual.Population.age.18." "Measure.Universe"
## [5] "Number"                              "Margin.of.Error...."
## [7] "Percent"                             "Percent.Margin.of.Error...."
```

```
# create percentages only data frame to combine later and rename columns
income_lost_percents <-orig_exp_income_loss_df%>%
  select(2,7)%>%
  rename(state= Area, pc_income_lost = Percent)
head(income_lost_percents)
```

```
##           state pc_income_lost
## 1 United States           35.2
## 2       Alabama           28.2
## 3        Alaska           32.0
## 4       Arizona           37.7
## 5      Arkansas           27.9
## 6    California           46.3
```

```
# EXPECTED EVICTION CHANGES
# read data file on people who anticipated eviction/foreclosure
orig_exp_eviction_df <-read.csv('final_project/eviction_likely.csv', skip = 1)
head(orig_exp_eviction_df)
```

```
##   Week          Area Total.Individual.Population.age.18. Measure.Universe
## 1   28 United States                           250265449         12793569
## 2   28       Alabama                             3737637           243389
## 3   28        Alaska                              525308            32759
## 4   28       Arizona                             5753909           204778
## 5   28      Arkansas                             2264877           142503
## 6   28    California                            29807656          1631596
##    Number Margin.of.Error.... Percent Percent.Margin.of.Error....
## 1 3918446              418124    30.6                         3.0
## 2   85192               48840    35.0                        17.4
## 3   14943                6462    45.6                        15.1
## 4   67667               41341    33.0                        16.9
## 5   53699               33220    37.7                        18.7
## 6  567283              181192    34.8                        10.2
```

```
colnames(orig_exp_eviction_df)
```

```
## [1] "Week"                                "Area"
## [3] "Total.Individual.Population.age.18." "Measure.Universe"
## [5] "Number"                              "Margin.of.Error...."
## [7] "Percent"                             "Percent.Margin.of.Error...."
```

```
# create percentages only data frame to combine later and rename columns
exp_eviction_percents <- orig_exp_eviction_df%>%
  select(2,7)%>%
  rename(state = Area, pc_exp_eviction = Percent)
head(exp_eviction_percents)
```

```
##           state pc_exp_eviction
## 1 United States            30.6
## 2       Alabama            35.0
## 3        Alaska            45.6
## 4       Arizona            33.0
## 5      Arkansas            37.7
## 6    California            34.8
```

```
# DELAYED MEDICAL CARE CHANGES
# read data file on people who delayed receiving medical care due to Covid
orig_delayed_med_df <- read.csv('final_project/delayed_med.csv', skip = 1)
head(orig_delayed_med_df)
```

```
##   Week          Area Total.Individual.Population.age.18. Measure.Universe
## 1   12 United States                           249170916        222316858
## 2   12       Alabama                             3717378          3164100
## 3   12        Alaska                              524925           475598
## 4   12       Arizona                             5597268          4888731
## 5   12      Arkansas                             2246527          2012016
## 6   12    California                            29939021         25827290
##     Number Margin.of.Error.... Percent Percent.Margin.of.Error....
## 1 89159211             1395159    40.1                         0.6
## 2  1410571              138735    44.6                         4.2
## 3   211725               20359    44.5                         4.2
## 4  1901081              179989    38.9                         3.5
## 5   744708               78204    37.0                         3.9
## 6 10634751              598567    41.2                         2.2
```

```
colnames(orig_delayed_med_df)
```

```
## [1] "Week"                                "Area"
## [3] "Total.Individual.Population.age.18." "Measure.Universe"
## [5] "Number"                              "Margin.of.Error...."
## [7] "Percent"                             "Percent.Margin.of.Error...."
```

All of variables chosen to be combined are expressed as percentages of respondents that answered yes to particular questions. These were combined into a single data frame of census variables only, called “my_data”. After some changes were made, other issues were discovered. Each data frame from the census survey includes rows of data on metro cities as opposed to the state. To address this, metro cities was separated out from the state data and set aside for potential use.This was done using filter functions and slicing functions. I do not have election data at this level of measurement, however, so any analysis of city metro populations would not involve election variables. Data included from the census also has a first row including the United States as a whole. This was also taken out and set aside for potential reference. Combining data was relatively easy as all rows were organized by state name.

```
#create percentages only data frame to combine later and rename columns
delayed_med_percents <- orig_delayed_med_df%>%
  select(2,7)%>%
  rename(state = Area, pc_delayed_med = Percent)
head(delayed_med_percents)
```

```
##           state pc_delayed_med
## 1 United States           40.1
## 2       Alabama           44.6
## 3        Alaska           44.5
## 4       Arizona           38.9
## 5      Arkansas           37.0
## 6    California           41.2
```

```
# Combine data frames into one, check, and tidy
my_data <-cbind(vaccine_willing_percents,
                exp_income_loss_percents,
                income_lost_percents,
                exp_eviction_percents,
                delayed_med_percents)
colnames(my_data)
```

```
##  [1] "state"              "pc_willing"         "state"
##  [4] "pc_exp_income_loss" "state"              "pc_income_lost"
##  [7] "state"              "pc_exp_eviction"    "state"
## [10] "pc_delayed_med"
```

```
# Take out duplicate state columns
my_data<-my_data%>%
  select(-3, -5, -7, -9 )
head(my_data)
```

```
##           state pc_willing pc_exp_income_loss pc_income_lost pc_exp_eviction
## 1 United States       47.7               35.2           35.2            30.6
## 2       Alabama       36.0               28.2           28.2            35.0
## 3        Alaska       23.6               32.0           32.0            45.6
## 4       Arizona       40.7               37.7           37.7            33.0
## 5      Arkansas       37.5               27.9           27.9            37.7
## 6    California       59.3               46.3           46.3            34.8
##   pc_delayed_med
## 1           40.1
## 2           44.6
## 3           44.5
## 4           38.9
## 5           37.0
## 6           41.2
```

```
# Separate out city metro data into separate file
library(stringr)
metro_data <- my_data%>%
  filter(str_detect(state, "Metro"))%>%
  rename(location = state)
head(metro_data)
```

```
##                                          location pc_willing pc_exp_income_loss
## 1 Atlanta-Sandy Springs-Alpharetta, GA Metro Area       43.3               31.0
## 2       Boston-Cambridge-Newton, MA-NH Metro Area       67.9               30.6
## 3   Chicago-Naperville-Elgin, IL-IN-WI Metro Area       58.9               40.2
## 4      Dallas-Fort Worth-Arlington, TX Metro Area       44.2               42.8
## 5          Detroit-Warren-Dearborn, MI Metro Area       47.2               37.6
## 6 Houston-The Woodlands-Sugar Land, TX Metro Area       48.7               46.2
##   pc_income_lost pc_exp_eviction pc_delayed_med
## 1           31.0            29.0           40.3
## 2           30.6            27.1           41.3
## 3           40.2            17.1           45.5
## 4           42.8            12.7           45.3
## 5           37.6            25.8           50.0
## 6           46.2            25.0           39.8
```

```
# Slice out metro data from my_data
nrow(my_data)
```

```
## [1] 68
```

```
my_data <- slice(my_data,c(1:52))

# Separate out United States level of observation
us_census_data <- my_data%>%
  filter(state == "United States")
head(us_census_data)
```

```
##           state pc_willing pc_exp_income_loss pc_income_lost pc_exp_eviction
## 1 United States       47.7               35.2           35.2            30.6
##   pc_delayed_med
## 1           40.1
```

```
# Slice out United States level of observations so only data on 51 states remains
my_data <- slice(my_data, c(2:52))
nrow(my_data)
```

```
## [1] 51
```

The election data retrieved from Kaggle.com is given at the county level. To get state percentages of vote by any candidate the total vote for each state much be tallied. Once grouped by state, total votes per state can be gained. After that votes for Donald Trump can be filtered and totaled. Dividing votes for Donald Trump by the total votes gains a percentage of state presidential votes for Donald Trump. By doing this, the data can share measurement scale at both the state level, and as percentages of values.

```
# ELECTION DATA CHANGES

# PRESIDENTIAL DATA
election2020_state_and_county <- read.csv('final_project/president_county_candidate.csv')
head(election2020_state_and_county)
```

```
##      state            county     candidate party total_votes   won
## 1 Delaware       Kent County     Joe Biden   DEM       44552  True
## 2 Delaware       Kent County  Donald Trump   REP       41009 False
## 3 Delaware       Kent County  Jo Jorgensen   LIB        1044 False
## 4 Delaware       Kent County Howie Hawkins   GRN         420 False
## 5 Delaware New Castle County     Joe Biden   DEM      195034  True
## 6 Delaware New Castle County  Donald Trump   REP       88364 False
```

```
# Get total pres votes by state
election2020<-election2020_state_and_county%>%
  group_by(state)%>%
  summarise_at(vars(total_votes), list(total_votes = sum))

# get republican pres votes by state
rep_votes <- election2020_state_and_county%>%
  filter(candidate == "Donald Trump")%>%
  group_by(state)%>%
  summarise_at(vars(total_votes), list(trump_votes = sum))%>%
  select(trump_votes)

head(rep_votes)
```

```
## # A tibble: 6 x 1
##   trump_votes
##         <int>
## 1     1441168
## 2      189892
## 3     1661686
## 4      760647
## 5     6005961
## 6     1364607
```

```
# Combine columns: state,total presidential votes, and total presidential votes
election2020 <- cbind(election2020,rep_votes)

# Create percentage republican presidential votes column
election2020$trump_percentage <- (election2020$trump_votes / election2020$total_votes) * 100

#Rename column to specify presidential total votes
election2020 <- election2020%>%
  rename(total_pres_votes = total_votes)
head(election2020)
```

```
##        state total_pres_votes trump_votes trump_percentage
## 1    Alabama          2323304     1441168         62.03097
## 2     Alaska           391346      189892         48.52279
## 3    Arizona          3387326     1661686         49.05598
## 4   Arkansas          1219069      760647         62.39573
## 5 California         17495906     6005961         34.32781
## 6   Colorado          3256953     1364607         41.89827
```

```
# Explore table
summary(election2020)
```

```
##     state            total_pres_votes    trump_votes      trump_percentage
##  Length:51          Min.   :  276765   Min.   :  18586   Min.   : 5.397
##  Class :character   1st Qu.:  840923   1st Qu.: 473638   1st Qu.:40.814
##  Mode  :character   Median : 2148062   Median :1020280   Median :49.056
##                     Mean   : 3129573   Mean   :1462465   Mean   :49.095
##                     3rd Qu.: 3859516   3rd Qu.:1791400   3rd Qu.:57.835
##                     Max.   :17495906   Max.   :6005961   Max.   :69.936
```

After further digging, I also found issues with election data for non-presidential elections. As re-elections vary due to term limits and other reasons, it was not possible to gather complete republican election percentages at other levels. I have yet decided how best to address this, so for now, data regarding republican party dominance by state will not have the nuance of including other offices of power, such as house seats and senate seats gained during the November election of 2020.

Final Data Set

With all the variables combined the complete data set contains the following variables for analysis:

Variable Name	Variable Meaning
state	state
pc_willing	percentage of individuals planning or willing to vaccinate once able
pc_exp_income_loss	percentage of individuals that anticipated a loss of income in the next 4 weeks
pc_income_lost	percentage of households where someone had a loss in employment income in the last 7 days
pc_ex_eviction	percentage of individuals that expected eviction or home foreclosure in the next two months
trump_percentage	percentage of votes that were won by Donald Trump out of all presidential votes cast

```
# COMBINE ALL VARIABLES AT THE STATE LEVEL INTO ONE DATA FRAME

nrow(my_data)
```

```
## [1] 51
```

```
nrow(election2020)
```

```
## [1] 51
```

```
combined_data <- merge(x = my_data, y = election2020)
head(combined_data)
```

```
##        state pc_willing pc_exp_income_loss pc_income_lost pc_exp_eviction
## 1    Alabama       36.0               28.2           28.2            35.0
## 2     Alaska       23.6               32.0           32.0            45.6
## 3    Arizona       40.7               37.7           37.7            33.0
## 4   Arkansas       37.5               27.9           27.9            37.7
## 5 California       59.3               46.3           46.3            34.8
## 6   Colorado       55.6               34.5           34.5            31.5
##   pc_delayed_med total_pres_votes trump_votes trump_percentage
## 1           44.6          2323304     1441168         62.03097
## 2           44.5           391346      189892         48.52279
## 3           38.9          3387326     1661686         49.05598
## 4           37.0          1219069      760647         62.39573
## 5           41.2         17495906     6005961         34.32781
## 6           41.1          3256953     1364607         41.89827
```

```
combined_data <- combined_data%>%
  select(-trump_votes,-total_pres_votes)
# Head final table
head(combined_data)
```

```
##        state pc_willing pc_exp_income_loss pc_income_lost pc_exp_eviction
## 1    Alabama       36.0               28.2           28.2            35.0
## 2     Alaska       23.6               32.0           32.0            45.6
## 3    Arizona       40.7               37.7           37.7            33.0
## 4   Arkansas       37.5               27.9           27.9            37.7
## 5 California       59.3               46.3           46.3            34.8
## 6   Colorado       55.6               34.5           34.5            31.5
##   pc_delayed_med trump_percentage
## 1           44.6         62.03097
## 2           44.5         48.52279
## 3           38.9         49.05598
## 4           37.0         62.39573
## 5           41.2         34.32781
## 6           41.1         41.89827
```

```r
#Rename columns to avoid issues calling up the data by shared starting text
combined_data <- combined_data%>%
  rename(wiling_pc = pc_willing,
                        exp_income_loss_pc = pc_exp_income_loss,
                        income_lost_pc = pc_income_lost,
                        exp_eviction_pc = pc_exp_eviction,
                        delayed_med_pc = pc_delayed_med)

head(combined_data)
```

```
##        state wiling_pc exp_income_loss_pc income_lost_pc exp_eviction_pc
## 1    Alabama      36.0               28.2           28.2            35.0
## 2     Alaska      23.6               32.0           32.0            45.6
## 3    Arizona      40.7               37.7           37.7            33.0
## 4   Arkansas      37.5               27.9           27.9            37.7
## 5 California      59.3               46.3           46.3            34.8
## 6   Colorado      55.6               34.5           34.5            31.5
##   delayed_med_pc trump_percentage
## 1           44.6         62.03097
## 2           44.5         48.52279
## 3           38.9         49.05598
## 4           37.0         62.39573
## 5           41.2         34.32781
## 6           41.1         41.89827
```

Questions for future steps

While my initial data sets were very large, by any measure, my approach has left me with 51 rows of data. As such, any analysis is at a great disadvantage. Furthermore, any single outlier in state data will bring my small sample down again, if removed. If I can find a way to combine all the data on the county level I will have ample data. To do so would require some careful cleaning to separate counties by name and match them. I am confident the variables from the survey and the election to not share the same amount of specified counties. So my questions largely rest there, learning how to correctly separate the strings to match and merge.

What information is not self-evident?

While I have information on all the sample sizes used to obtain data, they are not included in this final data set. That may pose issues. Margins of error are given in the census data, and may be more accurate than those I would obtain on the data I see. While I initially intended to include variables of race and ethnicity, but I have chosen to not include them at this time.

What are different ways you could look at this data?

I think it would be to my benefit to try and gain values for my variables at the county level to expand my sample for analysis. Including variables for state populations in the final data frame could lend some insights. Seeing relationships between variables, such as income loss and expected eviction could lend insight into the severity of problems state populations reported.

How could you summarize your data to answer key questions?

Maximum, minimum, median, and mean values would all lend insights into the distributions and shape of the frequencies of each variable. Any regressions will benefit from summary output as well. Multiple regression analysis would be the optimal way to summarize the current data set. Sharing findings from the summary function to add and compare models as parameters are added would be appropriate. Offering the R squared and adjusted R squared statistics would also be appropriate.

What types of plots and tables will help you to illustrate the findings to your questions?

Distributions of variables, visually will be informative. Residual plots for the predictors would be useful. Correlation plots of each variable, or at least any that show significance would be illustrative. I plan to include my table of variables to better explain what each measures.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

If I gain confidence in using machine learning techniques, I may conclude they would be useful to employ. However, given the small number of variables, that may not be necessary.

Questions for future steps

Is my data set of no use at this size? Does it even conform to the assignment constraints even though they were obtained from much larger set? I may find this out on my own shortly, but I would like to know. How will changing my scope to the county level create issues for me down the line. If the data obtained regarding elections and survey variables are not from the same exact source of individuals, how do i have to adjust my analyses. Are there other measures for political influence on willingness to vaccinate that might serve well?