Data Wrangling

Anand Patel

7/31/2021

Libraries

```r
library(tidyverse)
library(magrittr)
library(ggplot2)
library(patchwork)
library(sandwich)
library(lmtest)
```

Wrangle the Trips Data

```r
# Trips data source
# https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv/data

# Define column classes based on the column description from the source.
column_classes <- c("character", "Date", "character", "character",
                    "character", "character", "integer", "integer", "integer",
                    "integer", "integer", "integer", "integer",
                    "integer", "integer", "integer", "integer",
                    "integer", "integer", "character")

# Read the input file with the trips information in the states of California,
# Oregon and Washington.
alltrips <- read.csv('Trips_by_Distance.csv', header = TRUE, colClasses= column_classes,
                     stringsAsFactors = FALSE)

# Description of alltrips
str(alltrips)
```

```
## 'data.frame':    125392 obs. of  20 variables:
##  $ Level                         : chr  "State" "State" "State" "State" ...
##  $ Date                          : Date, format: "2019-01-01" "2019-01-01" ...
##  $ State.FIPS                    : chr  "41" "06" "53" "41" ...
##  $ State.Postal.Code             : chr  "OR" "CA" "WA" "OR" ...
##  $ County.FIPS                   : chr  "" "" "" "" ...
##  $ County.Name                   : chr  "" "" "" "" ...
##  $ Population.Staying.at.Home    : int  1033821 9212440 1664296 851784 7563889 1367232 772617
4 877120 1400573 863168 ...
##  $ Population.Not.Staying.at.Home: int  3144080 30223696 5848264 3326117 31872247 6145328 317
09962 3300781 6111987 3314733 ...
##  $ Number.of.Trips               : int  12028695 111648618 21452760 14972698 139079128 262017
89 140219864 14871791 26438994 15176367 ...
##  $ Number.of.Trips..1            : int  3152087 33567702 5419053 3792335 39632283 6263947 397
18114 3747006 6327637 3782732 ...
##  $ Number.of.Trips.1.3           : int  3334320 28725797 5458958 4191488 35482817 6770840 351
92591 4191579 6879347 4340350 ...
##  $ Number.of.Trips.3.5           : int  1473205 12723636 2536653 1887100 16064280 3238321 161
19910 1857005 3231389 1904944 ...
##  $ Number.of.Trips.5.10          : int  1641658 14685031 3261166 2089635 19235040 4063990 196
78629 2046310 4064734 2076731 ...
##  $ Number.of.Trips.10.25         : int  1555258 13372854 3146301 2030027 18622807 3961430 191
29041 2012876 3988409 2016570 ...
##  $ Number.of.Trips.25.50         : int  501785 5318558 1036112 598755 6717318 1273888 6926857
606314 1293678 623561 ...
##  $ Number.of.Trips.50.100        : int  220696 2111397 332766 230847 2240304 366933 2330842 2
43698 377715 259140 ...
##  $ Number.of.Trips.100.250       : int  99421 750377 169758 104962 729924 173412 766626 11652
9 180804 121008 ...
##  $ Number.of.Trips.250.500       : int  21330 178478 32116 23120 177674 35714 180217 25299 39
726 26021 ...
##  $ Number.of.Trips...500         : int  28935 214788 59877 24429 176681 53314 177037 25175 55
555 25310 ...
##  $ Row.ID                        : chr  "41-00000-20190101" "06-00000-20190101" "53-00000-201
90101" "41-00000-20190102" ...
```

```
# Unique values in Level
unique(alltrips$Level)
```

```
## [1] "State"  "County"
```

```
# Subset the data set to only county rows as we are interested at the county level
# and by the required date range
alltripscounty <-
  alltrips %>%
  filter(Level == "County" & Date >= as.Date("2021-05-14") & Date <= as.Date("2021-05-21"))

# summary of all trips by county
summary(alltripscounty)
```

```
##     Level                Date             State.FIPS        State.Postal.Code
##  Length:1064        Min.   :2021-05-14   Length:1064        Length:1064
##  Class :character   1st Qu.:2021-05-15   Class :character   Class :character
##  Mode  :character   Median :2021-05-17   Mode  :character   Mode  :character
##                     Mean   :2021-05-17
##                     3rd Qu.:2021-05-19
##                     Max.   :2021-05-21
##
##  County.FIPS        County.Name        Population.Staying.at.Home
##  Length:1064        Length:1064        Min.   :    192
##  Class :character   Class :character   1st Qu.:   4969
##  Mode  :character   Mode  :character   Median :  16554
##                                        Mean   :  85899
##                                        3rd Qu.:  63174
##                                        Max.   :2361054
##                                        NA's   :8
##  Population.Not.Staying.at.Home Number.of.Trips    Number.of.Trips..1
##  Min.   :   1094                Min.   :    2689   Min.   :    222
##  1st Qu.:  20331                1st Qu.:   87803   1st Qu.:  19459
##  Median :  67414                Median :  272253   Median :  69376
##  Mean   : 303069                Mean   : 1158599   Mean   : 331461
##  3rd Qu.: 226627                3rd Qu.:  971100   3rd Qu.: 266408
##  Max.   :7888585                Max.   :30316438   Max.   :9561245
##  NA's   :8                      NA's   :8          NA's   :8
##  Number.of.Trips.1.3 Number.of.Trips.3.5 Number.of.Trips.5.10
##  Min.   :      0     Min.   :      0     Min.   :      0
##  1st Qu.:  21713     1st Qu.:   9326     1st Qu.:  10750
##  Median :  70059     Median :  30911     Median :  35578
##  Mean   : 281527     Mean   : 134081     Mean   : 165254
##  3rd Qu.: 244997     3rd Qu.: 109945     3rd Qu.: 138734
##  Max.   :7261507     Max.   :3647228     Max.   :4500309
##  NA's   :8           NA's   :8           NA's   :8
##  Number.of.Trips.10.25 Number.of.Trips.25.50 Number.of.Trips.50.100
##  Min.   :      0       Min.   :     51       Min.   :     0
##  1st Qu.:  14438       1st Qu.:   6256       1st Qu.:  2292
##  Median :  38832       Median :  19598       Median :  5861
##  Mean   : 160618       Mean   :  56025       Mean   : 19516
##  3rd Qu.: 141227       3rd Qu.:  45632       3rd Qu.: 15305
##  Max.   :4366899       Max.   :1391426       Max.   :453743
##  NA's   :8             NA's   :8             NA's   :8
##  Number.of.Trips.100.250 Number.of.Trips.250.500 Number.of.Trips...500
##  Min.   :     0.0        Min.   :    0.0         Min.   :    0.0
##  1st Qu.:   801.8        1st Qu.:   84.0         1st Qu.:   48.0
##  Median :  2248.0        Median :  299.5         Median :  176.5
##  Mean   :  7616.6        Mean   : 1437.3         Mean   : 1062.1
##  3rd Qu.:  7319.2        3rd Qu.: 1083.0         3rd Qu.:  762.0
##  Max.   :256650.0        Max.   :62134.0         Max.   :28076.0
##  NA's   :8               NA's   :8               NA's   :8
##     Row.ID
##  Length:1064
##  Class :character
##  Mode  :character
##
```

```
##
##
##
```

```r
# a function which takes in a column as input and provides a vector of positions with NA
# napositions <- function(df, column) {
#   navalues <- which(is.na(df$column))
#   return(navalues, df$column)
#   #return(df$column)
# }

# Check all columns and list vector of NA positions
for (i in 1:ncol(alltripscounty)){
  print(colnames(alltripscounty)[i])
  print(which(is.na(alltripscounty[,i])))
}
```

```
## [1] "Level"
## integer(0)
## [1] "Date"
## integer(0)
## [1] "State.FIPS"
## integer(0)
## [1] "State.Postal.Code"
## integer(0)
## [1] "County.FIPS"
## integer(0)
## [1] "County.Name"
## integer(0)
## [1] "Population.Staying.at.Home"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Population.Not.Staying.at.Home"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips..1"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips.1.3"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips.3.5"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips.5.10"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips.10.25"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips.25.50"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips.50.100"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips.100.250"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips.250.500"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Number.of.Trips...500"
## [1]    2  135  323  401  589  797  855 1063
## [1] "Row.ID"
## integer(0)
```

```
# show the NA rows
dropped_rows_df_trips <- alltripscounty[is.na(alltripscounty$Number.of.Trips.50.100),]

# NA positions are same in the different columns
# Drop rows with NA value based on one of the columns
alltripscounty <- alltripscounty %>%
  drop_na(Number.of.Trips.50.100)

# We can see there are no more NA values
summary(alltripscounty)
```

```
##     Level                Date             State.FIPS        State.Postal.Code
##  Length:1056        Min.   :2021-05-14   Length:1056        Length:1056
##  Class :character   1st Qu.:2021-05-15   Class :character   Class :character
##  Mode  :character   Median :2021-05-17   Mode  :character   Mode  :character
##                     Mean   :2021-05-17
##                     3rd Qu.:2021-05-19
##                     Max.   :2021-05-21
##  County.FIPS        County.Name        Population.Staying.at.Home
##  Length:1056        Length:1056        Min.   :    192
##  Class :character   Class :character   1st Qu.:   4969
##  Mode  :character   Mode  :character   Median :  16554
##                                        Mean   :  85899
##                                        3rd Qu.:  63174
##                                        Max.   :2361054
##  Population.Not.Staying.at.Home Number.of.Trips    Number.of.Trips..1
##  Min.   :   1094                Min.   :    2689   Min.   :    222
##  1st Qu.:  20331                1st Qu.:   87803   1st Qu.:  19459
##  Median :  67414                Median :  272253   Median :  69376
##  Mean   : 303069                Mean   : 1158599   Mean   : 331461
##  3rd Qu.: 226627                3rd Qu.:  971100   3rd Qu.: 266408
##  Max.   :7888585                Max.   :30316438   Max.   :9561245
##  Number.of.Trips.1.3 Number.of.Trips.3.5 Number.of.Trips.5.10
##  Min.   :      0     Min.   :      0     Min.   :      0
##  1st Qu.:  21713     1st Qu.:   9326     1st Qu.:  10750
##  Median :  70059     Median :  30911     Median :  35578
##  Mean   : 281527     Mean   : 134081     Mean   : 165254
##  3rd Qu.: 244997     3rd Qu.: 109945     3rd Qu.: 138734
##  Max.   :7261507     Max.   :3647228     Max.   :4500309
##  Number.of.Trips.10.25 Number.of.Trips.25.50 Number.of.Trips.50.100
##  Min.   :      0       Min.   :     51       Min.   :     0
##  1st Qu.:  14438       1st Qu.:   6256       1st Qu.:  2292
##  Median :  38832       Median :  19598       Median :  5861
##  Mean   : 160618       Mean   :  56025       Mean   : 19516
##  3rd Qu.: 141227       3rd Qu.:  45632       3rd Qu.: 15305
##  Max.   :4366899       Max.   :1391426       Max.   :453743
##  Number.of.Trips.100.250 Number.of.Trips.250.500 Number.of.Trips...500
##  Min.   :     0.0        Min.   :    0.0         Min.   :    0.0
##  1st Qu.:   801.8        1st Qu.:   84.0         1st Qu.:   48.0
##  Median :  2248.0        Median :  299.5         Median :  176.5
##  Mean   :  7616.6        Mean   : 1437.3         Mean   : 1062.1
##  3rd Qu.:  7319.2        3rd Qu.: 1083.0         3rd Qu.:  762.0
##  Max.   :256650.0        Max.   :62134.0         Max.   :28076.0
##     Row.ID
##  Length:1056
##  Class :character
##  Mode  :character
##
##
##
```

```
# Checking the sample size of number of counties
length(unique(alltripscounty$County.FIPS))
```

```
## [1] 132
```

```
# Creating a new column to sum the total long distance trips
# All trips greater than 50 miles are considered long distance trips
alltripscounty$Number.of.Long.Trips <-
  alltripscounty$Number.of.Trips.50.100 + alltripscounty$Number.of.Trips.100.250 +
  alltripscounty$Number.of.Trips.250.500 + alltripscounty$Number.of.Trips...500

# Check all column names
for (i in colnames(alltripscounty)){
  print(i)
}
```

```
## [1] "Level"
## [1] "Date"
## [1] "State.FIPS"
## [1] "State.Postal.Code"
## [1] "County.FIPS"
## [1] "County.Name"
## [1] "Population.Staying.at.Home"
## [1] "Population.Not.Staying.at.Home"
## [1] "Number.of.Trips"
## [1] "Number.of.Trips..1"
## [1] "Number.of.Trips.1.3"
## [1] "Number.of.Trips.3.5"
## [1] "Number.of.Trips.5.10"
## [1] "Number.of.Trips.10.25"
## [1] "Number.of.Trips.25.50"
## [1] "Number.of.Trips.50.100"
## [1] "Number.of.Trips.100.250"
## [1] "Number.of.Trips.250.500"
## [1] "Number.of.Trips...500"
## [1] "Row.ID"
## [1] "Number.of.Long.Trips"
```

```
# Creating a dataframe with only the required columns
alltripscountyfinal <- alltripscounty %>%
  select(County.FIPS,
         County.Name,
         Date,
         Number.of.Long.Trips
  )

# Taking 8-day average of the long trips by each county
# Dataset for merging with Covid Vaccination percentage
County.Trip.Covid <-
  aggregate(
    Number.of.Long.Trips ~ County.FIPS + County.Name, data = alltripscountyfinal,
    FUN=mean)

# Round the mean trips with 0 decimal places
County.Trip.Covid$Number.of.Long.Trips =
  round(County.Trip.Covid$Number.of.Long.Trips,0)

head(County.Trip.Covid)
```

```
##   County.FIPS    County.Name Number.of.Long.Trips
## 1       53001   Adams County                 6129
## 2       06001 Alameda County                89244
## 3       06005  Amador County                 4951
## 4       53003  Asotin County                 1839
## 5       41001   Baker County                 2477
## 6       41003  Benton County                 7016
```

```
nrow(County.Trip.Covid)
```

```
## [1] 132
```

We have 132 counties in our dataset thus far. We drop the following rows for Alpine County in California since no number of trips was reported.

```
# Dropped rows from this trips dataframe
dropped_rows_df_trips
```

```
##       Level       Date State.FIPS State.Postal.Code County.FIPS   County.Name
## 2    County 2021-05-14         06                CA       06003 Alpine County
## 135  County 2021-05-15         06                CA       06003 Alpine County
## 323  County 2021-05-16         06                CA       06003 Alpine County
## 401  County 2021-05-17         06                CA       06003 Alpine County
## 589  County 2021-05-18         06                CA       06003 Alpine County
## 797  County 2021-05-19         06                CA       06003 Alpine County
## 855  County 2021-05-20         06                CA       06003 Alpine County
## 1063 County 2021-05-21         06                CA       06003 Alpine County
##      Population.Staying.at.Home Population.Not.Staying.at.Home Number.of.Trips
## 2                            NA                             NA              NA
## 135                          NA                             NA              NA
## 323                          NA                             NA              NA
## 401                          NA                             NA              NA
## 589                          NA                             NA              NA
## 797                          NA                             NA              NA
## 855                          NA                             NA              NA
## 1063                         NA                             NA              NA
##      Number.of.Trips..1 Number.of.Trips.1.3 Number.of.Trips.3.5
## 2                    NA                  NA                  NA
## 135                  NA                  NA                  NA
## 323                  NA                  NA                  NA
## 401                  NA                  NA                  NA
## 589                  NA                  NA                  NA
## 797                  NA                  NA                  NA
## 855                  NA                  NA                  NA
## 1063                 NA                  NA                  NA
##      Number.of.Trips.5.10 Number.of.Trips.10.25 Number.of.Trips.25.50
## 2                      NA                    NA                    NA
## 135                    NA                    NA                    NA
## 323                    NA                    NA                    NA
## 401                    NA                    NA                    NA
## 589                    NA                    NA                    NA
## 797                    NA                    NA                    NA
## 855                    NA                    NA                    NA
## 1063                   NA                    NA                    NA
##      Number.of.Trips.50.100 Number.of.Trips.100.250 Number.of.Trips.250.500
## 2                        NA                      NA                      NA
## 135                      NA                      NA                      NA
## 323                      NA                      NA                      NA
## 401                      NA                      NA                      NA
## 589                      NA                      NA                      NA
## 797                      NA                      NA                      NA
## 855                      NA                      NA                      NA
## 1063                     NA                      NA                      NA
##      Number.of.Trips...500            Row.ID
## 2                       NA 06-06003-20210514
## 135                     NA 06-06003-20210515
## 323                     NA 06-06003-20210516
## 401                     NA 06-06003-20210517
## 589                     NA 06-06003-20210518
## 797                     NA 06-06003-20210519
## 855                     NA 06-06003-20210520
## 1063                    NA 06-06003-20210521
```

Wrangle Covid Vaccine Data

Add the covid vaccine data to our trips data.

```
## Covid Vaccine Data Processing ----------------------------------------------

# Load the Covid vaccination percentage file for CA, OR & WA
covvac <- read.csv('COVID-19_Vaccinations_CA_OR_WA.csv', header = TRUE,
                   stringsAsFactors = FALSE)

# Creating a dataframe with only the required columns
covvac <- covvac %>%
  select(Date, FIPS, Recip_County, Recip_State, Series_Complete_Pop_Pct,
         Series_Complete_Yes)

# Format the date column
covvac <- covvac %>%
  mutate(Date=as.Date(Date, format = "%m/%d/%Y"))

# Rename FIPS column to County.FIPS
covvac <- covvac %>%
  rename(County.FIPS = FIPS)

str(covvac)
```

```
## 'data.frame':    29774 obs. of  6 variables:
##  $ Date                   : Date, format: "2021-07-19" "2021-07-19" ...
##  $ County.FIPS            : chr  "53069" "41053" "06011" "06097" ...
##  $ Recip_County           : chr  "Wahkiakum County" "Polk County" "Colusa County" "Sonoma Cou
nty" ...
##  $ Recip_State            : chr  "WA" "OR" "CA" "CA" ...
##  $ Series_Complete_Pop_Pct: num  41.2 49.9 41 60.6 33 38.4 52.9 54.5 48.2 38.6 ...
##  $ Series_Complete_Yes    : int  1849 42988 8831 299623 25983 293037 68414 461108 215579 3005
4 ...
```

```
# Filter only the vaccination data on date 5/1/2021
covvac <- covvac %>%
  filter(Date == as.Date("2021-05-01"))

# Inner Join County Trip dataframe and Covid vaccination dataframe by County.FIPS
County.Trip.Covid <- dplyr::inner_join(County.Trip.Covid, covvac, by = "County.FIPS")

head(County.Trip.Covid)
```

```
##   County.FIPS    County.Name Number.of.Long.Trips       Date   Recip_County
## 1       53001   Adams County                 6129 2021-05-01   Adams County
## 2       06001 Alameda County                89244 2021-05-01 Alameda County
## 3       06005  Amador County                 4951 2021-05-01  Amador County
## 4       53003  Asotin County                 1839 2021-05-01  Asotin County
## 5       41001   Baker County                 2477 2021-05-01   Baker County
## 6       41003  Benton County                 7016 2021-05-01  Benton County
##   Recip_State Series_Complete_Pop_Pct Series_Complete_Yes
## 1          WA                    24.1                4819
## 2          CA                    37.9              633709
## 3          CA                    27.9               11076
## 4          WA                    22.9                5171
## 5          OR                    48.6                7836
## 6          OR                    35.6               33116
```

```
summary(County.Trip.Covid)
```

```
##  County.FIPS        County.Name        Number.of.Long.Trips
##  Length:132         Length:132         Min.   :   160      
##  Class :character   Class :character   1st Qu.:  3323      
##  Mode  :character   Mode  :character   Median :  8778      
##                                        Mean   : 29632      
##                                        3rd Qu.: 28154      
##                                        Max.   :596013      
##       Date            Recip_County       Recip_State       
##  Min.   :2021-05-01   Length:132         Length:132        
##  1st Qu.:2021-05-01   Class :character   Class :character  
##  Median :2021-05-01   Mode  :character   Mode  :character  
##  Mean   :2021-05-01                                        
##  3rd Qu.:2021-05-01                                        
##  Max.   :2021-05-01                                        
##  Series_Complete_Pop_Pct Series_Complete_Yes
##  Min.   : 0.00           Min.   :      0    
##  1st Qu.:24.90           1st Qu.:   7531    
##  Median :28.55           Median :  23703    
##  Mean   :28.34           Mean   : 118979    
##  3rd Qu.:33.30           3rd Qu.:  94994    
##  Max.   :51.40           Max.   :3165827
```

```
# Since 0% vaccinated data in a county will most likely mean no data available
# We exclude those rows

dropped.County.Trip.Covid <- County.Trip.Covid %>%
  filter(Series_Complete_Pop_Pct == 0.00)

# Also the vaccine data might not be reported b/c CA doesn't if the county has less than 20,000
 people
County.Trip.Covid <- County.Trip.Covid %>%
  filter(Series_Complete_Pop_Pct != 0.00)


# Calculating the county population using percent vaccinated and
# total number of vaccinated people
County.Trip.Covid$County.POP =
  County.Trip.Covid$Series_Complete_Yes*100/County.Trip.Covid$Series_Complete_Pop_Pct

# Round population to 0 decimal places
County.Trip.Covid$County.POP = round(County.Trip.Covid$County.POP,0)

# Check if both county column names are exactly the same for 125 remaining
sum(County.Trip.Covid$County.Name==County.Trip.Covid$Recip_County)==125
```

```
## [1] TRUE
```

```
# Drop date and one of the county names column after merge
County.Trip.Covid$Date <- NULL
County.Trip.Covid$County.Name <- NULL

summary(County.Trip.Covid)
```

```
##  County.FIPS        Number.of.Long.Trips Recip_County       Recip_State
##  Length:125         Min.   :   287       Length:125         Length:125
##  Class :character   1st Qu.:  4352       Class :character   Class :character
##  Mode  :character   Median : 10052       Mode  :character   Mode  :character
##                     Mean   : 31168
##                     3rd Qu.: 31466
##                     Max.   :596013
##  Series_Complete_Pop_Pct Series_Complete_Yes   County.POP
##  Min.   :18.10           Min.   :    402     Min.   :    1333
##  1st Qu.:25.20           1st Qu.:   8387     1st Qu.:   30533
##  Median :29.10           Median :  27941     Median :   87382
##  Mean   :29.92           Mean   : 125642     Mean   :  410173
##  3rd Qu.:33.40           3rd Qu.: 115177     3rd Qu.:  347190
##  Max.   :51.40           Max.   :3165827     Max.   :10050244
```

```
nrow(County.Trip.Covid)
```

```
## [1] 125
```

We now have 125 counties in our County.Trip.Covid dataset. This dataset currently includes columns for the number of long trips (over 50 miles from home), vaccination rate, and county population. We drop any counties in our dataset with 0% vaccination rate, but there are no counties. Some counties in California are missing in the vaccine dataset because the data collection effort excluded the reporting of California county vaccination rate if the population was below 20,000 people. This is why our final dataset is 125 counties, down from our previous 132.

```
dropped.County.Trip.Covid
```

```
##   County.FIPS     County.Name Number.of.Long.Trips       Date    Recip_County
## 1       06027     Inyo County                 2592 2021-05-01     Inyo County
## 2       06043 Mariposa County                 2177 2021-05-01 Mariposa County
## 3       06049    Modoc County                 2404 2021-05-01    Modoc County
## 4       06051     Mono County                 1663 2021-05-01     Mono County
## 5       06063   Plumas County                 4910 2021-05-01   Plumas County
## 6       06091   Sierra County                  160 2021-05-01   Sierra County
## 7       06105  Trinity County                 1632 2021-05-01  Trinity County
##   Recip_State Series_Complete_Pop_Pct Series_Complete_Yes
## 1          CA                       0                   0
## 2          CA                       0                   0
## 3          CA                       0                   0
## 4          CA                       0                   0
## 5          CA                       0                   0
## 6          CA                       0                   0
## 7          CA                       0                   0
```

Wrangling Median Income Data

Create a new dataframe that joins median income data for our 125 counties in our County.Trip.Covid dataset.

```
## Median Income Dataset -------------------------------------------------------

# Load the Data for County Median Income. First create the datatype for the csv else the FIPS ge
ts loaded as integer rather than character
df_income_datatype <- c("character", "character", "character",
                    "integer", "numeric", "integer", "numeric")

df_income = read.csv("Median_Income.csv", header = TRUE, colClasses= df_income_datatype,
                     stringsAsFactors = FALSE)

# Join Median Income with County.Trip.Covid dataframe
df_county_ot_cov1_3 <- dplyr::inner_join(County.Trip.Covid, df_income, by = "County.FIPS")

# Sanity Check Data
# head(df_income)
# str(df_income)
#length(unique(df_county_ot_cov1_3$County.FIPS))
```

Wrangling Party Affiliation Data

Load in, clean, and process county voting data from 2020 presidential elections. Create a party affiliation dataset that returns 1 if county voted Republican (if votes exceed those for Democratic presidential candidate), or 0 if county voted Democrat (if votes exceed those for Republican presidential candidate). Join with our dataset on county FIPS.

```
## Party Affiliation ------------------------------------------------------------

# Load the Data for Party Affiliation
df_PreferredParty_datatype <- c("character", "character", "character",
                        "integer", "integer", "integer", "numeric", "numeric", "numeric")

df_PreferredParty = read.csv("Party_Inclination_County_v2.csv", header = TRUE, colClasses= df_Pr
eferredParty_datatype,
                     stringsAsFactors = FALSE)

# Rename df_PreferredParty$county_fips column to County.FIPS
df_PreferredParty <- df_PreferredParty %>%
  rename(County.FIPS = county_fips)

str(df_PreferredParty)
```

```
## 'data.frame':    133 obs. of  10 variables:
##  $ state_name    : chr  "CA" "CA" "CA" "CA" ...
##  $ County.FIPS   : chr  "06001" "06003" "06005" "06007" ...
##  $ county_name   : chr  "Alameda County" "Alpine County" "Amador County" "Butte County" ...
##  $ votes_gop     : int  136309 244 13585 48730 16518 4554 152877 6461 61838 164464 ...
##  $ votes_dem     : int  617659 476 8153 50426 10046 3234 416386 4677 51621 193025 ...
##  $ total_votes   : int  769864 741 22302 102042 27164 7951 581230 11452 116138 364809 ...
##  $ diff          : num  -481350 -232 5432 -1696 6472 ...
##  $ per_gop       : num  0.177 0.329 0.609 0.478 0.608 ...
##  $ per_dem       : num  0.802 0.642 0.366 0.494 0.37 ...
##  $ per_point_diff: chr  "-0.62524" "-0.31309" "0.243566" "-0.016621" ...
```

```
# Create a new column for Party Affiliation DF and run the logic to identify the party inclinati
on parameter
df_PreferredParty <- df_PreferredParty %>%
  select (
    state_name, County.FIPS, county_name, votes_gop, votes_dem, total_votes, diff,
    per_gop, per_dem, per_point_diff
  ) %>%
  mutate(
    party_affiliate = case_when(
      votes_gop > votes_dem ~ "1",
      TRUE                  ~ "0"
    )
  )

# Sanity Check Data
head(df_PreferredParty)
```

```
##   state_name County.FIPS      county_name votes_gop votes_dem total_votes
## 1         CA       06001   Alameda County    136309    617659      769864
## 2         CA       06003    Alpine County       244       476         741
## 3         CA       06005    Amador County     13585      8153       22302
## 4         CA       06007     Butte County     48730     50426      102042
## 5         CA       06009 Calaveras County     16518     10046       27164
## 6         CA       06011    Colusa County      4554      3234        7951
##      diff  per_gop  per_dem per_point_diff party_affiliate
## 1 -481350 0.177056 0.802296       -0.62524               0
## 2    -232 0.329285 0.642375       -0.31309               0
## 3    5432 0.609138 0.365573       0.243566               1
## 4   -1696 0.477548 0.494169      -0.016621               0
## 5    6472 0.608084 0.369828       0.238257               1
## 6    1320 0.572758 0.406741       0.166017               1
```

```
str(df_PreferredParty)
```

```
## 'data.frame':    133 obs. of  11 variables:
##  $ state_name     : chr  "CA" "CA" "CA" "CA" ...
##  $ County.FIPS    : chr  "06001" "06003" "06005" "06007" ...
##  $ county_name    : chr  "Alameda County" "Alpine County" "Amador County" "Butte County" ...
##  $ votes_gop      : int  136309 244 13585 48730 16518 4554 152877 6461 61838 164464 ...
##  $ votes_dem      : int  617659 476 8153 50426 10046 3234 416386 4677 51621 193025 ...
##  $ total_votes    : int  769864 741 22302 102042 27164 7951 581230 11452 116138 364809 ...
##  $ diff           : num  -481350 -232 5432 -1696 6472 ...
##  $ per_gop        : num  0.177 0.329 0.609 0.478 0.608 ...
##  $ per_dem        : num  0.802 0.642 0.366 0.494 0.37 ...
##  $ per_point_diff : chr  "-0.62524" "-0.31309" "0.243566" "-0.016621" ...
##  $ party_affiliate: chr  "0" "0" "1" "0" ...
```

```
length(unique(df_PreferredParty$County.FIPS))
```

```
## [1] 133
```

```
# Join Party Affiliation with Previous Dataframe
df_county_ot_cov1_2_3 <- dplyr::inner_join(df_county_ot_cov1_3, df_PreferredParty, by = "County.
FIPS")

# Validate if any rows got dropped.
length(unique(County.Trip.Covid$County.FIPS))
```

```
## [1] 125
```

```
length(unique(df_county_ot_cov1_3$County.FIPS))
```

```
## [1] 125
```

```
length(unique(df_county_ot_cov1_2_3$County.FIPS))
```

```
## [1] 125
```

```
str(df_county_ot_cov1_2_3)
```

```
## 'data.frame':    125 obs. of  23 variables:
##  $ County.FIPS                       : chr  "53001" "06001" "06005" "53003" ...
##  $ Number.of.Long.Trips              : num  6129 89244 4951 1839 2477 ...
##  $ Recip_County                      : chr  "Adams County" "Alameda County" "Amador County"
"Asotin County" ...
##  $ Recip_State.x                     : chr  "WA" "CA" "CA" "WA" ...
##  $ Series_Complete_Pop_Pct           : num  24.1 37.9 27.9 22.9 48.6 35.6 27.2 29.4 28.2 39.7
...
##  $ Series_Complete_Yes               : int  4819 633709 11076 5171 7836 33116 55515 64397 129
50 30616 ...
##  $ County.POP                        : num  19996 1672055 39699 22581 16123 ...
##  $ Recip_State.y                     : chr  "WA" "CA" "CA" "WA" ...
##  $ Recip_County_name                 : chr  "Adams County" "Alameda County" "Amador County"
"Asotin County" ...
##  $ County_Median_Income              : int  53535 107589 62640 54776 48530 69148 72847 58394
68248 59838 ...
##  $ Income_CountyMedian_vs_StateMedian: num  0.68 1.34 0.78 0.7 0.73 1.03 0.93 0.73 0.85 0.76
...
##  $ Recip_State_Median_Income         : int  78674 80423 80423 78674 66955 66955 78674 80423 8
0423 78674 ...
##  $ unemployment_pct_2020             : num  7.3 8.8 9.1 5.2 7.2 5.6 8.2 9.2 7.6 8.4 ...
##  $ state_name                        : chr  "WA" "CA" "CA" "WA" ...
##  $ county_name                       : chr  "Adams County" "Alameda County" "Amador County"
"Asotin County" ...
##  $ votes_gop                         : int  3907 136309 13585 7319 7352 14878 60365 48730 165
18 22746 ...
##  $ votes_dem                         : int  1814 617659 8153 4250 2346 35827 38706 50426 1004
6 19349 ...
##  $ total_votes                       : int  5862 769864 22302 11951 9932 52799 103033 102042
27164 43306 ...
##  $ diff                              : num  2093 -481350 5432 3069 5006 ...
##  $ per_gop                           : num  0.666 0.177 0.609 0.612 0.74 ...
##  $ per_dem                           : num  0.309 0.802 0.366 0.356 0.236 ...
##  $ per_point_diff                    : chr  "0.357045" "-0.62524" "0.243566" "0.256799" ...
##  $ party_affiliate                   : chr  "1" "0" "1" "1" ...
```

```
# Remove all the unwanted column to create the final dataframe
df_aftercleanup <- df_county_ot_cov1_2_3 %>%
  select (County.FIPS, Number.of.Long.Trips, Recip_County, Recip_State.x, Series_Complete_Pop_Pc
t, Series_Complete_Yes,
          County.POP, County_Median_Income, Income_CountyMedian_vs_StateMedian, Recip_State_Medi
an_Income, party_affiliate, unemployment_pct_2020)

# rename party_affiliate to isRepublican
df_aftercleanup <- df_aftercleanup %>%
  rename(isRepublican = party_affiliate)

str(df_aftercleanup)
```

```
## 'data.frame':    125 obs. of  12 variables:
##  $ County.FIPS                       : chr  "53001" "06001" "06005" "53003" ...
##  $ Number.of.Long.Trips              : num  6129 89244 4951 1839 2477 ...
##  $ Recip_County                      : chr  "Adams County" "Alameda County" "Amador County"
"Asotin County" ...
##  $ Recip_State.x                     : chr  "WA" "CA" "CA" "WA" ...
##  $ Series_Complete_Pop_Pct           : num  24.1 37.9 27.9 22.9 48.6 35.6 27.2 29.4 28.2 39.7
...
##  $ Series_Complete_Yes               : int  4819 633709 11076 5171 7836 33116 55515 64397 129
50 30616 ...
##  $ County.POP                        : num  19996 1672055 39699 22581 16123 ...
##  $ County_Median_Income              : int  53535 107589 62640 54776 48530 69148 72847 58394
68248 59838 ...
##  $ Income_CountyMedian_vs_StateMedian: num  0.68 1.34 0.78 0.7 0.73 1.03 0.93 0.73 0.85 0.76
...
##  $ Recip_State_Median_Income         : int  78674 80423 80423 78674 66955 66955 78674 80423 8
0423 78674 ...
##  $ isRepublican                      : chr  "1" "0" "1" "1" ...
##  $ unemployment_pct_2020             : num  7.3 8.8 9.1 5.2 7.2 5.6 8.2 9.2 7.6 8.4 ...
```

Wrangling County Median Age data

Load, clean, and join our county median age data to our 125 counties in our study.

```
## Median Age ------------------------------------------------------------------

## Load in age data
df_AgebyCounty = read.csv("CC-EST2020-AGESEX_CA-OR-WA.csv", header = TRUE,
                             stringsAsFactors = FALSE)

# Subset the data set to only county for Year = 13 (2020) and get row for every county in CA, O
R, Wa
df_AgebyCounty <-
  df_AgebyCounty %>%
  filter(YEAR == 13)

# select only the relevant columns
df_AgebyCounty <- df_AgebyCounty %>%
  select (STNAME, CTYNAME, POPESTIMATE, AGE18PLUS_TOT, AGE65PLUS_TOT, MEDIAN_AGE_TOT)

# Rename df_AgebyCounty$CTYNAME column to Recip_County
df_AgebyCounty <- df_AgebyCounty %>%
  rename(Recip_County = CTYNAME)

summary(df_AgebyCounty)
```

```
##     STNAME          Recip_County        POPESTIMATE      AGE18PLUS_TOT
##  Length:133         Length:133         Min.   :   1119   Min.   :    907
##  Class :character   Class :character   1st Qu.:  25105   1st Qu.:  20164
##  Mode  :character   Mode  :character   Median :  82109   Median :  65166
##                                        Mean   : 385738   Mean   : 300642
##                                        3rd Qu.: 282249   3rd Qu.: 231875
##                                        Max.   :9943046   Max.   :7843569
##  AGE65PLUS_TOT     MEDIAN_AGE_TOT
##  Min.   :    287   Min.   :25.40
##  1st Qu.:   5617   1st Qu.:37.00
##  Median :  16269   Median :40.40
##  Mean   :  60252   Mean   :41.95
##  3rd Qu.:  55595   3rd Qu.:47.50
##  Max.   :1444480   Max.   :59.80
```

```
# Join Age df with our df_aftercleanup
df_aftercleanup_age_joined <- dplyr::inner_join(df_aftercleanup, df_AgebyCounty, by = "Recip_Cou
nty")

# we have duplicate rows since some counties have the same name but belong to different states.
# take out all rows that have states mismatched after the join
df_aftercleanup2 <- subset(df_aftercleanup_age_joined,
        (df_aftercleanup_age_joined$Recip_State.x == "WA" & df_aftercleanup_age_joined$STNAME ==
"Washington") |
         (df_aftercleanup_age_joined$Recip_State.x == "OR" & df_aftercleanup_age_joined$STNAME =
= "Oregon") |
          (df_aftercleanup_age_joined$Recip_State.x == "CA" & df_aftercleanup_age_joined$STNAME
 == "California")) # Apply subset function


summary(df_aftercleanup2)
```

```
##  County.FIPS        Number.of.Long.Trips Recip_County       Recip_State.x
##  Length:125         Min.   :   287       Length:125         Length:125
##  Class :character   1st Qu.:  4352       Class :character   Class :character
##  Mode  :character   Median : 10052       Mode  :character   Mode  :character
##                     Mean   : 31168
##                     3rd Qu.: 31466
##                     Max.   :596013
##  Series_Complete_Pop_Pct Series_Complete_Yes   County.POP
##  Min.   :18.10           Min.   :    402     Min.   :    1333
##  1st Qu.:25.20           1st Qu.:   8387     1st Qu.:   30533
##  Median :29.10           Median :  27941     Median :   87382
##  Mean   :29.92           Mean   : 125642     Mean   :  410173
##  3rd Qu.:33.40           3rd Qu.: 115177     3rd Qu.:  347190
##  Max.   :51.40           Max.   :3165827     Max.   :10050244
##  County_Median_Income Income_CountyMedian_vs_StateMedian
##  Min.   : 39874       Min.   :0.5700
##  1st Qu.: 54555       1st Qu.:0.7300
##  Median : 60567       Median :0.8100
##  Mean   : 65916       Mean   :0.8672
##  3rd Qu.: 72285       3rd Qu.:0.9600
##  Max.   :135234       Max.   :1.6800
##  Recip_State_Median_Income isRepublican       unemployment_pct_2020
##  Min.   :66955             Length:125         Min.   : 4.30
##  1st Qu.:66955             Class :character   1st Qu.: 7.60
##  Median :78674             Mode  :character   Median : 8.60
##  Mean   :75999                                Mean   : 8.73
##  3rd Qu.:80423                                3rd Qu.: 9.50
##  Max.   :80423                                Max.   :22.50
##     STNAME           POPESTIMATE      AGE18PLUS_TOT     AGE65PLUS_TOT
##  Length:125         Min.   :   1387   Min.   :   1188   Min.   :    457
##  Class :character   1st Qu.:  30016   1st Qu.:  22988   1st Qu.:   7515
##  Mode  :character   Median :  88053   Median :  70504   Median :  17766
##                     Mean   : 409676   Mean   : 319268   Mean   :  63910
##                     3rd Qu.: 349204   3rd Qu.: 265368   3rd Qu.:  60460
##                     Max.   :9943046   Max.   :7843569   Max.   :1444480
##  MEDIAN_AGE_TOT
##  Min.   :25.40
##  1st Qu.:36.80
##  Median :40.20
##  Mean   :41.48
##  3rd Qu.:47.10
##  Max.   :59.80
```

```
length(unique(df_aftercleanup2$County.FIPS))
```

```
## [1] 125
```

```
str(df_aftercleanup2)
```

```
## 'data.frame':    125 obs. of  17 variables:
##  $ County.FIPS                       : chr  "53001" "06001" "06005" "53003" ...
##  $ Number.of.Long.Trips              : num  6129 89244 4951 1839 2477 ...
##  $ Recip_County                      : chr  "Adams County" "Alameda County" "Amador County"
"Asotin County" ...
##  $ Recip_State.x                     : chr  "WA" "CA" "CA" "WA" ...
##  $ Series_Complete_Pop_Pct           : num  24.1 37.9 27.9 22.9 48.6 35.6 27.2 29.4 28.2 39.7
...
##  $ Series_Complete_Yes               : int  4819 633709 11076 5171 7836 33116 55515 64397 129
50 30616 ...
##  $ County.POP                        : num  19996 1672055 39699 22581 16123 ...
##  $ County_Median_Income              : int  53535 107589 62640 54776 48530 69148 72847 58394
68248 59838 ...
##  $ Income_CountyMedian_vs_StateMedian: num  0.68 1.34 0.78 0.7 0.73 1.03 0.93 0.73 0.85 0.76
...
##  $ Recip_State_Median_Income         : int  78674 80423 80423 78674 66955 66955 78674 80423 8
0423 78674 ...
##  $ isRepublican                      : chr  "1" "0" "1" "1" ...
##  $ unemployment_pct_2020             : num  7.3 8.8 9.1 5.2 7.2 5.6 8.2 9.2 7.6 8.4 ...
##  $ STNAME                            : chr  "Washington" "California" "California" "Washingto
n" ...
##  $ POPESTIMATE                       : int  20027 1662323 40083 22820 16284 93239 206426 2127
44 46308 77574 ...
##  $ AGE18PLUS_TOT                     : int  12902 1327352 34043 18259 13062 78372 152238 1698
06 38486 59675 ...
##  $ AGE65PLUS_TOT                     : int  2346 245136 11232 5617 4417 16209 32470 39082 134
02 15669 ...
##  $ MEDIAN_AGE_TOT                    : num  28.2 38.2 50.3 46 47.9 33.6 36.2 36.8 52.3 40.2
...
```

```
df_aftercleanup2$STNAME <- NULL
# rename POPESTIMATE to POPESTIMATE_2020
df_aftercleanup2 <- df_aftercleanup2 %>%
  rename(POPESTIMATE_2020 = POPESTIMATE)

str(df_aftercleanup2)
```

```
## 'data.frame':    125 obs. of  16 variables:
##  $ County.FIPS                       : chr  "53001" "06001" "06005" "53003" ...
##  $ Number.of.Long.Trips              : num  6129 89244 4951 1839 2477 ...
##  $ Recip_County                      : chr  "Adams County" "Alameda County" "Amador County"
"Asotin County" ...
##  $ Recip_State.x                     : chr  "WA" "CA" "CA" "WA" ...
##  $ Series_Complete_Pop_Pct           : num  24.1 37.9 27.9 22.9 48.6 35.6 27.2 29.4 28.2 39.7
...
##  $ Series_Complete_Yes               : int  4819 633709 11076 5171 7836 33116 55515 64397 129
50 30616 ...
##  $ County.POP                        : num  19996 1672055 39699 22581 16123 ...
##  $ County_Median_Income              : int  53535 107589 62640 54776 48530 69148 72847 58394
68248 59838 ...
##  $ Income_CountyMedian_vs_StateMedian: num  0.68 1.34 0.78 0.7 0.73 1.03 0.93 0.73 0.85 0.76
...
##  $ Recip_State_Median_Income         : int  78674 80423 80423 78674 66955 66955 78674 80423 8
0423 78674 ...
##  $ isRepublican                      : chr  "1" "0" "1" "1" ...
##  $ unemployment_pct_2020             : num  7.3 8.8 9.1 5.2 7.2 5.6 8.2 9.2 7.6 8.4 ...
##  $ POPESTIMATE_2020                  : int  20027 1662323 40083 22820 16284 93239 206426 2127
44 46308 77574 ...
##  $ AGE18PLUS_TOT                     : int  12902 1327352 34043 18259 13062 78372 152238 1698
06 38486 59675 ...
##  $ AGE65PLUS_TOT                     : int  2346 245136 11232 5617 4417 16209 32470 39082 134
02 15669 ...
##  $ MEDIAN_AGE_TOT                    : num  28.2 38.2 50.3 46 47.9 33.6 36.2 36.8 52.3 40.2
...
```

```
nrow(df_aftercleanup2)
```

```
## [1] 125
```

Save the final data out to CSV

Our final columns available

```
str(df_aftercleanup2)
```

```
## 'data.frame':    125 obs. of  16 variables:
##  $ County.FIPS                       : chr  "53001" "06001" "06005" "53003" ...
##  $ Number.of.Long.Trips              : num  6129 89244 4951 1839 2477 ...
##  $ Recip_County                      : chr  "Adams County" "Alameda County" "Amador County"
"Asotin County" ...
##  $ Recip_State.x                     : chr  "WA" "CA" "CA" "WA" ...
##  $ Series_Complete_Pop_Pct           : num  24.1 37.9 27.9 22.9 48.6 35.6 27.2 29.4 28.2 39.7
...
##  $ Series_Complete_Yes               : int  4819 633709 11076 5171 7836 33116 55515 64397 129
50 30616 ...
##  $ County.POP                        : num  19996 1672055 39699 22581 16123 ...
##  $ County_Median_Income              : int  53535 107589 62640 54776 48530 69148 72847 58394
68248 59838 ...
##  $ Income_CountyMedian_vs_StateMedian: num  0.68 1.34 0.78 0.7 0.73 1.03 0.93 0.73 0.85 0.76
...
##  $ Recip_State_Median_Income         : int  78674 80423 80423 78674 66955 66955 78674 80423 8
0423 78674 ...
##  $ isRepublican                      : chr  "1" "0" "1" "1" ...
##  $ unemployment_pct_2020             : num  7.3 8.8 9.1 5.2 7.2 5.6 8.2 9.2 7.6 8.4 ...
##  $ POPESTIMATE_2020                  : int  20027 1662323 40083 22820 16284 93239 206426 2127
44 46308 77574 ...
##  $ AGE18PLUS_TOT                     : int  12902 1327352 34043 18259 13062 78372 152238 1698
06 38486 59675 ...
##  $ AGE65PLUS_TOT                     : int  2346 245136 11232 5617 4417 16209 32470 39082 134
02 15669 ...
##  $ MEDIAN_AGE_TOT                    : num  28.2 38.2 50.3 46 47.9 33.6 36.2 36.8 52.3 40.2
...
```

Our trips dataset provide `County.FIPS` , `Number.of.Long.Trips` .

Our vaccine rate dataset provide `Recip_County` , `Recip_State.x` , `Series_Complete_Pop_Pct` , `Series_Complete_Yes` , and our estimate for 2021 county population `County.POP` .

Our 2020 county median income datset provides `County_Median_Income` , `Income_CountyMedian_vs_StateMedian` , `Recip_State_Median_Income` , `unemployment_pct_2020` .

Our 2020 Presidential Election County Level dataset allows us to compute our `isRepublican` .

Our 2020 county age dataset gives us `POPESTIMATE_2020` , `AGE18PLUS_TOT` , `AGE65PLUS_TOT` , `MEDIAN_AGE_TOT` .

Save the file

```
write.csv(df_aftercleanup2,'final_data_v1.csv')
```