

COVID Vaccination Project

Adrian Cornejo and Victor Guillera

11/4/2021

Milestone 2

Section 1

This data set will relate to our question if being a homeowner status and available housing impacts vaccination rates per CA county. While not a perfect representation of SES of cases, this data set will help us elucidate if there is a trend via linear regression between the homeowner-related to renter rates and the vaccination rate per county.

We have two data sets, one is of COVID-19 vaccination progress throughout CA provided by the CDPH following vaccine dosage by zip code from 1/5/21 to about present 9/14/21, and the other is demographic info in CA across all counties from the US census data for 2014 to 2019. Below, we read in our libraries, data sets, explore our data sets, and then proceed to clean the data sets.

```
library(readr)
library(knitr)
library(tibble)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(stringr)
countydem <- read.csv("ca_county_demographics.csv")
covidvax <- read.csv("cov_vax_admin.csv")
```

To import the csv files, we used `read_csv` from base R and will inspect our imported files first before we clean our data set. Libraries we loaded in were `readr`, `dplyr`, and `tibble`. We are going to pull variables “persons_fully_vaccinated”, “persons_partially_vaccinated”, “zip_code_tabulation_area”, “age12_plus_population”, and “county” from the covidvax csv file. We will pull “name”, “renter_occ”, “owner_occ”, “hse_units” and “vacant” variables.

```
str(covidvax)
head(covidvax)
nrow(covidvax)
ncol(covidvax)
typeof(covidvax)
class(covidvax)
```

```
distinct(covidvax, as_of_date, .keep_all = F)
```

The structure of the “cov_vax_admin” file is a data frame with 11 columns and 65,268 rows. The data frame contains integer, number, and character data types and contains a redacted column that is inaccessible. There are also a total of 27 unique dates, each 7 weeks apart (weekly results).

```
str(countydem)
head(countydem)
nrow(countydem)
ncol(countydem)
typeof(countydem)
class(countydem)
```

The structure of “countydem” reveals a data frame containing 23 columns and 58 rows with integer, character and number data types.

Cleaning the data by removing unnecessary columns, rearranging the data, and removing “N/A” data points.

```
#COVID Vaccination Rates Dataset
covidvax2 <- select(covidvax, "county", "zip_code_tabulation_area", "as_of_date",
                    "age12_plus_population", "persons_fully_vaccinated",
                    "persons_partially_vaccinated")

names(covidvax2) <- c("county", "zip_code", "date", "total_pop_over_12",
                     "fully_vaccinated", "partially_vaccinated")

covidvax2[is.na(covidvax2)] <- 0
```

Using the function `str()` we see that the selected variables in `covidvax2` are `chr`, `int`, `chr`, `num`, `num`. We may need to convert the dates to numeric using the `as.numeric()` function to determine and compare vaccination rate. All other column types should be fine as is. Converting NA cells to 0 we allow ourselves to manipulate and describe entire numeric columns.

```
#Descriptions
#distinct(covidvax2, county, .keep_all = F)
#distinct(covidvax2, date, .keep_all = F)
#distinct(covidvax2, zip_code, .keep_all = F)

covidvax2 %>% group_by(zip_code) %>%
  summarise(fullyvax_range = range(fully_vaccinated),
            partiallyvax_range = range(partially_vaccinated))
```

‘`summarise()`’ has grouped output by ‘`zip_code`’. You can override using the ‘`.groups`’ argument.

```
## # A tibble: 3,528 x 3
## # Groups:   zip_code [1,764]
##   zip_code fullyvax_range partiallyvax_range
##   <int>      <dbl>          <dbl>
## 1  90001          0            0
## 2  90001     36923        18224
## 3  90002          0            0
## 4  90002     24049         6320
## 5  90003          0            0
## 6  90003     33150         8899
## 7  90004         19           893
## 8  90004     37672        10232
## 9  90005         14           402
## 10 90005     22653         6785
## # ... with 3,518 more rows
```

```
covidvax2 %>%
  summarise(total_pop_range = range(total_pop_over_12),
            total_pop_median = median(total_pop_over_12))
```

```
##   total_pop_range total_pop_median
## 1          0.0      13685.1
## 2      88556.7      13685.1
```

```
#County Demographics Dataset
```

```
countydem2 <- select(countydem, "name", "renter_occ", "owner_occ",  
  "vacant", "hse_units")
```

```
names(countydem2) <-c("county_name", "renters", "owners",  
  "vacant_housing", "total_housing_units")
```

```
countydem2[is.na(countydem2)] <- 0
```

Using the function `str()` we see that the selected variables in `countydem2` are `chr`, `int`, `int`, `int`, `int`. All column types should be fine as is for our study question. Converting NA cells to 0 we allow ourselves to manipulate and describe entire numeric columns. We also changed column names to allow for easier/obvious verbiage.

```
#Descriptions
```

```
#distinct(covidvax2, county, .keep_all = F)
```

```
#distinct(covidvax2, date, .keep_all = F)
```

```
#distinct(covidvax2, zip_code, .keep_all = F)
```

```
covidvax2 %>% group_by(zip_code) %>%  
  summarise(fullyvax_range = range(fully_vaccinated),  
    partiallyvax_range = range(partially_vaccinated))
```

Population Range and Median for COVID vaccinations for those over 12 years old.

```
covidvax2 %>% summarise(total_pop_range = range(total_pop_over_12),  
                        total_pop_median = median(total_pop_over_12))
```

Using the function str() we see that the selected variables in countydem2 are chr, int, int, int, int All column types should be fine as is for our study question Converting NA cells to 0 we allow ourselves to manipulate and describe entire numeric columns. Changed column names to easier/obvious verbiage

#Descriptions

```
#distinct(countydem2, county_name, .keep_all = F)
```

```
countydem2 %>% summarise(renters_mean = mean(renters),  
                        owners_mean = mean(owners),  
                        vacant_mean = mean(vacant_housing),  
                        total_mean = mean(total_housing_units))
```

```
##   renters_mean owners_mean vacant_mean total_mean  
## 1      95553.91    121299.5    19010.05    235863.5
```

```
countydem2 %>% summarise(renters_range = range(renters),  
                        owners_range = range(owners),  
                        vacant_range = range(vacant_housing),  
                        total_range = range(total_housing_units))
```

```
##   renters_range owners_range vacant_range total_range  
## 1           140           357           827          1760  
## 2        1696455        1544749        203872        3445076
```

Below, is the minimum, first quartile, median, third quartile, and max values.

```
fivenum(covidvax2$total_pop_over_12)
```

```
## [1]      0.00  1346.80 13685.10 31762.15 88556.70
```

```
fivenum(covidvax2$fully_vaccinated)
```

```
## [1]      0    189   1867 11637 67594
```

```
fivenum(covidvax2$partially_vaccinated)
```

```
## [1]      0     93   1033   3028 23195
```

```
#Descriptions
```

```
#We get the mean & range for all numeric and integer vectors in countydem2 below
```

```
distinct(countydem2, county_name, .keep_all = F)
```

```
countydem2 %>%  
  summarise(renters_mean = mean(renters),  
            owners_mean = mean(owners),  
            vacant_mean = mean(vacant_housing),  
            total_mean = mean(total_housing_units))
```

```
##   renters_mean owners_mean vacant_mean total_mean  
## 1    95553.91    121299.5    19010.05    235863.5
```

```
countydem2 %>%  
  summarise(renters_range = range(renters),  
            owners_range = range(owners),  
            vacant_range = range(vacant_housing),  
            total_range = range(total_housing_units))
```

```
##   renters_range owners_range vacant_range total_range  
## 1         140         357         827         1760  
## 2    1696455    1544749    203872    3445076
```

Below, we get the five number summary for data of interest in countydem2.

```
fivenum(countydem2$renters)
```

```
## [1]      140      5878     25140     87737    1696455
```

```
fivenum(countydem2$owners)
```

```
## [1]      357     12629     39306    123646    1544749
```

```
fivenum(countydem2$vacant_housing)
```

```
## [1]      827.0     3328.0     8580.5    18747.0    203872.0
```

```
fivenum(countydem2$total_housing_units)
```

```
## [1]     1760.0     23910.0     76183.5    233755.0    3445076.0
```

Milestone 3

#Covid Vaccination by County, Data Manipulation (covidvax files)

```
unique(covidvax2$county)
```

```
## [1] "ORANGE" "SAN BERNARDINO" "IMPERIAL" "RIVERSIDE"
## [5] "LOS ANGELES" "SAN DIEGO" "TRINITY" "SAN FRANCISCO"
## [9] "TULARE" "MARIN" "CONTRA COSTA" "KERN"
## [13] "VENTURA" "SANTA BARBARA" "SAN MATEO" "SOLANO"
## [17] "FRESNO" "MONTEREY" "SONOMA" "NAPA"
## [21] "ALAMEDA" "MADERA" "KINGS" "INYO"
## [25] "SACRAMENTO" "SAN LUIS OBISPO" "SANTA CLARA" "EL DORADO"
## [29] "GLENN" "YUBA" "BUTTE" "PLACER"
## [33] "AMADOR" "SUTTER" "MONO" "LAKE"
## [37] "YOLO" "HUMBOLDT" "SAN JOAQUIN" "TUOLUMNE"
## [41] "CALAVERAS" "SHASTA" "O" "SISKIYOU"
## [45] "LASSEN" "MERCED" "SANTA CRUZ" "SAN BENITO"
## [49] "MODOC" "STANISLAUS" "MENDOCINO" "SIERRA"
## [53] "TEHAMA" "PLUMAS" "MARIPOSA" "DEL NORTE"
## [57] "NEVADA" "ALPINE" "COLUSA"
```

```
covidvax2 <- covidvax2 %>% filter(county!="O")
```

```
unique(covidvax2$date)
```

```
## [1] "2021-01-05" "2021-01-12" "2021-01-19" "2021-01-26" "2021-02-02"
## [6] "2021-02-09" "2021-02-16" "2021-02-23" "2021-03-02" "2021-03-09"
## [11] "2021-03-16" "2021-03-23" "2021-03-30" "2021-04-06" "2021-04-13"
## [16] "2021-04-20" "2021-04-27" "2021-05-04" "2021-05-11" "2021-05-18"
## [21] "2021-05-25" "2021-06-01" "2021-06-08" "2021-06-15" "2021-06-22"
## [26] "2021-06-29" "2021-07-06" "2021-07-13" "2021-07-20" "2021-07-27"
## [31] "2021-08-03" "2021-08-10" "2021-08-17" "2021-08-24" "2021-08-31"
## [36] "2021-09-07" "2021-09-14"
```

```
covidvax3 <- covidvax2 %>%
  mutate(month = case_when(
    str_detect(date, "-01-") == T ~ "January",
    str_detect(date, "-02-") == T ~ "February",
    str_detect(date, "-03-") == T ~ "March",
    str_detect(date, "-04-") == T ~ "April",
    str_detect(date, "-05-") == T ~ "May",
    str_detect(date, "-06-") == T ~ "June",
    str_detect(date, "-07-") == T ~ "July",
    str_detect(date, "-08-") == T ~ "August",
    str_detect(date, "-09-") == T ~ "September"
  ))
```

```
unique(covidvax3$month)
```

```
## [1] "January" "February" "March" "April" "May" "June"
## [7] "July" "August" "September"
```

```
covidvax4 <- covidvax3 %>%
  group_by(county, month) %>%
  mutate(county_total_pop_over12 = sum(total_pop_over_12)) %>%
```



```

mutate(county_fully_vaccinated = sum(fully_vaccinated)) %>%
mutate(county_partially_vaccinated = sum(partially_vaccinated))

covidvax4 <- covidvax4 %>%
  select(-"zip_code",-"total_pop_over_12",-"fully_vaccinated",
        -"partially_vaccinated")

#Removing duplicate columns after combining the zip codes for each county
#and making column totals for each . Then we check to ensure we have 58 counties
#for each month.

covidvax5 <- covidvax4 %>% distinct(county,month, .keep_all=T)

covidvax5 %>% group_by(month) %>% summarise(n())

## # A tibble: 9 x 2
##   month      'n()'
##   <chr>      <int>
## 1 April        58
## 2 August       58
## 3 February    58
## 4 January     58
## 5 July        58
## 6 June        58
## 7 March       58
## 8 May        58
## 9 September   58

covidvax5 <- covidvax5 %>%
mutate(county_percent_fullvax =
(county_fully_vaccinated)/(county_total_pop_over12)) %>%
mutate(county_percent_partialvax =
(county_partially_vaccinated)/(county_total_pop_over12))

#Checking that our percentages make sense.
summary(covidvax5$county_percent_fullvax)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000 0.1246 0.3662 0.3440 0.5118 0.8609

summary(covidvax5$county_percent_partialvax)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.06396 0.08420 0.09282 0.12036 0.24445

```

#County Demographic Data Manipulation (countydem files)

```
countydem3 <- countydem2 %>%  
  mutate(total_in_use_housing = total_housing_units - vacant_housing) %>%  
  mutate(renter_housing_prop = (renters/(total_in_use_housing))) %>%  
  mutate(owner_housing_prop = (owners/(total_in_use_housing))) %>%  
  mutate(renter_owner_ratio = renters/owners)
```

Data Dictionary

```
str(countydem3)  
str(covidvax5)
```

County Demographics Data Set:

Name: county_name Data type: character Description: A list of all 58 counties in California.

Name:renters Data type: integer Description: Number of renters in each county in California.

Name:owners Data type: integer Description: Number of owners in each county in California.

Name:vacant_housing Data type: integer Description: Number of empty homes in each county in California.

Name:total_housing_units Data type: integer Description: Number of total homes in each county in California.

Name:total_in_use_housing Data type: integer Description: Number of total homes currently in use in each county in California.

Name:renter_housing_prop Data type: number Description: Proportion of renters living in occupied homes in each county in California.

Name:owner_housing_prop Data type: number Description: Proportion of home owners living in occupied homes in each county in California.

Name:renter_owner_ratio Data type: number Description: Renter to Home owner ratio for each county in California.

Covid Vaccination Rate Data Set:

Name:county Data type: character Description: A list containing the 58 counties in California.

Name:date Data type: character Description:The dates when the data was compiled, ranging from January 5, 2021 to September 14, 2021.

Name:month Data type: character Description: The month in which the testing took place. Created to more easily combine data from specific dates into monthly variables for each county in the future.

Name:county_total_pop_over12 Data type: number Description: Total population over the age of 12 years old in each county in California for a specific month.

Name:county_fully_vaccinated Data type: number Description: Total number of people fully vaccinated over the age of 12 in each county in California for a specific month.

Name:county_fully_vaccinated Data type: number Description: Total number of people fully vaccinated over the age of 12 in each county in California for a specific month.

Name:county_partially_vaccinated Data type: number Description: Total number of people partially vaccinated over the age of 12 in each county in California for a specific month.

Name:county_percent_fullvax Data type: number Description: Percentage of people fully vaccinated over the age of 12 over the total population over 12 years old in each county in California for a specific month.

Name:county_percent_partialvax Data type: number Description: Percentage of people partially vaccinated over the age of 12 over the total population over 12 years old in each county in California for a specific month.

Descriptive Statistics Tables

```

countydem4 <- countydem3 %>%
  summarise(total_in_use_housing_range = summary(total_in_use_housing),
            renter_housing_ratio_range = summary(renter_housing_prop),
            owner_housing_ratio_range = summary(owner_housing_prop),
            renter_owner_ratio_range = summary(renter_owner_ratio))

rownames(countydem4) <- c("Min", "1st Quartile", "Median", "Mean", "3rd Quartile", "Max")

countydem5 <- countydem3 %>%
  summarise(total_renters = sum(renters),
            total_owners = sum(owners))

rownames(countydem5) <- c("Total Count")

library(kableExtra)

##
## Attaching package: 'kableExtra'
## The following object is masked from 'package:dplyr':
##
##   group_rows

countydem_kable <- kable(countydem4, longtable=T, booktabs=T,
                        col.names = c("Total in Use Housing", "Renter/Housing Ratio",
                                      "Owner/Housing Ratio", "Renter/Owner Ratio"),
                        caption = "Summary of CA Housing Across All Counties in 2021")
countydem_kable

```

Table 1: Summary of CA Housing Across All Counties in 2021

	Total in Use Housing	Renter/Housing Ratio	Owner/Housing Ratio	Renter/Owner Ratio
Min	497.00	0.2311765	0.3575537	0.3006887
1st Quartile	19040.75	0.3426035	0.5753113	0.5211578
Median	70284.50	0.3854496	0.6145504	0.6272347
Mean	216853.41	0.3833366	0.6166634	0.6472471
3rd Quartile	207711.50	0.4246887	0.6573965	0.7381916
Max	3241204.00	0.6424463	0.7688235	1.7967828

```

countydem_kable2 <- kable(countydem5, longtable=T, booktabs=T,
                        col.names = c("Renters", "Owners"),
                        caption = "Total Number of Renters and Owners in CA in 2021")
countydem_kable2

```

Table 2: Total Number of Renters and Owners in CA in 2021

	Renters	Owners
Total Count	5542127	7035371

*#does not make sense to visualize covidvax dataset in a kable, will update visualization
#with graph using ggplot in milestone 4*