COVID-19 Linear Regression Model for Sonoma County

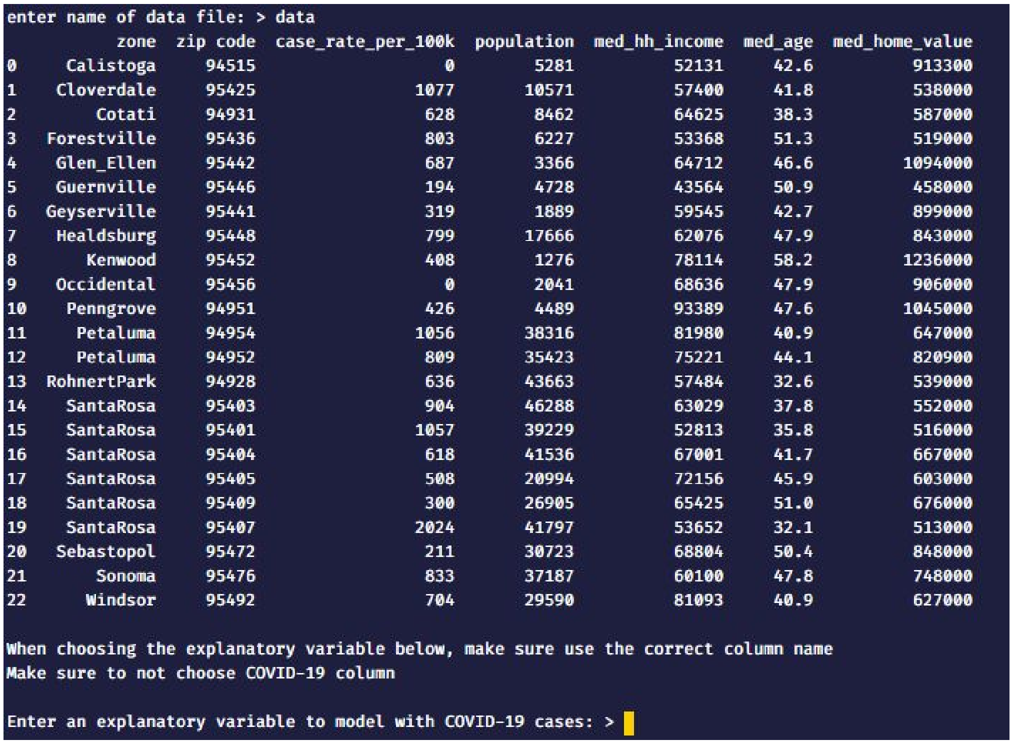
By Paris Osuch

Abstract

The research in this document can conclude that the population may have a strong correlation with covid cases per 100k people.

Introduction and Methods

This paper explores the correlation between a multitude of variables vs. COVID-19 cases per 100k people. The explanatory variables being explored in this model include: population, median household income, median age, and median house value. To determine the correlation between these variables, we were tasked to create a linear regression model. To create these linear regression models I wrote a program in python using Pandas, Matplotlib, and the standard python library. Both Pandas and Matplotlib are powerful data analysis tools that are used to create the tables and plot the data in the tables. To help me with the task of finding the standard deviation, mean, correlation, and plot the best fit line; I developed a number of functions that fit accordingly to the prior tasks at hand. These python files can be found in my git repository at: https://github.com/parisosuch-dev/COVID-19-Correlation-Model . In the stats.py file, you will find the stats functions I created for my analysis. In the main.py file, you will find the plotting and analysis portion of the research. In the beginning of the program, the user is prompted to enter the name of an explanatory variable they would like to find correlated with COVID-19 cases. See *Figure 1*



```
enter name of data file: > data
          zone  zip code  case_rate_per_100k  population  med_hh_income  med_age  med_home_value
0    Calistoga     94515                   0        5281          52131     42.6          913300
1   Cloverdale     95425                1077       10571          57400     41.8          538000
2       Cotati     94931                 628        8462          64625     38.3          587000
3  Forestville     95436                 803        6227          53368     51.3          519000
4    Glen_Ellen    95442                 687        3366          64712     46.6         1094000
5    Guernville    95446                 194        4728          43564     50.9          458000
6   Geyserville    95441                 319        1889          59545     42.7          899000
7    Healdsburg    95448                 799       17666          62076     47.9          843000
8      Kenwood     95452                 408        1276          78114     58.2         1236000
9    Occidental    95456                   0        2041          68636     47.9          906000
10    Penngrove    94951                 426        4489          93389     47.6         1045000
11     Petaluma    94954                1056       38316          81980     40.9          647000
12     Petaluma    94952                 809       35423          75221     44.1          820900
13  RohnertPark    94928                 636       43663          57484     32.6          539000
14    SantaRosa    95403                 904       46288          63029     37.8          552000
15    SantaRosa    95401                1057       39229          52813     35.8          516000
16    SantaRosa    95404                 618       41536          67001     41.7          667000
17    SantaRosa    95405                 508       20994          72156     45.9          603000
18    SantaRosa    95409                 300       26905          65425     51.0          676000
19    SantaRosa    95407                2024       41797          53652     32.1          513000
20   Sebastopol    95472                 211       30723          68804     50.4          848000
21       Sonoma    95476                 833       37187          60100     47.8          748000
22      Windsor    95492                 704       29590          81093     40.9          627000

When choosing the explanatory variable below, make sure use the correct column name
Make sure to not choose COVID-19 column

Enter an explanatory variable to model with COVID-19 cases: >
```

Figure 1

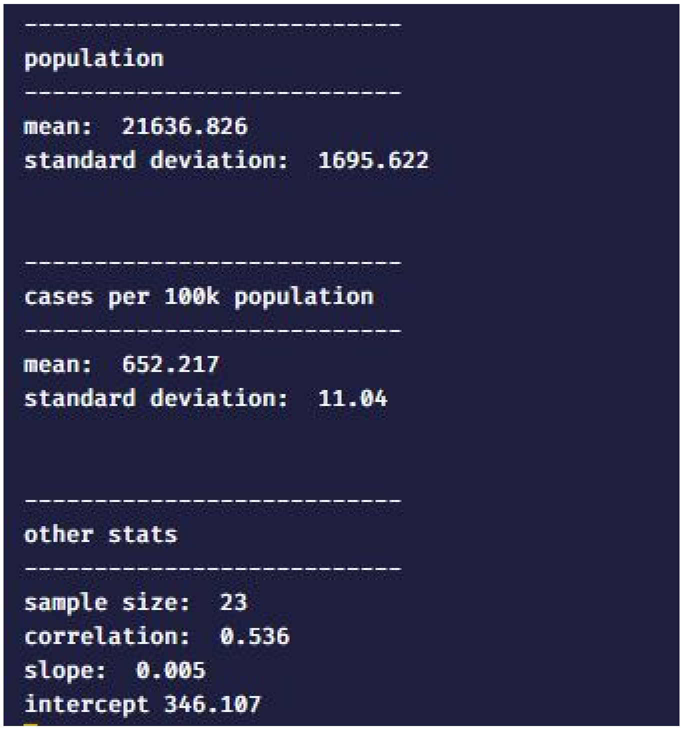
Initial Model

The data set being used is from Sonoma county official website as of 8/19/20 and from the website bestplaces.net, with the variables; zone, zip code, cases per 100k people, population, median household income, median age, and median home value. The model being created is a linear regression model analyzing COVID-19 cases vs. an explanatory variable. The equation for the best fit line is: *Y* = *a* +*bX*. The intercept of the line is *a* and the slope of the line is *b*. The correlation coefficient assesses the relationship between two variables and how related they are. However, it can not be used for causation. The formula for the correlation coefficient is:

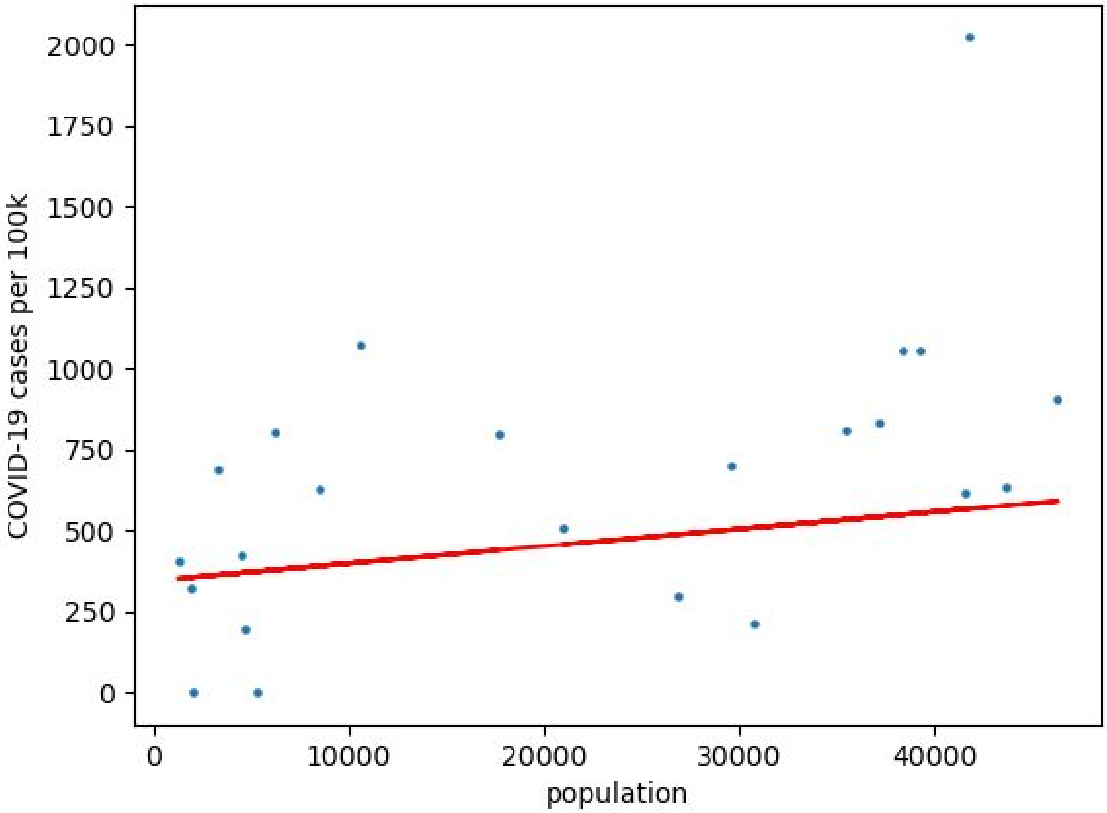
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

If *r* is less than 0, it has a negative correlation. If *r* is greater than 0, it has a positive correlation. As you can see from the model created below, the population has a 0.536 positive correlation with COVID-19 cases.

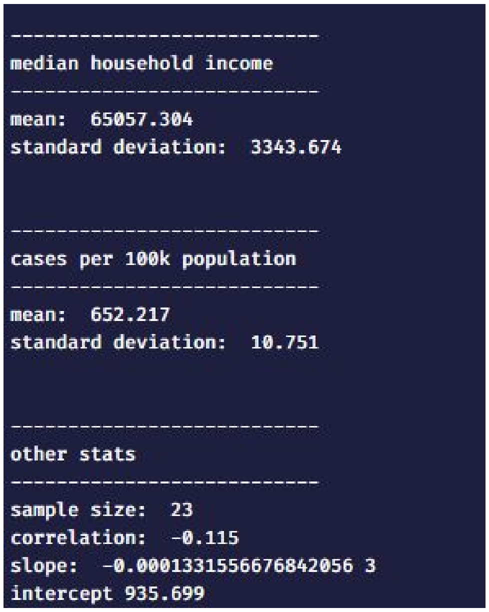
Initial Model - COVID-19 Cases vs Population



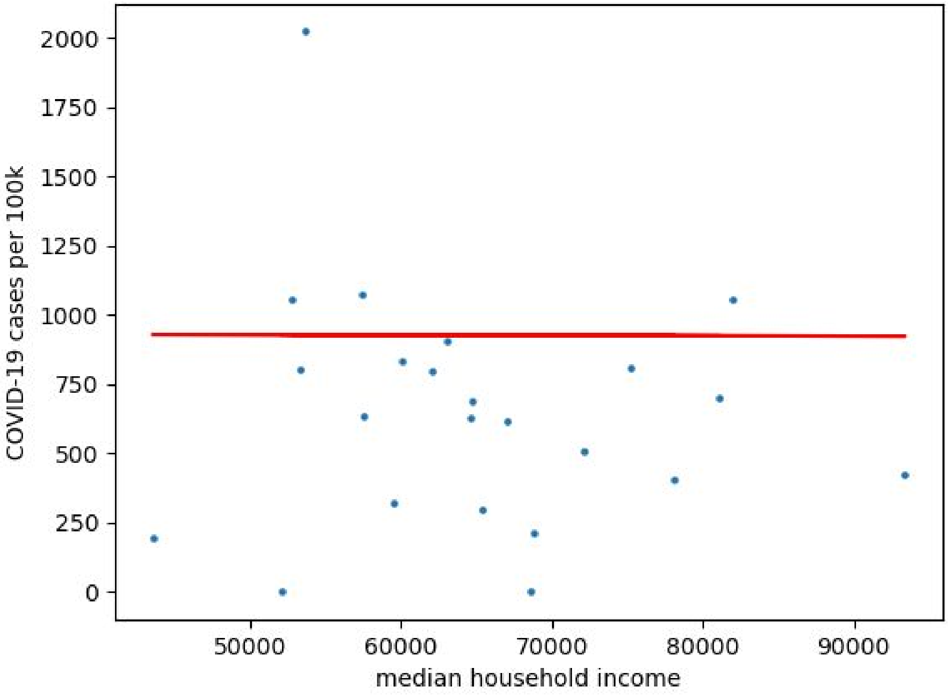
```
---------------------------
population
---------------------------
mean:  21636.826
standard deviation:  1695.622


---------------------------
cases per 100k population
---------------------------
mean:  652.217
standard deviation:  11.04


---------------------------
other stats
---------------------------
sample size:  23
correlation:  0.536
slope:  0.005
intercept 346.107
```


COVID-19 cases per 100k
2000
1750
1500
1250
1000
750
500
250
0
0
10000
20000
30000
40000
population

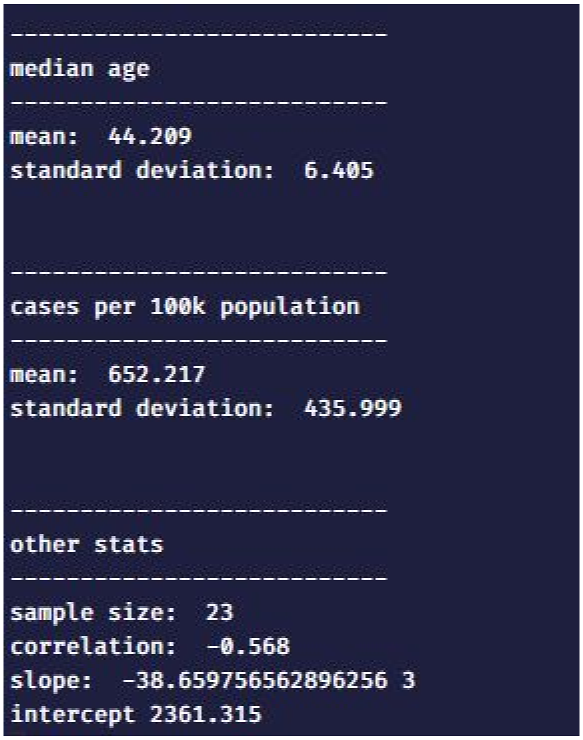
Initial Model - COVID-19 Cases vs Median Household Income



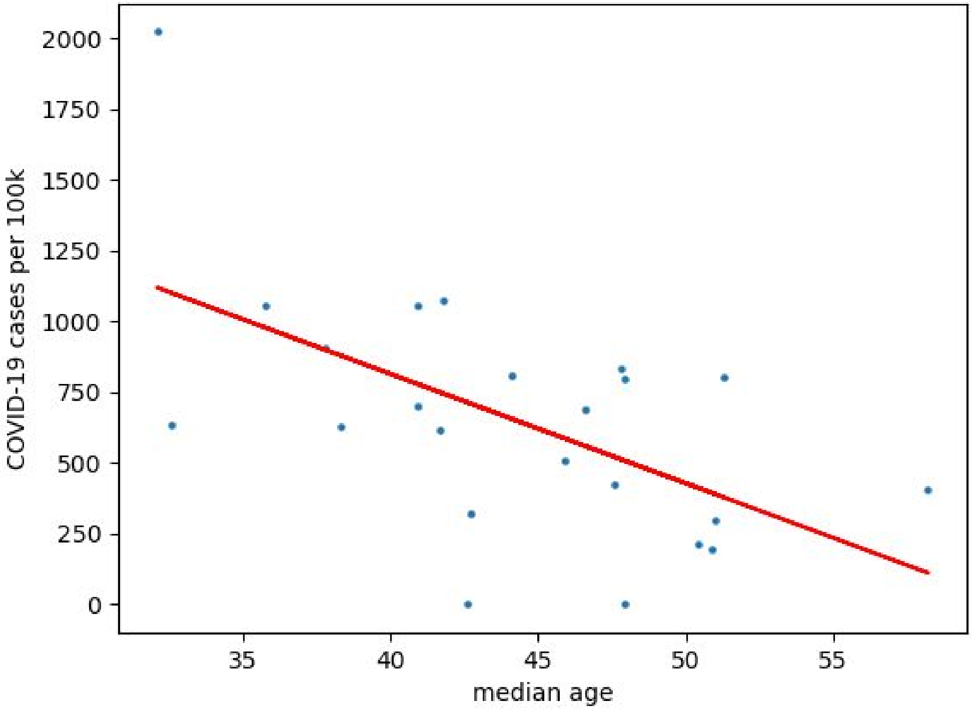
```
---------------------------
median household income
---------------------------
mean:  65057.304
standard deviation:  3343.674


---------------------------
cases per 100k population
---------------------------
mean:  652.217
standard deviation:  10.751


---------------------------
other stats
---------------------------
sample size:  23
correlation:  -0.115
slope:  -0.0001331556676842056 3
intercept 935.699
```



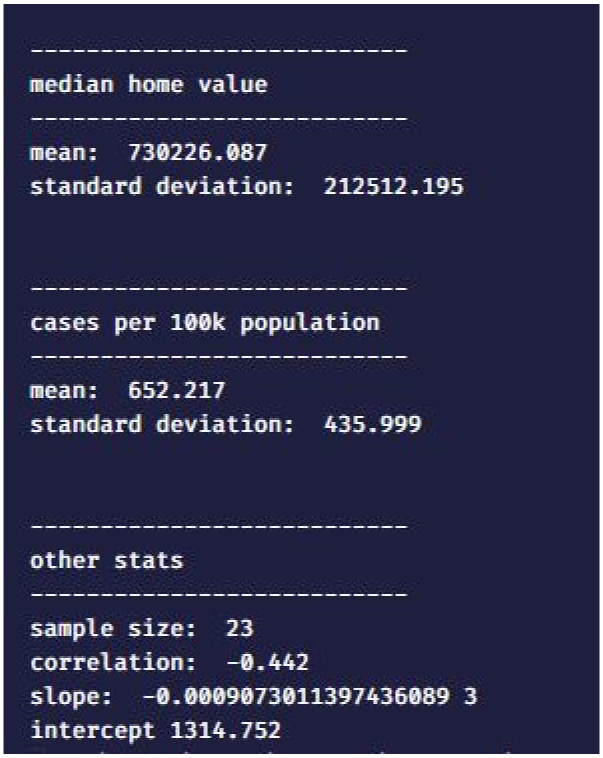
Initial Model - COVID-19 Cases vs Median Age



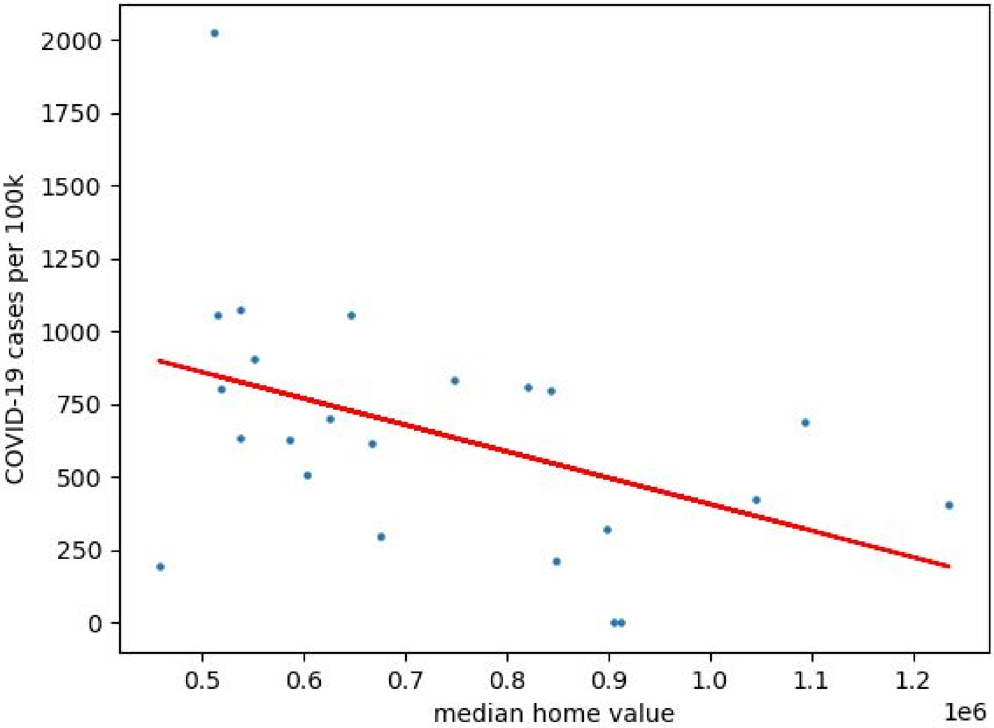
```
---------------------------
median age
---------------------------
mean:  44.209
standard deviation:  6.405


---------------------------
cases per 100k population
---------------------------
mean:  652.217
standard deviation:  435.999


---------------------------
other stats
---------------------------
sample size:  23
correlation:  -0.568
slope:  -38.659756562896256 3
intercept 2361.315
```



Initial Model - COVID-19 Cases vs Median Home Value

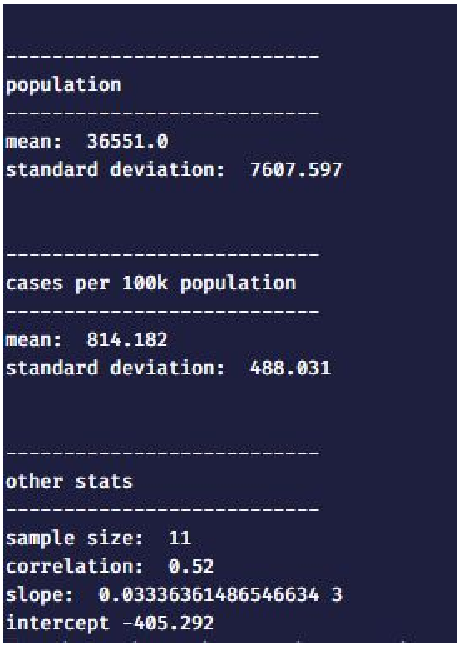


```
---------------------------
median home value
---------------------------
mean:  730226.087
standard deviation:  212512.195


---------------------------
cases per 100k population
---------------------------
mean:  652.217
standard deviation:  435.999


---------------------------
other stats
---------------------------
sample size:  23
correlation:  -0.442
slope:  -0.0009073011397436089 3
intercept 1314.752
```

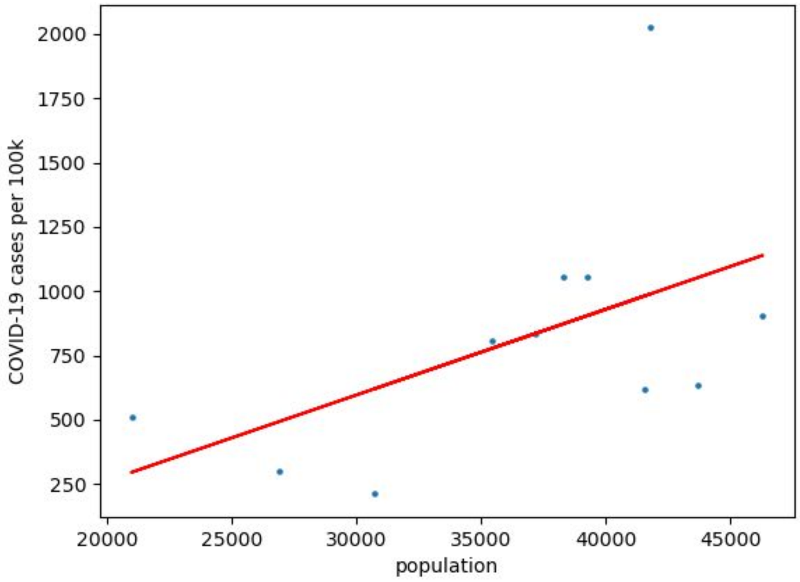


Refined Model

It seems like the population happens to be the only one with a positive association with cases so let’s focus on that. To refine this model, the data must be narrowed down to places that have higher population density. Because of how disease ultimately functions, higher density populations have higher cases due to it being easier to spread viruses in tighter living spaces. For this model, we will be focusing on regions that have populations higher than 30,000 which would make these cities larger. Santa Rosa, Sonoma, Rohnert Park, Sebastopol, and Petaluma will be the subset.



```
---------------------------
population
---------------------------
mean:  36551.0
standard deviation:  7607.597


---------------------------
cases per 100k population
---------------------------
mean:  814.182
standard deviation:  488.031


---------------------------
other stats
---------------------------
sample size:  11
correlation:  0.52
slope:  0.03336361486546634 3
intercept -405.292
```



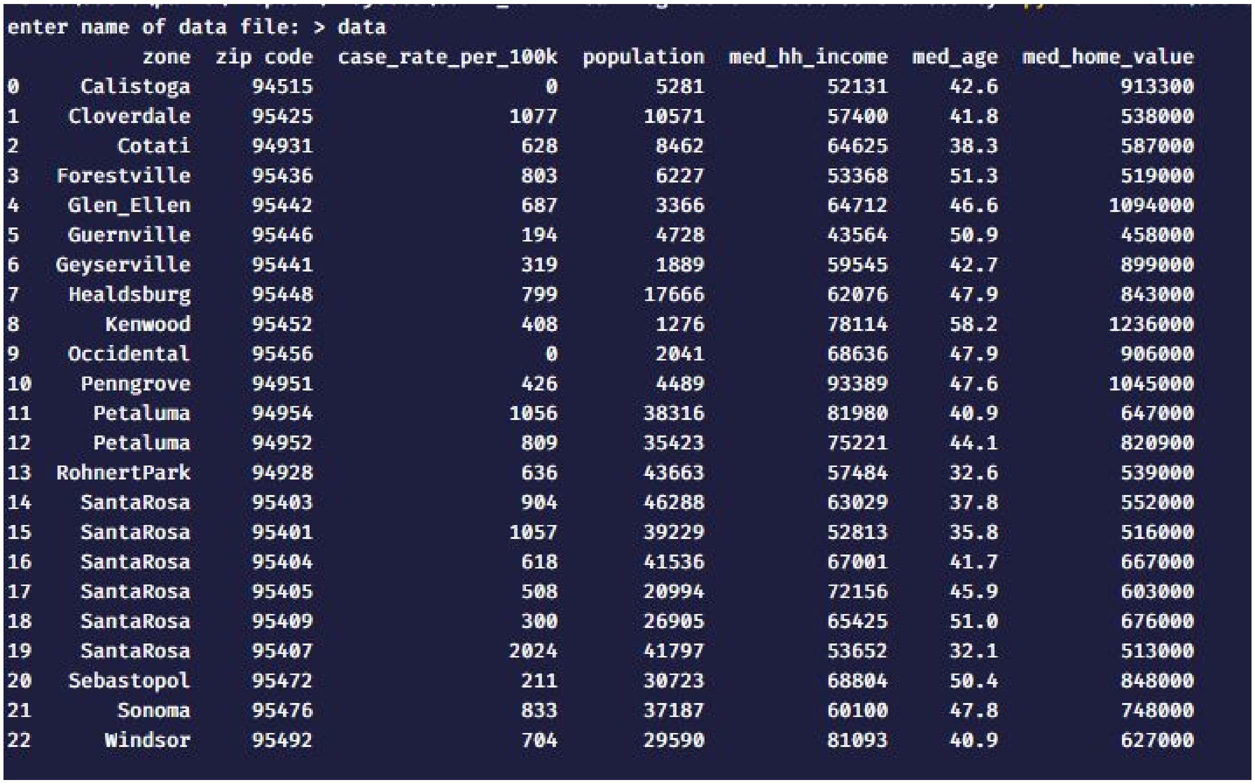
Confounding Variables

Two confounding variables found in this research are population and how dense the population is. This is because it is easier for a virus to spread in dense areas rather than rural areas. More people tend to come in contact with a multitude of others than people who live in smaller areas. You could collect data on population density for these areas and then compare them to cases.

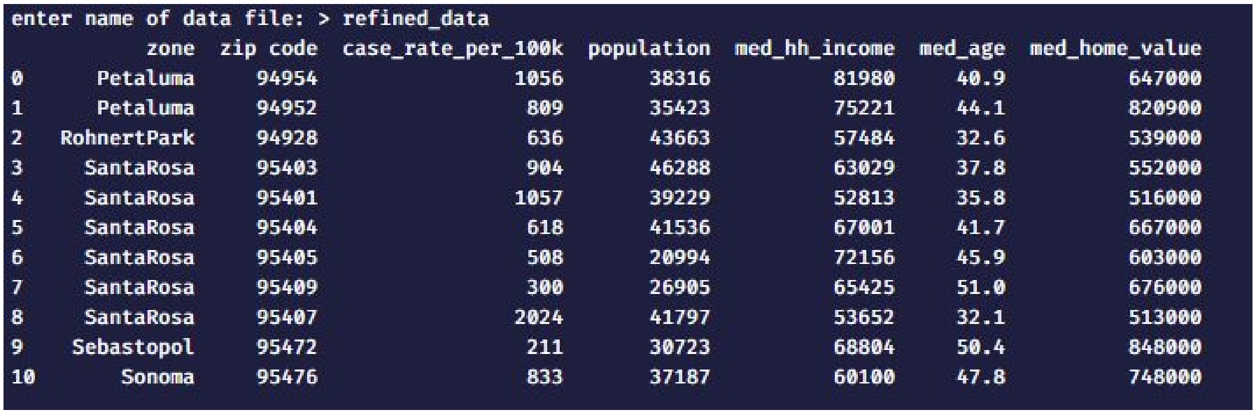
Personal Reflections

The hardest part of this project for me was creating some of the formulas for the slope and intercept. This project taught me that linear regression is helpful when trying to find correlation between variables. I personally believe projects reflect someone's ability more than just answering questions in a test. Especially when I get to use something I love, like programming.

Data Tables



```
enter name of data file: > data
           zone  zip code  case_rate_per_100k  population  med_hh_income  med_age  med_home_value
0     Calistoga     94515                   0        5281          52131     42.6          913300
1    Cloverdale     95425                1077       10571          57400     41.8          538000
2        Cotati     94931                 628        8462          64625     38.3          587000
3   Forestville     95436                 803        6227          53368     51.3          519000
4    Glen_Ellen     95442                 687        3366          64712     46.6         1094000
5    Guernville     95446                 194        4728          43564     50.9          458000
6   Geyserville     95441                 319        1889          59545     42.7          899000
7    Healdsburg     95448                 799       17666          62076     47.9          843000
8       Kenwood     95452                 408        1276          78114     58.2         1236000
9    Occidental     95456                   0        2041          68636     47.9          906000
10    Penngrove     94951                 426        4489          93389     47.6         1045000
11     Petaluma     94954                1056       38316          81980     40.9          647000
12     Petaluma     94952                 809       35423          75221     44.1          820900
13  RohnertPark     94928                 636       43663          57484     32.6          539000
14    SantaRosa     95403                 904       46288          63029     37.8          552000
15    SantaRosa     95401                1057       39229          52813     35.8          516000
16    SantaRosa     95404                 618       41536          67001     41.7          667000
17    SantaRosa     95405                 508       20994          72156     45.9          603000
18    SantaRosa     95409                 300       26905          65425     51.0          676000
19    SantaRosa     95407                2024       41797          53652     32.1          513000
20   Sebastopol     95472                 211       30723          68804     50.4          848000
21       Sonoma     95476                 833       37187          60100     47.8          748000
22      Windsor     95492                 704       29590          81093     40.9          627000
```



```
enter name of data file: > refined_data
           zone  zip code  case_rate_per_100k  population  med_hh_income  med_age  med_home_value
0      Petaluma     94954                1056       38316          81980     40.9          647000
1      Petaluma     94952                 809       35423          75221     44.1          820900
2   RohnertPark     94928                 636       43663          57484     32.6          539000
3     SantaRosa     95403                 904       46288          63029     37.8          552000
4     SantaRosa     95401                1057       39229          52813     35.8          516000
5     SantaRosa     95404                 618       41536          67001     41.7          667000
6     SantaRosa     95405                 508       20994          72156     45.9          603000
7     SantaRosa     95409                 300       26905          65425     51.0          676000
8     SantaRosa     95407                2024       41797          53652     32.1          513000
9    Sebastopol     95472                 211       30723          68804     50.4          848000
10       Sonoma     95476                 833       37187          60100     47.8          748000
```