Final Project

William Breckwoldt

Northeastern University

ALY 6010: Probability Theory and Introductory Statistics

Professor Thomas L. Goulding

December 12th, 2021

Introduction

For Milestone 1, I provided an exploratory analysis of the *United States COVID-19 Cases and Deaths by State over Time* CSV dataset provided by the CDC for public use. The dataset is composed of 15 columns and 39.4k rows, the columns include the state, the submission date, and aggregate COVID-19 data and has entries for all states from January 1st, 2020, to October 30th, 2021.

Topics of interest following the exploratory analysis, include inconsistencies among new cases and states. In Milestone 2, I merged a vaccinations by state dataset (https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/us_state_vaccinations.csv) with the dataset from the previous milestone.

There has been much debate on the factors that influence the spread of COVID-19. In my final project, I will provide some insight using correlation, regression, and inferential statistics. I have obtained and merged population data (https://worldpopulationreview.com/state-rankings/state-densities; https://data.ers.usda.gov/reports.aspx?ID=17827) with infection and vaccination data from the previous milestones. I also mutated columns to find the total infections as a percentage of the population, new cases per capita, and the percentage of population fully vaccinated.

Questions

What variables influence the percentage of population infected?

Comparatively, which states are suffering the most from COVID-19 and which are safe?

How is New England fairing against COVID-19?

Are vaccination rates rising in New England?

Are infection rates rising in in New England?

Hypothesis Test 1

Ho: There is a correlation among the percentage of population infected, the daily infections per capita, the percentage of population fully vaccinated, and daily death and daily infection totals.

Ha: There is no correlation among the percentage of population infected, the daily infections per capita, the percentage of population fully vaccinated, and daily death and daily infection totals.

Figure 1: Correlation Table

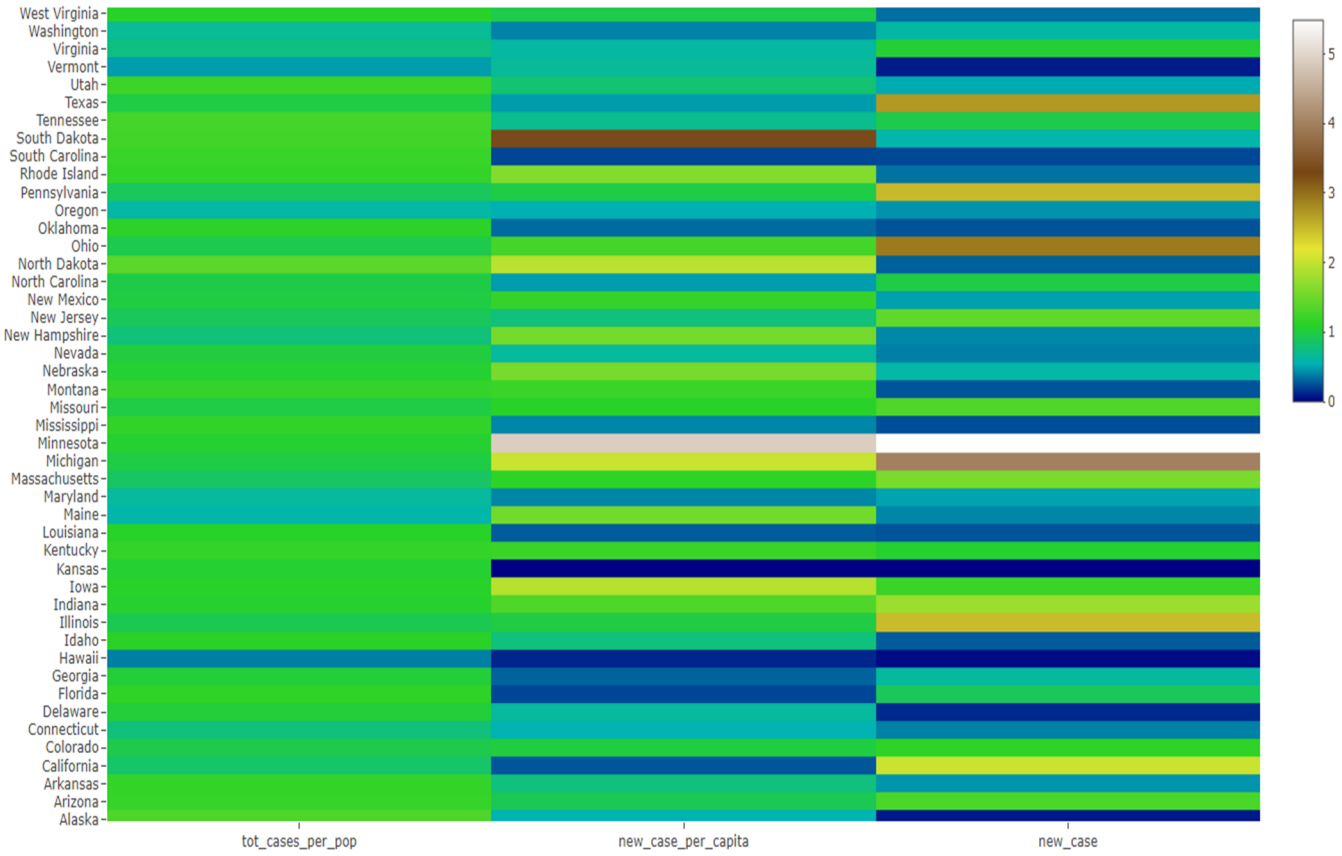
	tot_cases_per_pop	new_case_per_capita	full_vax_rate	new_death
tot_cases_per_pop	1.0000000	0.07483820	-0.65290150	-0.01636900
new_case_per_capita	0.0748382	1.00000000	0.09958389	0.13781557
full_vax_rate	-0.6529015	0.09958389	1.00000000	-0.02019544
new_death	-0.0163690	0.13781557	-0.02019544	1.00000000
new_case	-0.0371181	0.51182012	0.09977504	0.51100554

Using data from the 50 states from November 1st until now, I created this correlation table. Unfortunately, there is no clear evidence to argue if variable have strong correlation. It appears that most are unrelated as their absolute values are close to 0.

However, the largest absolute correlation (-0.6529015) is between total cases per capita (tot-cases_per_pop) and the percent of population fully vaccinated (full_vax_rate), followed by daily infections (new_case) and daily infections per capita (new_case_per_capita) with 0.51182012, then daily infections and daily deaths (new_death) with 0.51100554.

Regardless, we accept the null hypothesis here.

Figure 2: Correlation Heatmap



This interactive correlation heatmap provides insight into the impact COVID-19 has had on individual states as recorded on November 30th, 2021. The data is normalized to provide more clear comparisons.

Minnesota has suffered the most, as their new cases (z = 5.482691) and their new cases per capita (z = 4.890058) is significantly larger than most other states.

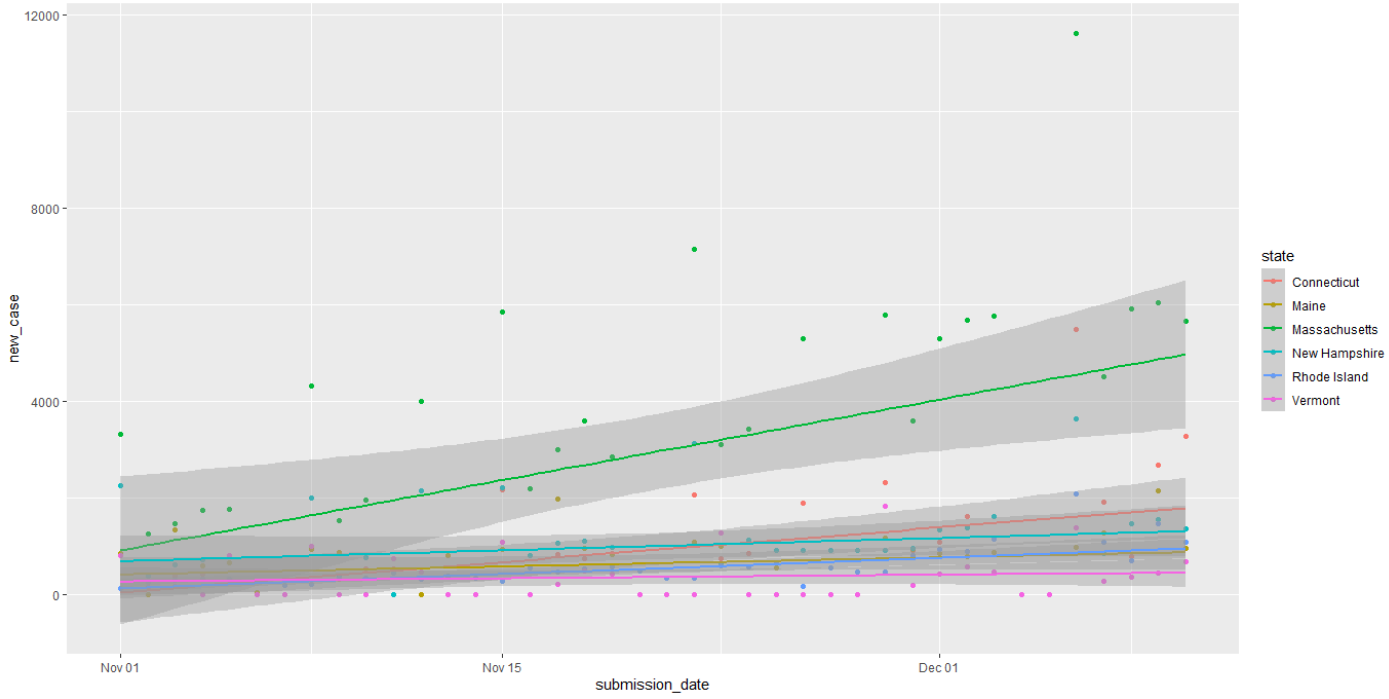
On the other hand, Hawaii has been safe along with Kansas, South Carolina, and South Dakota recently.

Hypothesis Test 2

Ho: New England daily infection rates are decreasing.

Ha: New England daily infection rates are not decreasing.

Figure 3: New England Daily Infections Linear Regression Scatterplot



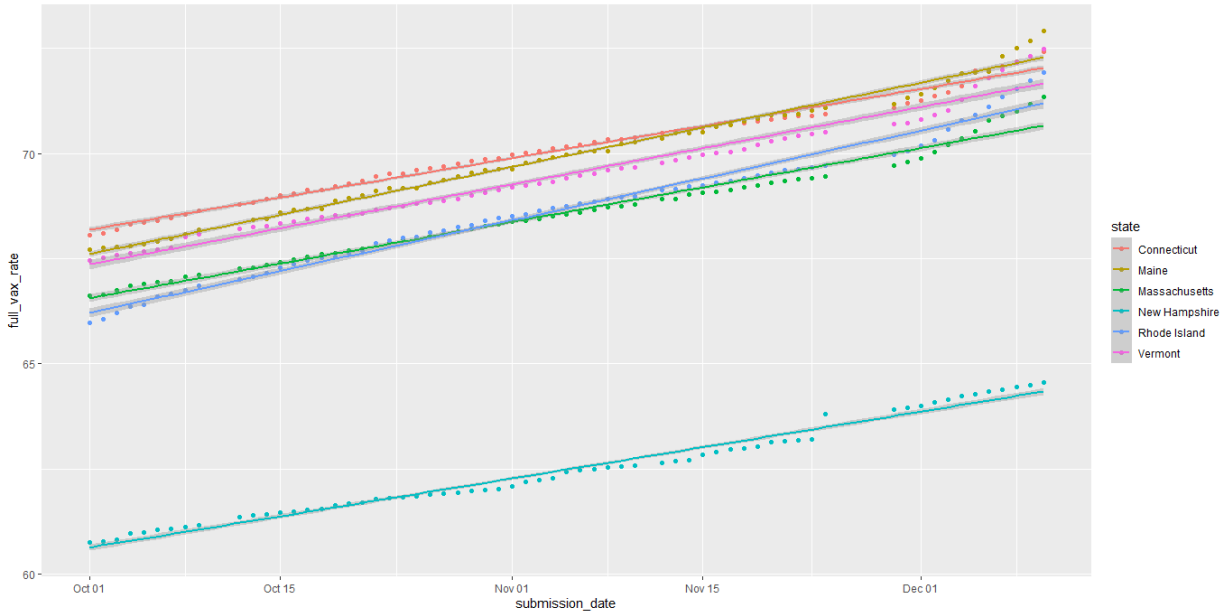
From this scatterplot and these regression lines, we can see that daily infections are increasing in all New England states but especially Massachusetts and Connecticut so we accept the alternative hypothesis. Fortunatlely, infection rates remain somewhat low and steady in the rural, less populated states of Maine and Vermont.

Hypothesis Test 2

Ho: The percentage of population fully vaccinated in Northeastern states is increasing.

Ha: The percentage of population fully vaccinated in Northeastern states is not increasing.

Figure 4: Percent of Population Fully Vaccinated Linear Regression Scatterplot



The percentage of population fully vaccinated in New England states are continuing to rise across the board, most notable in Maine, thus we accept the null hypothesis.

Hypothesis Test 4

Ho: Population density and the percentage of population fully vaccinated influence the percentage of population infected.

Ha: Population density and the percentage of population fully vaccinated does not influence the percentage of population infected.

Figure 5: Regression Linear Model Summary

```
Call:
lm(formula = tot_cases_per_pop ~ Density + full_vax_rate, data = df5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.08905 -0.01429  0.00022  0.01467  0.06204

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.290e-01  9.020e-03  36.476  < 2e-16 ***
Density        1.998e-05  4.683e-06   4.266 2.36e-05 ***
full_vax_rate -3.135e-03  1.636e-04 -19.165  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02462 on 529 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.4373,    Adjusted R-squared:  0.4352
F-statistic: 205.6 on 2 and 529 DF,  p-value: < 2.2e-16
```

This model suggests that there is a relationship between infections per capita and population density and between infections per capita and the percent of population fully vaccinated. This is because the positive test-values are much greater than 1.96 and the negative test-value is much less than -1.96. Furthermore, the p-values are significantly lower than the alpha value of 0.05.

Conclusion

From the results above, we have identified a potential relationship between percentage of population vaccinated and infections per capita. Furthermore, we identified states successfully mitigating COVID-19 infections and others that have recently been suffering. We also found that infection rates are rising in New England, despite the percentage of populations fully vaccinated also rising. It is important to note the lower percentage of population fully vaccinated in New Hampshire as they are experiencing a surge. Finally, the linear regression model suggested that there is a relationship between the daily infections per capita and population density and the percentage of population vaccinated.

Despite this report having inconclusive inferences, it did provide insight into the complication of COVID-19 and the factors at bay. This Final Project provides evidence that supports that samples with higher fully vaccinated rates tend to have less infections per capita. Furthermore,

References

CDC. (13 Nov 2021) *United States COVID-19 Cases and Deaths by State over Time.* Centers for Disease Control and Prevention. Data.CDC.gov. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

Chiluiza, D. (2021) *Dee_Chiluiza*. Northeastern University CPS. *RPubs*. https://rpubs.com/Dee_Chiluiza/800800

Owidbot (12 Dec 2021) *us_state_vaccinations.csv.* Our World in Data. https://data.ers.usda.gov/reports.aspx?ID=17827.

U.S. Census Bureau (8 Oct 2021) *Population.* Economic Research Services: U.S. Department of Agriculture. https://data.ers.usda.gov/reports.aspx?ID=17827.

World Population Review (2020) *United States by Density.* State Rankings. https://worldpopulationreview.com/state-rankings/state-densities.