235/40 very good!

Dennie Truong

Is there an association between US states GDP and COVID-19 deaths?

Introduction

What we now call the COVID pandemic was primarily centered in Washington and New York at the beginning of the US crisis; however, the outbreak quickly spread to the Northeast and South and finally reached the Midwest and West before early fall of 2020 (Hallas, et al.). Now, two years into the pandemic, all states continues to experience the wrath of this virus with over 960,000 COVID related deaths and nearly 80 million COVID confirmed cases reported nationwide (*Johns Hopkins*).

An observational study found a linear, positive association between total GDP (in millions of USD) and total confirmed cases of COVID-19 (during January 1 to May 31, 2020) on a logarithmic scale of 28 European countries with a regression coefficient of 0.7156 ($p < 0.001$) (Aycock). A similar observational study done in China (data up to February 16, 2020) on its 30 provinces also reveals GDP (in trillions of yuan) was positively associated with cumulative COVID-19 cases ($r = 0.69$, $p < 0.01$) (Mo, et al.). Both these studies show higher GDP nations or provinces are associated with more COVID cases and suggest economic growth and urbanization create many more opportunities to facilitate the spread of COVID-19 (Mo, et al.). Even though COVID cases have been remediated by the scientific ingenuity of the COVID vaccines, many experts are still surprised by how fast the virus evolved creating the many variants including Delta and Omicron (Runwal).

? Right at the beginning of the pandemic?

This leads us to ask the question: Do states with a higher Gross Domestic Product (GDP) have a significant difference in COVID reported cases (and ultimately deaths) than those with lower GDP from January 2020 to March 2022? Through understanding if an association for GDP of US states and COVID cases/deaths exists or not would allow the government, both state and federal, to better target and provide the appropriate resources for each state based on their specific needs.

Method

Our data is a combined dataset collected from a multitude of online sources including Worldometer, Ballotpedia, US Bureau of Economic Analysis (BEA), US Energy Information Administration (EIA) and the New York Times. Our cases are the 50 US states and the nation's capital, District of Columbia (DC). The primary explanatory variable of this study is each state's GDP (in billions). Other explanatory variables includes US states by regions (W=West, M=Midwest, N=Northeast, and S=South) and the political party of the governor in office from 2020 to 2021. One response variables we will be looking at is the total COVID cases categorically which has four levels: A: less than 20%, B: between 20% and 25%, C: between 25% and 33.33%, D: greater than 33.33% cases ~~per capita~~. The other response variable is the total COVID death numerical (per 1 million people), and also categorical with four levels: A: less than 0.15%, B: between 0.15% and 0.25%, C: between 0.25% and 0.37%, D: greater than 0.37%. Total cases and deaths is defined by the cummulative reported and detected cases since the start of the pandemic till April 3, 2022. This is an observational study that aims to determine and comment on associations between variables rather than imply causality.

These are % of the population.

Results

The US 50 states and DC vary tremendously in their annual GDP with the maximum GDP belonging to California's $2871.42 billion, the minimum GDP of $29.65 billion from Vermont, while the average of all states is $377.70 billion (Table 1). The shape of GDP's graph is skewed significantly to the right with 3 outliers being California, Texas and New York. As for total COVID cases, the data ranges from 167,036 cases to 341,251 cases per ~~100,000~~ 1 million people ~~per capita of the state's~~ ($\bar{x} = 247,756$) (Table 1). The graph is roughly symmetric with 4 outliers being Hawaii, Maryland, Oregon and Rhode Island. For total COVID deaths, the data ranges from 974 to 4,166 deaths per ~~100,000~~ 1 million people ~~per capita of the state's~~ ($\bar{x} = 2873$) (Table 1). The graph is

slightly skewed to the left with no outliers. The following table provides a 5 number summary of GDP (in billions), total COVID cases and deaths (per 1 million people).

	n	Mean	SD	Min	Q1	Median	Q3	Max
GDP	51	377.70	500.5453	29.65	88.47	220.69	482.98	2871.42
COVID cases	51	247,756	39,097.51	167,036	230,251	250,809	270,153	341,251
COVID deaths	51	2873	808.1615	974	2256	3031	3476	4166

Table 1. Descriptive statistics of US states and DC's GDP, Covid cases, and COVID deaths.

First, we are investigating the potential correlation between the log(GDP) and total COVID death by performing a t-test for correlation (Figure 1). We decided to log GDP because otherwise an exponential model better fits the observed data. After checking the conditions, we find that two of the conditions fail; therefore, we turned to StatKey to perform a randomization test for correlation. The calculated correlation coefficient, r, is 0.195. After carrying out the test, we find a p-value of 0.172. This value is greater than 0.05, therefore, we fail to reject the null hypothesis of no correlation between COVID deaths and log(GDP).

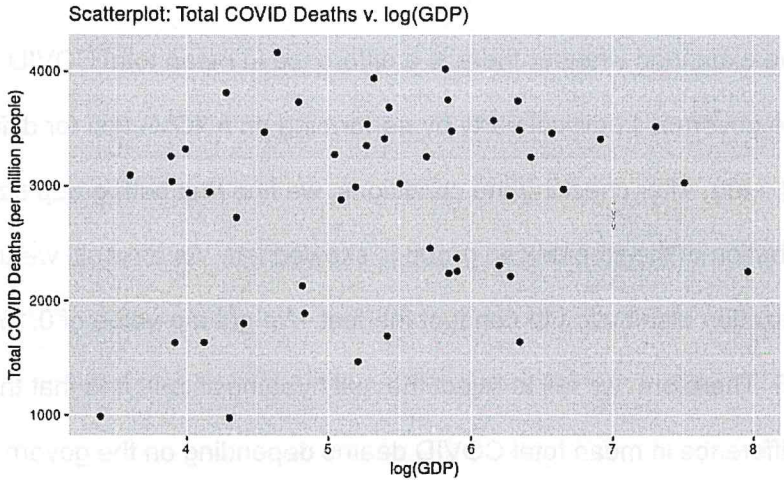


Figure 1. Scatterplot of total COVID deaths (per million people) vs log(GDP) (in billions).

In addition, we also examined for an association between US regions and total COVID cases by performing a Chi-square test for association (Figure 2, Left). After checking the conditions, we find that the majority of the expected counts were less than 5. As a result, StatKey was used to create a randomization distribution. We got a p-value of 0.375 which is greater than 0.05. Therefore, we fail to reject the null hypothesis and find that there is no evidence for an association between US regions and total COVID cases. The other mosaic plot shows the relationship between US regions and total COVID deaths (Figure 2, Right).

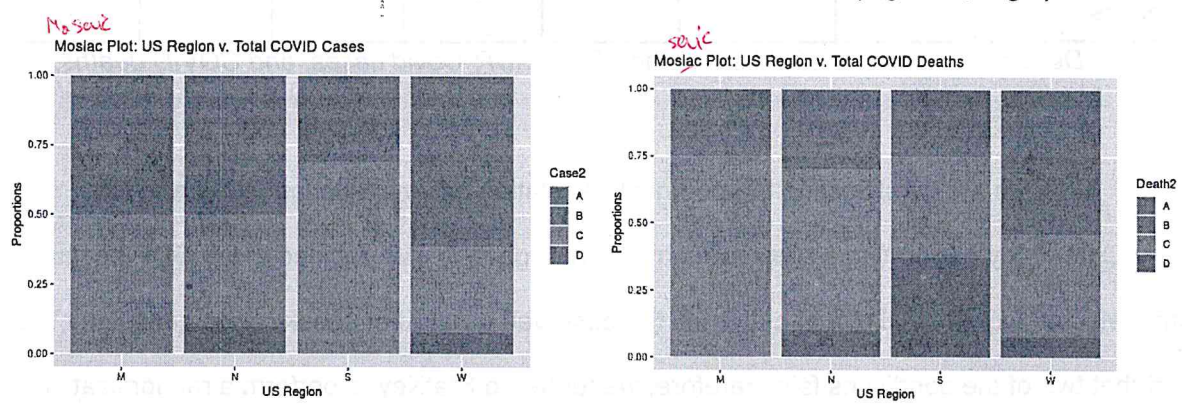


Figure 2. Left: Mosaic plot for an association between total COVID cases (cases per 100,000 people) and US region. Right: Mosaic plot for an association between total COVID deaths (deaths per 100,000 people) and US region.

Finally, we examined whether there is a difference in mean total COVID deaths depending on the governor's political party by performing an ANOVA test for difference in means (Figure 3, Left). After checking the conditions, we find that both group sizes are less than 30 and the distribution of the Republican group is skewed left. As a result, we used StatKey to create a randomization distribution to conduct the test. We got a p-value of 0.150 which is greater than 0.05. Therefore, we fail to reject the null hypothesis and find that there is no evidence for a difference in mean total COVID deaths depending on the governor's political affiliation. The other depicted side-by-side bar chart shows the relationship between US region and log of GDP (Figure 3, Right).

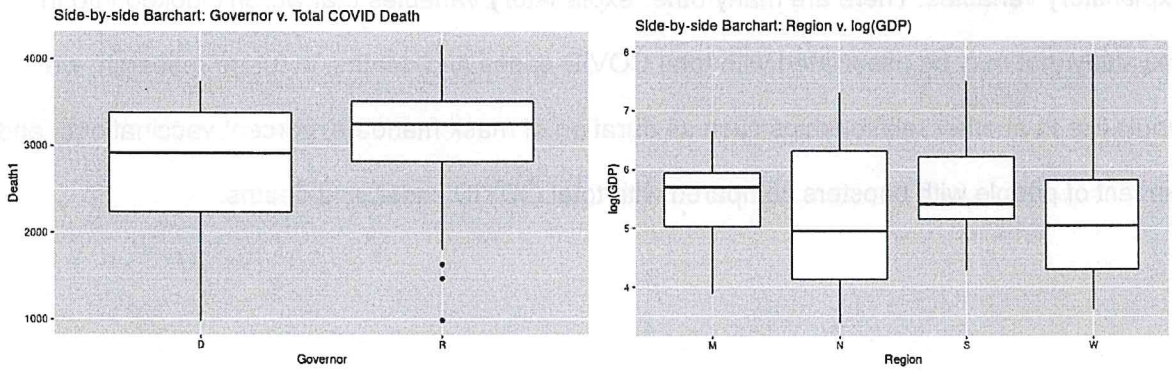


Figure 3. Left: Side-by-side boxplot of total COVID deaths (per million people) by the state's governor political party. Right: Side-by-side boxplot of log(GDP) (in billions) by US regions.

Discussion

maybe there is but it is insignificant

Overall, our statistical tests show that there is no association between a state's GDP and the total COVID cases or death. Furthermore, we also did not find a significant difference in mean total COVID deaths depending on the governor's political ideaology. This is an important result because it shows that the federal government for the most part is iterating a consistent message to all US states about COVID-19 and have provided the required assistance to all states to combat the pandemic (Ballotpedia). Furthermore, by finding no evidence that a state may have higher COVID case or death based on a state's wealth (measured by GDP), where the state is located (in terms of US region), or the political affiliation of the state's governor (or mayor in the case of District of Columbia), it shows that there is no evidence that the federal and state governments' decision on public health are influence by outisde factors that we examined in our study. Our study is limited to only US states (and DC) as well as confined to data from the start of the pandemic until Apr 3, 2022. As a result, our findings is not localized to the community level nor the global level and is only representative of a snapshot of the COVID pandemic history.

Isn't this surprising though? It contradicts the initial findings which led you to study this? Can you explain what the two different findings? Maybe the fact that the urbanized areas were hit first and the two studies were early in the pandemic?

-0.25

Even though we have found no significant results, we only explore three possible explanatory variables. There are many other explanatory variables that weren't looked into in this study that may be associated with total COVID cases and deaths. In future research, we would like to look at other relationships such as duration of mask mandate, percent vaccinations, and percent of people with boosters compared with total COVID cases and deaths.

Bibliography

Aycock, Lauren, and Xinguang Chen. "Levels of Economic Developement and the Spread of Coronavirus Disease 2019 (COVID-19) in 50 U.S. States and Territories and 28 European Countries: An Association Analysis of Aggregated Data." *Global Health Journal*, vol. 5, no. 1, Mar. 2021, pp. 24–30., https://doi.org/10.1016/j.glohj.2021.02.006.

Ballotpedia. (n.d.). Federal government responses to the coronavirus (COVID-19) pandemic, 2020-2022. Retrieved April 29, 2022, from https://ballotpedia.org/Federal_government_responses_to_the_coronavirus_(COVID-19)_pandemic,_2020-2022

Hallas, Laura, et al. *Variation in Government Responses to Covid-19*. Blavatnik School of Government, 17 Dec. 2020, https://www.bsg.ox.ac.uk/sites/default/files/2020-12/BSG-WP-2020-032-v10.pdf.

Mo, Qiqing, et al. "Levels of Economic Growth and Cross-Province Spread of the Covid-19 in China." *Journal of Epidemiology and Community Health*, vol. 75, no. 9, 2021, pp. 824–828., https://doi.org/10.1136/jech-2020-214169.

Runwal, Priyanka. "Two Years Later, Coronavirus Evolution Still Surprises Experts. Here's Why." *Science*, National Geographic, 11 Mar. 2022, https://www.nationalgeographic.com/science/article/two-years-into-the-pandemic-covid-19-still-surprises-experts?loggedin=true.

"United States - COVID-19 Overview - Johns Hopkins." *Johns Hopkins Coronavirus Resource Center*, Johns Hopkins University of Medicine, 2022, https://coronavirus.jhu.edu/region/united-states.

Appendix

1. *Randomization test for correlation: Total COVID Death v ln(GDP)*

Parameters: Let ρ be the correlation between total COVID Deaths (in deaths per million people) and the natural log of GDP of the 50 US states and the nation's capital

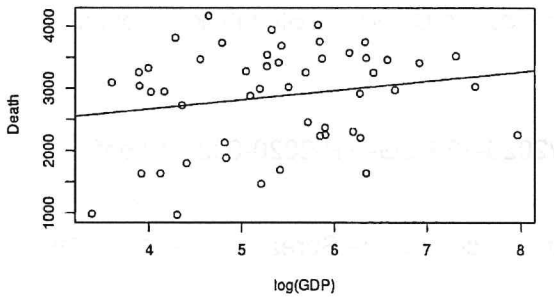
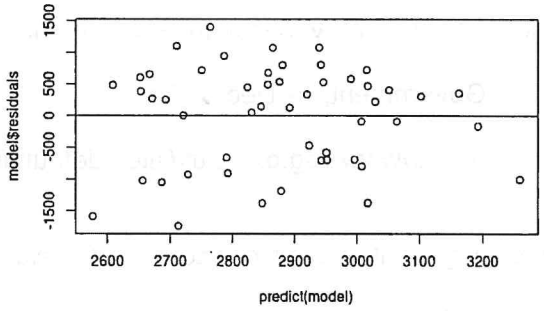
Hypothesis Test:

H_0: ρ = 0 (There is no correlation)

H_a: ρ ≠ 0 (There is a correlation)

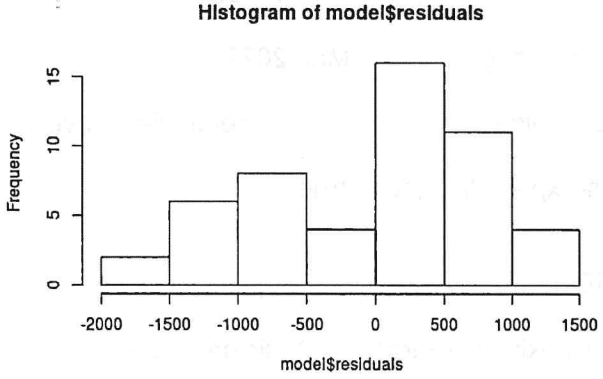
Check Conditions:

1. Linearity: A line best fit the data in the scatterplot. Ther is no trend in the RVF plot. ☑

Scatterplot	Residual Plot
	

2. Independence: Yes, each state is independent from other states. ☑
3. ✗ Normality:

For a given value of log(GDP) the # deaths for each state are independent of one another.

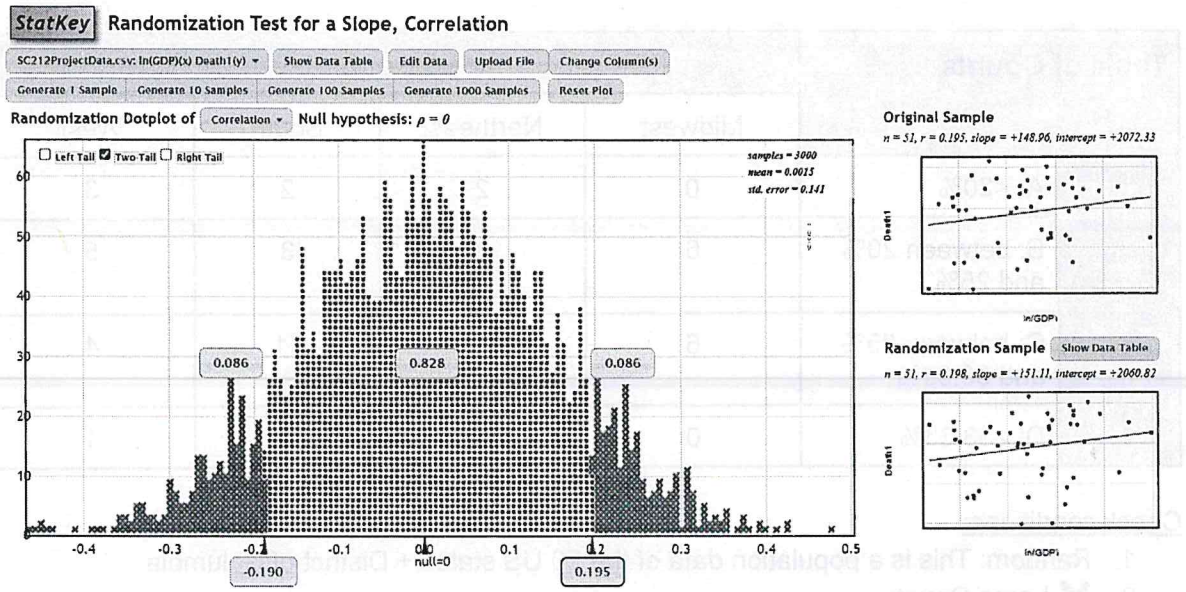


Histogram of the residuals is not normally distributed, has two peaks/bimodal.

Looks ok.

4. ✗ Equal Variance: There is less variability as you move from left to right of the RVF plot.
5. Random: This is a population data of the 50 US states + District of Columbia

Because not all conditions are met, we have to rely on StatKey to perform a randomization test for correlation.



- p-value = 2(0.086) = 0.172 > 0.05 ✓
- Because p-value = 0.172 > 0.05, we fail to reject the null hypothesis. We do not have evidence that there is a correlation between total COVID Deaths (in deaths per million people) and the natural log of GDP of the 50 US states and the nation's capital.

Linear regression Model:

$$\hat{Deaths} = 2072.33 + 148.96 \bullet ln(GDP)$$

Don't need this

2. *Chi-square Test for Association: Total COVID cases (categorical) v. US Region*

Hypothesis Test:

H_0: Total COVID cases and US regions are not associated.

H_a: Total COVID cases and US regions are associated.

Table of Counts		US Region			
		Midwest	Northeast	South	West
Total COVID Cases per capita	A: <20%	0	2	2	3
	B: between 20% and 25%	6	3	3	5
	C: between 25% and 33.33%	6	4	11	4
	D: >33.33%	0	1	0	1

Check conditions:

1. Random: This is a population data of the 50 US states + District of Columbia
2. ✗ Large Count:

$$\text{Expected Count} = \frac{Row\ Total \times Column\ Total}{Sample\ size}$$

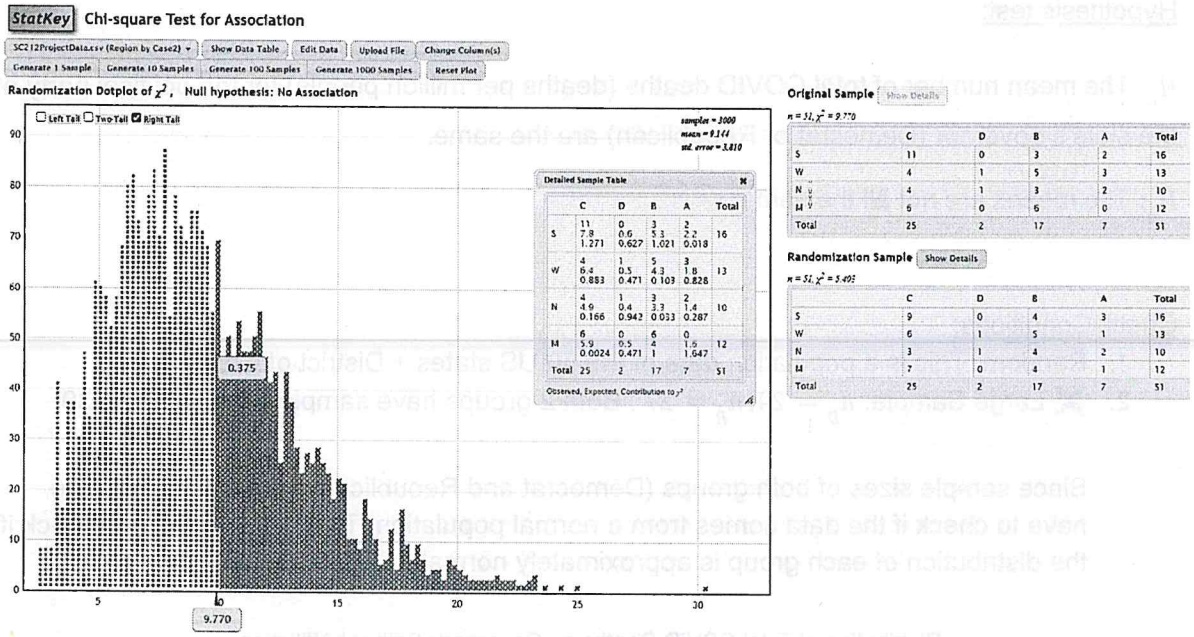
Table of Expected Counts:

	Midwest	Northeast	South	West
A	$\frac{7\times12}{51} = 1.647$	$\frac{7\times10}{51} = 1.372$	$\frac{7\times16}{51} = 2.196$	$\frac{7\times13}{51} = 1.784$
B	$\frac{17\times12}{51} = 4$	$\frac{17\times10}{51} = 3.333$	$\frac{17\times16}{51} = 5.333$	$\frac{17\times13}{51} = 4.333$
C	$\frac{25\times12}{51} = 5.882$	$\frac{25\times10}{51} = 4.902$	$\frac{25\times16}{51} = 7.843$	$\frac{25\times13}{51} = 6.373$
D	$\frac{2\times12}{51} = 0.471$	$\frac{2\times10}{51} = 0.392$	$\frac{2\times16}{51} = 0.627$	$\frac{2\times13}{51} = 0.510$

Not all expected counts are greater than 5.

$$\chi^2 = \sum_{categories} \frac{(Observed - Expected)^2}{Expected} \approx 9.770 \text{ (from StatKey)}$$

Because not all conditions are met, we have to rely on StatKey to perform a randomization test for Chi-sqaure Test for Association.



p-value = 0.375 > 0.05

∴ We fail to reject the H_0. We do not have evidence that the total COVID cases (in cases per 100,000 people) and US regions are associated.

3. *ANOVA for Difference in Means: Governor and Total COVID Deaths*

Parameters:

Let μ_D = average total COVID deaths (per million people) under a Democratic governor

Let μ_R = average total COVID deaths (per million people) under a Republican governor

Hypothesis test:

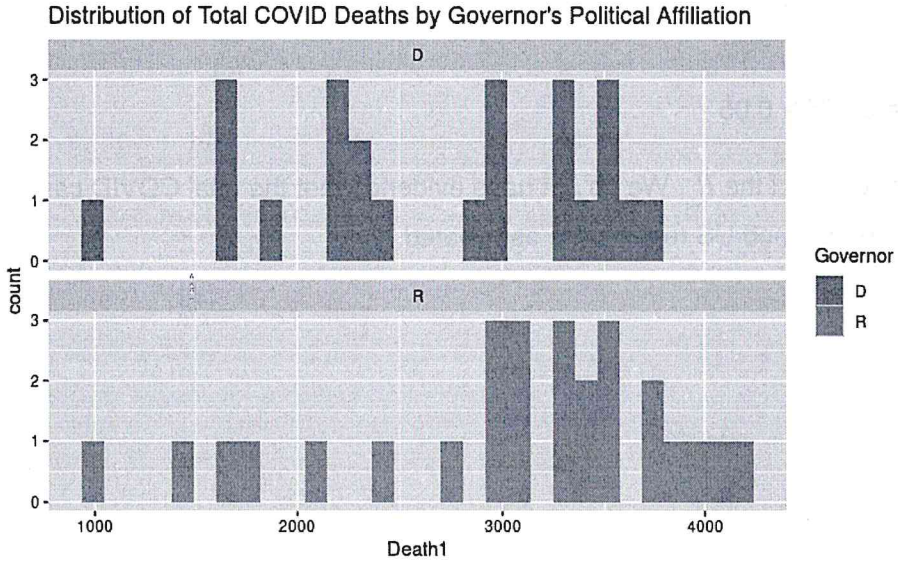
H_0: The mean number of total COVID deaths (deaths per million people) for the political party of the state's governor (Democrat or Republican) are the same.

H_a: The means are not all the same.

Check Conditions:

1. Random: This is a population data of the 50 US states + District of Columbia.
2. ✗ Large Sample: $n_D = 24, n_R = 27$. Both 2 groups have sample size less than 30.

   Since sample sizes of both groups (Democrat and Republican) are less than 30, we have to check if the data comes from a normal population. To do so, we want to check if the distribution of each group is approximately normal.



   Based on the distribution above, the Democrat group is approximately normal; however, the Republican group is skewed to the left.

3. Independent sample: Yes, there are 2 separate political parties. ☑

4. Equal variance: $s_{Largest} < 2s_{Smallest}$

$815 < 2(781.4)$ ← True, so condition is satisfy ☑

$Overall\ mean = \bar{x} = 2872.6$

$$SSG = \sum_{groups} n_i(\bar{x}_i - \bar{x})^2 = 24(2700.6 - 2872.6)^2 + 27(3025.5 - 2872.6)^2$$
$$= 1{,}340{,}873.8$$

$df_{Group} = k - 1 = 2 - 1 = 1$

$MSG = \frac{SSG}{k-1} = \frac{1{,}340{,}873.8}{1} = 1{,}340{,}873.8$

$$SSE = \sum_{groups} (n_i - 1)s_i^2 = (24 - 1)(781.4)^2 + (27 - 1)(815)^2$$
$$= 31{,}315{,}380.4$$

Didn't need to include all of this just the ANOVA table.

$n = 51$

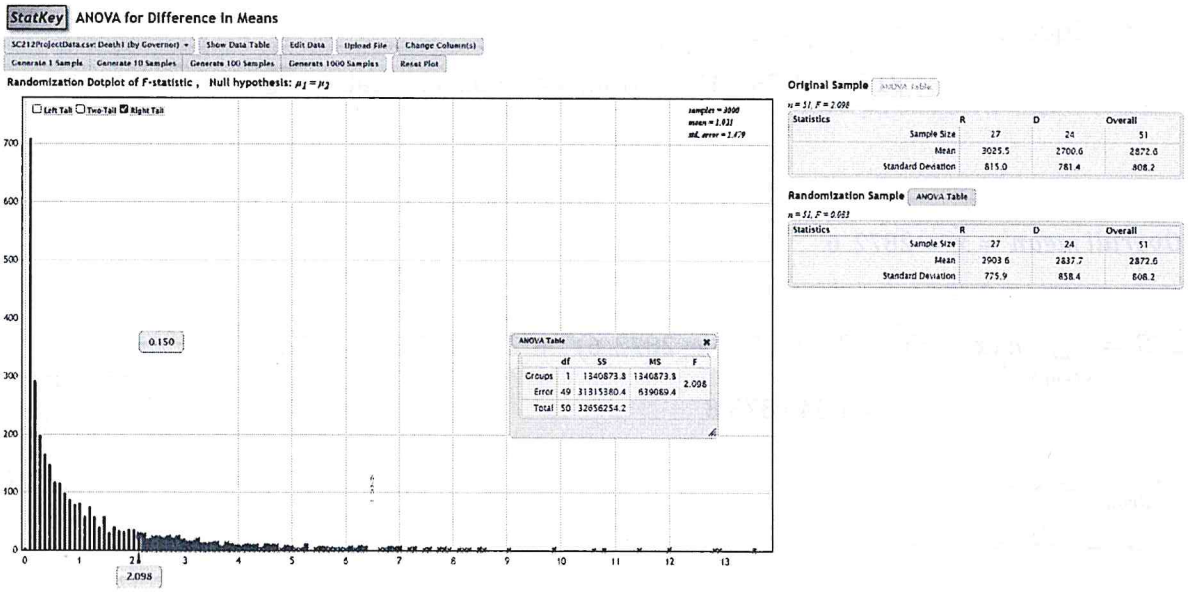
$df_{Error} = n - k = 51 - 2 = 49$

$MSE = \frac{SSE}{n-k} = \frac{31{,}315{,}380.4}{49} = 639089.4$

$$F = \frac{MSG}{MSE} = \frac{1{,}340{,}873.8}{639089.4} \approx 2.098$$

ANOVA Table

	df	SS	MS	F
Groups	1	1340873.8	1340873.8	2.098
Error	49	31315380.4	639089.4	
Total	50	32656254.2		



p-value = 0.150

Because p-value = 0.150 > 0.05, we fail to reject the null hypothesis. Therefore, we do not have evidence of a difference in mean number of total COVID deaths (deaths per million people) depending on the political party of the state's governor.