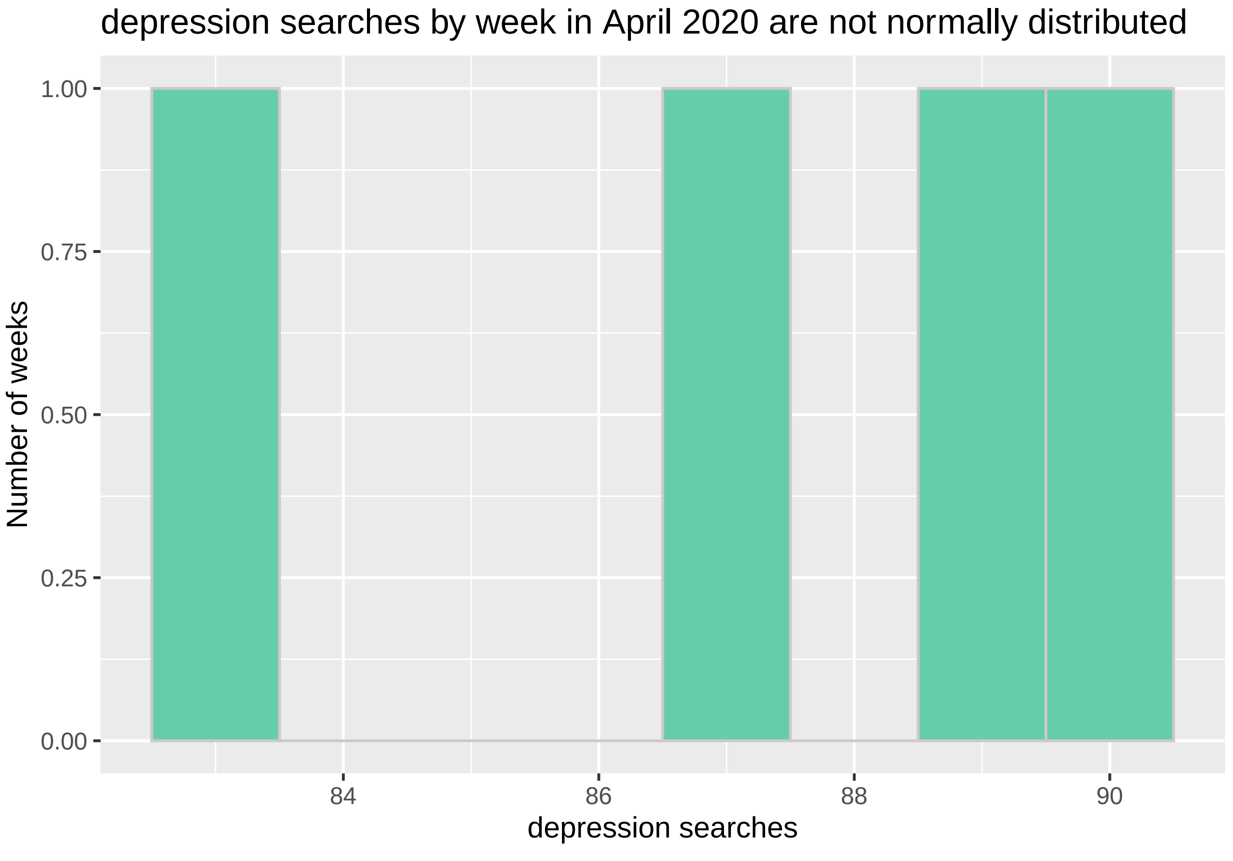
Final Project

Brenda Yang, Charlie Bonetti, Nour Kanaan

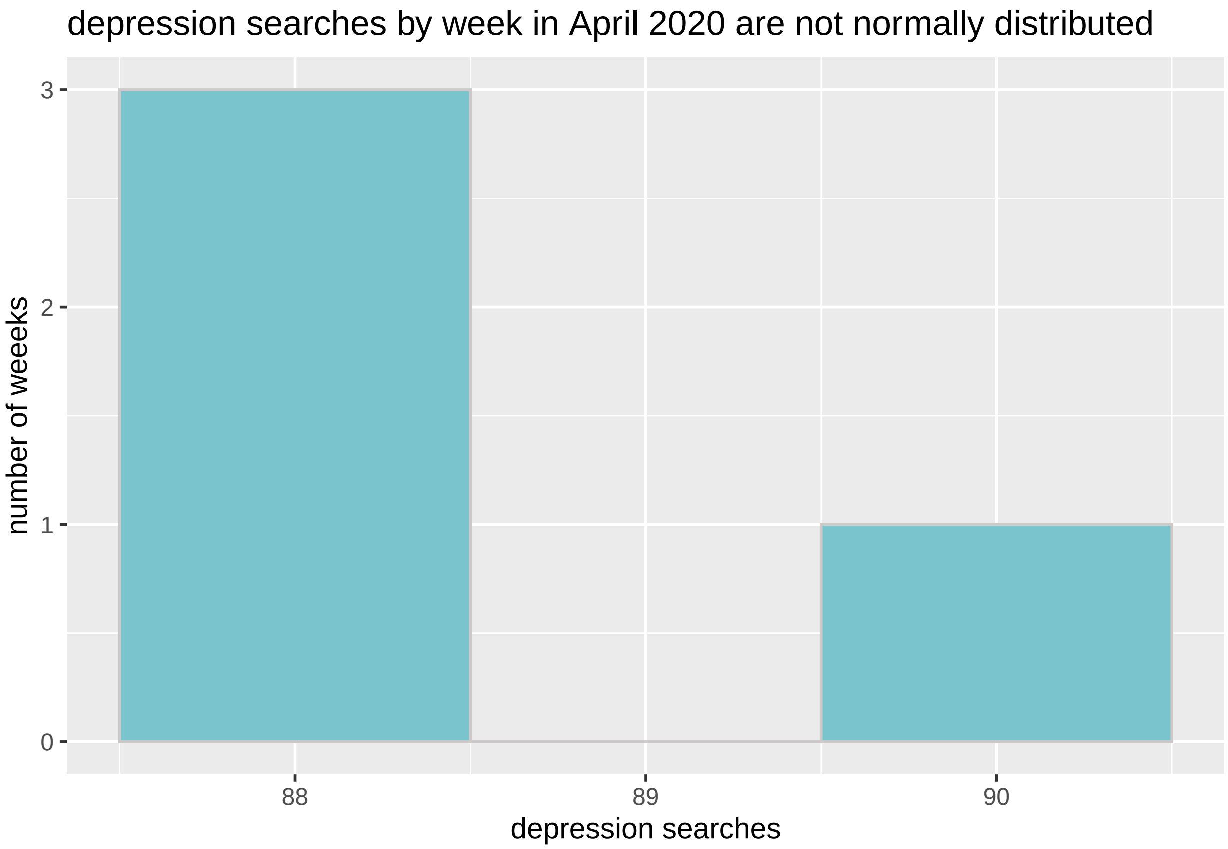
7/19/2020

Question 1: Does COVID have a significant effect on depression and anxiety levels in the US?

"Depression" searches in April 2018 vs. April 2020



n<30 and not normal distribution: assumption for t-test not satisfied



n<30 and not normal distribution: assumption for t-test not satisfied

The assumptions for the two-sample t-test is not satisfied, so we decided to try the Wilcox signed-rank test instead.

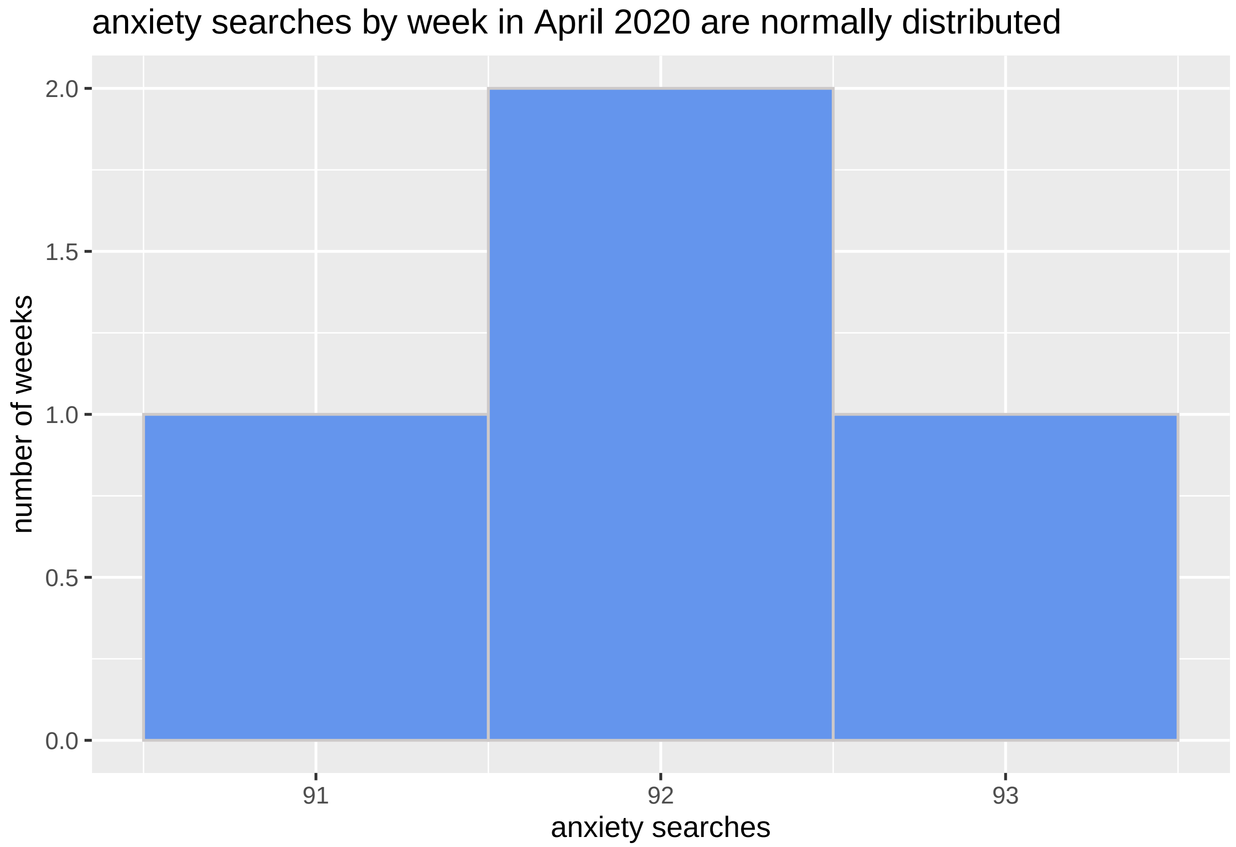
```
## Warning in wilcox.test.default(d2020, d2018, alternative = "two.sided", : cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  d2020 and d2018
## V = 5.5, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

The null hypothesis is that there is no difference in the median number of depression searches in the US between the times of April 2020 and April 2018. The alternate hypothesis is that there is a difference between the two means. Assuming that the null hypothesis is true, the model follows a z-distribution. The z-statistic is 5.5. This corresponds to a p-value of 1. We cannot reject the null at the alpha = 0.05 level. We do not have enough evidence to claim that there is a difference in the median amount of depression searches in the US between the times of April 2020 and April 2018.

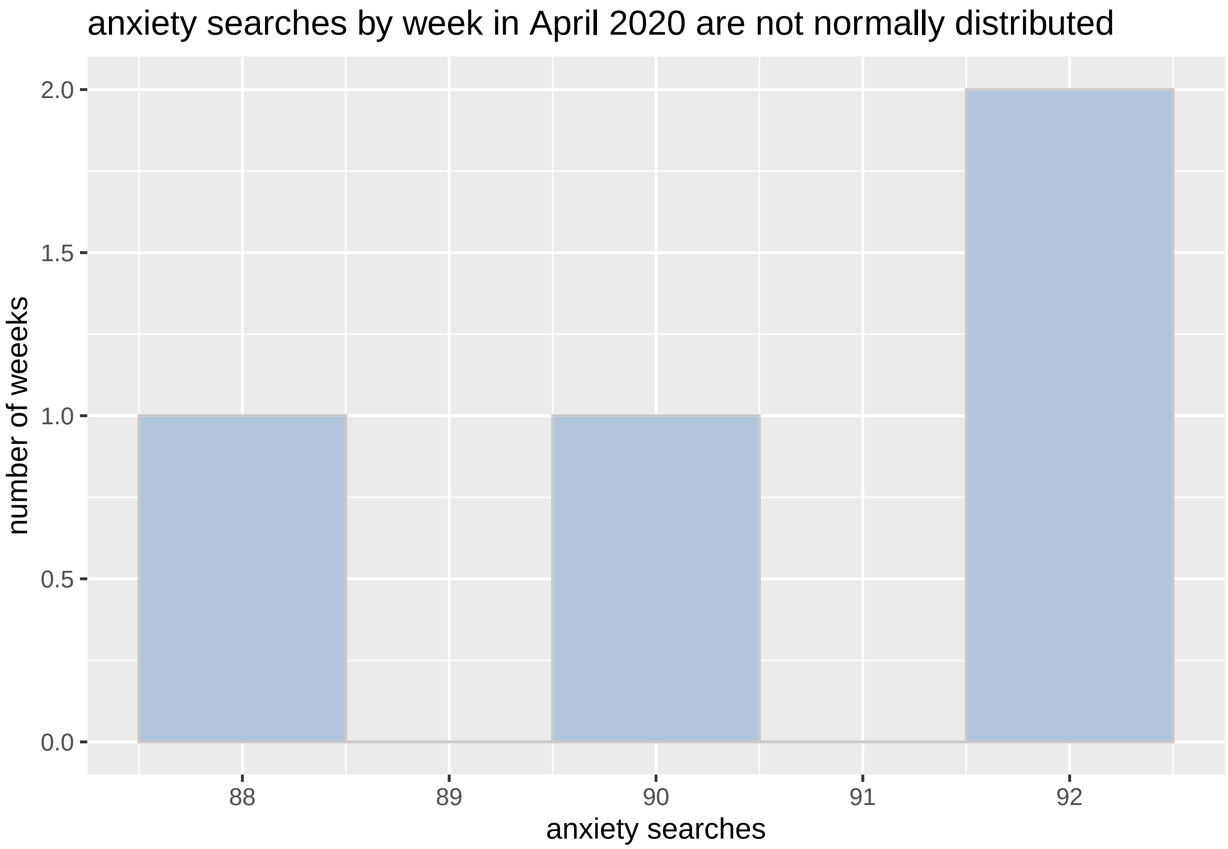
"Anxiety" searches in April 2018 vs. April 2020

```
ggplot(data = atrends2018, mapping = aes(x = anxiety)) +
  geom_histogram(color = "snow3", fill = "cornflowerblue", binwidth = 1)+
  labs(title = "anxiety searches by week in April 2020 are normally distributed",
       x = "anxiety searches",
       y = "number of weeeks")
```



n<30, but has a normal distribution : assumption satisfied

```
ggplot(data = atrends2020, mapping = aes(x = anxiety)) +
  geom_histogram(color = "snow3", fill = "lightsteelblue", binwidth = 1)+
  labs(title = "anxiety searches by week in April 2020 are not normally distributed",
       x = "anxiety searches",
       y = "number of weeeks")
```



n<30 and not normal distribution: assumption not satisfied

The assumptions are not satisfied, so we used the Wilcoxon signed test.

```
## Warning in wilcox.test.default(a2020, a2018, alternative = "two.sided", : cannot
## compute exact p-value with zeroes
```

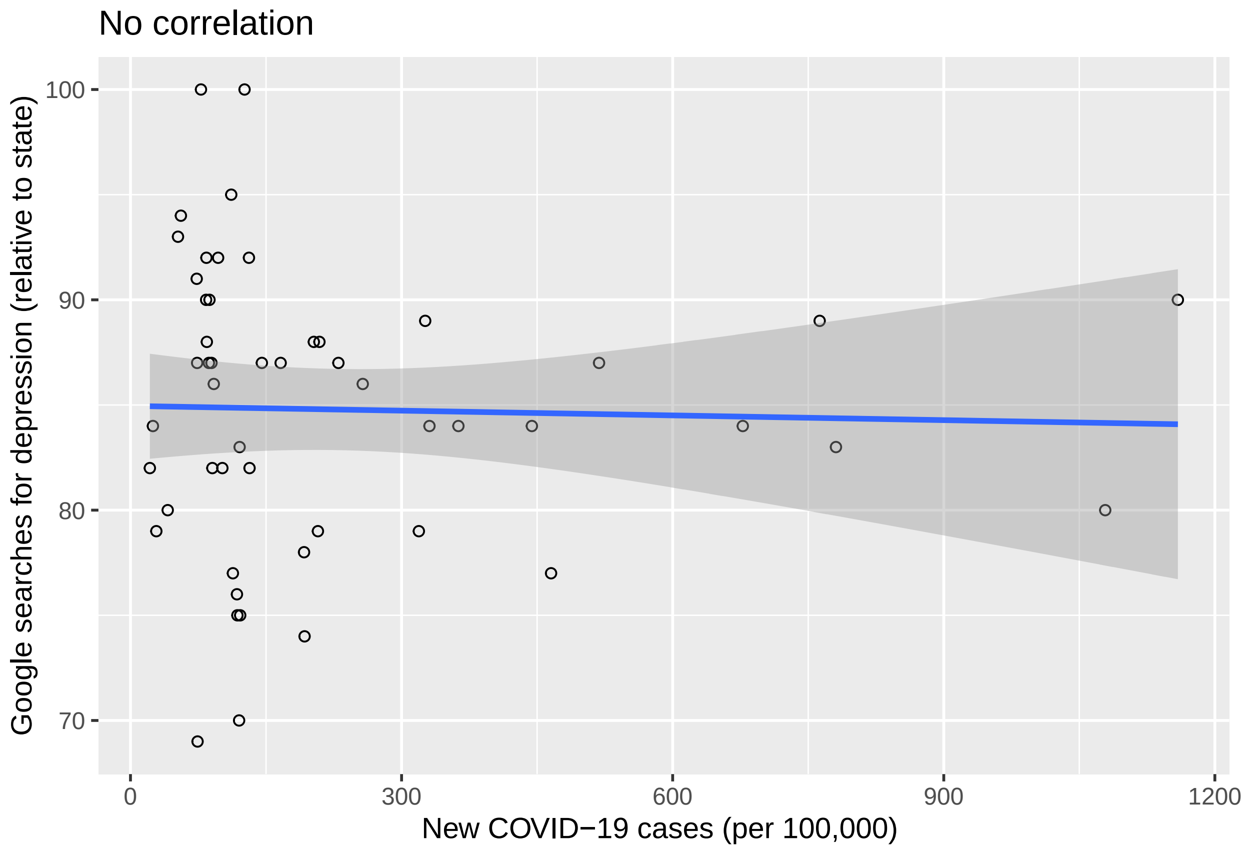
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  a2020 and a2018
## V = 1, p-value = 0.4227
## alternative hypothesis: true location shift is not equal to 0
```

The null hypothesis is that there is no difference in the median amount of anxiety searches in the US between the times of April 2020 and April 2018. The alternate hypothesis is that there is a difference between the two medians. Assuming that the null hypothesis is true, the model follows a z-distribution. The z-statistic is 1. This corresponds to a p-value of 0.4227. We cannot reject the null at the alpha = 0.05 level. We do not have enough evidence to claim that there is a difference in the median amount of anxiety searches in the US between the times of April 2020 and April 2018.

Question 2: Does the level of COVID cases per US state have an effect on depression and anxiety levels?

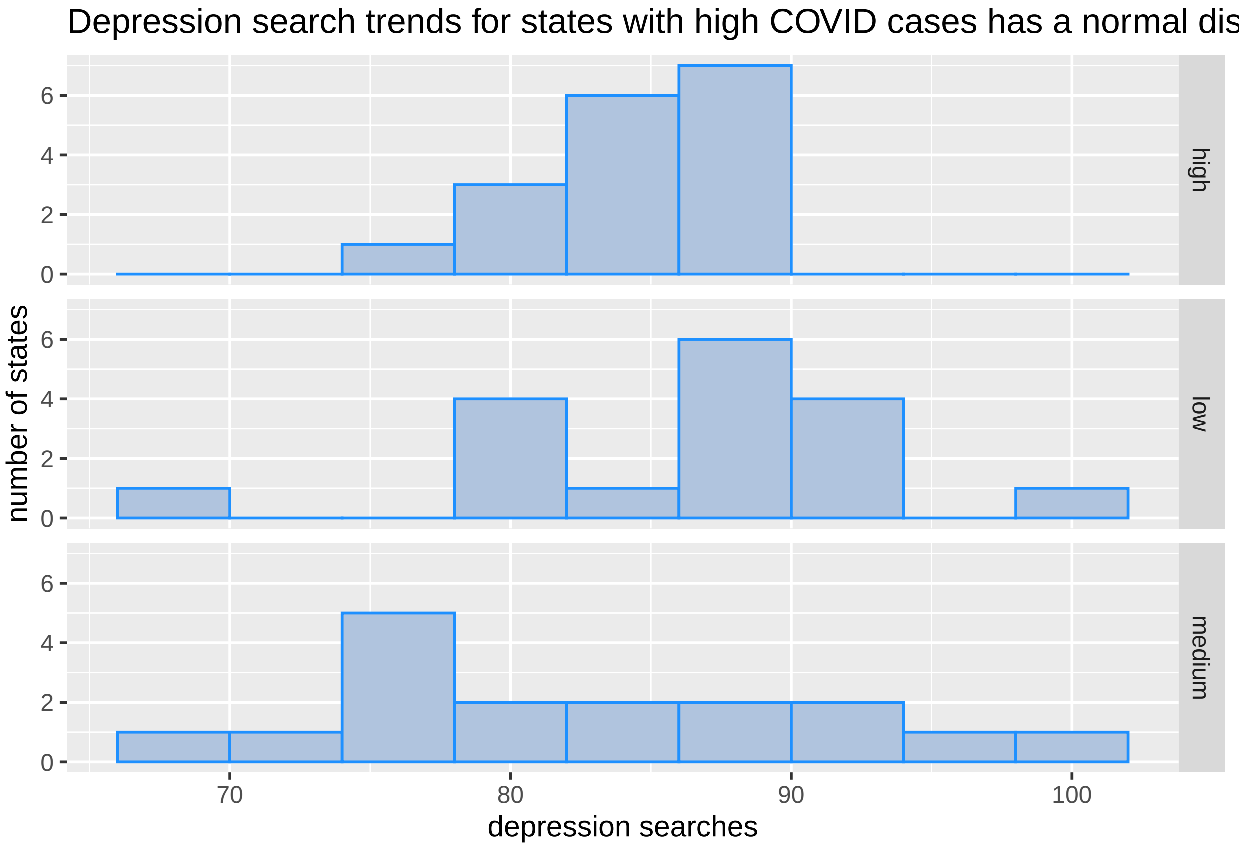
Association between "depression" searches and new COVID-19 cases per 100,000 residents in April 2020

```
## `geom_smooth()` using formula 'y ~ x'
```



The above scatterplot shows the relative google searches for depression among the 50 American states as a function of new COVID-19 cases per 100,000 during the month of april, 2020. The line of best fit (in blue) is horizontal with R = approximately zero. Thus, we cannot see any correlation between the new COVID-19 cases and the Google searches for depression for the month of April.

COVID cases vs. depression rate


Depression search trends for states with high COVID cases has a normal dis

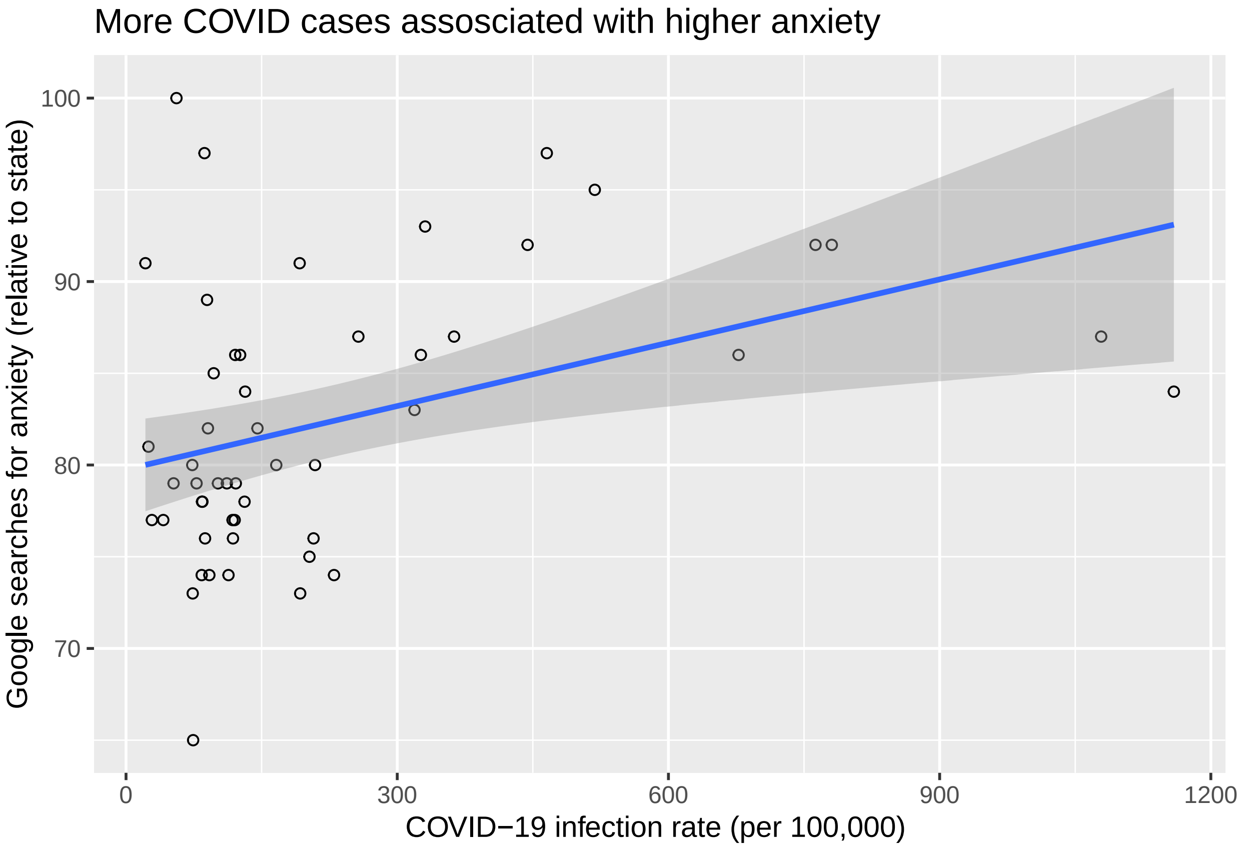
The outcomes within each group is not normal. The depression search trends for states with medium numbers of COVID cases and low numbers of COVID cases do not have a normal distribution, and n<30. Therefore, this assumption is not satisfied. By looking at the graphs, it also seems that there is not equal variance among each group, not satisfying the assumption of homoscedastic variance. In addition, these samples may not all be independent. Some states may have the same values/cultures as others, causing the people who live in each state to react to the virus similarly to each other and affecting the depression searches within those states. Because the assumptions for ANOVA is not satisfied, we decided to use the Kruskal-Wallis test despite the data not being independent.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  depression by case_cat
## Kruskal-Wallis chi-squared = 3.4137, df = 2, p-value = 0.1814
```

The null is that there is no significant difference between the median depression trends of states with low COVID cases, medium COVID cases, and high COVID cases. The alternate hypothesis is that there exists at least one median that is different. Assuming the null hypothesis is true, the model follows a chi square distribution with a df of 2. The chi square statistic is 3.4137, and the corresponding p-value is 0.1814. Therefore, we can not reject the null under the alpha = 0.05 significance level. There is not enough evidence to suggest that there is at least one difference in median depression trends of states with low, medium, and high COVID cases.

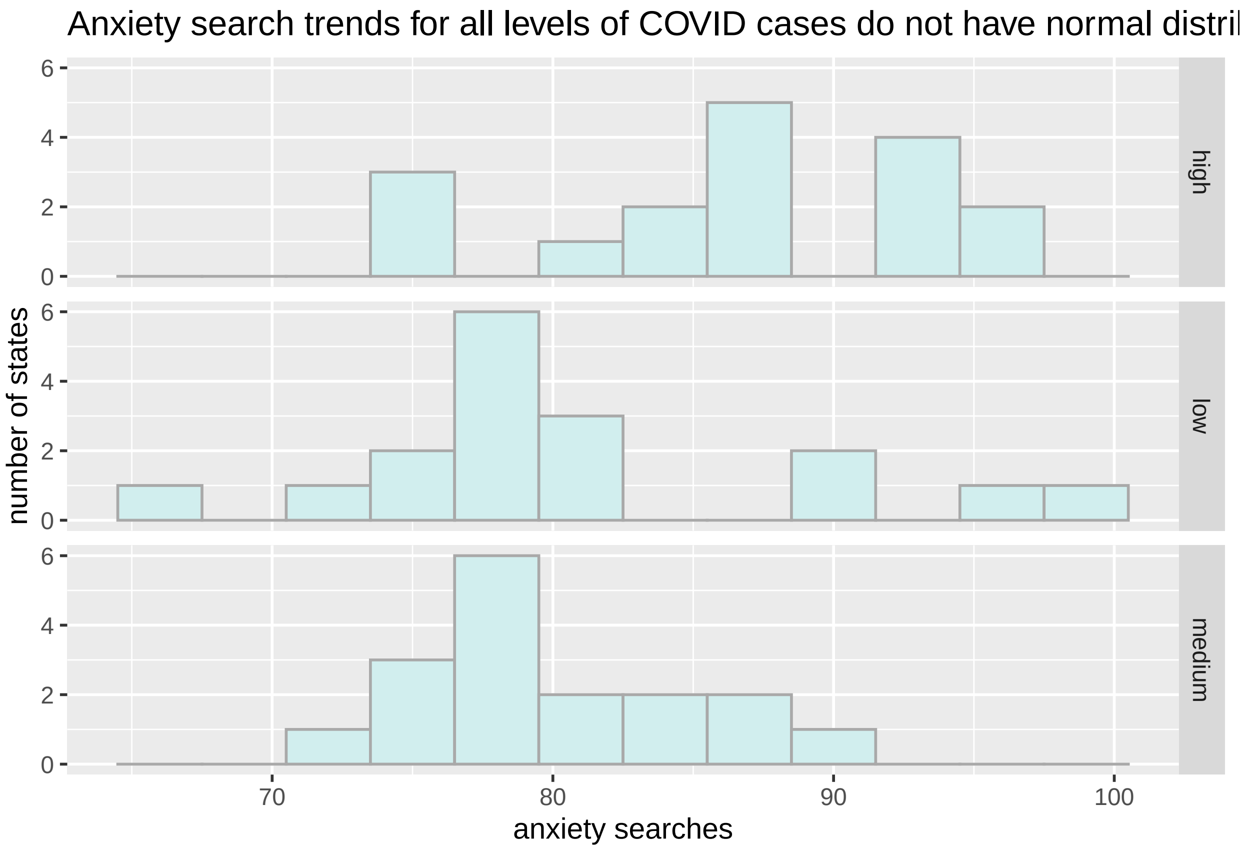
Association between "anxiety" searches and new COVID-19 cases per 100,000 residents in April 2020

```
## `geom_smooth()` using formula 'y ~ x'
```



The above scatterplot shows the relative google searches for anxiety among the 50 american states as a function of new COVID-19 cases per 100,000 during the month of april, 2020. The line of best fit (in blue) has a positive slope. Thus, we can conclude that there's a correlation between anxiety searches and new COVID cases during the month of april. The more new COVID cases a state records, the higher anxiety searches it's expected to have.

COVID cases vs. anxiety rate



Looking at the graphs, outcomes within groups are not normally distributed for any level of COVID cases, so this assumption is not satisfied. It also looks like the within-group variance among all groups is not the same, so the assumption for homoscedastic variance is not satisfied. The samples are also not independent because states with similar values that live close to each other my have similar anxiety search trends. The assumptions for ANOVA are not satisfied. Therefore, we used the Kruskal-Wallis test.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  anxiety by case_cat
## Kruskal-Wallis chi-squared = 7.4355, df = 2, p-value = 0.02429
```

The null is that there is no significant difference between the mean anxiety trends of states with low COVID cases, medium COVID cases, and high COVID cases. The alternate hypothesis is that there exists at least one mean that is different. Assuming the null hypothesis is true, the model follows an F distribution with a df of 2. The chi square statistic is 3.826, and the corresponding p-value is 0.0287. Therefore, we can reject the null under the alpha = 0.05 significance level. There is enough evidence to suggest that there is at least one difference in mean anxiety trends of states with low, medium, and high COVID cases.

Since the overall Kruskal-Wallis test was significant, we then performed step down to identify where the differences are. The appropriate stepdown test is the Wilcox rank sum test. To account for multiple comparison, we will perform the Bonferroni correction and assess our results.

```
low_covid <- trends %>%
  filter(case_cat == "low") %>%
  select(anxiety) %>%
  pull()
```

```
med_covid <- trends %>%
  filter(case_cat == "medium") %>%
  select(anxiety) %>%
  pull()

high_covid <- trends %>%
  filter(case_cat == "high") %>%
  select(anxiety) %>%
  pull()

wilcox.test(low_covid, med_covid, data = licorice,
            alternative = "two.sided",
            paired = FALSE,
            conf.level = 0.95)
```

```
## Warning in wilcox.test.default(low_covid, med_covid, data = licorice,
## alternative = "two.sided", : cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  low_covid and med_covid
## W = 145.5, p-value = 0.9862
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(low_covid, high_covid, data = licorice,
            alternative = "two.sided",
            paired = FALSE,
            conf.level = 0.95)
```

```
## Warning in wilcox.test.default(low_covid, high_covid, data = licorice,
## alternative = "two.sided", : cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  low_covid and high_covid
## W = 88, p-value = 0.05346
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(med_covid, high_covid, data = licorice,
            alternative = "two.sided",
            paired = FALSE,
            conf.level = 0.95)
```

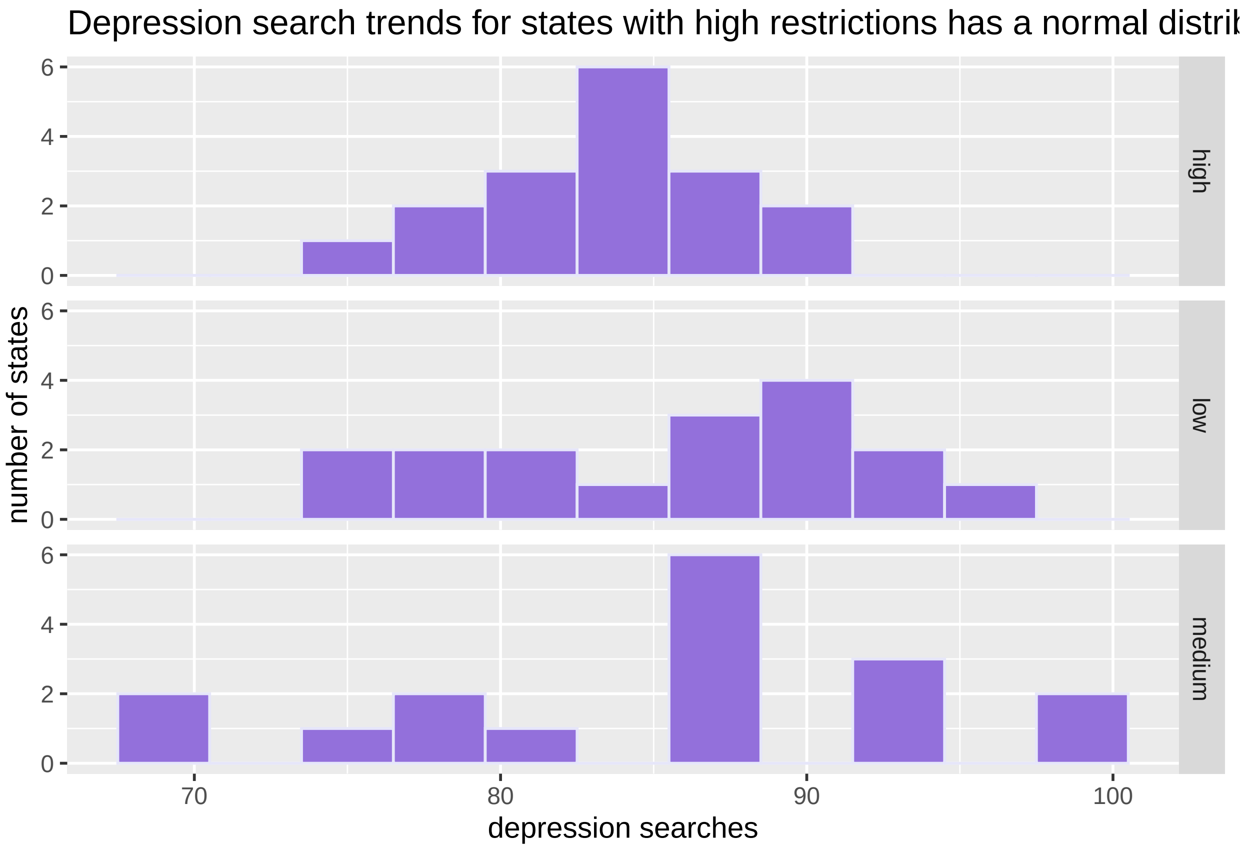
```
## Warning in wilcox.test.default(med_covid, high_covid, data = licorice,
## alternative = "two.sided", : cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  med_covid and high_covid
## W = 65.5, p-value = 0.006728
## alternative hypothesis: true location shift is not equal to 0
```

We find that the only pairwise difference in medians that is significant at the adjusted siginficance level is between medium COVID-19 case states and high COVID-19 case states. The null hypothesis for this test would be that there is no difference in the median anxiety searches of states with medium COVID-19 cases and high COVID-19 cases. The alternate hypothesis is there is a difference in the two medians. The p-value = 0.00672, and the adjusted significance level is alpha = 0.05/3 = 0.0167. We reject the null at the adjusted significance level, and conclude that there is enough evidence to suggest that there is a difference in median anxiety searches of states with medium COVID-19 cases and high COVID-19 cases.

Question3: Does the level of restriction in US states have an effect on depression and anxiety levels?

Restrictions vs. state depression trends

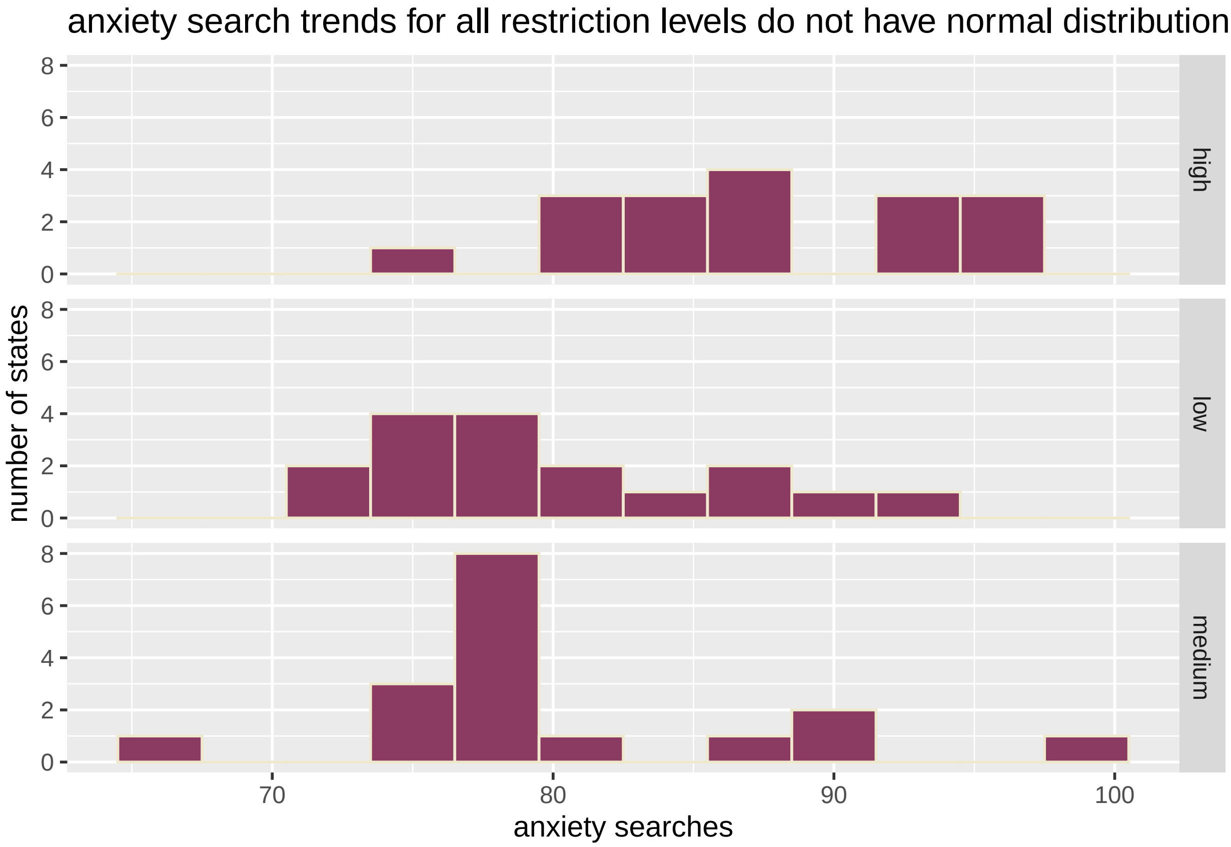


Looking at the graphs, outcomes within groups are not normally distributed for low and medium restriction level states, so this assumption is not satisfied. It also looks like the within-group variance among all groups is not the same, so the assumption for homoscedastic variance is not satisfied. The samples are also not independent because states with similar values that live close to each other may have similar anxiety search trends. The assumptions for ANOVA are not satisfied. We therefore decided to use the Kruskal-Wallis test.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  depression by restriction_cat
## Kruskal-Wallis chi-squared = 1.723, df = 2, p-value = 0.4225
```

The null is that there is no significant difference between the median depression trends of states with low restrictions, medium restrictions, and high restrictions. The alternate hypothesis is that there exists at least one median that is different. Assuming the null hypothesis is true, the model follows a chi-square distribution with a df of 2. The chi-square statistic is 1.723 with a df = 2, and the corresponding p-value is 0.4225. Therefore, we can not reject the null under the alpha = 0.05 significance level. There is not enough evidence to suggest that there is at least one difference in median depression trends of states with low, medium, and high restrictions.

Restrictions vs. state anxiety trends



Looking at the graphs, outcomes within groups are not normally distributed for any level of COVID cases, so this assumption is not satisfied. It also looks like the within-group variance among all groups is not the same, so the assumption for homoscedastic variance is not satisfied. The samples are also not independent because states with similar values that live close to each other may have similar anxiety search trends. The assumptions for ANOVA are not satisfied. We therefore decide to use the Kruskal-Wallis test.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  anxiety by restriction_cat
## Kruskal-Wallis chi-squared = 12.719, df = 2, p-value = 0.00173
```

The null is that there is no significant difference between the median anxiety trends of states with low restrictions, medium restrictions, and high restrictions. The alternate hypothesis is that there exists at least one median that is different. Assuming the null hypothesis is true, the model follows a chi square distribution with a df of 2. The chi square statistic is 12.719, and the corresponding p-value is 0.00173. Therefore, we reject the null under the alpha = 0.05 significance level. There is enough evidence to suggest that there is at least one difference in median anxiety trends of states with low, medium, and high restrictions.

Since the overall Kruskal-Wallis test was significant, we then performed step down to identify where the differences are. The appropriate stepdown test is the Wilcox rank sum test. To account for multiple comparison, we will perform the Bonferroni correction and assess our results.

```
low_rest <- trends %>%
  filter(restriction_cat == "low") %>%
  select(anxiety) %>%
  pull()

med_rest <- trends %>%
  filter(restriction_cat == "medium") %>%
  select(anxiety) %>%
  pull()

high_rest <- trends %>%
  filter(restriction_cat == "high") %>%
  select(anxiety) %>%
  pull()

wilcox.test(low_rest, med_rest, data = trends,
            alternative = "two.sided",
            paired = FALSE,
            conf.level = 0.95)
```

```
## Warning in wilcox.test.default(low_rest, med_rest, data = trends, alternative =
## "two.sided", : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  low_rest and med_rest
## W = 138, p-value = 0.8356
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(low_rest, high_rest, data = trends,
            alternative = "two.sided",
            paired = FALSE,
            conf.level = 0.95)
```

```
## Warning in wilcox.test.default(low_rest, high_rest, data = trends, alternative =
## "two.sided", : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  low_rest and high_rest
## W = 55, p-value = 0.002134
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(med_rest, high_rest, data = trends,
            alternative = "two.sided",
            paired = FALSE,
            conf.level = 0.95)
```

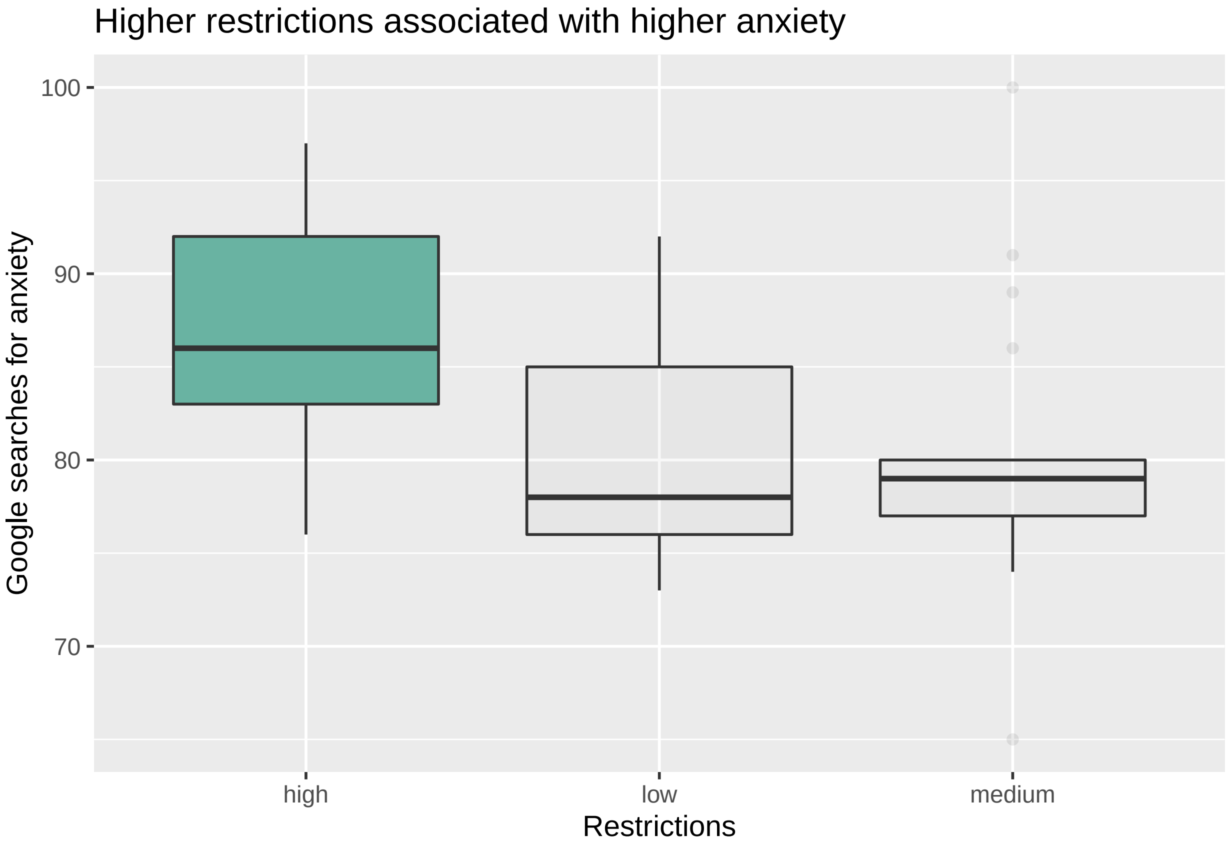
```
## Warning in wilcox.test.default(med_rest, high_rest, data = trends, alternative =
## "two.sided", : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  med_rest and high_rest
## W = 56, p-value = 0.002396
## alternative hypothesis: true location shift is not equal to 0
```

We find that there are two pairwise differences in medians that is significant at the adjusted siginficance level: low restriction states and high restrictions staes, and medium restriction states vs. high restriction states. For the Wilcoxon rank sum test for low restriction vs. high restriction states, the p-value = 0.00213. With the adjusted significance level is alpha = 0.05/3 = 0.0167, we reject the null at the adjusted significance level, and conclude that there is enough evidence to suggest that there is a difference in median anxiety searches of states with low restrictions and high restrictions. Similarly, the Wilcoxon rank sum test for medium restriction vs. high restriction states had a p-value = 0.00239. We reject the null at the adjusted significance level alpha = 0.05/3, and conclude that there is enough evidence to suggest that there is a difference in median anxiety searches for medium restriction vs. high restriction states.

Association between "anxiety" searches and restriction severity in April 2020



The above boxplot shows that the states with the highest COVID-related restrictions also record the highest google searches for anxiety compared to the states with low or moderate COVID-related restrictions. We notice that the highlighted box corresponding to the high restrictions states shows the highest median (approximately 87 compared to 78 and 79) and the highest maximum (approximately 97) and minimum (Approximately 76) upon excluding the extreme outliers in the medium group.

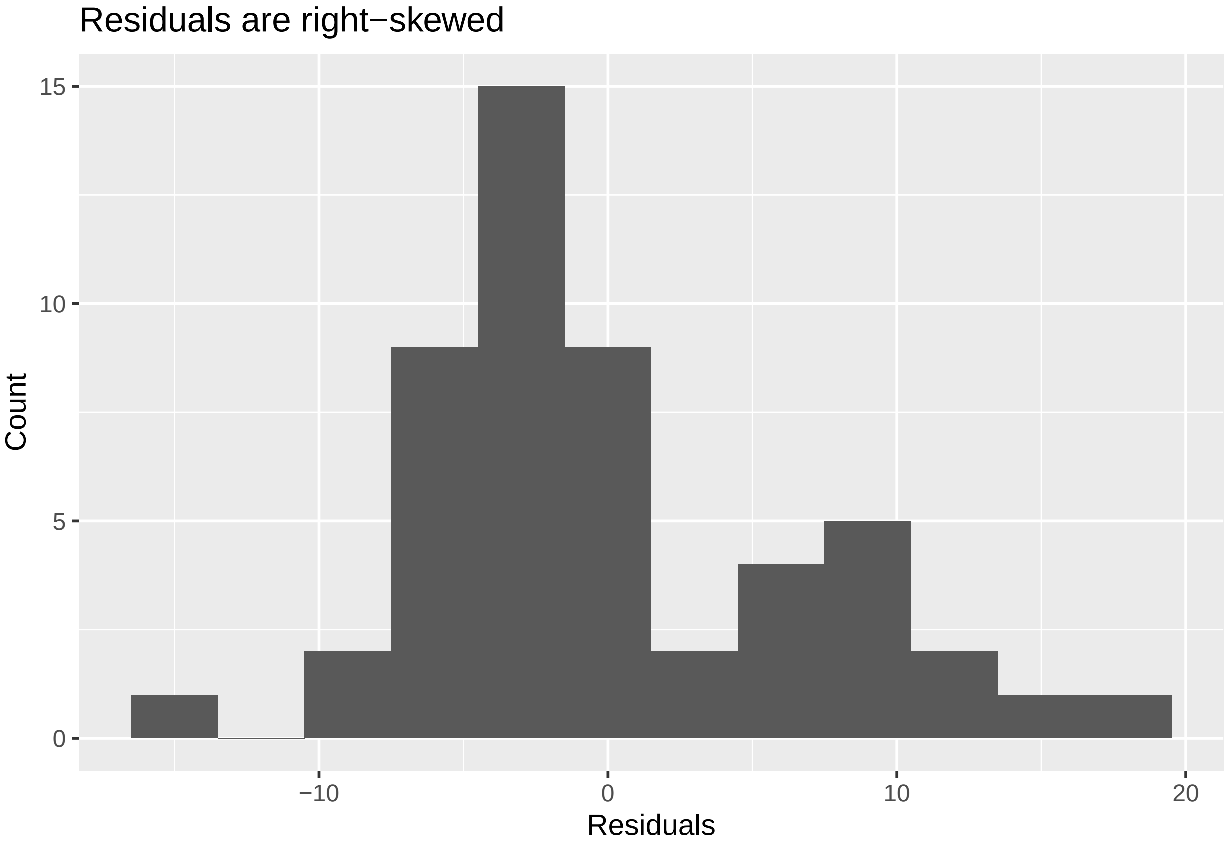
Verifying results

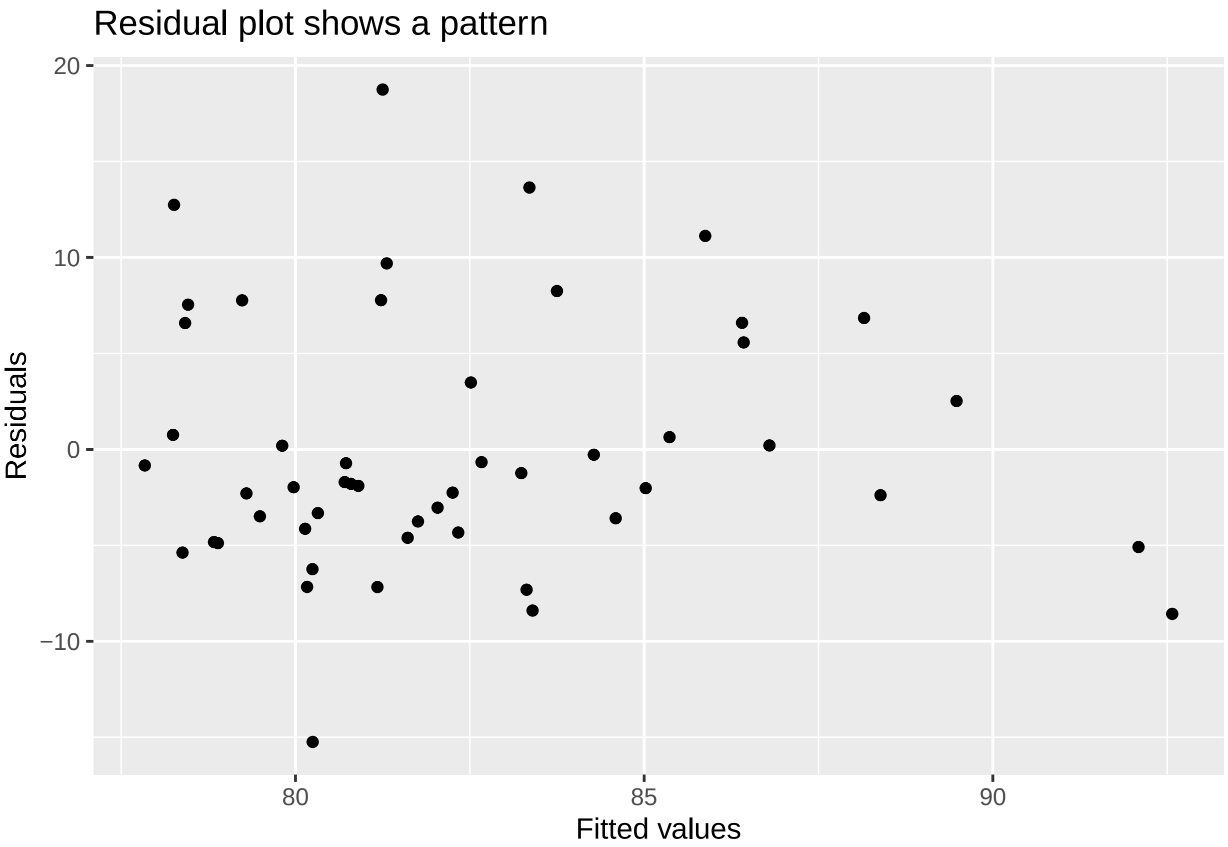
Assumptions for the regression model

1. Independence: the independence assumption does not hold in the above model because one observation directly affects others; for instance, states that share a similar culture, for example states of the Deep South, record similar anxiety rates. Also, since COVID-19 is an infectious disease, states in the same geographic region affect one another. In addition, states of the same political affiliation probably have similar restriction laws.

2. Normal distribution of residuals: The histogram of the residuals demonstrates that normality is violated, since the distribution is right-skewed.

```
## # A tibble: 3 x 5
##   term                                 estimate std.error statistic  p.value
##   <chr>                                   <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                          77.1       1.83        42.2  1.34e-39
## 2 New.COVID.cases.per.100.000.in.April  0.00779   0.00416      1.88 6.68e- 2
## 3 restriction                           0.143     0.0704       2.04 4.73e- 2
```





3.Linearity 4. Constant variance

The residual plot suggests that both the linearty and constant variance assumptions are violated; the residual plot is not symmetric about the x-axis and the dots are closer to each other below the x-axis and are more scattered above it.

Regression Model

The plots above show a positive correlation between both the new COVID-19 cases and the severity of restrictions with google searches for anxiety relative to state. However, we're yet to find out whether this correlation is significant.

```
## # A tibble: 3 x 5
##   term                                 estimate std.error statistic  p.value
##   <chr>                                   <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                          77.1       1.83        42.2  1.34e-39
## 2 New.COVID.cases.per.100.000.in.April  0.00779   0.00416      1.88 6.68e- 2
## 3 restriction                           0.143     0.0704       2.04 4.73e- 2
```

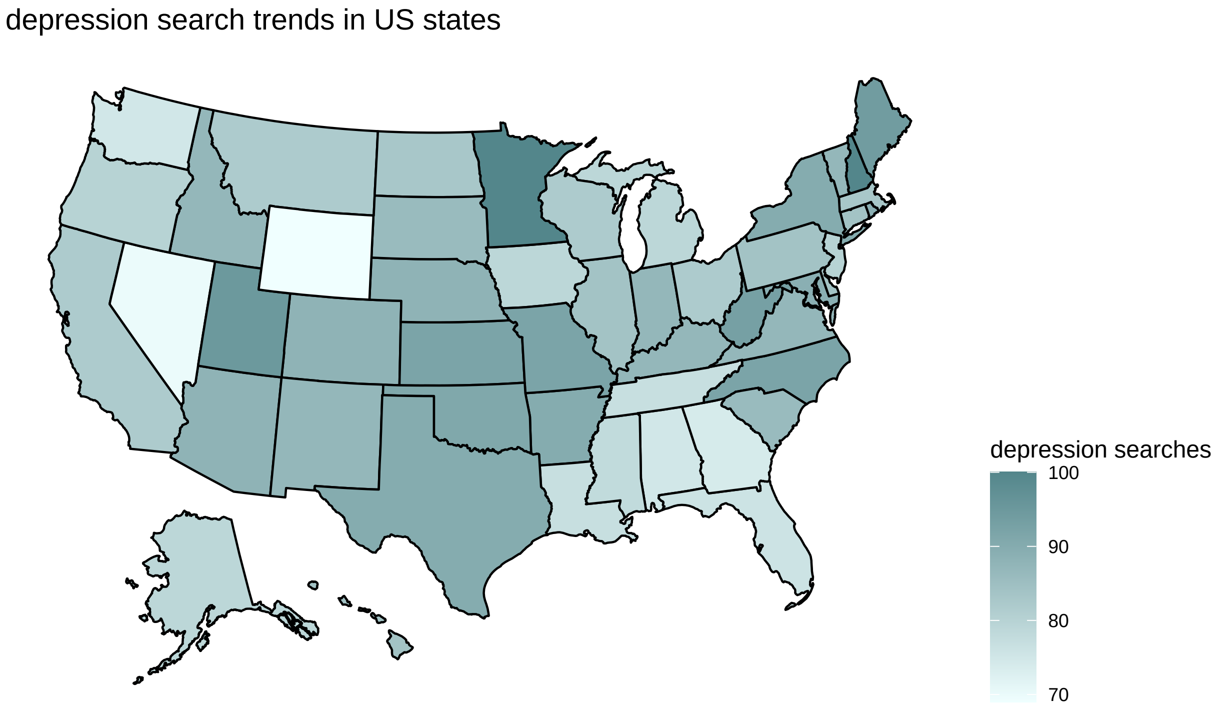
The above chart supports the positive correlations we found earlier. The slope estimate corresponding to new covid cases per 100,000 in the month of april is 0.007794587. This means that for each additional increase in new COVID-19 cases, we would expect a 0.007794587 unit increase in anxiety score, holding restrictions constant. Similarly, the slope estimate corresponding to restrictions is also positive 0.143335319. This means that for each additional increase restrictions rank, we would expect a 0.143335319 unit increase in anxiety score, holding new COVID-19 cases constant. The p-value corresponding to the estimated slope of restrictions is approximately 0.047, which is less than our significant level of 0.05. Thus, the positive correlation between restrictions and google searches corresponding to anxiety is statistically significant. On the other hand, the p-value corresponding to the estimated slope of new COVID-19 cases per 100,000 is approximately 0.067, which is higher than our significant level of 0.05. Thus, the positive correlation between new COVID-19 cases per 100,000 and google searches corresponding to anxiety is not statistically significant.

Depression rate in each state map

```
## Warning: Use of `map_df$x` is discouraged. Use `x` instead.
```

```
## Warning: Use of `map_df$y` is discouraged. Use `y` instead.
```

```
## Warning: Use of `map_df$group` is discouraged. Use `group` instead.
```



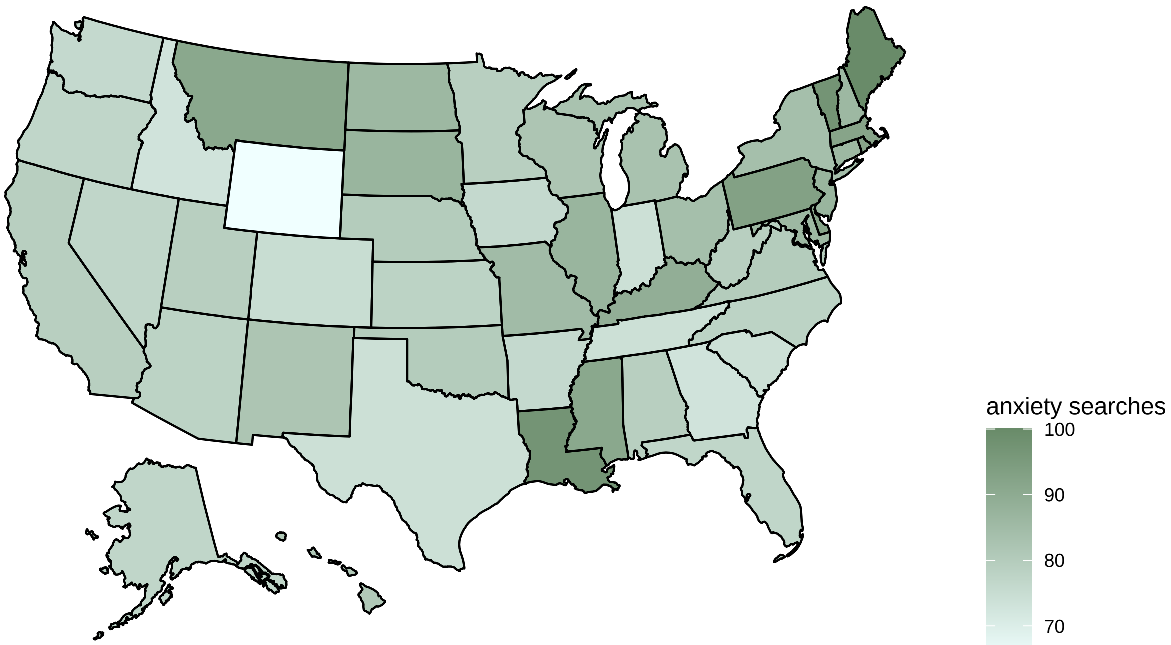
Depression rate in each state map

```
## Warning: Use of `map_df$x` is discouraged. Use `x` instead.
```

```
## Warning: Use of `map_df$y` is discouraged. Use `y` instead.
```

```
## Warning: Use of `map_df$group` is discouraged. Use `group` instead.
```

anxiety search trends in US states


anxiety searches
100
90
80
70