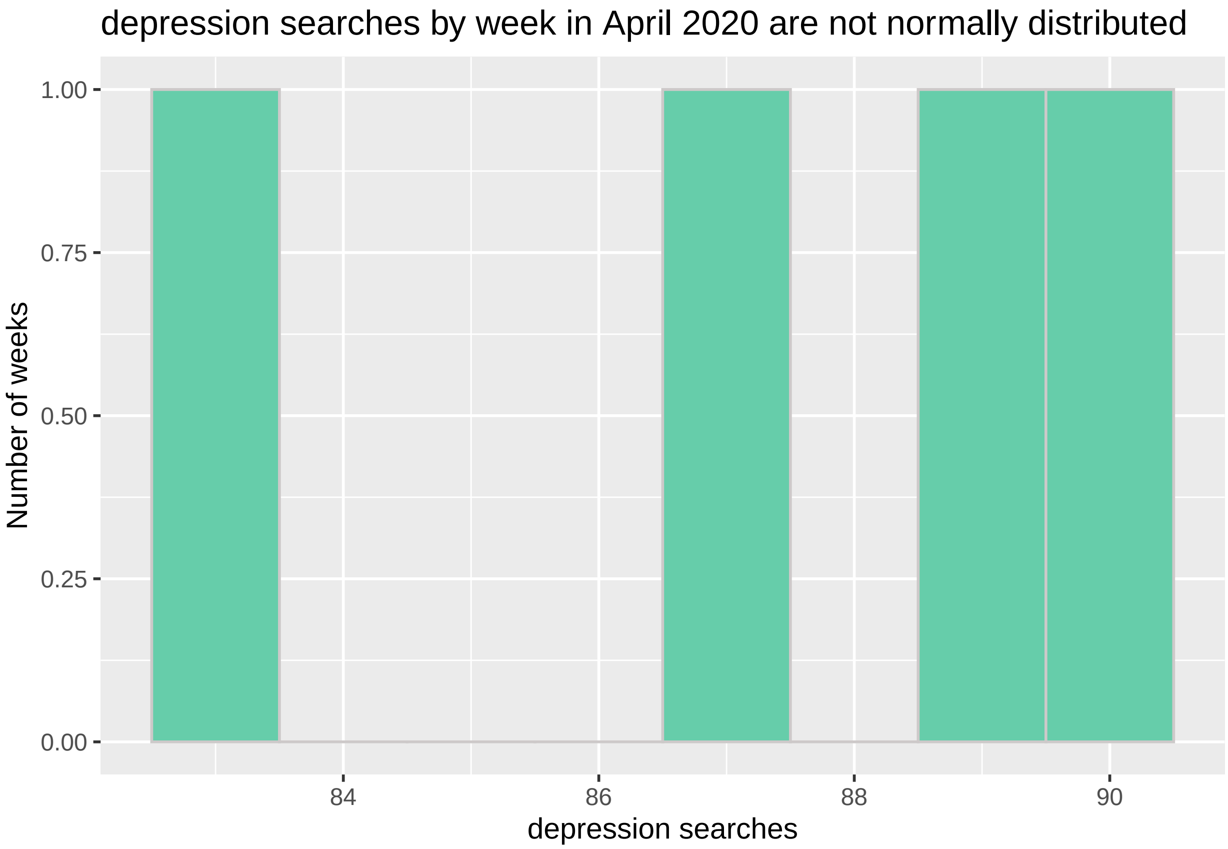
Final Project

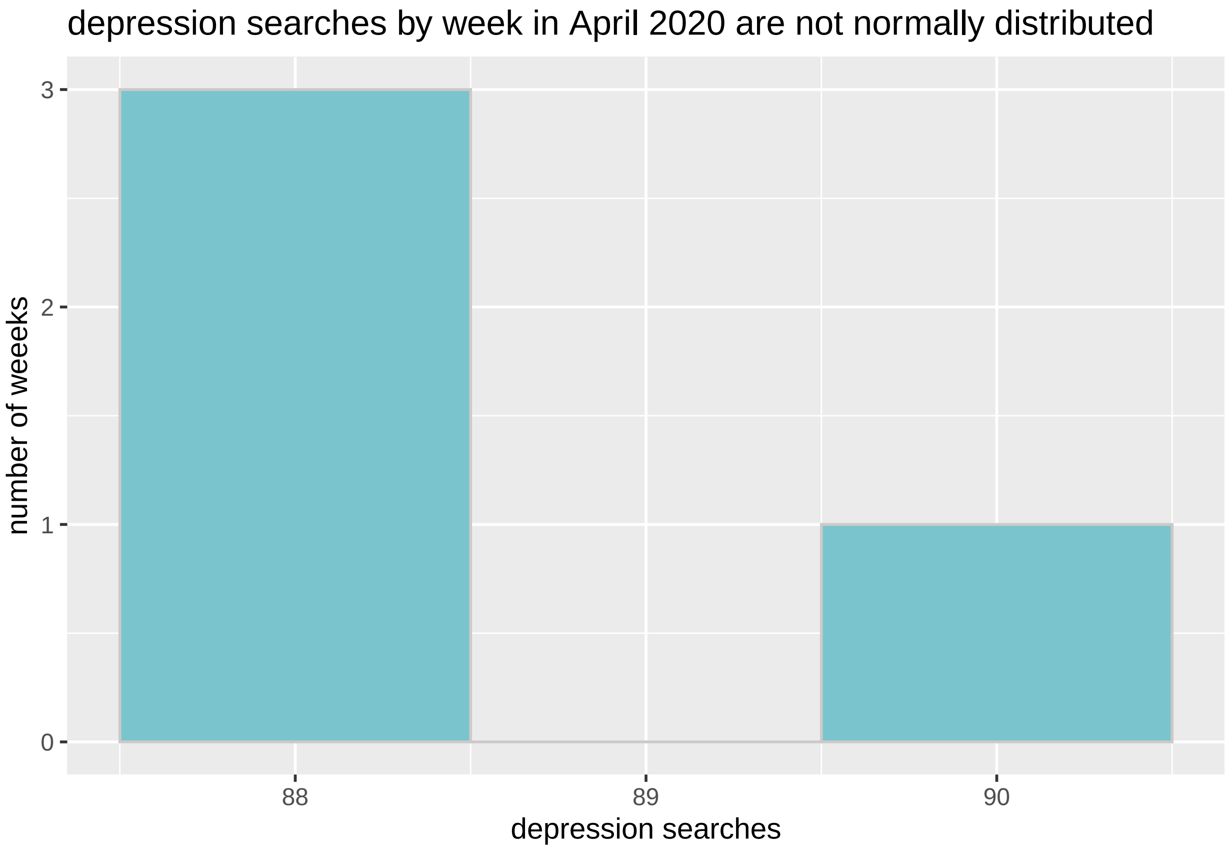
Brenda Yang, Charlie Bonetti, Nour Kanaan

7/19/2020

Comparing depression trends between April 2020 and April 2018



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

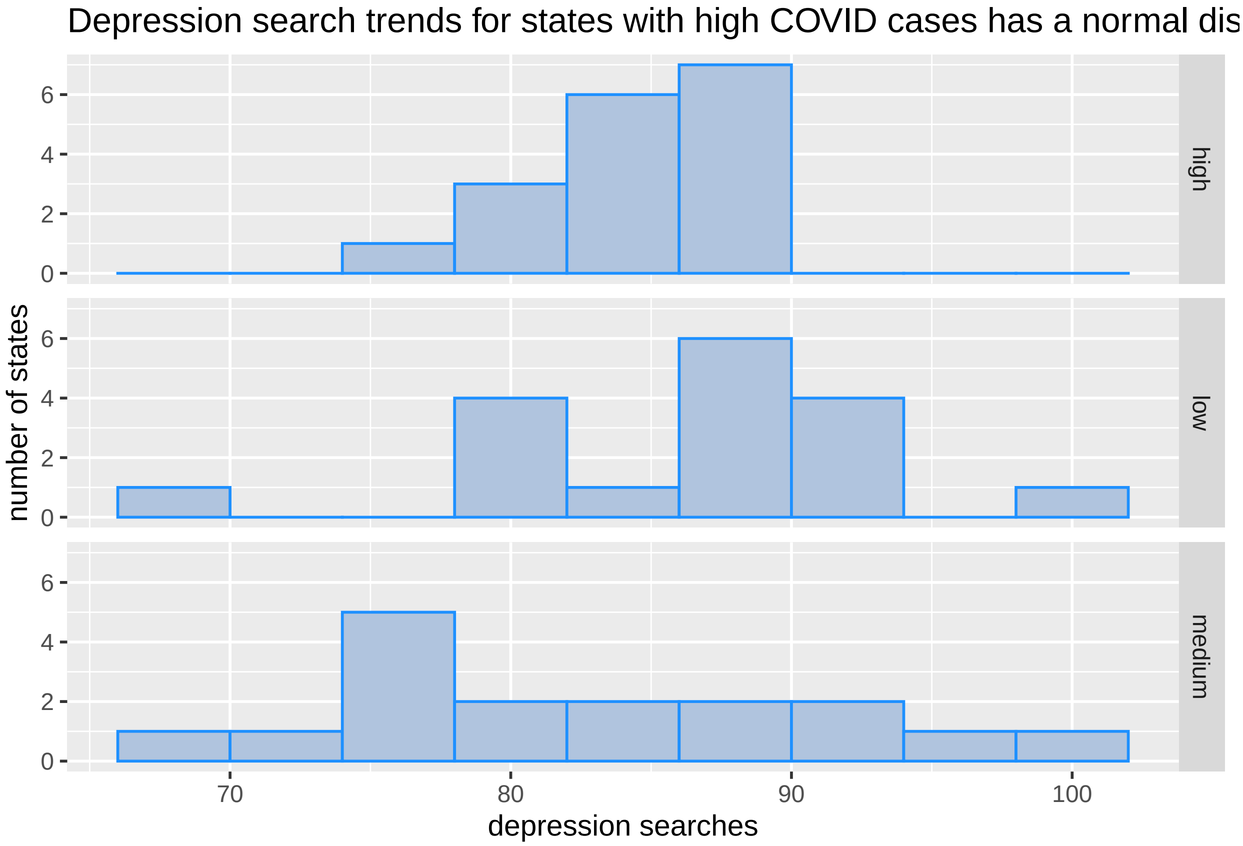


n<30 and not normal distribution: assumption for t-test not satisfied

```
##
##  Paired t-test
##
## data:  d2020 and d2018
## t = 0.62017, df = 3, p-value = 0.5791
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.164426  7.664426
## sample estimates:
## mean of the differences
##                    1.25
```

The null hypothesis is that there is no difference in the mean amount 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 t-distribution. The t-statistic is 0.755 and the df = 29. This corresponds to a p-value of 0.4561. 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 mean amount of depression searches in the US between the times of April 2020 and April 2018.

COVID cases vs. depression rate

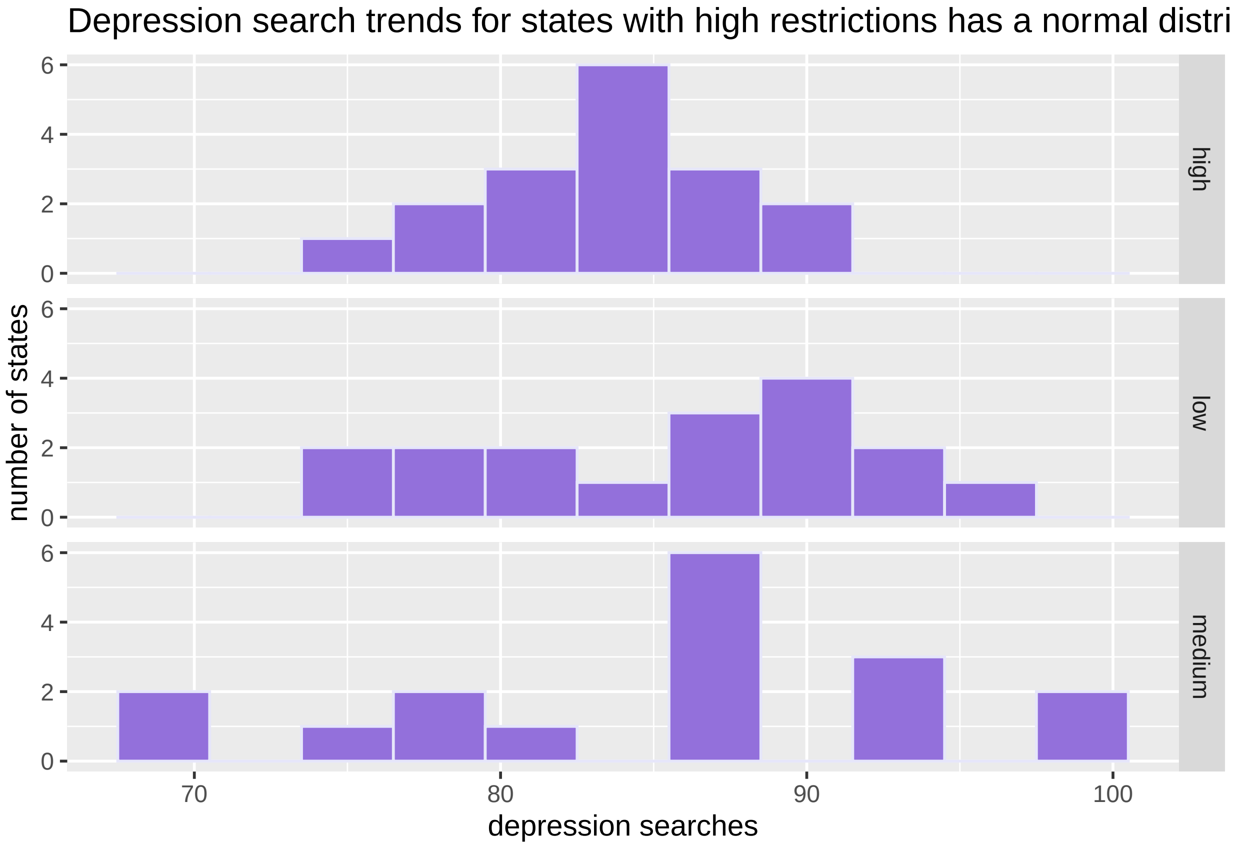


The outcomes within each group is not normal. The depression search trends for states with medium numbers of COVID cases and low numbers of COVID cases do not have a normal distribution, and n<30. Therefore, this assumption is not satisfied. By looking at the graphs, it also seems that there is not equal 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. Therefore, the assumptions for ANOVA are not satisfied.

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## case_cat     2  121.5   60.73   1.334  0.273
## Residuals   48 2185.2   45.52
```

The null is that there is no significant difference between the mean 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 mean that is different. Assuming the null hypothesis is true, the model follows an F distribution with a df of 2. The F-statistic is 1.334, and the corresponding p-value is 0.273. 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 mean depression trends of states with low, medium, and high COVID cases.

Restrictions vs. depression



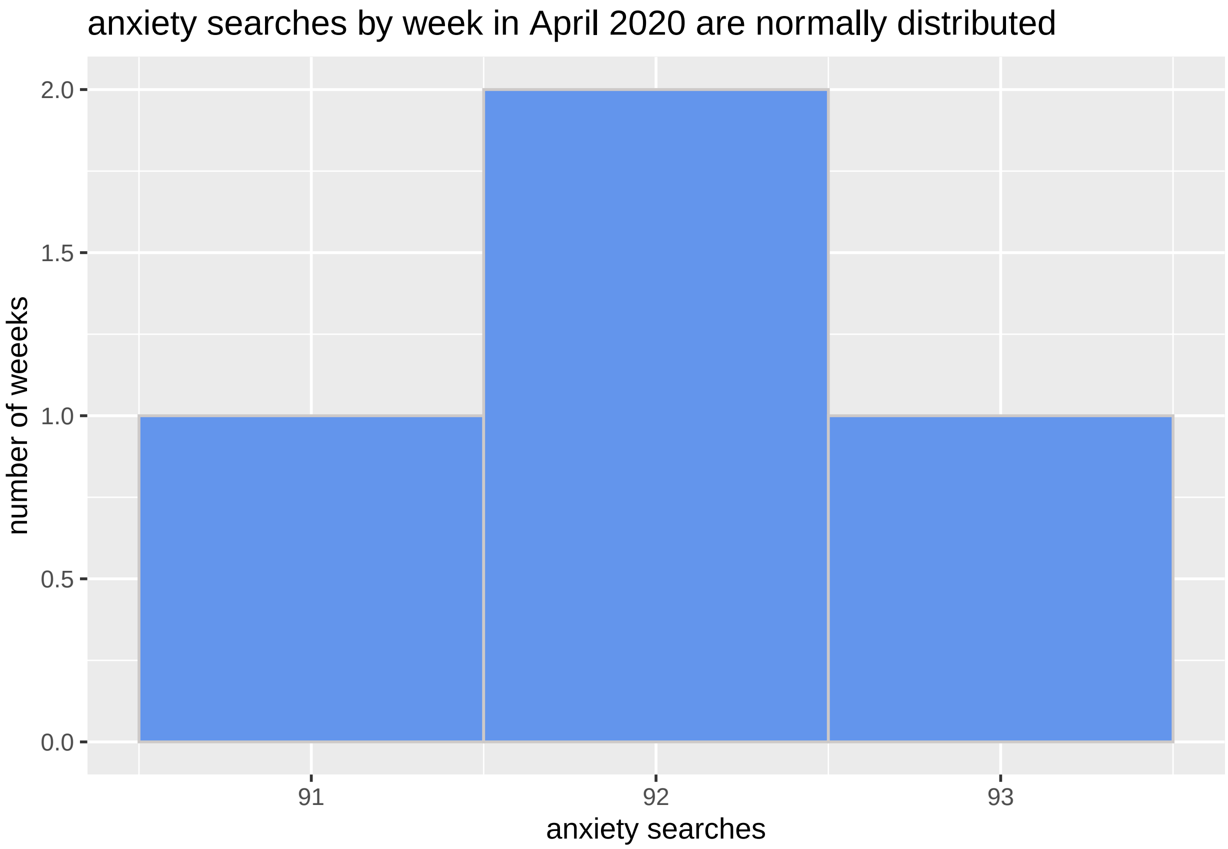
Looking at the graphs, outcomes within groups are not normally distributed for low and medium restriction level states, so this assumption is not satisfied. It also looks like the within-group variance among all groups is not the same, so the assumption for homoscedastic variance is not satisfied. The 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.

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## restriction_cat  2   48.5   24.25   0.516    0.6
## Residuals       48 2258.1   47.04
```

The null is that there is no significant difference between the mean depression trends of states with low restrictions, medium restrictions, and high restrictions. 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 F-statistic is 0.516, and the corresponding p-value is 0.6. 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 mean depression trends of states with low, medium, and high restrictions.

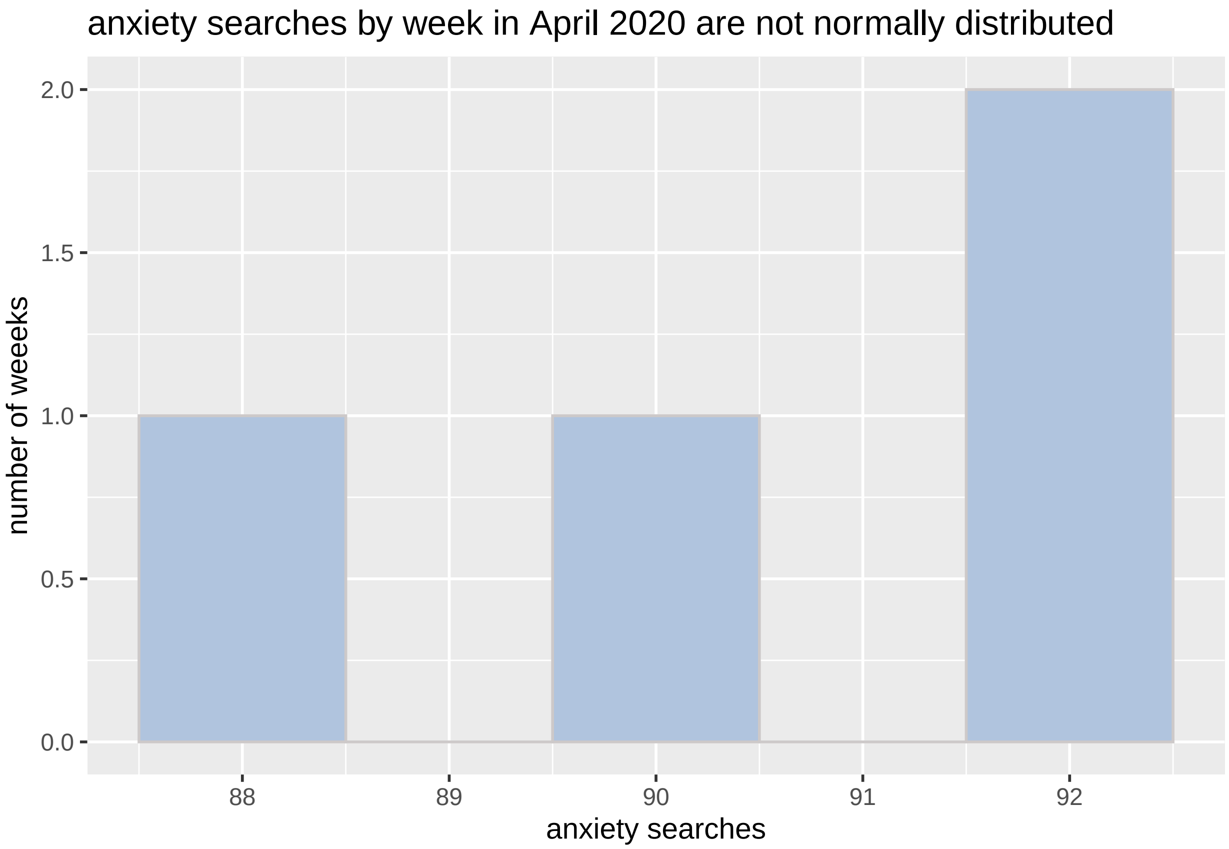
Comparing anxiety trends in April 2020 to April 2018

```
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")
```

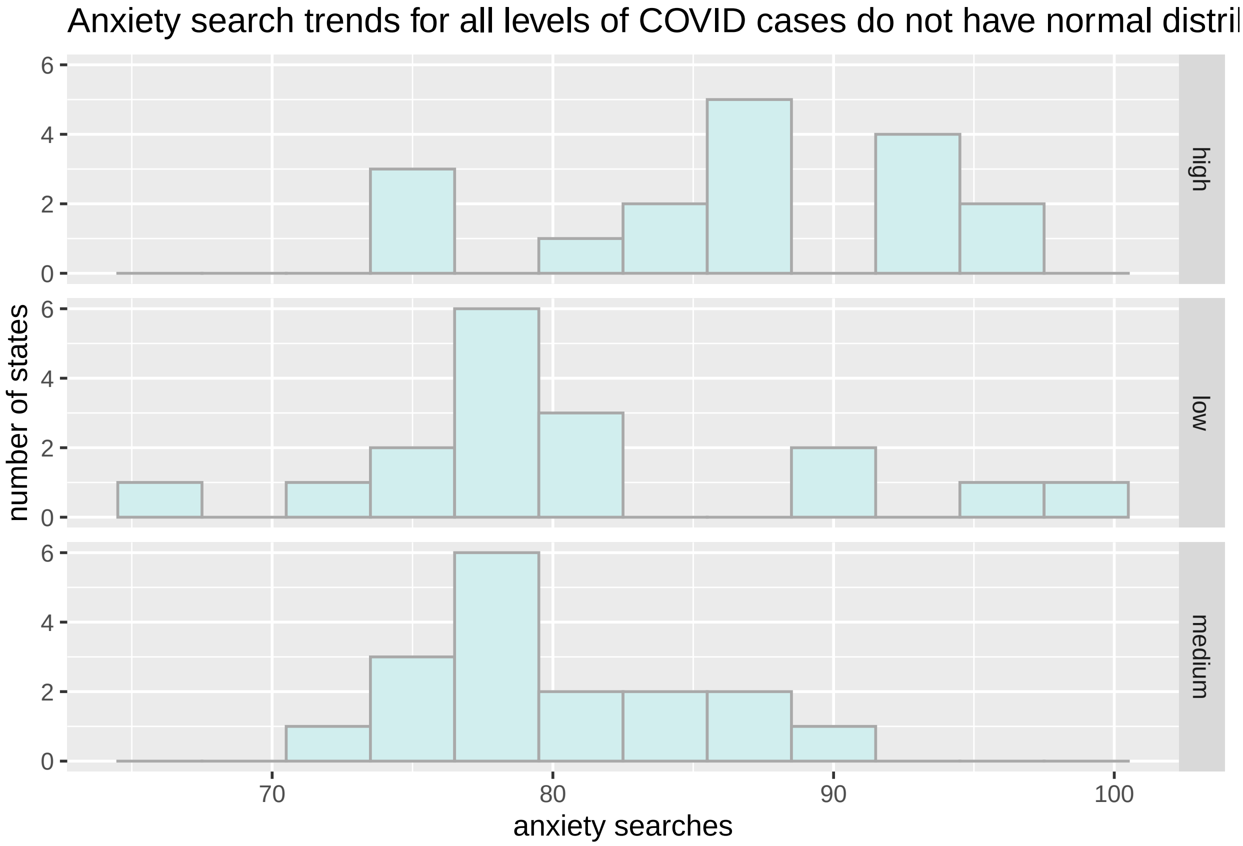


n<30 and not normal distribution: assumption not satisfied

```
##
##  Paired t-test
##
## data:  a2020 and a2018
## t = -1.2603, df = 3, p-value = 0.2967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.287869  2.287869
## sample estimates:
## mean of the differences
##                    -1.5
```

The null hypothesis is that there is no difference in the mean 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 means. Assuming that the null hypothesis is true, the model follows a t-distribution. The t-statistic is 1.66 and the df = 29. This corresponds to a p-value of 0.1086. 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 mean amount of anxiety searches in the US between the times of April 2020 and April 2018.

COVID cases vs. anxiety rate

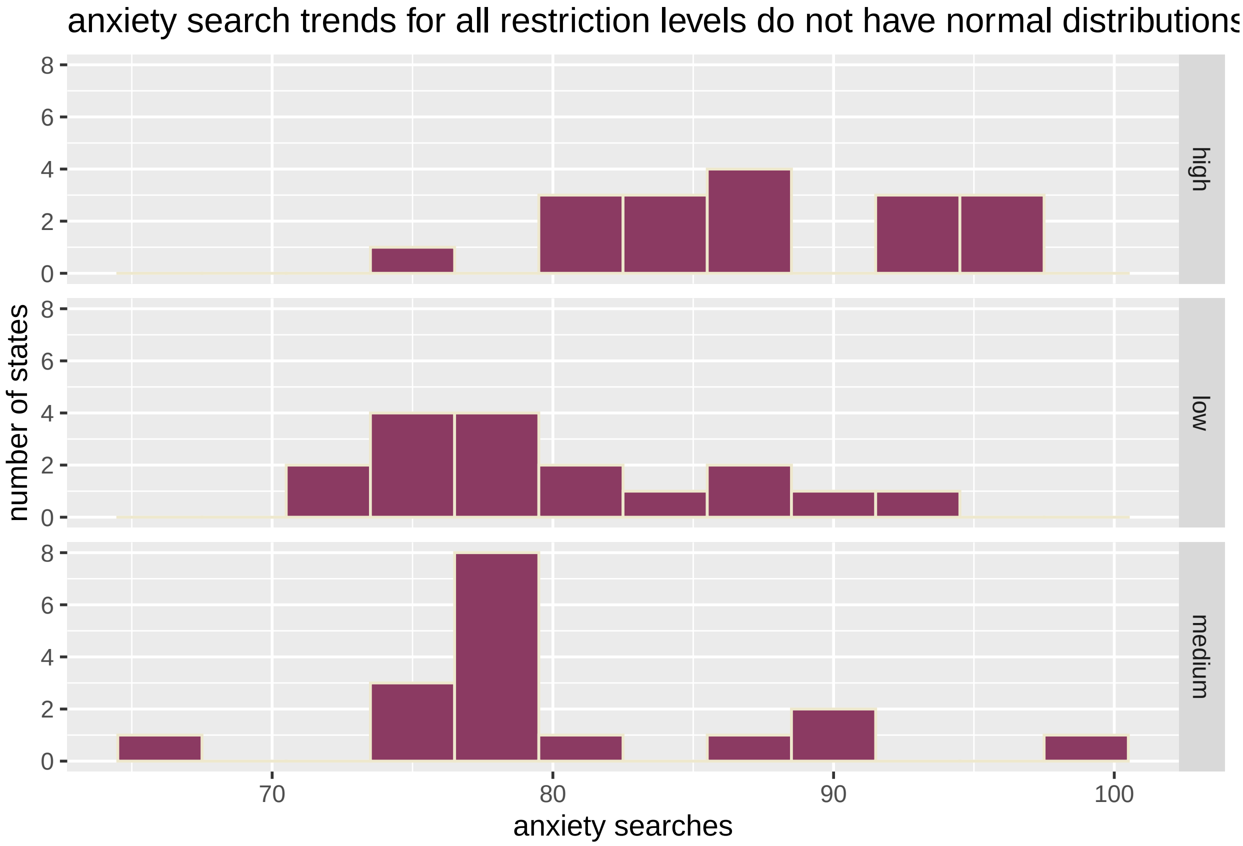


Looking at the graphs, outcomes within groups are not normally distributed for any level of COVID cases, so this assumption is not satisfied. It also looks like the within-group variance among all groups is not the same, so the assumption for homoscedastic variance is not satisfied. The samples are also 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.

```
##             Df Sum Sq Mean Sq F value Pr(>F)  
## case_cat     2  384.2  192.08   3.826 0.0287 *
## Residuals   48 2410.0   50.21                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 F-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.

Restrictions vs. anxiety



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.

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)   
## restriction_cat  2  612.9  306.43   6.743 0.00262 **
## Residuals       48 2181.3   45.44                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

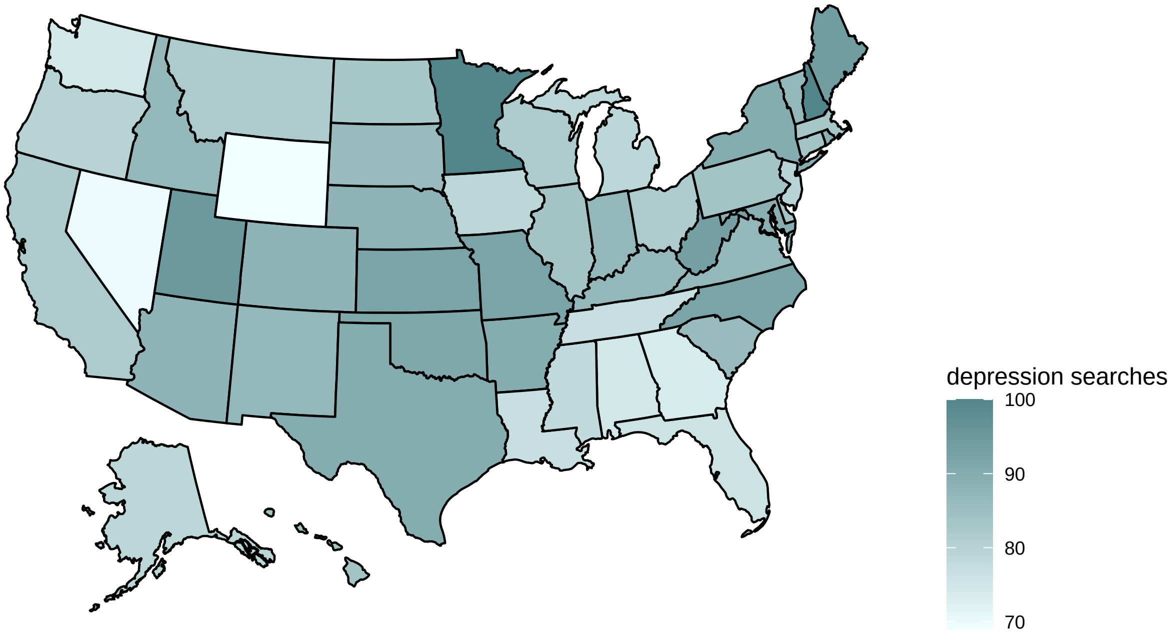
The null is that there is no significant difference between the mean anxiety trends of states with low restrictions, medium restrictions, and high restrictions. The alternate hypothesis is that there exists at least one mean that is different. Assuming the null hypothesis is true, the model follows an F distribution with a df of 2. The F-statistic is 6.746, and the corresponding p-value is 0.00262. 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 mean anxiety trends of states with low, medium, and high restrictions.

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.
```

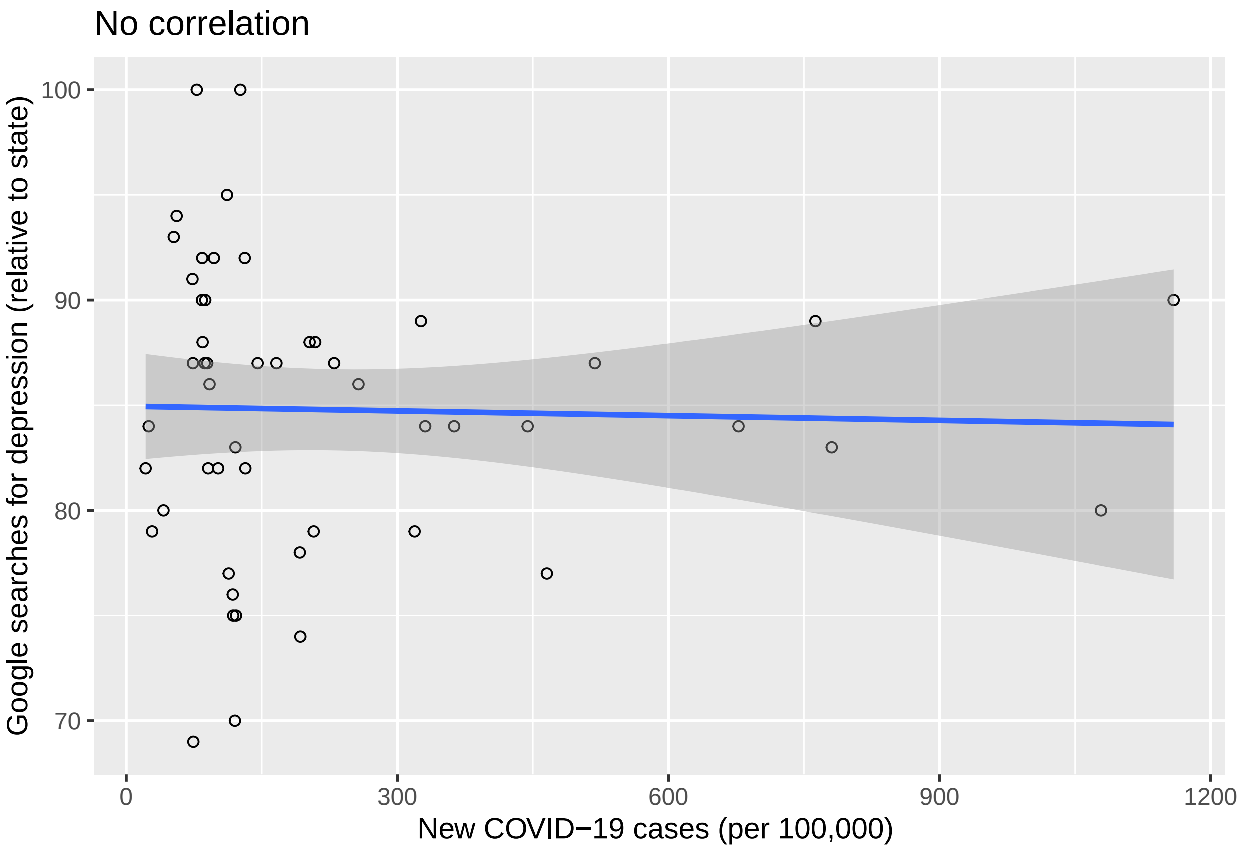


Nour Visuals

```
ggplot(trends, aes(New.COVID.cases.per.100.000.in.April, depression)) +
  geom_point(shape=1) +
   geom_smooth(method=lm) +
  labs(title = "No correlation",
       x = "New COVID-19 cases (per 100,000) ",
       y = "Google searches for depression (relative to state)" )
```

Effect of COVID on depression rates in each State

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

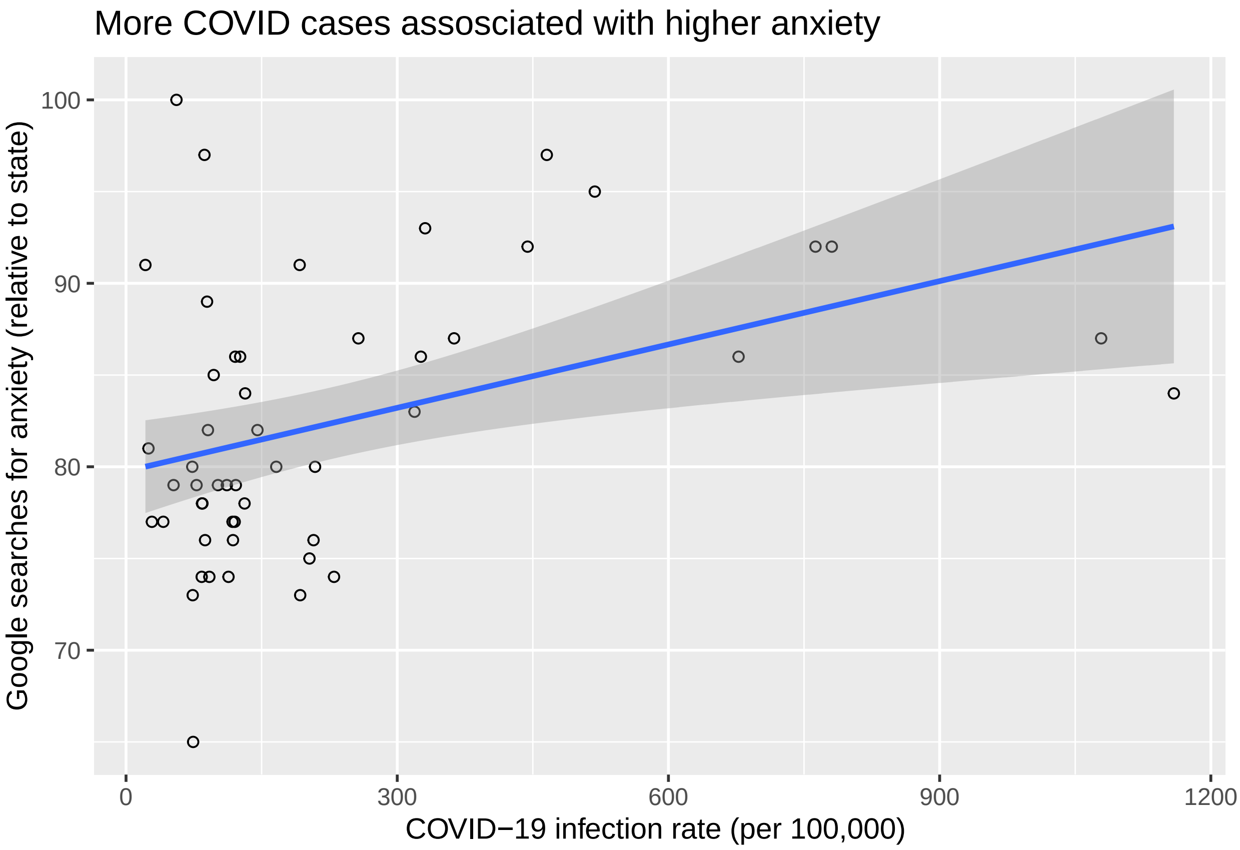


The above scatterplot shows the relative google searches for depression among the 50 american states as a function of new COVID-19 cases per 100,000 during the month of april, 2020. The line of best fit (in blue) is horizontal with R = approximately zero. Thus, we cannot see any correlation between the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

```r
ggplot(trends, aes(New.COVID.cases.per.100.000.in.April, anxiety)) +
  geom_point(shape=1) +
   geom_smooth(method=lm) +
  labs(title = "More COVID cases assosciated with higher anxiety",
       x = "COVID-19 infection rate (per 100,000) ",
       y = "Google searches for anxiety (relative to state)" )
```

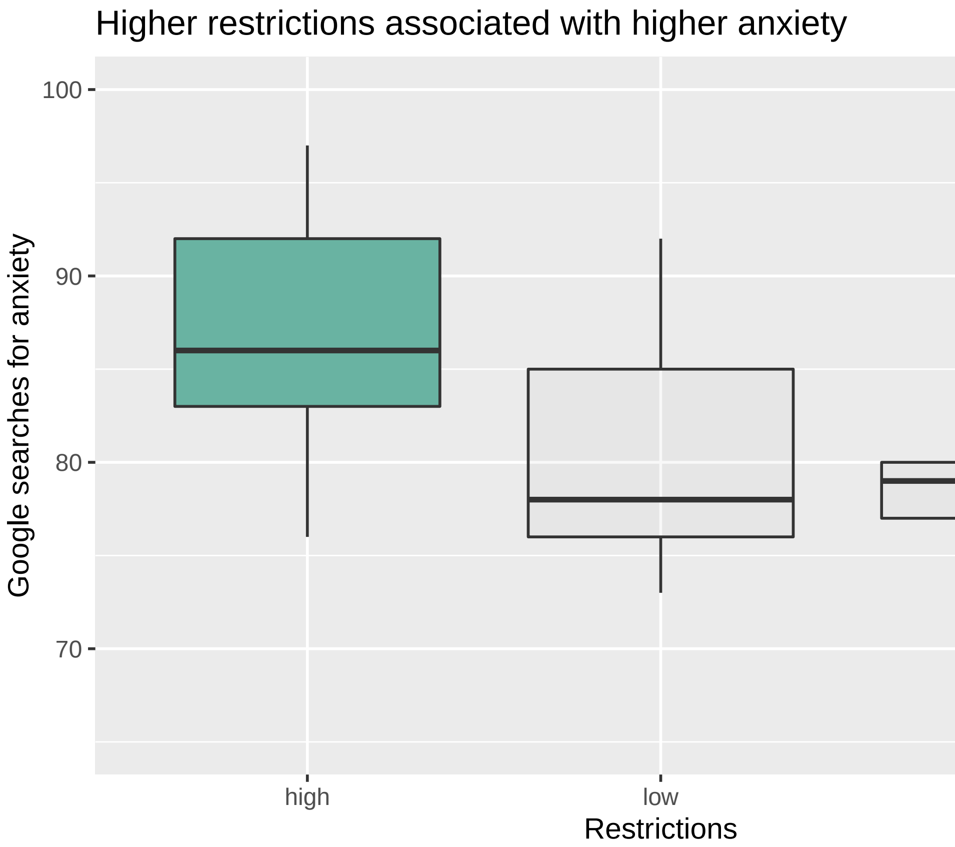
Effect of COVID on anxiety rates in each State

```
## `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.

```
trends %>%
  mutate( type=ifelse(restriction_cat=="high","Highlighted","Normal")) %>% ggplot( aes(x=restriction_ca
    geom_boxplot() +
    scale_fill_manual(values=c("#69b3a2", "grey")) +
    scale_alpha_manual(values=c(1,0.1)) +
    theme(legend.position = "none") +
    labs(title = "Higher restrictions associated with higher anxiety",
       x = "Restrictions",
       y = "Google searches for anxiety")
```



Restrictions effect on anxiety rates

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.

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 which variable have the stronger correlation. In other words, are people more anxious of new COVID cases or COVID-related restrictions?

```
Anxiety_regression <- lm(anxiety ~ New.COVID.cases.per.100.000.in.April + restriction, data = trends)
tidy(Anxiety_regression)
```

```
## # 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
```

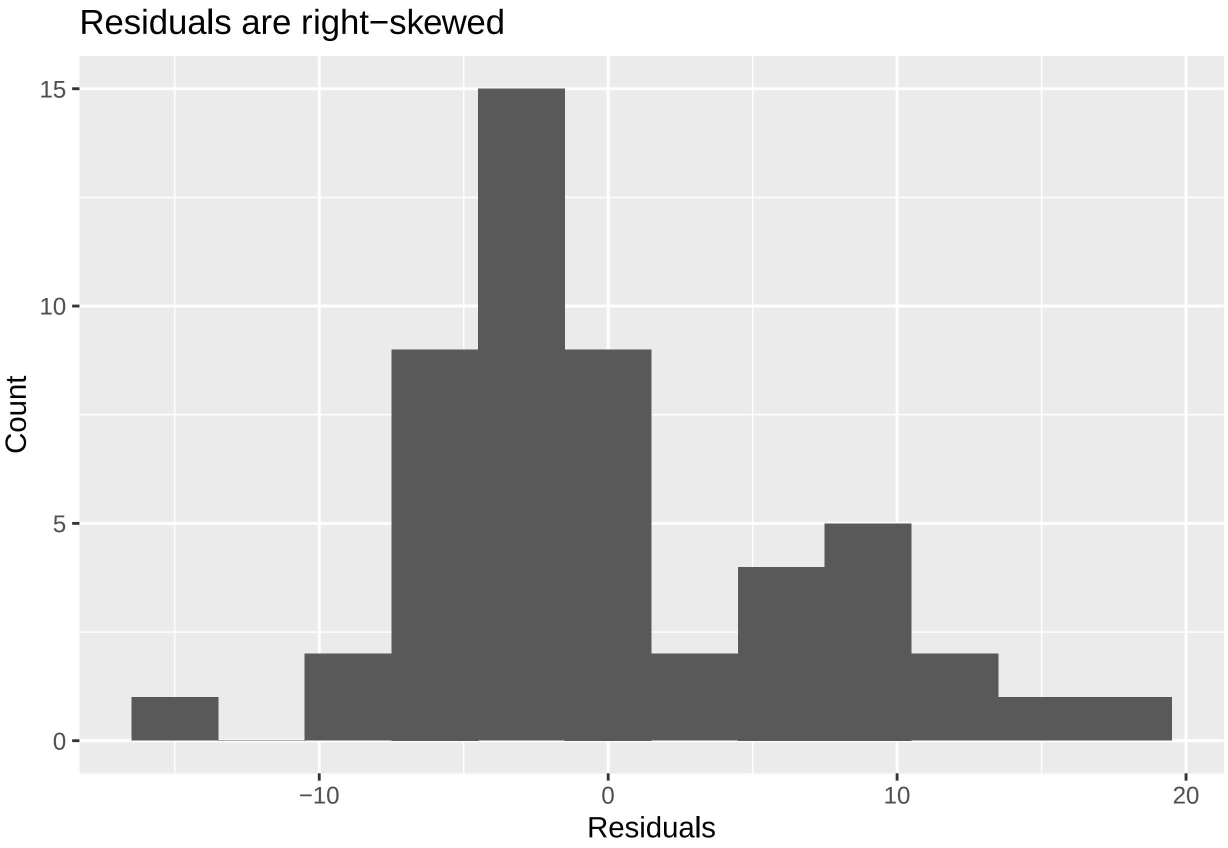
The above chart supports the positive correlations we found earlier. The slope estimate corresponding to new covid cases per 100,000 in the month of april is 0.007794587. This means that for each additional increase in new COVID-19 cases, we would expect a 0.007794587 unit increase in anxiety score, 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.

Assumptions for the above regression model

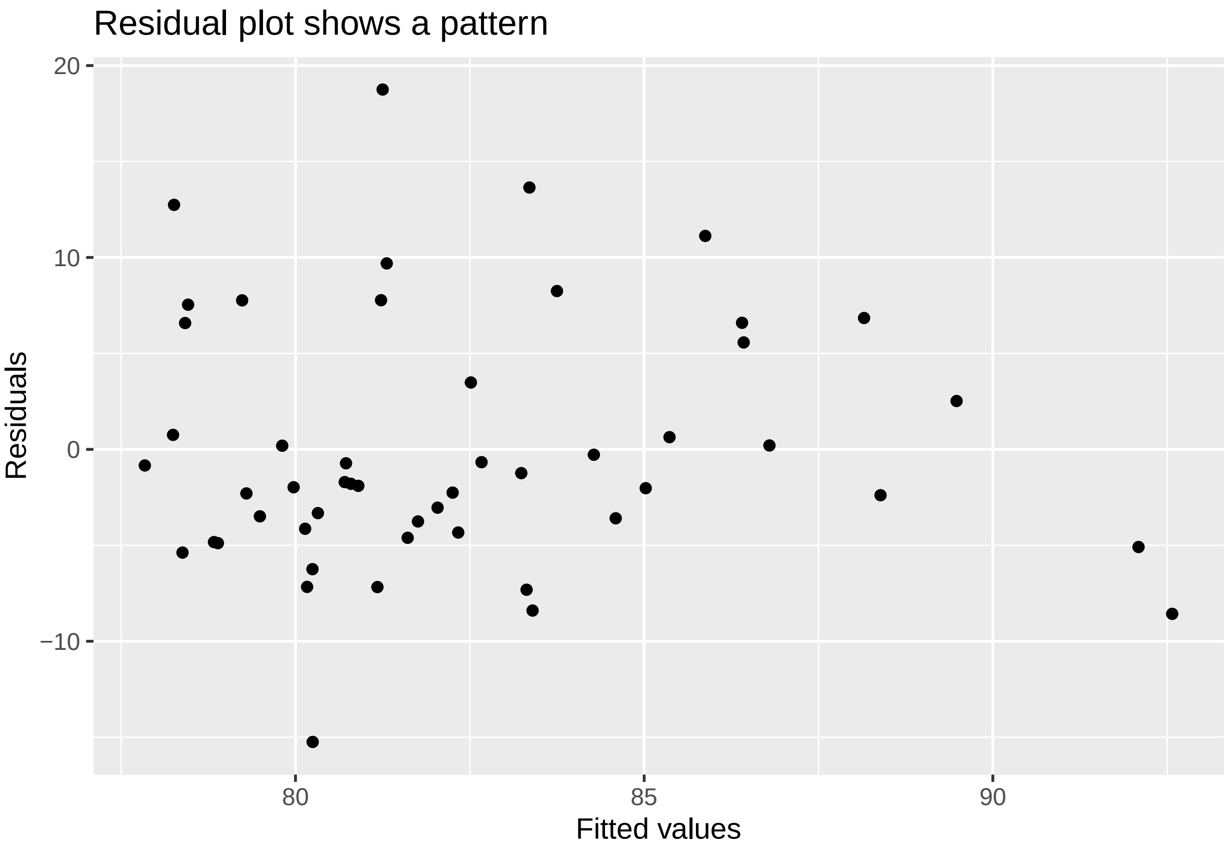
1. Independence: the independence assumption does not hold in the above model because one observation directly affects others; for instance, states that share a similar culture, for example states of the Deep South, record similar anxiety rates. Also, since COVID-19 is an infectious disease, states in 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.

```
ggplot(data = Anxiety_regression, mapping = aes(x = .resid)) +
  geom_histogram(binwidth = 3) +
  labs(x = "Residuals", y = "Count",
       title = "Residuals are right-skewed")
```



```
ggplot(data = Anxiety_regression, mapping = aes(x = .fitted, y = .resid)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residuals",
       title = "Residual plot shows a pattern")
```



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.