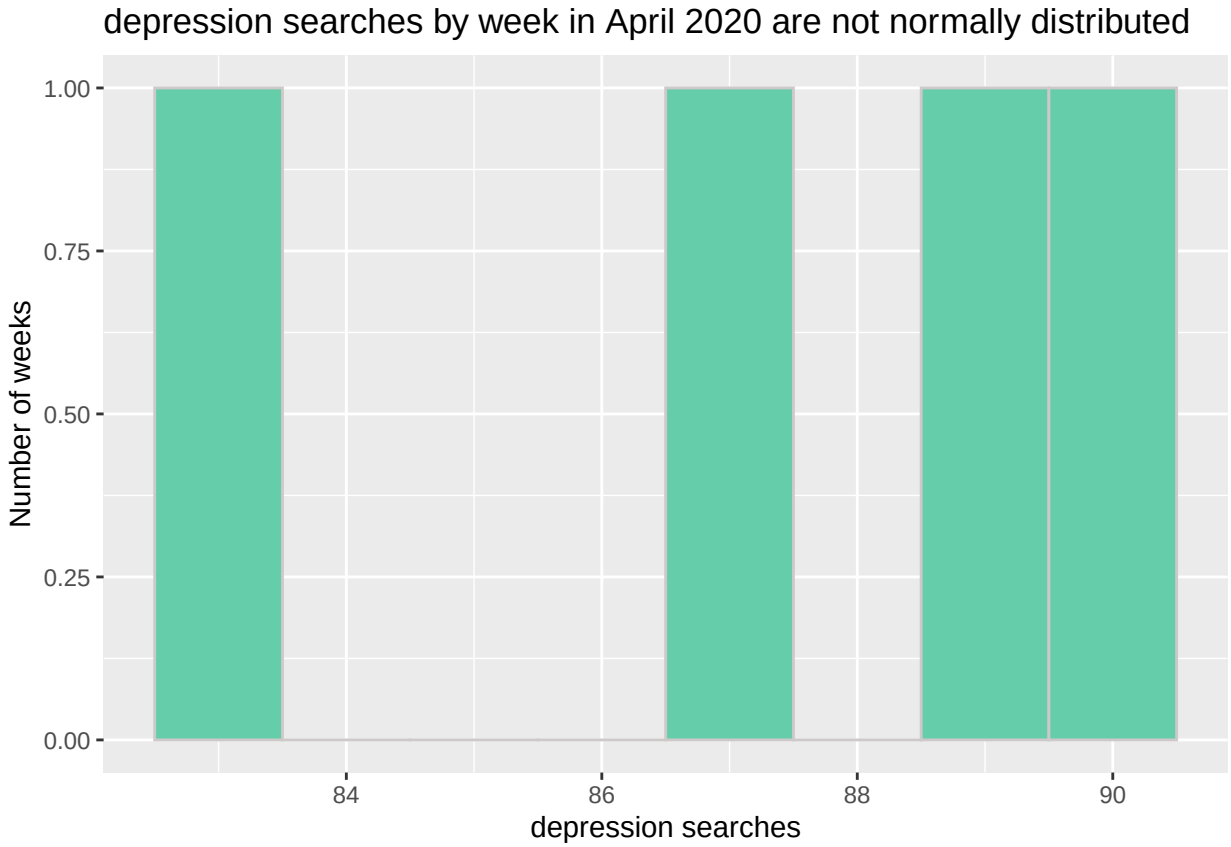
Final Project

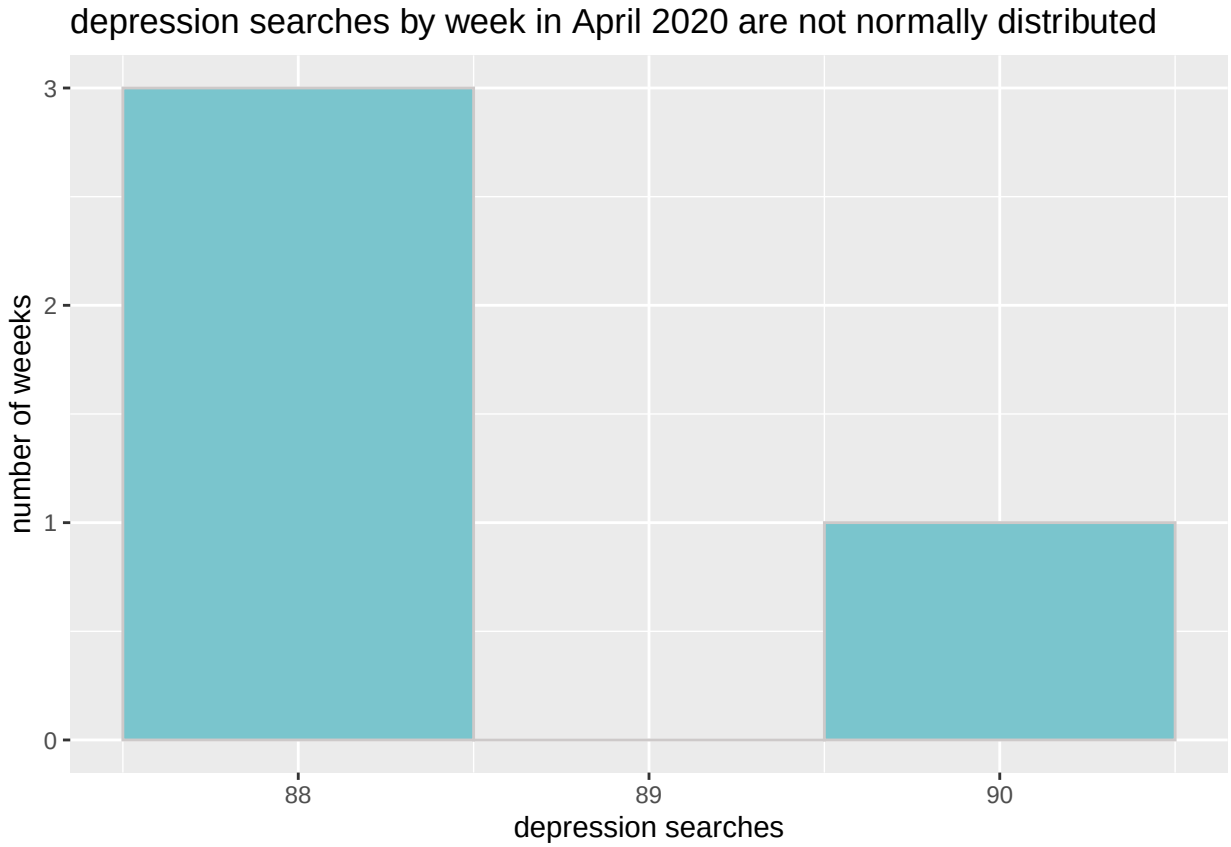
Brenda Yang, Charlie Bonetti, Nour Kanaan

7/19/2020

Comparing depression trends between April 2020 and April 2018



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

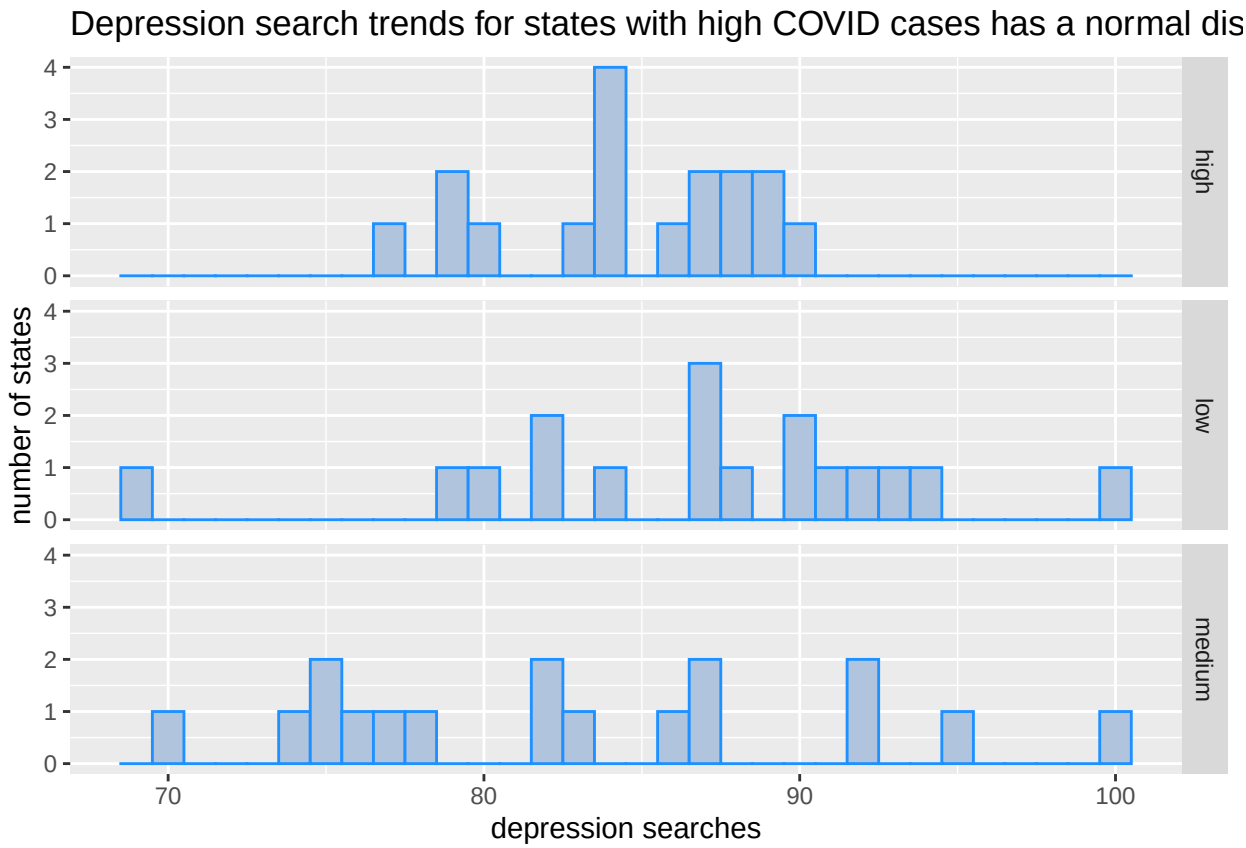


n<30 and not normal distribution: assumption for t-test not satisfied

```
##
##  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

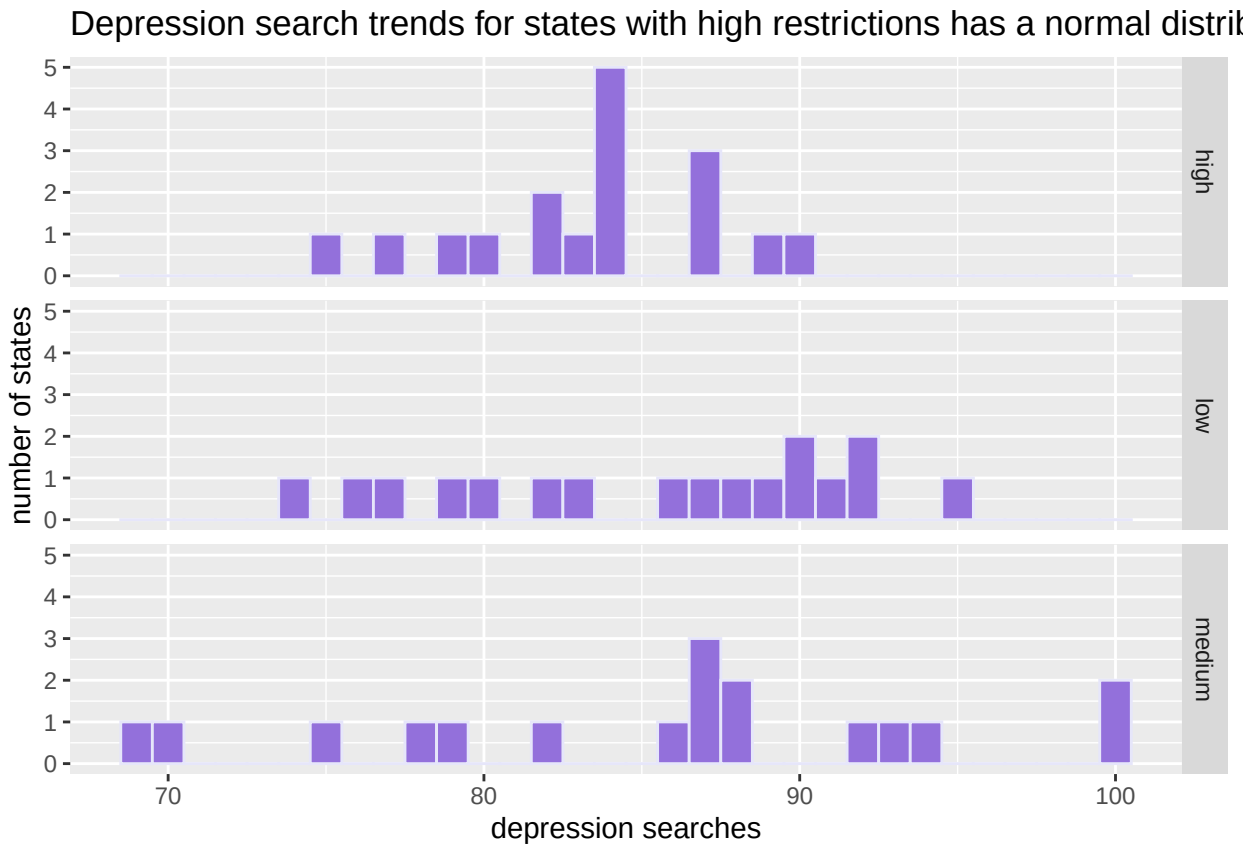


The outcomes within each group is not normal. The depression search trends for states with medium numbers of COVID cases and low numbers of COVID cases do not have a normal distribution, and n<30. Therefore, this assumption is not satisfied. By looking at the graphs, it also seems that there is not equal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



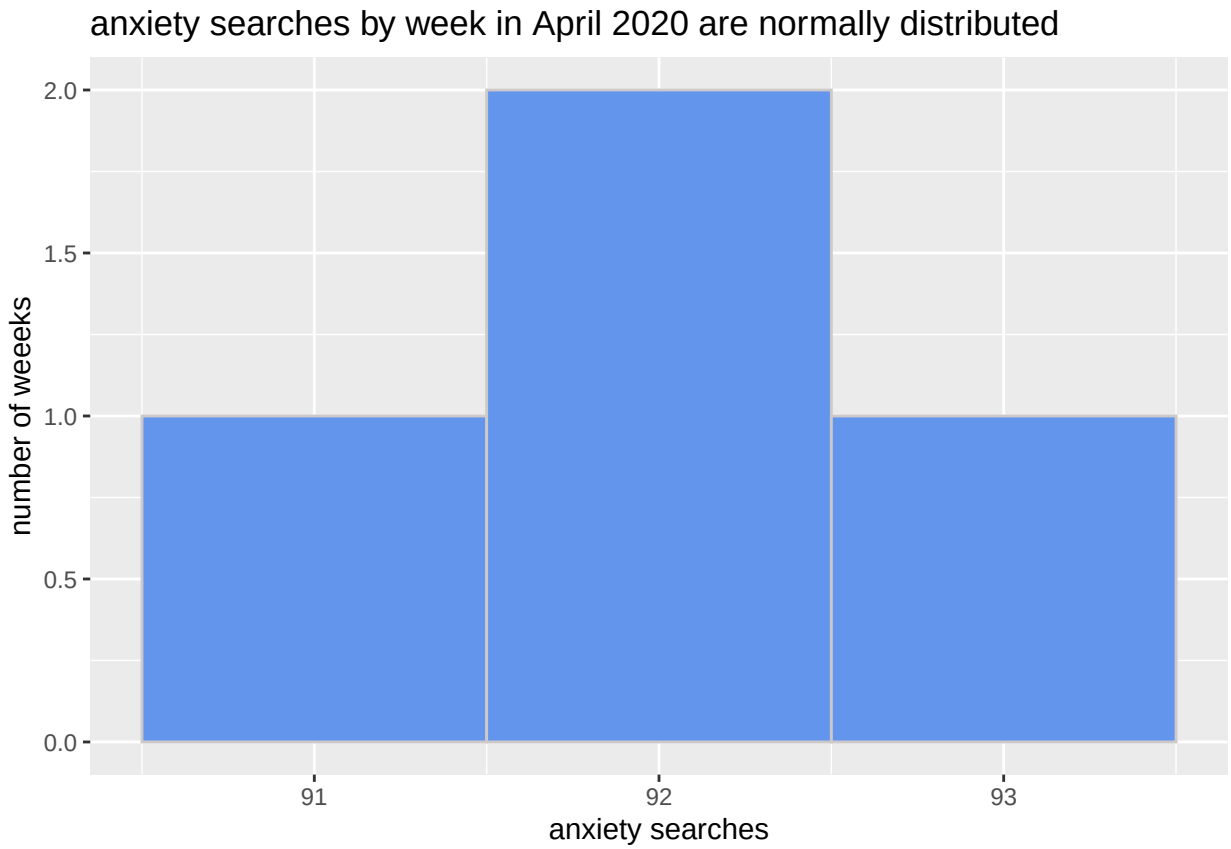
Looking at the graphs, outcomes within groups are not normally distributed for low and medium restriction level states, so this assumption is not satisfied. It also looks like the within-group variance among all groups is not the same, so the assumption for homoscedastic variance is not satisfied. The samples are also not independent because states with similar values that live close to each other my 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.

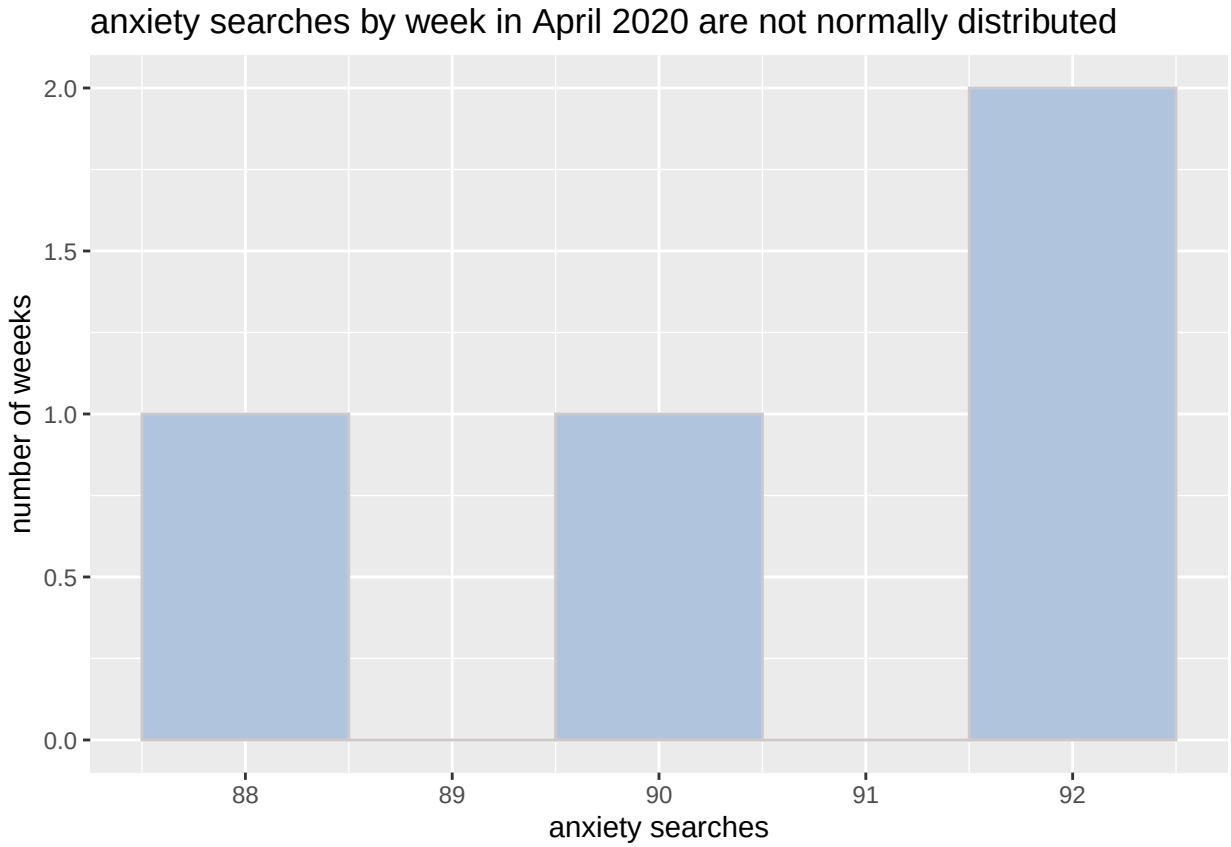
Comparing anxiety trends in April 2020 to April 2018

```
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")
```

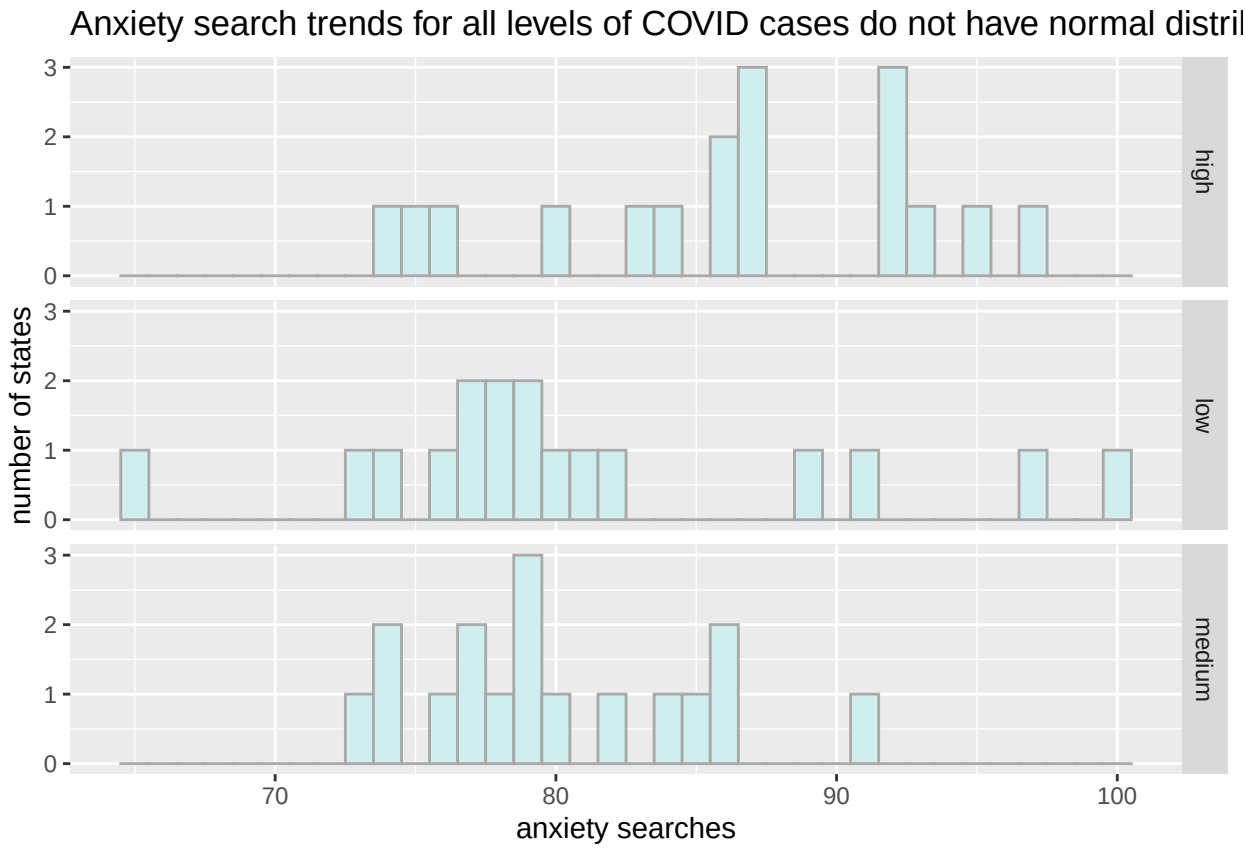


n<30 and not normal distribution: assumption not satisfied

```
##
##  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

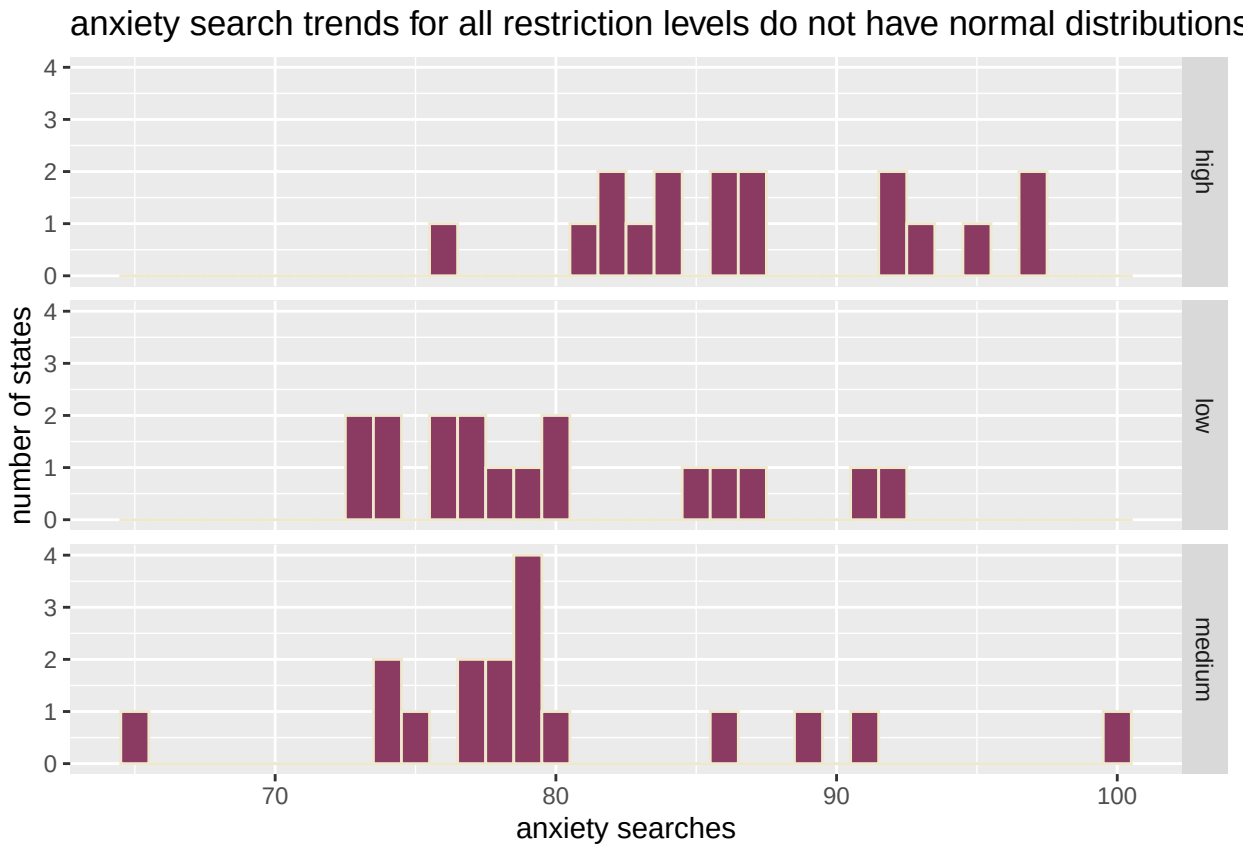


Looking at the graphs, outcomes within groups are not normally distributed for any level of COVID cases, so this assumption is not satisfied. It also looks like the within-group variance among all groups is not the same, so the assumption for homoscedastic variance is not satisfied. The samples are also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



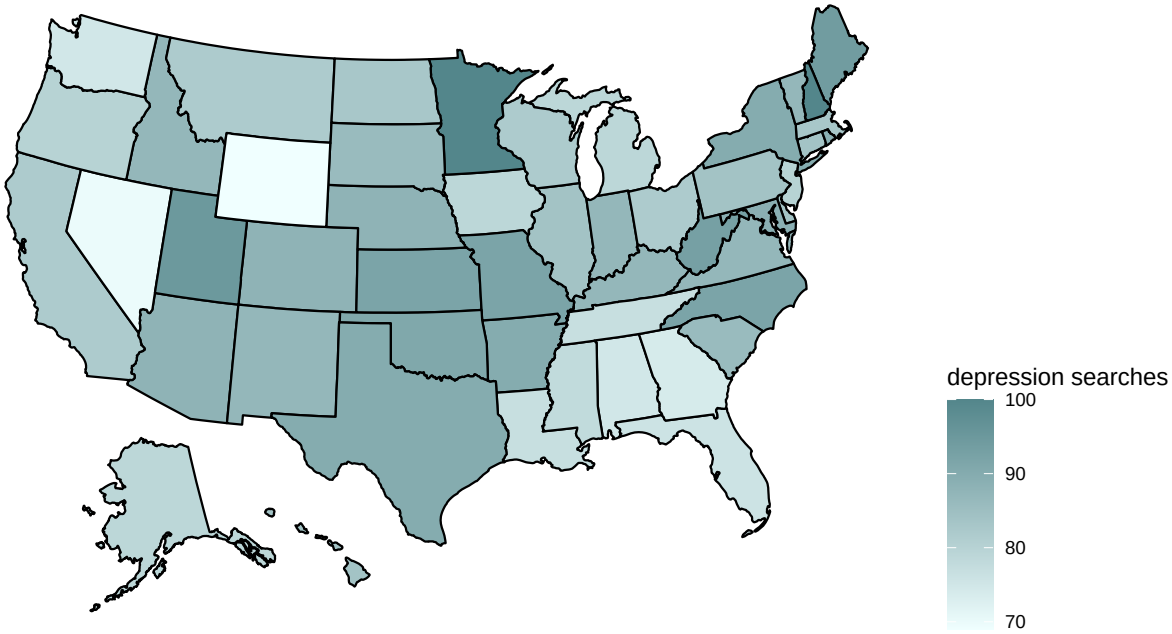
Looking at the graphs, outcomes within groups are not normally distributed for any level of COVID cases, so this assumption is not satisfied. It also looks like the within-group variance among all groups is not the same, so the assumption for homoscedastic variance is not satisfied. The samples are also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

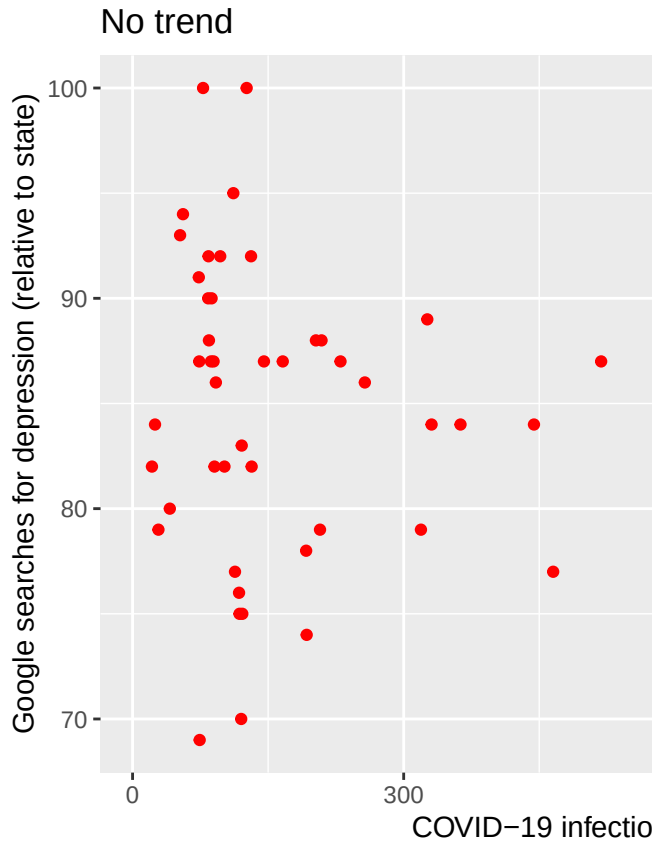
Depression rate in each state map

```
## 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.
```



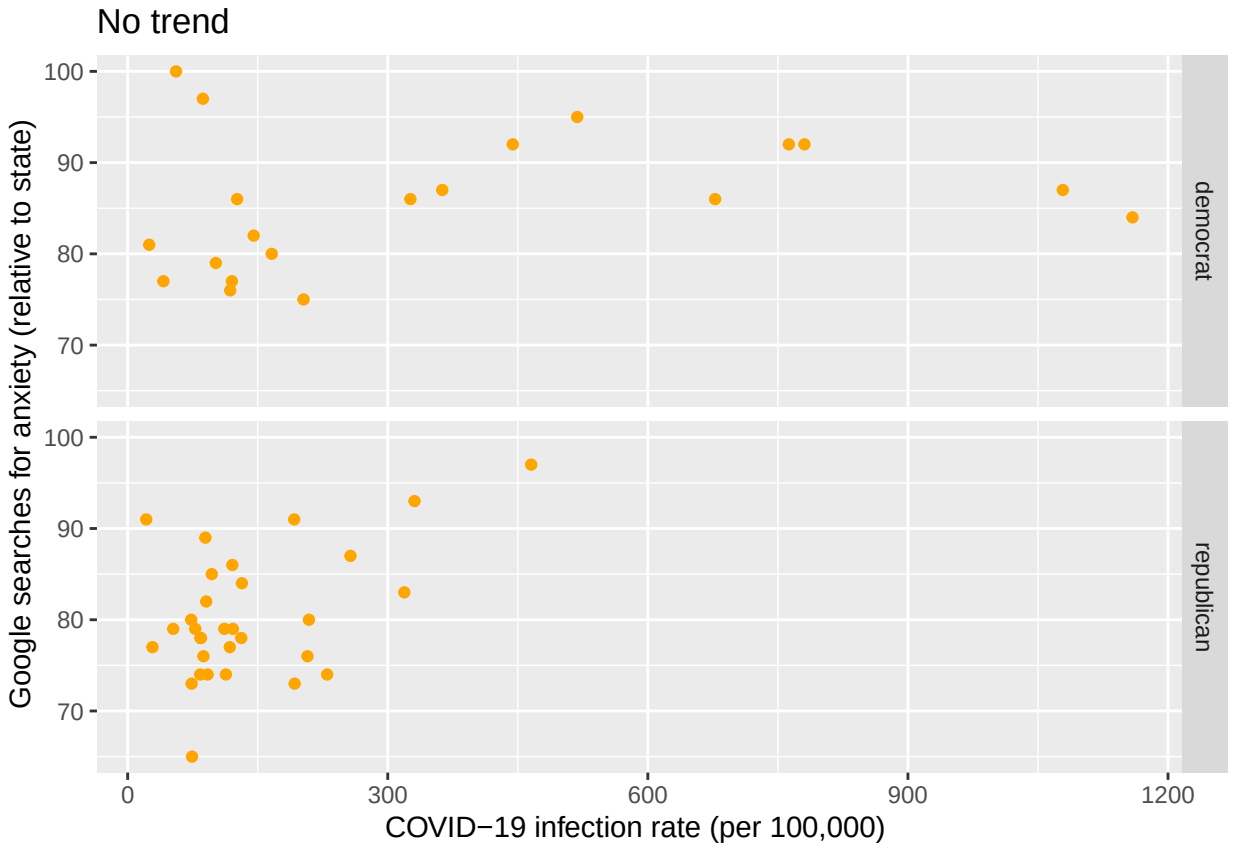
Nour Visuals

```
ggplot(data = covidrate, mapping = aes(x = New.COVID.cases.per.100.000.in.April, y = depression)) +
  geom_point(color= "red") +
  labs(title = "No trend",
       x = "COVID-19 infection rate (per 100,000) ",
       y = "Google searches for depression (relative to state)" )
```

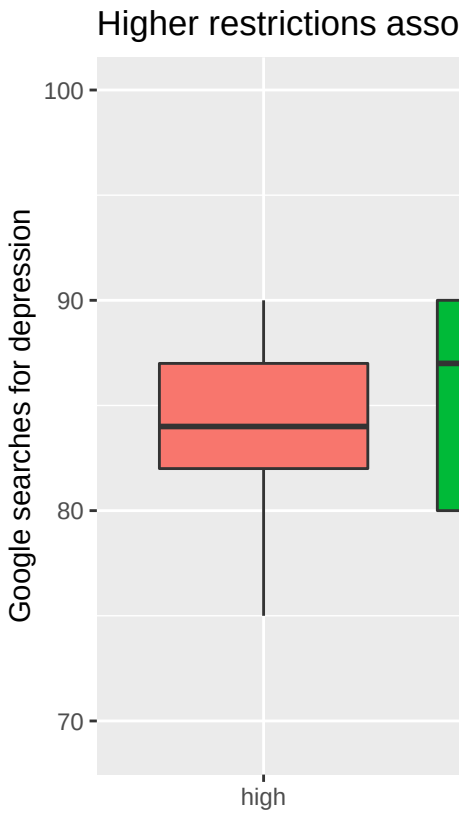


Effect of COVID on depression/anxiety rates in each State

```
ggplot(data = covidrate, mapping = aes(x = New.COVID.cases.per.100.000.in.April, y = anxiety)) +
  geom_point(color= "orange") + facet_grid(party~.) +
  labs(title = "No trend",
       x = "COVID-19 infection rate (per 100,000) ",
       y = "Google searches for anxiety (relative to state)" )
```

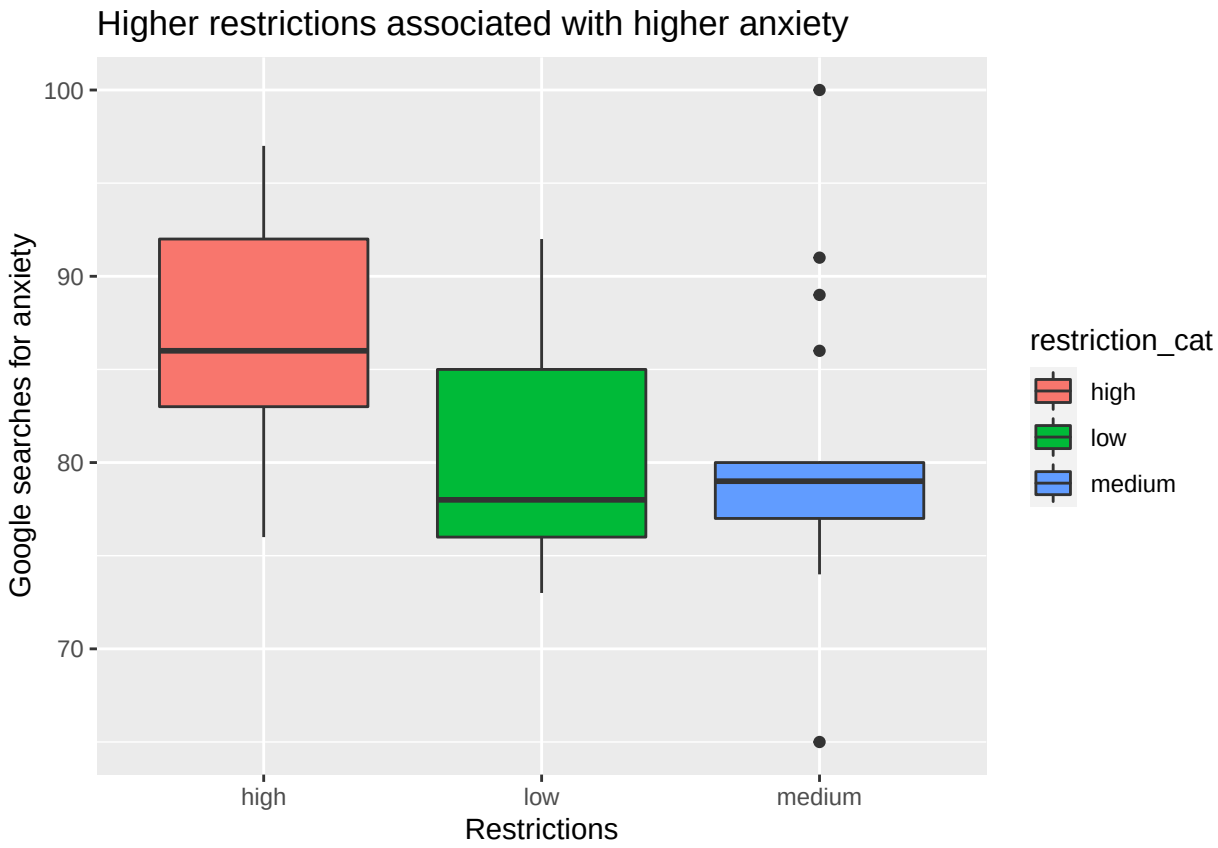



```
ggplot(data = covidrate, mapping = aes(x = restriction_cat, y = depression, fill= restriction_cat)) +
  geom_boxplot() +
  labs(title = "Higher restrictions associated with lower depression",
       x = "Restrictions",
       y = "Google searches for depression")
```



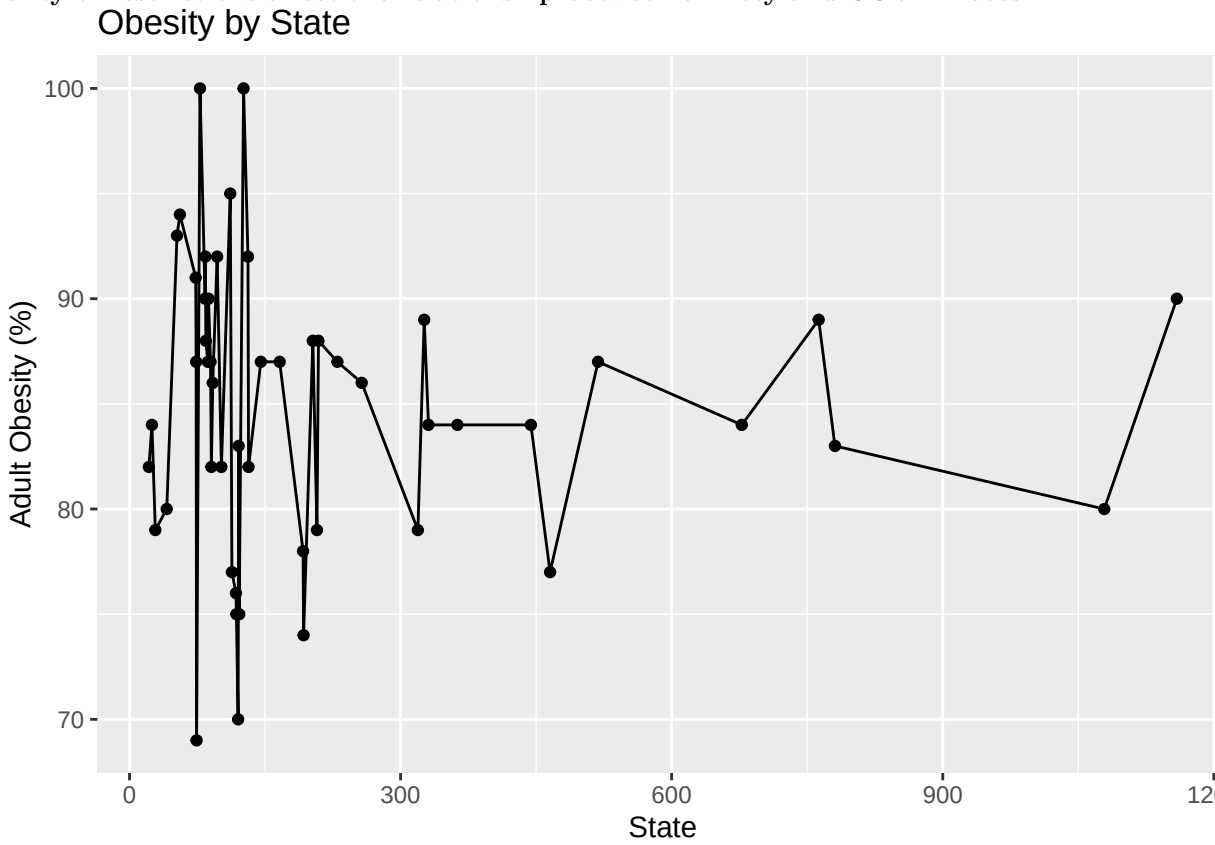
Effect of severity of restrictions on depression/anxiety rates in each State

```
ggplot(data = covidrate, mapping = aes(x = restriction_cat, y = anxiety, fill= restriction_cat)) +
  geom_boxplot() +
  labs(title = "Higher restrictions associated with higher anxiety",
       x = "Restrictions",
       y = "Google searches for anxiety")
```




```
ggplot(data = covidrate, mapping = aes(x = New.COVID.cases.per.100.000.in.April, y = depression)) +
  geom_line() +
  geom_point() +
  labs(title = "Obesity by State",
       x = "State",
       y = "Adult Obesity (%)")
```

How does the severity of restrictions affect the relationship between anxiety and COVID rates



in the 50 states