---

title: “Final Report” subtitle: “due November 16, 2021 by 11:59 PM” author: “Harris Upchurch, Biniam Garomsa, Philemon Hailemariam” date: “11/16/2021” output: pdf_document

Outline:

#Introduction

Starting in 2020, public health measures against the spread of respiratory disease were taken in an effort to slow or prevent the spread of COVID-19. These measures were much more significant than in previous years, as the primary strategies against the spread of respiratory disease in the United States are typically vaccination and encouraging frequent hand-washing, but measures against COVID-19 included shutting down large parts of society and moving online, as well as enforcement or encouragement of face coverings and vaccinations when available. This report studies how these public health measures may have affected the spread of other infectious disease by focusing on influenza hospitalizations and deaths in the United States, as the flu is one of the most common and deadly diseases each year in the U.S., and shares many characteristics with COVID-19.

#Flu seasons during the COVID-19 Pandemic vs. past flu seasons

Have flu deaths and hospitalizations changed during the COVID-19 pandemic compared to past years? If so, is this difference statistically significant?

```
#data/libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v readr   2.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
ilinet <-read.csv("~/R/TBD2/data/ILINet.csv", header=TRUE)
flu<-readr::read_csv(file = '~/R/TBD2/data/State_Custom_Data.csv')
```

```
## Rows: 21996 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): AREA, SUB AREA, AGE GROUP, SEASON, PERCENT COMPLETE
## dbl (4): WEEK, PERCENT P&I, PERCENT PIC, NUM INFLUENZA DEATHS

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
oxford_lockdown<-readr::read_csv(file = '~/R/TBD2/data/OxCGRT_US_latest.csv')
```

```
## Rows: 33800 Columns: 72

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (26): CountryName, CountryCode, RegionName, RegionCode, Jurisdiction, C1...
## dbl (45): Date, C1_School closing, C1_Flag, C2_Workplace closing, C2_Flag, C...
## lgl  (1): M1_Wildcard

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
flu_sec<-readr::read_csv(file = '~/R/TBD2/data/State_Custom_Data.csv')
```

```
## Rows: 21996 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): AREA, SUB AREA, AGE GROUP, SEASON, PERCENT COMPLETE
## dbl (4): WEEK, PERCENT P&I, PERCENT PIC, NUM INFLUENZA DEATHS

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ilinet_sec <- readr::read_csv("~/R/TBD2/data/ILINet.csv")
```

```
## Rows: 22952 Columns: 15

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): REGION TYPE, REGION, % WEIGHTED ILI, AGE 0-4, AGE 25-49, AGE 25-64,...
## dbl (6): YEAR, WEEK, %UNWEIGHTED ILI, ILITOTAL, NUM. OF PROVIDERS, TOTAL PAT...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

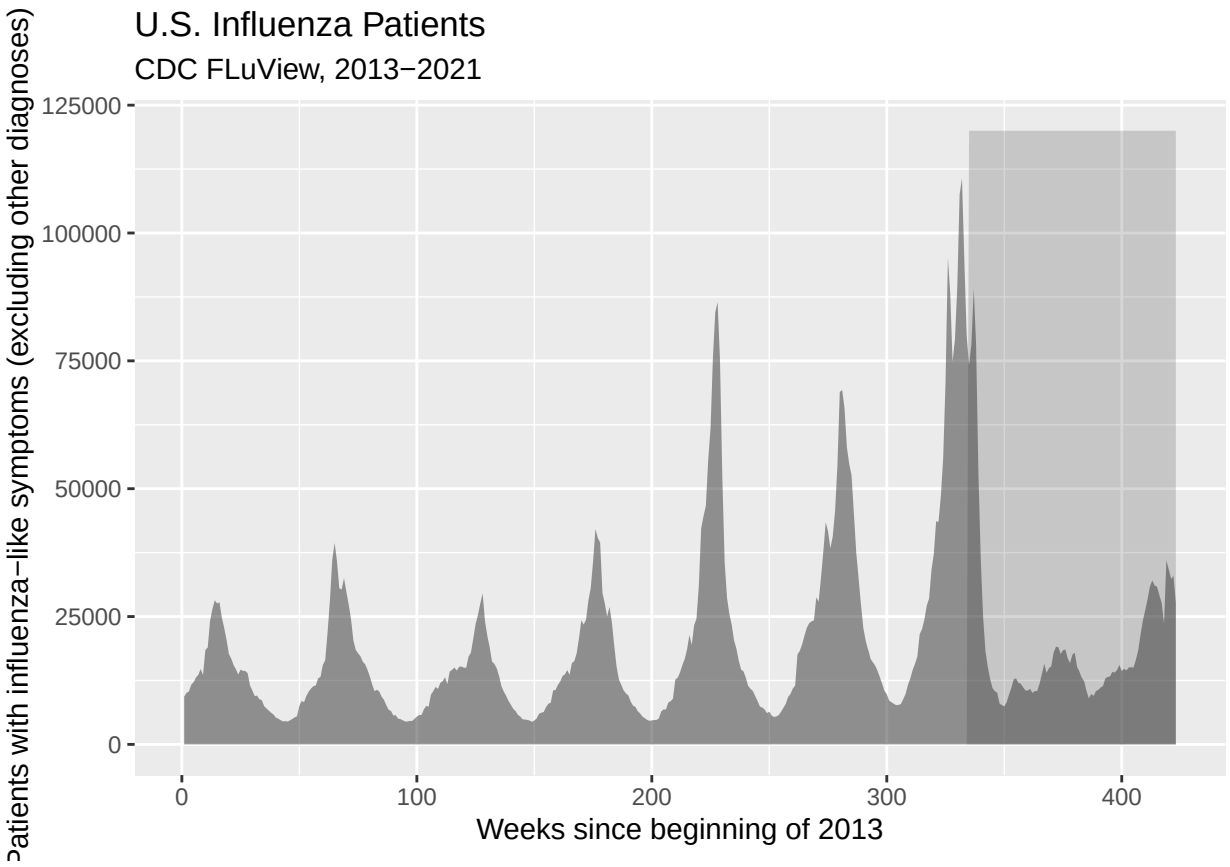
```
#data merging
flu2<-flu%>%
  mutate(Year = ifelse(WEEK>=40 & WEEK<=53, substr(SEASON,3,4),substr(SEASON,6,7)))%>%
  mutate(Year = paste("20",Year,sep=""))
flu2$WEEK<- as.numeric(flu2$WEEK)
flu2$Year<- as.numeric(flu2$Year)
ilinet$YEAR<-as.numeric(ilinet$YEAR)
ilinet$WEEK<- as.numeric(ilinet$WEEK)
flu_ili <- left_join(flu2,ilinet,by=c("SUB AREA" ="REGION","WEEK","Year"="YEAR"))
flu_ili<-flu_ili%>% rename("YEAR" = "Year","STATE" ="SUB AREA")
aggflu_ili <- flu_ili %>%
```

```
  mutate(year = as.character(YEAR), week = as.character(WEEK), total_flu = as.numeric(ILITOTAL)) %>%
  mutate(yearweek = ifelse(WEEK <= 9,paste(year, week, sep = '0'),paste(year, week, sep = ''))) %>%
  drop_na()%>%
  group_by(yearweek)%>%
  summarize(total_flu_patients = sum(total_flu),total_flu_deaths = sum(`NUM INFLUENZA DEATHS`))%>%
  transform(yearweek = as.numeric(yearweek))%>%
  mutate(covidstatus = ifelse(yearweek>=202009,1,0))%>%
  transform(yearweek = rank(yearweek, ties.method = 'min'))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

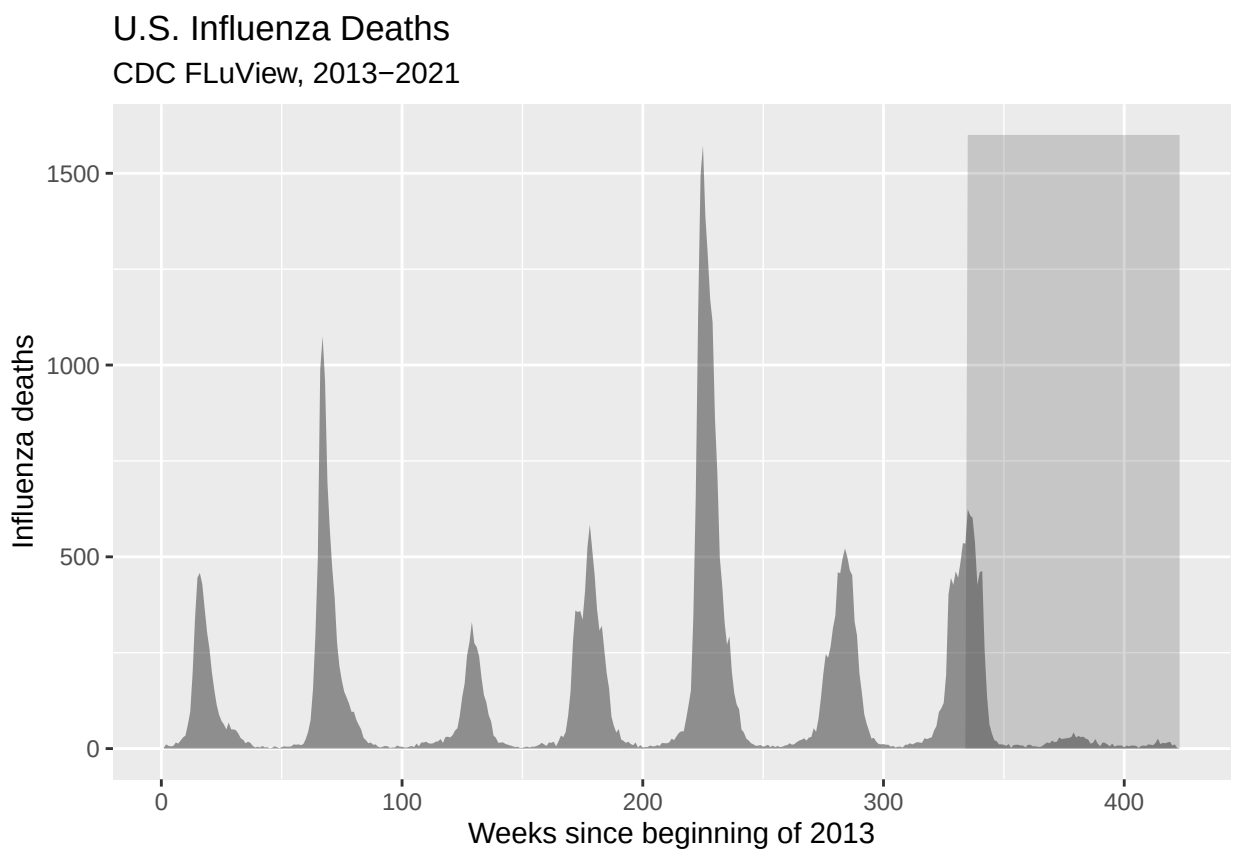
```
## Warning: One or more parsing issues, see `problems()` for details
```

```
#line graph
ggplot(aggflu_ili, aes(x = yearweek)) +
  geom_area(aes(y = total_flu_patients), alpha = 0.5) +
  geom_area(aes(y = covidstatus*120000), alpha = 0.2) +
  labs(title = 'U.S. Influenza Patients',
       subtitle = 'CDC FLuView, 2013-2021',
       x = 'Weeks since beginning of 2013',
       y = 'Patients with influenza-like symptoms (excluding other diagnoses)')
```




```
ggplot(aggflu_ili, aes(x = yearweek)) +
  geom_area(aes(y = total_flu_deaths), alpha = 0.5) +
  geom_area(aes(y = covidstatus*1600), alpha = 0.2) +
  labs(title = 'U.S. Influenza Deaths',
       subtitle = 'CDC FLuView, 2013-2021',
```

```
       x = 'Weeks since beginning of 2013',
       y = 'Influenza deaths')
```




```
#table
flu_table<-aggflu_ili %>%
  mutate(season = ifelse(yearweek >=1 & yearweek <= 52, '2013-2014',
                  ifelse(yearweek >=53 & yearweek <= 105, '2014-2015',
                  ifelse(yearweek >=106 & yearweek <= 157, '2015-2016',
                  ifelse(yearweek >=158 & yearweek <= 209, '2016-2017',
                  ifelse(yearweek >=210 & yearweek <= 261, '2017-2018',
                  ifelse(yearweek >=262 & yearweek <= 313, '2018-2019',
                  ifelse(yearweek >=314 & yearweek <= 365, '2019-2020',
                  ifelse(yearweek >=366 & yearweek <= 418, '2020-2021','2021-2022'
  ))))))))))
flu_table %>%
  group_by(season) %>%
  summarize(total_flu_deaths = sum(total_flu_deaths),
            total_flu_patients = sum(total_flu_patients))%>%
  print()
```

```
## # A tibble: 9 x 3
##   season    total_flu_deaths total_flu_patients
##   <chr>                <dbl>              <dbl>
## 1 2013-2014             4255             649058
## 2 2014-2015             7921             751851
## 3 2015-2016             3237             644170
```

```
## 4 2016-2017             6672             802314
## 5 2017-2018            14968            1215966
## 6 2018-2019             6835            1384249
## 7 2019-2020             8914            1976525
## 8 2020-2021              839             917038
## 9 2021-2022               39             163186
```

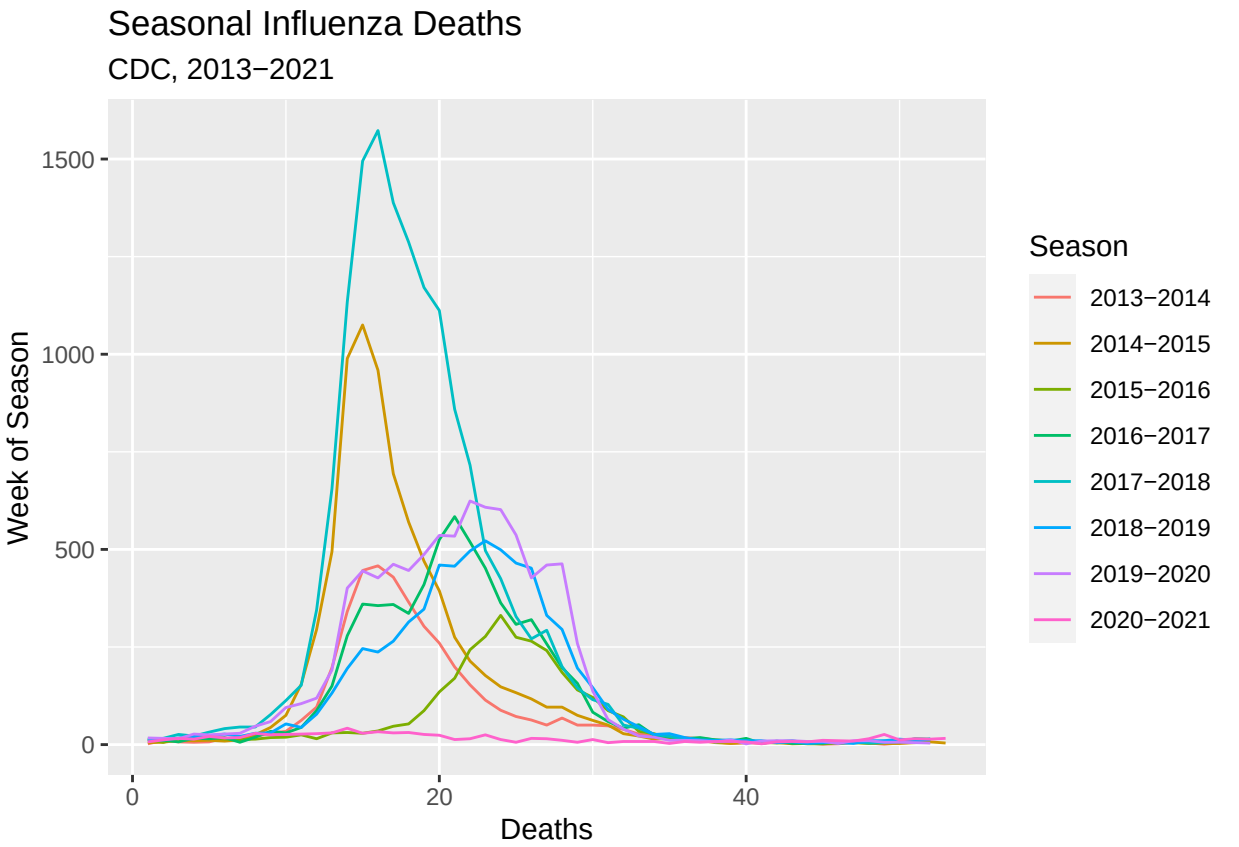
Based on the data, the hospitalizations for the 2020-2021 flu season were similar to several seasons in past years, but noticeably lower than the most recent 2017-2019 seasons, with a total of 917,038 hospitalizations, which was somewhat higher than the 2013-2017 seasons, but about half of the most recent season. The 2021-2022 season is not complete, but appears to be on track to be similar to years before the 2017-2019 period, like the 2020-2021 season. The deaths of both the 2020-2021 season and the 2021-2022 season were and have been noticeably less than any of the previous years used for this comparison. The 2020-2021 season had 25.9% of the deaths of the next least deadly year and 11.1% of the average deaths (7,543.14) of previous seasons, and the 2021-2022 season is on track to follow a similar pattern. These findings indicate some noticeable differences between influenza effects before and during COVID-19, but statistical tests should be done to prove and quantify their difference. Due to the variance between weeks within season, a test should be used which tests for significance either across seasons in total or mean or within analogous weeks across seasons. However, because the peak and distribution within each season is not consistent, comparing along weekly lines causes some issues. This significantly reduces the sample size that may be used within a test from approximately 400 weeks to about 9 seasons. One potential way to adjust for the seasonal patterns is to select the most impactful weeks from each season, so that there are more data points for each category than only using seasons, but the range of values does not include many weeks where the flu is not present. Additionally, the 2021-2022 season must be adjusted to be able to be fairly compared and used for statistical significance. The mean will not allow comparison on its own, due to the peak pattern of each season, so the best way to compare 2021-2022 to other seasons is most likely to cut off each season to the first few weeks, which would further reduce sample size. The statistical test that is used will test for how the binary variable for whether or not the COVID-19 pandemic is occurring explains the observations for the numerical variables of influenza hospitalizations and deaths.

```
library(broom)
library(modelr)
```

```
##
## Attaching package: 'modelr'
```

```
## The following object is masked from 'package:broom':
##
##     bootstrap
```

```
library(purrr)
covid_flu_table <- flu_table %>%
  mutate(covidstatus = ifelse(season == '2020-2021' | season == '2021-2022', 1, 0))%>%
  filter(season !='2021-2022')%>%
  group_by(season)%>%
  mutate(seasonweek = (rank(yearweek)))%>%
  ungroup()%>%
  group_by(seasonweek)
covid_flu_table %>%
  ggplot( aes(x = seasonweek, color = season)) +
  geom_line(aes(y = total_flu_deaths, group = season)) +
  labs(title = 'Seasonal Influenza Deaths',
       subtitle = 'CDC, 2013-2021',
       x = 'Deaths',
       y = 'Week of Season',
       color = 'Season')
```



In the line graph above, some of the issues posed by the data in finding statistical significance can be observed. Peaks and patterns do not line up consistently across years and the current definition of seasons includes many weeks that are not actually affected by the flu. Restricting the definition of a season by only taking the top weeks for deaths or hospitalization in each season can allow for more effective significance testing.

```
covid_flu_table<-covid_flu_table%>%
  group_by(season)%>%
  slice_max(order_by = total_flu_deaths, n = 10)%>%
  ungroup()
lm(total_flu_deaths ~ covidstatus, data = covid_flu_table)%>%
tidy(conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)     528.      36.2     14.6  3.34e-24     456.      600.
## 2 covidstatus    -498.      94.7     -5.26 1.19e- 6    -686.     -309.
```

```
lm(total_flu_patients ~ covidstatus, data = covid_flu_table)%>%
tidy(conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)   42691.     2619.     16.3  3.77e-27   37480.    47903.
## 2 covidstatus  -25191.     6846.     -3.68 4.22e- 4  -38814.   -11568.
```

Based on the model, there is a statistically significant difference between the number of deaths at the peak 10 weeks of each season when comparing weeks in non-COVID times with weeks in COVID-affected times.

The p value for covidstatus is 1.19x10^-6, which is less than .05 and the confidence interval of -686.46 to -309.41 does not include the null hypothesis of 0, which would indicate no difference in the data. The model indicates that moving from non-COVID to COVID time correlates with a decrease in deaths of 497.94 from the estimate of 527.69. Based on this data, something that has happened during the pandemic has most likely caused a decrease in deaths due to the flu. The model also indicates that there was a statistically significant decrease in hospitalizations when compared to the norm, by 25,190.74, from an estimate of 42,691.49, with a p value of 4.22x10^-4, which is less than 0.05, and a confidence interval of -38,813.74 to -11,567.73, which does not include the null hypothesis of 0. This is less than the difference in deaths, in terms of significance, but it is interesting because the hospitalizations were similar to past hospitalization levels, but the model still determine that its difference was significance due to the upward trend in hospitalization in recent years. The data for deaths was not similar to anything recorded in the past for the time period tested. #Comparison of lockdowns and flu levels by state/comparison of flu levels by lockdown stringency (?) compare states/lockdown and mask levels (stringency index) to flu hospitalizations or deaths consider using a model test which model fits best test for and explain statistical significance talk about why flu levels may be lower more protective measures against respiratory illnesses possibly more issues going to hospital for flu/collecting data is the population most affected by flu being reduced by COVID?

```
# t-test for number of Influenza Death pre-covid and during covid
t.test(covid_flu_table$total_flu_deaths~covid_flu_table$covidstatus,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  covid_flu_table$total_flu_deaths by covid_flu_table$covidstatus
## t = 12.754, df = 69.148, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  420.0507 575.8207
## sample estimates:
## mean in group 0 mean in group 1
##        527.6857         29.7500
```

```
# t-test for number of Influenza Hospitalizations pre-covid and during covid
t.test(covid_flu_table$total_flu_patients~covid_flu_table$covidstatus,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  covid_flu_table$total_flu_patients by covid_flu_table$covidstatus
## t = 8.0363, df = 77.744, p-value = 8.165e-12
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  18949.89 31431.58
## sample estimates:
## mean in group 0 mean in group 1
##        42691.49        17500.75
```

Description about the t-test values after the modeling - Harry Writes

Our hypothesis is that the change in number of deaths and hospitalizations due to influenza is highly related to the restrictions enforced as a result of the COVID-19 pandemic. In order to examine this hypothesis, a variable that quantitatively describes the restriction policies imposed during the pandemic was used.This variable is called `Stringency Index` and is calculated as a mean of sub-indices, which describe restriction policies such as school closures, work space closing, cancellation of public events, public and international travel controls and mask and vaccination policy. ’’’ The detailed calculation of this variable can be found in this reference.https://www.bsg.ox.ac.uk/sites/default/files/Calculation%20and%20presentation%20of%20the%20Stringency%20Index.p
’’’

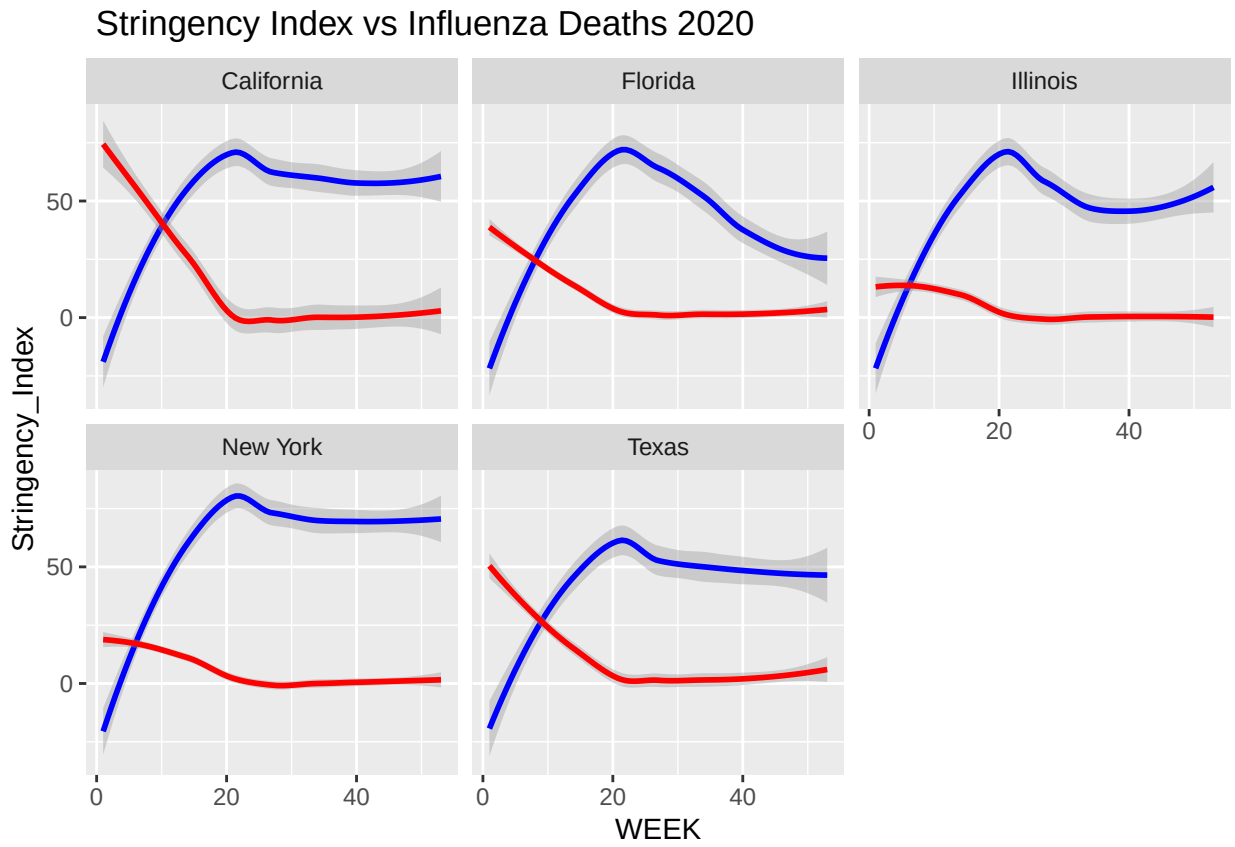
```
## `summarise()` has grouped output by 'CountryName', 'CountryCode', 'RegionName', 'RegionCode', 'Year'
## Adding missing grouping variables: `CountryName`, `CountryCode`, `STATE`, `RegionCode`
```

```
merged_5_states%>%
ggplot()+
geom_smooth(aes(x = WEEK, y = Stringency_Index, color = as.factor(YEAR)))+
  facet_wrap(~STATE)+
  labs(title = 'Stringency Index',
  y = 'Stringency Index',
    color ='Year')
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


```
merged_5_states %>%
  filter(YEAR == 2020) %>%
ggplot()+
geom_smooth(aes(x = WEEK, y = Stringency_Index), color = 'blue')+
  geom_smooth(aes(x = WEEK, y = `NUM INFLUENZA DEATHS`), color ='red')+
  facet_wrap(~STATE)+
  labs(title =' Stringency Index vs Influenza Deaths 2020',
  )
```

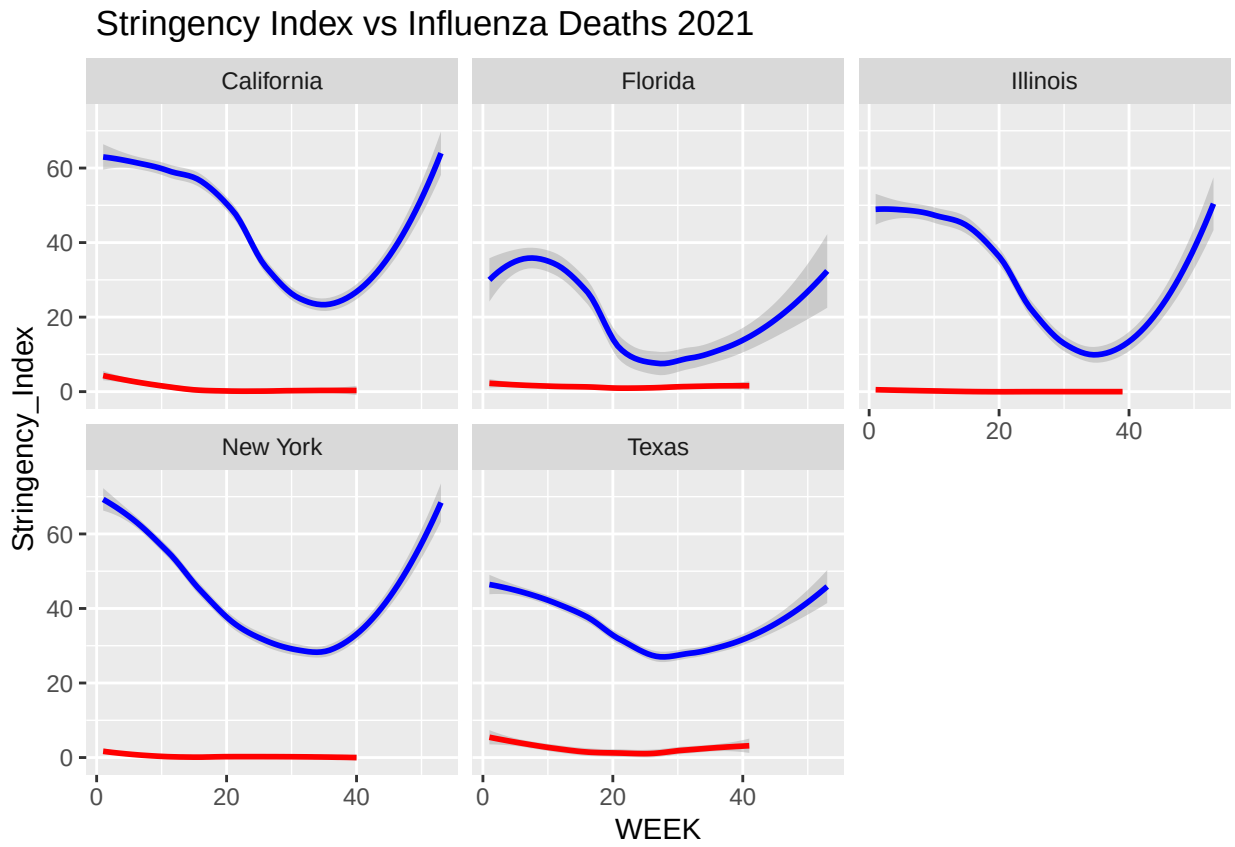
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
merged_5_states %>%
  filter(YEAR == 2021) %>%
ggplot()+
geom_smooth(aes(x = WEEK, y = Stringency_Index), color = 'blue')+
  geom_smooth(aes(x = WEEK, y = `NUM INFLUENZA DEATHS`), color ='red')+
  facet_wrap(~STATE)+
  labs(title =' Stringency Index vs Influenza Deaths 2021',
  )
```

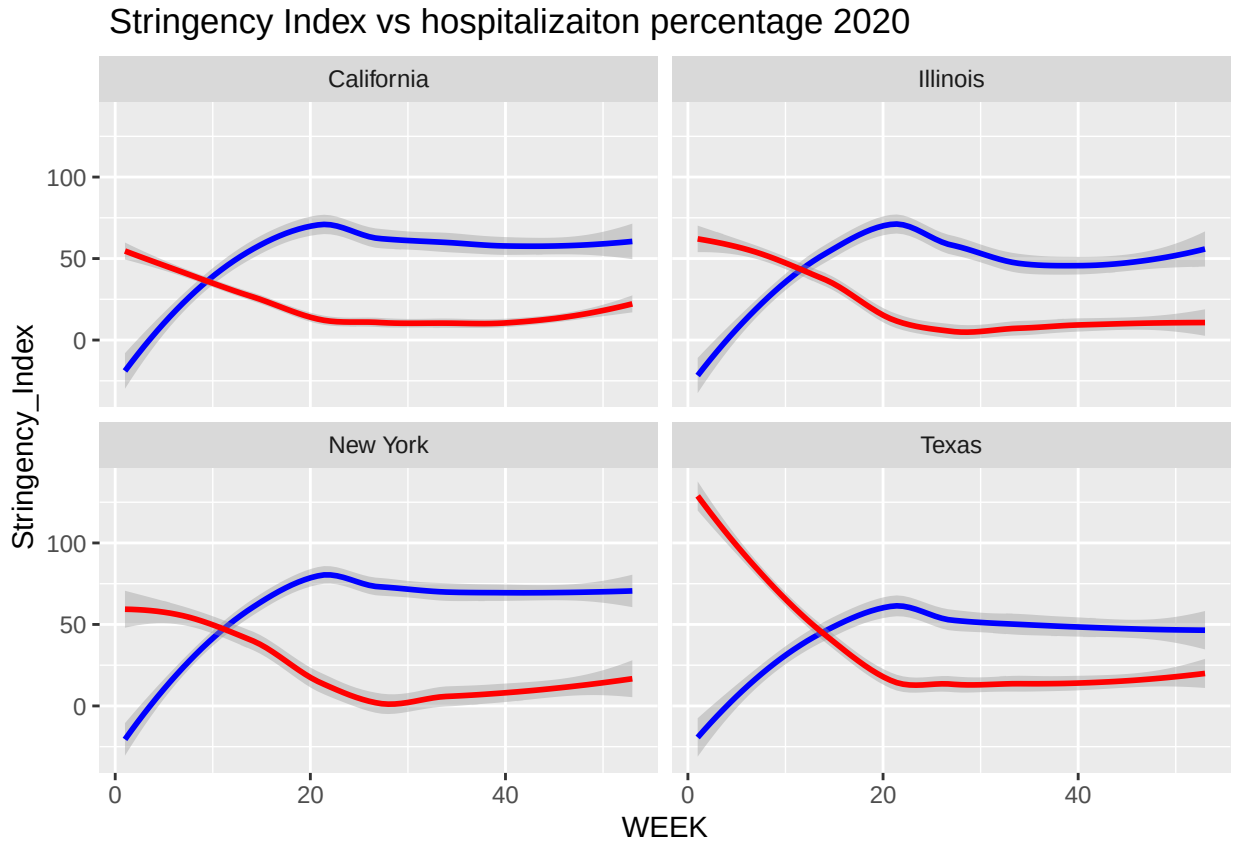
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
```



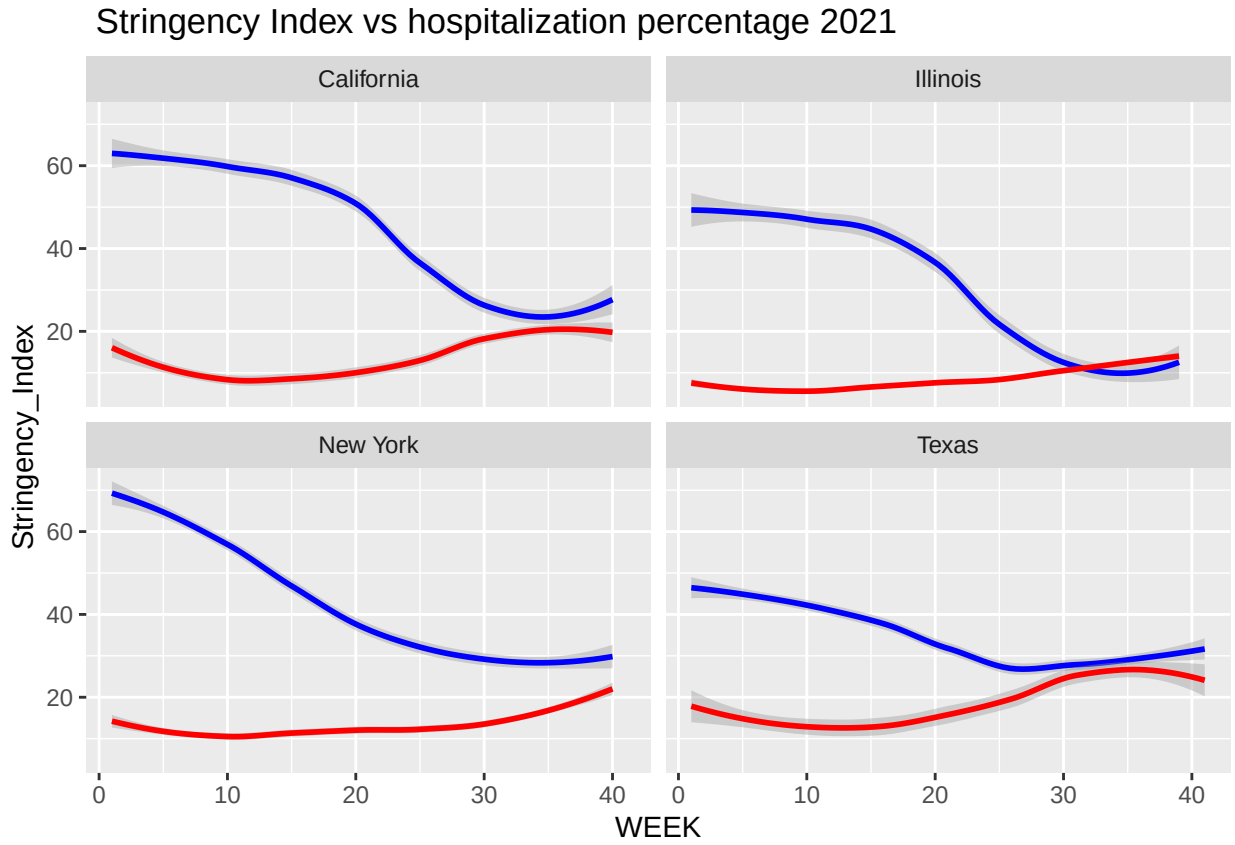

```
merged_5_states %>%
  filter(YEAR == 2020 & !is.na(`%UNWEIGHTED ILI`)) %>%
ggplot()+
geom_smooth(aes(x = WEEK, y = Stringency_Index), color = 'blue')+
  geom_smooth(aes(x = WEEK, y = `%UNWEIGHTED ILI`*10), color ='red')+
  facet_wrap(~STATE)+
  labs(title =' Stringency Index vs hospitalizaiton percentage 2020',
  )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



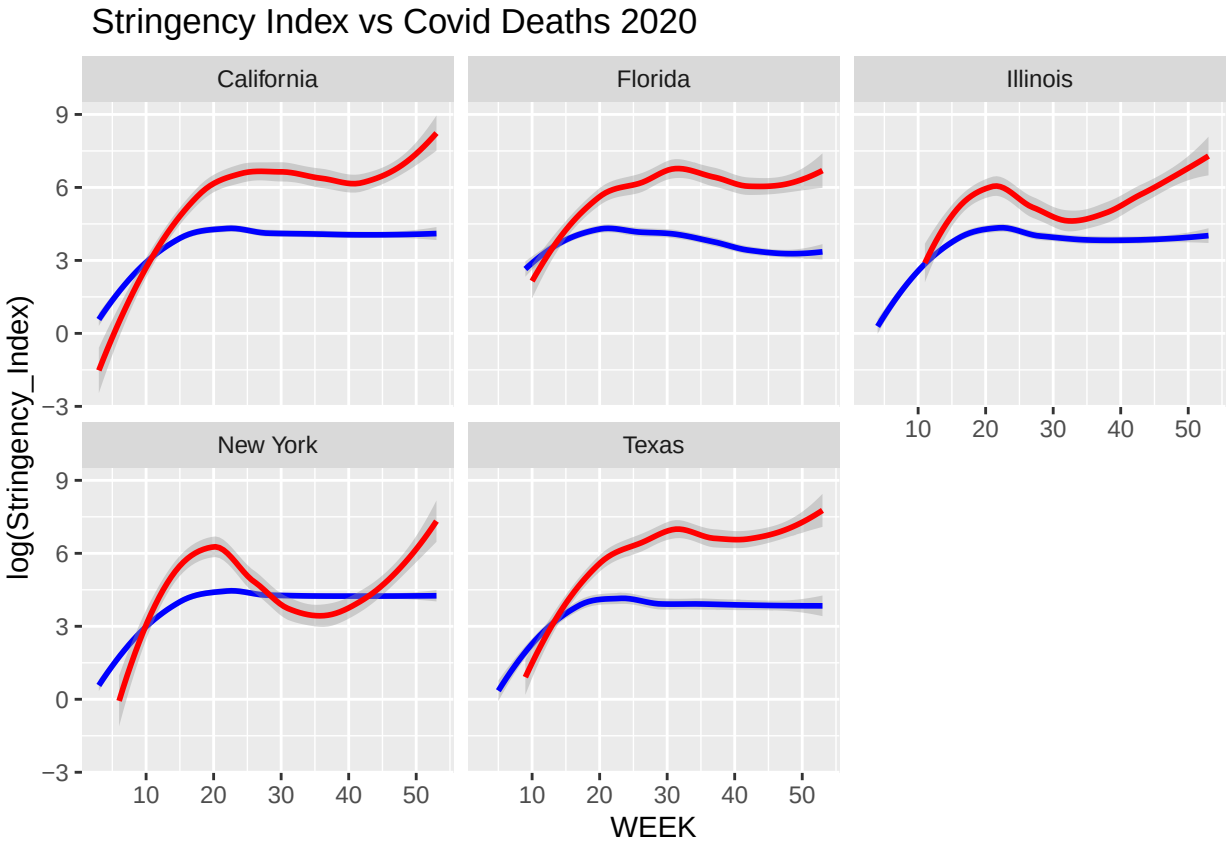

```
merged_5_states %>%
  filter(YEAR == 2021 & !is.na(`%UNWEIGHTED ILI`) & STATE !='Florida') %>%
ggplot()+
geom_smooth(aes(x = WEEK, y = Stringency_Index), color = 'blue')+
  geom_smooth(aes(x = WEEK, y = `%UNWEIGHTED ILI`*10), color ='red')+
  facet_wrap(~STATE)+
  labs(title =' Stringency Index vs hospitalization percentage 2021')
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
merged_5_states %>%
  filter(YEAR == 2020) %>%
ggplot()+
geom_smooth(aes(x = WEEK, y = log(Stringency_Index)), color = 'blue')+
  geom_smooth(aes(x = WEEK, y = log(`NUM COVID-19 DEATHS`)), color ='red')+
  facet_wrap(~STATE)+
  labs(title =' Stringency Index vs Covid Deaths 2020')
```

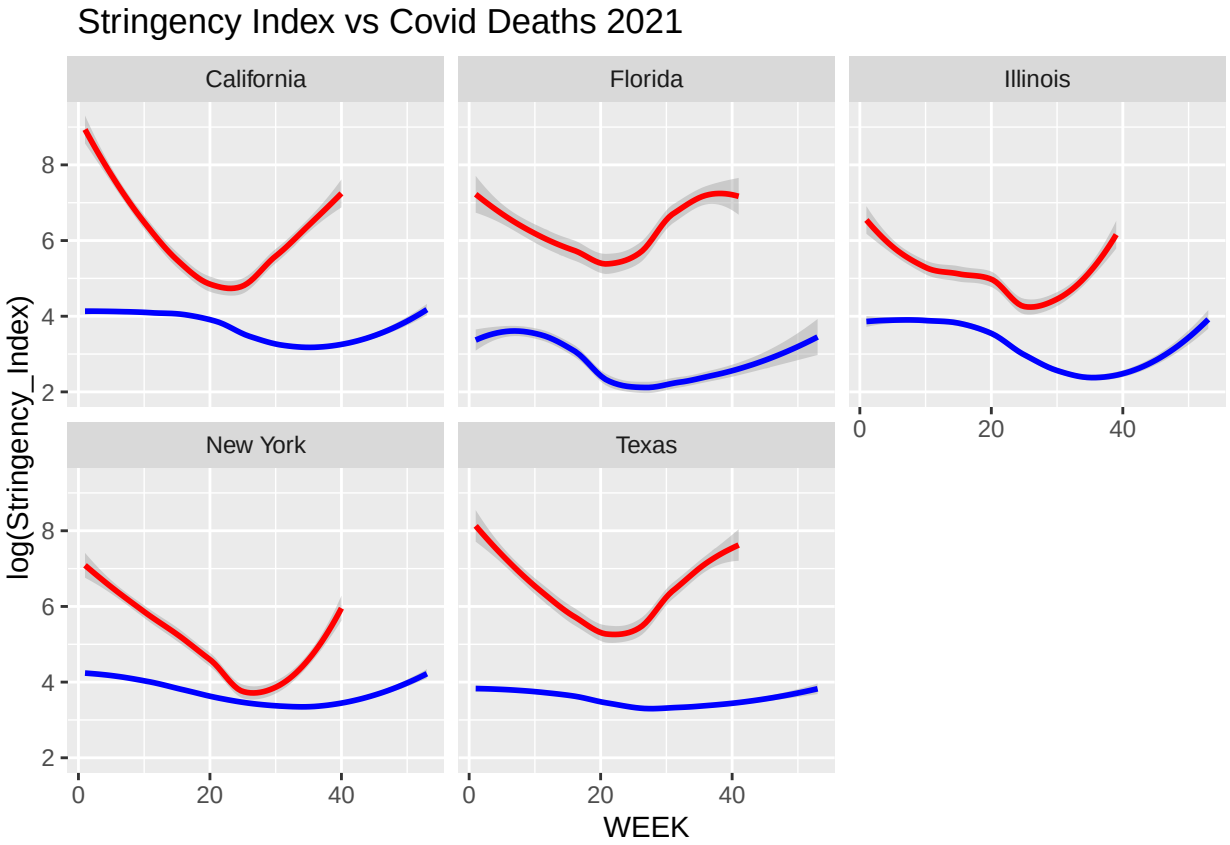
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## Warning: Removed 19 rows containing non-finite values (stat_smooth).
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## Warning: Removed 42 rows containing non-finite values (stat_smooth).
```




```
merged_5_states %>%
  filter(YEAR == 2021) %>%
ggplot()+
geom_smooth(aes(x = WEEK, y = log(Stringency_Index)), color = 'blue')+
  geom_smooth(aes(x = WEEK, y = log(`NUM COVID-19 DEATHS`)), color ='red')+
  facet_wrap(~STATE)+
  labs(title =' Stringency Index vs Covid Deaths 2021')
```

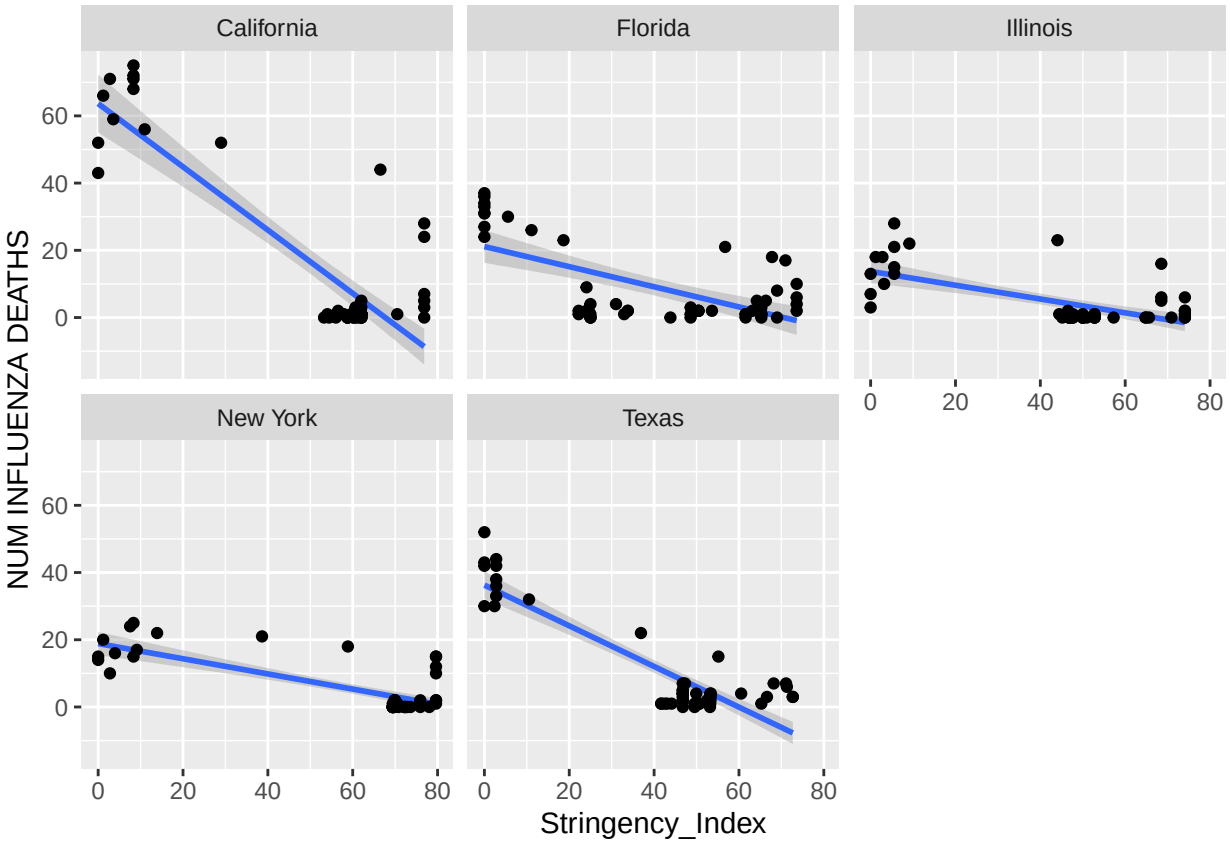
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
```




```
merged_5_states%>%
  filter(YEAR == 2020)%>%
  ggplot(aes(`Stringency_Index`, `NUM INFLUENZA DEATHS`))+
  geom_smooth(method = "lm")+
  geom_point()+
  facet_wrap(~STATE)
```

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

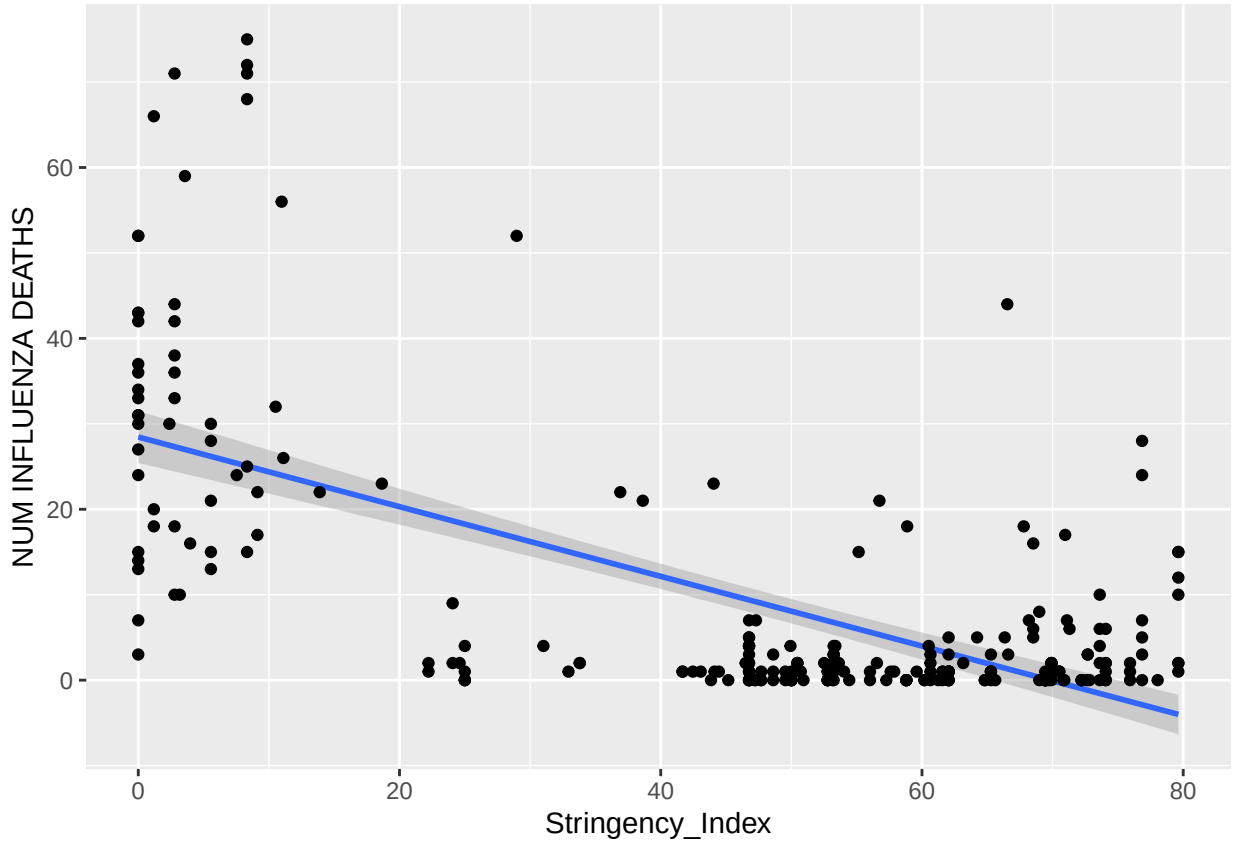



```
lm(`NUM INFLUENZA DEATHS`~`Stringency_Index`, data = merged_5_states)%>%
  tidy(conf.int=TRUE)
```

```
## # A tibble: 2 x 7
##   term             estimate std.error statistic  p.value conf.low conf.high
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)        15.4      1.10        14.0 2.50e-37   13.2      17.5
## 2 Stringency_Index   -0.232    0.0231     -10.1 1.24e-21   -0.278    -0.187
```

```
merged_5_states%>%
  filter(YEAR == 2020)%>%
  ggplot(aes(`Stringency_Index`, `NUM INFLUENZA DEATHS`))+
  geom_smooth(method = "lm")+
  geom_point()
```

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




```
fl<-flu_ili%>%
  mutate(afterCovid = ifelse((YEAR==2021 | YEAR ==2020), 1,0))
t.test(fl$`NUM INFLUENZA DEATHS`~fl$afterCovid,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  fl$`NUM INFLUENZA DEATHS` by fl$afterCovid
## t = 6.5347, df = 8207.7, p-value = 6.753e-11
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.5978802 1.1102887
## sample estimates:
## mean in group 0 mean in group 1
##        2.756135        1.902051
```

```
t.test(fl$`NUM PNEUMONIA DEATHS`~fl$afterCovid,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  fl$`NUM PNEUMONIA DEATHS` by fl$afterCovid
## t = -21.581, df = 5322.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -75.26668 -62.73112
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        66.22487       135.22377
```

```
t.test(fl$`%UNWEIGHTED ILI`~fl$afterCovid,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  fl$`%UNWEIGHTED ILI` by fl$afterCovid
## t = 2.1169, df = 8248, p-value = 0.0343
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.004725211 0.122987174
## sample estimates:
## mean in group 0 mean in group 1
##        1.807872        1.744016
```