# Relationship between the occurrence of female depression and intensity of COVID-19 pandemic.

Filipp Trubin. SRI Interview

1/11/2022

**The observation parallels the most recent article by SRI on medium.com:**
*Link: SRI research points to a tripling of depression risk in emerging adults during the pandemic*

This observation can be useful both for people experiencing depression and for researchers as another point of view.

**Steps:**
1. Datasets loading and preparation.
2. EDA: build time series charts.
3. Select null hypothesis. Define model.
4. Calculate the correlation coefficient, build linear regression model.
5. Summary.
6. Conclusion.
7. References.

### 1. Datasets loading and preparation

**Load dataset** "Indicators of Anxiety or Depression Based on Reported Frequency of Symptoms During Last 7 Days".
*Link:*Dataset/Variables description (source: data.cdc.gov)

```
library(dplyr, warn.conflicts = FALSE)
library(lubridate, warn.conflicts = FALSE)
library(ggplot2)
library(shiny)
options(scipen=999)

data_survey <- read.csv("https://data.cdc.gov/api/views/8pt5-q6wp/rows.csv?accessType=DOWNLOAD")
head(data_survey)
```

```
##                           Indicator             Group         State      Subgroup
## 1 Symptoms of Depressive Disorder National Estimate United States United States
## 2 Symptoms of Depressive Disorder            By Age United States 18 - 29 years
## 3 Symptoms of Depressive Disorder            By Age United States 30 - 39 years
## 4 Symptoms of Depressive Disorder            By Age United States 40 - 49 years
## 5 Symptoms of Depressive Disorder            By Age United States 50 - 59 years
## 6 Symptoms of Depressive Disorder            By Age United States 60 - 69 years
##   Phase Time.Period    Time.Period.Label Time.Period.Start.Date
## 1     1           1 Apr 23 - May 5, 2020             04/23/2020
## 2     1           1 Apr 23 - May 5, 2020             04/23/2020
## 3     1           1 Apr 23 - May 5, 2020             04/23/2020
## 4     1           1 Apr 23 - May 5, 2020             04/23/2020
## 5     1           1 Apr 23 - May 5, 2020             04/23/2020
## 6     1           1 Apr 23 - May 5, 2020             04/23/2020
##   Time.Period.End.Date Value Low.CI High.CI Confidence.Interval Quartile.Range
## 1           05/05/2020  23.5   22.7    24.3         22.7 - 24.3
## 2           05/05/2020  32.7   30.2    35.2         30.2 - 35.2
## 3           05/05/2020  25.7   24.1    27.3         24.1 - 27.3
## 4           05/05/2020  24.8   23.3    26.2         23.3 - 26.2
## 5           05/05/2020  23.2   21.5    25.0         21.5 - 25.0
## 6           05/05/2020  18.4   17.0    19.7         17.0 - 19.7
```

**Filter down focus group: Var.1: "Depression Value" Var.2: "Female".**

```
data_survey_1 <- data_survey[which(data_survey$Indicator == 'Symptoms of Depressive Disorder'
          & data_survey$Group == 'By Sex'
          & data_survey$Subgroup == 'Female'),]
```

**Convert Date field / drop NA.**

```
data_survey_2 <- data_survey_1 %>% dplyr::select(Time.Period.End.Date, Value)
data_survey_2$Time.Period.End.Date <- mdy(data_survey_2$Time.Period.End.Date)
data_survey_3 <- data_survey_2[complete.cases(data_survey_2), ]
```

**Load dataset** "United States COVID-19 Cases and Deaths by State over Time."
*Link:* Dataset/Variables description (source: data.cdc.gov)

```
data_COVID19 <- read.csv("https://data.cdc.gov/api/views/9mfq-cb36/rows.csv?accessType=DOWNLOAD")
head(data_survey)
```

```
##                         Indicator             Group         State      Subgroup
## 1 Symptoms of Depressive Disorder National Estimate United States United States
## 2 Symptoms of Depressive Disorder            By Age United States 18 - 29 years
## 3 Symptoms of Depressive Disorder            By Age United States 30 - 39 years
## 4 Symptoms of Depressive Disorder            By Age United States 40 - 49 years
## 5 Symptoms of Depressive Disorder            By Age United States 50 - 59 years
## 6 Symptoms of Depressive Disorder            By Age United States 60 - 69 years
##   Phase Time.Period    Time.Period.Label Time.Period.Start.Date
## 1     1           1 Apr 23 - May 5, 2020             04/23/2020
## 2     1           1 Apr 23 - May 5, 2020             04/23/2020
## 3     1           1 Apr 23 - May 5, 2020             04/23/2020
## 4     1           1 Apr 23 - May 5, 2020             04/23/2020
## 5     1           1 Apr 23 - May 5, 2020             04/23/2020
## 6     1           1 Apr 23 - May 5, 2020             04/23/2020
##   Time.Period.End.Date Value Low.CI High.CI Confidence.Interval Quartile.Range
## 1           05/05/2020  23.5   22.7    24.3         22.7 - 24.3
## 2           05/05/2020  32.7   30.2    35.2         30.2 - 35.2
## 3           05/05/2020  25.7   24.1    27.3         24.1 - 27.3
## 4           05/05/2020  24.8   23.3    26.2         23.3 - 26.2
## 5           05/05/2020  23.2   21.5    25.0         21.5 - 25.0
## 6           05/05/2020  18.4   17.0    19.7         17.0 - 19.7
```

**Filter down focus group: Var.1: “Date” Var.2: “Number of cases”.**

```
data_COVID19_2 <- data_COVID19 %>% dplyr::select(submission_date, new_case)
data_COVID19_2$submission_date <- mdy(data_COVID19_2$submission_date)
data_COVID19_2 <- data_COVID19_2 %>% filter(new_case > 0 & submission_date <= as.Date('2021-12-13'))
```

**Grouping number of cases by day.**

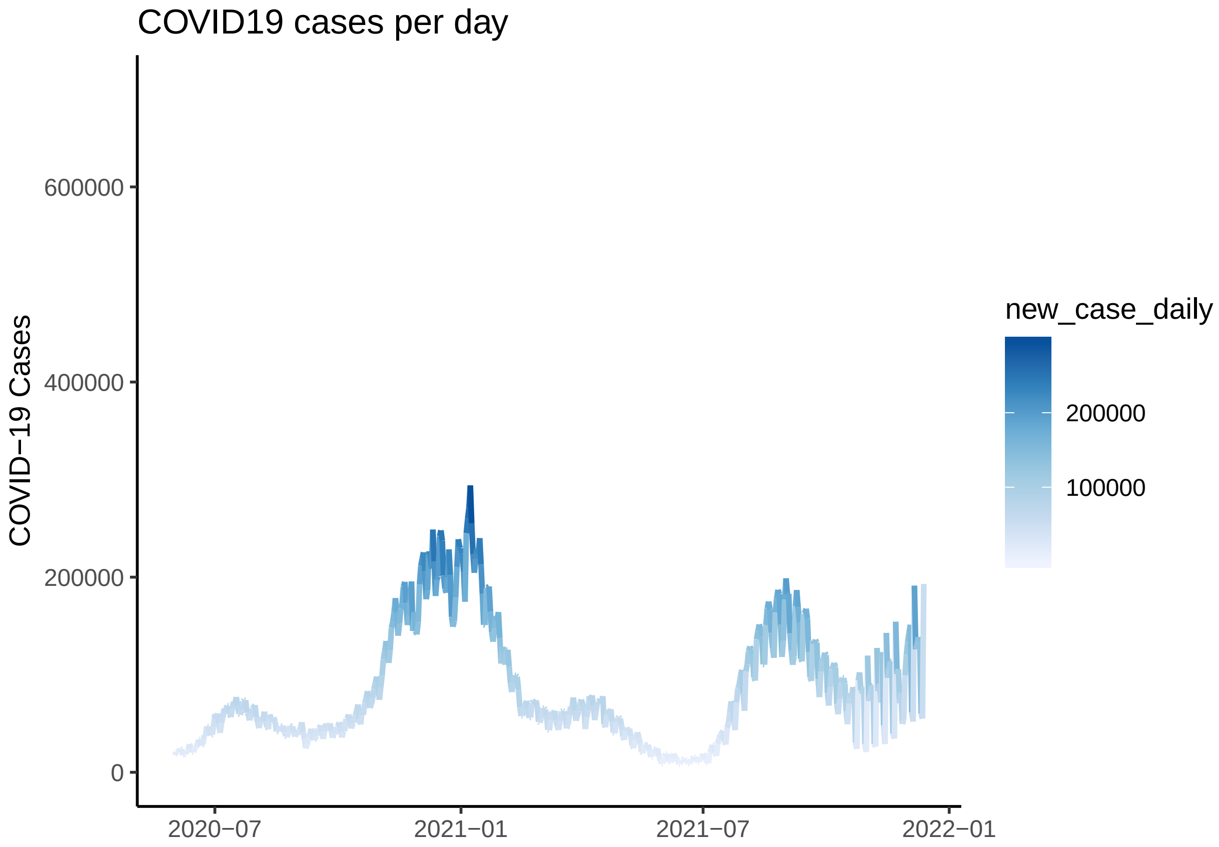
```
data_COVID19_3 <- data_COVID19_2 %>%
  group_by(submission_date) %>%
  summarize(new_case_daily = sum(new_case, na.rm = TRUE))
```

## 2. EDA: build time series charts.

**EDA: build graph to reflect number of cases daily (06/2020 - 12/2021).**

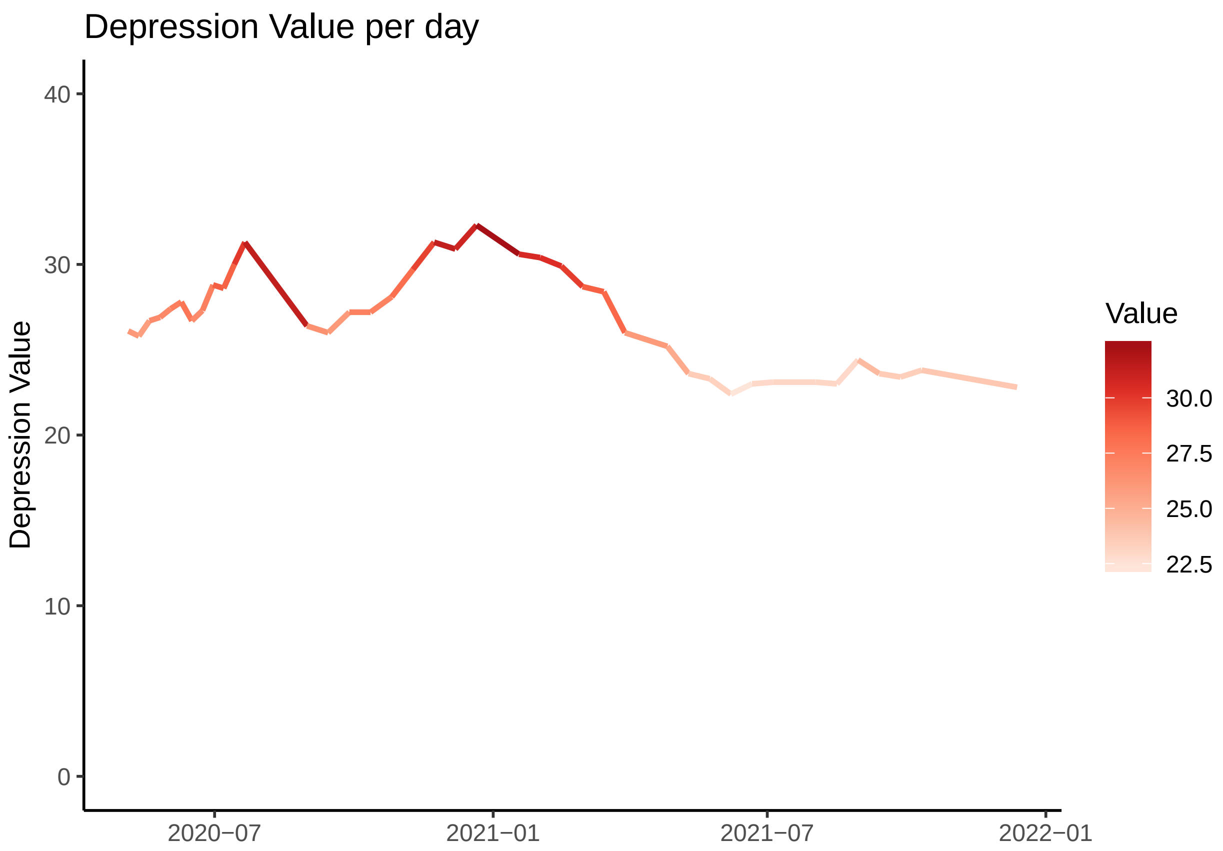
```
ggplot(data_COVID19_3) +
  aes(x = submission_date, y = new_case_daily, colour = new_case_daily) +
  geom_line(size = 1L) +
  scale_color_distiller(palette = "Blues", direction = 1) +
  labs(title = "COVID19 cases per day",x = NULL, y = 'COVID-19 Cases') +
  ylim(0, 700000) +
  scale_x_date(limit=c(as.Date('2020-06-01'), as.Date('2021-12-13'))) +
  theme_classic()
```

```
## Warning: Removed 114 rows containing missing values (geom_path).
```



**EDA: build graph to reflect Depression Value daily (05/2020 - 12-2021).**

```
ggplot(data_survey_3) +
  aes(x = Time.Period.End.Date, y = Value, colour = Value) +
  geom_line(size = 1L) +
  scale_color_distiller(palette = 'Reds', direction = 1) +
  labs(title = 'Depression Value per day',x = NULL,y = 'Depression Value') +
  theme_classic() +
  ylim(0, 40)
```



**3. Select null hypothesis. Define model.**

**Null hypothesis** is female Depression appearance or increasing doesn't have a relationship with intensity of COVID-19 pandemic.
To **accept** or **reject** that hypothesis linear regression model is applied.

**4. Calculate correlation coefficient of Depression Value and COVID-19 cases per day.**

Build linear regression model.
**Merge datasets. Number of cases and Depression Value aggregated by day.**

```
data_joined <- merge(x = data_survey_3, y = data_COVID19_3, by.x=c('Time.Period.End.Date'), by.y=c('subm
```

**Add new column: rolling Correlation coefficient (Pearson method).**

```
data_joined[ , 'Correlation_Coef'] <- NA
for (i in 1:nrow(data_joined)) {
  data_joined[i, 4] = as.numeric(cor(data_joined[1:i, 2], data_joined[1:i, 3]))}
data_joined[1, 4] = 0
data_joined$Correlation_Coef <- round(data_joined$Correlation_Coef, digits = 2)
```

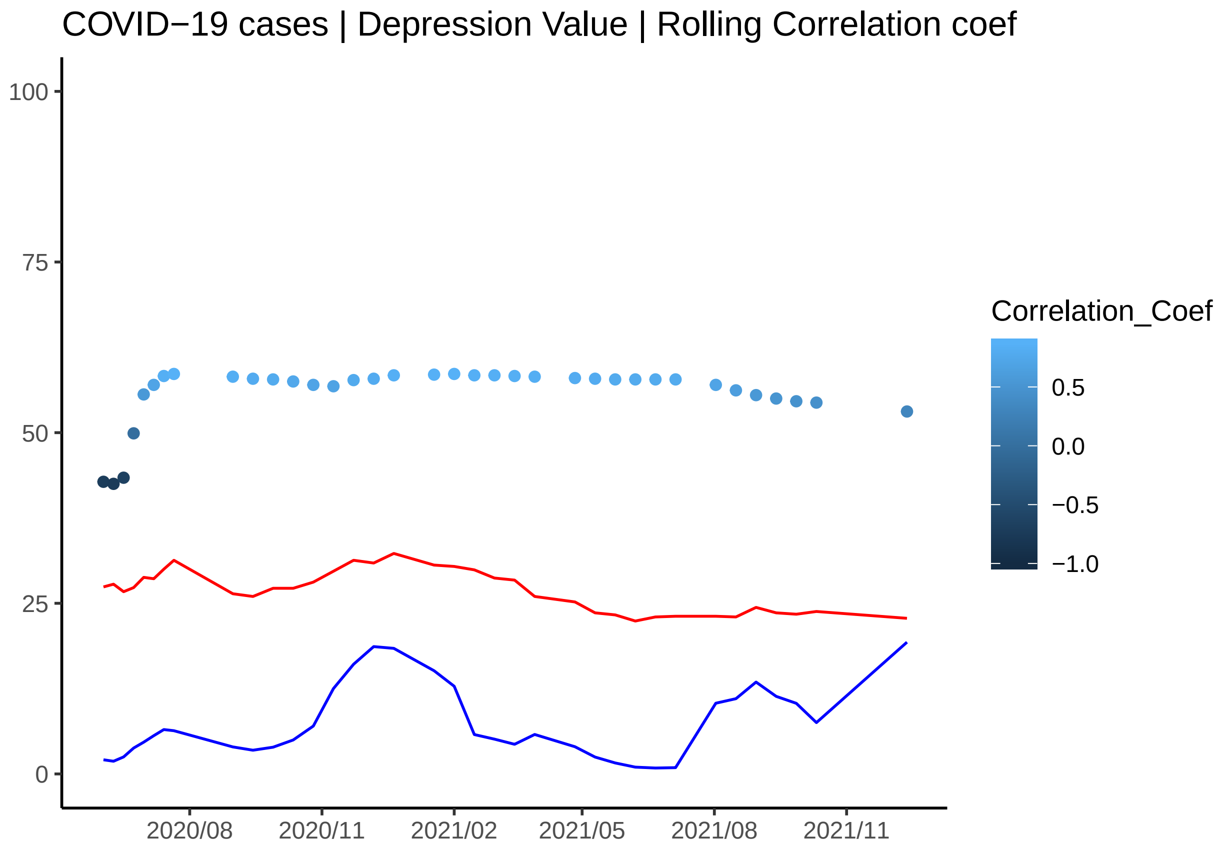
**EDA: Combine graphs: Depression Value + COVID19 cases + rolling Corr. coef.**

```
ggplot(data = data_joined) +
  geom_line(aes(x = Time.Period.End.Date, y = new_case_daily / 10000), color='blue') +
  geom_line(aes(x = Time.Period.End.Date, y = Value), color='red') +
  aes(x = Time.Period.End.Date, y = Correlation_Coef * 10 + 50, color = Correlation_Coef) +
  labs(x = NULL, y = NULL, title = 'COVID-19 cases | Depression Value | Rolling Correlation coef') +
  geom_point(shape = 'circle') +
  lims(y = c(0, 100)) +
  scale_x_date(limit=c(as.Date('2020-06-01'), as.Date('2021-12-13')), date_breaks = '3 month', date_labe
  theme_classic()
```

```
## Warning: Removed 4 rows containing missing values (geom_path).

## Warning: Removed 4 rows containing missing values (geom_path).

## Warning: Removed 4 rows containing missing values (geom_point).
```



Overlaying graphs under a common scale (Y-axis) allows you to see the relationship between variables until mid-2021. Further, the correlation decreases.

**Build LR model.**

```
linear_regression = lm(Value ~ new_case_daily, data = data_joined)
summary(linear_regression)
```

```
##
```

```
## Call:
## lm(formula = Value ~ new_case_daily, data = data_joined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0507 -2.6301  0.7231  2.0666  4.6061
##
## Coefficients:
##                    Estimate   Std. Error t value            Pr(>|t|)
## (Intercept)    25.640996735  0.706207121  36.308 <0.0000000000000002 ***
## new_case_daily  0.000016628  0.000008143   2.042              0.0481 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.739 on 38 degrees of freedom
## Multiple R-squared:  0.09888,    Adjusted R-squared:  0.07516
## F-statistic:  4.17 on 1 and 38 DF,  p-value: 0.04814
```

### 5. Summary.

**P-value** (0.04814) is less than 0.05 which means the **relationship is statistically significant** and indicates strong evidence against the null hypothesis.

### 6. Conslusion.

There is a relationship between the occurrence of female depression and the intensity of the COVID-19 pandemic. Based on the rolling correlation coefficient, the strongest relationship was observed in the first year of the pandemic. Since the second half of 2021, the relationship has been gradually decreasing. This can be interpreted as an acquired tolerance to the pandemic.

### 6. References.

Datasets, support files, .PDF version of the observation are available in the *Link:* Github repo