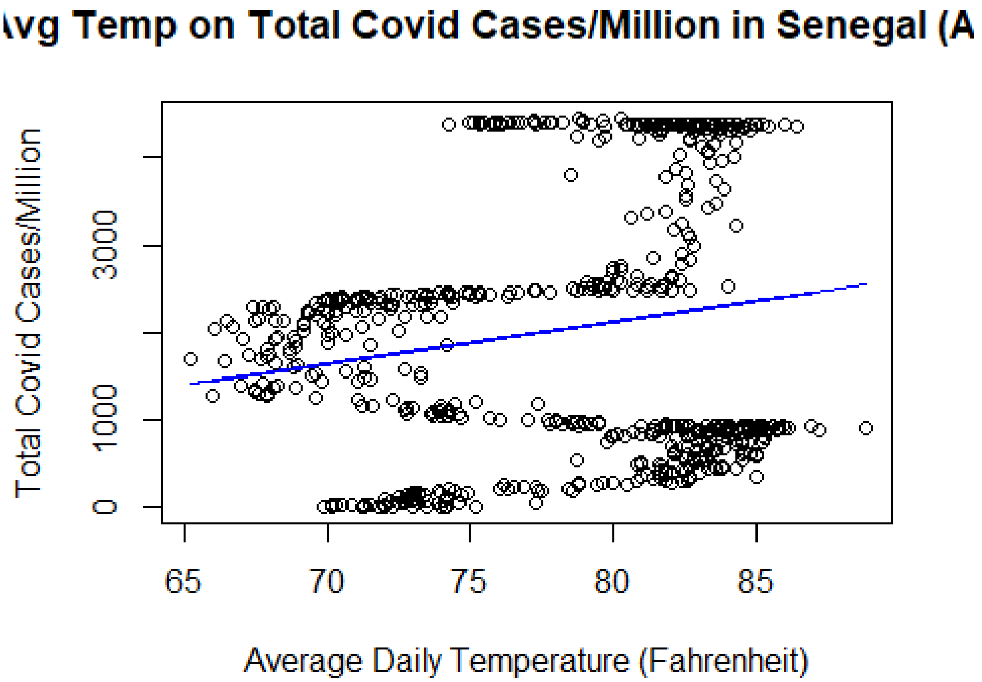
stats_challenge

2022-09-14

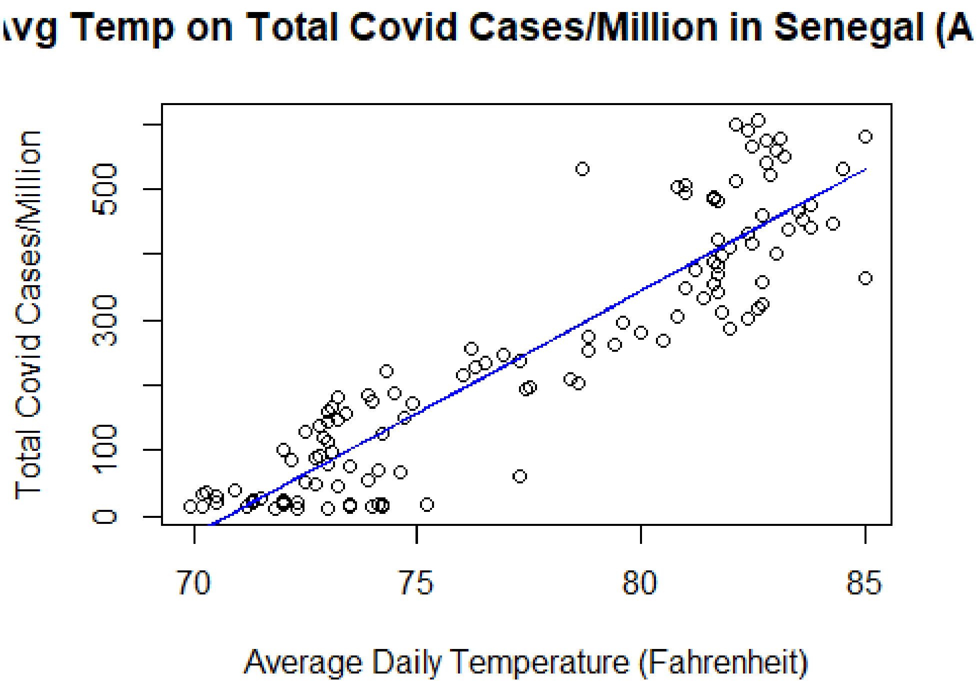
Goal:

We are interested in understanding how the average daily temperature in Senegal relates to the total number of COVID-19 cases per million residents.

```
#setwd
setwd("C:/Users/san2031/OneDrive - med.cornell.edu/Desktop/Biosciences
Bootcamp")

#read in covid data
covid <- read.csv("owid-covid-data.csv", stringsAsFactors = FALSE)

#subset covid data to senegal
covid_sen <- subset(covid, covid$location == "Senegal")

#read in weather data for senegal
weather <- read.csv("Senegal_Temp_Data.csv", stringsAsFactors = FALSE)

#merge on date field
merged <- merge(covid_sen, weather, by=c("date"))

#model the data to a lin reg
model1 <- lm(merged$total_cases_per_million ~ merged$avgtemp)
summary(model1) # p-value: 7.78e-06 (less than 0.0001 so significant)

##
## Call:
## lm(formula = merged$total_cases_per_million ~ merged$avgtemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1999.8 -1482.9   227.8  1162.2  2536.2
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1729.22     837.20  -2.065   0.0393 *
## merged$avgtemp     48.14      10.68   4.508 7.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1494 on 638 degrees of freedom
## Multiple R-squared:  0.03087,    Adjusted R-squared:  0.02935
## F-statistic: 20.32 on 1 and 638 DF,  p-value: 7.783e-06
```



```
#cherry pick partial data to only include a season where weather had a
seemingly strong relation to covid cases
subset <- subset(merged, merged$date < "2020-08-01")
model2 <- lm(subset$total_cases_per_million ~ subset$avgtemp)
summary(model2)

##
## Call:
## lm(formula = subset$total_cases_per_million ~ subset$avgtemp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -182.811  -53.519   -3.604   54.233  236.624
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2643.134    117.234  -22.55   <2e-16 ***
## subset$avgtemp    37.343      1.513   24.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.96 on 120 degrees of freedom
## Multiple R-squared:  0.8355, Adjusted R-squared:  0.8341
## F-statistic: 609.4 on 1 and 120 DF,  p-value: < 2.2e-16
```



Subjective Conclusion:

We found a statistically significant positive relationship between total covid cases per million people and the average daily temperature in Senegal (p value = 7.783e-06). This would indicate that warmer weather is causing the increase of Covid 19 cases. This is illustrated perfectly in the Spring and Summer seasons of 2020; as the daily temperature steadily increased, the number of total covid cases per million people also increased daily, with an R squared value of 0.84. In this period of time, the significance of the relationship also increased as the p-value dropped to < 2.2e-16. This shows that the rise in temperature is having an increasing effect on the number of total covid cases.

Explanation:

Of course, the average daily temperature does not actually have a significant increasing effect on the number of covid cases and the correlation we see here is mostly likely due to the fact that the beginning of the pandemic happened to start at the end of Winter, so as time went on, both temperature and cases were going to increase independently. Cherry-picking the data to only include the Spring and Summer months of 2020, is a good example of p-hacking and manipulation as the p-value decreased by a lot and the R squared value completely changed from 0.03 for the whole data set, to 0.84. After these summer months, as the temperature cools, the total covid cases still increase but the temperature decreases which creates a lot less strong of a correlation and statistical significance. ```