BIA 652

Final Team Project

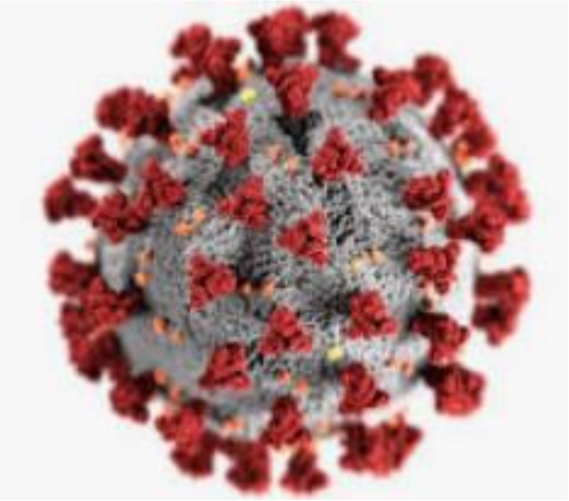
COVID-19 Impact / Effect

Multivariate Data Analysis

Dec 9, 2020

Team COVID-19 Impact

- Julia Suzuki
- Kevin Ho
- Kendra Rusinek
- Michael Weiss



Coronavirus (COVID-19)

Agenda

Topics	
Introduction	Goal & Objectives
Trend Analysis	COVID-19
	US Stock Market (S&P 500 Stock Prices)
	Mental Health (Depressed Feelings)
Statistical Analysis for Question #1	Statistical Methods & Techniques
	Analysis Results
	Conclusion & Next Steps
Statistical Analysis for Question #2	Statistical Methods & Techniques
	Analysis Results
	Conclusion & Next Steps

COVID-19 is having a profound impact on the US Economy and Mental Health now and in the future.

- The global economic impact of COVID-19 and stock market seem to be disconnected. And many might think that a rising market is an evidence of a strong economy; however, experts say market is not the economy; there is **no correlation between economy (i.e. GDP) and stock market (i.e. S&P 500)**
- Mental health experts state that there is an **association between the COVID-19 pandemic and mental health**

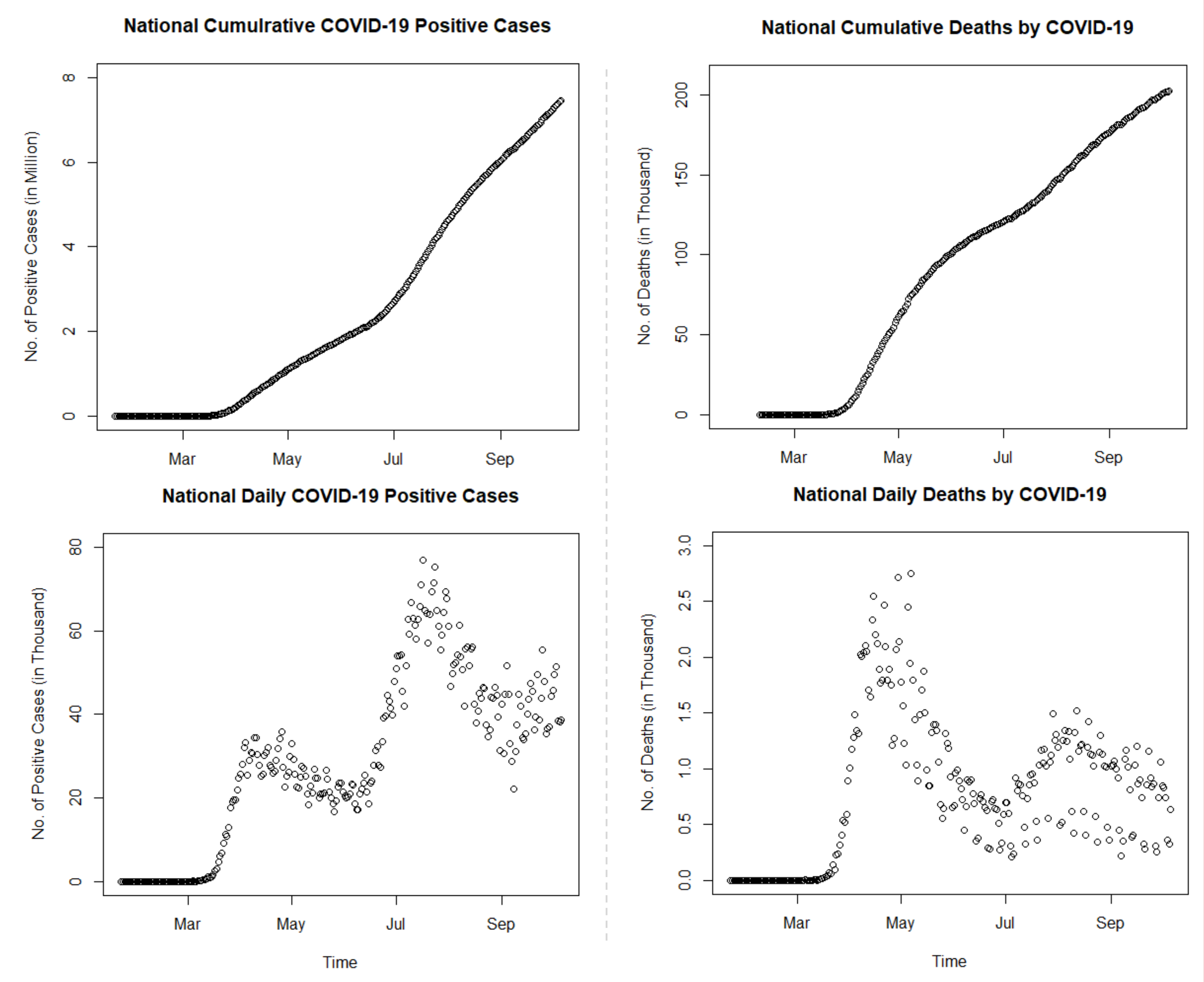
Experts warn of urgent need for Covid-19 mental health research

America's mental health Covid-19 recovery needs to start now

Stock Market Is Not the Economy by An...
bloomberg.com

Our goal is to answer Economic and Mental Health questions by analyzing data using statistical methods and techniques learned in the class.

Initial analysis showed the numbers of **Positive Cases and Death increased over time**. In addition, **Daily Positive Cases and Daily Deaths spiked in May and August** respectively.

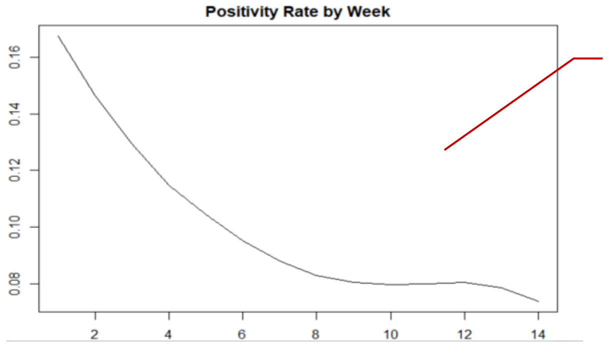


National Cumulrative COVID-19 Positive Cases

National Cumulative Deaths by COVID-19

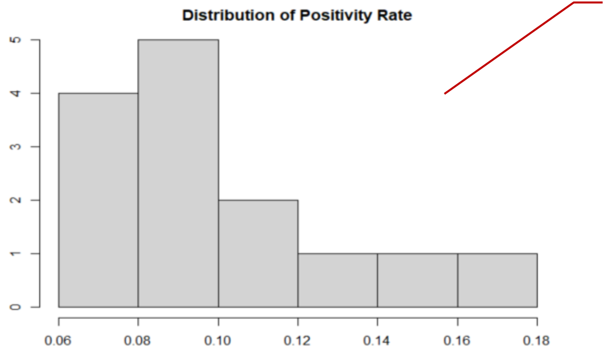
National Daily COVID-19 Positive Cases

National Daily Deaths by COVID-19



Positivity Rate by Week

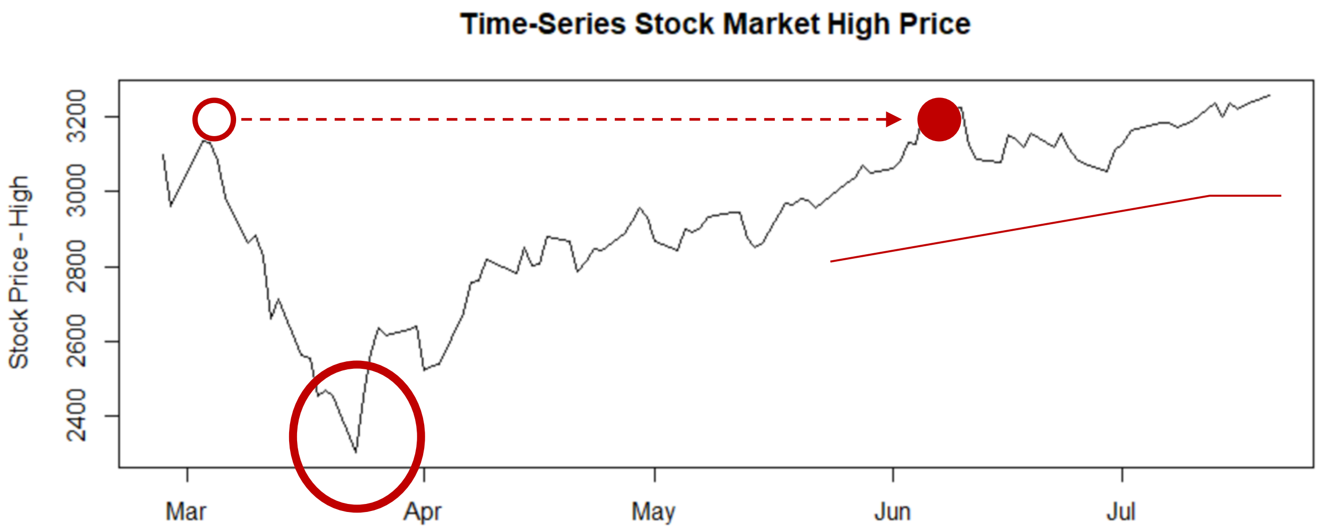
Positivity Rate decreased over time. This aligned with the increased number of tests



Distribution of Positivity Rate

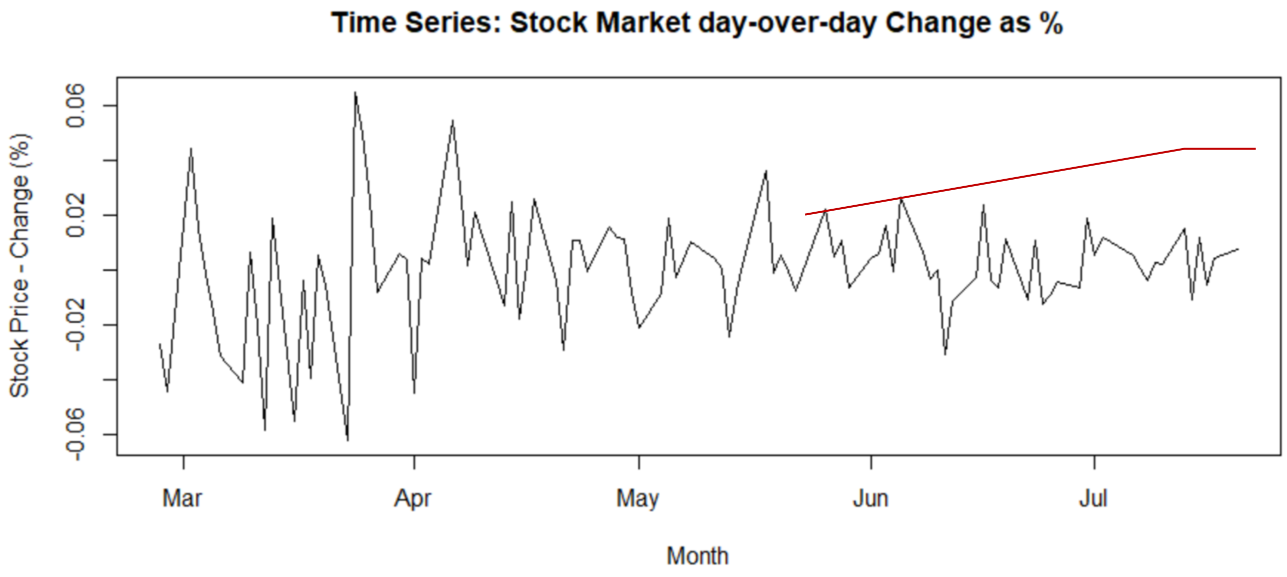
The distribution of Positivity Rate is skewed to the right, showing that there is a lean towards less daily positive cases per daily test cases

Plotted US Stock Market Prices (high) and Day-Over-Day Change over time and observed trends



Time-Series Stock Market High Price

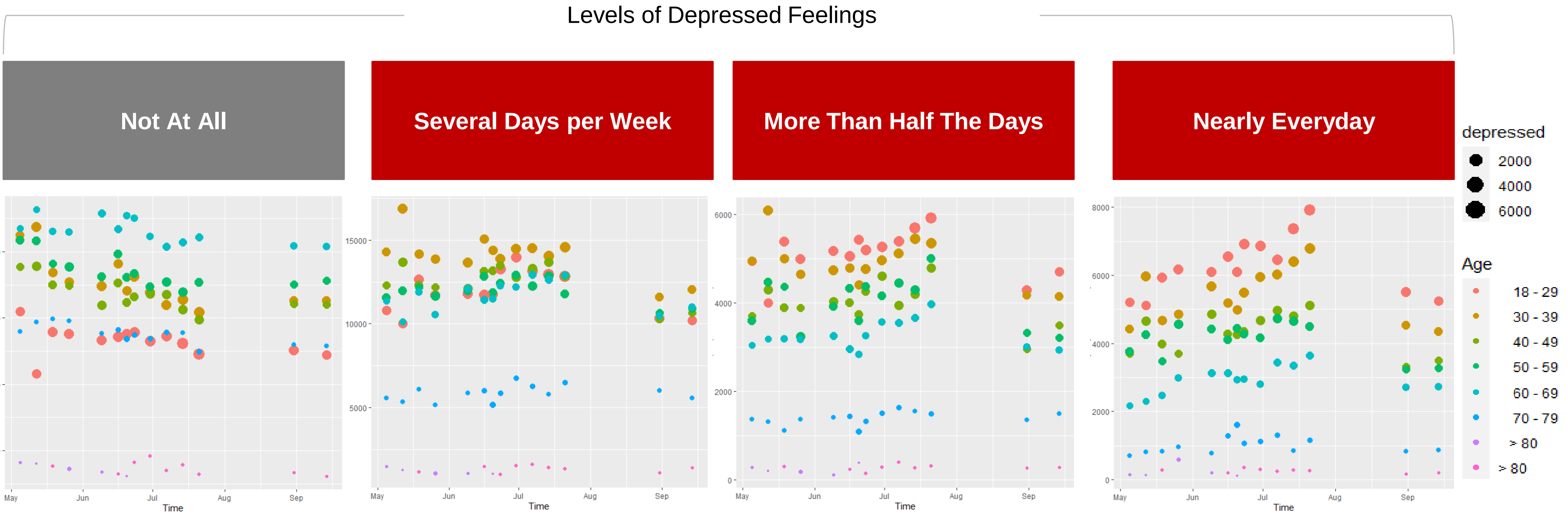
Stock Price observed a sharp dip in March which corresponds with initial lockdowns; however, grew to initial levels by August



Time Series: Stock Market day-over-day Change as %

Day-over-day change did not show a significant trend and the % of change oscillated between positive and negative with high volatility observed in March and .

Scatter Plots show Varying Levels of Depressed Feelings felt by Different Age Groups. Younger Populations (age between 18 and 49) are More Impacted by COVID-19 than Older Populations (age over 70 and above).



Levels of Depressed Feelings

Not At All

Several Days per Week

More Than Half The Days

Nearly Everyday

Note: Each dot represents the size of a population who felt a certain level of depressed feelings during a time frame (a week or two weeks)

QUESTION #1

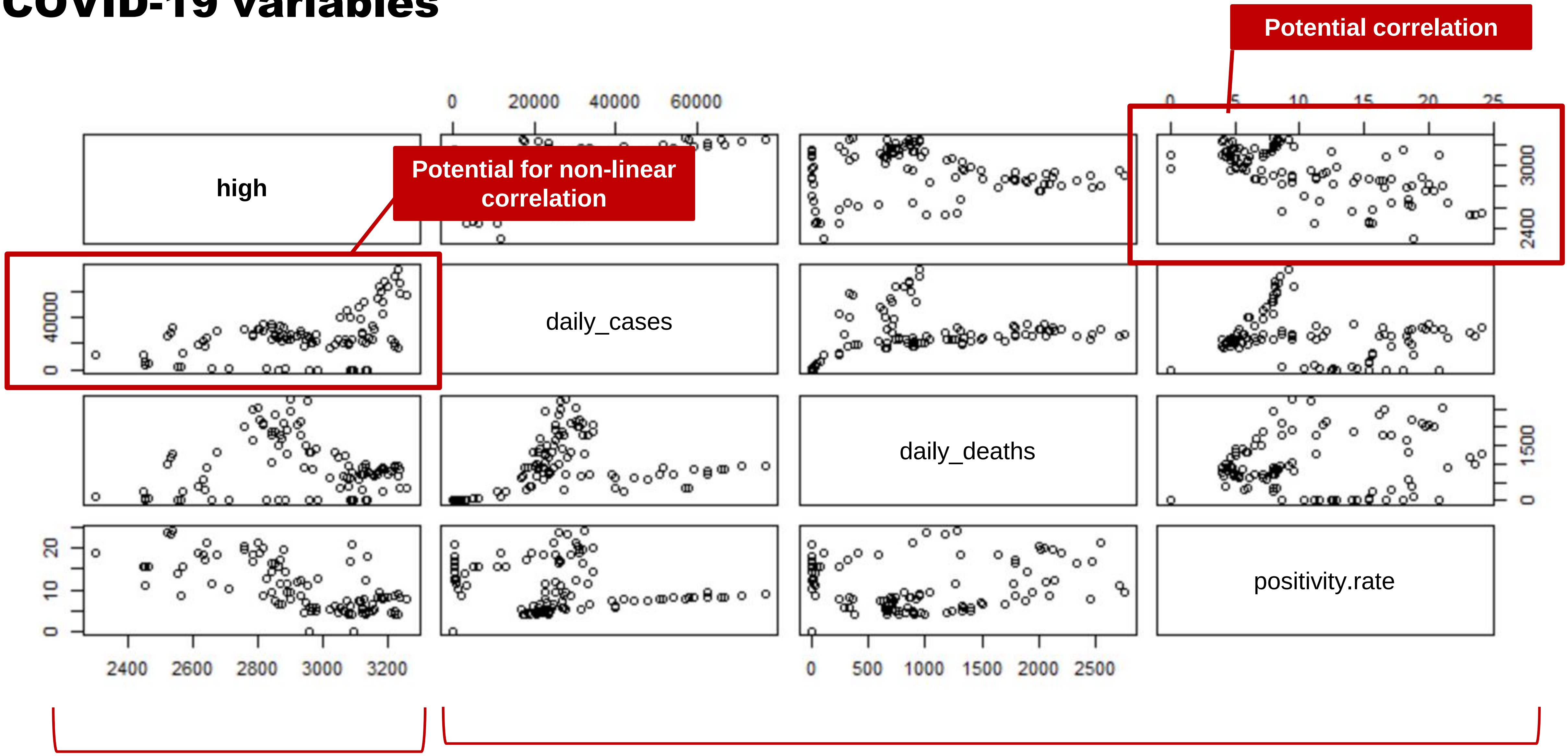
Does **COVID-19** have an effect on the **US Stock Market**?

More specifically, does **COVID-19** (represented by the **Positivity Rate, Positive Cases** and **Deaths**) have a negative impact on **S&P 500 Stock Prices?**

As part of our project, we created **3 Models** on the **Impact of COVID-19 on the S&P 500**

1. Multivariate regression based on the **daily high price** of **S&P 500**
2. Multivariate regression based on the **day-over-day changes** of **S&P 500**
3. Simple Linear regression based on the **daily returns** of **S&P 500**

Analyzed the correlation between S&P 500 High Stock Price & COVID-19 variables



Y - Variable

X - Variables

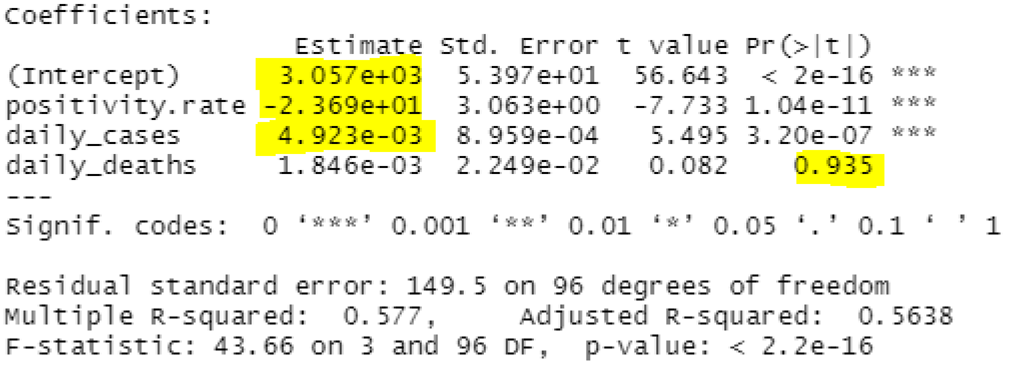
Defined the Equation for the Multiple Regression Line

$$\hat{y} = 3057 - 2.36x_1 + 0.005x_2$$

1 Variables

- COVID-19 Positivity Rate (Predictor Variable; x_1)
- COVID-19 Positive Cases per Day (Predictor Variable; x_2)
- COVID-19 Deaths per Day (Predictor Variable; x_3)
- COVID-19 Tests per Day – **excluded from model due to potential influence on other x-variables**
- S&P500 daily high stock price (Response Variable)

2 Summary Statistics



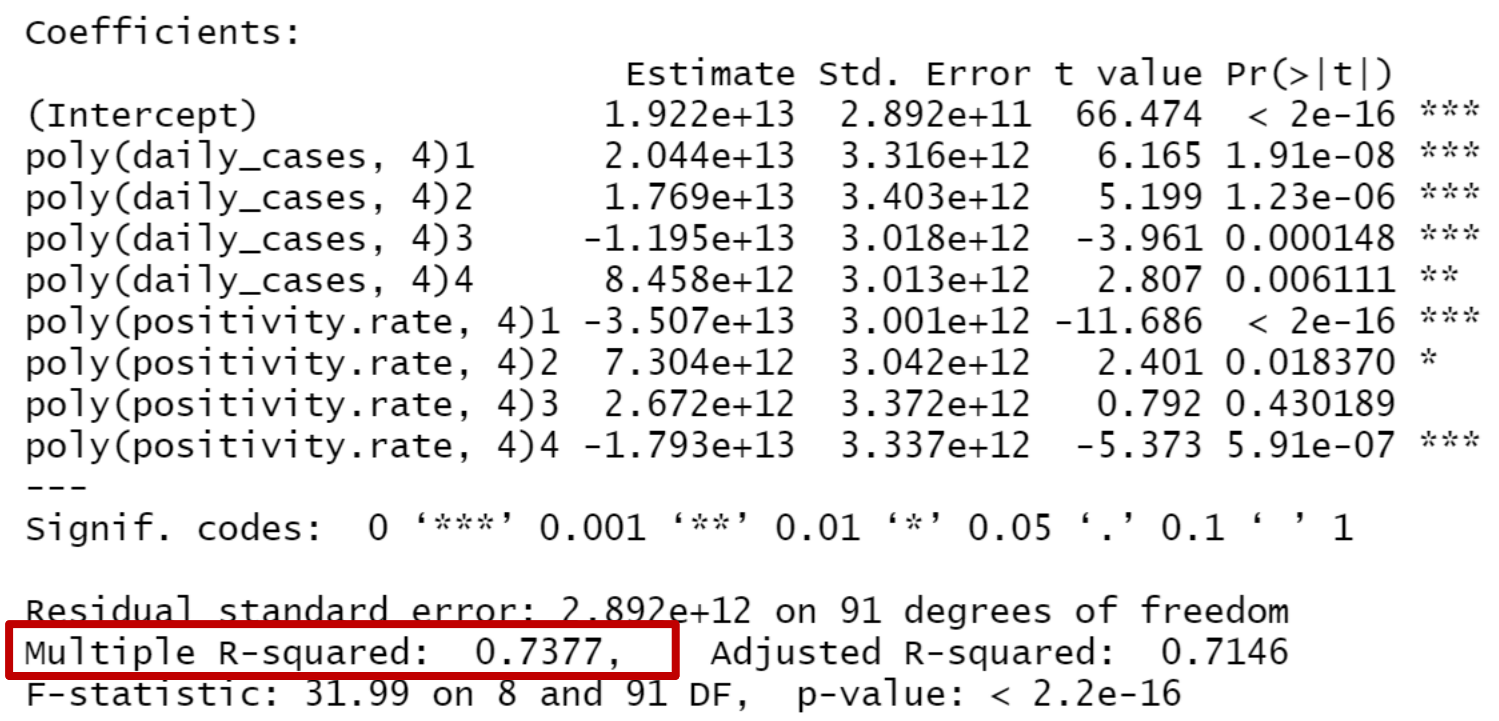
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.057e+03  5.397e+01  56.643  < 2e-16 ***
positivity.rate -2.369e+01  3.063e+00  -7.733 1.04e-11 ***
daily_cases     4.923e-03  8.959e-04   5.495 3.20e-07 ***
daily_deaths    1.846e-03  2.249e-02   0.082    0.935    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 149.5 on 96 degrees of freedom
Multiple R-squared:  0.577,	Adjusted R-squared:  0.5638 
F-statistic: 43.66 on 3 and 96 DF,  p-value: < 2.2e-16
```

Key Observations:

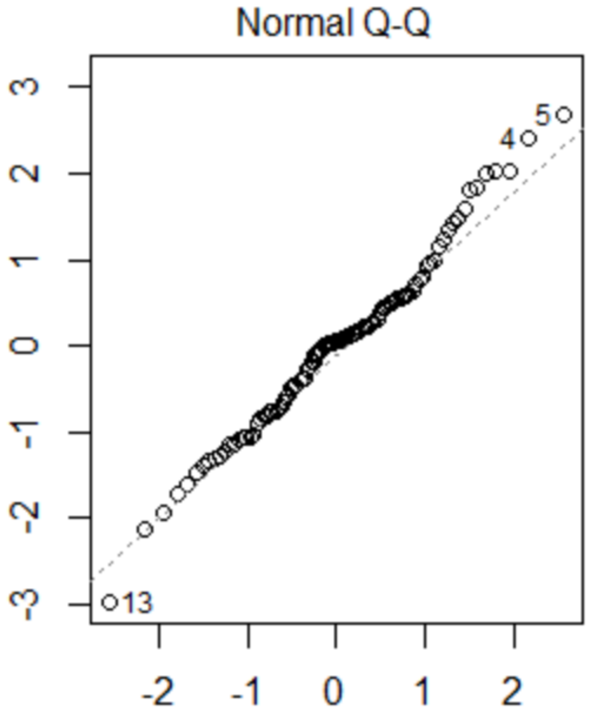
- Our regression shows that COVID-19 deaths per day (x3) is not significant; therefore we have excluded it from the equation.
- Strong, **negative** correlation among variables is observed (r: - 0.76).
- Our model explains about **57%** of the variance (r-squared: 0.577).
- Residuals plot (refer to appendix) indicates a **transformation** may be helpful

Achieved improvement in model using Box Cox Transformation & Polynomial Regression

If we apply a transformation of y using 1/5*(y^5-1) and apply polynomial regression, we can improve our R^2 significantly, as well as improve normality. We would exclude the x_2^2 and x_2^3 terms as they have minimal significance compared with other terms.



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                1.922e+13  2.892e+11  66.474  < 2e-16 ***
poly(daily_cases, 4)1      2.044e+13  3.316e+12   6.165 1.91e-08 ***
poly(daily_cases, 4)2      1.769e+13  3.403e+12   5.199 1.23e-06 ***
poly(daily_cases, 4)3     -1.195e+13  3.018e+12  -3.961 0.000148 ***
poly(daily_cases, 4)4      8.458e+12  3.013e+12   2.807 0.006111 ** 
poly(positivity.rate, 4)1 -3.507e+13  3.001e+12 -11.686  < 2e-16 ***
poly(positivity.rate, 4)2  7.304e+12  3.042e+12   2.401 0.018370 *  
poly(positivity.rate, 4)3  2.672e+12  3.372e+12   0.792 0.430189    
poly(positivity.rate, 4)4 -1.793e+13  3.337e+12  -5.373 5.91e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.892e+12 on 91 degrees of freedom
Multiple R-squared:  0.7377,	Adjusted R-squared:  0.7146 
F-statistic: 31.99 on 8 and 91 DF,  p-value: < 2.2e-16
```

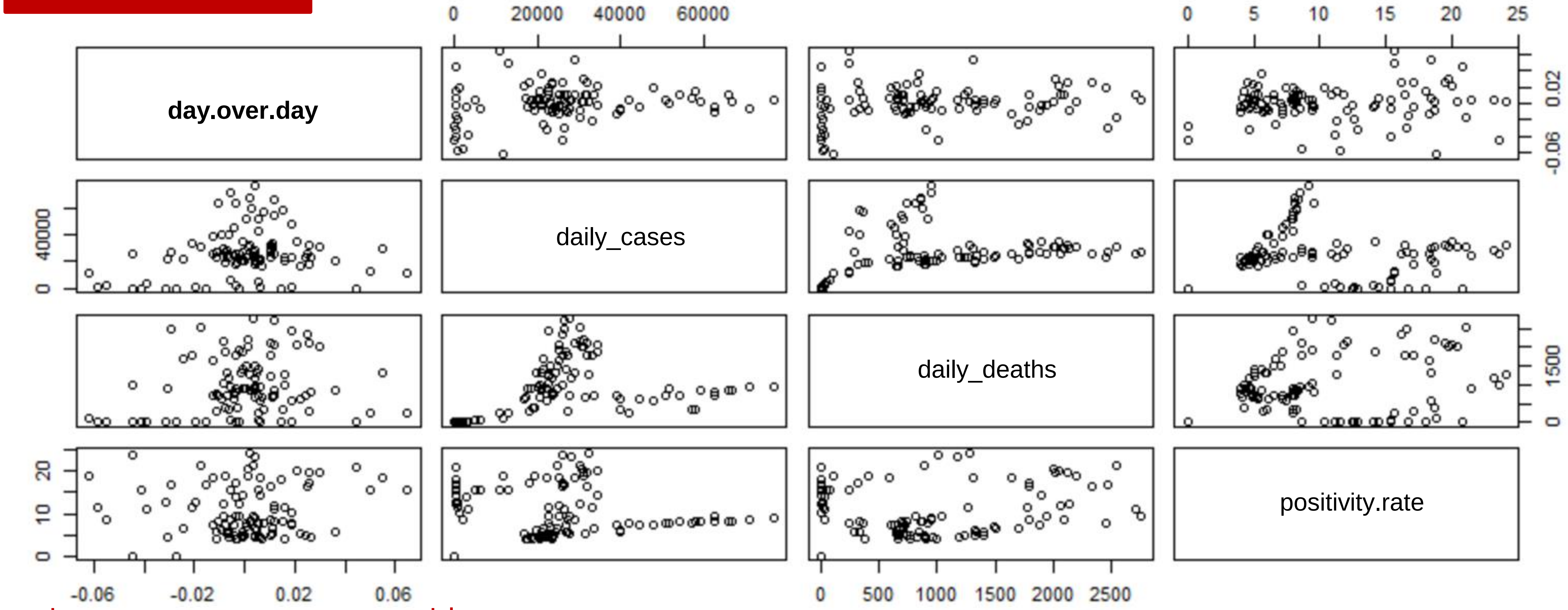


As part of our project, we created 3 Models on the Impact of COVID-19 on the S&P 500

1. Multivariate regression based on the **daily high price** of **S&P 500**
2. Multivariate regression based on the **day-over-day changes** of **S&P 500**
3. Simple Linear regression based on the **daily returns** of **S&P 500**

Analyzed the correlation between Day-Over-Day Stock Price Change & COVID-19 variables

There is no correlation observed

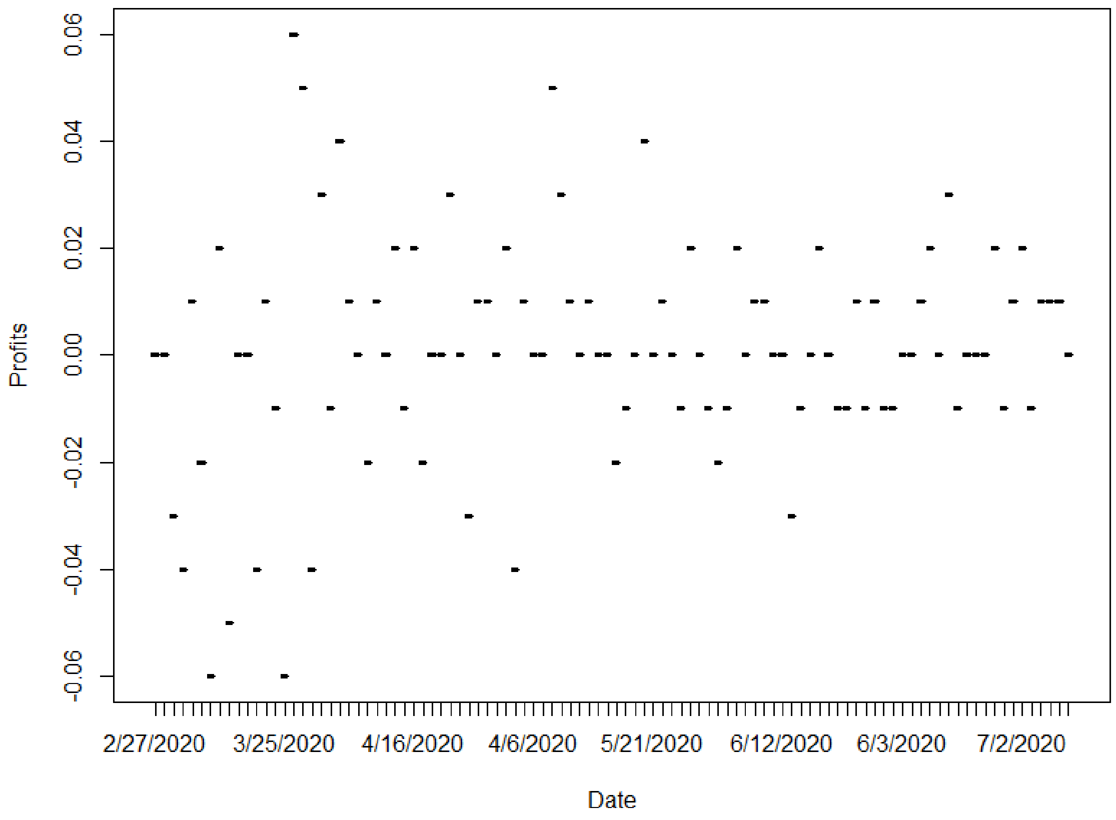


Y - Variable

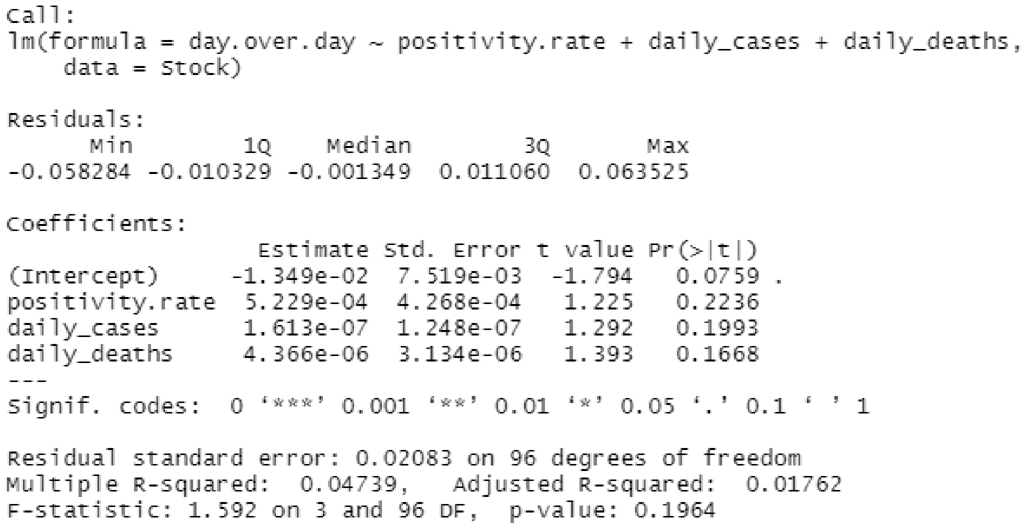
X - Variables

Linear relationship does not exist between Day-Over-Day Stock Price Change and COVID-19 variables

Day-Over-Day Profits Over Time



Summary Statistics for the Multivariate Regression Model based on the day-over-day changes of S&P500



```
Call:
lm(formula = day.over.day ~ positivity.rate + daily_cases + daily_deaths,
    data = Stock)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.058284 -0.010329 -0.001349  0.011060  0.063525 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)  
(Intercept)     -1.349e-02  7.519e-03  -1.794   0.0759 .
positivity.rate  5.229e-04  4.268e-04   1.225   0.2236  
daily_cases      1.613e-07  1.248e-07   1.292   0.1993  
daily_deaths     4.366e-06  3.134e-06   1.393   0.1668  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02083 on 96 degrees of freedom
Multiple R-squared:  0.04739,	Adjusted R-squared:  0.01762 
F-statistic: 1.592 on 3 and 96 DF,  p-value: 0.1964
```

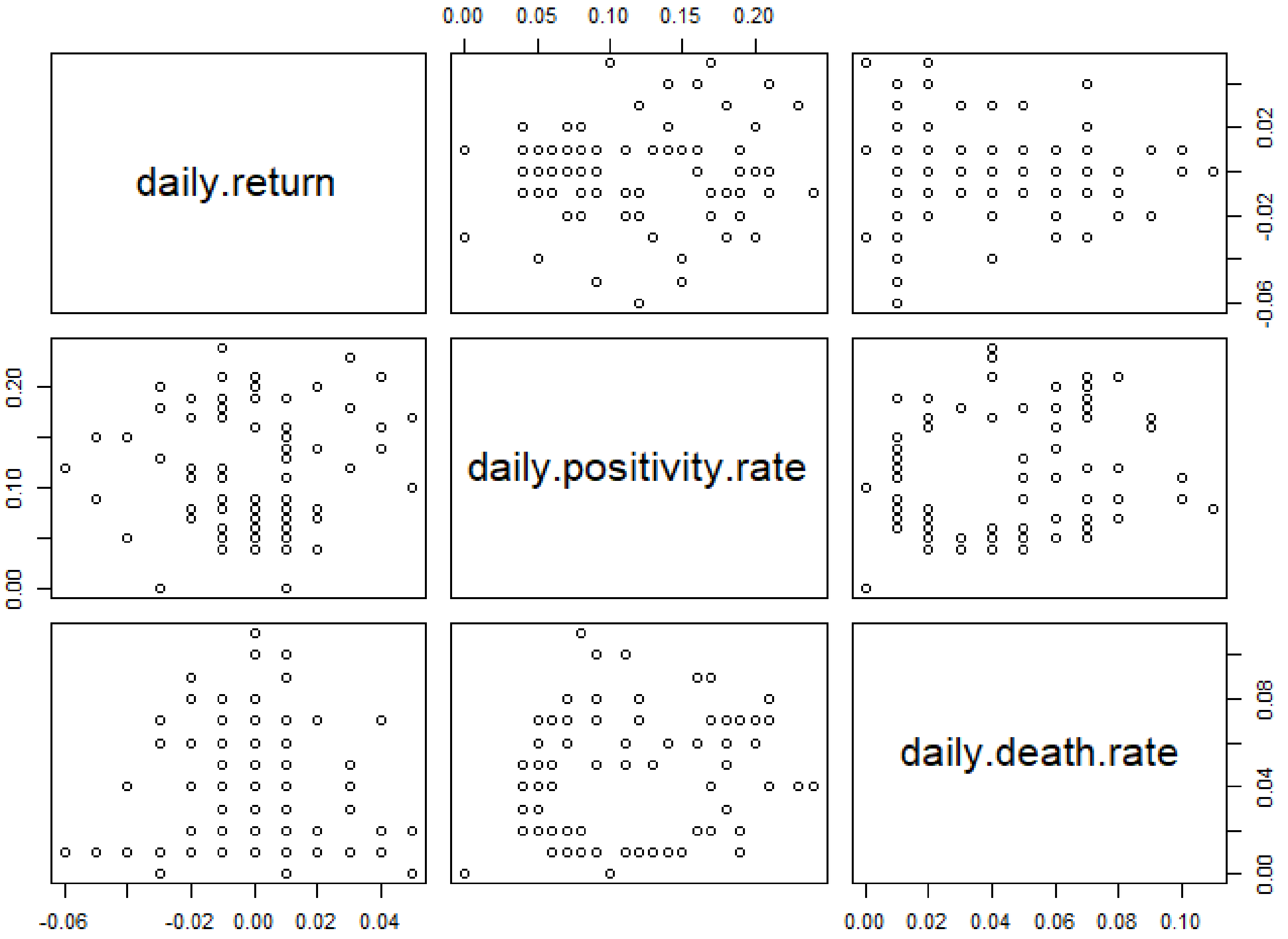
Key Observations:

- None of our variables show any significance and our R-squared is also very low and close to zero.
- This model suggests there is no impact on the S&P500 daily profits

As part of our project, we created 3 Models on the Impact of COVID-19 on the S&P 500

1. Multivariate regression based on the **daily high price** of **S&P 500**
2. Multivariate regression based on the **daily returns** of **S&P 500**
3. Simple Linear regression based on the **daily returns** of **S&P 500**

Analyzed the correlation between S&P 500 Daily Return & COVID-19 variables using Correlation Matrix and Correlation Tests

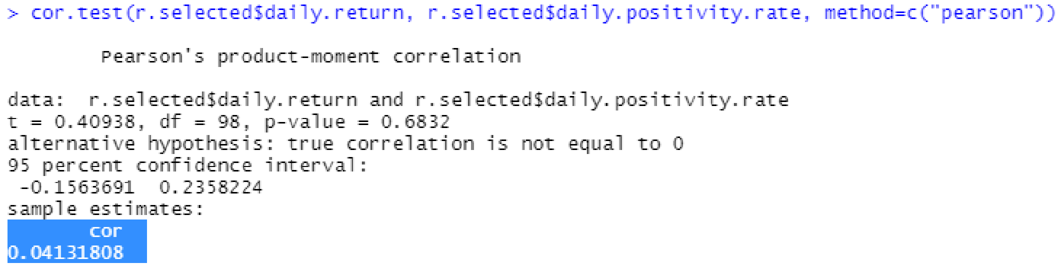


Y - Variable

X - Variables

Correlation test showed no relationship between daily returns & daily positivity rate

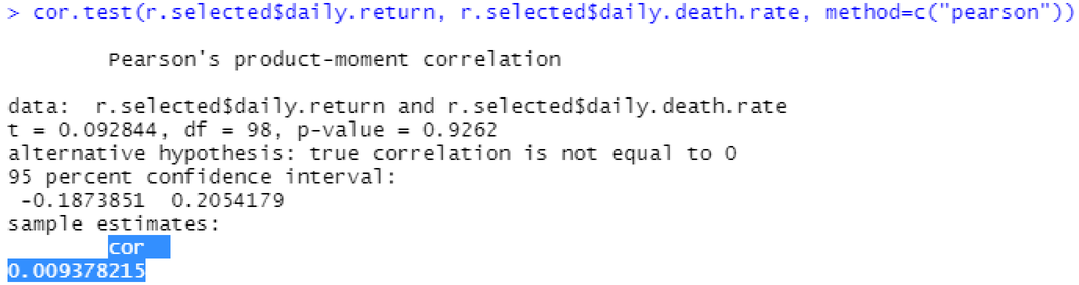
- R: 0.04
- P-Value: 0.68



```
> cor.test(r.selected$daily.return, r.selected$daily.positivity.rate, method=c("pearson"))

	Pearson's product-moment correlation

data:  r.selected$daily.return and r.selected$daily.positivity.rate
t = 0.40938, df = 98, p-value = 0.6832
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1563691  0.2358224
sample estimates:
       cor 
0.04131808 
```

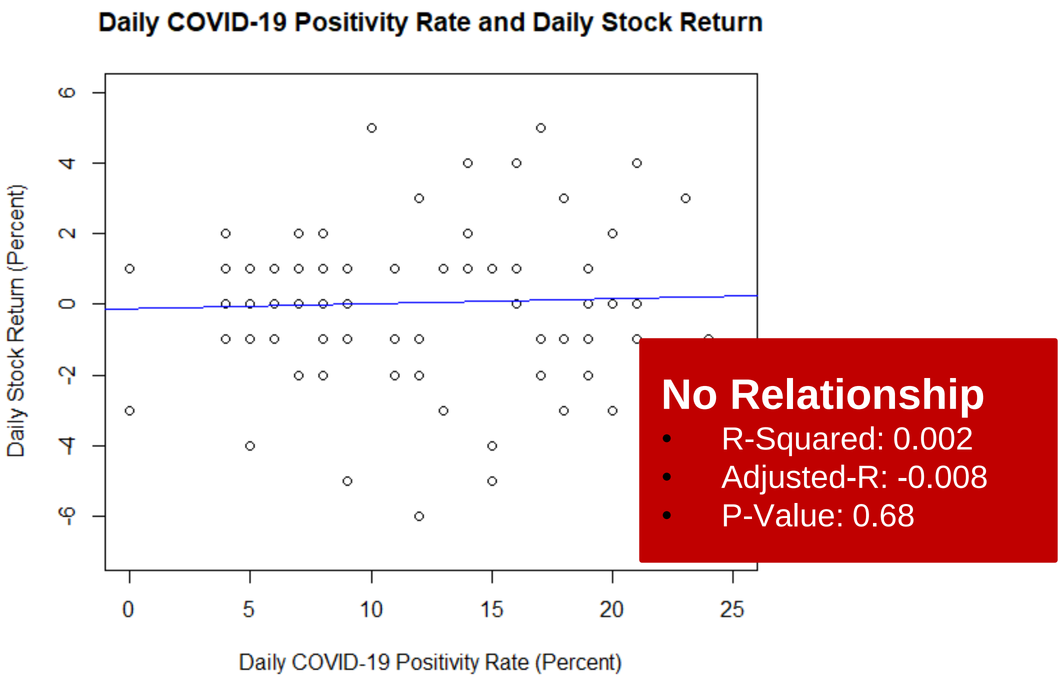
Correlation test showed no relationship between daily returns & daily death rate

- R: 0.01
- P-Value: 0.93



```
> cor.test(r.selected$daily.return, r.selected$daily.death.rate, method=c("pearson"))

	Pearson's product-moment correlation

data:  r.selected$daily.return and r.selected$daily.death.rate
t = 0.092844, df = 98, p-value = 0.9262
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1873851  0.2054179
sample estimates:
        cor 
0.009378215 
```

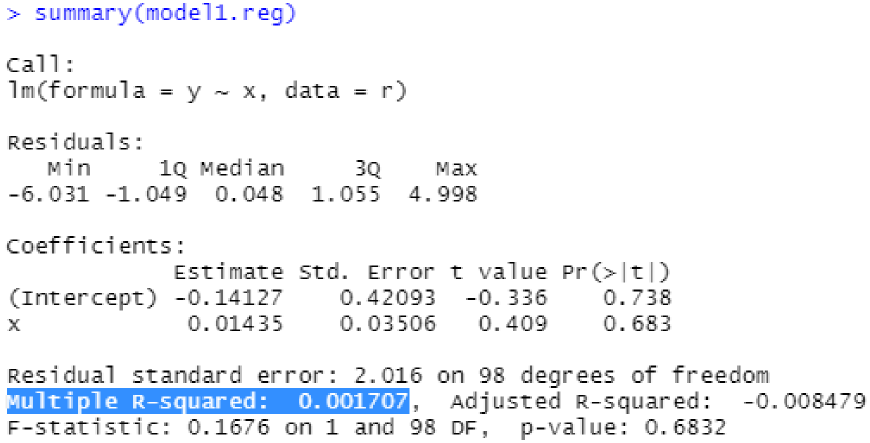
Linear relationship does not exist between S&P 500 returns and COVID-19 variables



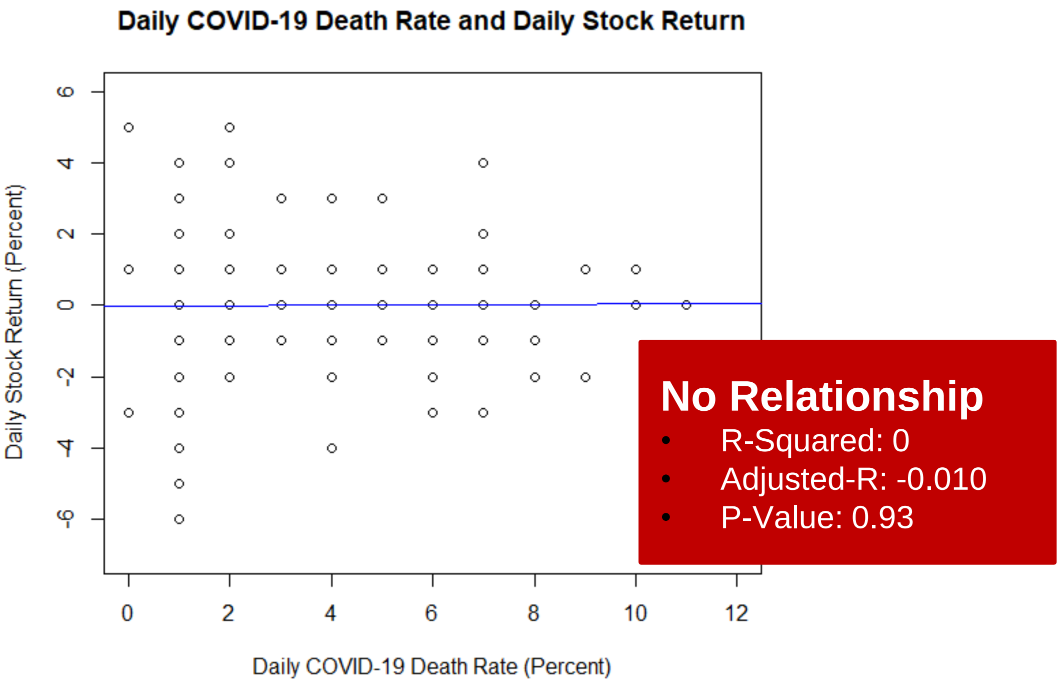
Daily COVID-19 Positivity Rate and Daily Stock Return

No Relationship

- R-Squared: 0.002
- Adjusted-R: -0.008
- P-Value: 0.68



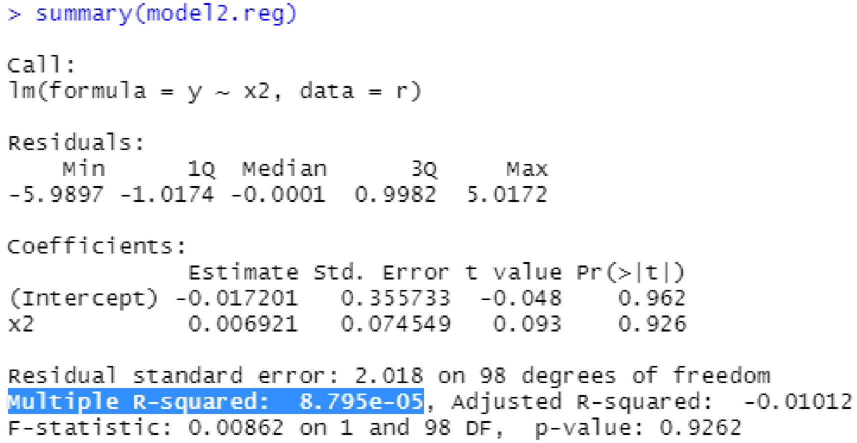
```
> summary(model1.reg)

Call:
lm(formula = y ~ x, data = r)

Residuals:
   Min     1Q Median     3Q    Max 
-6.031 -1.049  0.048  1.055  4.998 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.14127    0.42093  -0.336    0.738
x            0.01435    0.03506   0.409    0.683

Residual standard error: 2.016 on 98 degrees of freedom
Multiple R-squared:  0.001707,	Adjusted R-squared:  -0.008479 
F-statistic: 0.1676 on 1 and 98 DF,  p-value: 0.6832
```



Daily COVID-19 Death Rate and Daily Stock Return

No Relationship

- R-Squared: 0
- Adjusted-R: -0.010
- P-Value: 0.93



```
> summary(model2.reg)

Call:
lm(formula = y ~ x2, data = r)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9897 -1.0174 -0.0001  0.9982  5.0172 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.017201   0.355733  -0.048    0.962
x2           0.006921   0.074549   0.093    0.926

Residual standard error: 2.018 on 98 degrees of freedom
Multiple R-squared:  8.795e-05,	Adjusted R-squared:  -0.01012 
F-statistic: 0.00862 on 1 and 98 DF,  p-value: 0.9262
```

CONCLUSION

- We initially assumed that there would be a **negative relationship between COVID-19 cases** (represented by the positivity rate) **and the S&P 500**. When looking at the daily high price of the S&P 500, this **assumption is confirmed.**

- However, after **changing our Y-variable from daily high price to Day-Over-Day Profits**, this hypothesis was overturned. We **do not see any relationship** now.

Next Steps:

- Determine why both models **provide such different results**
- **Determine if there is multicollinearity** between variables in the first model

QUESTION #2

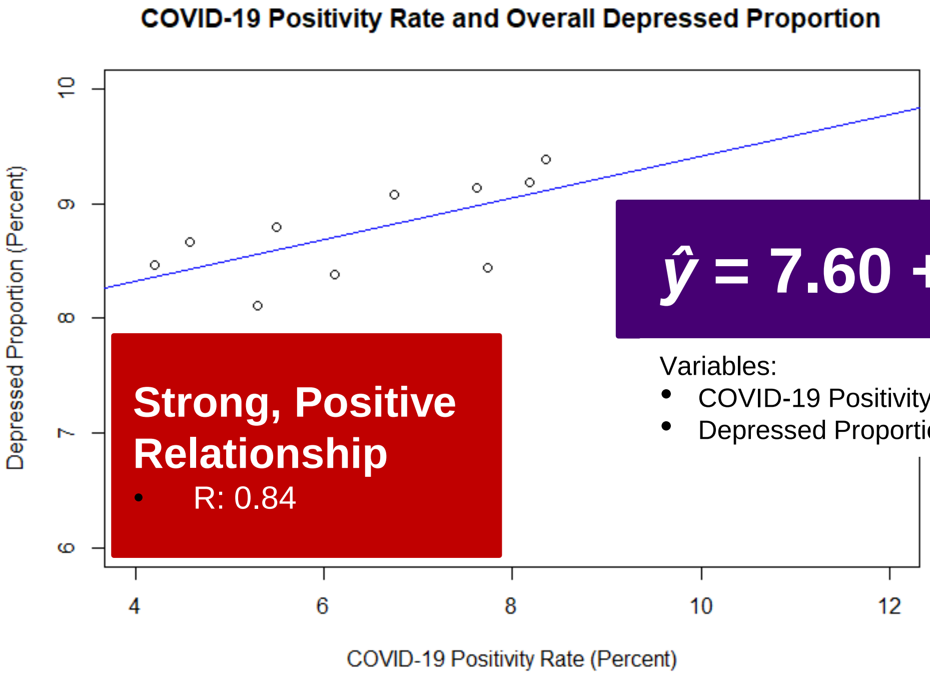
Does **COVID-19** have effect on **Mental Health**?

More specifically, does **COVID-19** (represented by the **Positivity Rate** and **Death Rate**) have influence on **Depression**?

Analyzed the Impact of COVID-19 on Depressed Feelings by creating 2 models and transforming one of the models. Also, **Tested the Relationship between Two Factors** using Two-Table Analysis method and the Chi-Square Test.

1. Simple liner regression with **only 1 COVID-19 explanatory variable**
2. Multivariate regression with **2 COVID-19 explanatory variables**
3. Two-table analysis between **age and depression**

Simple Linear Regression Models show there are moderately strong linear relationships between Depression and 1 COVID-19 Variable



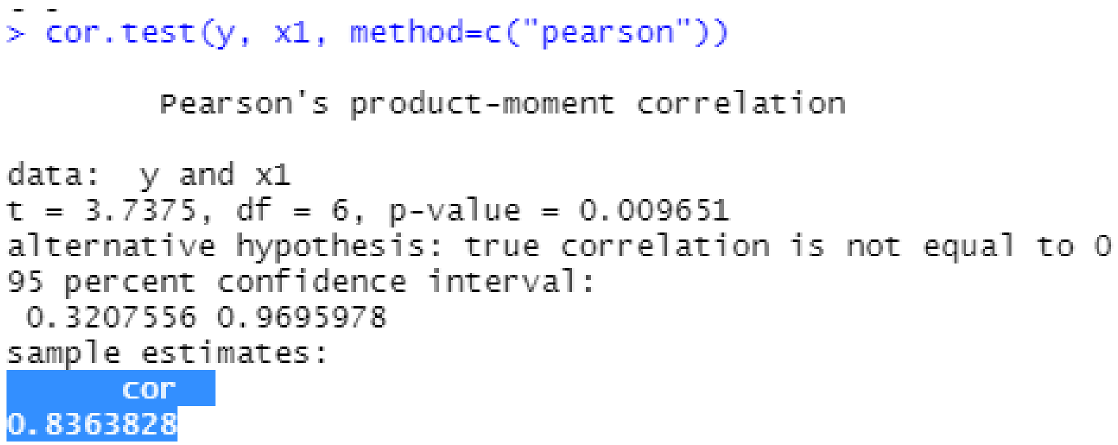
$\hat{y} = 7.60 + 0.18x$

Variables:

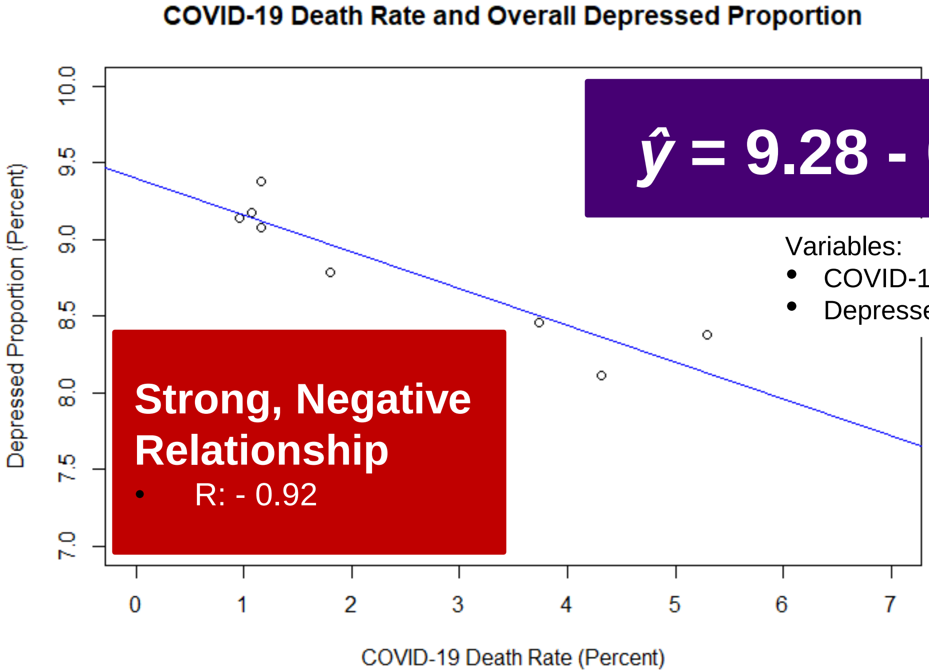
- COVID-19 Positivity Rate (x)
- Depressed Proportion ($\hat{y}$)

Strong, Positive Relationship

- R: 0.84



```
> cor.test(y, x1, method=c("pearson"))

        Pearson's product-moment correlation

data:  y and x1
t = 3.7375, df = 6, p-value = 0.009651
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3207556 0.9695978
sample estimates:
      cor
0.8363828
```



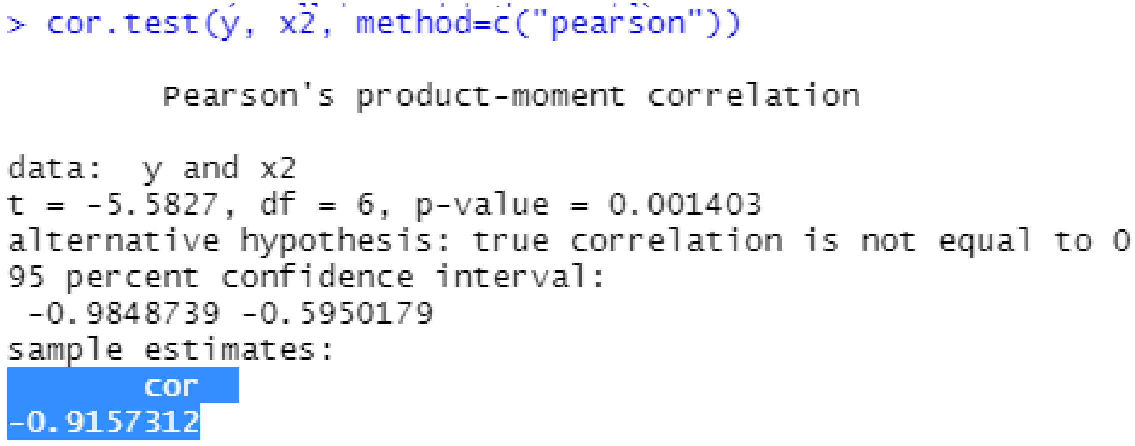
$\hat{y} = 9.28 - 0.18x$

Variables:

- COVID-19 Death Rate (x)
- Depressed Proportion ($\hat{y}$)

Strong, Negative Relationship

- R: - 0.92



```
> cor.test(y, x2, method=c("pearson"))

        Pearson's product-moment correlation

data:  y and x2
t = -5.5827, df = 6, p-value = 0.001403
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9848739 -0.5950179
sample estimates:
       cor
-0.9157312
```

Analyzed the Impact of COVID-19 on Depressed Feelings by creating 2 models and transforming one of the models. Also, **Tested the Relationship between Two Factors** using Two-Table Analysis method and the Chi-Square Test.

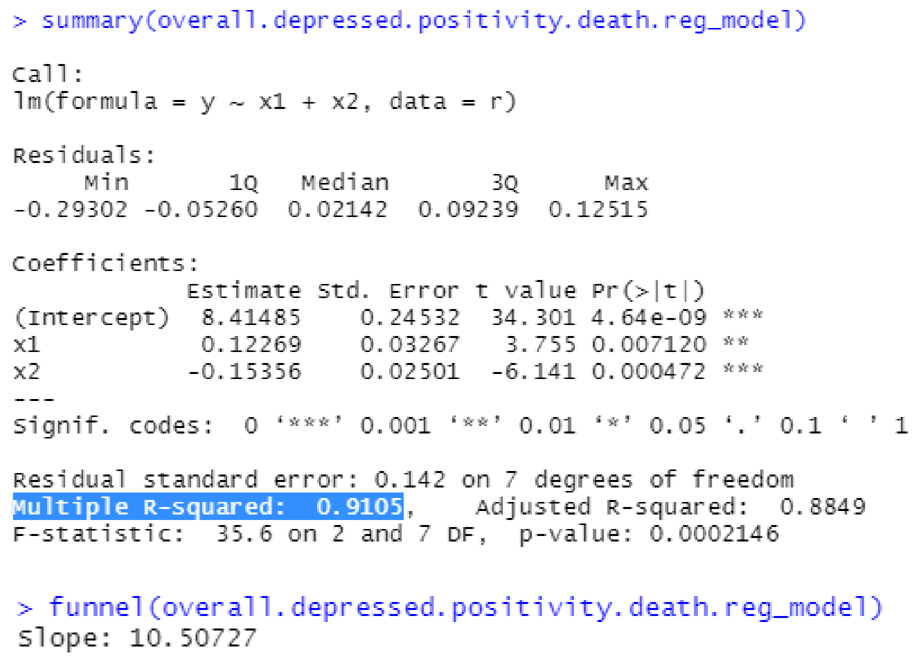
1. Simple liner regression with **only 1 COVID-19 explanatory variable**
2. Multivariate regression with **2 COVID-19 explanatory variables**
3. Two-table analysis between **age and depression**

Multivariate Regression Model showed there is a strong linear relationship between Depression and multiple COVID-19 Variables. In addition, Equality of Variance Test suggested the Transformation of y-variable to improve the model.

$$\hat{y} = 8.41 + 0.12x_1 - 0.15x_2$$

Variables:

- COVID-19 Positivity Rate (x_1)
- COVID-19 Death Rate (x_2)
- Depressed Proportion ($\hat{y}$)



```
> summary(overall.depressed.positivity.death.reg_model)

Call:
lm(formula = y ~ x1 + x2, data = r)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.29302 -0.05260  0.02142  0.09239  0.12515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.41485    0.24532  34.301 4.64e-09 ***
x1           0.12269    0.03267   3.755 0.007120 ** 
x2          -0.15356    0.02501  -6.141 0.000472 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.142 on 7 degrees of freedom
Multiple R-squared:  0.9105,	Adjusted R-squared:  0.8849 
F-statistic:  35.6 on 2 and 7 DF,  p-value: 0.0002146

> funnel(overall.depressed.positivity.death.reg_model)
Slope: 10.50727
```

Strong, Positive Relationship

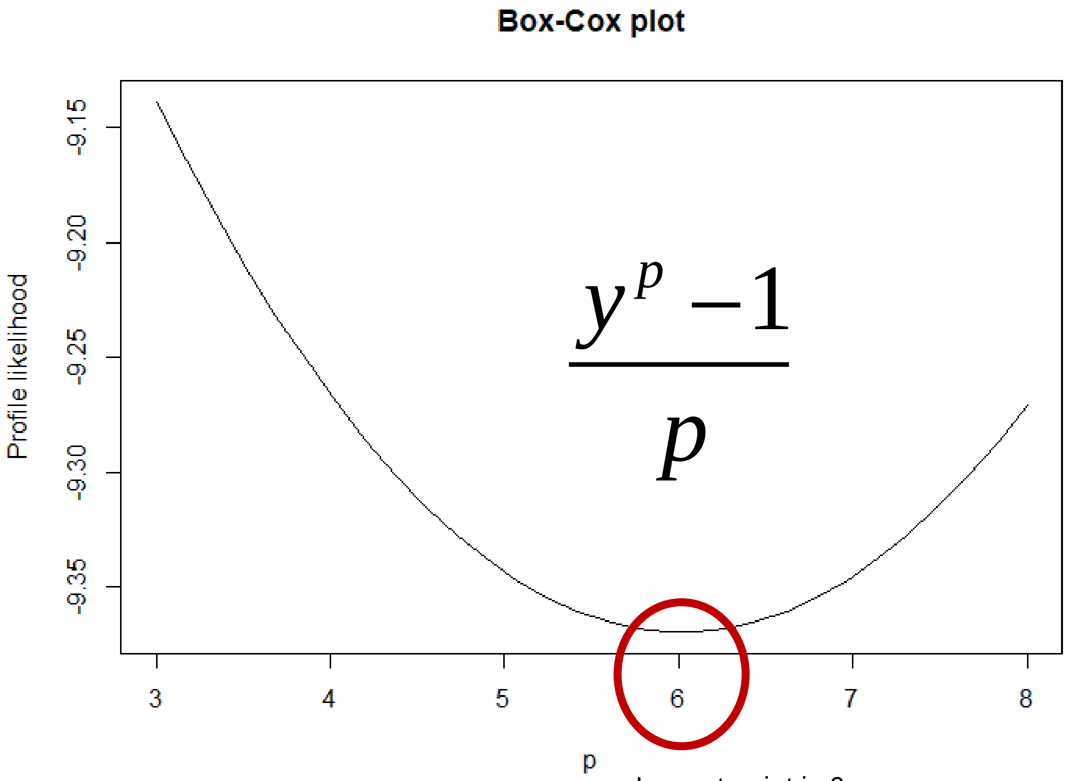
- R-Squared: 0.91
- Adjusted-R: 0.89
- P-Value: 0.0002

Equality of Variance Test showed the model can benefit the transformation of y-variable

- Slope: 10.50

UPDATED

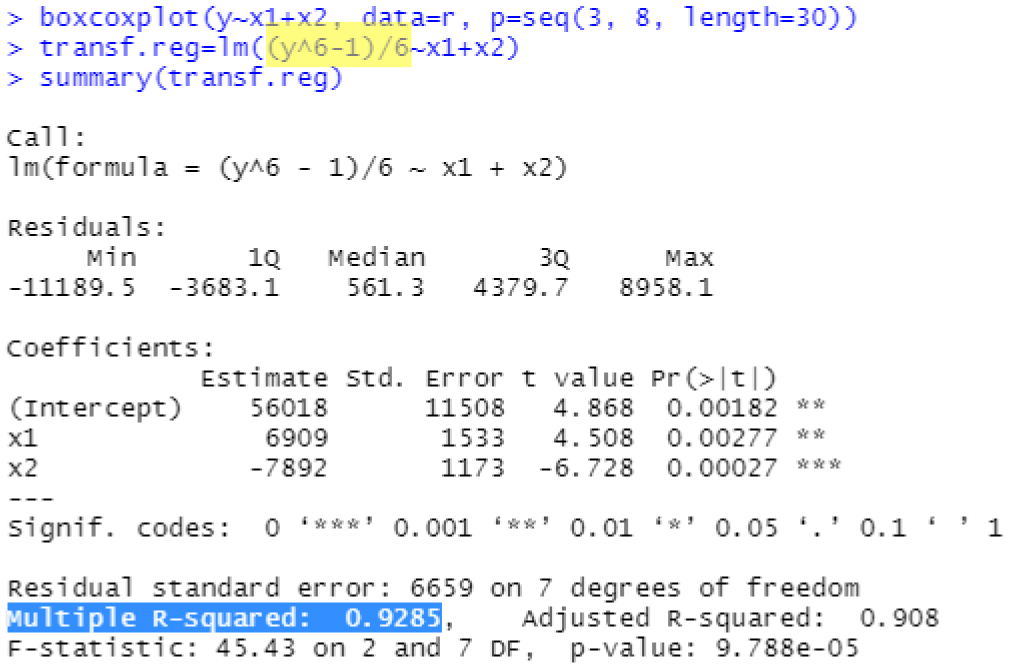
Box-Cox Transformation improved the Multivariate Regression Model with R-Squared of 0.93



$$\frac{y^{p}-1}{p}$$

Lowest point is 6

boxcoxplot(y~x1+x2, data=r, p=seq(3, 8, length=30))



```
> boxcoxplot(y~x1+x2, data=r, p=seq(3, 8, length=30))
> transf.reg=lm((y^6-1)/6~x1+x2)
> summary(transf.reg)

Call:
lm(formula = (y^6 - 1)/6 ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-11189.5  -3683.1    561.3   4379.7   8958.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    56018      11508   4.868  0.00182 ** 
x1              6909       1533   4.508  0.00277 ** 
x2             -7892       1173  -6.728  0.00027 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6659 on 7 degrees of freedom
Multiple R-squared:  0.9285,	Adjusted R-squared:  0.908 
F-statistic: 45.43 on 2 and 7 DF,  p-value: 9.788e-05
```

The model has improved !

- R-Squared: 0.93
- Adjusted R: 0.91
- P-Value: 9.788e-05

UPDATED

Comparison matrix below shows that the Multivariate Regression Model with the Box-Cox Transformation is the Best Model because of its optimized key metrics

				Key Metrics		
Regression Model	**Transformation**	**Explanatory Variables (x)**	**Response Variable (y)**	**R-Squared**	**Adjusted-R**	**P-value**
Simple Linear	None	• COVID-19 Positivity Rate	Depressed Proportion	0.43	0.36	0.0401
Simple Linear	None	• COVID-19 Death Rate		0.73	0.70	0.0016
Multivariate	**None**	**• COVID-19 Positivity Rate** **• COVID-19 Death Rate**		**0.91**	**0.89**	**0.0002**
Multivariate	• Polynomial			0.99	0.94	0.1829
Multivariate	**• Box Cox**			**0.93**	**0.91**	**9.788e-05**
Multivariate	**• Polynomial** **• Box Cox**			**0.99**	**0.93**	**0.1899**

NEXT BEST MODEL

BEST MODEL

Interesting.
P-value has increased.

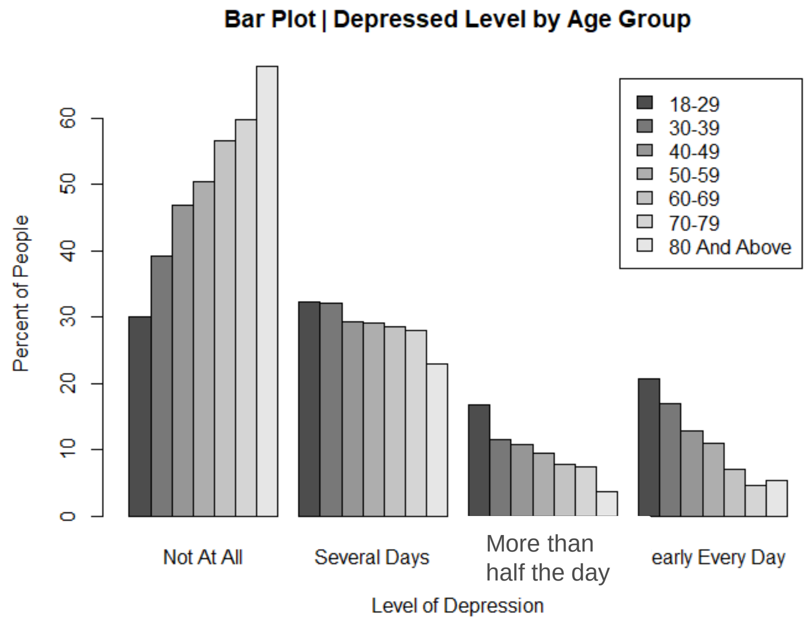
Analyzed the Impact of COVID-19 on Depressed Feelings by creating 2 models and transforming one of the models. Also, **Tested the Relationship between Two Factors** using Two-Table Analysis method and the Chi-Square Test.

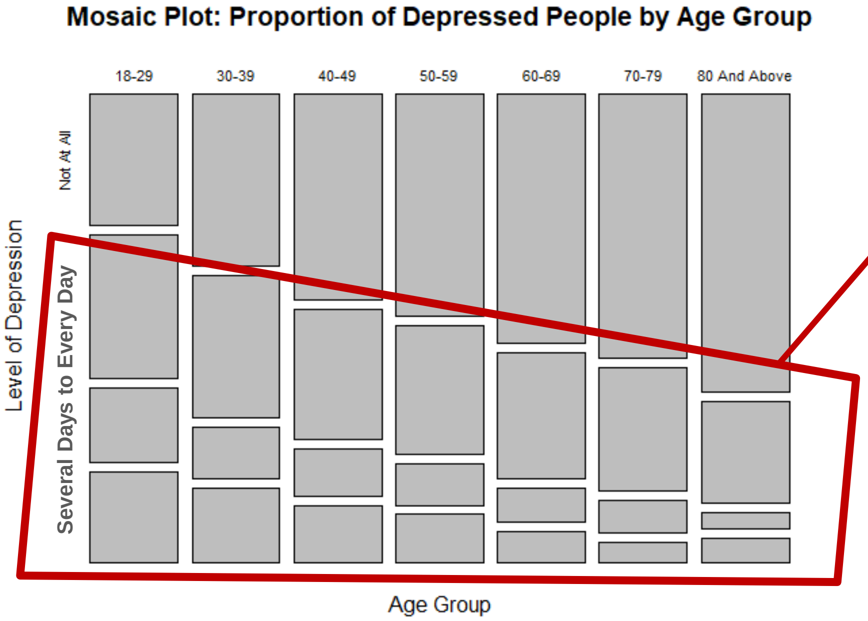
1 Simple liner regression with **only 1 COVID-19 explanatory variable**

2 Multivariate regression with **2 COVID-19 explanatory variables**

3 Two-Way Table analysis between **age and depression**

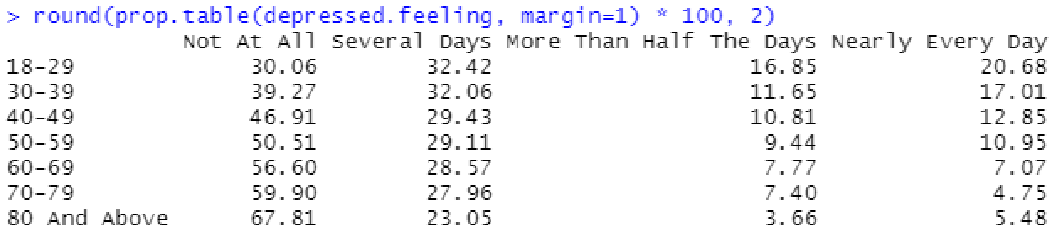
Two-Way Table Analysis using the Conditional Distributions showed that younger populations are more impacted by COVID-19



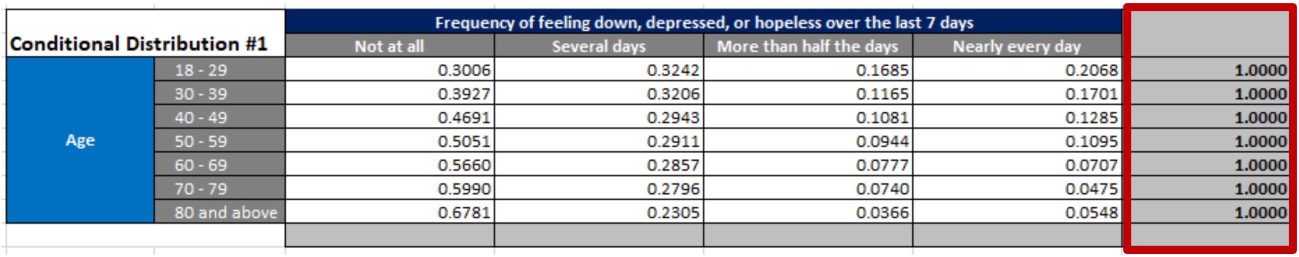


Younger people are feeling more depressed by COVID-19 than older people

PROPORTION (Depressed Level by Age Group)



```
> round(prop.table(depressed.feeling, margin=1) * 100, 2)
              Not At All Several Days More Than Half The Days Nearly Every Day
18-29              30.06        32.42                   16.85            20.68
30-39              39.27        32.06                   11.65            17.01
40-49              46.91        29.43                   10.81            12.85
50-59              50.51        29.11                    9.44            10.95
60-69              56.60        28.57                    7.77             7.07
70-79              59.90        27.96                    7.40             4.75
80 And Above       67.81        23.05                    3.66             5.48
```



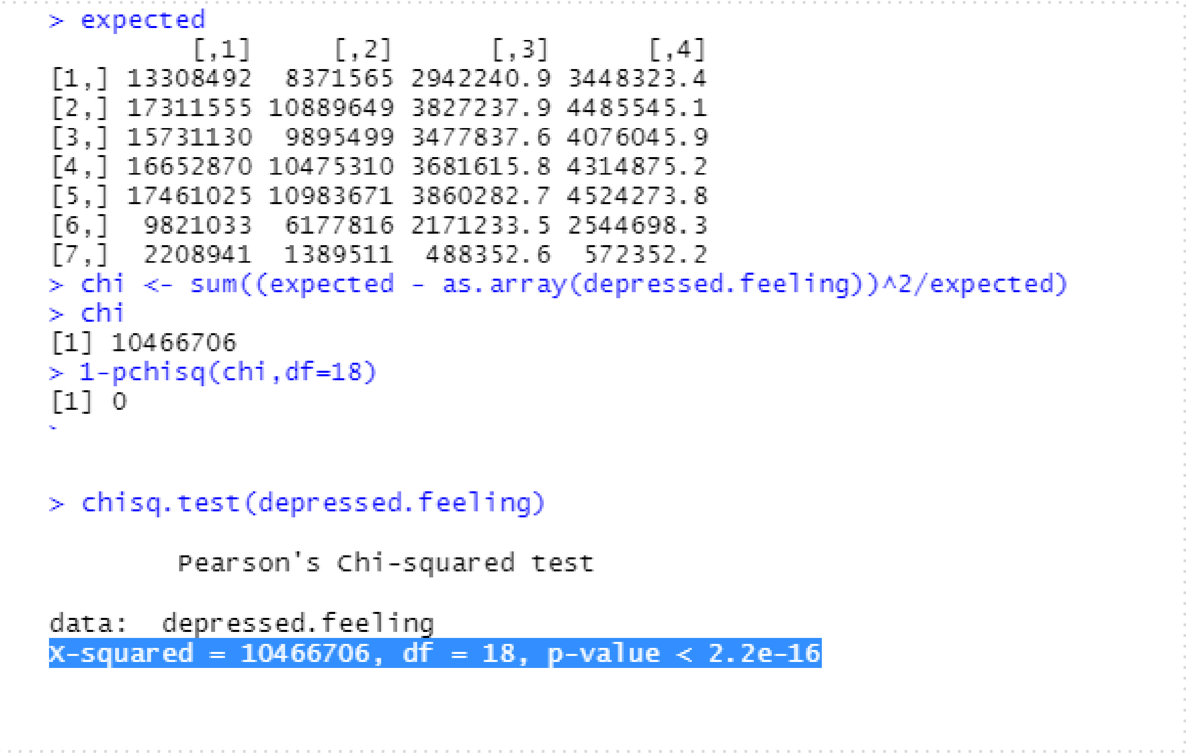
Conditional Distribution #1		Frequency of feeling down, depressed, or hopeless over the last 7 days				
		Not at all	Several days	More than half the days	Nearly every day	
Age	18 - 29	0.3006	0.3242	0.1685	0.2068	1.0000
	30 - 39	0.3927	0.3206	0.1165	0.1701	1.0000
	40 - 49	0.4691	0.2943	0.1081	0.1285	1.0000
	50 - 59	0.5051	0.2911	0.0944	0.1095	1.0000
	60 - 69	0.5660	0.2857	0.0777	0.0707	1.0000
	70 - 79	0.5990	0.2796	0.0740	0.0475	1.0000
	80 and above	0.6781	0.2305	0.0366	0.0548	1.0000

Cross checked the numbers between the two

The Chi-Square Test showed there is a Relationship between Age and Depression

H_0 : There is no relationship between Age and Depression

H_a : Age and Depression are related



Method 1

```
> expected
         [,1]     [,2]      [,3]      [,4]
[1,] 13308492  8371565 2942240.9 3448323.4
[2,] 17311555 10889649 3827237.9 4485545.1
[3,] 15731130  9895499 3477837.6 4076045.9
[4,] 16652870 10475310 3681615.8 4314875.2
[5,] 17461025 10983671 3860282.7 4524273.8
[6,]  9821033  6177816 2171233.5 2544698.3
[7,]  2208941  1389511  488352.6  572352.2
> chi <- sum((expected - as.array(depressed.feeling))^2/expected)
> chi
[1] 10466706
> 1-pchisq(chi,df=18)
[1] 0
```

Method 2

```
> chisq.test(depressed.feeling)

        Pearson's Chi-squared test

data:  depressed.feeling
X-squared = 10466706, df = 18, p-value < 2.2e-16
```

P-value is very very small. Based on the p-value, we rejected the null hypothesis and accepted the alternative hypothesis.

CONCLUSION

- We explored ways to improve regression models and concluded that the **Multivariate Regression Model with the Box-Cox Transformation is the best model with R-Squared of 0.93.**

- **Two-Way Table Analysis** result was statistically significant and supported the alternative hypothesis, "**Age and Depression are related**".

Next Steps:

- **Calculate AIC**, one of the important metrics used to assess the model
- **Test Multicollinearity** between COVID-19 variables (positivity rate and death rate)

THANK YOU