Research Paper

A Correlation Analysis of COVID-19 Cases and Miami Twitter Messages

Giovanny Ruiz

Miami Dade College

School of Science: STEM Research Institute

Mentor: Dr. Sean Mondesire

Spring 2021

A Correlation Analysis of COVID-19 Cases and Miami Twitter Messages

# Introduction

Twitter is a social media platform that allows people to express their emotions in a specific form of blogging. The users mostly interact with each other using messages, also known as “tweets”. When identifying with specific topics, people tend to use hashtags to enable relevant keywords to stand out and create a network. A topic becomes more popular through repeated mentions, when a certain keyword reaches a high number of recurrences it becomes a trend. This trend produces a huge amount of user-generated content that demonstrates people’s behavior. During this unprecedented time, data collected from Twitter is important to analyze how COVID-19 evolves in our society.

This research paper discusses the relevance of Twitter as an indicator of COVID-19 status in Florida, focusing on Miami, an attractive tourist destination in which a high concentration of people increases the chance of infections. The hypothesis is that tweets containing hashtag Miami *(#miami*) have a positive impact on variables such as number of positive cases of COVID-19, occupied beds in hospitals, and new Florida cases. The rationale is that the more people that use hashtag Miami, the more active they are in public and social media, which leads to a higher chance of infection. This research examines the correlation between the number of #miami tweets and the number of new COVID-19 hospitalizations, cases, and positive percentage test results.

# Background

The relationship between two variables is called *correlation*. The most common practice in statistics is using a correlation coefficient that goes from negative one to positive one, to represent how much one variable follow the numerical trend of the other variable [correlation reference]. If one variable increases in the same number of instances as the second one, it is a perfect positive correlation, and the coefficient is close to one, they are directly proportional. However, the perfect negative correlation, represented by the negative one is the complete opposite, they are inversely proportional. When they are not related, the coefficient is zero, and the correlation is neutral.

In this work, several plots are used to graph the results. One such plot is a scatterplot. A *scatterplot* is useful when representing the correlation graphicly. A variable is assigned to each axis and therefore, a pair of coordinates is in the Cartesian plane with points that could be connected in a linear pattern. The strength of the relationship between the variables is an important property of the correlation, represented by the slope of the line, which also represents the coefficient. Equation 1, table 1, and fig 1 describes correlations, strengths and its coefficients.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Equation 1 | |  |  | | --- | --- | | Type of correlation | Coefficient | | Strong Negative Correlation | - 1 to -0.75 | | Weak Negative Correlation | -0.74 to -0.5 | | No Correlation | -0.49 to 0.49 | | Weak Positive Correlation | 0.50 to 0.74 | | Strong Positive Correlation | 0.75 to 1 |   Table 1 |

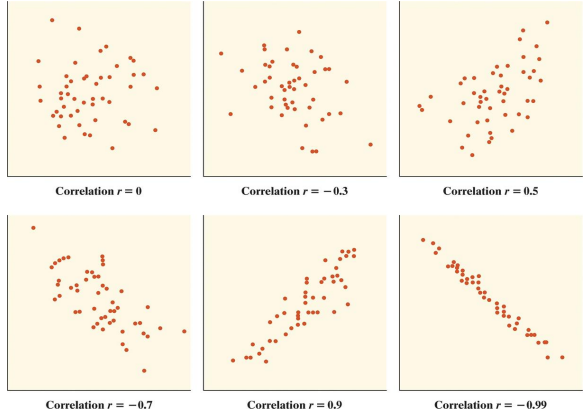


Figure . Correlation coefficients plotted by type

# Methodology

This research examines the correlation between the number social media messages, (#miami specifically) and the number of new COVID-19 hospitalizations, cases, and positive percentage test results. To get the data, one has to request a developer account from Twitter. That process takes some time while Twitter processes the researchers’ Developer application. Afterwards, the developer is grant access to twitter API, which enable the use of programmatic endpoints to retrieve information from the social media. There are limits to retrieve tweet data information per user per day and, it is only possible to get information for the past week. Therefore, to get each study day’s tweets, the developer must make a request for 5,000 tweets per day, then wait 15 minutes to repeat the same request for the next day.

Using Google’s Cloud development environment *Colab*, a Python computer program was created to scrape data from Twitter and appended each day to a list of tweets, that information is display in fig. 2. After scraping the tweets, the next step was aggregate the tweets to count total tweets per day in list 2. A third step was using the dates to merge the twitter data with COVID-19 data found on the COVID tracker from FAU’s public dashboard (see list 3).

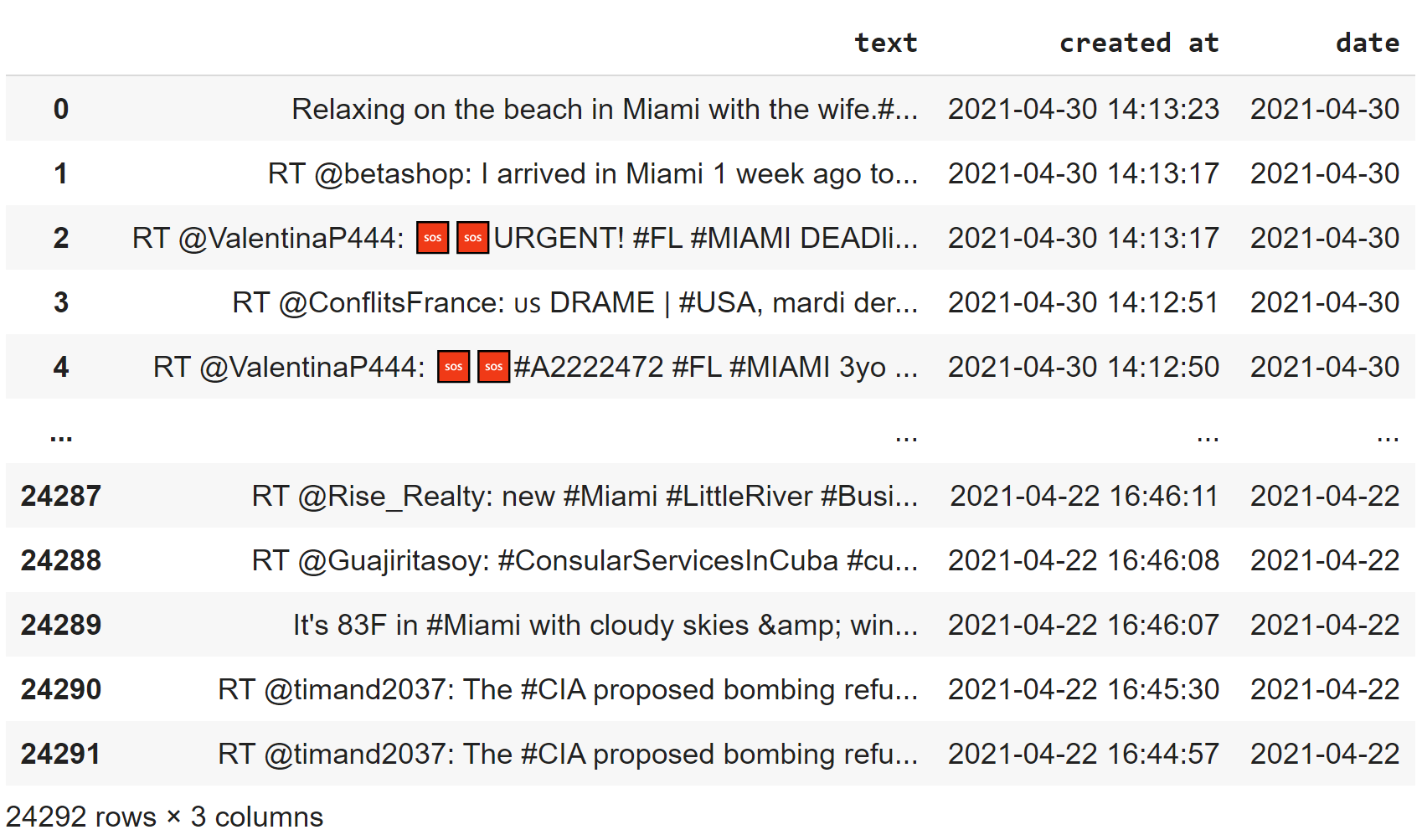


Figure . Tweet data from April 22-30, 2021

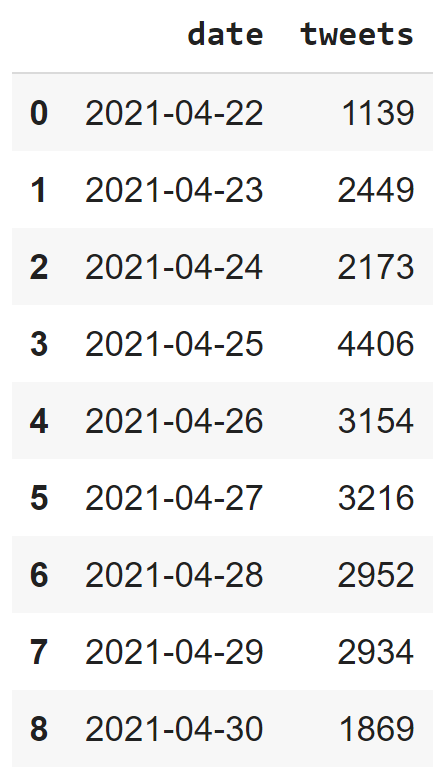


Figure . Aggregated Tweets per Day of #miami

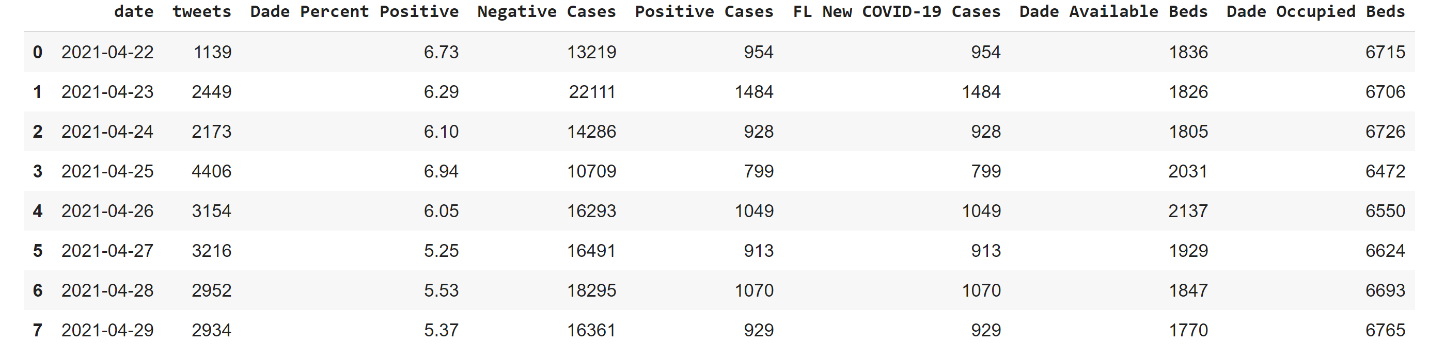
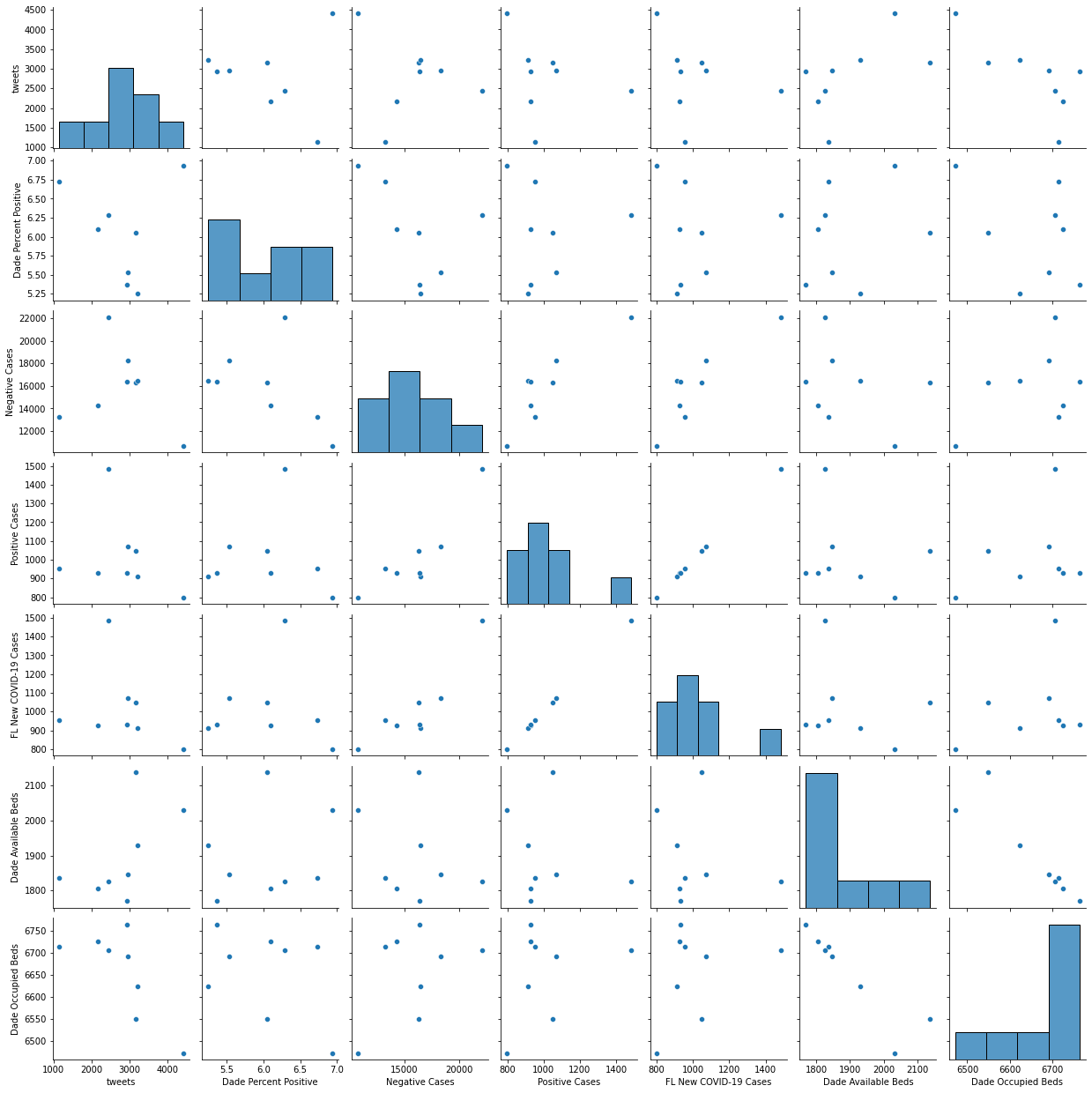


Figure . COVID-19 Statistics for April 22-30, 2021

Figure 5. Pair plot of aggregated tweets and COVID-19 statistics.

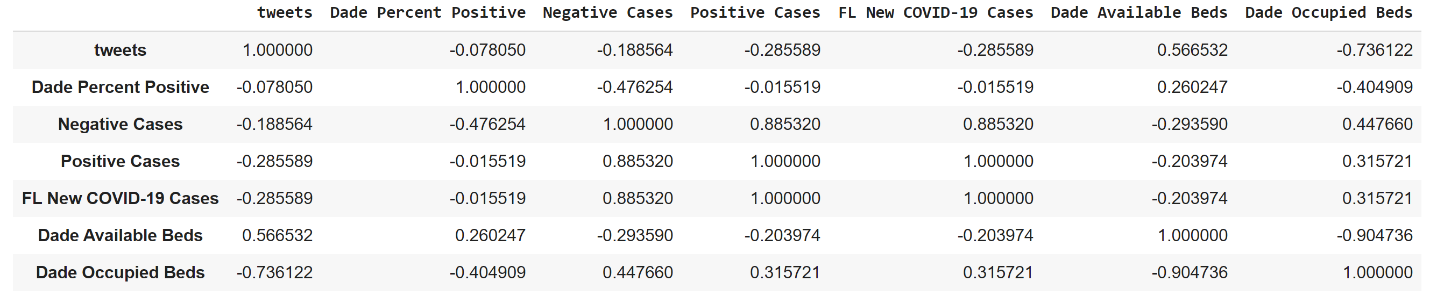
The pair plot (fig. 5) displays all the data collected graphically. As pointed previously, the goal is comparing the tweets with all the other variables, thus, an easy way to look at it is just focusing on the first column (tweets). When looking closely, there is difficult to find any kind of correlation if any between the tweets and all the other variables, the histogram is the plot that correlates tweets to itself, the higher the bar, more tweets using #Miami where found.

# Results

We have individual plots for the comparison between the variables (individual plots 1,2,3,4). The first one shows a line where the number of cases is compared to the number of tweets per day. The second one, the Florida positive cases and the number of tweets. The third one, the positive cases in Miami-Dade County against the tweets. The fourth one, the number of beds occupied in hospitals in Miami-Dade County against tweets. It is not possible to determine any kind of correlation between those variables graphically.

The correlation table (table 1) is the ultimate resource find out if there is any correlation between the variables. The diagonal of the table is clearly one simply because there is a strong positive correlation between any variable and itself. There is no significant correlation among most of the variables except for two of them: first, a weak correlation between tweets and available beds (correlation coefficient: 0.56); second, a stronger negative correlation between tweets and occupied beds in Miami-Dade. The correlations are unexpected, in fact are opposite to what the hypothesis expressed, the cause of that kind of correlation between number of tweets and the bed variables could vary: first, the mention of the #Miami referring to activities that do not involve unsafe behavior by users, such as vaccination; second, the time frame is not as popular as other dates that could show more infections, for example spring break; third, the success of vaccination in the country makes the beds less likely to be occupied by a large amount of individuals.

Table . Correlation coefficients of tweets and COVID-19 variables.



The limited 7-day time frame was an unexpected obstacle to tackle, the ideal time frame for this kind of study is at least three months, in order to have enough data related to COVID-19 variables and tweets. The limitation using the twitter API where data extracted was limited to one week in the past makes difficult to retrieve all the information needed for the research.

It is important to have in mind that correlation do not mean causation. Correlation measures the strength and direction of the relation between two variables, but causation refers to a deterministic way to verify the effect that one variable produces on another one. Correlation describes the link shared between two variables, but does not explain the reason of that link, because there could be data that is not shown but has been unaccounted for in the study in causing that effect.

This study will be extended in the future by including more data in a larger time frame, such effort may help predicting when and where future outbreaks are likely to happen, all based on social media analysis. Because social media is a nest of a large amount of user-generated data, the examination and usage of such platforms may help prevent problems and shed some light on people’s behavior.

# Conclusion

In conclusion, this study found that the hospitalization variables have strong correlations with number of tweets, but number of cases and positivity rate of people tested positive for COVID-19 do not. Because correlation does not imply causation, this study is worth further investigation to determine the correlation of the variables using a larger data frame. With the extension, social media may prove to be a useful tool to alert public health officials and prevent future outbreaks.

# Personal Accomplishments

The usage of the coding language Python and some of its libraries was enjoyable, many of them were new for me and the fact that I could extend the use of social media beyond leisure made the idea of the usage of social media platforms wider. Social media could be a distressful tool for many, but is also an important source of raw material for data analysis.

# References

[1] Mindrila D., and Balentyne P., Scatterplots and Correlation. <https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf>. Date accessed: April 30, 2021.

[2] Center for Systems Science and Engineering (CSSE), Johns Hopkins University (JHU). COVID-19 Dashboard <https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6>. Date accessed: April 30, 2021.

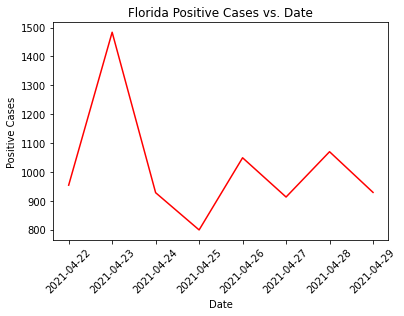
[3] Cagliero L., Fiori A. (2012) Analyzing Twitter User Behaviors and Topic Trends by Exploiting Dynamic Rules. In: Cao L., Yu P. (eds) Behavior Computing. Springer, London. <https://doi.org/10.1007/978-1-4471-2969-1_17>

[4] Pandas, Pandas documentation. <https://pandas.pydata.org/docs/>. Date accessed: April 30, 2021.

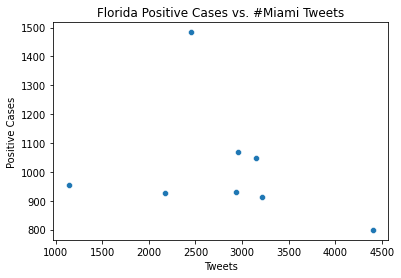
[5] Parker E., COVID-19 Tracker. <https://shiny.rstudio.com/gallery/covid19-tracker.html>. Date accessed: April 30, 2021.

[6] Twitter, Use Cases, Tutorials, & Documentation | Twitter Developer. <https://developer.twitter.com/en>. Date accessed: April 30, 2021.

[7] Tweepy, Tweepy library usage. <https://www.tweepy.org/>. Date accessed: April 30, 2021.



Individual plot 2



Individual plot 3



Individual plot 4

