ITEC 621 Project Deliverable 2

James Lindahl, Minh Nguyen, Shaun and Reina Villanueva-Unger

April 9th, 2020

PREPARING THE WORLD WITH TWEETS

**The Business Case**

We are currently in the middle of the CORVID-19 pandemic that has affected, as of April 7, 2020, more than 200 countries and territories. More than 1,000,000 people have been diagnosed, more than 50,000 have died as a result, entire countries are on forced quarantine and, unfortunately, there seems to be no foreseeable end. News reports were few at first when the first case was reported on December 31, 2019, in the Wuhan Province in China. However, by January 30, 2020 it had grown to nearly 8000, with 180 deaths. By February 29, that number had grown ten-fold to 85403. By March 15, that number was 153,517. By March 30, that number was 750,890, globally. The long-term effects are still terrifyingly uncertain, with millions laid off, supply chains irreparably disrupted, and major corporations that have existed in twice our lifetimes are shuttering their doors, and that’s just in the United States. It is inevitable for tragedy to strike, but large-scale events such as CORVID-19 require a swift and efficient response.

**The Business Question**

Social media is an extremely useful tool, connecting people and developing global relationships, both personal and professional, and offers a wider perspective of the world. But the current state of the world forces the question: are we using it to its greatest benefit? It is undeniably one of the most powerful tools we have now available to communicate the spread of the disease, news reports of relief efforts, and informing how those not infected can help. But understanding the data is key. Cambridge Analytica is probably a fond, if distant, memory, but while they used their power nefariously, we could now see this as an opportunity for good, and more importantly, helping. We are faced with the same question that Cambridge Analytica asked themselves, only turned on its head: how can we use social media to help people now that we *all* need it the most?

**The Analytics Question**

The purpose of this exercise is to use the most common use of communication for instant news in the US, Twitter. Twitter has 330 million global active users who generate an astounding 6,000 tweets *per second*. We are hoping to see how best to use Twitter to calculate the human and infrastructural damage, and if it’s possible to analyze the tweets to better understand what influences those who are actively spreading word and contributing to the effort. With that information we hope to form a better understanding of people, what they will most likely respond to, positively and negatively, and, hopefully, in the event of any future crisis how best to use social media to promote collective relief efforts. If disaster strikes, we should have at our disposal those tools that will most swiftly meet the needs of response. Assuming that mobile users comprise Twitter’s main user base response time could be near-instantaneous. The question, then, is how to best utilize Twitter data to see what words or phrases will generate the most retweets and thereby spreading awareness, particularly in areas particularly affected. In order to gain a better understanding we are analyzing Twitter feeds, gauging response through text mining in the COVID-19 pandemic.

**The Dataset**

Tweets were scraped using Twitter’s proprietary API, with data scraped between March 1 and March 14, 2020, amounting to 3,000 tweets per day, 42,000 observations total. The data scraped include: date, name, source, text, favorites, and retweets. To better assess the effect of each tweet or retweet total, we selected commonly used keywords and created variables that tally their number of occurrences with the expectation it would explain whether or not they generated response, and which held the most influence.

COVID-19 data was cloned from John Hopkins’s GitHub repository. The two datasets utilized were the *time\_series\_19\_covid-Confirmed* and *time\_series\_19\_covid-Deaths*. From this data we collected the number of cases that occurred within countries reported by the World Health Organization. Because China, Italy, and the US are considered epicenters of the crisis we merged the datasets by these countries, with all other countries tallied under “other,” adding eight new variables. These data were set against the Twitter data to see if there was correlation.

**Defining the Predictors**

The data was chosen to determine if there were predictors that could explain how emergencies, social media – in this instance Twitter – and social reaction correlate. Specifically, the selected predictors capture different text strains from Twitter. From that data we determined how many people were tweeting about the COVID-19 crisis. From there, we analyzed the frequency at which certain keywords would appear in trending tweets, the number of responses, and what compels users to reply, retweet, or both; with that data it might be possible to see if social media can accurately track how the virus is spreading.

The first predictor captures the source of the post, tracking the time and location a tweet is sent, ensuring the data stays within our specified locations. This data tracks user postings on the same subject. The second predictor gathers the metrics of the tweet, e.g. time of retweet, text length, content, and whether or not it was favorited. These observations form an understanding of how post content generates response and activity over certain topics, words, or phrases, in turn providing a better understanding of user reaction. The third predictor highlights the previous by focusing on certain keywords contained within the tweet to understand and predict patterns of behavior during a crisis; the isvirus, ispandemic, iscovid, isdeath, iscdc, isquarantine variables help to define which issues generate a response. The final set of predictors demonstrate how many cases of COVID-19 are confirmed and by whom, the corresponding number of deaths reported in that locality, if those deaths have been confirmed, and gauge reaction and response on social media. The four predictors form a gestalt of understanding motivation and how to harness that energy into positive collective response.

**Brief Conclusion Statements and What Models we would have potentially used**

Our initial hypothesis was that generating tweets and retweets will help emphasize the seriousness of a crisis, and are more likely to reach decision makers and government agencies. To develop a predictive model, we began with a GLM and, because the response value (i.e. retweets) was not continuous it was truncated at 0. Perhaps the most notable interaction was that once something has been retweeted – either a native tweet, a retweet, or a published article, report, or video – its retweet counts increases by 6.546 . Significant positive correlations to retweets include using words that relate to current criss like virus (+1.358), world health organization (+2.319), ban (+1.502), coronavirus (+5.704), and China (+4.201), as well as reports of deaths in both China (+4.929) and Italy (+3.917). Conversely, and in most cases counterintuitively, variables “racist” or “racism”, “NIH”, “COVID-19”, or “disease” saw negative correlation to the response variable. Furthermore, the longer the tweet, the less likely it is to be retweeted. And most perplexingly, it appears that the model indicates that as reports of confirmed cases in the U.S. increases the number of retweets decrease (-5.654).[[1]](#footnote-0) From the regression tree, the data suggests that during the relevant period, retweets were the highest where the confirmed cases in China is greater than 80,938.5, and when mentioning Corona Virus, between 132.5-133.5 characters, but *did not* mention COVID-19.[[2]](#footnote-1)

From a preliminary analysis the data needs to be pre-processed further, with potential alternative models. When performing a visual inspection of the residual plots for our two OLS models, we noticed the errors seemed to grow systemically along the outcome variable in the OLS predicting retweets. The Bruesch-Pagan test confirmed that heteroskedasticity was present based on its significant p-value. To correct for heteroskedasticity, we created a WLS model to predict retweets. Next, we ran multicollinearity tests on both models, and found that the CI for both was severe. Upon inspecting the VIFs, we found that for variables ch.confirmed, it.confirmed, ot.confirmed, us.confirmed, ch.deaths, it.deaths, ot.deaths, and us.deaths, the variance was too high. In order to correct for this, we will have to consider dropping these variables or trying other approaches such as PCR, PLS, Ridge, or LASSO.

However, despite this we were able to determine that there is a possible threshold of awareness that should be explored further; note that in the boxplot for us.deaths the upper quartile jumps significantly after 40.[[3]](#footnote-2) As well, the data did confirm our assumption that most activity would come from mobile users, most noticeably in the various boxplots differentiating source.[[4]](#footnote-3) In sum, while the data requires further examination, it does offer promise that harnessing the technology and social media in crisis response can be of tremendous benefit.

**Appendix 1**

**Dataset Description**

**Twitter Data**

* date – date and time of post
* name – name/Twitter handle/username
* sources – device used (converted to factor with 4 levels – iPad, iPhone, Android and Web)
* isretweeted – retweet or not (TRUE = retweet, FALSE = not a retweet)
* text – text/content
* length – character count
* favorites – the number of favorites received
* retweets – the number of retweets

Below are the keywords selected for text mining and analysis:

* ishealth – # of occurrences for “health”
* ispandemic – # of occurrences for “pandemic”
* isvirus – # of occurrences for “virus”
* isemergency – # of occurrences for “emergency”
* isdeaths – # of occurrences for “dead” and “deaths”
* iswho – # of occurrences for “who” and “world health organization”
* iscdc – # of occurrences for “cdc” and “centers for disease control”
* isnih – # of occurrences for “nih” and “national institutes of health”
* isdisease – # of occurrences for “disease”
* isquarantine – # of occurrences for “quarantine”
* isrecover – # of occurrences for “recover”
* isban – # of occurrences for “ban”
* iscoronavirus – # of occurrences for “coronavirus”
* iscovid19 – # of occurrences for “covid19”
* iswash – # of occurrences for “wash”
* isracist – # of occurrences for “racist” and “racism”
* isasian – # of occurrences for “asian”
* ischinese – # of occurrences for “chinese” and “china”
* isinfectious – # of occurrences for “infectious” and “infections”

**John Hopkins’s GitHub [[5]](#footnote-4)**

COVID-19 data was cloned from John Hopkins’s GitHub repository. The two datasets used are *time\_series\_19\_covid-Confirmed*, and *time\_series\_19\_covid-Deaths*. These datasets include: Province/State, Country/Region, Latitude, Longitude, and collects data from between 1/22/2020 - 3/16/2020.

Because China, Italy and the US are considered the epicenters of the crisis, they are specified in the two time series datasets, with all other countries grouped as “other”.

* ch.confirmed – total # of confirmed cases in China
* it. confirmed – total # of confirmed cases in Italy
* ot.confirmed – total # of confirmed cases in other countries
* us.confirmed – total # of confirmed cases in United States
* ch.deaths – total # of deaths in China
* it. deaths – total # of deaths in Italy
* ot.deaths – total # of deaths cases in other countries
* us.deaths – total # of deaths in United States

The repository data sources include:

* World Health Organization (WHO)
* DXY.cn. Pneumonia. 2020
* BNO News
* National Health Commission of the People’s Republic of China (NHC)
* China CDC (CCDC)
* Hong Kong Department of Health
* Macau Government
* Taiwan CDC
* U.S. CDC
* Government of Canada
* Australia Government Department of Health
* European Centre for Disease Prevention and Control (ECDC)
* Ministry of Health Singapore (MOH)
* Italy Ministry of Health

**Appendix 2**

**Generalized Linear Model (GLM)**

The following predictors were significant in the first model:

* isretweetedTrue - On average, holding all else constant, retweeted tweets (TRUE) increases retweet total by 6.546 when compared to non-retweeted tweets (FALSE = ref).
* length - On average, holding all else constant, an increase in tweet length decreases retweets total by -5.138 (longer tweets generate less retweet and reachability).
* isvirus - On average, holding all else constant, an increase in usage of the word “virus” increases the retweet total by 1.358.
* isemergency - On average, holding all else constant, an increase in usage of the word “emergency” decreases the retweet total by -4.448.
* isdeaths - On average, holding all else constant, an increase in usage of the word “dead” or “deaths”  decreases the retweet total by -3.387.
* iswho - On average, holding all else constant, an increase in usage of the word “WHO” or “World Health Organization” increases the retweet total by 2.319.
* iscdc - On average, holding all else constant, an increase in usage of the word “CDC” or “Centers for Disease Control” decreases the retweet total by -3.676.
* isnih - On average, holding all else constant, an increase in usage of the word “NIH” or “National Institutes of Health” decreases the retweet total by -6.797.
* isdisease - On average, holding all else constant, an increase in usage of the word “disease” decreases the retweet total by -4.519.
* isquarantine - On average, holding all else constant, an increase in usage of the word “quarantine” decreases the retweet total by -3.845.
* isban - On average, holding all else constant, an increase in usage of the word “ban” increases the retweet total by 1.502.
* iscoronavirus - On average, holding all else constant, an increase in usage of the word “coronavirus” increases the retweet total by 5.704.
* iscovid19 - On average, holding all else constant, an increase in usage of the word “covid19” decreases the retweet total by -4.556.
* iswash - On average, holding all else constant, an increase in usage of the word “quarantine” decreases the retweet total by -2.697.
* isracist - On average, holding all else constant, an increase in usage of the word “racist” or “racism” decreases the retweet total by -8.455.
* ischinese - On average, holding all else constant, an increase in usage of the word “chinese” or “china” increases the retweet total by 4.201.
* ch.confirmed - On average, holding all else constant, when confirmed cases in China increases by 1, retweet total decreases by -1.806.
* it.confirmed - On average, holding all else constant, when confirmed cases in Italy increases by 1, retweet total decreases by -1.475.
* ot.confirmed - On average, holding all else constant, when confirmed cases in others (all other countries not China, Italy, or US) increases by 1, retweet total increases by 1.124.
* us.confirmed - On average, holding all else constant, when confirmed cases in US increases by 1, retweet total decreases by -5.654.
* ch.deaths - On average, holding all else constant, when the number of deaths in China increases by 1, retweet total increases by 4.929.
* it.deaths - On average, holding all else constant, when the number of deaths in Italy increases by 1, retweet total increases by 3.917.
* ot.deaths - On average, holding all else constant, when the number of deaths in others (all other countries not China, Italy, or US) increases by 1, retweet total increases by -6.725.

In addition, the following predictors were marginally significant; although not significant as their p-values > .05, there is still an effect in this first mode worth mentioning:

* ispandemic (p-value = 0.0952) - On average, holding all else constant, an increase in usage of the word “ispandemic” decreases the retweet total by -1.525.
* isrecover (p-value = 0.0933) -  On average, holding all else constant, an increase in usage of the word “recover” decreases the retweet total by -3.071.

**GLM Summary Output**

Call:

glm(formula = retweets ~ isretweeted + length + favorites + ishealth +

    ispandemic + isvirus + isemergency + isdeaths + iswho + iscdc +

    isnih + isdisease + isquarantine + isrecover + isban + iscoronavirus +

    iscovid19 + iswash + isracist + isasian + ischinese + isinfectious +

    ch.confirmed + it.confirmed + ot.confirmed + us.confirmed +

    ch.deaths + it.deaths + ot.deaths + us.deaths, data = covid19current)

Deviance Residuals:

    Min       1Q   Median       3Q      Max

-139722    -7264    -2056     2839   187104

Coefficients:

                  Estimate Std. Error t value Pr(>|t|)

(Intercept)      1.299e+07  5.800e+05  22.387  < 2e-16 \*\*\*

isretweetedTRUE  6.546e+03  3.160e+02  20.713  < 2e-16 \*\*\*

length          -5.138e+01  6.945e+00  -7.398 1.40e-13 \*\*\*

favorites        8.145e-01  4.204e+00   0.194 0.846379

ishealth        -4.176e+02  5.016e+02  -0.833 0.405118

ispandemic      -1.525e+03  9.141e+02  -1.669 0.095180 .

isvirus          1.358e+03  3.725e+02   3.646 0.000267 \*\*\*

isemergency     -4.448e+03  1.301e+03  -3.419 0.000630 \*\*\*

isdeaths        -3.387e+03  8.674e+02  -3.905 9.46e-05 \*\*\*

iswho            2.319e+04  5.130e+02  45.218  < 2e-16 \*\*\*

iscdc           -3.676e+03  1.008e+03  -3.645 0.000268 \*\*\*

isnih           -6.797e+03  2.430e+03  -2.797 0.005167 \*\*

isdisease       -4.519e+03  1.164e+03  -3.883 0.000103 \*\*\*

isquarantine    -3.845e+03  8.403e+02  -4.575 4.77e-06 \*\*\*

isrecover       -3.071e+03  1.830e+03  -1.678 0.093328 .

isban            1.502e+03  7.469e+02   2.011 0.044285 \*

iscoronavirus    5.704e+03  4.150e+02  13.744  < 2e-16 \*\*\*

iscovid19       -4.556e+03  2.076e+02 -21.950  < 2e-16 \*\*\*

iswash          -2.697e+03  8.856e+02  -3.046 0.002322 \*\*

isracist        -8.455e+03  3.713e+03  -2.277 0.022795 \*

isasian         -3.421e+03  2.978e+03  -1.149 0.250661

ischinese        4.201e+03  6.736e+02   6.237 4.51e-10 \*\*\*

isinfectious    -2.736e+03  2.172e+03  -1.259 0.207878

ch.confirmed    -1.806e+02  8.150e+00 -22.157  < 2e-16 \*\*\*

it.confirmed    -1.475e+01  6.582e-01 -22.414  < 2e-16 \*\*\*

ot.confirmed     1.124e+01  3.724e-01  30.190  < 2e-16 \*\*\*

us.confirmed    -5.654e+01  4.096e+00 -13.802  < 2e-16 \*\*\*

ch.deaths        4.929e+02  2.713e+01  18.167  < 2e-16 \*\*\*

it.deaths        3.917e+01  5.369e+00   7.296 3.02e-13 \*\*\*

ot.deaths       -6.725e+01  1.311e+01  -5.131 2.89e-07 \*\*\*

us.deaths       -2.440e+02  1.523e+02  -1.603 0.109003

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 458727180)

    Null deviance: 2.3711e+13  on 41999  degrees of freedom

Residual deviance: 1.9252e+13  on 41969  degrees of freedom

AIC: 956870

Number of Fisher Scoring iterations: 2

**Appendix 3**

**Regression Tree**

Regression tree:

tree(formula = retweets ~ isretweeted + length + favorites +

    ishealth + ispandemic + isvirus + isemergency + isdeaths +

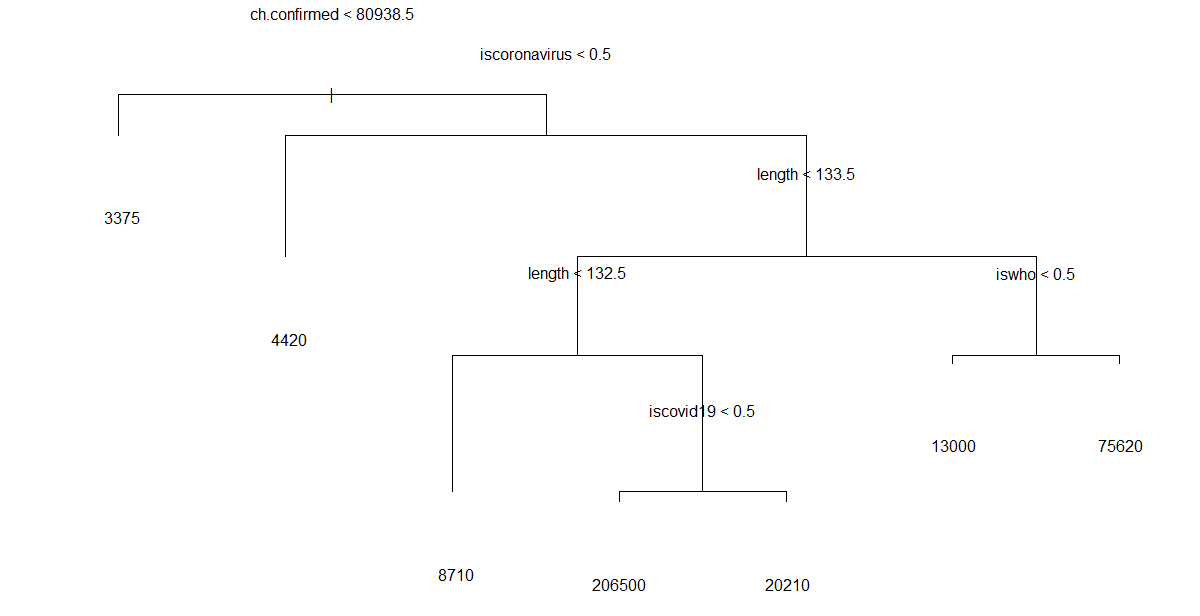
    iswho + iscdc + isnih + isdisease + isquarantine + isrecover +

    isban + iscoronavirus + iscovid19 + iswash + isracist + isasian +

    ischinese + isinfectious + ch.confirmed + it.confirmed +

    ot.confirmed + us.confirmed + ch.deaths + it.deaths + ot.deaths +

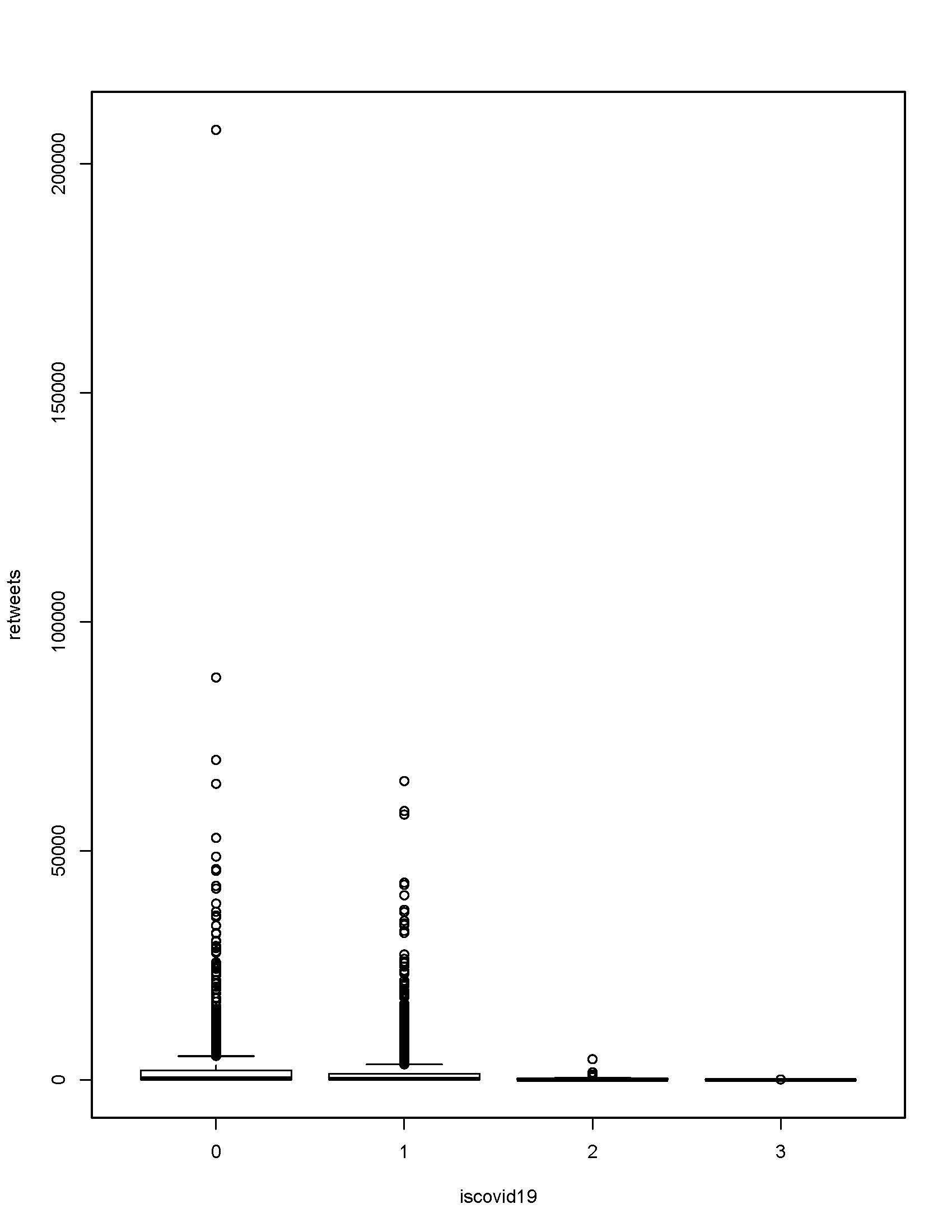
    us.deaths, data = covid19current)



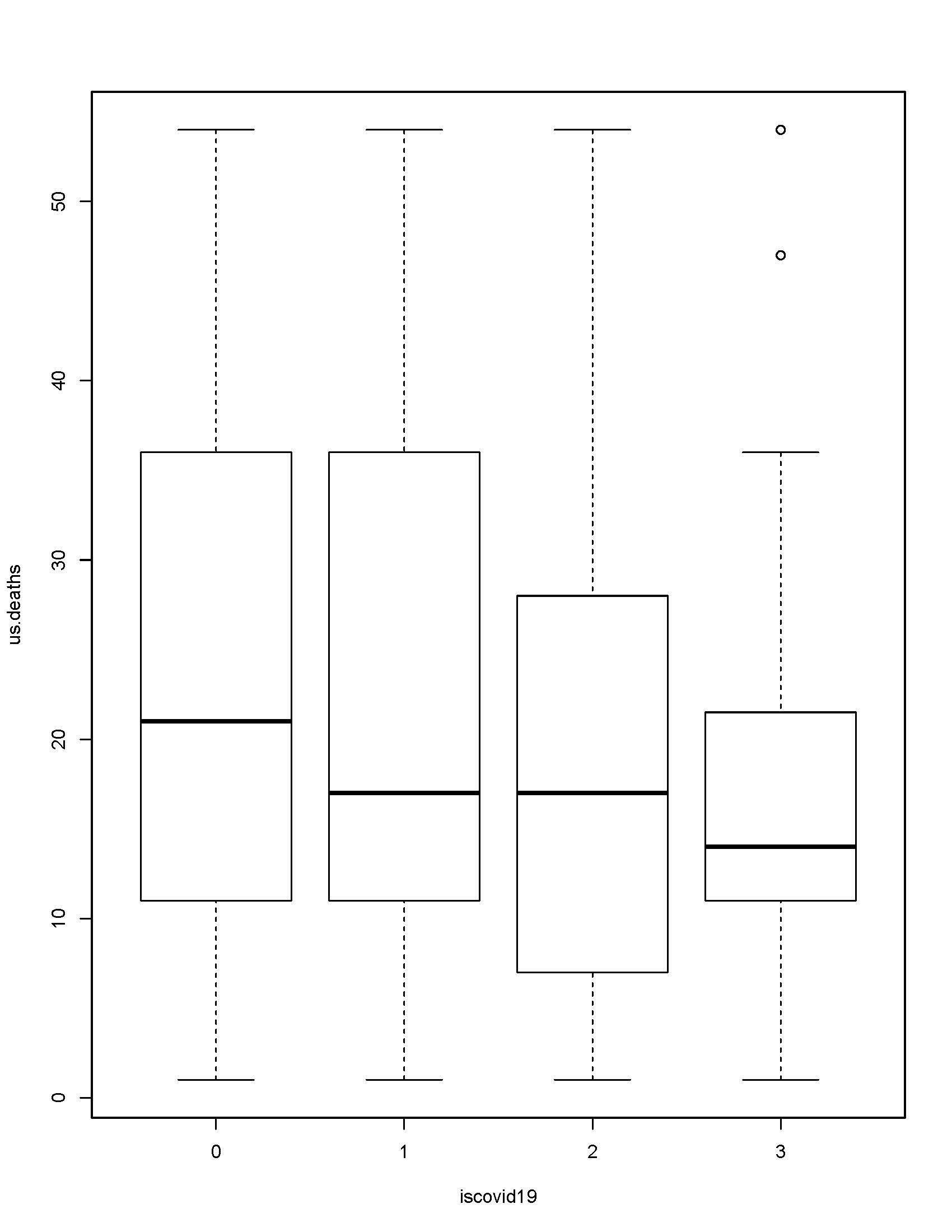
**Appendix 4**

**Boxplots**

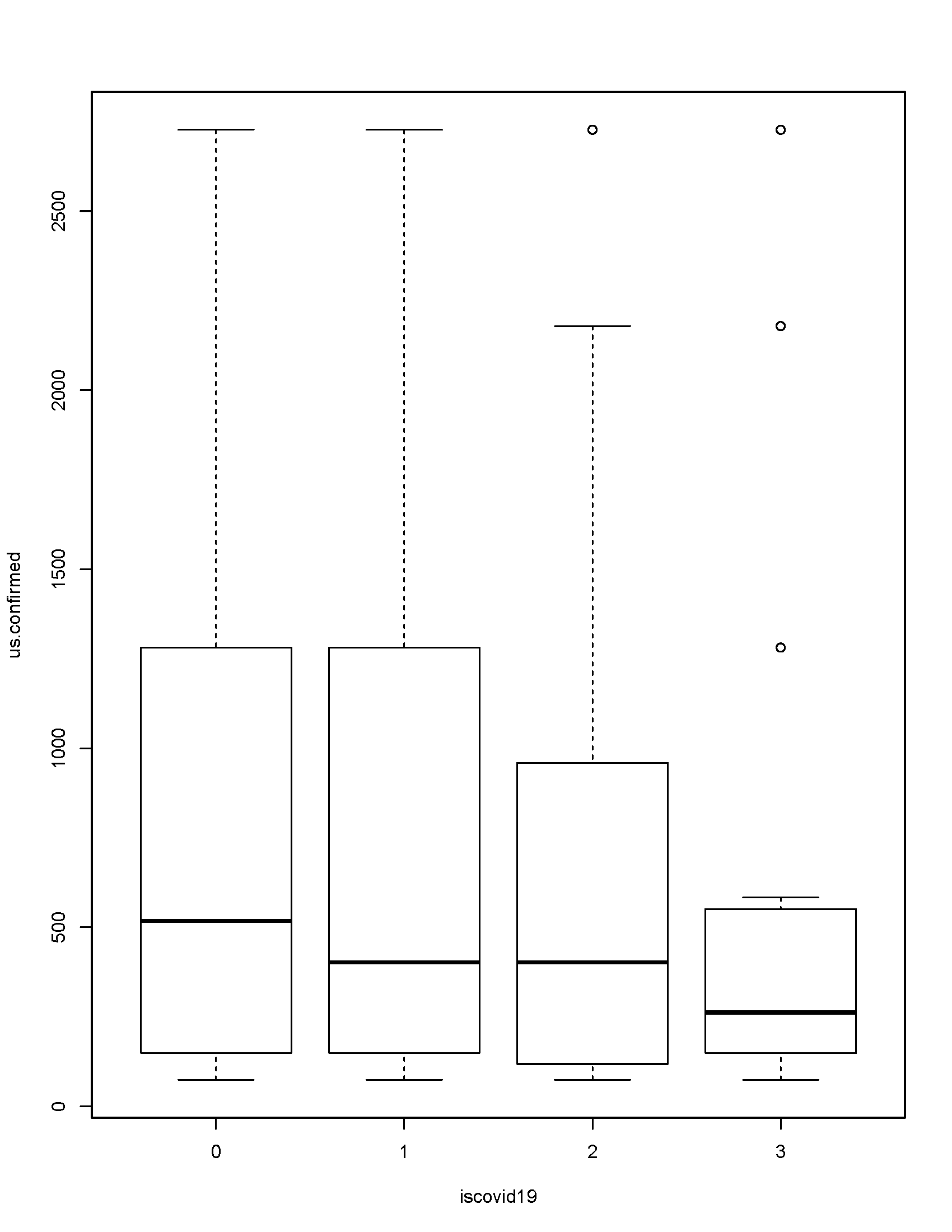
*Figure 1: COVID-19 Retweets by Source*



*Figure 2: COVID-19 and US Deaths by Source*



*Figure 3: COVID-19 and U.S. Confirmed Cases by Source*



1. For more detailed analysis of the GLM *see* Appendix 2 [↑](#footnote-ref-0)
2. Appendix 3 [↑](#footnote-ref-1)
3. Appendix 4, *Figure 1* [↑](#footnote-ref-2)
4. Appendix 4, *Figure 2* [↑](#footnote-ref-3)
5. GitHub, & Johns Hopkins CCSE. (2020, April 24). 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE. Retrieved March 17, 2020, from https://github.com/CSSEGISandData/COVID-19 [↑](#footnote-ref-4)