**Data Analytics Capstone Topic Approval Form**

**Student Name:** Peter Harmazinski

**Student ID: #**733267

**Capstone Project Name:** Statistical Analysis of COVID-19 and Twitter Sentiment

**Project Topic**: Statistical Analysis on COVID -19 Metrics and Tweet Sentiment

**Research Question:** Do COVID -19 metrics affect the sentiment of COVID -19 tweets?

**Hypotheses**: H0 : COVID -19 metrics have no effect on tweet sentiment. H1: COVID -19 metrics do influence tweet sentiment.

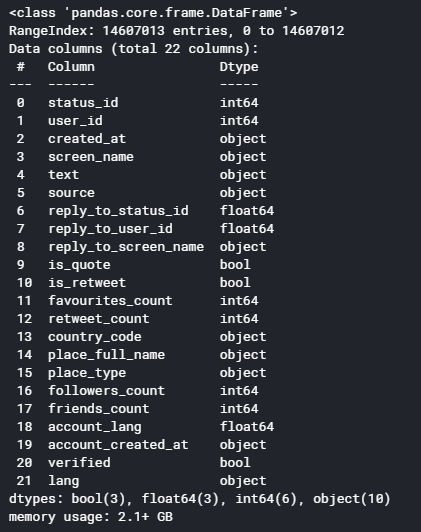
**Context:**

COVID-19 is ravaging human life. People are dying, economies are struggling, and the world is changing. 14 years ago, a social media platform, Twitter, was created that enables people around the globe to express themselves. This study examines if COVID-19 metrics have an impact on the sentiment of related tweets. Since the data is not normally distributed, nonparametric methods will be used for the analysis. There has also been related work to predict COVID-19 outbreaks, based on Twiiter data (Jahanbin & Rahmanian, 2020).

**Data:** Datasets are retrieved from Kaggle, but the data originates from Twitter and the COVID-19 Tracking Project.

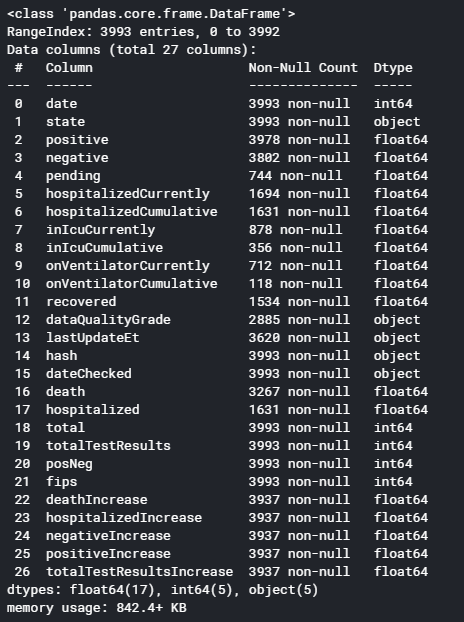
Twitter:

* <https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april>
* <https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april>



COVID-19 Tracking Project:

* <https://www.kaggle.com/sudalairajkumar/covid19-in-usa>

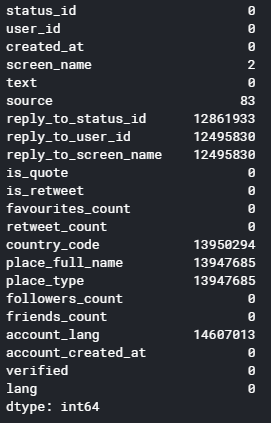
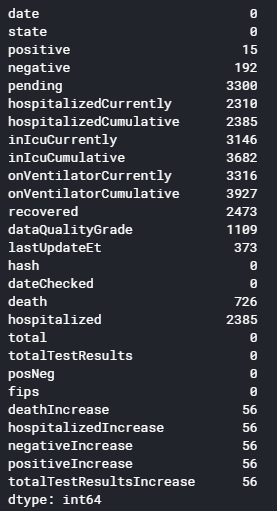


The Twitter dataset has 22 features and 14,607,013 observations. 18 of the features are qualitative and 4 are quantitative. The COVID-19 dataset has 27 features and 3,993 observations. 6 features are qualitative and 21 are quantitative. There are several limitations to the dataset. First, the Twitter data timespan is approximately a month. Next, sentiment analysis is not yet a precise science. There are also linguistic factors to consider. Second, COVID-19 tracking has not been precise due to the nature of the disease.

Despite these limitations, this study approaches an important topic: mental health during a pandemic. The COVID-19 outbreak can lead to increased stress and mental health problems Coping with Stress”, 2020). Since Twitter is a popular outlet for expression, it may facilitate the expression of struggles or positive messages to encourage others.

**Data Gathering:**

Twitter null values: COVID-19 null values:

Both datasets have a significant number of null values. Most of the Twitter features are irrelevant to the sentiment analysis. However, country\_code is not irrelevant and it has 12,950,294 null values. Since this analysis is location-specific, those observations will be excluded. For the COVID-19 dataset, any null values will be assumed to have a value of 0 and will be imputed with the same (LeDoux, 2019).

**Data Analytics Tools and Techniques**:

Since the data is nonparametric, nonparametric methods will be used. Kruskal-Wallis will be used to identify group differences. K-means cluster analysis will be used to identify similar groups. Interaction terms will be identified and dimensionality will be reduced through PCA. There are 3 potential classes: negative, neutral, and positive sentiment. Since this is a multiclass classification problem, I will compare 3 types of models: Random Forest, K-nearest Neighbors and Multinomial Logistic Regression (“Multiclass algorithms”, n.d.). Each has its respective benefits, therefore all 3 will be compared to determine the feasibility of predicting Twitter sentiment from COVID-19 metrics. Models will be compared with a goal of obtaining the highest accuracy. The analysis and visualizations will be completed with Python.

**Justification of Tools/Techniques:**

Python will be used for data extraction, prep, and analysis. Python is a free, open source programming language with a robust ecosystem of libraries and support. It’s also easy to understand and scalable if a production deployment is required (DataCamp Team, 2020).

**Project Outcomes**: Insight into whether COVID-19 metrics have an impact on related tweet sentiment.

**Projected Project End Date**: **6/19/2020**

**Reference Sources**:

*Coping with Stress* (2020, April 20). Centers for Disease Control and Prevention. <https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/managing-stress-anxiety.html>.

DataCamp Team (2020, January 9). *Choosing Python or R for Data Analysis? An Infographic*. DataCamp. <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>.

Jahanbin, K. & Rahmanian, V. (2020, March). *Using twitter and web news mining to predict COVID-19 outbreak*. ResearchGate. <https://www.researchgate.net/profile/Kia_Jahanbin2/publication/339770709_Using_twitter_and_web_news_mining_to_predict_COVID-19_outbreak/links/5e84d4db4585150839b508b7/Using-twitter-and-web-news-mining-to-predict-COVID-19-outbreak.pdf>.

LeDoux, J. (2019, June 1). *Impute Missing Values*. <https://jamesrledoux.com/code/imputation>.

*Multiclass and multilabel algorithms* (n.d.). scikit-learn. <https://scikit-learn.org/stable/modules/multiclass.html>.

**Institutional Review Board Quiz and Approval**

Have you read and understood the “Human Subjects FAQ” page and completed the “Human Subjects FAQ Quiz” at the WGU Institutional Review Board (IRB) website? (<https://irb.wgu.edu/info/Pages/Home.aspx>)

Yes, I have read and understood the “Human Subjects FAQ” and have provided email proof of my completed quiz in appendix A. (<https://irb.wgu.edu/info/Pages/Human-Subjects-FAQ-Quiz.aspx>)

No, I have not completed the Human Subjects FAQ quiz.

Assess whether your capstone proposal complies with WGU’s IRB standards for exemption status. Explain why you believe the proposed project complies with the standards for exemption status. If it does not, make arrangements with a course mentor and the IRB for approval.

The research complies with WGU’s IRB exemption status because:

* Research involving the collection or study of freely available existing data.
* Research that does not employ methodology on human subjects.
* User IDs and screen names will not be used in the analysis.

The research requires approval from WGU’s IRB because:

Yes, I would like to schedule a conference to discuss my project.

To be filled out by a course mentor:

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Mentor’s Approval Status: Select one

Date: Click here to enter a date.

Reviewed by: Click here to enter text.

Comments: Click here to enter text.

**APPENDIX A:** Email showing completed “Human Subjects FAQ Quiz”

# Human Subjects FAQ Quiz

This exam consists of 6 questions to test your knowledge of the human subjects FAQs. You will be prompted to repeat the question whenever you enter an incorrect or partially-correct answer. Upon completion of the quiz, YOU WILL BE EMAILED A COPY OF YOUR RESPONSES as evidence that you successfully completed the quiz.   
  
• \*\*\*\*\* NOTE (STUDENTS): Your FAQ Quiz results (copy of your responses) will need to be submitted through your course.   
Here are the instructions: [https://westerngovernorsuniversity.sharepoint.com/sites/IRB2/info/PublishingImages/Pages/Human-Subjects-FAQ-Quiz/Completing%20HS%20FAQ%20Quiz%20for%20submission.pdf](https://www.google.com/url?q=https://westerngovernorsuniversity.sharepoint.com/sites/IRB2/info/PublishingImages/Pages/Human-Subjects-FAQ-Quiz/Completing%2520HS%2520FAQ%2520Quiz%2520for%2520submission.pdf&sa=D&ust=1589656566372000&usg=AFQjCNGGIVjCluP70XCe75OE-oNd2rGh6A) \*\*\*\*\*

Your email address ([**pharmaz@wgu.edu**](mailto:pharmaz@wgu.edu)) was recorded when you submitted this form.

Top of Form

## Name

**Please enter your first and last name below \***

Peter Harmazinski

## Question 1

**Who needs to first approve a capstone proposal of a student collecting data from WGU students, alumni, employers of graduates, faculty, or staff? \***

*  Office of Institutional Research
*  Student Services
*  Institutional Review Board (IRB)
*  Capstone Course Mentors in the relevant College

## Question 2

**Which one of these capstone projects may need IRB approval for use of human subjects? \***

*  An analysis of public data where identities of subjects is unknown.
*  Statistics about the prevalence of a problem gathered from quality assurance records.
*  A survey and interviews of employees about the effectiveness of a new technology
*  A comprehensive review of Internet sites and academic literature to expose trends and issues.

## Question 3

**Which one of these capstone projects may need IRB approval because it involves research? \***

*  The number and type of service calls are measured to compare the usability of two competing IT products for a business.
*  Instructional effectiveness is measured through pre- and post-test scores in a regular school or adult education program.
*  Findings from the project are to be presented at a national convention and submitted to a referred journal
*  Nurses complete an anonymous survey to assist in a clinic’s quality improvement initiative.

## Question 4

**Which one of these capstone projects may need IRB approval because a vulnerable population or situation is involved? \***

*  If the boss finds out which employees gave their opinions they could be fired or demoted if the results reveal organizational deficiencies.
*  Participants are to be paid a large incentive if the results of an organizational opinion survey are positive.
*  Research involving prisoners, homeless, elderly, decisionally-impaired, pregnant women, hospitalized patients, or undocumented individuals as research subjects.
*  The owner’s son wants to interview top leadership about what makes their employees effective in this family-owned business.
*  All of the above

## Question 5

**Which one these capstone projects may need IRB approval because data are gathered in atypical ways, subjects, or settings? \***

*  Data gathered through use of voice, video, or image recordings of human subjects.
*  Data gathered about such topics as sexuality, mental health, interpersonal violence, or illegal activities
*  Data are collected through deception about the real project purpose.
*  Use of surveys, interviews, or participant observations in a school setting.
*  Action research of training effectiveness in business, industry, health care, military, or a community organization.
*  All of the above

## Question 6

**How long is the IRB review process if the project does not qualify for "exempt" status? \***

*  48 hours for an expedited review by an IRB member, and 72 hours for a review by the full IRB.
*  Three business days for an expedited review by an IRB member, and five to ten days for a review by the full IRB.
*  None of the above; the timing depends on dates of the IRB quarterly meetings.
*  Three weeks for an expedited review by an IRB member, and two to three months for a review by the full IRB.

## Congratulations! Once you click "SUBMIT" below, you will have completed the FAQ/Flowchart Quiz!

YOU WILL BE EMAILED A COPY OF YOUR RESPONSES, VERIFYING COMPLETION OF THIS QUIZ. You must click the "SUBMIT" button to have your responses recorded and emailed to you. • \*\*\*\*\* NOTE (STUDENTS): Your FAQ Quiz results (copy of your responses) will need to be submitted through your course. Here are instructions: [https://irb.wgu.edu/info/PublishingImages/Pages/Human-Subjects-FAQ-Quiz/Completing%20HS%20FAQ%20Quiz%20for%20submission.pdf](https://www.google.com/url?q=https://irb.wgu.edu/info/PublishingImages/Pages/Human-Subjects-FAQ-Quiz/Completing%2520HS%2520FAQ%2520Quiz%2520for%2520submission.pdf&sa=D&ust=1589656566377000&usg=AFQjCNG1uc17t7RFz4VvPNXkLNTRYcEiAA) \*\*\*\*\*

Bottom of Form

**APPENDIX B:** IRB approval (if applicable)