HW08 — Team final project proposal — STAT/CS 287  
TEAM NAME: SPINNING MIXER  
DATE: 11/30/2021

# Project title

Twitter Sentiment Analysis, Purchasing Power, and Mental Health in the United States

# Abstract

The increase of online presence on social media platforms has led to a plethora of data being available to researchers looking to answer a variety of questions regarding human health. This project aims at analyzing whether or not changes in sentiment analysis of tweets can be correlated with shifts in mental health by state. Additionally, this project aims at analyzing how strongly correlated mental health and purchasing power are in comparison to the Kessler Psychological Distress Scale (K6). Lastly, this project aims at analyzing how accurately one can build a prediction model to predict shifts in mental health in a particular state according to tweets from that state. To address these analyses, we will be consuming tweets via Twitter publicly available API service, mental health metrics via the Substance Abuse and Mental Health Data Archive and the Household Pulse Survey, and US financial data via the 2021/W17: Regional Price Parity Per State dataset.

# Overview and motivation of research questions

**Motivation:**

The mental health crisis facing the United States is often viewed with a different lens than other public health crises, just as mental health has historically been treated as separate from physical health. This perspective is diminutive and dangerous, considering that suicide is one of the major causes of death across multiple age groups in this country. Clearly, much more action is needed to address this problem. One of the major issues in addressing this crisis is the lack of concrete methods to predict changes in mental health prevalence and severity in a given population, which would allow for more effective and efficient allocation of relevant resources, such as funding for access to mental health professionals, the availability of mental health crisis response teams, and public outreach and awareness campaigns.

Never before in human history have we had access to the daily thoughts and correspondence of so many individuals. Over the past two decades, the rise and subsequent widespread adoption of social media has allowed for this unprecedented look into the mental state of billions of users worldwide. This provides researchers with a unique opportunity in examining social media correspondence in relation to numerous societal factors. Performing sentiment analysis on this text provides concrete metrics on positivity and negativity, factors that could potentially be explained by a user's mental status and health. One could easily envision how a user suffering from depression, anxiety, PTSD, or other mental health issues might exhibit more negative sentiment than a user not afflicted by mental health issues.

This project makes heavy use of predictive modeling and sentiment analysis, two key skills developed in this course and provides us with the opportunity to apply skills we learned from homework assignments in UVM’s CS-287 Data Science I course to speculate on, through the lens of data science, the correlations of twitter sentiment analysis, purchasing power, and mental health across various states in the United States of America.

**Research Questions:**

1. How strongly are mental health and compound tweet sentiment correlated?
2. How strongly are purchasing power and mental health correlated?
3. How useful a predictor of mental health is social media correspondence?

# Datasets to be used

* Twitter API (<https://developer.twitter.com/en/doc>)
  + The Twitter API is a set of endpoints that can be used (programatically) to understand or create conversation on Twitter
* Substance Abuse and Mental Health Data Archive - National Survey on Drug Use and Health (<https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001>)
  + This dataset is the leading source of statistical information on mental health issues in the United States. It tracks trends in specific mental health illness measures and treatment for these disorders. The population is the general civilian population ages 12 and older. It includes questions from the Diagnostic and Standard Manual of Mental Disorders that allows diagnostic criteria to be applied.
  + Our variables of interest are the Kessler K6 score, year, and state. We have recoded counts for the number of individuals in each state at each level of the score (0-24) for years from 2002-2019.
  + According to Prochaska (2012), “The widely-used Kessler K6 non-specific distress scale screens for severe mental illness defined as a K6 score ≥ 13, estimated to afflict about 6% of US adults.” (Abstract). It is the result of a six question psychological distress survey, with each question having a possible response of 0 to 4.
  + According to Hedden (2012), these questions “correspond to how nervous, hopeless, restless or fidgety, sad or depressed, or worthless the respondent felt and to what extent everything felt like an effort to the respondent.” (Section 2.1.2)
  + While the Kessler K6 score has traditionally been used to screen for severe mental illness, which is defined as a K6 score of at least 13, subsequent analysis of the scale has determined that it is also a valid indicator of moderate mental distress with a threshold score of 5 on the scale. (Prochaska, 2012)
* National Center for Health Statistics - Household Pulse Survey (<https://www.cdc.gov/nchs/covid19/pulse/mental-health.htm>)
  + This dataset is the result of a partnership between the Centers for Disease Control and the Census Bureau, producing an experimental data system which aims to rapidly monitor recent changes in mental health in the US.
  + It was designed to provide information about the impact of the coronavirus pandemic. Data was collected during 3 separate phases, and is broken down into 2 week intervals.
  + The data shows estimates of the percentage of adults who report symptoms of anxiety or depression, and are presented for every state in the US.
* 2021/W17: Regional Price Parity Per State (<https://data.world/makeovermonday/2021w17>)
  + This dataset compares regional prices in America from 2008-2019 to gauge the purchasing power across states in the ~10 year time frame

# Methods

This project aims to better understand the relationship between sentiment extracted from tweets and mental health issues in the US. To explore this relationship, we will utilize Python’s Flask framework to provide a web service that uses the Twitter API in conjunction with multiple US-centric mental health datasets. Additionally, this project aims at building a mental health prediction tool using tweets associated with a particular location. Likewise, the web service will contrast the accuracy of various NLP algorithms in attempts to predict movement of mental health by location according to tweets in said region. The web service will provide both a front-end experience for those interested in visual analyses, as well as a ReSTFUL back-end API service (and documentation) for those who are solely interested in using this tool as a wrapper to easily pull any potential data of interest presented (or used) in the front-end analysis.

Our predictive model will be trained on compound sentiment from VADER sentiment analysis on geotagged tweets in the US during the time period of interest (2012-2019), as well as data on personal finances by state during this time period. We will collect sentiment scores for each tweet in our dataset by state over the time period. We will build at least two different models in terms of targets. One model will use the Kessler K6 score from the NSDUH dataset, and incorporate tweet sentiment over multiple years. The goal of this model will be to predict longer term trends in mental health prevalence and severity. The other model will be more short term focused, and use the targets of depression and anxiety prevalence taken from the Household Pulse Survey. This will also be trained using tweet sentiment, but the moving average will obviously be taken over a shorter time scale. We will experiment with different lengths of time over which to take the moving average for both models. Our initial hypothesis is that we will achieve the best results with a monthly or weekly average in the long term model, and a daily or hourly average in the short term model.

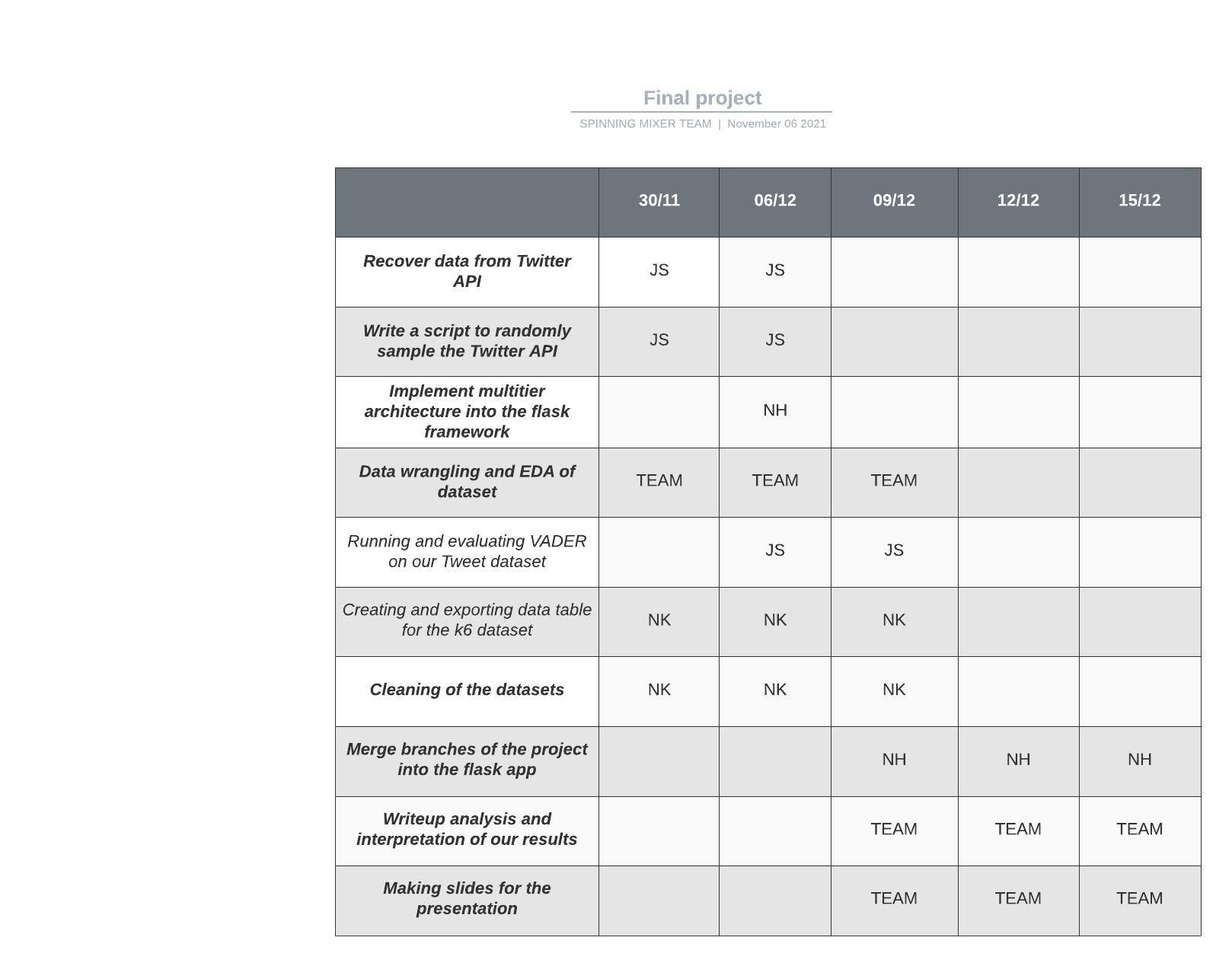


Figure 1: Gantt chart: a roadmap to success! This organizational structure will enable us to be more efficient and flexible in tackling the tasks required for our project.

# Bibliography

Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. *ICWSM*.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60.<https://doi.org/10.3115/v1/W14-3207>

Fischer, C. S. (2008). What wealth-happiness paradox? A short note on the American case. *Journal of Happiness Studies*, *9*(2), 219–226.<https://doi.org/10.1007/s10902-007-9047-4>

Hedden, S., Gfroerer, J., Barker, P., Smith, S., Pemberton, M. R., Saavedra, L. M., Forman-Hoffman, V. L., Ringeisen, H., & Novak, S. P. (2012). Comparison of NSDUH Mental Health Data and Methods with Other Data Sources. In *CBHSQ Data Review*. Substance Abuse and Mental Health Services Administration (US).<http://www.ncbi.nlm.nih.gov/books/NBK390286/>

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976.<https://doi.org/10.1017/S0033291702006074>

Kim, H., Jang, S. M., Kim, S.-H., & Wan, A. (2018). Evaluating Sampling Methods for Content Analysis of Twitter Data. *Social Media + Society*, *4*(2), 2056305118772836.<https://doi.org/10.1177/2056305118772836>

Prochaska, J. J., Sung, H.-Y., Max, W., Shi, Y., & Ong, M. (2012). Validity study of the K6 scale as a measure of moderate mental distress based on mental health treatment need and utilization. *International Journal of Methods in Psychiatric Research*, *21*(2), 88–97.<https://doi.org/10.1002/mpr.1349>