

Name: Mark Febrizio, Shumel Siraj, Alex Thiersch, Xuan Zou

Date: 04/24/2022

DATS 6101

Team 6

Final Project Topic Proposal

The results of our midterm project indicated that there was a relationship between income inequality and mental health, which varied geographically across regions of the U.S. In this next stage, we intend to use statistical models to better understand the strength of the relationship.

We plan to build a linear regression model that assesses the strength of this relationship. Our independent variable for [measuring income inequality](#) is the “ratio of household income at the 80th percentile to income at the 20th percentile.” We have two alternative measures of mental health that can be used as dependent variables: the *rate of frequent mental distress* and the *average number of poor mental health days reported in the past 30 days*. We also plan to include additional variables that are expected to affect mental health, including median household income, the percentage of households with severe housing problems, the unemployment rate, the child poverty rate, and county-level demographic information. Additionally, regression tree and random forest models will be used to determine variable importance and better predict the prevalence of mental health issues.

This discussion provokes the following SMART questions:

- *Is the relationship between mental health and income inequality across U.S. counties robust when including other economic variables?*
- *Does the relationship differ across regions of the U.S.?*
- *Does the relationship depend on the measure used for mental health?*

Our dataset combines annual datasets tracking county-level public health data, containing approximately 19,000 observations in total. The specific timeframe of 2016-2021 provides the most complete data for our dependent and independent variables.

Data source: RWJF County Health Rankings; [2016-2019](#); [2020-2021](#)

GitHub repo: <https://github.com/mfebrizio/DATS-6101-group-6>