**Title:** Credit Score and Default Risk
**Team members:** Spencer Hanson, William Brickowski, Ben Miller, Max Moghadam

- **Description:** Study of loan, macroeconomic, and crime rate data to devise a credit score that reflects risk of defaulting. We hope to come up with a similar or more accurate credit score than the ones that have been computed by Freddie Mac.
- **Research Questions:**
  - How do local societal and monetary factors, like GDP and crime rate, affect default rates?
  - What attributes make it likely that a person will default on their loan? (e.g. age, gender, previous home ownership, ect.)
- **Prior Work:** There is a lot of prior work involving risk quantification with regard to loan defaults. We are primarily interested in the drawbacks of common statistical methods used to create credit scores, but are also interested in general methods for quantifying risk. The following papers discuss some of these:
  - Credit Scoring and Scoring of Risk
    - https://www.brookings.edu/wp-content/uploads/2016/07/cbo-credit-score-background-paper.pdf
  - Credit Scoring and Loan Default
    - https://files.stlouisfed.org/files/htdocs/wp/2011/2011-040.pdf
  - What's the Point of Credit Scoring?
    - https://pdfs.semanticscholar.org/4ccd/81d64e04ac7cadd9936a703543075fa24846.pdf
  - Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files
    - http://onlinelibrary.wiley.com/doi/10.1111/1540-6229.00811/epdf
- **Datasets:**
  - Mac Single Family Loan-Level Dataset
    - http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.html
  - FBI Crime Data
    - https://ucr.fbi.gov/
  - **Census Data**
    - https://usa.ipums.org/usa/
- **Proposed work**: What do we need to do?
  - Data cleansing: We will need to convert the multiple data-sets into compatible forms. This includes changing n/a values in all sets to a uniform number or symbol, removing observations with many missing attributes, ect.
  - Data preprocessing: We will need to filter out useless data columns, which we will pick according to how irrelevant we consider an arbitrary column to be with regard to loan default.

    - Data integration: We will need to match monthly loan data to the individual loan data by merging the two datasets according to the consumer ID. Additionally, we will need to match the locations that the loan was made with locations related to GDP and crime rates.
    - Data regression and correlation: We will need to formulate regressions that will predict a probability of a customer defaulting on their loan. This can be done with hazard rate modeling, reduced form credit models, weight of evidence models, linear or logistic regression.
        - Correlation coefficient between default rate and varying attributes to determine a weighted credit score
        - Probit/Logit regression models to predict the probability of defaulting on loan
- **List of tool(s):** Python, Pandas, Numpy, SciKit-Learn
- **Evaluation:** Were we able to calculate loan default probability to a similar or greater degree than Freddie-Mac credit score? If so, we achieved our goal.