My project proposal will be to look evaluate the U.S. Chronic Disease Indicators: Cardiovascular Disease from the CDC.

My data can be found here:

<https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Cardiovascular-Dise/232j-jiq5>

Data Description:

34 columns

109K rows

**Features that I am interested in:**

* Rate of occurrence (continuous):
  + Question (nominal):
    - Mortality from Cardio vascular
    - Hospitality from cardio vascular
    - Pneumonia, influenza death and hospitalization related to cardiovascular (these will be ignored to simplify analysis)
  + Reported in %, cases in 1000 and cases in 10,000 and raw number
    - will need to be converted to 1 number
  + Rate is reported in:
    - Crude Prevalence and Crude Rate
      * will be combine to 1 metric
    - Age adjusted prevalence and age adjusted rate
      * will be combine to 1 metric
    - raw count number
      * raw count number will also be ignored
  + Whether or not the rate is age adjusted or crude will be coded to a feature
* Stratification Category:
  + Gender (nominal)
  + Race (nominal)
  + Gender and Race does not cross as features and thus MUST be stratified and cannot be analyzed together. Race is chosen as the larger data set.
* Location (nominal): State

**Questions I want to answer with machine learning:**

1. Linear Regression (manual from scratch): Given, race and Location, model will learn to predict rate of hospitalization and mortality from cardio vascular causes.
2. Classification: Location, rate of hospitalization, and rate of mortality, can we build a model to predict race. In other word, how is cardiovascular issues impact different ethnicity.