TEDS Data Cleaning

* Original Dataset
  + TEDS-D
    - 2017
    - 2018
    - 2019
  + The original dataset contained roughly 1.5 million rows for each file
* Initial Cleaning
  + TedsMulticollinearity\_Datacleaning.ipynb
    - Read in each of the TEDS files (2017, 2018, 2019) and combine them
    - Columns Initially dropped
      * DISYR, CASEID, CBSA2010 , LIVARAG , DAYWAIT , DIVISION, REGION, HLTHINS , PRIMPAY, DETNLF , DETNLF\_D
    - Check for Multicollinearity based off correlations
    - Drop highly correlated variables
      * 'ROUTE2','ROUTE3','FRSTUSE3','FRSTUSE2','PREG','ROUTE1','FREQ3\_D','FRSTUSE1','SUB3\_D','SUB2\_D','SUB1\_D','SERVICES\_D'
    - Drop STFIPS
    - Reassigning numerical values to actual values (some are in ranges, then middle value is chosen)
      * Education
      * Age
      * Gender
      * Race
      * Ethnicity
      * Marital Status
      * Employment
      * VET
      * Living Arrangement
      * Primary Income
      * Arrests
      * Services
      * Opiod Therapy
      * Target Variable (Reason)
      * Length of Stay
      * Referral
      * Criminal
      * Prior Treatment
      * Substance
      * Frequency
      * IDU
      * Alcohol Drug Use
      * Diagnosis
      * Psychiatric Problems
      * Frequency of self attendance
    - Savings as TEDS\_cleaned\_data\_categorical\_field\_values\_state.csv
* Handling Missing values and Categorical Variables
  + TEDS\_cleaned\_code\_110521
    - Read in “TEDS\_cleaned\_data\_categorical\_field\_values.csv”
  + Columns are dropped if the number of missing data is > 60%
  + Create a separate dataframe for categorical variables
  + Replace Missing
    - If the column contains ‘None’, ‘Nan’, ‘’ , replace with ‘Missing’
      * Done for categorical dataframe
      * Done for numerical dataframe
    - Replace missing categorical values with the mode values
    - Replace missing numerical values with the mean values
  + Combine categorical dataframe and numerical dataframe together
  + Get dummies for categorical dataframe
  + Save as ‘tedcleanedcsv11521.csv’