Virginia Hebert

DAT-119: Python 1

Fall 2020

# Project Plan

## Overview

Plan partner: Rachael Shockey

My plan is to use Python to clean up and pull information from the earthquakes file found on Kaggle and supplemented with some USGS GIS data.

Breakdown data by fault type and run summary statistics on it for some general information, then create plots comparing various variables using just csv reader and pandas. References used to create the functions are included in the function descriptions.

## Inputs

* CSV file (some file clean-up beforehand, due to the earthquake/fault merge done in GIS)
* Some input from user as to what they’d like to do (stats/export data)

## Outputs

* Subsetted file(s) by fault type
* Multiple/various statistics on earthquake and fault attributes

## Steps the program needs to take

1. Open the file
2. Read in the information
3. Loop through and cut down to pertinent columns/rows
4. Run statistics/functions on data
5. Create crossplots of select columns
6. Export data based on user selection

## Functions I expect to need

* clean\_file() – loop through and remove unnecessary columns/rows
* subset\_data() – subset the according to fault type (Normal, Thrust, Right Lateral, Left Lateral, Strike Slip)
* read\_file() – by row and by header
* tally() – tally of each fault type
* avg\_depth/magnitude/azimuth/age() – function for each
* basic\_plot() – simple cross plot using pandas
* boxplots() – boxplot of attributes using pandas
* pearson\_matrix() – matrix of correlation coefficients
* all\_stats() – list of basic stats for all the attributes
* main()

# Process Log

* Download earthquake csv from Kaggle (<https://www.kaggle.com/usgs/earthquake-database>)
  + 23412 rows
* Create shapefile in ArcGIS Pro using lat/long information in csv
* Spatial proximity analysis to join earthquake data to USGS fault data (based on the assumption that the fault was the likely source of the earthquake)
* Pull database file from shapefile and delete any rows that did not have an earthquake/fault join
  + 131 rows left
* Delete any columns of obviously not useful data for Python analysis (ArcGIS specific information, null columns, columns with minimal data, columns with web links)