Final Assignment

Due November 5, 2025

For this assignment, we will be using Jupyter Notebooks to accomplish two tasks:
(1) generate some simulation data using numpy and one of the built-in methods from the random number generator object
(2) explore a new dataset and provide some summary statistics (mean, median, etc.)

We will be using the following libraries:

- numpy for simulations
  - use the random number generator object, `numpy.random.default_rng()`
- scipy for loading the data
- pandas for data manipulation
- matplotlib for creating figures and for plotting
- seaborn for plotting

Within your notebook, please include examples of using the following:

- for loop
- boolean expression
  - could be with an if statement
  - could be with indexing a dataframe
- create a function and use it within your report
  - e.g. take an axis as input and change the linestyle of a plot
- use of Markdown cells to explain the steps in your analysis

Explore pandas features and its classes, `DataFrame` and `Series` and use some of their methods:

- use groupby and agg to generate your own descriptive statistics
- use describe for some general descriptive statistics
- use the mean method of the `Series`
- subselect some of your dataframe using indexing

Dataset to use for the final project
https://www.openml.org/search?type=data&sort=runs&status=active&id=287

This dataset is stored in the `.arff` format. This format is not readily useful in `pandas`. However, `scipy` has a built-in file reader for this data file. To access the data, you can use the following commands:

```
from scipy.io import arff
import pandas


arrs = arff.loadarff('/Users/teonbrooks/Downloads/wine_quality.arff')
df = pd.DataFrame(arrs[0])
```

For the final project, we will be presented our notebooks. You will each get roughly five minutes to walk us through your notebook and highlight some of the findings from this assignment.