# Comp 4433: Assignment 2

Oftentimes it is helpful to modularize specialized plotting functionality for use on a common or recurring task. In other words, if we find ourselves consistently in need of producing the same plot with a standard set of parameters and/or modifications then building a function to implement that plot can speed up our work and offer a mechanism to quickly generate consistent output.

For this exercise you are asked to build a function utilizing the object-oriented matplotlib interface. The function should generate a histogram with an overlay of the empirical cumulative distribution. The purpose of the function is to be applied to features that are potentially heavily skewed or that may have extreme values. In such cases our distributions will be visually compressed, preventing us from seeing how the bulk of the data are distributed. To this end, your function should implement some approach for empirically identifying a floor and ceiling for the displayed values. These thresholds will then be used within the function to limit the scope of plotted data.

The function should generalize to any dataset and feature.

Your function should support being called with different binning strategies but should default to something other than using k bins (sqrt is a good choice).

The function should take, as arguments, the following:

- A pandas DataFrame
- A feature name
- One or more parameter dictionaries to be passed into the mpl plotting functions. Note that these dictionaries can be unpacked into the plotting functions and may include plot-specific arguments as well as supported **kwargs.

The function should return a figure and an axes object, such that it may be called as follows:

**fig, ax = my_func(df, ‘my_feature’, param_dict1, param_dict2)**

The resulting plot should also contain a text box with relevant information about the truncation that occurred. Specifically, please include the following:

- The empirical floor and ceiling
- The proportion of records not displayed
- The minimum and maximum observed values

Note that depending on your input it may not be necessary to truncate your displayed data. Additionally, you may only need to truncate extreme high values or extreme low values.

Generated plots should have a title or label that indicates which variable is being displayed. The y axes of the plot should be appropriately labeled. Consideration should be given to producing plots that have simple and uncluttered aesthetic properties.

To test your function please use fall enrollment data from the Integrated Postsecondary education Data System (IPEDS). These data express fall enrollments for US colleges and universities. These values are typically highly positively skewed. Use a selection of the available enrollment features to assess your function.

You may retrieve the required file with the following command:

```
df=pd.read_csv('https://nces.ed.gov/ipeds/datacenter/data/EF2022A.zip',
               compression='zip',
                encoding="ISO-8859-1")
```

A data dictionary for the file can be found at the following location.

https://nces.ed.gov/ipeds/datacenter/data/EF2022A_Dict.zip

### Deliverable

Submit a notebook file with your final function and at least one test call of the function using the data referenced above.