Lab 12 – Collaboration

Working with 1 to 2 other people create a Python program to accomplish the following and document that program in a lab report:

We are going to generate an analysis of four datasets:

1. The Average Wage Index of the United States from 1973 – 2017 (Source: Social Security Administration)
2. The average cost of fresh vegetables from 1947 – 2015 (Source: US Bureau of Labor Statistics)
3. The average cost of steers and heifers from 1947 – 2015 (Source: US Bureau of Labor Statistics)
4. The average cost of College from 1971 – 2018 (Source: College Board)

For the years in common amongst the four data sets, calculate the following:

1. If monthly numbers are given, an annual average
2. Percent increase from first year to final year of the analysis
3. Average annual increase
4. The linear trend line for the four datasets

Write these pieces of information into an attractively formatted file

Generate a single graph showing the four pieces of data along with their corresponding linear trendlines, include descriptive axes and a legend. To make a useful graph, a scaling factor may need to be added to normalize the data sets. These pieces of data are on different scales (example: the vegetables data has a range of 34 – 120, yet the wage data has a range of 7,000 – 50,000). Normalize the data by dividing all the dependent data (y values) by the first data point and then multiplying by 100. This will give data points which will graph reasonably together. The trendlines will need to be recalculated using the normalized data points.

A technically correct solution is insufficient to obtain full points. The solution must logically partition the problem into functions and utilize data types (dictionaries, lists, objects, etc.) which are reasonable to hold the various pieces of data being analyzed. Note: there is not one right way to do this, the approach must be reasonable.

Do not edit or change the input files.

Equations to generate a linear trendline:

Recall that the y-intercept form of the equation for a line is y = mx + b, where m is the slope and b is the y-intercept.

For a set of data where every data point is in the form (x,y) (this is called a linear regression)

$$m = \frac{n(\sum xy) - \sum x \sum y}{n(\sum(x^2)) - (\sum x)^2} \qquad b = \frac{\sum y(\sum(x^2)) - \sum x \sum xy}{n(\sum(x^2)) - (\sum x)^2}$$

Recall that the x is the independent variable. To calculate the y values for the linear trendline, first find m & b for the dataset, then calculate y for each x using the generated equation. This will generate a set of points which all reside on the linear trendline and can be graphed.