# Arrays-Order to the Universe

The types of data structures we have been using are (

**Lists**-We have worked with these and they are nice because they are mutable and easy to work with. But they are not great with numbers.

**Tuples**-We are not spending time on them. They are like lists with a funny name but they are immutable. So they can be good if you need something not to change. But we don't use them because they don't do much with numbers.

**Dictionaries**-We use these a little. They are thought of key:value pairs and are created with {}. We have used these already when defining our "props" on the graphs. It lets you pass a few keyword values at once. We won't use these much but you will come across them.

**Numpy arrays**-we have started using these. We have seen they are easy to plot and to do math with. But they are not great with large datasets with lots of different columns and with missing data. But they are the basis for a lot of things in python so you always build off of numpy.

Pandas Dataframes-We have started these, these are like supercharged numpy arrays that give you a lot more information. If you could imagine that you could name the rows and columns in a numpy array it starts to get you there. Sort of like an excel sheet in the computer memory but more powerful. Plus they are good with dates and reordering. So these are good for complex datasets where we want to name variables. We will mainly be using these and building everything off of them.

### But how do we think about data?

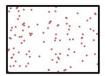
When we usually think about data we think about tabular data. This is an excel sheet. Just a table of the data we have, this is the dataframe we read in. But there are really three data types we might run into

- Tabular Data
- Vector Data
- Raster Data

The picture below explain them

# Data types

- Vector
  - Points
  - Lines
  - Polygons
- Raster
  - Digital Elevation Models (DEM)
  - Ortho imagery (aerial photography)
  - Satellite imagery
- Tabular data
  - Attributes
  - Databases

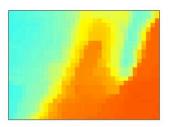


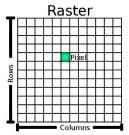


п	FID	Strape."	AREA	STATE NAME	STATE MPS	SUS RECEIP
		Polygon	67293 054	Washington	53	Pacific
	- 1	Phlygen	547294 653		200	Mrs
	2	Pulygon	22161.025	Neina	22	NEIG
		Pelypos	70612,056	NO TH CONOTE	58	NY N Ces
	- 4	Polygon	77185-055	South Cavora	46	KI N Cen
		Phlysen	97993 199	Wygren	56	Mr.
		Polygon	56003.178	Wespen	55	EN Cox
	7	Pelypon	03343,643	Ideho	10	Mon
		Felyton	9663,272	Vernont	50	Nicro
	9	Phlyspn	541270.49	Venuects	27	SV N Cen
	10	Privates	97973.594	Oregon	41	Paping
	- 11	Polyppa	9259.527	New Hampahire	22	NEto
	12	Pelyton	00257,905	lows .	15	10 N Cen
	120	Relygen	6172.561	Vaccachusetts.	28	N ting
	14	Patroon	77530.258	Nationalse	31	N N Con

# Raster data

- Areas broken into "pixels" or cells
- Each cell contains data
- · Good at representing dense data:
  - Land cover
  - Elevation





Vector data are used with GIS. They are for putting lines, points, or polygons on a map. For example putting a shoreline on a map or a road or a lake. We are not going to use them much this semester.

But first today we are going to talk about Raster data and two dimensional arrays. Raster data is really a 2d array.

Today lets just work with two dimensional numpy arrays. You can have arrays of as many dimensions as you want but I have trouble comprehending at three and more dimensions.

```
In [1]: %matplotlib ipympl
        import matplotlib.pylab as plt
        import numpy as np
        from scipy import stats
        import pandas as pd
In [2]: oneD=np.array([1,2,3,4,5,6,7,8,9,10]) # does anyone remember that band?
In [3]: oneD
Out[3]: array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
In [4]: #review-what will this print?
        #oneD[0:10:2]
In [5]: twoD=np.array([[1,2,3,4],[5,6,7,8]])
In [6]: twoD
Out[6]: array([[1, 2, 3, 4],
                [5, 6, 7, 8]])
        Do you see what I just did? It is 2 one dimensional arrays together to make a 2-dimensional array or table!
In [7]: type(twoD)
Out[7]: numpy.ndarray
```

In [8]: len(twoD)

```
Out[8]: 2
 In [9]: twoD.shape
 Out[9]: (2, 4)
In [10]: np.shape(twoD)
Out[10]: (2, 4)
In [11]: twoD.size
Out[11]: 8
          You can see what you can do with the np array by typing twoD. and then tab and you see the functions available. try
In [108... twoD.
          Now lets try slicing. Remember it is rows and then columns. See if you can guess before uncommenting and running.
          This picture is just to help you and you don't need to load it.
In [77]: from IPython.display import Image
          Image(filename='array-axes.png',width=400)
                                       axis 1
Out[77]:
                                                       2
                               0
                                           1
                      0
                              0,0
                                          0,1
                                                      0,2
         axis 0
                      1
                              1,0
                                          1,1
                                                      1, 2
                      2
                              2,0
                                          2, 1
                                                      2, 2
In [12]: twoD[0,0]
Out[12]: 1
In [13]: twoD[1,0]
Out[13]: 5
In [14]: twoD[:,:]
Out[14]: array([[1, 2, 3, 4],
                  [5, 6, 7, 8]])
In [16]: #twoD[:,0]
```

In [111... #twoD[:,1]

In [112... #twoD[0,:]

In [113... #twoD[1,:]

# Now lets do some more slicing!

remember. For numpy it is

### [start:stop:skip]

In [21]: twoD[3,3]=100

In [22]: twoD

if you list multiple items and leaving one out assumes the last one is missing

so that means [1::] is one to the end by 1.

If you have a 2d array it will be [start:stop:skip,start:stop:skip] for the rows and then the columns

```
Out[22]: array([[ 1,
                                 4],
                           7,
                                 8],
                [5,6,
                [ 9, 10, 11, 12],
                [ 13, 14, 15, 100]])
In [23]: twoD[1:3,1:3]=55
In [24]: twoD
                                 4],
Out[24]: array([[ 1,
                       2,
                            3,
                           55,
                                 8],
                [ 5, 55,
                [ 9, 55, 55, 12],
                [ 13, 14, 15, 100]])
         you can use a function to set numbers!
In [25]: twoD[0,:]=np.arange(10,14)
In [26]: twoD
Out[26]: array([[ 10,
                      11,
                           12,
                                13],
                [ 5,
                      55,
                           55,
                                8],
                [ 9,
                      55, 55, 12],
                [ 13, 14, 15, 100]])
         you can reshape the array if you want to.
In [27]: print (twoD.reshape(16,1))
        [[ 10]
         [ 11]
         [ 12]
         [ 13]
         [ 5]
         [ 55]
         [ 55]
         [ 8]
         [ 9]
         [ 55]
         [ 55]
         [ 12]
        [ 13]
         [ 14]
         [ 15]
         [100]]
In [28]: print (np.reshape(twoD,(1,16)))
        [[ 10 11 12 13 5 55 55 8
                                          9 55 55 12 13 14 15 100]]
In [29]: print (twoD.reshape(8,2))
        [[ 10 11]
        [ 12 13]
         [ 5 55]
         [ 55
              8]
         [ 9 55]
         [55 12]
         [ 13 14]
         [ 15 100]]
In [30]: print (twoD)
        [[ 10 11 12 13]
        [ 5
              55
                  55
                      8]
         [ 9
              55
                  55 12]
         [ 13 14 15 100]]
```

The shape is back to how we had it because we never changed because we only printed it. we never set it.

### Now this is where we intersect with Raster Data

#### Add Color!

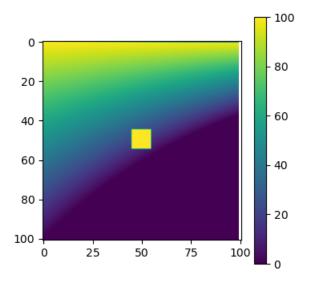
You can visualize the whole array! This just colors the array/grid we have by its values

This is raster data. It is like satelite data.

I added fig.set\_size\_inches(4,4) to save space printing. You do not need to add it.

```
In [78]: fig.ax=plt.subplots()
   fig.set_size_inches(4,4)
   cax=ax.imshow(twoD)
   fig.colorbar(cax)
```

Out[78]: <matplotlib.colorbar.Colorbar at 0x1353724f0> Figure

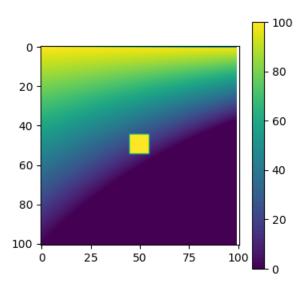


If you want to make the edges of each box smooth we need to interpolate the data. I chose interpolation='bilinear' but there are many options https://matplotlib.org/gallery/images\_contours\_and\_fields/interpolation\_methods.html

```
In [79]: fig.ax=plt.subplots()
   fig.set_size_inches(4,4)
   cax=ax.imshow(twoD,interpolation='bilinear')
   fig.colorbar(cax)
```

Out[79]: <matplotlib.colorbar.Colorbar at 0x13542b190>

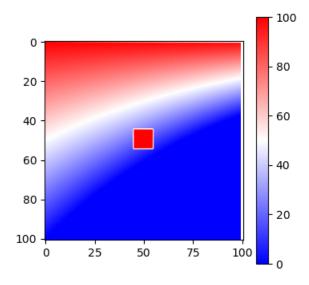
Figure



We can change the colorbar. This is another keyword argument

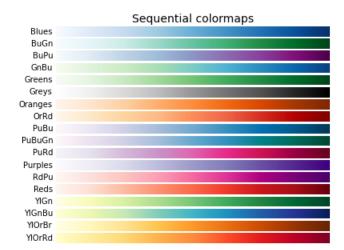
```
In [80]: fig.ax=plt.subplots()
    fig.set_size_inches(4,4)
    cax=ax.imshow(twoD,interpolation='bilinear',cmap='bwr')
    fig.colorbar(cax)
```

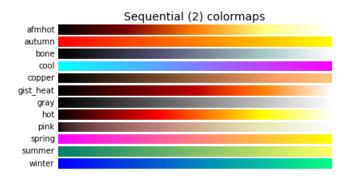
Out[80]: <matplotlib.colorbar.Colorbar at 0x1354eaa30> Figure

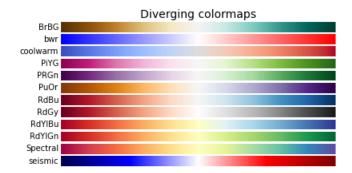


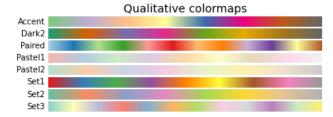
Now make your own. Here is just one list. You can google colormaps python.

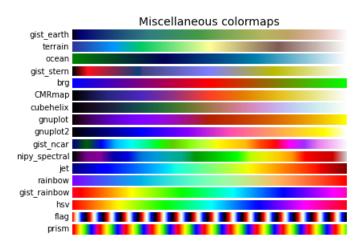
In [140... #See below for the colormap code. I also just ran the code off the web to make them ourselves.









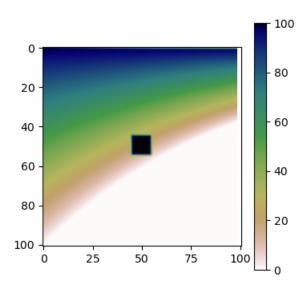


- Plot the same plot with a color bar of your choice.
- reverse the color bar by adding \_r to the name.

```
In [81]: fig,ax=plt.subplots()
    fig.set_size_inches(4,4)
    cax=ax.imshow(twoD,cmap='gist_earth_r',interpolation='gaussian')
    fig.colorbar(cax)
```

Out[81]: <matplotlib.colorbar.Colorbar at 0x1355a9fa0>

Figure



# Now we are going to intersect back with pandas and dataframes.

We are going to

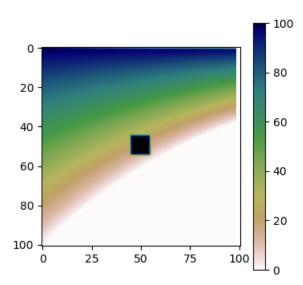
- read in a csv using pandas.
- this gives us a big dataframe which we can convert to a 2 dimensional array that is 100x100 is in size
- then we can plot/map it.

https://github.com/bmaillou/BigDataPython/blob/master/my\_first\_csv.csv

```
In [36]: df=pd.read_csv('my_first_csv.csv',header=None)
In [37]: twoD=df.values # this takes the dataframe and turns it into an array
In [82]: fig,ax=plt.subplots()
    fig.set_size_inches(4,4)
        cax=ax.imshow(twoD,cmap='gist_earth_r',interpolation='gaussian')
    fig.colorbar(cax)
```

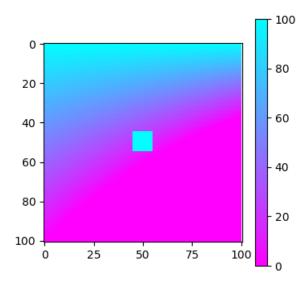
Out[82]: <matplotlib.colorbar.Colorbar at 0x1356661c0>

Figure



Remember we can alter the array. I am going to add a square in the middle of value 100...

Out[83]: <matplotlib.colorbar.Colorbar at 0x13573cb50> Figure



In []:

# Now you need to Read in Brian.csv.

It is a 2d array. Plot it with imshow and tell me what it looks like. you will be shocked at what it shows......

https://github.com/bmaillou/BigDataPython/blob/master/Brian.csv

In [ ]:

Homework hint: Think about how the data from Brian.csv is stored and how you can change it....

#### Now back to Tabular data.

For most of our data we usually work with tabular data. One example is if the first column is X and the remainding columns are all differnt Y's. Here is an example. Read in the oneX\_manyY.csv file. print it to see it. Then plot all the values. This is really how we think of excel. But we give the columns nicer names

X Value	First Y Value	Second Y Value	Third Y Value	Fourth Y Value
1.0	1.0	10.0	4.0	6.0
2.0	2.0	9.0	4.0	6.0
3.0	3.0	8.0	4.0	6.0
4.0	4.0	7.0	4.0	6.0

https://github.com/bmaillou/BigDataPython/blob/master/oneX\_manyY.csv

```
In [41]: manyY=pd.read_csv('oneX_manyY.csv')
manyY
```

Out[41]:		x_values	first_y_values	second_y_values	third_y_values	fourth_y_values
	0	1	1	10	4	6
	1	2	2	9	4	6
	2	3	3	8	4	6
	3	4	4	7	4	6
	4	5	5	6	4	6
	5	6	6	5	4	6
	6	7	7	4	4	6
	7	8	8	3	4	6
	8	9	9	2	4	6
	9	10	10	1	4	6

To show you how data sets/types are related we could strip off the column titles so it becomes a 2d array for numpy

```
In [42]: manyY=manyY.values
In [43]: manyY
Out[43]: array([[ 1,
                                 6],
                [ 2,
                      2, 9,
                                 6],
                [ 3,
                     3, 8,
                             4, 6],
                [ 4,
                     4, 7, 4, 6],
                [5,
                     5, 6,
                             4, 6],
                         5,
                [6,
                     6,
                                6],
                     7,
                [7,
                         4,
                                 6],
                                 6],
                [8,
                     8,
                         3,
                             4,
                     9,
                [ 9,
                         2,
                             4,
                                 6],
                                 6]])
```

Now you can do all of your array nomenclature. For example lets look at the first column

```
In [44]: manyY[:,0]
Out[44]: array([ 1,  2,  3,  4,  5,  6,  7,  8,  9,  10])
```

now we can plot the data. For plotting you just keep on listing the x and y pairs.

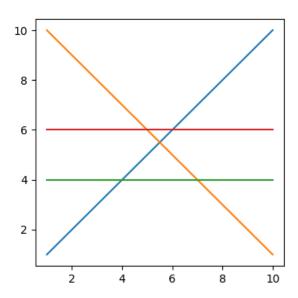
#### So plot

- column 0 versus column 1, X Value versus First Y Value
- column 0 versus column 2, X Value versus Second Y Value
- column 0 versus column 3, X Value versus Third Y Value
- column 0 versus column 4, X Value versus Fourth Y Value

```
In [84]: fig,ax=plt.subplots()
   fig.set_size_inches(4,4)
   ax.plot(manyY[:,0],manyY[:,1])
   ax.plot(manyY[:,0],manyY[:,2])
   ax.plot(manyY[:,0],manyY[:,3])
   ax.plot(manyY[:,0],manyY[:,4])
```

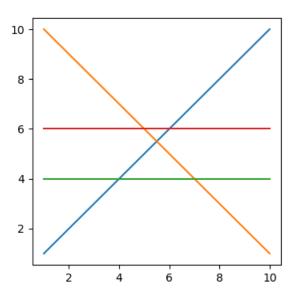
Out[84]: [<matplotlib.lines.Line2D at 0x13581ba00>]

Figure



or you can do it in one call to ax.plot by listing x and y pairs but I find this hard to follow

Figure

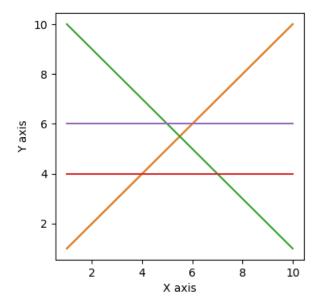


If you wanted to get fancy we could program a for loop to loop over the columns and plot them

```
In [87]: fig,ax=plt.subplots()
   fig.set_size_inches(4,4)
   for i in np.arange(5):
        ax.plot(manyY[:,0],manyY[:,i])
   ax.set_xlabel('X axis')
   ax.set_ylabel('Y axis')
```

Out[87]: Text(0, 0.5, 'Y axis')

Figure

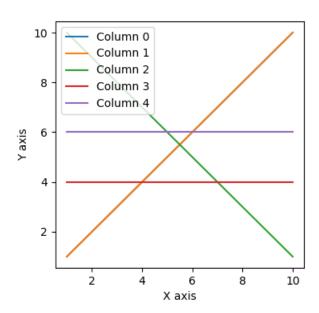


And you can add a legend

```
ax.plot(manyY[:,0],manyY[:,i],label=labeltext)
ax.legend(loc='best')
ax.set_xlabel('X axis')
ax.set_ylabel('Y axis')
```

Out[88]: Text(0, 0.5, 'Y axis')

Figure



# Compare a 2d array to a Pandas Dataframe

But now lets do it in pandas and see how it compares

In [49]:	<pre>df_manyY=pd.read_csv('oneX_manyY.csv')</pre>
	df_manyY

Out[49]:		x_values	first_y_values	second_y_values	third_y_values	fourth_y_values
	0	1	1	10	4	6
	1	2	2	9	4	6
	2	3	3	8	4	6
	3	4	4	7	4	6
	4	5	5	6	4	6
	5	6	6	5	4	6
	6	7	7	4	4	6
	7	8	8	3	4	6
	8	9	9	2	4	6
	9	10	10	1	4	6

Pandas is like an upgraded numpy array with column names.

This is going to make keeping track of data much nicer

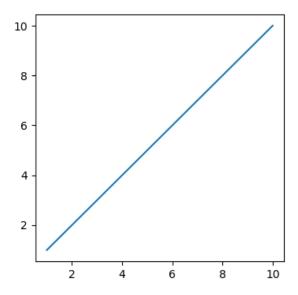
using pandas we can use the column names

```
In [50]: df_manyY.columns
```

```
Out[50]: Index(['x_values', 'first_y_values', 'second_y_values', 'third_y_values',
                 'fourth_y_values'],
                dtype='object')
In [51]: df_manyY['x_values']
Out[51]:
         0
                1
                2
          2
                3
          3
                4
          4
                5
                6
                7
          7
                8
                9
          9
               10
         Name: x_values, dtype: int64
         making the plot in pandas
In [89]: fig,ax=plt.subplots()
         fig.set_size_inches(4,4)
         ax.plot(df_manyY['x_values'],df_manyY['first_y_values'])
```

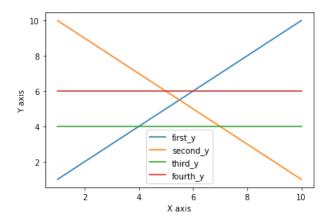
Out[89]: [<matplotlib.lines.Line2D at 0x135a130d0>]





Can you add the other columns and make a legend? Using Pandas?

```
In [86]:
Out[86]: Text(0, 0.5, 'Y axis')
```

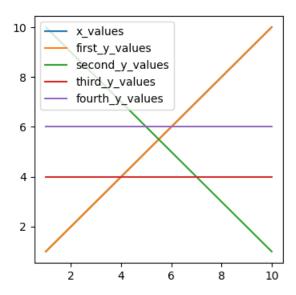


Pandas does some things to make your life easy. You can for loop over the columns. So the for loop returns the column name to col and you can pass that to ax.plot. We are going to be doing a lot more of this the next few weeks. So this is a sneak peak.

```
In [90]: fig.ax=plt.subplots()
fig.set_size_inches(4,4)

for col in df_manyY:
    ax.plot(df_manyY['x_values'],df_manyY[col],label=col)
ax.legend()
```

 $\label{eq:continuous} \mbox{Out[90]: <matplotlib.legend.Legend at 0x1359caee0>} \\ \mbox{Figure}$ 



But this plots the first column versus itself. So we just need to only call from after the first column. We will learn how to do this next time. But here it is

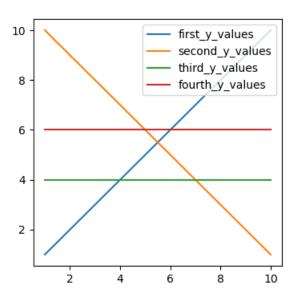
```
In [91]: fig,ax=plt.subplots()
fig.set_size_inches(4,4)

for col in df_manyY.iloc[:,1:]:
          ax.plot(df_manyY['x_values'],df_manyY[col],label=col)

ax.legend()
```

Out[91]: <matplotlib.legend.Legend at 0x135a7a520>

Figure



# Mystery file

The file mystery.csv contains data in columns. Use what you know and plot the data! The first column is x values. The others are y values

In [143...

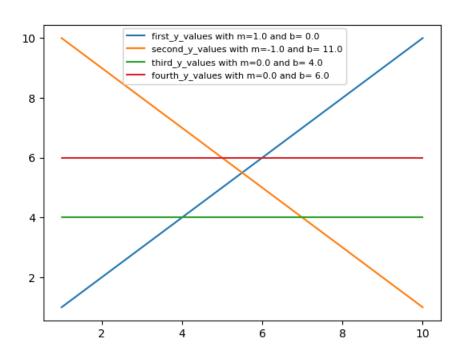
### Bonus.

If you got through this quickly see if you can go back to oneX\_manyY.csv and get the equations for each line. You could do this in a for loop and adding each equation for a line to the legend...

In [74]:

Out[74]: <matplotlib.legend.Legend at 0x13342e6d0>

Figure



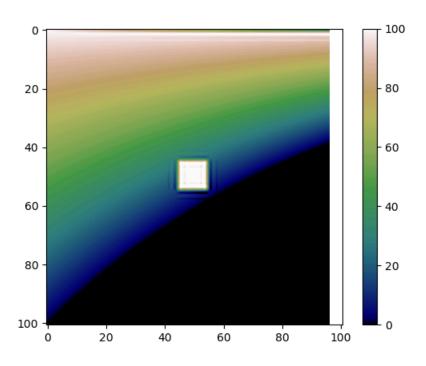
# **Answers**

# My own colorbar

```
In [60]: fig,ax=plt.subplots()
  cax=ax.imshow(twoD,cmap='gist_earth',interpolation='bessel')
  fig.colorbar(cax)
```

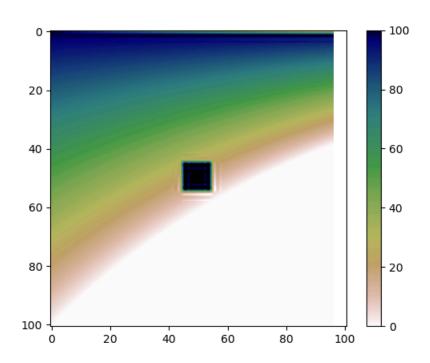
Out[60]: <matplotlib.colorbar.Colorbar at 0x132d5fbe0>

Figure



### now reversed

```
In [61]:
    fig,ax=plt.subplots()
    cax=ax.imshow(twoD,cmap='gist_earth_r',interpolation='bessel')
    fig.colorbar(cax)
```



#### Brian result

```
In [62]: Brian=pd.read_csv('Brian.csv',header=None)
Brian=Brian.values
fig,ax=plt.subplots()
ax.imshow(Brian,cmap='gnuplot',interpolation='none')
```

Out[62]: <matplotlib.image.AxesImage at 0x132d2e970>

Figure

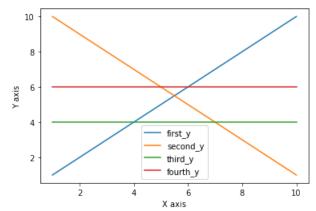


```
In [99]: fig,ax=plt.subplots()

ax.plot(df_manyY['x_values'],df_manyY['first_y_values'],label='first_y')
ax.plot(df_manyY['x_values'],df_manyY['second_y_values'],label='second_y')
ax.plot(df_manyY['x_values'],df_manyY['third_y_values'],label='third_y')
ax.plot(df_manyY['x_values'],df_manyY['fourth_y_values'],label='fourth_y')

ax.legend(loc='best')
ax.set_xlabel('X axis')
ax.set_ylabel('Y axis')
```

Out[99]: Text(0, 0.5, 'Y axis')



#### Mystery File

```
In [65]: df_mystery=pd.read_csv('mystery.csv')
    df_mystery.columns

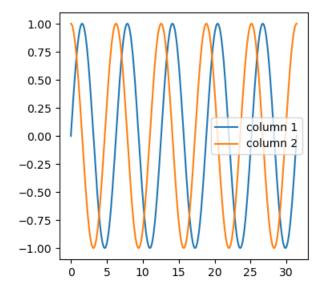
Out[65]: Index(['Column_0', 'Column_1', 'Column_2'], dtype='object')

In [92]: df_mystery=pd.read_csv('mystery.csv')
    fig,ax=plt.subplots()
    fig.set_size_inches(4,4)

    ax.plot(df_mystery['Column_0'],df_mystery['Column_1'],label='column 1')
    ax.plot(df_mystery['Column_0'],df_mystery['Column_2'],label='column 2')
    ax.legend()
```

Out[92]: <matplotlib.legend.Legend at 0x135b58460>





#### Using a for loop to get all the equations for the lines. Use linregress

```
In [73]: df_manyY=pd.read_csv('oneX_manyY.csv')
fig,ax=plt.subplots()

for col in df_manyY.iloc[:,1:]:
    x=df_manyY['x_values']
    y=df_manyY[col]

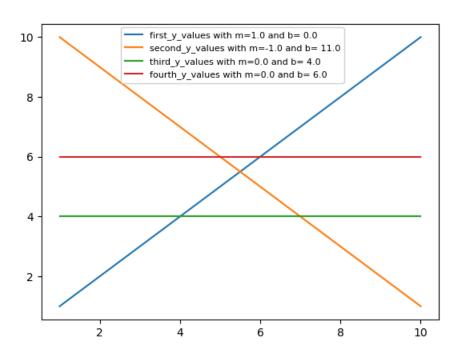
    slope, intercept, r_value,p_value,stderr = stats.linregress(x,y)
    label='{} with m={} and b= {}'.format(col,slope,intercept)

    ax.plot(df_manyY['x_values'],df_manyY[col],label=label)

ax.legend(loc=(.2,.8),fontsize=8)
```

Out[73]: <matplotlib.legend.Legend at 0x133390430>

**Figure** 



#### Code I copied from the web to show all the colormaps

```
In []:
        Reference for colormaps included with Matplotlib.
        This reference example shows all colormaps included with Matplotlib. Note that
        any colormap listed here can be reversed by appending "_r" (e.g., "pink_r").
        These colormaps are divided into the following categories:
        Sequential:
            These colormaps are approximately monochromatic colormaps varying smoothly
            between two color tones---usually from low saturation (e.g. white) to high
            saturation (e.g. a bright blue). Sequential colormaps are ideal for
            representing most scientific data since they show a clear progression from
            low-to-high values.
        Diverging:
            These colormaps have a median value (usually light in color) and vary
            smoothly to two different color tones at high and low values. Diverging
            colormaps are ideal when your data has a median value that is significant
            (e.g. 0, such that positive and negative values are represented by
            different colors of the colormap).
        Oualitative:
            These colormaps vary rapidly in color. Qualitative colormaps are useful for
            choosing a set of discrete colors. For example::
                color_list = plt.cm.Set3(np.linspace(0, 1, 12))
            gives a list of RGB colors that are good for plotting a series of lines on
            a dark background.
        Miscellaneous:
            Colormaps that don't fit into the categories above.
        .....
        import numpy as np
        import matplotlib.pyplot as plt
```

```
cmaps = [('Sequential',
                                     ['Blues', 'BuGn', 'BuPu',
                                      'GnBu', 'Greens', 'Greys', 'Oranges', 'OrRd', 'PuBu', 'PuBuGn', 'PuRd', 'Purples', 'RdPu', 'Reds', 'YlGn', 'YlGnBu', 'YlOrBr', 'YlOrRd']),
           ('Sequential (2)', ['afmhot', 'autumn', 'bone', 'cool', 'copper', 'gist_heat', 'gray', 'hot', 'pink', 'spring', 'summer', 'winter']),
                                     ['BrBG', 'bwr', 'coolwarm', 'PiYG', 'PRGn', 'PuOr',
'RdBu', 'RdGy', 'RdYlBu', 'RdYlGn', 'Spectral',
            ('Diverging',
                                      'seismic']),
                                     ['Accent', 'Dark2', 'Paired', 'Pastel1',
    'Pastel2', 'Set1', 'Set2', 'Set3']),
['gist_earth', 'terrain', 'ocean', 'gist_stern',
            ('Qualitative',
            ('Miscellaneous',
                                      'brg', 'CMRmap', 'cubehelix',
                                      'gnuplot', 'gnuplot2', 'gist_ncar',
'nipy_spectral', 'jet', 'rainbow',
'gist_rainbow', 'hsv', 'flag', 'prism'])]
nrows = max(len(cmap_list) for cmap_category, cmap_list in cmaps)
gradient = np.linspace(0, 1, 256)
gradient = np.vstack((gradient, gradient))
def plot_color_gradients(cmap_category, cmap_list):
     fig, axes = plt.subplots(nrows=nrows)
     fig.subplots_adjust(top=0.95, bottom=0.01, left=0.2, right=0.99)
     axes[0].set_title(cmap_category + ' colormaps', fontsize=14)
     for ax, name in zip(axes, cmap_list):
          ax.imshow(gradient, aspect='auto', cmap=plt.get_cmap(name))
          pos = list(ax.get_position().bounds)
          x_{text} = pos[0] - 0.01
          y_{\text{text}} = pos[1] + pos[3]/2.
          fig.text(x_text, y_text, name, va='center', ha='right', fontsize=10)
     # Turn off *all* ticks & spines, not just the ones with colormaps.
     for ax in axes:
          ax.set_axis_off()
for cmap category, cmap list in cmaps:
     plot_color_gradients(cmap_category, cmap_list)
plt.show()
```

```
In [ ]:
In [ ]:
In []:
```