

Plotting with Pandas

**Name :** Jann Moises Nyll B. De los Reyes  
**Section :** CPE22S3  
**Submitted to:** Engr. Roman Richard  
**Date :** March 28, 2024

The `plot()` method is available on `Series` and `DataFrame` objects. Many of the parameters get passed down to `matplotlib`. The `kind` argument let's us vary the plot type.

About the data

In this notebook, we will be working with 2 datasets:

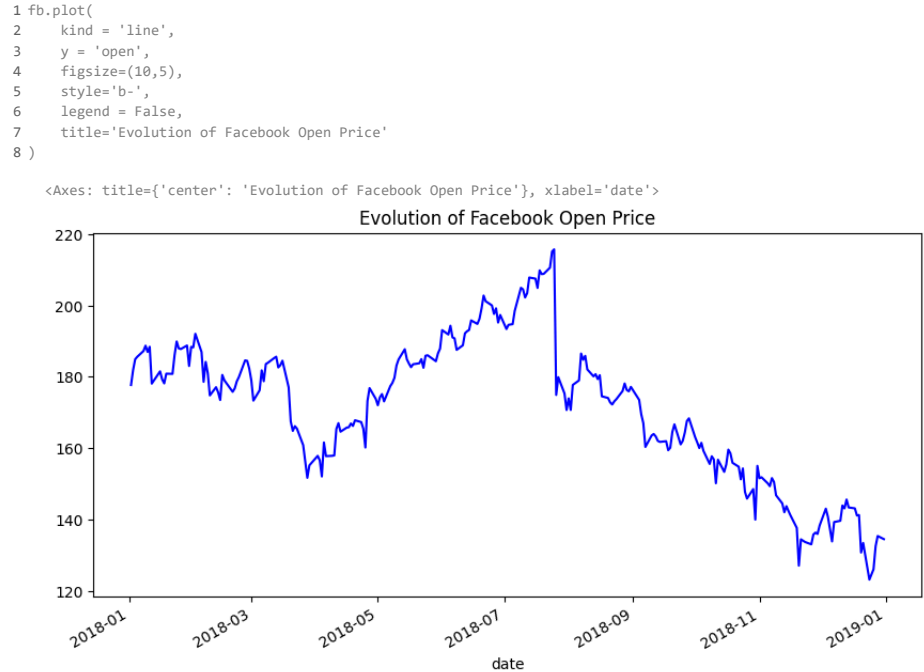
- Facebook's stock price throughout 2018 (obtained using the [stock\\_analysis package](#))
- Earthquake data from September 18, 2018 - October 13, 2018 (obtained from the US Geological Survey (USGS) using the [USGS API](#))

Setup

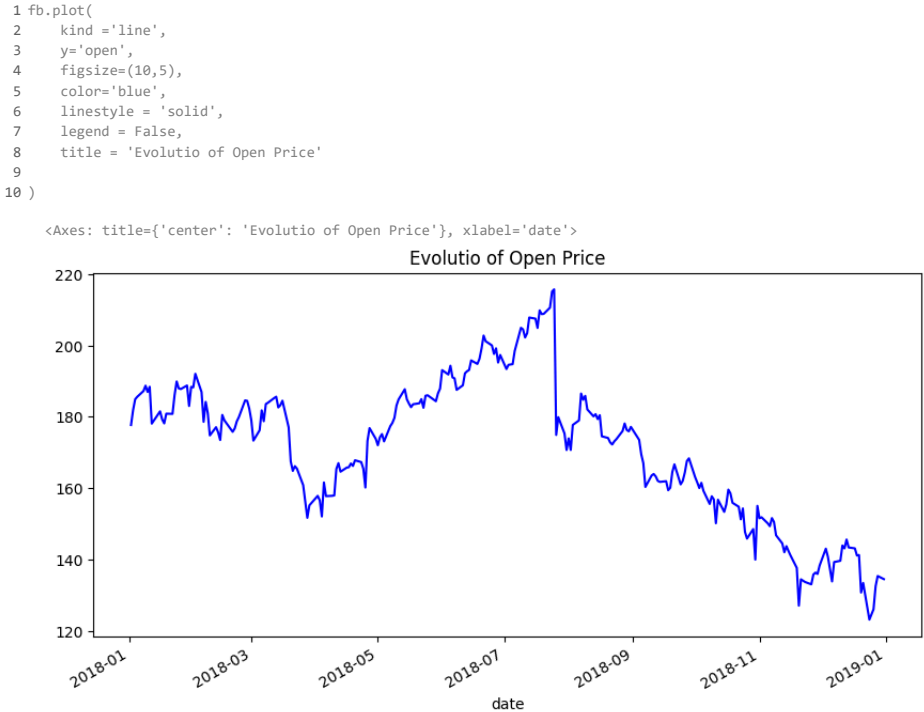
```
1 %matplotlib inline
2 import matplotlib.pyplot as plt
3 import numpy as np
4 import pandas as pd
5
6 fb = pd.read_csv('/content/drive/MyDrive/Module 9: Data Visualization using Pandas, Matplotlib and Seaborn/fb_stock_prices_2018.csv',
7                 index_col = 'date', parse_dates = True
8 )
9 quakes = pd.read_csv('/content/drive/MyDrive/Module 9: Data Visualization using Pandas, Matplotlib and Seaborn/earthquakes.csv')
```

Evolution over time

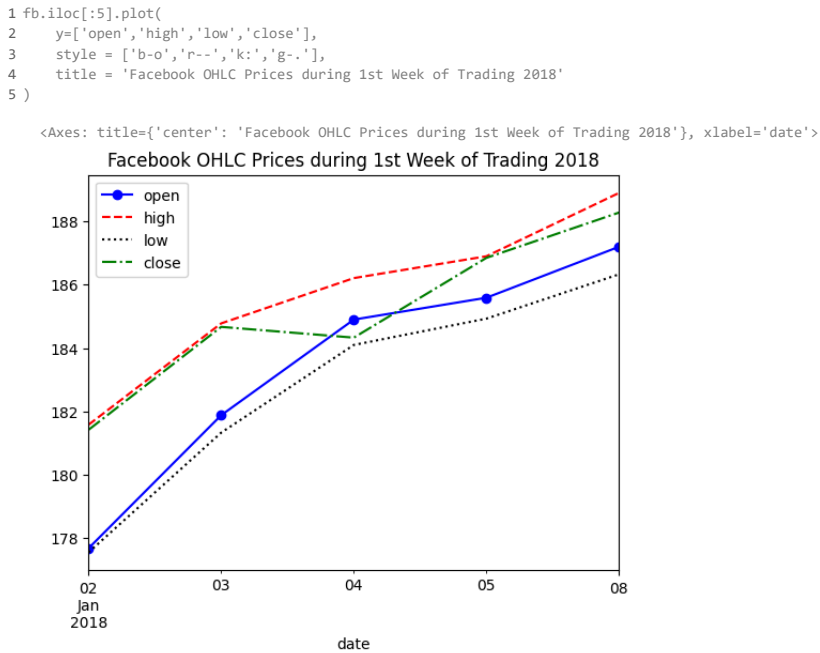
Line plots help us see how a variable changes over time. They are the default for the `kind` argument, but we can pass `kind='line'` to be explicit in our intent



We provided the `style` argument in the previous example; however, we can use the `color` and `linestyle` arguments to get the same result:



We can also plot many lines at once by simply passing a list of the columns to plot:



Creating subplots

When plotting with pandas, creating subplots is simply a matter of passing `subplots=True` to the `plot()` method, and (optionally) specifying the layout in a tuple of (rows, columns) :

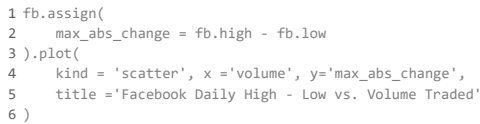


Note that we didn't provide a specific column to plot and pandas plotted all of them for us.

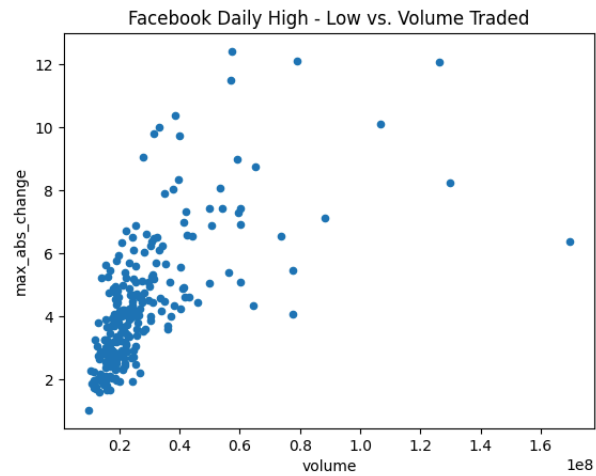
Visualizing relationship between variables

Scatter plots

We make scatter plots to help visualize the relationship between two variables. Creating scatter plots requires we pass in `kind='scatter'` along with a column for the x-axis and a column for the y-axis



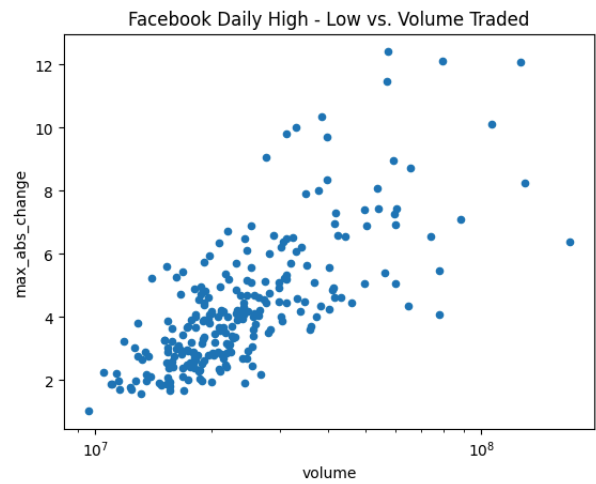
```
<Axes: title={'center': 'Facebook Daily High - Low vs. Volume Traded'}, xlabel='volume',
ylabel='max_abs_change'>
```



The relationship doesn't seem to be linear, but we can try a log transform on the x-axis since the scales of the axes are very different. With pandas, we simply pass in `logx=True` :

```
1 fb.assign(
2     max_abs_change = fb.high - fb.low
3 ).plot(
4     kind = 'scatter', x ='volume', y='max_abs_change',
5     title ='Facebook Daily High - Low vs. log(Volume Traded)',
6     logx= True
7 )

<Axes: title={'center': 'Facebook Daily High - Low vs. Volume Traded'}, xlabel='volume',
ylabel='max_abs_change'>
```



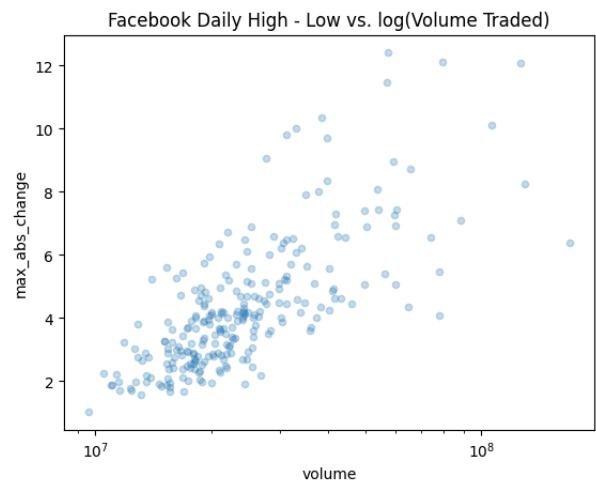
With matplotlib, we could use `plt.xscale('log')` to do the same thing.

▼ Adding Transparency to Plots with `alpha`

Sometimes our plots have many overlapping values, but this can be impossible to see. This can be addressed by increasing the transparency of what we are plotting using the `alpha` parameter. It is a float on `[0, 1]` where 0 is completely transparent and 1 is completely opaque. By default this is 1, so let's put in a lower value and re-plot the scatter plot:

```
1 fb.assign(
2     max_abs_change = fb.high - fb.low
3 ).plot(
4     kind = 'scatter', x ='volume', y='max_abs_change',
5     title ='Facebook Daily High - Low vs. log(Volume Traded)',
6     logx= True,alpha = 0.25
7 )

<Axes: title={'center': 'Facebook Daily High - Low vs. log(Volume Traded)'}, xlabel='volume',
ylabel='max_abs_change'>
```

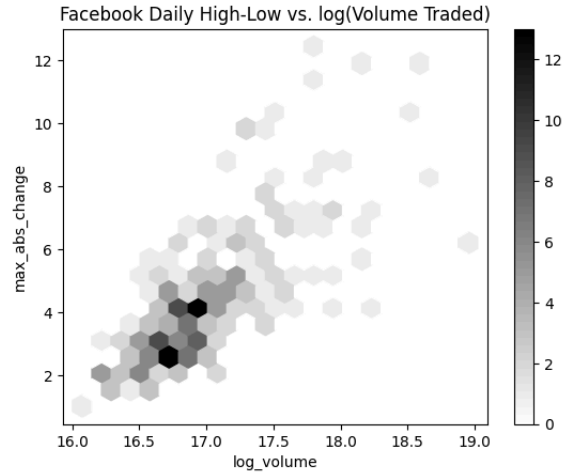


▼ Hexbins

In the previous example, we can start to see the overlaps, but it is still difficult. Hexbins are another plot type that divide up the plot into hexagons, which are shaded according to the density of points there. With pandas, this is the `hexbin` value for the `kind` argument. It can also be important to tweak the `gridsize` , which determines the number of hexagons along the y-axis:

```
1 fb.assign(
2     log_volume = np.log(fb.volume),
3     max_abs_change= fb.high - fb.low
4 ).plot(
5     kind = 'hexbin',
6     x = 'log_volume',
7     y = 'max_abs_change',
8     title = 'Facebook Daily High-Low vs. log(Volume Traded)',
9     colormap = 'gray_r',
10    gridsize = 20,
11    sharex = False # we have to pass this to see the x-axis due to a bug in this version of pandas
12 )
```

<Axes: title={'center': 'Facebook Daily High-Low vs. log(Volume Traded)'}, xlabel='log\_volume', ylabel='max\_abs\_change'>

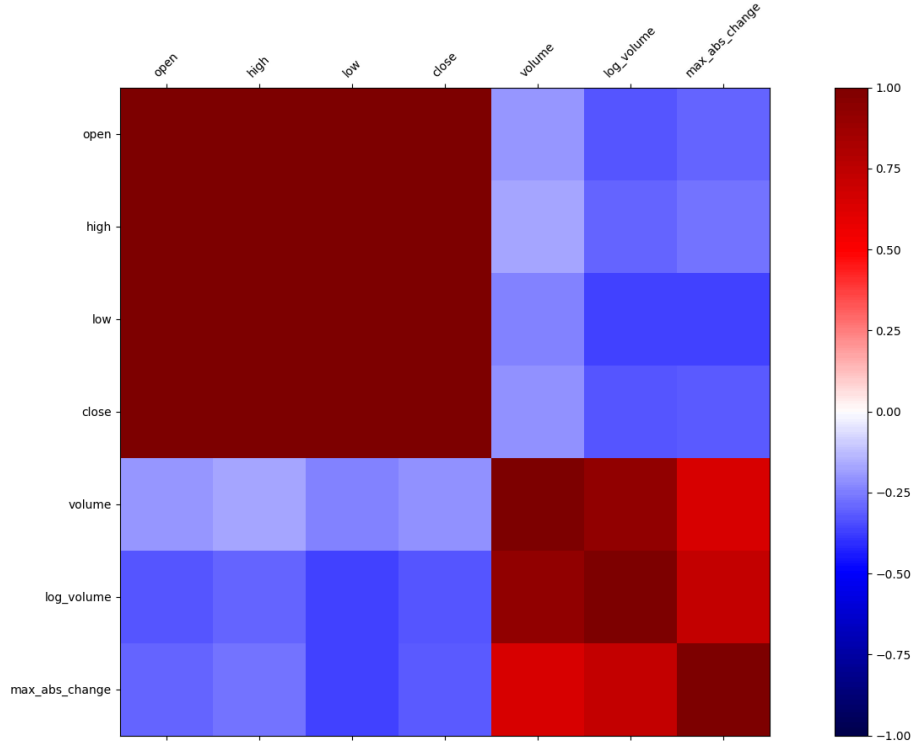


Visualizing Correlations with Heatmaps

Pandas doesn't offer heatmaps; however, if we are able to get our data into a matrix, we can use `matshow()` from `matplotlib`:

```
1 fig, ax =plt.subplots(figsize=(20,10))
2
3 fb_corr = fb.assign(
4     log_volume = np.log(fb.volume),
5     max_abs_change= fb.high - fb.low
6 ).corr()
7
8 im = ax.matshow(fb_corr, cmap='seismic')
9 fig.colorbar(im)
10
11 im.set_clim(-1,1)
12
13 labels = [col.lower() for col in fb_corr.columns]
14 ax.set_xticklabels(['']+ labels, rotation = 45)
15 ax.set_yticklabels(['']+labels)
16
```

<ipython-input-50-52d544446afb>:14: UserWarning: FixedFormatter should only be used together with FixedLocator  
ax.set\_xticklabels(['']+ labels, rotation = 45)  
<ipython-input-50-52d544446afb>:15: UserWarning: FixedFormatter should only be used together with FixedLocator  
ax.set\_yticklabels(['']+labels)  
[Text(0, -1.0, ''),  
Text(0, 0.0, 'open'),  
Text(0, 1.0, 'high'),  
Text(0, 2.0, 'low'),  
Text(0, 3.0, 'close'),  
Text(0, 4.0, 'volume'),  
Text(0, 5.0, 'log\_volume'),  
Text(0, 6.0, 'max\_abs\_change'),  
Text(0, 7.0, '')]



```
1 fb_corr.loc['max_abs_change',['volume', 'log_volume']]
```

```
volume      0.642027
log_volume  0.731542
Name: max_abs_change, dtype: float64
```

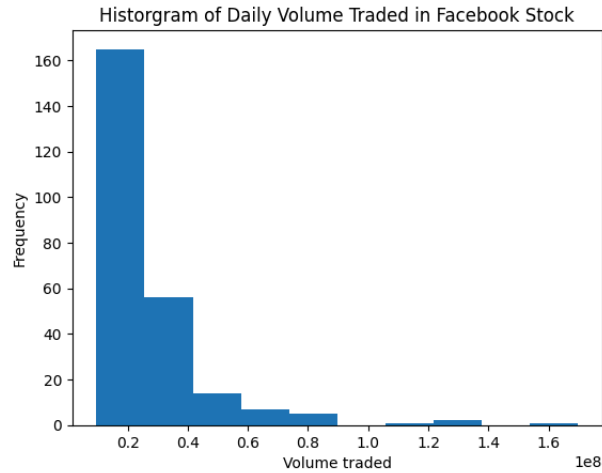
Visualizing distributions

Histograms

With the pandas `plot()` method, making histograms is as easy as passing in `kind='hist'` :

```
1 fb.volume.plot(
2     kind='hist',
3     title='Histogram of Daily Volume Traded in Facebook Stock'
4 )
5 plt.xlabel('Volume traded') #label  the x-axis (discussed in chapter 6)

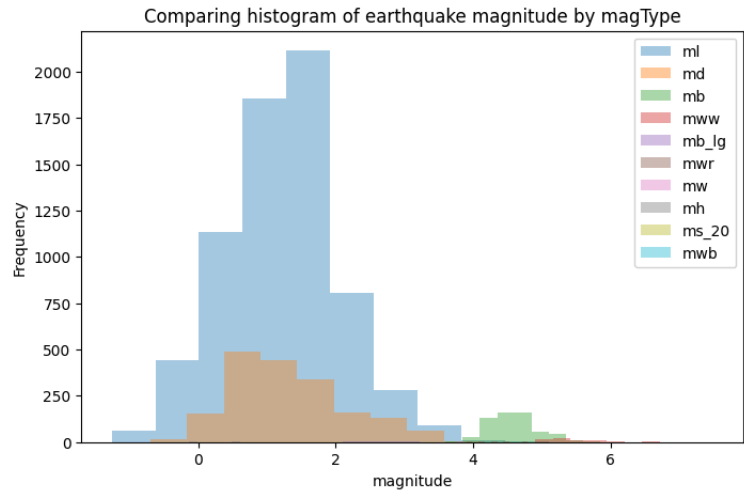
Text(0.5, 0, 'Volume traded')
```



We can overlap histograms to compare distributions provided we use the `alpha` parameter. For example, let's compare the usage and magnitude of the various `magTypes` in the data:

```
1 fig, axes = plt.subplots(figsize=(8,5))
2 for magtype in quakes.magType.unique():
3     data = quakes.query(f'magType == "{magtype}"').mag
4     if not data.empty:
5         data.plot(
6             kind='hist', ax = axes, alpha= 0.4,
7             label = magtype, legend = True,
8             title='Comparing histogram of earthquake magnitude by magType'
9         )
10 plt.xlabel('magnitude') #label the x-axis (discussed in chapter 6)

Text(0.5, 0, 'magnitude')
```

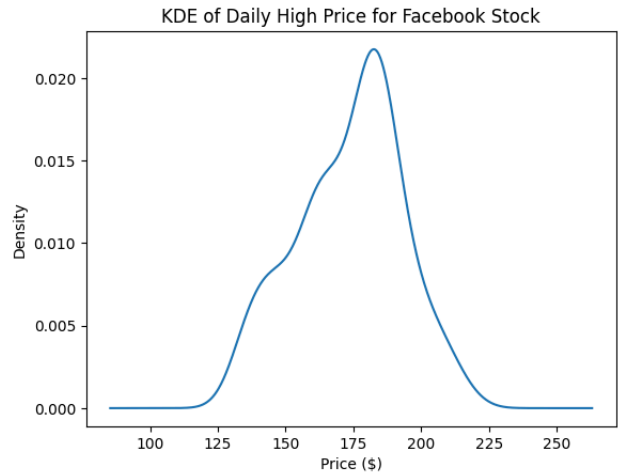


Kernel Density Estimation (KDE)

We can pass `kind='kde'` for a probability density function (PDF), which tells us the probability of getting a particular value

```
1 fb.high.plot(
2     kind = 'kde',
3     title = 'KDE of Daily High Price for Facebook Stock'
4 )
5 plt.xlabel('Price ($)') # label the axis  (discussed in chapter 6)

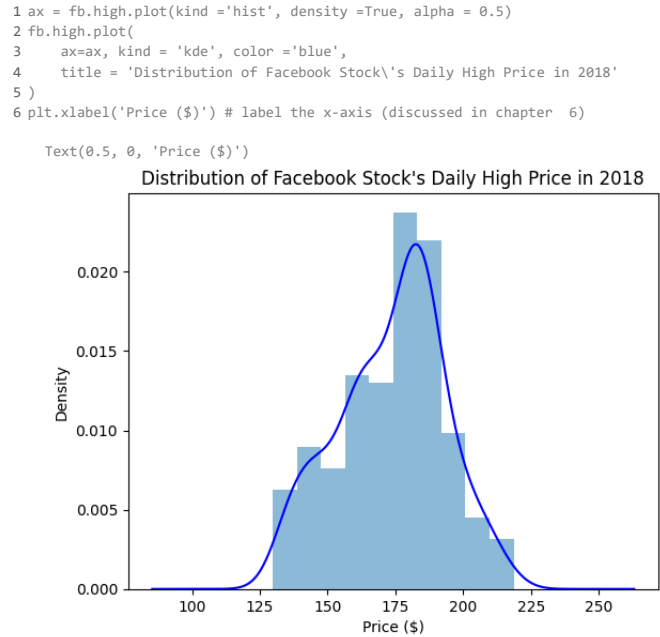
Text(0.5, 0, 'Price ($)')
```



▼ Adding to the result of plot()

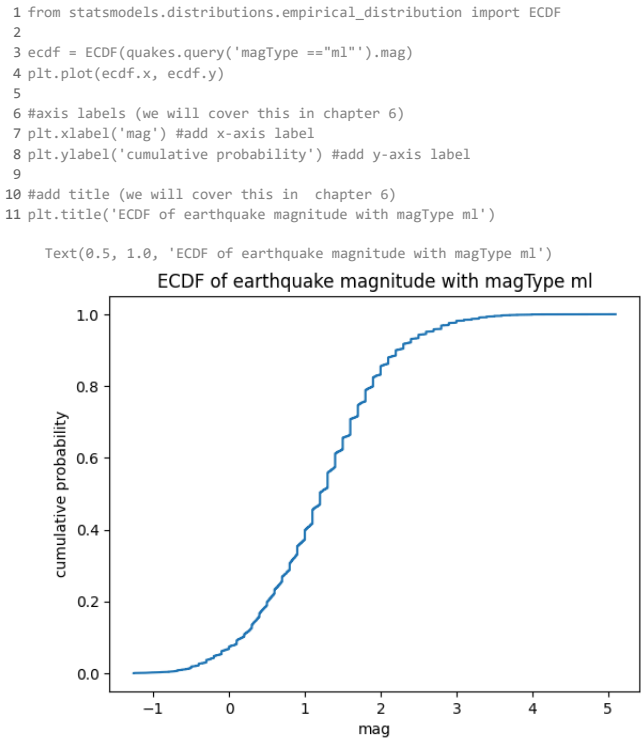
The `plot()` method returns a matplotlib `Axes` object. We can store this for additional customization of the plot, or we can pass this into another call to `plot()` as the `ax` argument to add to the original plot.

It can often be helpful to view the KDE superimposed on top of the histogram, which can be achieved with this strategy:

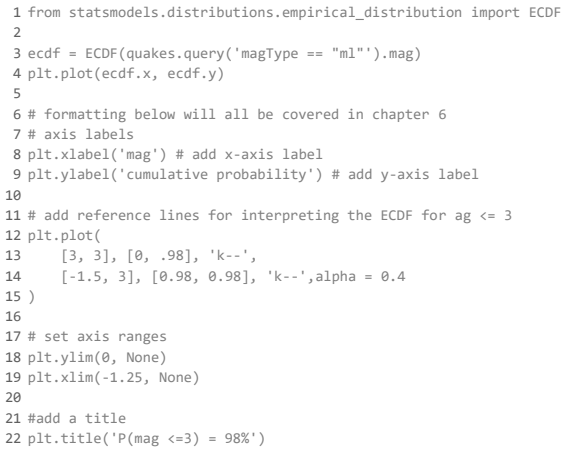


▼ Plotting the ECDF

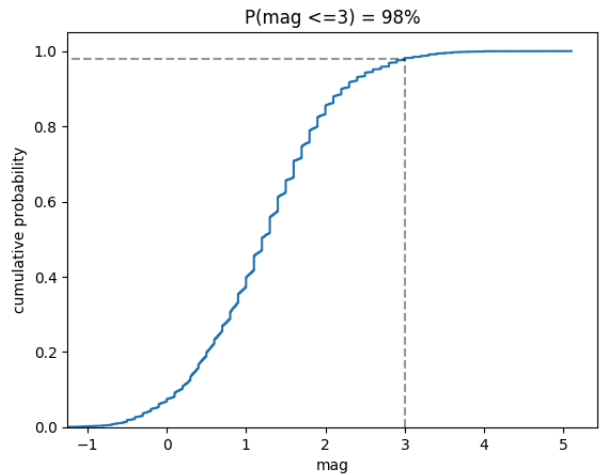
In some cases, we are more interested in the probability of getting less than or equal to that value (or greater than or equal), which we can see with the cumulative distribution function (CDF). Using the `statsmodels` package, we can estimate the CDF giving us the empirical cumulative distribution function (ECDF):



This ECDF tells us the probability of getting an earthquake with magnitude of 3 or less using the `ml` scale is 98%:



```
Text(0.5, 1.0, 'P(mag <=3) = 98%')
```

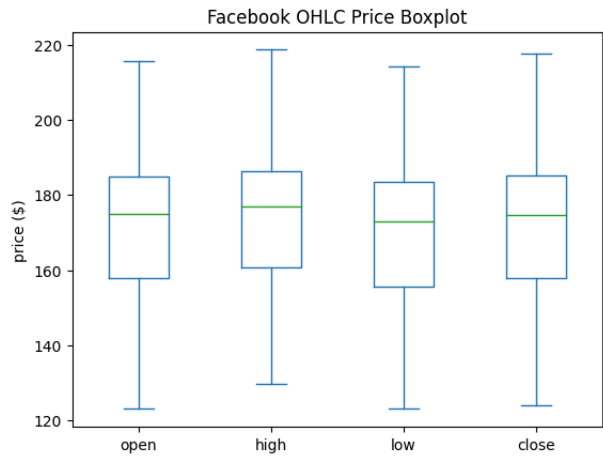


Box plots

To make box plots with pandas, we pass `kind='box'` to the `plot()` method:

```
1 fb.iloc[:,4].plot(kind='box', title='Facebook OHLC Price Boxplot')
2 plt.ylabel('price ($)') # label the y-axis (discussed in chapter 6)
```

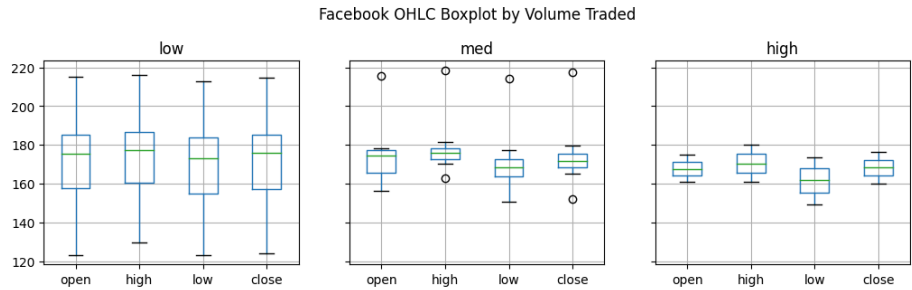
```
Text(0, 0.5, 'price ($)')
```



This can also be combined with a `groupby()`:

```
1 fb.assign(
2     volume_bin=pd.cut(fb.volume, 3, labels=['low', 'med', 'high'])
3 ).groupby('volume_bin').boxplot(
4     column = ['open', 'high', 'low', 'close'],
5     layout = (1,3), figsize=(12,3)
6 )
7 plt.suptitle('Facebook OHLC Boxplot by Volume Traded', y = 1.1)
```

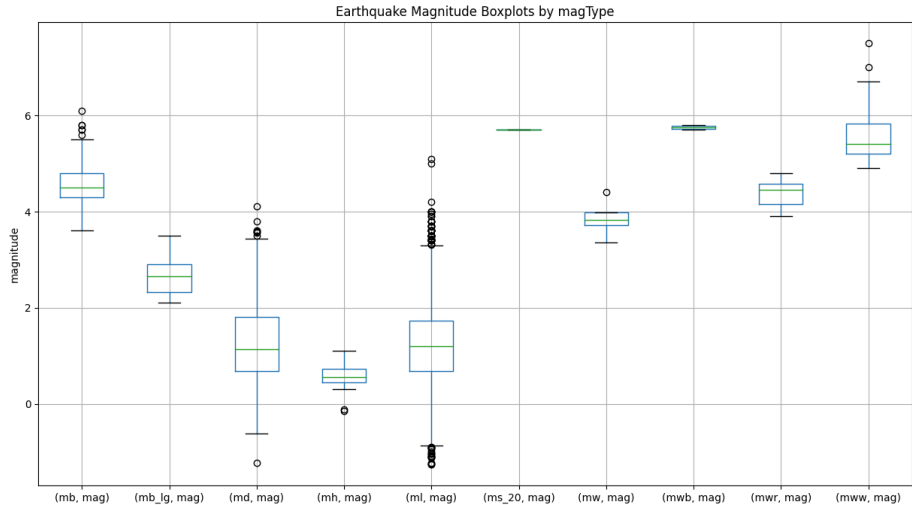
```
Text(0.5, 1.1, 'Facebook OHLC Boxplot by Volume Traded')
```



We can use this to see the distribution of magnitudes across the different measurement methods for earthquakes:

```
1 quakes[['mag', 'magType']].groupby('magType').boxplot(
2     figsize=(15, 8), subplots=False
3 )
4 plt.title('Earthquake Magnitude Boxplots by magType')
5 plt.ylabel('magnitude') # label the y-axis (discussed in chapter 6)
```

Text(0, 0.5, 'magnitude')



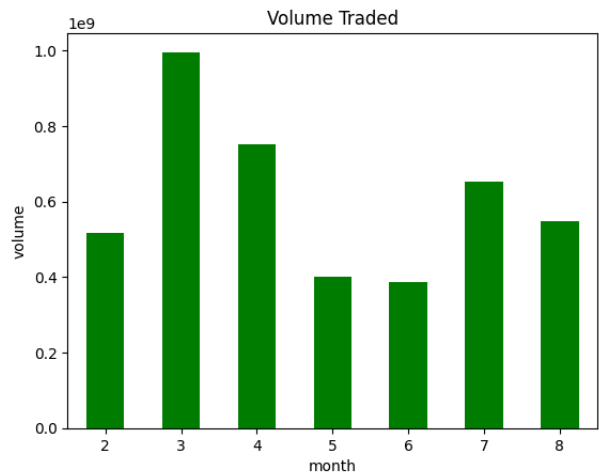
Counts and frequencies

Bar Charts

With pandas, we have the option of using the `kind` argument or using `plot.<kind>()` . Let's use `plot.bar()` here to show the evolution of monthly volume traded in Facebook stock over time:

```
1 fb['2018-02':'2018-08'].assign(  
2     month=lambda x: x.index.month  
3 ).groupby('month').sum().volume.plot.bar(  
4     color='green', rot=0, title='Volume Traded'  
5 )  
6 plt.ylabel('volume') # label the y-axis (discussed in chapter 6)
```

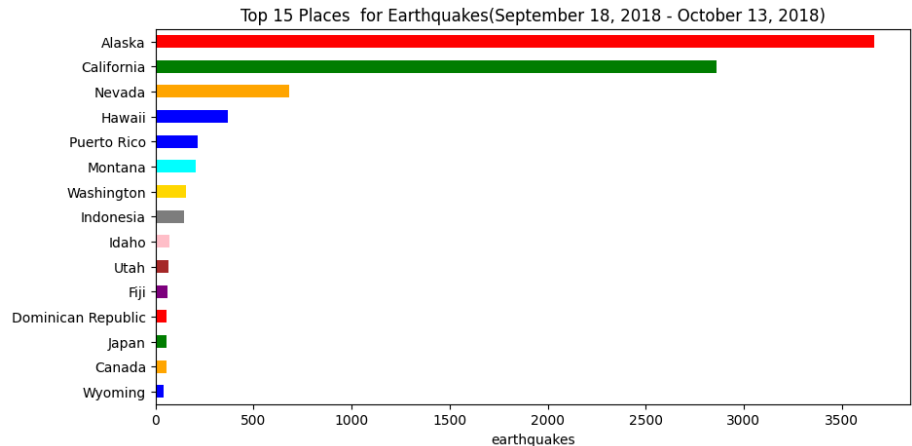
Text(0, 0.5, 'volume')



We can also change the orientation of the bars. Passing `kind='barh'` gives us horizontal bars instead of vertical ones. Let's use this to look at the top 15 places for earthquakes in our data:

```
1 bar_color = [ 'blue','orange','green','red','purple','brown','pink','gray','gold','cyan','blue']  
2 quakes.parsed_place.value_counts().iloc[14::-1].plot(  
3     kind='barh', figsize=(10,5), color = bar_color,  
4     title = 'Top 15 Places for Earthquakes'\n5     '(September 18, 2018 - October 13, 2018)'  
6 )  
7  
8 plt.xlabel('earthquakes') #label the x-axis (discussed in chapter 6)
```

Text(0.5, 0, 'earthquakes')



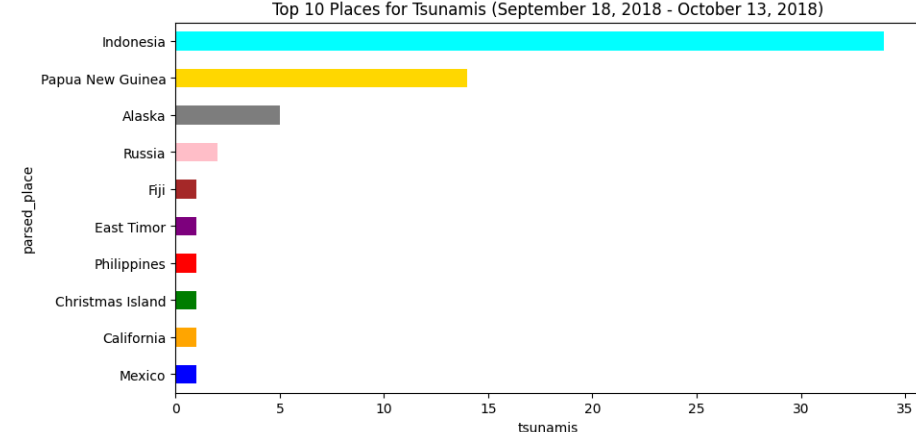
We also have data on whether earthquakes were accompanied by tsunamis. Let's see what the top places for tsunamis are:



```
1 quakes.groupby('parsed_place').tsunami.sum().sort_values().iloc[-10:],].plot(
2     kind='barh', figsize=(10,5),color = bar_color,
3     title='Top 10 Places for Tsunamis '\
4         '(September 18, 2018 - October 13, 2018)'
5 )
6 plt.xlabel('tsunamis') #label the x-axis (discussed in chapter 6)

Text(0.5, 0, 'tsunamis')
```

Top 10 Places for Tsunamis (September 18, 2018 - October 13, 2018)



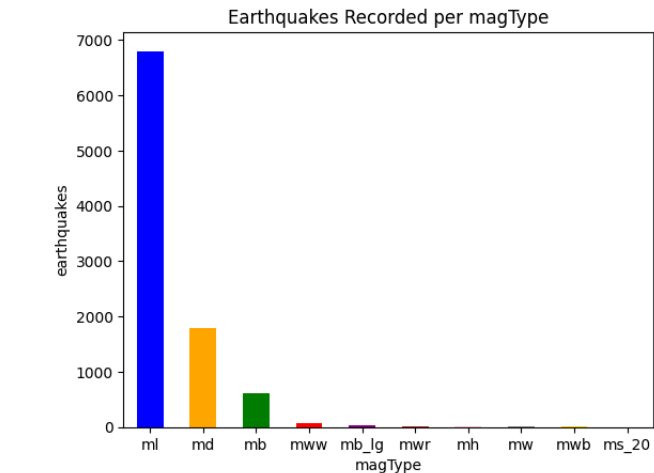
parsed_place	tsunamis
Indonesia	34
Papua New Guinea	14
Alaska	5
Russia	2
Fiji	1
East Timor	1
Philippines	1
Christmas Island	1
California	1
Mexico	1

```
1 indonesia_quakes = quakes.query('parsed_place == "Indonesia"').assign(
2     time=lambda x: pd.to_datetime(x.time, unit='ms'),
3     earthquake=1
4 ).set_index('time').resample('1D').sum()
5
6 indonesia_quakes.index = indonesia_quakes.index.strftime('%b\n%d')
7
8 indonesia_quakes.plot(
9     y=['earthquake', 'tsunami'], kind='bar', figsize=(15, 3), rot=0,
10    label=['earthquakes', 'tsunamis'],
11    title='Earthquakes and Tsunamis in Indonesia '\
12        '(September 18, 2018 - October 13, 2018)'
13 )
14
15 # label the axes (discussed in chapter 6)
16 plt.xlabel('date')
17 plt.ylabel('count')
```

<ipython-input-72-5d51f544148a>:4: FutureWarning: The default value of numeric\_only in DataFrameGroupBy.sum is d
).set\_index('time').resample('1D').sum()
Text(0, 0.5, 'count')

Using the kind argumnet for vertical bars when the labels for each bar are shorter:

```
1 quakes.magType.value_counts().plot(
2     kind='bar', title='Earthquakes Recorded per magType', rot=0,color =bar_color
3 )
4
5 # label the axes (discussed in chapter 6)
6 plt.xlabel('magType')
7 plt.ylabel('earthquakes')
8
Text(0, 0.5, 'earthquakes')
```



Double-click (or enter) to edit

Top 4 places with earthquakes:

```
Text(0, 0.5, 'earthquakes')
```



Text(0, 0.5, 'earthquakes')



```
1 normalized_pivot.pivot.fillna(0).apply(lambda x: x/x.sum(), axis=1)
2 ax = normalized_pivot.plot.bar(
3     stacked=True, rot=0, figsize=(10, 5),
4     title='Percentage of earthquakes by integer magnitude for each magType'
5 )
6 ax.legend(bbox_to_anchor=(1, 0.8)) # move legend to the right of the plot
7 plt.ylabel('percentage') # label the axes (discussed in chapter 6)
```

```
➡ Text(0, 0.5, 'percentage')
```



Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit