Plotting with Pandas

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

In [2]:

```
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

fb = pd.read_csv('fb_stock_prices_2018.csv', index_col='date', parse_dates=True)
quakes = pd.read_csv('earthquakes-1.csv')
```

Evolution over time

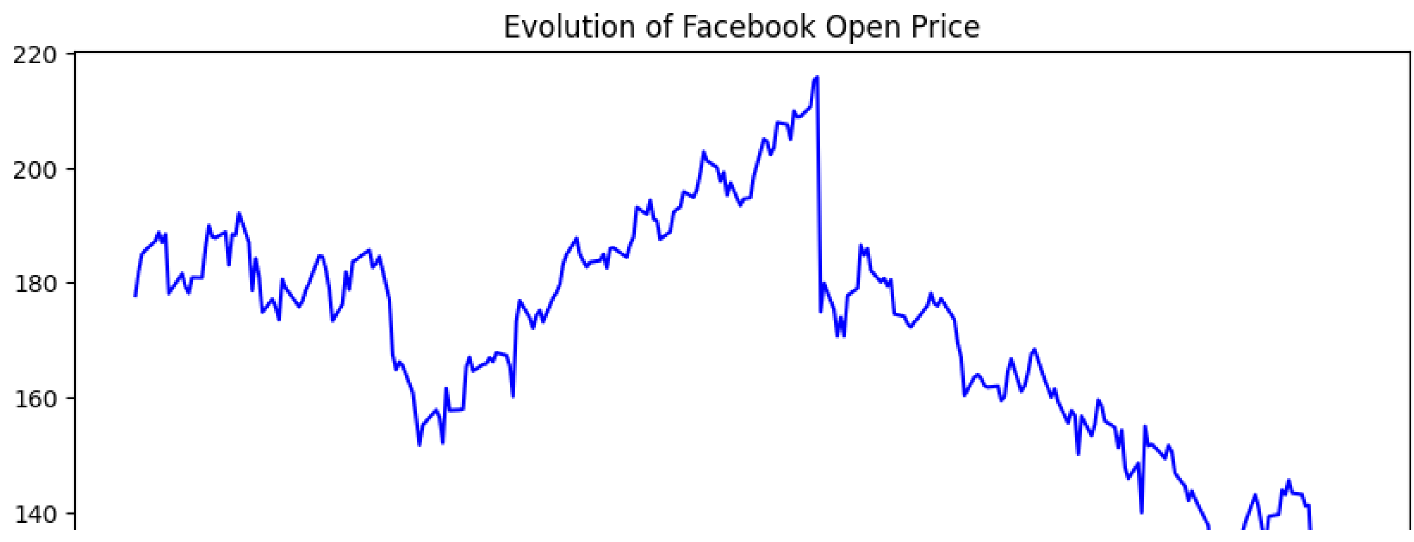
Line plots help us see how a variable changes over time. They are the default for the kind argument, but we can pass kind='line' to be explicit in our intent:

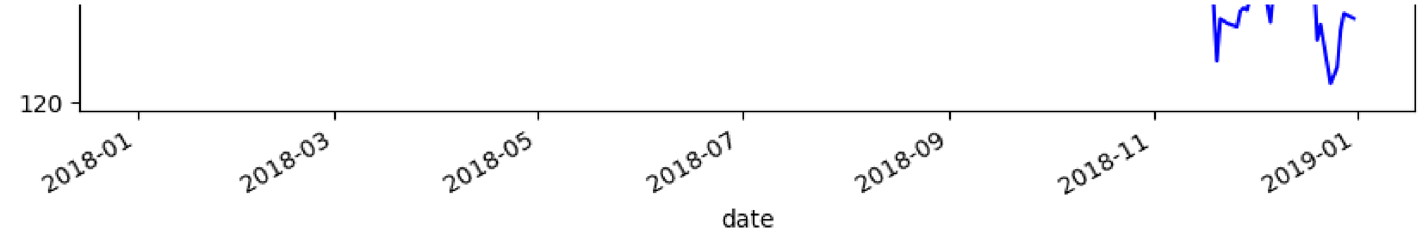
In [3]:

```
fb.plot(
  kind='line',
  y='open',
  figsize=(10, 5),
  style='b-',
  legend=False,
  title='Evolution of Facebook Open Price'
)
```

Out[3]:

```
<Axes: title={'center': 'Evolution of Facebook Open Price'}, xlabel='date'>
```





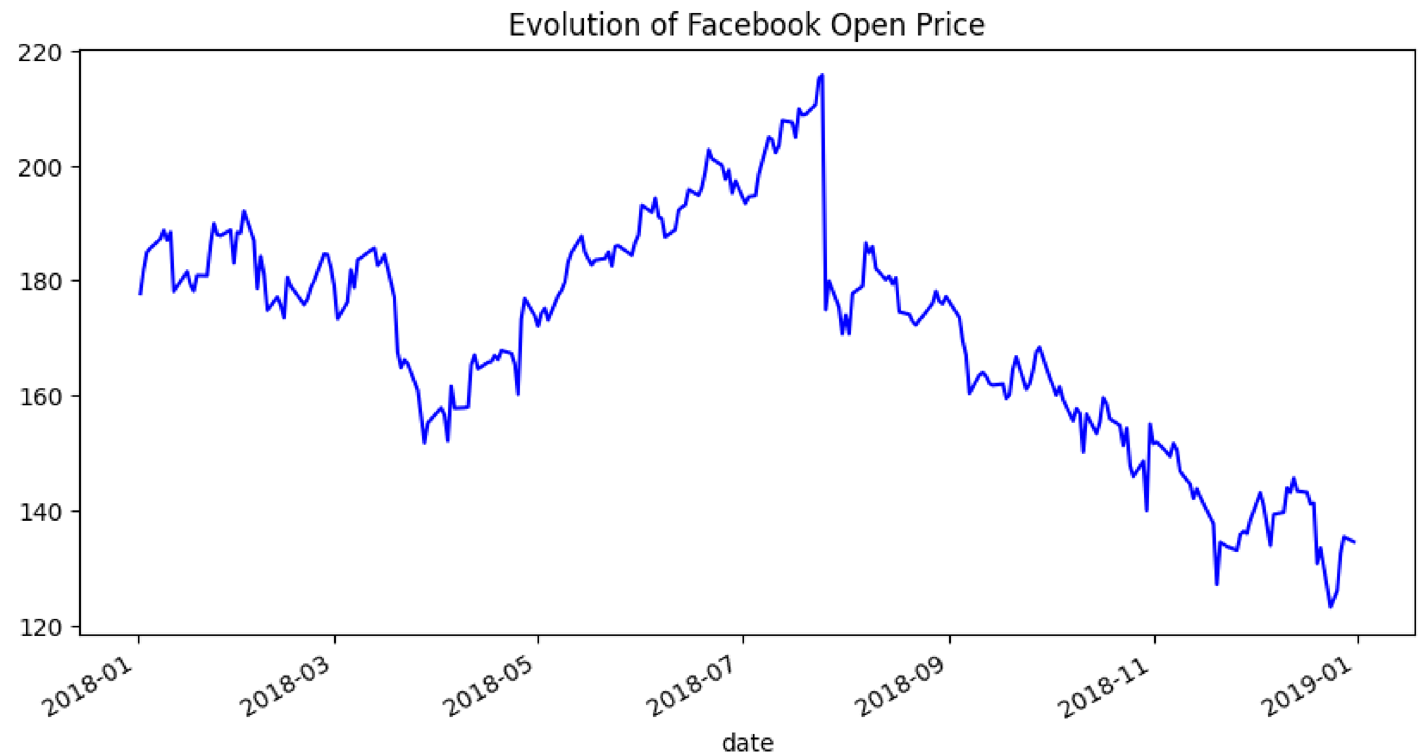
We provided the style argument in the previous example; however, we can use the color and linestyle arguments to get the same result:

In [4]:

```
fb.plot(
  kind='line',
  y='open',
  figsize=(10, 5),
  color='blue',
  linestyle='solid',
  legend=False,
  title='Evolution of Facebook Open Price'
)
```

Out[4]:

```
<Axes: title={'center': 'Evolution of Facebook Open Price'}, xlabel='date'>
```



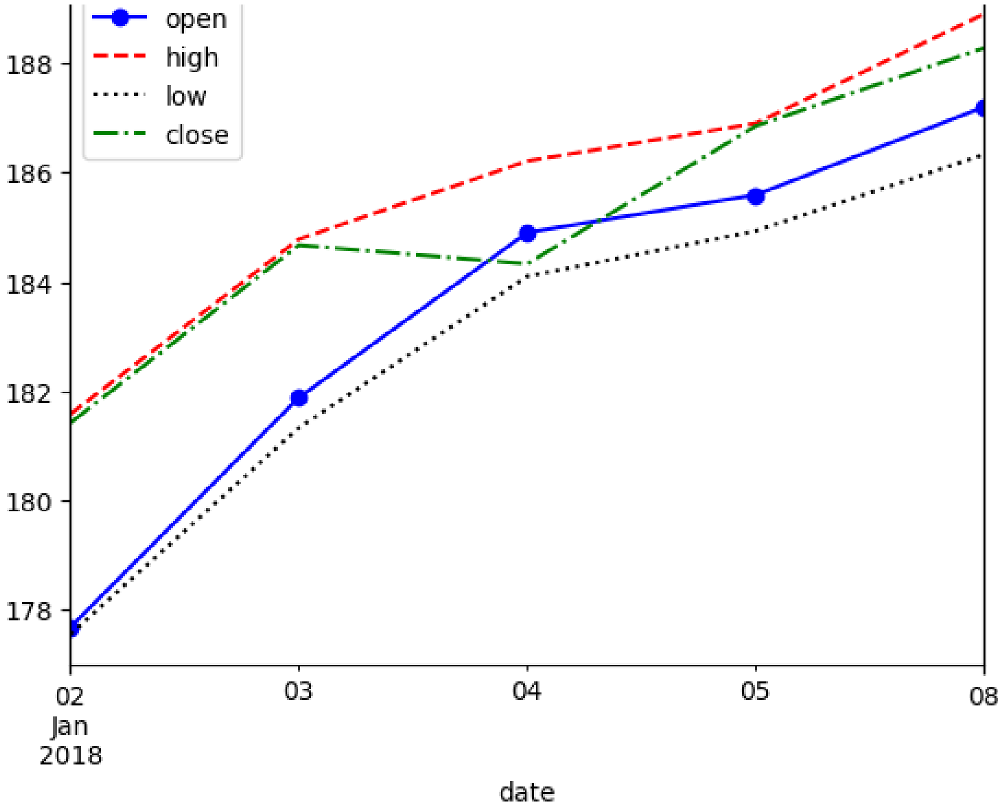
We can also plot many lines at once by simply passing a list of the columns to plot:

In [5]:

```
fb.iloc[:5,].plot(
  y=['open', 'high', 'low', 'close'],
  style=['b-o', 'r--', 'k:', 'g-.'],
  title='Facebook OHLC Prices during 1st Week of Trading 2018'
)
```

Out[5]:

```
<Axes: title={'center': 'Facebook OHLC Prices during 1st Week of Trading 2018'}, xlabel='
date'>
```



Creating subplots

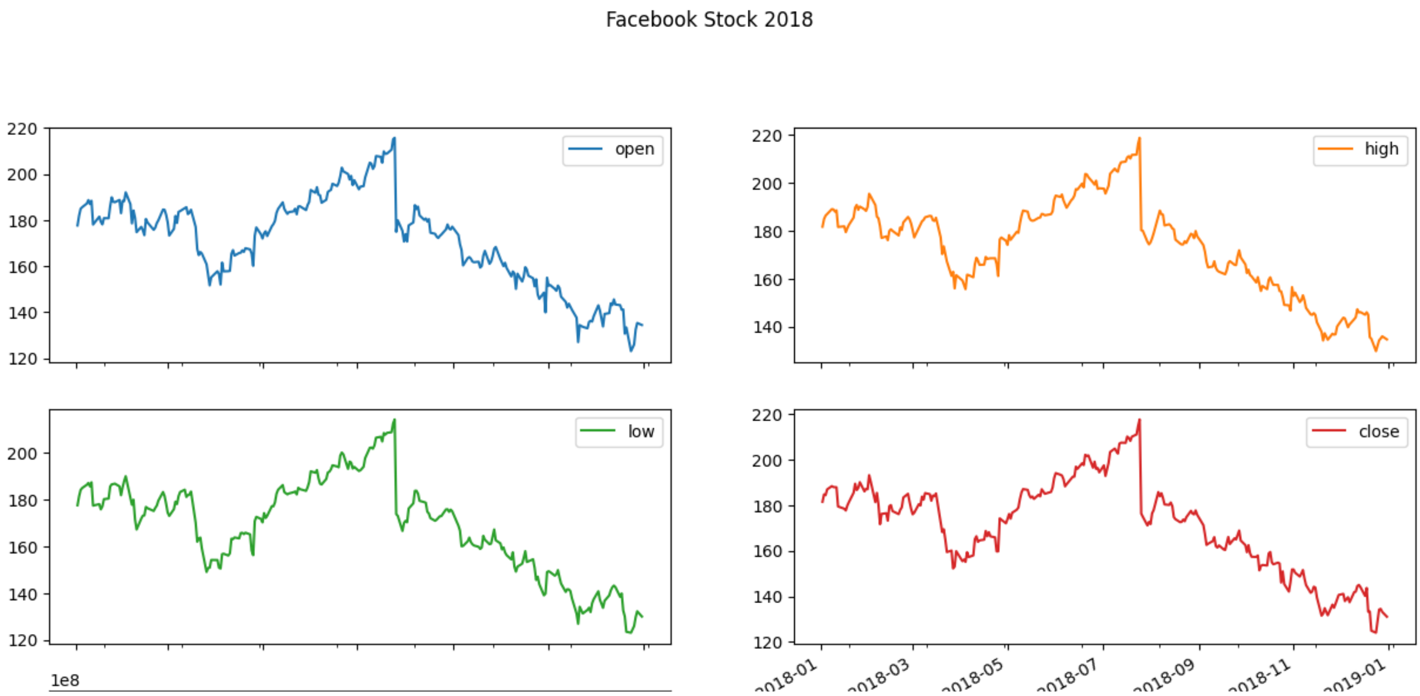
When plotting with pandas, creating subplots is simply a matter of passing subplots=True to the plot() method, and (optionally) specifying the layout in a tuple of (rows, columns)

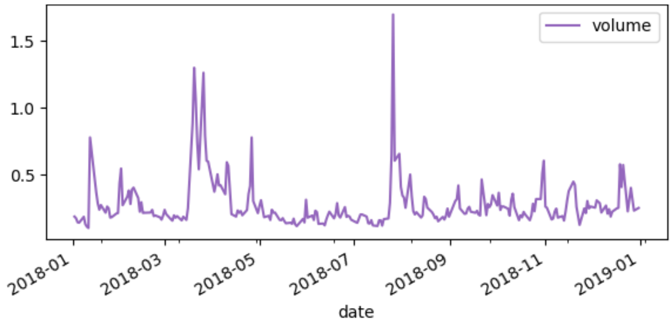
In [6]:

```
fb.plot(
  kind='line',
  subplots=True,
  layout=(3,2),
  figsize=(15,10),
  title='Facebook Stock 2018'
)
```

Out[6]:

```
array([[<Axes: xlabel='date'>, <Axes: xlabel='date'>],
       [<Axes: xlabel='date'>, <Axes: xlabel='date'>],
       [<Axes: xlabel='date'>, <Axes: xlabel='date'>]], dtype=object)
```





date

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

Visualizing relationships between variables

Scatter plots

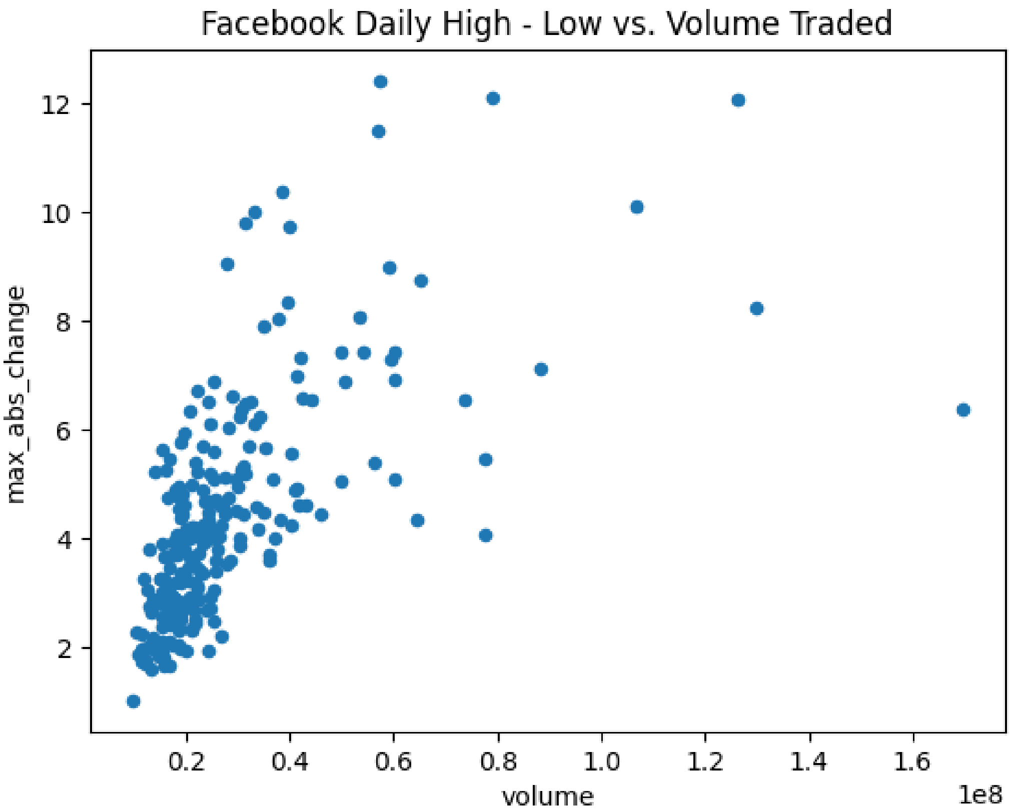
We make scatter plots to help visualize the relationship between two variables. Creating scatter plots requires we pass in kind='scatter' along with a column for the x- axis and a column for the y-axis:

In [7]:

```
fb.assign(
  max_abs_change=fb.high - fb.low
).plot(
  kind='scatter', x='volume', y='max_abs_change',
  title='Facebook Daily High - Low vs. Volume Traded'
)
```

Out[7]:

```
<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 :

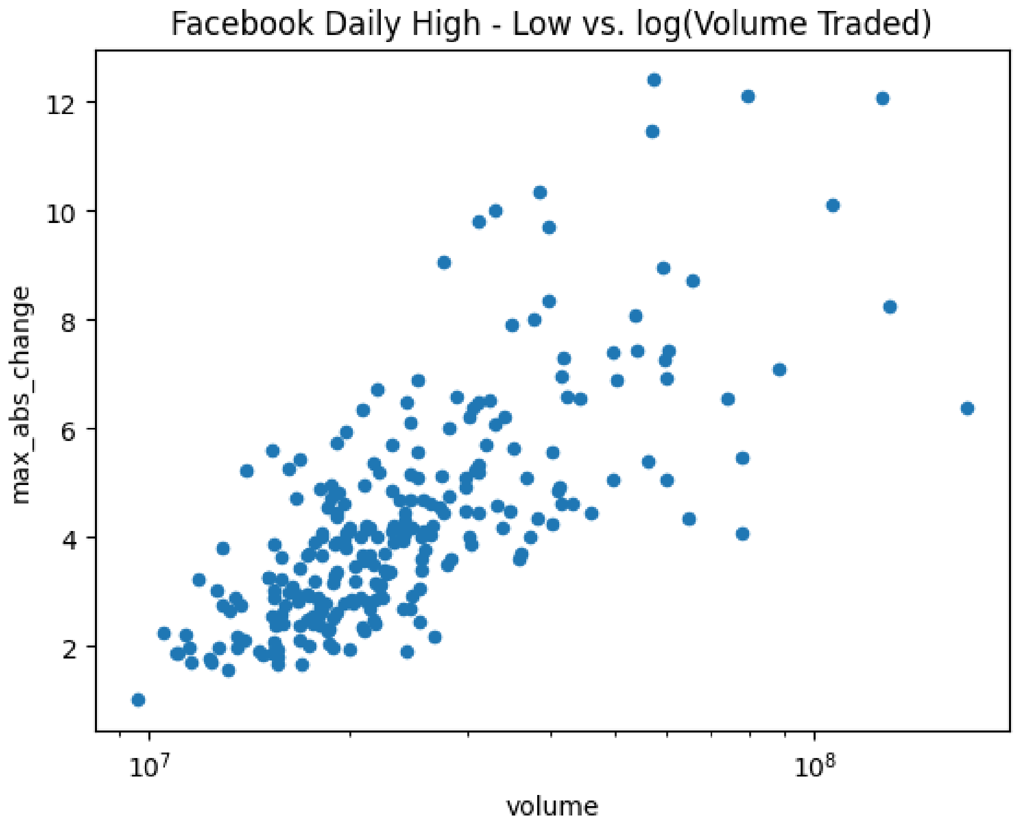
In [8]:

```
fb.assign(
```

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

Out[8]:

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



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

Adding Transparency to Plots with alpha

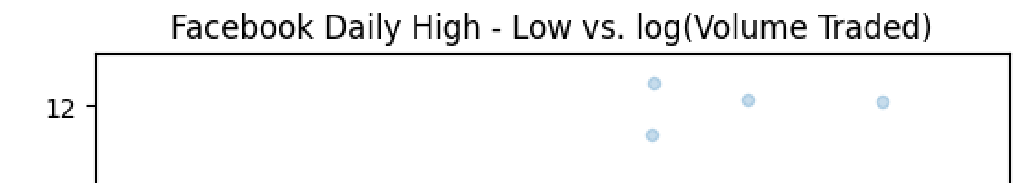
Sometimes our plots have many overlapping values, but this can be impossible to see. This can be addressed by increasing the transparency of what we are plotting using the alpha parameter. It is a float on [0, 1] where 0 is completely transparent and 1 is completely opaque. By default this is 1, so let's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

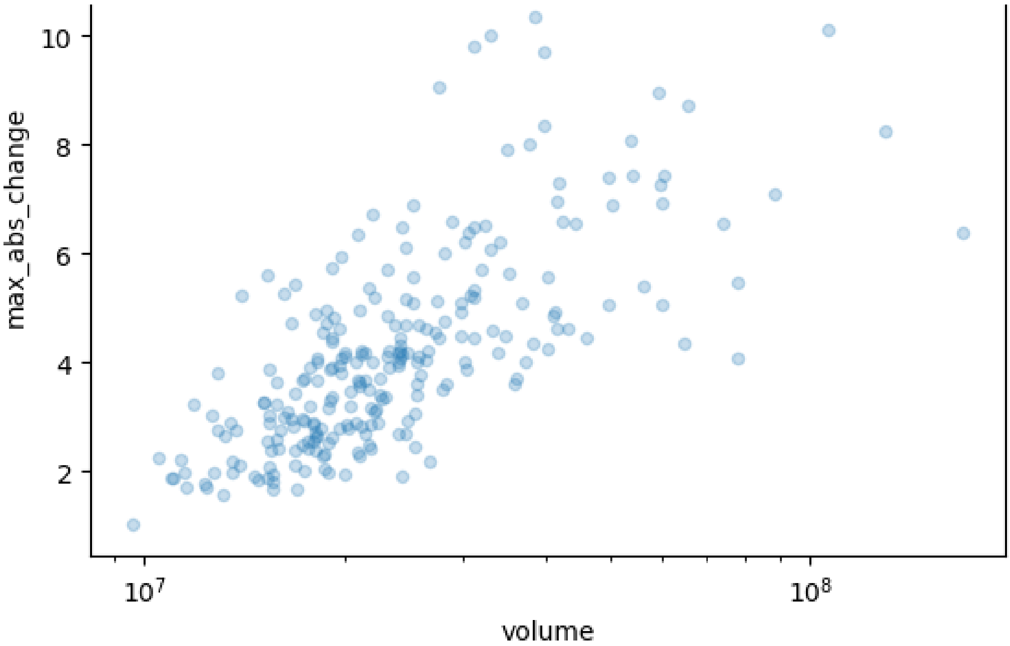
In [9]:

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

Out[9]:

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





Hexbins

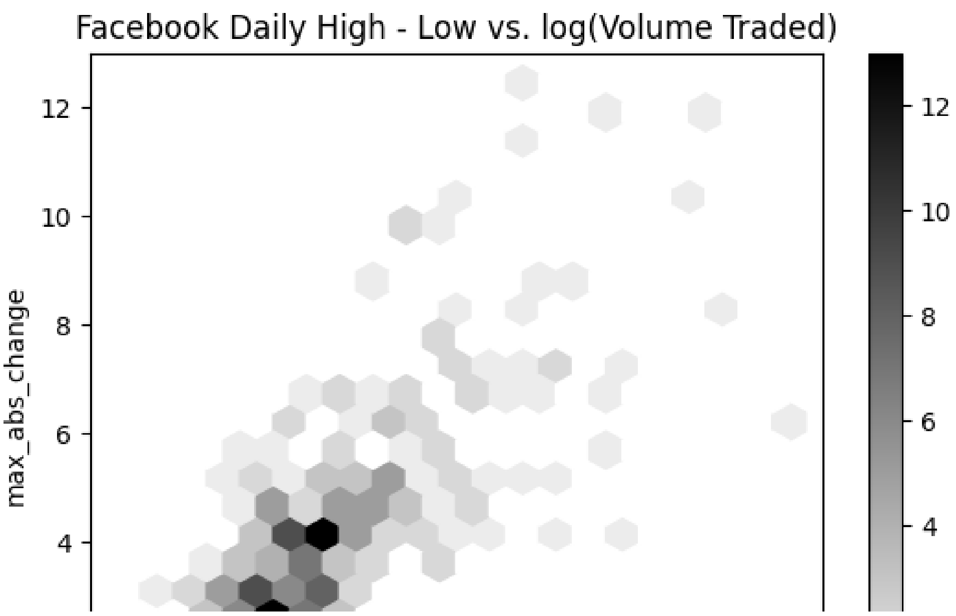
In the previous example, we can start to see the overlaps, but it is still difficult. Hexbins are another plot type that divide up the plot into hexagons, which are shaded according to the density of points there. With pandas, this is the hexbin value for the kind argument. It can also be important to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

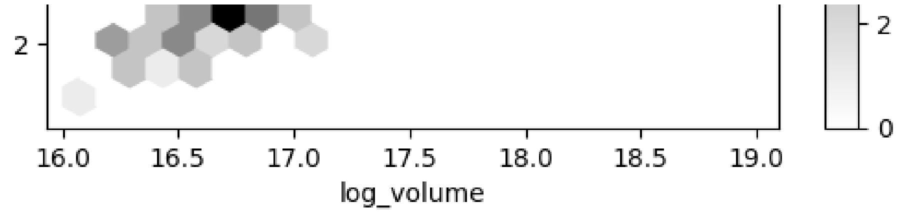
In [10]:

```
fb.assign(
log_volume=np.log(fb.volume),
max_abs_change=fb.high - fb.low
).plot(
kind='hexbin',
x='log_volume',
y='max_abs_change',
title='Facebook Daily High - Low vs. log(Volume Traded)',
colormap='gray_r',
gridsize=20,
sharex=False # we have to pass this to see the x-axis due to a bug in this version of pan
das
)
```

Out[10]:

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





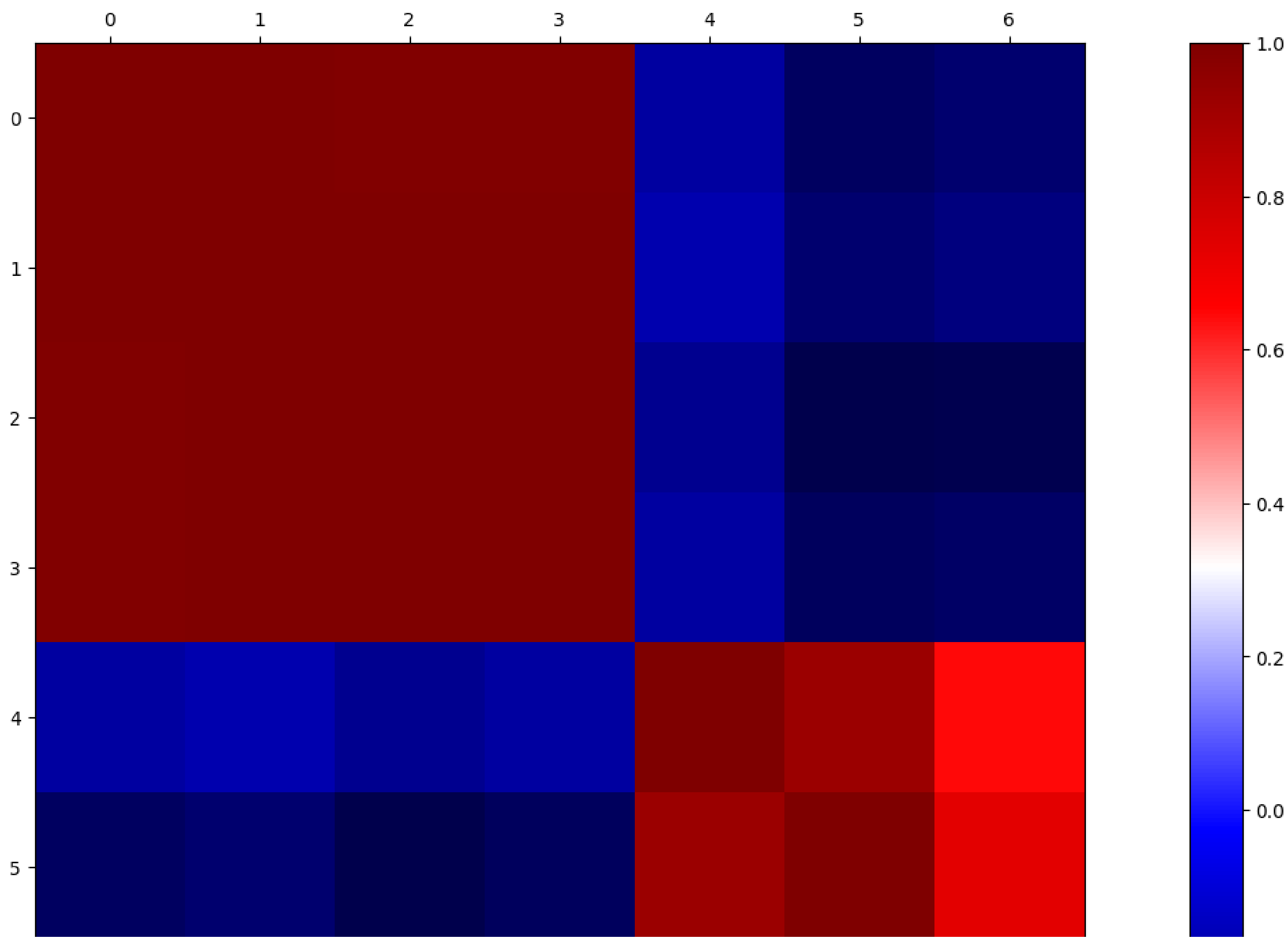
Visualizing Correlations with Heatmaps

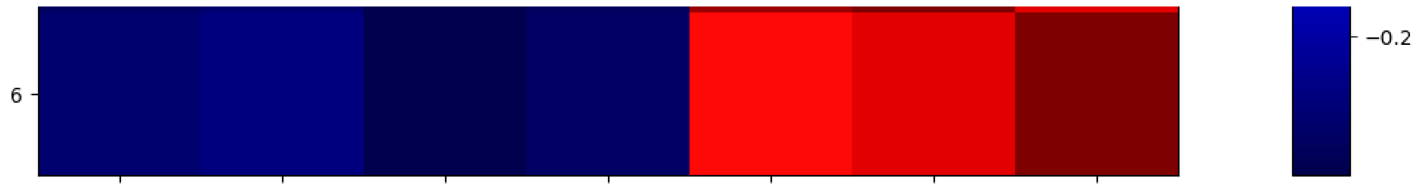
Pandas doesn't offer heatmaps; however, if we are able to get our data into a matrix, we can use matshow() from matplotlib:

In [13]:

```
fig, ax = plt.subplots(figsize=(20, 10))
fb_corr = fb.assign(
log_volume=np.log(fb.volume),
max_abs_change=fb.high - fb.low
).corr()
im = ax.matshow(fb_corr, cmap='seismic')
fig.colorbar(im).set_clim(-1, 1)
labels = [col.lower() for col in fb_corr.columns]
ax.set_xticklabels([''] + labels, rotation=45)
ax.set_yticklabels([''] + labels)
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-13-247f2a08d0d6> in <cell line: 0>()
      5 ).corr()
      6 im = ax.matshow(fb_corr, cmap='seismic')
----> 7 fig.colorbar(im).set_clim(-1, 1)
      8 labels = [col.lower() for col in fb_corr.columns]
      9 ax.set_xticklabels([''] + labels, rotation=45)

AttributeError: 'Colorbar' object has no attribute 'set_clim'
```





In [14]:

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

Out[14]:

	max_abs_change
volume	0.642027
log_volume	0.731542

dtype: float64

Visualizing distributions

Histograms

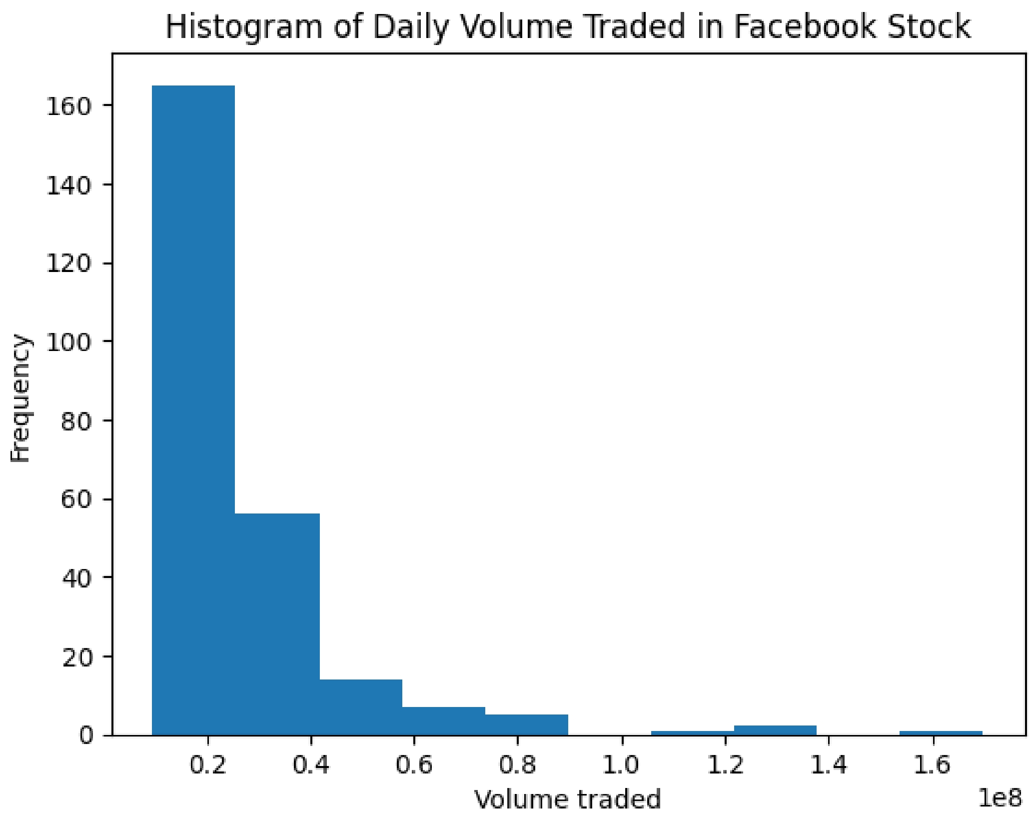
With the pandas plot() method, making histograms is as easy as passing in kind='hist' :

In [15]:

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

Out[15]:

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



We can overlap histograms to compare distributions provided we use the alpha parameter. For example, let's

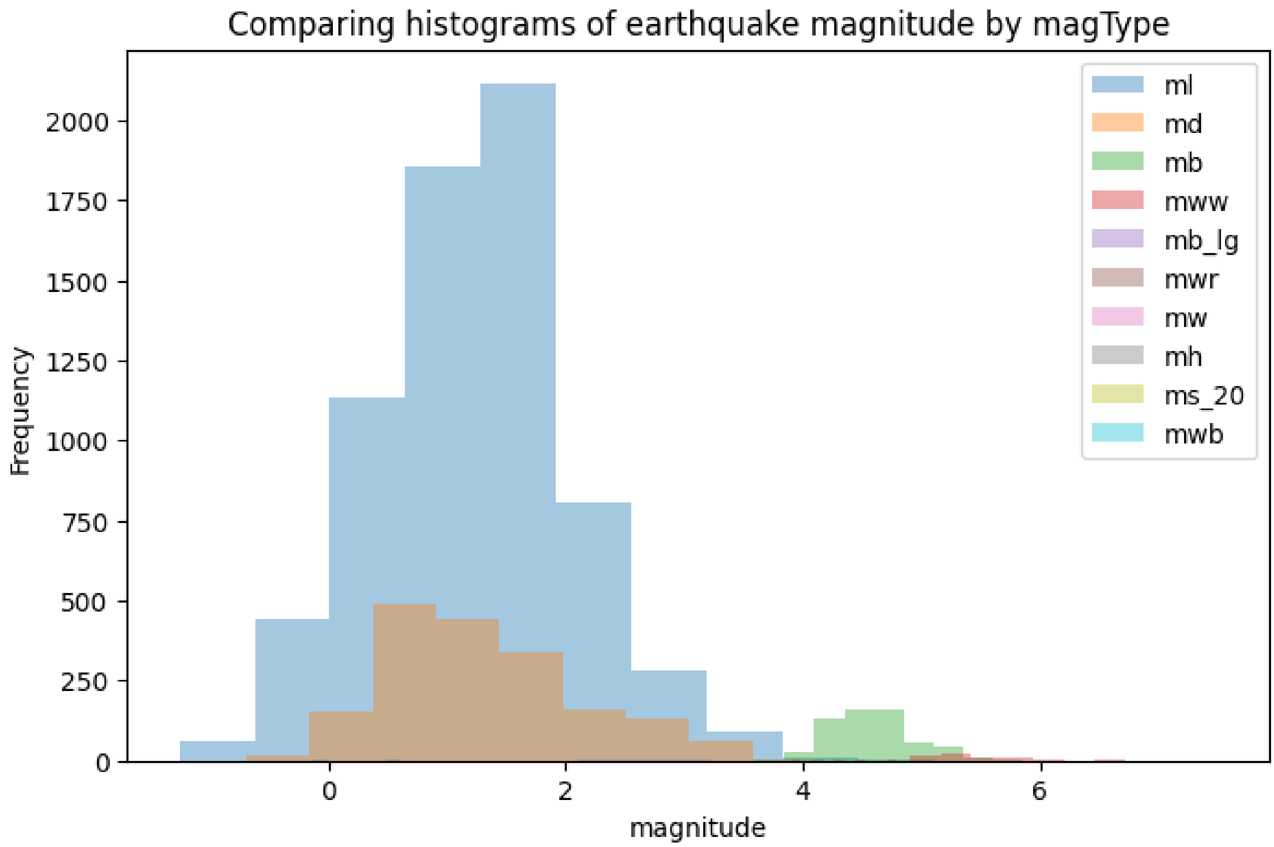
compare the usage and magnitude of the various magTypes in the data:

In [16]:

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

Out[16]:

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



Kernel Density Estimation (KDE)

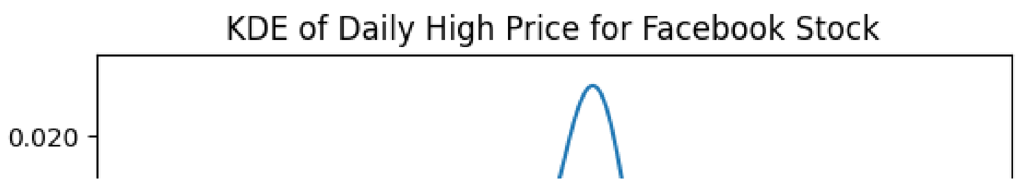
We can pass kind='kde' for a probability density function (PDF), which tells us the probability of getting a particular value:

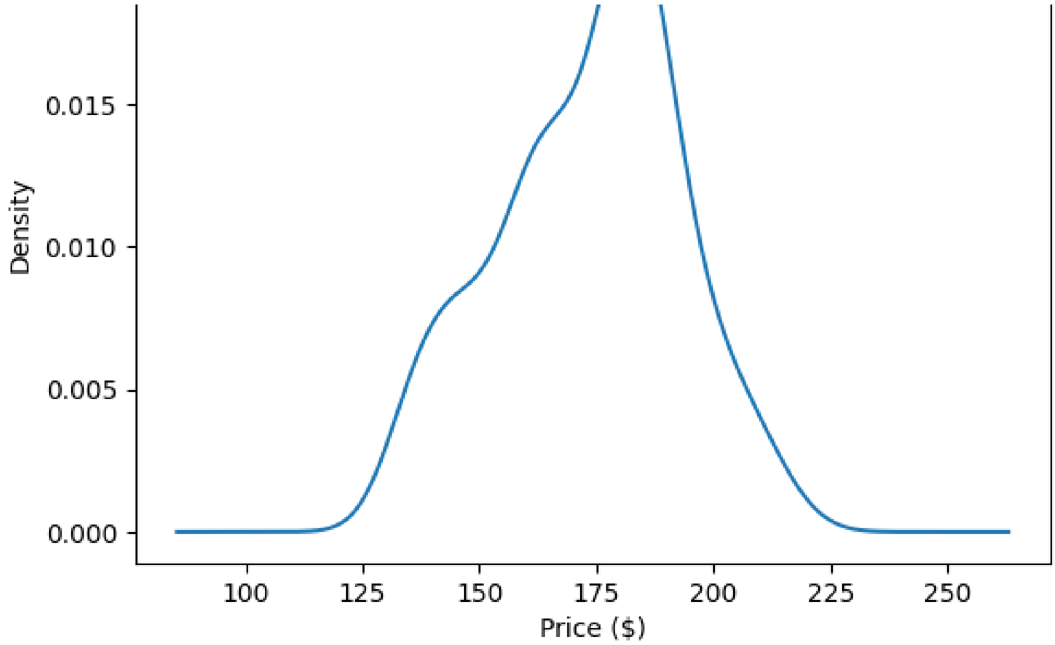
In [17]:

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

Out[17]:

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





Adding to the result of plot()

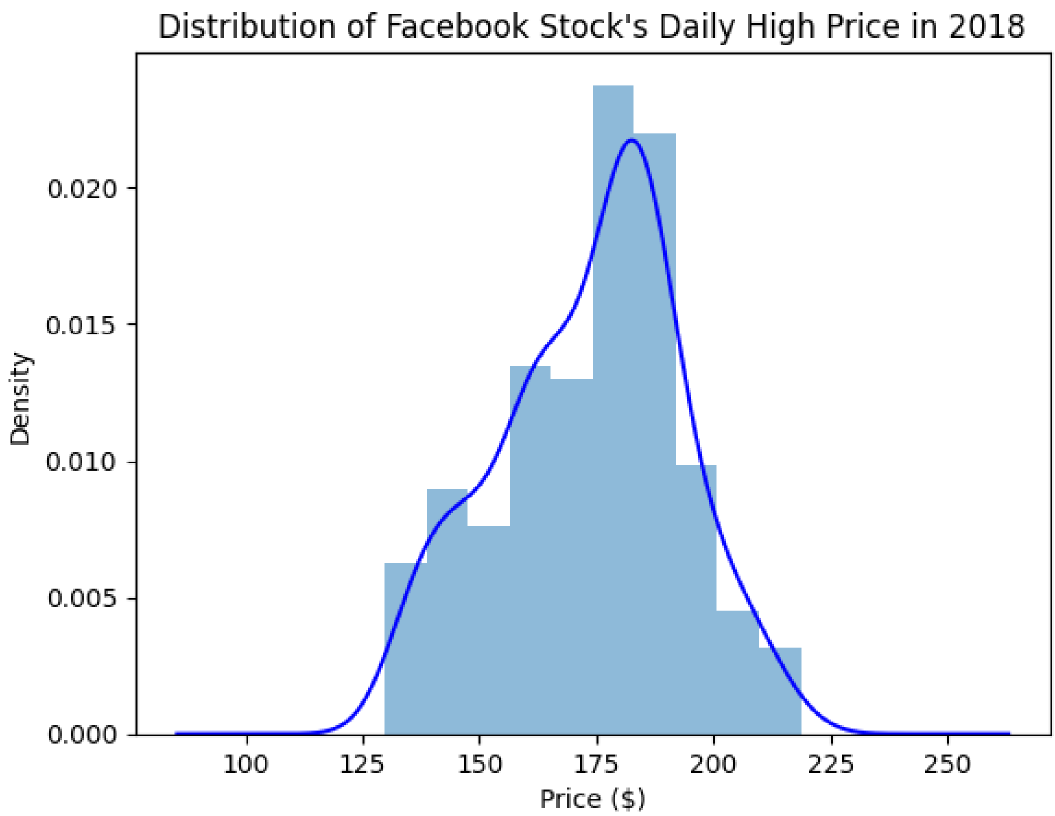
The plot() method returns a matplotlib Axes object. We can store this for additional customization of the plot, or we can pass this into another call to plot() as the ax argument to add to the original plot. It can often be helpful to view the KDE superimposed on top of the histogram, which can be achieved with this strategy:

In [18]:

```
ax = fb.high.plot(kind='hist', density=True, alpha=0.5)
fb.high.plot(
  ax=ax, kind='kde', color='blue',
  title='Distribution of Facebook Stock\'s Daily High Price in 2018'
)
plt.xlabel('Price ($)') # label the x-axis (discussed in chapter 6)
```

Out[18]:

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



Plotting the ECDF

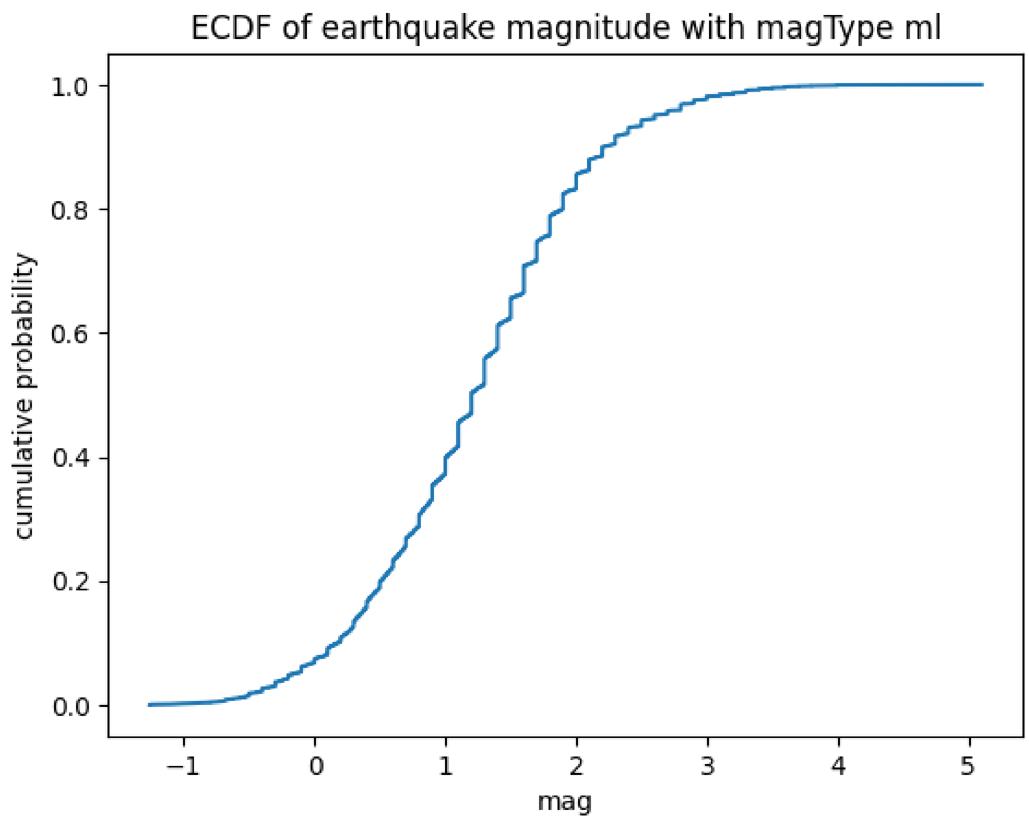
In some cases, we are more interested in the probability of getting less than or equal to that value (or greater than or equal), which we can see with the cumulative disribution function (CDF). Using the statsmodels package, we can estimate the CDF giving us the empirical cumulative distribution function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

In [19]:

```
from statsmodels.distributions.empirical_distribution import ECDF

ecdf = ECDF(quakes.query('magType == "ml"').mag)
plt.plot(ecdf.x, ecdf.y)

# axis labels (we will cover this in chapter 6)
plt.xlabel('mag') # add x-axis label
plt.ylabel('cumulative probability') # add y-axis label

# add title (we will cover this in chapter 6)
plt.title('ECDF of earthquake magnitude with magType ml')
```

Out[19]:

```
Text(0.5, 1.0, 'ECDF of earthquake magnitude with magType ml')
```



Box plots

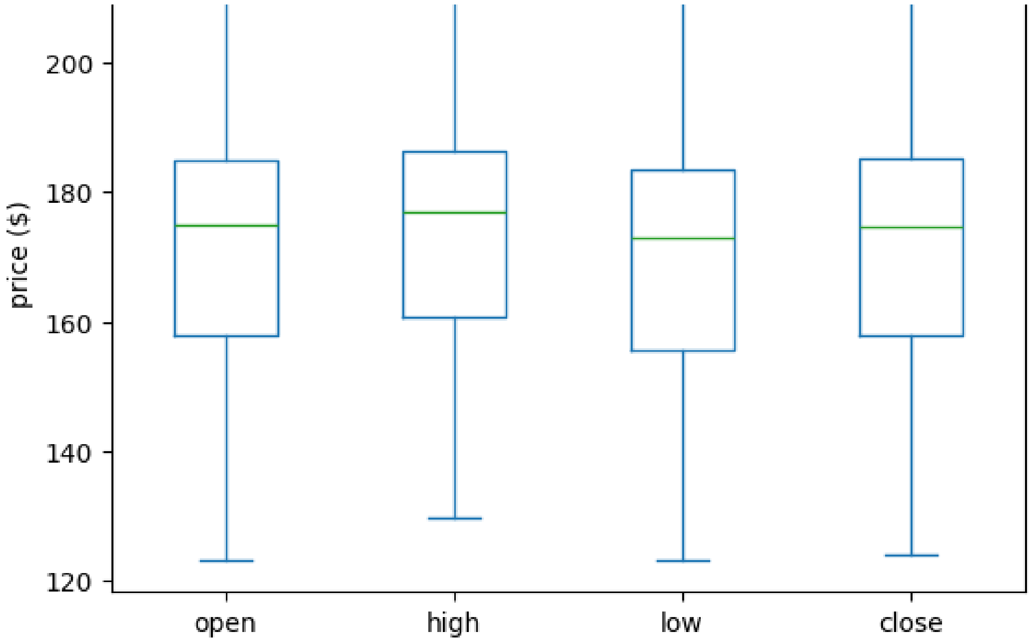
To make box plots with pandas, we pass kind='box' to the plot() method:

In [20]:

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

Out[20]:

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



This can also be combined with a groupby() :

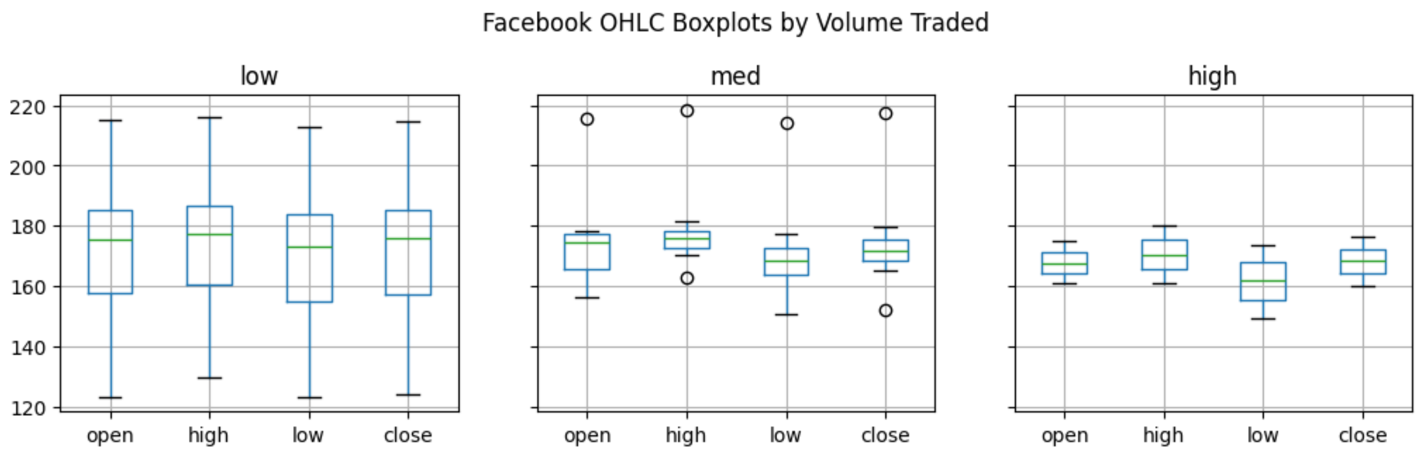
In [21]:

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

```
<ipython-input-21-981746d17a35>:3: FutureWarning: The default of observed=False is deprec
ated and will be changed to True in a future version of pandas. Pass observed=False to re
tain current behavior or observed=True to adopt the future default and silence this warni
ng.
  ).groupby('volume_bin').boxplot(
```

Out[21]:

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



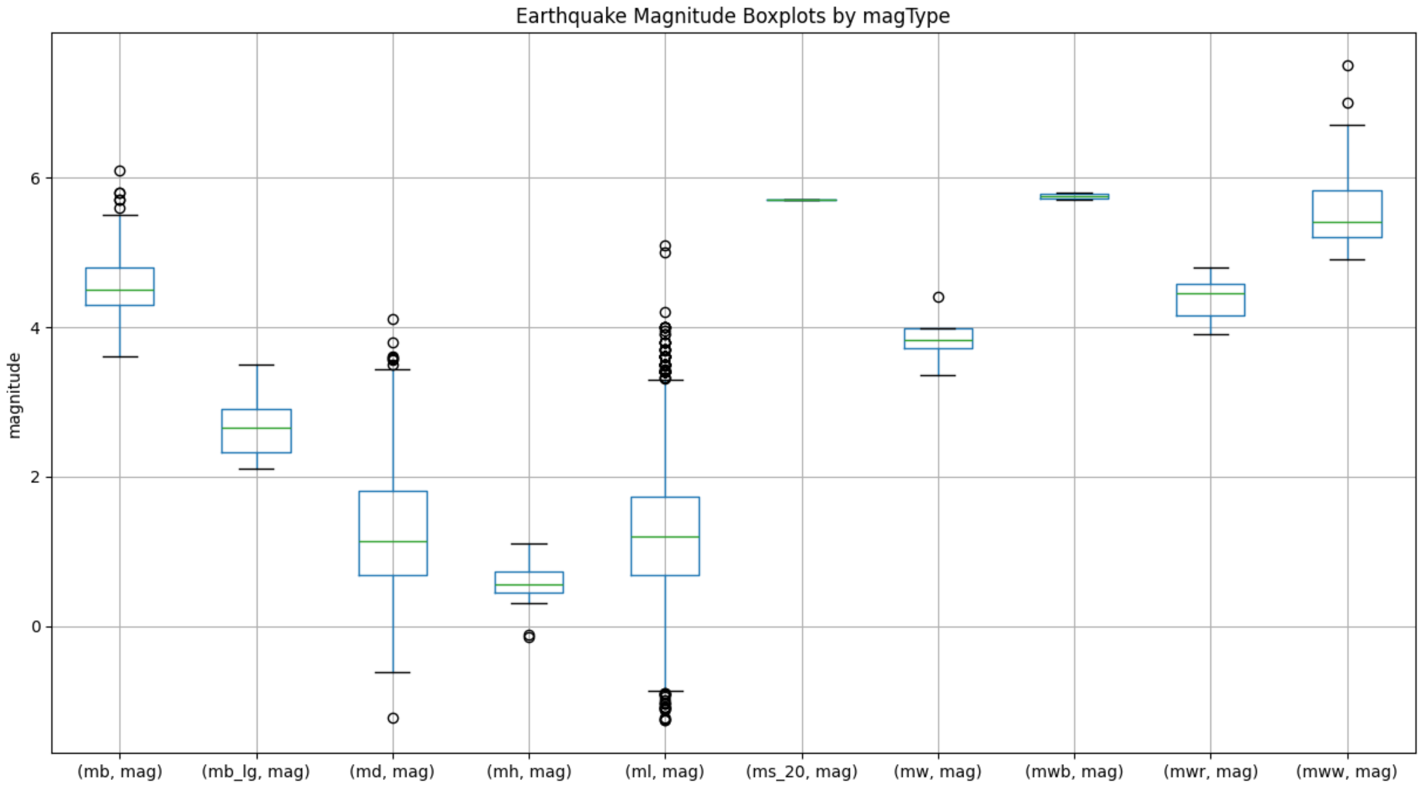
We can use this to see the distribution of magnitudes across the different measurement methods for earthquakes:

In [22]:

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

Out[22]:

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



Counts and frequencies

Bar charts

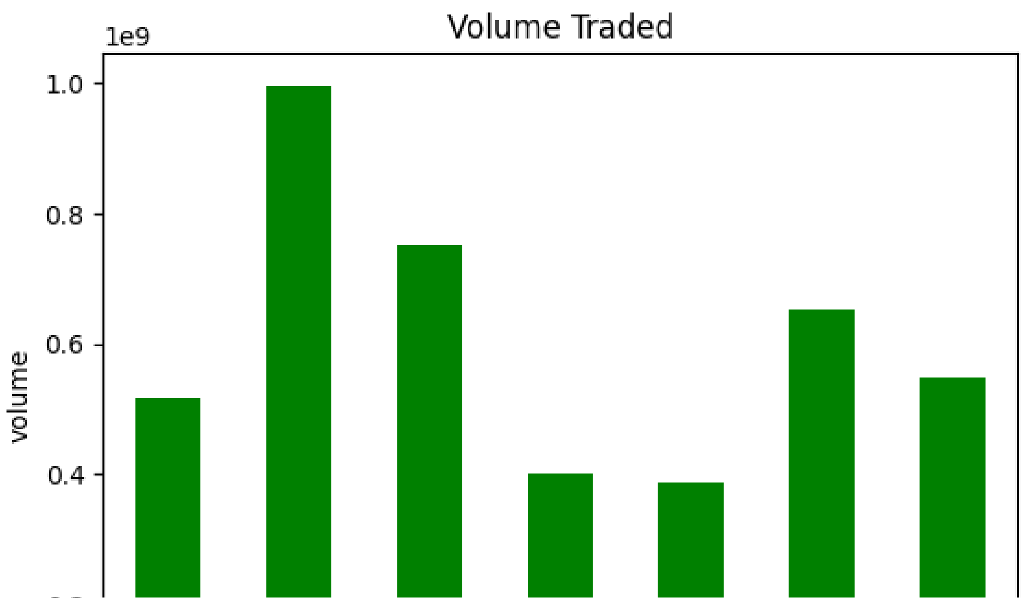
With pandas, we have the option of using the kind argument or using plot. () . Let's use plot.bar() here to show the evolution of monthly volume traded in Facebook stock over time:

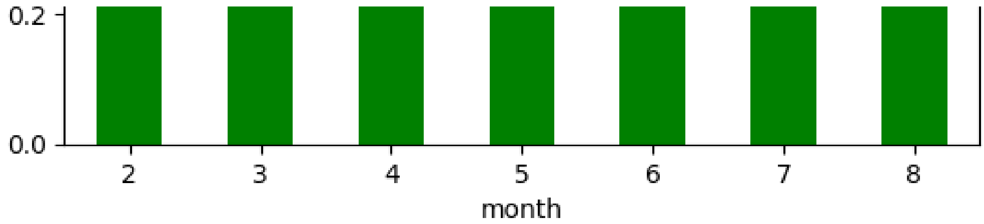
In [23]:

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

Out[23]:

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





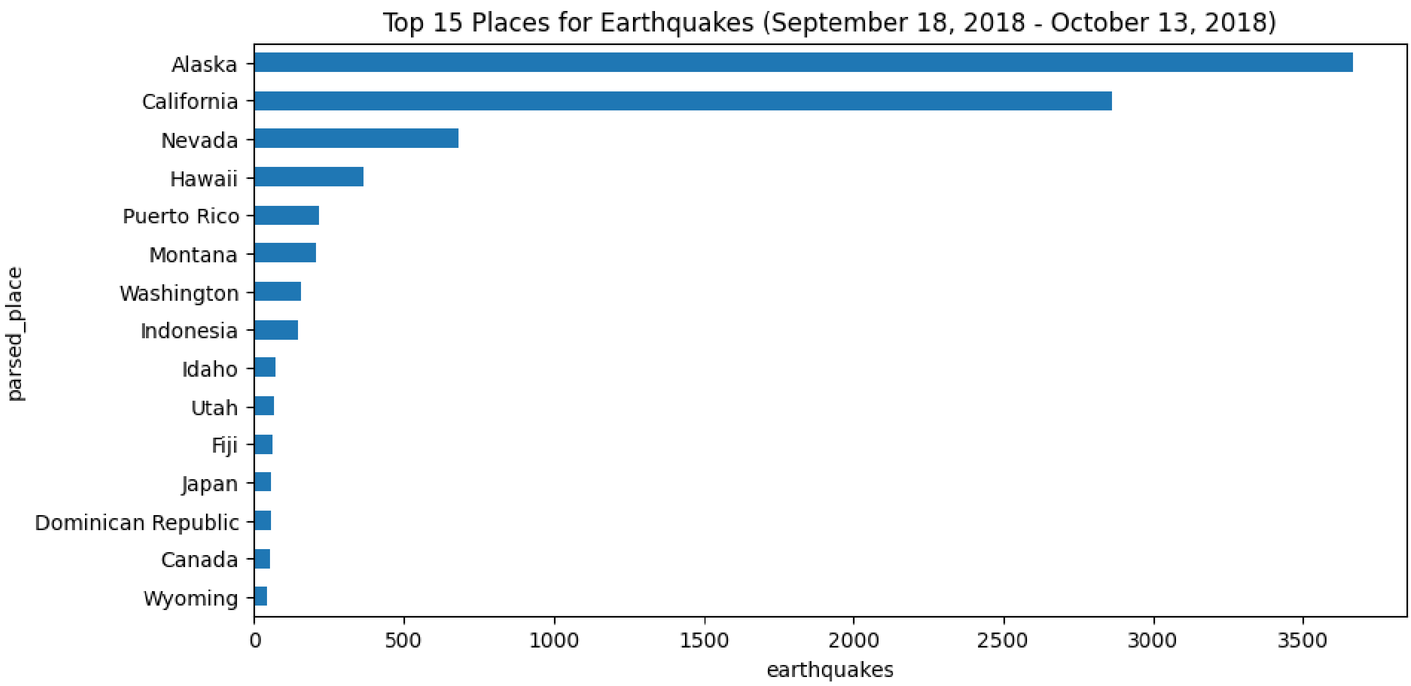
We can also change the orientation of the bars. Passing kind='barh' gives us horizontal bars instead of vertical ones. Let's use this to look at the top 15 places for earthquakes in our data:

In [24]:

```
quakes.parsed_place.value_counts().iloc[14::-1,].plot(
  kind='barh', figsize=(10, 5),
  title='Top 15 Places for Earthquakes '\
        '(September 18, 2018 - October 13, 2018)'
)
plt.xlabel('earthquakes') # label the x-axis (discussed in chapter 6)
```

Out[24]:

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



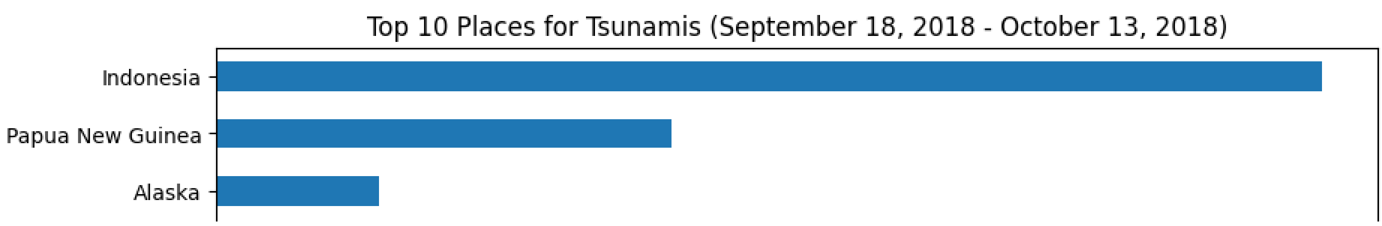
We also have data on whether earthquakes were accompanied by tsunamis. Let's see what the top places for tsunamis are:

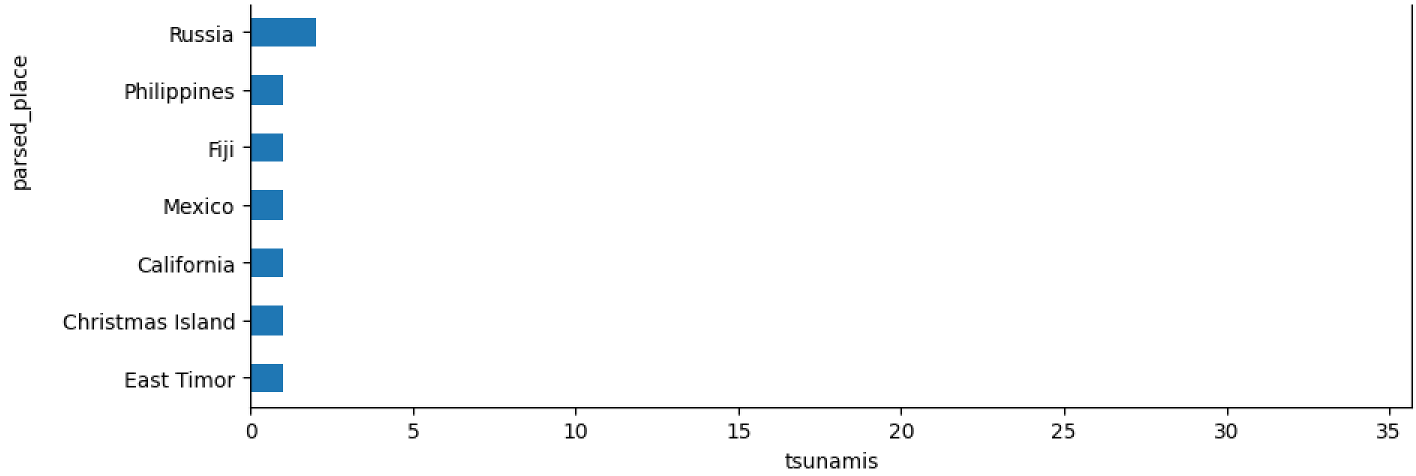
In [25]:

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

Out[25]:

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





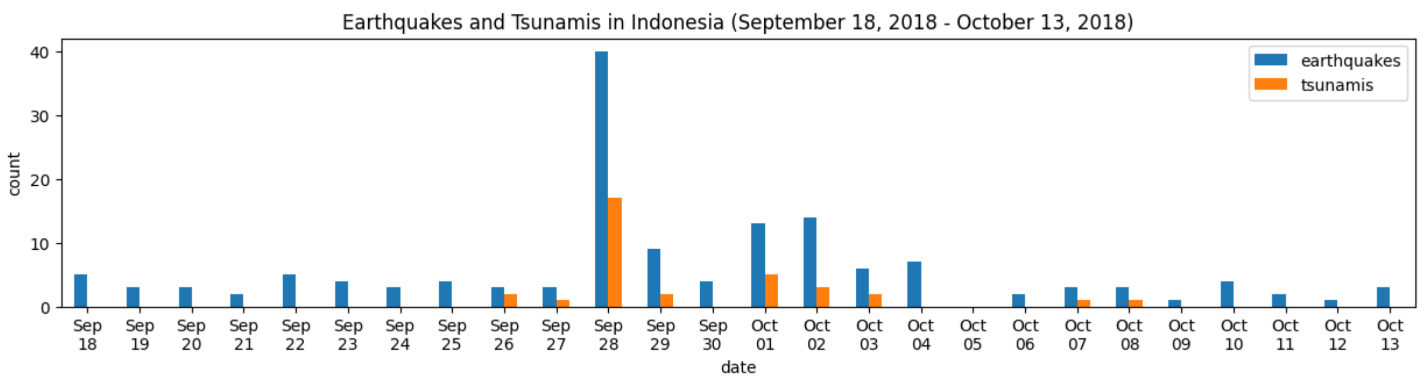
Seeing that Indonesia is the top place for tsunamis during the time period we are looking at, we may want to look how many earthquakes and tsunamis Indonesia gets on a daily basis. We could show this as a line plot or with bars; since this section is about bars, we will use bars here:

In [26]:

```
indonesia_quakes = quakes.query('parsed_place == "Indonesia"').assign(
time=lambda x: pd.to_datetime(x.time, unit='ms'),
earthquake=1
).set_index('time').resample('1D').sum()

indonesia_quakes.index = indonesia_quakes.index.strftime('%b\n%d')

indonesia_quakes.plot(
  y=['earthquake', 'tsunami'], kind='bar', figsize=(15, 3), rot=0,
  label=['earthquakes', 'tsunamis'],
  title='Earthquakes and Tsunamis in Indonesia '\
          '(September 18, 2018 - October 13, 2018)'
)
# label the axes (discussed in chapter 6)
plt.xlabel('date')
plt.ylabel('count')
```

Out[26]:

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



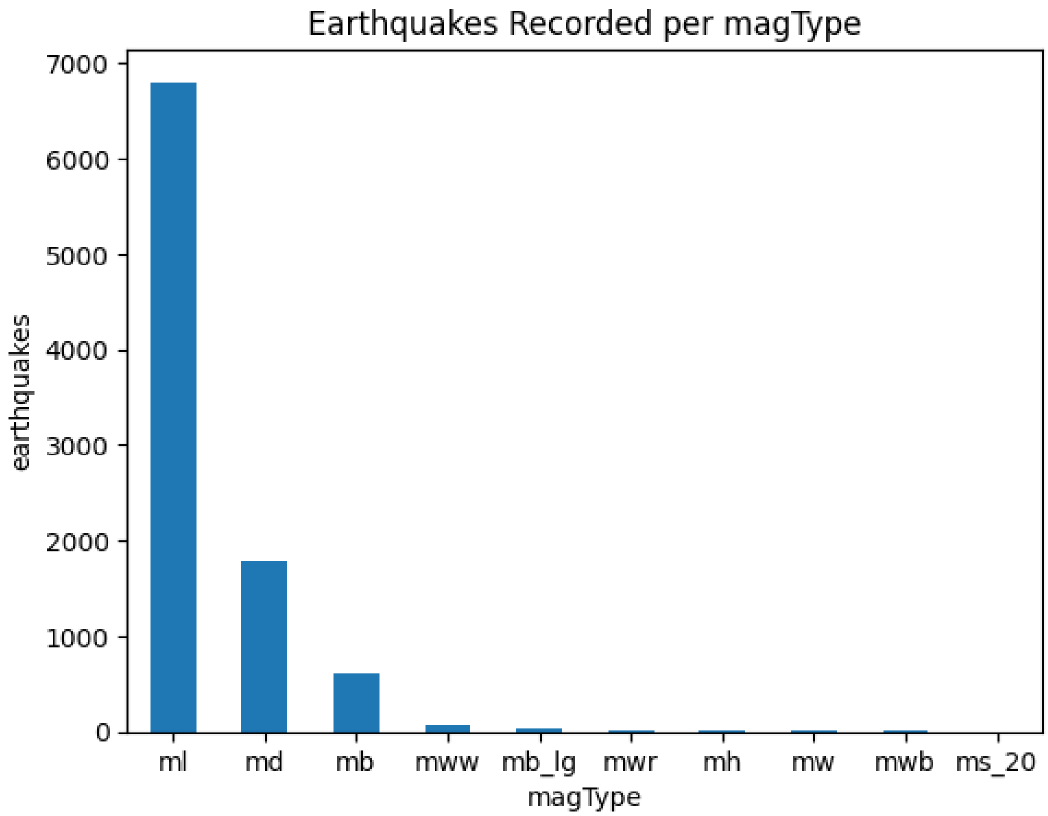
Using the kind arugment for vertical bars when the labels for each bar are shorter:

In [27]:

```
quakes.magType.value_counts().plot(
  kind='bar', title='Earthquakes Recorded per magType', rot=0
)
# label the axes (discussed in chapter 6)
plt.xlabel('magType')
plt.ylabel('earthquakes')
```

Out[27]:

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



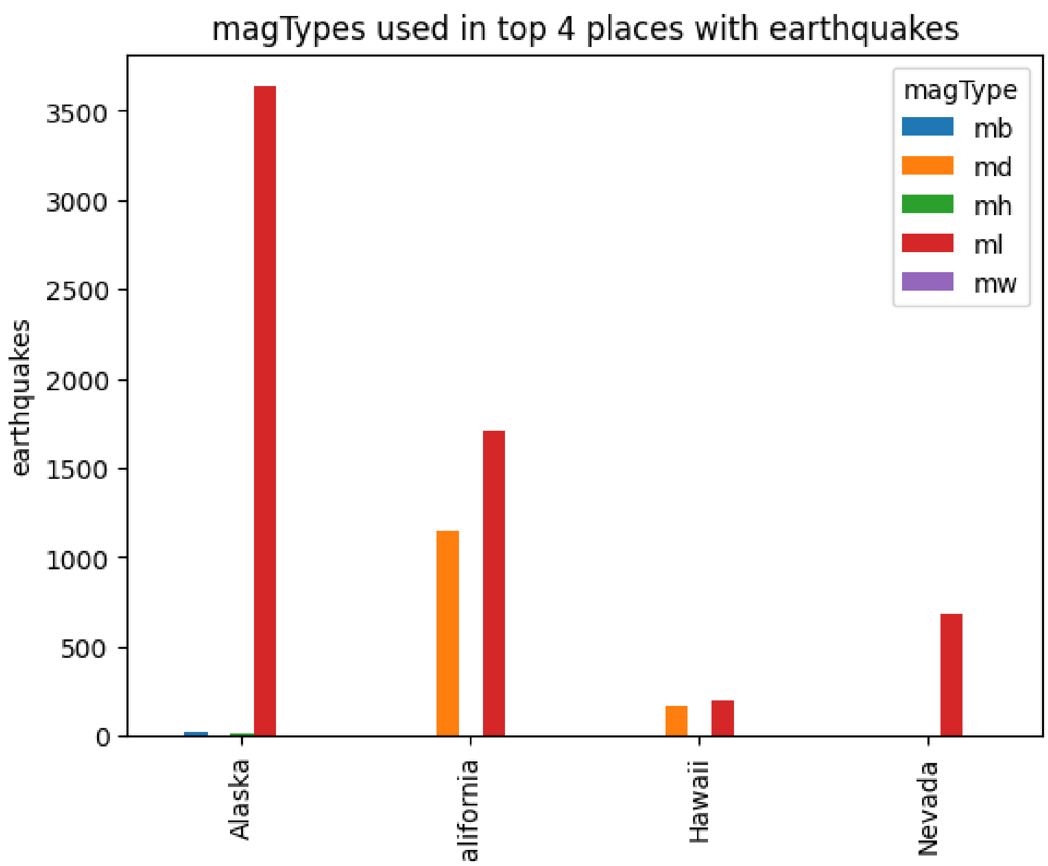
Top 4 places with earthquakes:

In [28]:

```
quakes[
  quakes.parsed_place.isin(['California', 'Alaska', 'Nevada', 'Hawaii'])
].groupby(['parsed_place', 'magType']).mag.count().unstack().plot.bar(
  title='magTypes used in top 4 places with earthquakes'
)
plt.ylabel('earthquakes') # label the axes (discussed in chapter 6)
```

Out[28]:

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



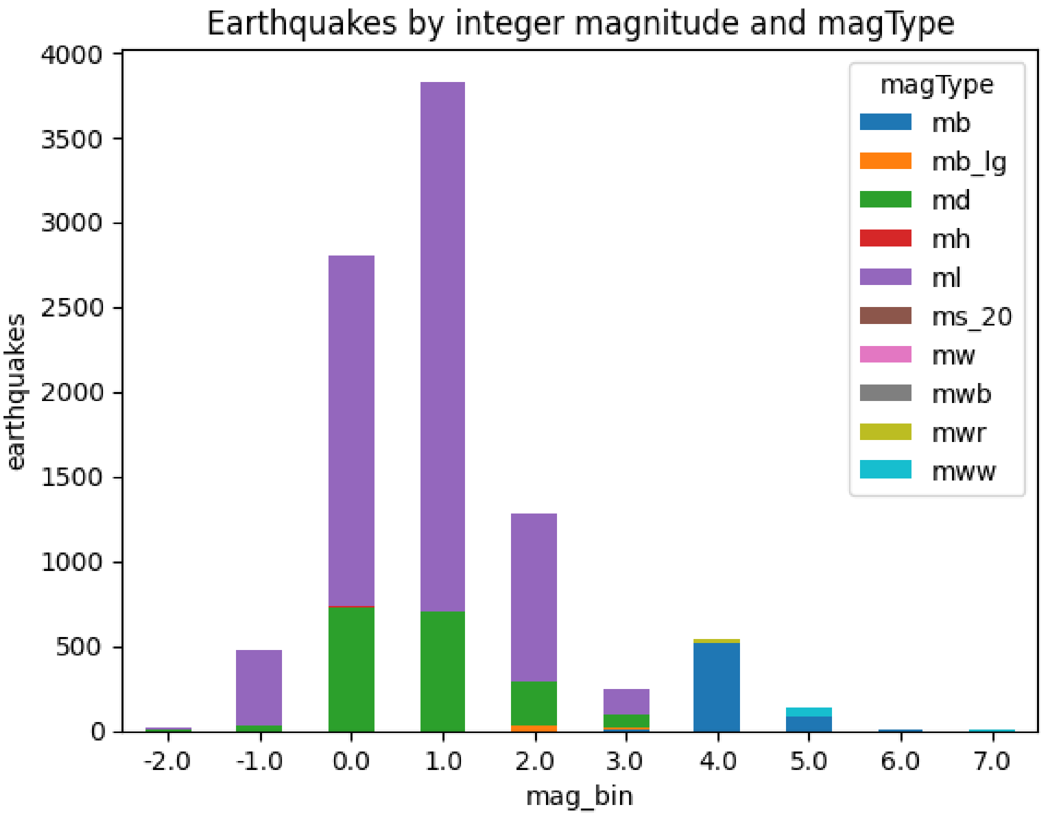
Stacked bar chart

In [29]:

```
pivot = quakes.assign(
  mag_bin=lambda x: np.floor(x.mag)
).pivot_table(
  index='mag_bin', columns='magType', values='mag', aggfunc='count'
)
pivot.plot.bar(
  stacked=True, rot=0,
  title='Earthquakes by integer magnitude and magType'
)
plt.ylabel('earthquakes') # label the axes (discussed in chapter 6)
```

Out[29]:

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



Normalized stacked bars

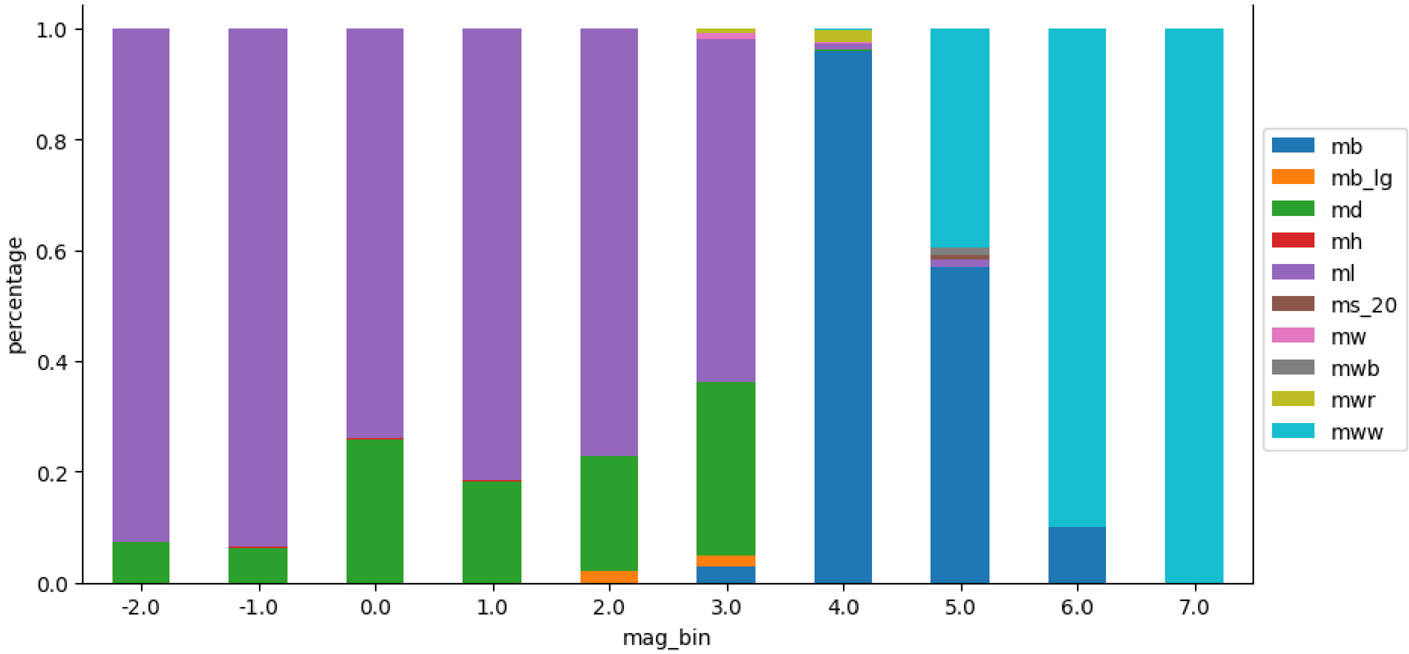
Plot the percentages to be better able to see the different magTypes .

In [30]:

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

Out[30]:

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



1.0
0.8
0.6
0.4
0.2
0.0
percentage
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
mag_bin
mb
mb_lg
md
mh
ml
ms_20
mw
mwb
mwr
mww

In [ ]: