DataFrame Operations

Background on the weather data

Data meanings:

AWND : average wind speed

PRCP : precipitation in millimeters

SNOW : snowfall in millimeters

SNWD : snow depth in millimeters

TMAX : maximum daily temperature in Celsius

TMIN : minimum daily temperature in Celsius

Setup

```
import numpy as np
import pandas as pd

weather = pd.read_csv('/content/nyc_weather_2018.csv', parse_dates=['date'])
weather.head() # Displaying the first few rows of the dataframe
```

	date	datatype	station	attributes	value
0	2018-01-01	PRCP	GHCND:US1CTFR0039	,,N,0800	0.0
1	2018-01-01	PRCP	GHCND:US1NJBG0015	,,N,1050	0.0
2	2018-01-01	SNOW	GHCND:US1NJBG0015	,,N,1050	0.0
3	2018-01-01	PRCP	GHCND:US1NJBG0017	,,N,0920	0.0
4	2018-01-01	SNOW	GHCND:US1NJBG0017	,,N,0920	0.0

Next steps: View recommended plots

```
fb = pd.read_csv('/content/fb_2018.csv', index_col='date', parse_dates=True)
fb.head() # Displaying the first few rows of the dataframe
```

date	open	high	low	close	volume
2018-01-02	177.68	181.58	177.5500	181.42	18151903
2018-01-03	181.88	184.78	181.3300	184.67	16886563
2018-01-04	184.90	186.21	184.0996	184.33	13880896
2018-01-05	185.59	186.90	184.9300	186.85	13574535
2018-01-08	187.20	188.90	186.3300	188.28	17994726

Next steps: View recommended plots

Arithmetic and statistics

We already saw that we can use mathematical operators like + and / with dataframes directly. However, we can also use methods, which allow us to specify the axis to perform the calculation over. By default this is per column. Let's find the z-scores for the volume traded and look at the days where this was more than 3 standard deviations from the mean:

```
fb.assign(
    abs_z_score_volume=lambda x: x.volume.sub(x.volume.mean()).div(x.volume.std()).abs()
    ).query('abs_z_score_volume > 3')
# Adding absolute Z-score of 'volume' column and filtering values with Z-score > 3
```

date	open	high	low	close	volume	abs_z_score_volume
2018-03-19	177.01	177.17	170.06	172.56	88140060	3.145078
2018-03-20	167.47	170.20	161.95	168.15	129851768	5.315169
2018-03-21	164.80	173.40	163.30	169.39	106598834	4.105413
2018-03-26	160.82	161.10	149.02	160.06	126116634	5.120845
2018-07-26	174.89	180.13	173.75	176.26	169803668	7.393705

We can use rank() and pct_change() to see which days had the largest change in volume traded from the day before:

```
fb.assign(
    volume_pct_change=fb.volume.pct_change(),
    pct_change_rank=lambda x: x.volume_pct_change.abs().rank(
        ascending=False
        )
    ).nsmallest(5, 'pct_change_rank')
# Calculating percentage change in volume and ranking by absolute value, then selecting the 5 smallest ranks
```

date	open	high	low	close	volume	volume_pct_change	pct_change_rank
2018-01-12	178.06	181.48	177.40	179.37	77551299	7.087876	1.0
2018-03-19	177.01	177.17	170.06	172.56	88140060	2.611789	2.0
2018-07-26	174.89	180.13	173.75	176.26	169803668	1.628841	3.0

January 12th was when the news that Facebook changed its news feed product to focus more on content from a users' friends over the brands they follow. Given that Facebook's advertising is a key component of its business (nearly 89% in 2017), many shares were sold and the price dropped in panic:

```
fb['2018-01-11':'2018-01-12']
```

date	open	high	low	close	volume
2018-01-11	188.40	188.40	187.38	187.77	9588587
2018-01-12	178.06	181.48	177.40	179.37	77551299

```
fb['2018-01-11':'2024-01-01']
```

date	open	high	low	close	volume
2018-01-11	188.40	188.40	187.38	187.77	9588587
2018-01-12	178.06	181.48	177.40	179.37	77551299
2018-01-16	181.50	181.75	178.04	178.39	36183842
2018-01-17	179.26	179.32	175.80	177.60	27992376
2018-01-18	178.13	180.98	177.08	179.80	23304901
...	...	...	...	...	...
2018-12-24	123.10	129.74	123.02	124.06	22066002
2018-12-26	126.00	134.24	125.89	134.18	39723370
2018-12-27	132.44	134.99	129.67	134.52	31202509
2018-12-28	135.34	135.92	132.20	133.20	22627569
2018-12-31	134.45	134.64	129.95	131.09	24625308

244 rows × 5 columns

Throughout 2018, Facebook's stock price never had a low above $215:

```
(fb > 215).any()
```

```
open       True
high       True
low       False
close      True
volume     True
dtype: bool
```

Facebook's OHLC (open, high, low, and close) prices all had at least one day they were at $215 or less:

```
(fb > 215).all()
```

```
open      False
high      False
low       False
close     False
volume     True
dtype: bool
```

Binning and thresholds

When working with the volume traded, we may be interested in ranges of volume rather than the exact values. No two days have the same volume traded:

```
(fb.volume.value_counts() > 1).sum() # Counting the number of unique volume values that occur more than once
```

```
0
```

We can use pd.cut() to create 3 bins of even an even range in volume traded and name them. Then we can work with low, medium, and high volume traded categories:

```
volume_binned = pd.cut(fb.volume, bins=3, labels=['low', 'med', 'high'])  # Binning the 'volume' column into three categories: low, medium, ar
volume_binned.value_counts()  # Counting the occurrences of each bin
```

```
low     240
med       8
high      3
Name: volume, dtype: int64
```

```
fb[volume_binned == 'high'].sort_values(
    'volume', ascending=False
) # Filtering rows where volume is categorized as 'high' and sorting them by volume in descending order
```

	open	high	low	close	volume
date					
2018-07-26	174.89	180.13	173.75	176.26	169803668
2018-03-20	167.47	170.20	161.95	168.15	129851768
2018-03-26	160.82	161.10	149.02	160.06	126116634

July 25th Facebook announced disappointing user growth and the stock tanked in the after hours:

```
fb['2018-07-25':'2018-07-26']
```

	open	high	low	close	volume
date					
2018-07-25	215.715	218.62	214.27	217.50	64592585
2018-07-26	174.890	180.13	173.75	176.26	169803668

Cambridge Analytica scandal broke on Saturday March 17th, so we look to the Monday for the numbers:

```
fb['2018-03-16':'2018-03-20']
```

	open	high	low	close	volume
date					
2018-03-16	184.49	185.33	183.41	185.09	24403438
2018-03-19	177.01	177.17	170.06	172.56	88140060
2018-03-20	167.47	170.20	161.95	168.15	129851768

Since most days have similar volume, but a few are very large, we have very wide bins. Most of the data is in the low bin. Note: visualizations will be covered in chapters 5 and 6.

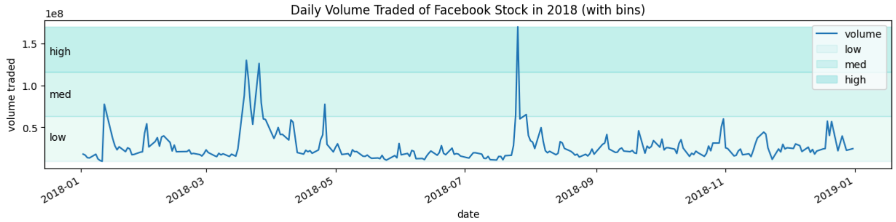
```
import matplotlib.pyplot as plt
```

```
import matplotlib.pyplot as plt
import pandas as pd

# Plotting the volume data with specified parameters
fb.plot(y='volume', figsize=(15, 3), title='Daily Volume Traded of Facebook Stock in 2018 (with bins)')

# Iterating over bins and their properties
for bin_name, alpha, bounds in zip(
    ['low', 'med', 'high'], [0.1, 0.2, 0.3], pd.cut(fb.volume, bins=3).unique().categories.values
    ):
  # Adding shaded areas to represent volume categories
  plt.axhspan(bounds.left, bounds.right, alpha=alpha, label=bin_name, color='mediumturquoise')
  # Annotating the bin names at a specific position
  plt.annotate(bin_name, xy=('2017-12-17', (bounds.left + bounds.right)/2.1))

# Adding labels and legend
plt.ylabel('volume traded')
plt.legend()
plt.show()
```



If we split using quantiles, the bins will have roughly the same number of observations. For this, we use qcut() . We will make 4 quartiles:

```
# Creating quartile bins for volume
volume_qbinned = pd.qcut(fb.volume, q=4, labels=['q1', 'q2', 'q3', 'q4'])

# Counting occurrences in each quartile
volume_qbinned.value_counts()
```

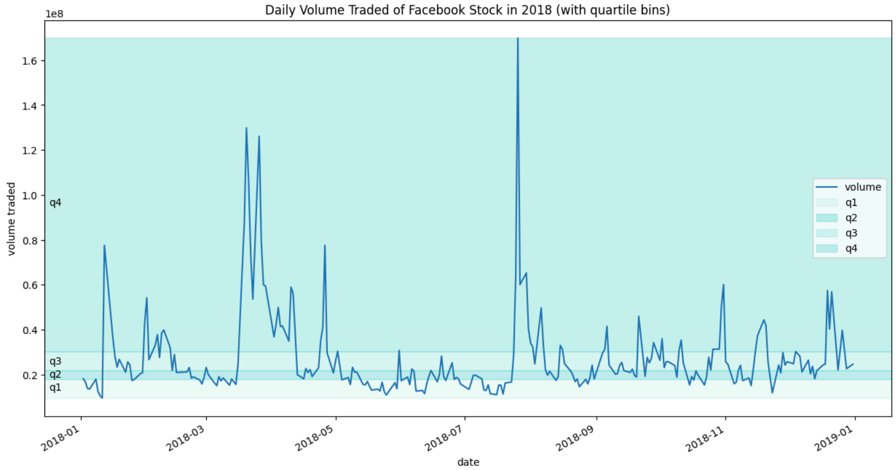
```
q1    63
q2    63
q4    63
q3    62
Name: volume, dtype: int64
```

Notice the bins don't cover ranges of the same size anymore:

```
# Plotting the volume data with specified parameters
fb.plot(y='volume', figsize=(15, 8), title='Daily Volume Traded of Facebook Stock in 2018 (with quartile bins)')

# Iterating over quartile bins and their properties
for bin_name, alpha, bounds in zip(
    ['q1', 'q2', 'q3', 'q4'], [0.1, 0.35, 0.2, 0.3], pd.qcut(fb.volume, q=4).unique().categories.values
    ):
  # Adding shaded areas to represent quartile bins
  plt.axhspan(bounds.left, bounds.right, alpha=alpha, label=bin_name, color='mediumturquoise')
  # Annotating the quartile bin names at a specific position
  plt.annotate(bin_name, xy=('2017-12-17', (bounds.left + bounds.right)/2.1))

# Adding labels and legend
plt.ylabel('volume traded')
plt.legend()
plt.show()
```



Sometimes we don't want to make bins, but rather cap values at a threshold. Before we look at an example, let's pivot our weather data for the Central Park station:

```
# Filtering weather data for Central Park
central_park_weather = weather.query(
    'station == "GHCND:USW00094728"'
).pivot(index='date', columns='datatype', values='value')
```

Say we don't care how much snow their was, just that it snowed in Central Park. However, we don't want to make a Boolean column since we need to preserve the data type of float. We can use clip() to replace values above a upper threshold with the threshold and replace values below a lower threshold with the lower threshold. This means we can use clip(0, 1) to change all the snow values of one or more to 1, which easily shows us the days snow was recorded in Central Park. Preserving the data type will save some work later on if we are building a model:

```
central_park_weather.SNOW.clip(0, 1).value_counts()
# Note: the clip() method can also be called on the dataframe itself.
```

```
0.0    354
1.0     11
Name: SNOW, dtype: int64
```

Applying Functions

We can use the apply() method to run the same operation on all columns (or rows) of the dataframe. Let's calculate the z-scores of the TMIN, TMAX, and PRCP observations in Central Park in October 2018:

```
# Calculate z-scores for temperature minimum, temperature maximum, and precipitation for October 2018
oct_weather_z_scores = central_park_weather.loc[
    '2018-10', ['TMIN', 'TMAX', 'PRCP']
    ].apply(lambda x: x.sub(x.mean()).div(x.std()))
oct_weather_z_scores.describe().T # Calculate summary statistics for z-scores
```

datatype	count	mean	std	min	25%	50%	75%	max
TMIN	31.0	-1.790682e-16	1.0	-1.339112	-0.751019	-0.474269	1.065152	1.843511
TMAX	31.0	1.951844e-16	1.0	-1.305582	-0.870013	-0.138258	1.011643	1.604016

October 27th rained much more than the rest of the days:

```
# Selecting rows from the dataframe where precipitation (PRCP) is greater than 3
oct_weather_z_scores.query('PRCP > 3')
```

date	TMIN	TMAX	PRCP
2018-10-27	-0.751019	-1.201045	3.936167

Indeed, this day was much higher than the rest:

```
# Getting descriptive statistics for precipitation (PRCP) in Central Park for October 2018
central_park_weather.loc['2018-10', 'PRCP'].describe()
```

```
count    31.000000
mean      2.941935
std       7.458542
min       0.000000
25%       0.000000
50%       0.000000
75%       1.150000
max      32.300000
Name: PRCP, dtype: float64
```

When the function we want to apply isn't vectorized, we can: use np.vectorize() to vectorize it (similar to how map() works) and then use it with apply() use applymap() and pass it the non-vectorized function directly Say we wanted to count the digits of the whole numbers for the Facebook data. len() is not vectorized:

```
import numpy as np
fb.apply(
    lambda x: np.vectorize(lambda y: len(str(np.ceil(y))))(x)
    ).astype('int64').equals(
        fb.applymap(lambda x: len(str(np.ceil(x))))
        )
# check if applying a lambda function followed by vectorized operation on 'fb' df
# the same result as applying a lambda function directly to each element of 'fb' df
```

```
True
```

A simple operation of addition to each element in a series grows linearly in time complexity when using iteritems() , but stays near 0 when using vectorized operations. iteritems() and related methods should only be used if there is no vectorized solution:

```
import time
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

np.random.seed(0)
vectorized_results = {}
iteritems_results = {}

# Iterate over different sizes
for size in [10, 100, 1000, 10000, 100000, 500000, 1000000, 5000000, 10000000]:
    # Generate random test data
    test = pd.Series(np.random.uniform(size=size))

    # Time vectorized operation
    start = time.time()
    x = test + 10
```

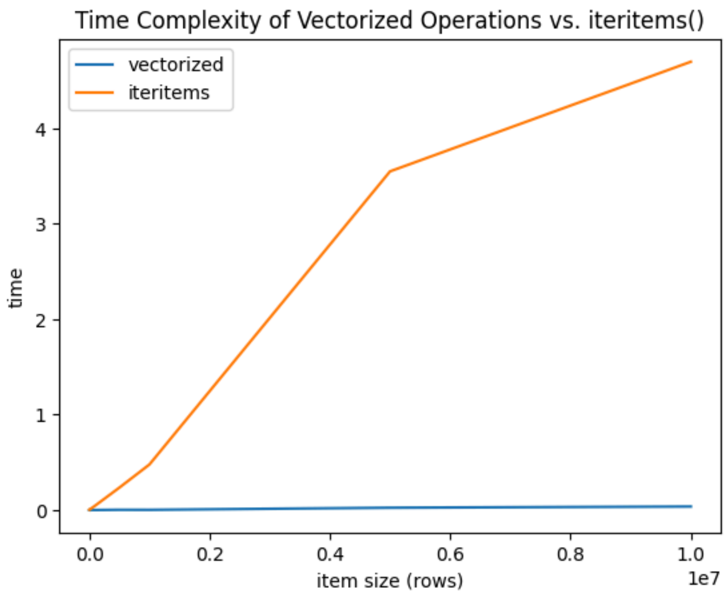
```
    end = time.time()
    vectorized_results[size] = end - start

    # Time iteritems operation
    start = time.time()
    x = []
    for i, v in test.iteritems():
        x.append(v + 10)
    x = pd.Series(x)
    end = time.time()
    iteritems_results[size] = end - start

# Create DataFrame to plot results
df_results = pd.DataFrame({'vectorized': pd.Series(vectorized_results), 'iteritems': pd.Series(iteritems_results)})

# Plot the time complexity comparison
df_results.plot(title='Time Complexity of Vectorized Operations vs. iteritems()')
plt.xlabel('item size (rows)')
plt.ylabel('time')
plt.show()
```

```
<ipython-input-39-feeaf17f583a>:24: FutureWarning: iteritems is deprecated and will be r
  for i, v in test.iteritems():
```



Window Calculations

Consult the understanding windows calculation notebook for interactive visualizations to help understand window calculations.

The rolling() method allows us to perform rolling window calculations. We simply specify the window size (3 days here) and follow it with a call to an aggregation function (sum here):

```
central_park_weather['2018-10'].assign(
    rolling_PRCP=lambda x: x.PRCP.rolling('3D').sum()
)[['PRCP', 'rolling_PRCP']].head(7).T
# Calculate rolling sum of precipitation (PRCP) over a 3-day window for October 2018
```

```
<ipython-input-25-6bb479c3b729>:1: FutureWarning: Indexing a DataFrame with a datetimeli
  central_park_weather['2018-10'].assign(
```

date	2018-10-01	2018-10-02	2018-10-03	2018-10-04	2018-10-05	2018-10-06	2018-10-07
datatype							
PRCP	0.0	17.5	0.0	1.0	0.0	0.0	0.0
rolling_PRCP	0.0	17.5	17.5	18.5	1.0	1.0	0.0

We can also perform the rolling calculations on the entire dataframe at once. This will apply the same aggregation function to each column:

```
central_park_weather['2018-10'].rolling('3D').mean().head(7).iloc[:,:6]
# Calculate rolling mean of Central Park weather data for October 2018 over a 3-day window
```

```
<ipython-input-26-3a953c56f33e>:1: FutureWarning: Indexing a DataFrame with a datetimeli
  central_park_weather['2018-10'].rolling('3D').mean().head(7).iloc[:,:6]
```

datatype date	ADPT	ASLP	ASTP	AWBT	AWND	PRCP
2018-10-01	172.000000	10247.000000	10200.000000	189.000000	0.900000	0.000000
2018-10-02	180.500000	10221.500000	10176.000000	194.500000	0.900000	8.750000
2018-10-03	172.333333	10205.333333	10159.000000	187.000000	0.966667	5.833333
2018-10-04	176.000000	10175.000000	10128.333333	187.000000	0.800000	6.166667
2018-10-05	155.666667	10177.333333	10128.333333	170.333333	1.033333	0.333333
2018-10-06	157.333333	10194.333333	10145.333333	170.333333	0.833333	0.333333
2018-10-07	163.000000	10217.000000	10165.666667	177.666667	1.066667	0.000000

We can use different aggregation functions per column if we use agg() instead. We pass in a dictionary mapping the column to the aggregation to perform on it:

```
central_park_weather['2018-10-01':'2018-10-07'].rolling('3D').agg(
    {'TMAX': 'max', 'TMIN': 'min', 'AWND': 'mean', 'PRCP': 'sum'}
    ).join( # join with original data for comparison
           central_park_weather[['TMAX', 'TMIN', 'AWND', 'PRCP']],
            lsuffix='_rolling'
            ).sort_index(axis=1) # sort columns so rolling calcs are next to originals
```

datatype date	AWND	AWND_rolling	PRCP	PRCP_rolling	TMAX	TMAX_rolling	TMIN	TMIN_rolling
2018-10-01	0.9	0.900000	0.0	0.0	24.4	24.4	17.2	17.2
2018-10-02	0.9	0.900000	17.5	17.5	25.0	25.0	18.3	17.2
2018-10-03	1.1	0.966667	0.0	17.5	23.3	25.0	17.2	17.2
2018-10-04	0.4	0.800000	1.0	18.5	24.4	25.0	16.1	16.1
2018-10-								

Rolling calculations (rolling()) use a sliding window. Expanding calculations (expanding()) however grow in size. These are equivalent to cumulative aggregations like cumsum() ; however, we can specify the minimum number of periods required to start calculating (default is 1):

```
central_park_weather.PRCP.expanding().sum().equals(central_park_weather.PRCP.cumsum())
# Checking if the expanding sum of precipitation (PRCP) is equal to the cumulative sum of precipitation
```

```
False
```

Separate expanding aggregations per column. Note that agg() will accept numpy functions too:

```
central_park_weather['2018-10-01':'2018-10-07'].expanding().agg(
    {'TMAX': np.max, 'TMIN': np.min, 'AWND': np.mean, 'PRCP': np.sum}
    ).join(
        central_park_weather[['TMAX', 'TMIN', 'AWND', 'PRCP']],
        lsuffix='_expanding'
        ).sort_index(axis=1)
# Calculate expanding aggregations for specific columns over the period from October 1st, 2018 to October 7th, 2018
# Then join with original data for those specific columns
```

datatype	AWND	AWND_expanding	PRCP	PRCP_expanding	TMAX	TMAX_expanding	TMIN	TMIN_e
date								
2018-10-01	0.9	0.900000	0.0	0.0	24.4	24.4	17.2	
2018-10-02	0.9	0.900000	17.5	17.5	25.0	25.0	18.3	
2018-10-03	1.1	0.966667	0.0	17.5	23.3	25.0	17.2	
2018-10-04	0.4	0.825000	1.0	18.5	24.4	25.0	16.1	
2018-10-								

We can calculate the exponentially weighted moving average as follows. Note that span here is the periods to use:

```
fb.assign(
    close_ewma=lambda x: x.close.ewm(span=5).mean()
).tail(10)[['close', 'close_ewma']]
# Calculate exponentially weighted moving average (EWMA) for the 'close' column with a span of 5
# Then select the last 10 rows and specific columns
```

	close	close_ewma
date		
2018-12-17	140.19	142.235433
2018-12-18	143.66	142.710289
2018-12-19	133.24	139.553526
2018-12-20	133.40	137.502350
2018-12-21	124.95	133.318234
2018-12-24	124.06	130.232156
2018-12-26	134.18	131.548104
2018-12-27	134.52	132.538736
2018-12-28	133.20	132.759157
2018-12-31	131.09	132.202772

Consult the understanding_window_calculations.ipynb notebook for interactive visualizations to help understand window calculations.

Pipes

Pipes all use to apply any function that accepts our data as the first argument and pass in any additional arguments. This makes it easy to chain steps together regardless of if they are methods or functions:

We can pass any function that will accept the caller of pipe() as the first argument:

```
def get_info(df):
    return '%d rows and %d columns and max closing z-score was %d' % (*df.shape, df.close.max())

# Applying z-score normalization to the '2018-Q1' subset of the DataFrame and getting info
subset_info = fb['2018-Q1'].apply(lambda x: (x - x.mean())/x.std()).pipe(get_info)

# Applying z-score normalization to the entire DataFrame and getting info
full_info = get_info(fb.apply(lambda x: (x - x.mean())/x.std()))

# Comparing the outputs
subset_info == full_info
```

```
<ipython-input-31-a344417d98ca>:5: FutureWarning: Indexing a DataFrame with a datetimelike index using a single string to slice the rows
  subset_info = fb['2018-Q1'].apply(lambda x: (x - x.mean())/x.std()).pipe(get_info)
False
```

For example, passing pd.DataFrame.rolling to pipe() is equivalent to calling rolling() directly on the dataframe, except we have more flexiblity to change this:

```
fb.pipe(pd.DataFrame.rolling, '20D').mean().equals(fb.rolling('20D').mean())
# Checking if two rolling means calculated using different approaches are equal
```

```
True
```

The pipe takes the function passed in and calls it with the object that called pipe() as the first argument. Positional and keyword arguments are passed down:

```
pd.DataFrame.rolling(fb, '20D').mean().equals(fb.rolling('20D').mean())
# Checking if two rolling means calculated using different approaches are equal
```

```
True
```

We can use a pipe to make a function that we can use for all our window calculation needs:

```
from window_calc import window_calc
window_calc??
```

```
Signature: window_calc(df, func, agg_dict, *args, **kwargs)
Source:
def window_calc(df, func, agg_dict, *args, **kwargs):
    """
    Run a window calculation of your choice on a DataFrame.

    Parameters:
        - df: The DataFrame to run the calculation on.
        - func: The window calculation method that takes df
                as the first argument.
        - agg_dict: Information to pass to `agg()`, could be a
                    dictionary mapping the columns to the aggregation
                    function to use, a string name for the function,
                    or the function itself.
        - args: Positional arguments to pass to `func`.
        - kwargs: Keyword arguments to pass to `func`.

    Returns:
        - A new DataFrame object.
    """
    return df.pipe(func, *args, **kwargs).agg(agg_dict)
File:      /content/window_calc.py
Type:      function
```

We can use the same interface to calculate various window calculations now. Let's find the expanding median for the Facebook data:

```
window_calc(fb, pd.DataFrame.expanding, np.median).head()
# Calculating median using expanding window
```

date	open	high	low	close	volume
2018-01-02	177.68	181.580	177.5500	181.420	18151903.0
2018-01-03	179.78	183.180	179.4400	183.045	17519233.0
2018-01-04	181.88	184.780	181.3300	184.330	16886563.0
2018-01-05	183.39	185.495	182.7148	184.500	15383729.5
2018-01-08	184.90	186.210	184.0996	184.670	16886563.0

Using the exponentially weighted moving average requires we pass in a keyword argument