```
In [26]: pd.set_option('expand_frame_repr', False)

In [27]: wide_frame
Out[27]:
          0         1         2         3         4         5         6         7    ␣
↪    8         9        10        11        12        13        14        15
0 -0.548702  1.467327 -1.015962 -0.483075  1.637550 -1.217659 -0.291519 -1.745505 -0.
↪263952  0.991460 -0.919069  0.266046 -0.709661  1.669052  1.037882 -1.705775
1 -0.919854 -0.042379  1.247642 -0.009920  0.290213  0.495767  0.362949  1.548106 -1.
↪131345 -0.089329  0.337863 -0.945867 -0.932132  1.956030  0.017587 -0.016692
2 -0.575247  0.254161 -1.143704  0.215897  1.193555 -0.077118 -0.408530 -0.862495  1.
↪346061  1.511763  1.627081 -0.990582 -0.441652  1.211526  0.268520  0.024580
3 -1.577585  0.396823 -0.105381 -0.532532  1.453749  1.208843 -0.080952 -0.264610 -0.
↪727965 -0.589346  0.339969 -0.693205 -0.339355  0.593616  0.884345  1.591431
4  0.141809  0.220390  0.435589  0.192451 -0.096701  0.803351  1.715071 -0.708758 -1.
↪202872 -1.814470  1.018601 -0.595447  1.395433 -0.392670  0.007207  1.928123
```

The width of each line can be changed via 'line_width' (80 by default):

```
pd.set_option('line_width', 40)

wide_frame
```

Updated PyTables support

Docs for PyTables `Table` format & several enhancements to the api. Here is a taste of what to expect.

```
In [41]: store = pd.HDFStore('store.h5')

In [42]: df = pd.DataFrame(np.random.randn(8, 3),
   ....:                   index=pd.date_range('1/1/2000', periods=8),
   ....:                   columns=['A', 'B', 'C'])

In [43]: df
Out[43]:
                   A         B         C
2000-01-01 -2.036047  0.000830 -0.955697
2000-01-02 -0.898872 -0.725411  0.059904
2000-01-03 -0.449644  1.082900 -1.221265
2000-01-04  0.361078  1.330704  0.855932
2000-01-05 -1.216718  1.488887  0.018993
2000-01-06 -0.877046  0.045976  0.437274
2000-01-07 -0.567182 -0.888657 -0.556383
2000-01-08  0.655457  1.117949 -2.782376

[8 rows x 3 columns]

# appending data frames
In [44]: df1 = df[0:4]

In [45]: df2 = df[4:]

In [46]: store.append('df', df1)

In [47]: store.append('df', df2)
```

(continues on next page)

(continued from previous page)

```
In [48]: store
Out[48]:
<class 'pandas.io.pytables.HDFStore'>
File path: store.h5
/df            frame_table  (typ->appendable,nrows->8,ncols->3,indexers->[index])

# selecting the entire store
In [49]: store.select('df')
Out[49]:
                   A         B         C
2000-01-01 -2.036047  0.000830 -0.955697
2000-01-02 -0.898872 -0.725411  0.059904
2000-01-03 -0.449644  1.082900 -1.221265
2000-01-04  0.361078  1.330704  0.855932
2000-01-05 -1.216718  1.488887  0.018993
2000-01-06 -0.877046  0.045976  0.437274
2000-01-07 -0.567182 -0.888657 -0.556383
2000-01-08  0.655457  1.117949 -2.782376

[8 rows x 3 columns]
```

```
In [50]: wp = pd.Panel(np.random.randn(2, 5, 4), items=['Item1', 'Item2'],
   ....:               major_axis=pd.date_range('1/1/2000', periods=5),
   ....:               minor_axis=['A', 'B', 'C', 'D'])

In [51]: wp
Out[51]:
<class 'pandas.core.panel.Panel'>
Dimensions: 2 (items) x 5 (major_axis) x 4 (minor_axis)
Items axis: Item1 to Item2
Major_axis axis: 2000-01-01 00:00:00 to 2000-01-05 00:00:00
Minor_axis axis: A to D

# storing a panel
In [52]: store.append('wp', wp)

# selecting via A QUERY
In [53]: store.select('wp', [pd.Term('major_axis>20000102'),
   ....:                     pd.Term('minor_axis', '=', ['A', 'B'])])
   ....:
Out[53]:
<class 'pandas.core.panel.Panel'>
Dimensions: 2 (items) x 3 (major_axis) x 2 (minor_axis)
Items axis: Item1 to Item2
Major_axis axis: 2000-01-03 00:00:00 to 2000-01-05 00:00:00
Minor_axis axis: A to B

# removing data from tables
In [54]: store.remove('wp', pd.Term('major_axis>20000103'))
Out[54]: 8

In [55]: store.select('wp')
Out[55]:
<class 'pandas.core.panel.Panel'>
Dimensions: 2 (items) x 3 (major_axis) x 4 (minor_axis)
Items axis: Item1 to Item2
```

(continues on next page)

(continued from previous page)

```
Major_axis axis: 2000-01-01 00:00:00 to 2000-01-03 00:00:00
Minor_axis axis: A to D

# deleting a store
In [56]: del store['df']

In [57]: store
Out[57]:
<class 'pandas.io.pytables.HDFStore'>
File path: store.h5
/wp            wide_table   (typ->appendable,nrows->12,ncols->2,indexers->[major_axis,
↪minor_axis])
```

Enhancements

- added ability to hierarchical keys

```
In [58]: store.put('foo/bar/bah', df)

In [59]: store.append('food/orange', df)

In [60]: store.append('food/apple', df)

In [61]: store
Out[61]:
<class 'pandas.io.pytables.HDFStore'>
File path: store.h5
/foo/bar/bah            frame        (shape->[8,3])
/food/apple             frame_table  (typ->appendable,nrows->8,ncols->3,
↪indexers->[index])
/food/orange            frame_table  (typ->appendable,nrows->8,ncols->3,
↪indexers->[index])
/wp                     wide_table   (typ->appendable,nrows->12,ncols->2,
↪indexers->[major_axis,minor_axis])

# remove all nodes under this level
In [62]: store.remove('food')

In [63]: store
Out[63]:
<class 'pandas.io.pytables.HDFStore'>
File path: store.h5
/foo/bar/bah            frame        (shape->[8,3])
/wp                     wide_table   (typ->appendable,nrows->12,ncols->2,
↪indexers->[major_axis,minor_axis])
```

- added mixed-dtype support!

```
In [64]: df['string'] = 'string'

In [65]: df['int'] = 1

In [66]: store.append('df', df)

In [67]: df1 = store.select('df')

In [68]: df1
```

(continues on next page)

(continued from previous page)

```
Out[68]:
                   A         B         C  string  int
2000-01-01 -2.036047  0.000830 -0.955697  string    1
2000-01-02 -0.898872 -0.725411  0.059904  string    1
2000-01-03 -0.449644  1.082900 -1.221265  string    1
2000-01-04  0.361078  1.330704  0.855932  string    1
2000-01-05 -1.216718  1.488887  0.018993  string    1
2000-01-06 -0.877046  0.045976  0.437274  string    1
2000-01-07 -0.567182 -0.888657 -0.556383  string    1
2000-01-08  0.655457  1.117949 -2.782376  string    1

[8 rows x 5 columns]

In [69]: df1.get_dtype_counts()
Out[69]:
float64    3
int64      1
object     1
dtype: int64
```

- performance improvements on table writing
- support for arbitrarily indexed dimensions
- `SparseSeries` now has a `density` property (GH2384)
- enable `Series.str.strip/lstrip/rstrip` methods to take an input argument to strip arbitrary characters (GH2411)
- implement `value_vars` in `melt` to limit values to certain columns and add `melt` to pandas namespace (GH2412)

Bug Fixes

- added `Term` method of specifying where conditions (GH1996).
- `del store['df']` now call `store.remove('df')` for store deletion
- deleting of consecutive rows is much faster than before
- `min_itemsize` parameter can be specified in table creation to force a minimum size for indexing columns (the previous implementation would set the column size based on the first append)
- indexing support via `create_table_index` (requires PyTables >= 2.3) (GH698).
- appending on a store would fail if the table was not first created via `put`
- fixed issue with missing attributes after loading a pickled dataframe (GH2431)
- minor change to select and remove: require a table ONLY if where is also provided (and not None)

Compatibility

0.10 of `HDFStore` is backwards compatible for reading tables created in a prior version of pandas, however, query terms using the prior (undocumented) methodology are unsupported. You must read in the entire file and write it out using the new format to take advantage of the updates.

N dimensional Panels (experimental)

Adding experimental support for Panel4D and factory functions to create n-dimensional named panels. Here is a taste of what to expect.

```
In [58]: p4d = Panel4D(np.random.randn(2, 2, 5, 4),
   ....:        labels=['Label1','Label2'],
   ....:        items=['Item1', 'Item2'],
   ....:        major_axis=date_range('1/1/2000', periods=5),
   ....:        minor_axis=['A', 'B', 'C', 'D'])
   ....:

In [59]: p4d
Out[59]:
<class 'pandas.core.panelnd.Panel4D'>
Dimensions: 2 (labels) x 2 (items) x 5 (major_axis) x 4 (minor_axis)
Labels axis: Label1 to Label2
Items axis: Item1 to Item2
Major_axis axis: 2000-01-01 00:00:00 to 2000-01-05 00:00:00
Minor_axis axis: A to D
```

See the *full release notes* or issue tracker on GitHub for a complete list.

Contributors

A total of 26 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- A. Flaxman +
- Abraham Flaxman
- Adam Obeng +
- Brenda Moon +
- Chang She
- Chris Mulligan +
- Dieter Vandenbussche
- Donald Curtis +
- Jay Bourque +
- Jeff Reback +
- Justin C Johnson +
- K.-Michael Aye
- Keith Hughitt +
- Ken Van Haren +
- Laurent Gautier +
- Luke Lee +
- Martin Blais
- Tobias Brandt +

- Wes McKinney
- Wouter Overmeire
- alex arsenovic +
- jreback +
- locojaydev +
- timmie
- y-p
- zach powers +

5.18 Version 0.9

5.18.1 v0.9.1 (November 14, 2012)

This is a bug fix release from 0.9.0 and includes several new features and enhancements along with a large number of bug fixes. The new features include by-column sort order for DataFrame and Series, improved NA handling for the rank method, masking functions for DataFrame, and intraday time-series filtering for DataFrame.

New features

- *Series.sort*, *DataFrame.sort*, and *DataFrame.sort_index* can now be specified in a per-column manner to support multiple sort orders (GH928)

```
In [2]: df = pd.DataFrame(np.random.randint(0, 2, (6, 3)),
   ...:                    columns=['A', 'B', 'C'])

In [3]: df.sort(['A', 'B'], ascending=[1, 0])

Out[3]:
   A  B  C
3  0  1  1
4  0  1  1
2  0  0  1
0  1  0  0
1  1  0  0
5  1  0  0
```

- *DataFrame.rank* now supports additional argument values for the *na_option* parameter so missing values can be assigned either the largest or the smallest rank (GH1508, GH2159)

```
In [1]: df = pd.DataFrame(np.random.randn(6, 3), columns=['A', 'B', 'C'])

In [2]: df.loc[2:4] = np.nan

In [3]: df.rank()
Out[3]:
     A    B    C
0  3.0  2.0  1.0
1  1.0  3.0  2.0
2  NaN  NaN  NaN
```

(continues on next page)

(continued from previous page)

```
3   NaN   NaN   NaN
4   NaN   NaN   NaN
5   2.0   1.0   3.0

[6 rows x 3 columns]

In [4]: df.rank(na_option='top')
Out[4]:
     A    B    C
0  6.0  5.0  4.0
1  4.0  6.0  5.0
2  2.0  2.0  2.0
3  2.0  2.0  2.0
4  2.0  2.0  2.0
5  5.0  4.0  6.0

[6 rows x 3 columns]

In [5]: df.rank(na_option='bottom')
Out[5]:
     A    B    C
0  3.0  2.0  1.0
1  1.0  3.0  2.0
2  5.0  5.0  5.0
3  5.0  5.0  5.0
4  5.0  5.0  5.0
5  2.0  1.0  3.0

[6 rows x 3 columns]
```

- DataFrame has new *where* and *mask* methods to select values according to a given boolean mask (GH2109, GH2151)

  DataFrame currently supports slicing via a boolean vector the same length as the DataFrame (inside the *[]*). The returned DataFrame has the same number of columns as the original, but is sliced on its index.

```
In [6]: df = DataFrame(np.random.randn(5, 3), columns = ['A','B','C'])

In [7]: df
Out[7]:
          A         B         C
0  0.276232 -1.087401 -0.673690
1  0.113648 -1.478427  0.524988
2  0.404705  0.577046 -1.715002
3 -1.039268 -0.370647 -1.157892
4 -1.344312  0.844885  1.075770

[5 rows x 3 columns]

In [8]: df[df['A'] > 0]
Out[8]:
          A         B         C
0  0.276232 -1.087401 -0.673690
1  0.113648 -1.478427  0.524988
2  0.404705  0.577046 -1.715002
```

(continues on next page)

(continued from previous page)

```
[3 rows x 3 columns]
```

If a DataFrame is sliced with a DataFrame based boolean condition (with the same size as the original DataFrame), then a DataFrame the same size (index and columns) as the original is returned, with elements that do not meet the boolean condition as *NaN*. This is accomplished via the new method *DataFrame.where*. In addition, *where* takes an optional *other* argument for replacement.

```
In [9]: df[df>0]
Out[9]:
          A         B         C
0  0.276232       NaN       NaN
1  0.113648       NaN  0.524988
2  0.404705  0.577046       NaN
3       NaN       NaN       NaN
4       NaN  0.844885  1.075770

[5 rows x 3 columns]

In [10]: df.where(df>0)
Out[10]:
          A         B         C
0  0.276232       NaN       NaN
1  0.113648       NaN  0.524988
2  0.404705  0.577046       NaN
3       NaN       NaN       NaN
4       NaN  0.844885  1.075770

[5 rows x 3 columns]

In [11]: df.where(df>0,-df)
Out[11]:
          A         B         C
0  0.276232  1.087401  0.673690
1  0.113648  1.478427  0.524988
2  0.404705  0.577046  1.715002
3  1.039268  0.370647  1.157892
4  1.344312  0.844885  1.075770

[5 rows x 3 columns]
```

Furthermore, *where* now aligns the input boolean condition (ndarray or DataFrame), such that partial selection with setting is possible. This is analogous to partial setting via *.ix* (but on the contents rather than the axis labels)

```
In [12]: df2 = df.copy()

In [13]: df2[ df2[1:4] > 0 ] = 3

In [14]: df2
Out[14]:
          A         B         C
0  0.276232 -1.087401 -0.673690
1  3.000000 -1.478427  3.000000
2  3.000000  3.000000 -1.715002
3 -1.039268 -0.370647 -1.157892
4 -1.344312  0.844885  1.075770
```

(continues on next page)

(continued from previous page)

```

[5 rows x 3 columns]
```

DataFrame.mask is the inverse boolean operation of *where*.

```
In [15]: df.mask(df<=0)
Out[15]: 
          A         B         C
0  0.276232       NaN       NaN
1  0.113648       NaN  0.524988
2  0.404705  0.577046       NaN
3       NaN       NaN       NaN
4       NaN  0.844885  1.075770

[5 rows x 3 columns]
```

- Enable referencing of Excel columns by their column names (GH1936)

```
In [16]: xl = pd.ExcelFile('data/test.xls')

In [17]: xl.parse('Sheet1', index_col=0, parse_dates=True,
   ....:          parse_cols='A:D')
   ....: 
Out[17]: 
                   A         B         C         D
2000-01-03  0.980269  3.685731 -0.364217 -1.159738
2000-01-04  1.047916 -0.041232 -0.161812  0.212549
2000-01-05  0.498581  0.731168 -0.537677  1.346270
2000-01-06  1.120202  1.567621  0.003641  0.675253
2000-01-07 -0.487094  0.571455 -1.611639  0.103469
2000-01-10  0.836649  0.246462  0.588543  1.062782
2000-01-11 -0.157161  1.340307  1.195778 -1.097007

[7 rows x 4 columns]
```

- Added option to disable pandas-style tick locators and formatters using *series.plot(x_compat=True)* or *pandas.plot_params['x_compat'] = True* (GH2205)
- Existing TimeSeries methods *at_time* and *between_time* were added to DataFrame (GH2149)
- DataFrame.dot can now accept ndarrays (GH2042)
- DataFrame.drop now supports non-unique indexes (GH2101)
- Panel.shift now supports negative periods (GH2164)
- DataFrame now support unary ~ operator (GH2110)

API changes

- Upsampling data with a PeriodIndex will result in a higher frequency TimeSeries that spans the original time window

```
In [1]: prng = pd.period_range('2012Q1', periods=2, freq='Q')

In [2]: s = pd.Series(np.random.randn(len(prng)), prng)

In [4]: s.resample('M')
Out[4]:
2012-01   -1.471992
2012-02         NaN
2012-03         NaN
2012-04   -0.493593
2012-05         NaN
2012-06         NaN
Freq: M, dtype: float64
```

- Period.end_time now returns the last nanosecond in the time interval (GH2124, GH2125, GH1764)

```
In [18]: p = pd.Period('2012')

In [19]: p.end_time
Out[19]: Timestamp('2012-12-31 23:59:59.999999999')
```

- File parsers no longer coerce to float or bool for columns that have custom converters specified (GH2184)

```
In [20]: import io

In [21]: data = ('A,B,C\n'
   ....:         '00001,001,5\n'
   ....:         '00002,002,6')
   ....:

In [22]: pd.read_csv(io.StringIO(data), converters={'A': lambda x: x.strip()})
Out[22]:
       A  B  C
0  00001  1  5
1  00002  2  6

[2 rows x 3 columns]
```

See the *full release notes* or issue tracker on GitHub for a complete list.

Contributors

A total of 11 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Brenda Moon +
- Chang She
- Jeff Reback +
- Justin C Johnson +
- K.-Michael Aye

- Martin Blais
- Tobias Brandt +
- Wes McKinney
- Wouter Overmeire
- timmie
- y-p

5.18.2 v0.9.0 (October 7, 2012)

This is a major release from 0.8.1 and includes several new features and enhancements along with a large number of bug fixes. New features include vectorized unicode encoding/decoding for *Series.str*, *to_latex* method to DataFrame, more flexible parsing of boolean values, and enabling the download of options data from Yahoo! Finance.

New features

- Add `encode` and `decode` for unicode handling to *vectorized string processing methods* in Series.str (GH1706)
- Add `DataFrame.to_latex` method (GH1735)
- Add convenient expanding window equivalents of all rolling_* ops (GH1785)
- Add Options class to pandas.io.data for fetching options data from Yahoo! Finance (GH1748, GH1739)
- More flexible parsing of boolean values (Yes, No, TRUE, FALSE, etc) (GH1691, GH1295)
- Add `level` parameter to `Series.reset_index`
- `TimeSeries.between_time` can now select times across midnight (GH1871)
- Series constructor can now handle generator as input (GH1679)
- `DataFrame.dropna` can now take multiple axes (tuple/list) as input (GH924)
- Enable `skip_footer` parameter in `ExcelFile.parse` (GH1843)

API changes

- The default column names when `header=None` and no columns names passed to functions like `read_csv` has changed to be more Pythonic and amenable to attribute access:

```
In [1]: import io

In [2]: data = ('0,0,1\n'
   ...:         '1,1,0\n'
   ...:         '0,1,0')
   ...:

In [3]: df = pd.read_csv(io.StringIO(data), header=None)

In [4]: df
Out[4]:
   0  1  2
0  0  0  1
1  1  1  0
```

(continues on next page)

(continued from previous page)

```
2  0  1  0

[3 rows x 3 columns]
```

- Creating a Series from another Series, passing an index, will cause reindexing to happen inside rather than treating the Series like an ndarray. Technically improper usages like `Series(df[col1], index=df[col2])` that worked before "by accident" (this was never intended) will lead to all NA Series in some cases. To be perfectly clear:

```
In [5]: s1 = pd.Series([1, 2, 3])

In [6]: s1
Out[6]:
0    1
1    2
2    3
Length: 3, dtype: int64

In [7]: s2 = pd.Series(s1, index=['foo', 'bar', 'baz'])

In [8]: s2
Out[8]:
foo   NaN
bar   NaN
baz   NaN
Length: 3, dtype: float64
```

- Deprecated `day_of_year` API removed from PeriodIndex, use `dayofyear` (GH1723)
- Don't modify NumPy suppress printoption to True at import time
- The internal HDF5 data arrangement for DataFrames has been transposed. Legacy files will still be readable by HDFStore (GH1834, GH1824)
- Legacy cruft removed: pandas.stats.misc.quantileTS
- Use ISO8601 format for Period repr: monthly, daily, and on down (GH1776)
- Empty DataFrame columns are now created as object dtype. This will prevent a class of TypeErrors that was occurring in code where the dtype of a column would depend on the presence of data or not (e.g. a SQL query having results) (GH1783)
- Setting parts of DataFrame/Panel using ix now aligns input Series/DataFrame (GH1630)
- `first` and `last` methods in `GroupBy` no longer drop non-numeric columns (GH1809)
- Resolved inconsistencies in specifying custom NA values in text parser. `na_values` of type dict no longer override default NAs unless `keep_default_na` is set to false explicitly (GH1657)
- `DataFrame.dot` will not do data alignment, and also work with Series (GH1915)

See the *full release notes* or issue tracker on GitHub for a complete list.

Contributors

A total of 24 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Chang She
- Christopher Whelan +
- Dan Miller +
- Daniel Shapiro +
- Dieter Vandenbussche
- Doug Coleman +
- John-Colvin +
- Johnny +
- Joshua Leahy +
- Lars Buitinck +
- Mark O'Leary +
- Martin Blais
- MinRK +
- Paul Ivanov +
- Skipper Seabold
- Spencer Lyon +
- Taavi Burns +
- Wes McKinney
- Wouter Overmeire
- Yaroslav Halchenko
- lenolib +
- tshauck +
- y-p +
- Øystein S. Haaland +

5.19 Version 0.8

5.19.1 v0.8.1 (July 22, 2012)

This release includes a few new features, performance enhancements, and over 30 bug fixes from 0.8.0. New features include notably NA friendly string processing functionality and a series of new plot types and options.

New features

- Add *vectorized string processing methods* accessible via Series.str (GH620)
- Add option to disable adjustment in EWMA (GH1584)
- *Radviz plot* (GH1566)
- *Parallel coordinates plot*
- *Bootstrap plot*
- Per column styles and secondary y-axis plotting (GH1559)
- New datetime converters millisecond plotting (GH1599)
- Add option to disable "sparse" display of hierarchical indexes (GH1538)
- Series/DataFrame's `set_index` method can *append levels* to an existing Index/MultiIndex (GH1569, GH1577)

Performance improvements

- Improved implementation of rolling min and max (thanks to Bottleneck !)
- Add accelerated `'median'` GroupBy option (GH1358)
- Significantly improve the performance of parsing ISO8601-format date strings with `DatetimeIndex` or `to_datetime` (GH1571)
- Improve the performance of GroupBy on single-key aggregations and use with Categorical types
- Significant datetime parsing performance improvements

Contributors

A total of 5 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Chang She
- Skipper Seabold
- Todd DeLuca +
- Vytautas Jancauskas
- Wes McKinney

5.19.2 v0.8.0 (June 29, 2012)

This is a major release from 0.7.3 and includes extensive work on the time series handling and processing infrastructure as well as a great deal of new functionality throughout the library. It includes over 700 commits from more than 20 distinct authors. Most pandas 0.7.3 and earlier users should not experience any issues upgrading, but due to the migration to the NumPy datetime64 dtype, there may be a number of bugs and incompatibilities lurking. Lingering incompatibilities will be fixed ASAP in a 0.8.1 release if necessary. See the *full release notes* or issue tracker on GitHub for a complete list.

Support for non-unique indexes

All objects can now work with non-unique indexes. Data alignment / join operations work according to SQL join semantics (including, if application, index duplication in many-to-many joins)

NumPy datetime64 dtype and 1.6 dependency

Time series data are now represented using NumPy's datetime64 dtype; thus, pandas 0.8.0 now requires at least NumPy 1.6. It has been tested and verified to work with the development version (1.7+) of NumPy as well which includes some significant user-facing API changes. NumPy 1.6 also has a number of bugs having to do with nanosecond resolution data, so I recommend that you steer clear of NumPy 1.6's datetime64 API functions (though limited as they are) and only interact with this data using the interface that pandas provides.

See the end of the 0.8.0 section for a "porting" guide listing potential issues for users migrating legacy code bases from pandas 0.7 or earlier to 0.8.0.

Bug fixes to the 0.7.x series for legacy NumPy < 1.6 users will be provided as they arise. There will be no more further development in 0.7.x beyond bug fixes.

Time series changes and improvements

Note: With this release, legacy scikits.timeseries users should be able to port their code to use pandas.

Note: See *documentation* for overview of pandas timeseries API.

- New datetime64 representation **speeds up join operations and data alignment**, **reduces memory usage**, and improve serialization / deserialization performance significantly over datetime.datetime
- High performance and flexible **resample** method for converting from high-to-low and low-to-high frequency. Supports interpolation, user-defined aggregation functions, and control over how the intervals and result labeling are defined. A suite of high performance Cython/C-based resampling functions (including Open-High-Low-Close) have also been implemented.
- Revamp of *frequency aliases* and support for **frequency shortcuts** like '15min', or '1h30min'
- New *DatetimeIndex class* supports both fixed frequency and irregular time series. Replaces now deprecated DateRange class
- New `PeriodIndex` and `Period` classes for representing *time spans* and performing **calendar logic**, including the *12 fiscal quarterly frequencies <timeseries.quarterly>*. This is a partial port of, and a substantial enhancement to, elements of the scikits.timeseries code base. Support for conversion between PeriodIndex and DatetimeIndex
- New Timestamp data type subclasses *datetime.datetime*, providing the same interface while enabling working with nanosecond-resolution data. Also provides *easy time zone conversions*.
- Enhanced support for *time zones*. Add *tz_convert* and `tz_localize` methods to TimeSeries and DataFrame. All timestamps are stored as UTC; Timestamps from DatetimeIndex objects with time zone set will be localized to local time. Time zone conversions are therefore essentially free. User needs to know very little about pytz library now; only time zone names as as strings are required. Time zone-aware timestamps are equal if and only if their UTC timestamps match. Operations between time zone-aware time series with different time zones will result in a UTC-indexed time series.
- Time series **string indexing conveniences** / shortcuts: slice years, year and month, and index values with strings

- Enhanced time series **plotting**; adaptation of scikits.timeseries matplotlib-based plotting code
- New `date_range`, `bdate_range`, and `period_range` *factory functions*
- Robust **frequency inference** function *infer_freq* and `inferred_freq` property of DatetimeIndex, with option to infer frequency on construction of DatetimeIndex
- to_datetime function efficiently **parses array of strings** to DatetimeIndex. DatetimeIndex will parse array or list of strings to datetime64
- **Optimized** support for datetime64-dtype data in Series and DataFrame columns
- New NaT (Not-a-Time) type to represent **NA** in timestamp arrays
- Optimize Series.asof for looking up **"as of" values** for arrays of timestamps
- Milli, Micro, Nano date offset objects
- Can index time series with datetime.time objects to select all data at particular **time of day** (`TimeSeries.at_time`) or **between two times** (`TimeSeries.between_time`)
- Add *tshift* method for leading/lagging using the frequency (if any) of the index, as opposed to a naive lead/lag using shift

Other new features

- New *cut* and `qcut` functions (like R's cut function) for computing a categorical variable from a continuous variable by binning values either into value-based (`cut`) or quantile-based (`qcut`) bins
- Rename `Factor` to `Categorical` and add a number of usability features
- Add *limit* argument to fillna/reindex
- More flexible multiple function application in GroupBy, and can pass list (name, function) tuples to get result in particular order with given names
- Add flexible *replace* method for efficiently substituting values
- Enhanced *read_csv/read_table* for reading time series data and converting multiple columns to dates
- Add *comments* option to parser functions: read_csv, etc.
- Add *dayfirst* option to parser functions for parsing international DD/MM/YYYY dates
- Allow the user to specify the CSV reader *dialect* to control quoting etc.
- Handling *thousands* separators in read_csv to improve integer parsing.
- Enable unstacking of multiple levels in one shot. Alleviate `pivot_table` bugs (empty columns being introduced)
- Move to klib-based hash tables for indexing; better performance and less memory usage than Python's dict
- Add first, last, min, max, and prod optimized GroupBy functions
- New *ordered_merge* function
- Add flexible *comparison* instance methods eq, ne, lt, gt, etc. to DataFrame, Series
- Improve *scatter_matrix* plotting function and add histogram or kernel density estimates to diagonal
- Add *'kde'* plot option for density plots
- Support for converting DataFrame to R data.frame through rpy2
- Improved support for complex numbers in Series and DataFrame

- Add *pct_change* method to all data structures
- Add max_colwidth configuration option for DataFrame console output
- *Interpolate* Series values using index values
- Can select multiple columns from GroupBy
- Add *update* methods to Series/DataFrame for updating values in place
- Add `any` and `all` method to DataFrame

New plotting methods

```
import pandas as pd
fx = pd.read_pickle('data/fx_prices')
import matplotlib.pyplot as plt
```

`Series.plot` now supports a `secondary_y` option:

```
plt.figure()

fx['FR'].plot(style='g')

fx['IT'].plot(style='k--', secondary_y=True)
```

Vytautas Jancauskas, the 2012 GSOC participant, has added many new plot types. For example, `'kde'` is a new option:

```
In [1]: s = pd.Series(np.concatenate((np.random.randn(1000),
   ...:                               np.random.randn(1000) * 0.5 + 3)))
   ...:

In [2]: plt.figure()
Out[2]: <Figure size 640x480 with 0 Axes>

In [3]: s.hist(density=True, alpha=0.2)
Out[3]: <matplotlib.axes._subplots.AxesSubplot at 0x7f52b7e0b590>

In [4]: s.plot(kind='kde')
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x7f52b7e0b590>
```

See *the plotting page* for much more.

Other API changes

- Deprecation of `offset`, `time_rule`, and `timeRule` arguments names in time series functions. Warnings will be printed until pandas 0.9 or 1.0.

Potential porting issues for pandas <= 0.7.3 users

The major change that may affect you in pandas 0.8.0 is that time series indexes use NumPy's `datetime64` data type instead of `dtype=object` arrays of Python's built-in `datetime.datetime` objects. `DateRange` has been replaced by `DatetimeIndex` but otherwise behaved identically. But, if you have code that converts `DateRange` or `Index` objects that used to contain `datetime.datetime` values to plain NumPy arrays, you may have bugs lurking with code using scalar values because you are handing control over to NumPy:

```
In [5]: import datetime

In [6]: rng = pd.date_range('1/1/2000', periods=10)

In [7]: rng[5]
Out[7]: Timestamp('2000-01-06 00:00:00', freq='D')

In [8]: isinstance(rng[5], datetime.datetime)
Out[8]: True

In [9]: rng_asarray = np.asarray(rng)

In [10]: scalar_val = rng_asarray[5]

In [11]: type(scalar_val)
Out[11]: numpy.datetime64
```

pandas's `Timestamp` object is a subclass of `datetime.datetime` that has nanosecond support (the `nanosecond` field store the nanosecond value between 0 and 999). It should substitute directly into any code that used `datetime.datetime` values before. Thus, I recommend not casting `DatetimeIndex` to regular NumPy arrays.

If you have code that requires an array of `datetime.datetime` objects, you have a couple of options. First, the `astype(object)` method of `DatetimeIndex` produces an array of `Timestamp` objects:

```
In [12]: stamp_array = rng.astype(object)

In [13]: stamp_array
Out[13]: 
Index([2000-01-01 00:00:00, 2000-01-02 00:00:00, 2000-01-03 00:00:00,
       2000-01-04 00:00:00, 2000-01-05 00:00:00, 2000-01-06 00:00:00,
       2000-01-07 00:00:00, 2000-01-08 00:00:00, 2000-01-09 00:00:00,
       2000-01-10 00:00:00],
      dtype='object')

In [14]: stamp_array[5]
Out[14]: Timestamp('2000-01-06 00:00:00', freq='D')
```

To get an array of proper `datetime.datetime` objects, use the `to_pydatetime` method:

```
In [15]: dt_array = rng.to_pydatetime()

In [16]: dt_array
Out[16]: 
array([datetime.datetime(2000, 1, 1, 0, 0),
       datetime.datetime(2000, 1, 2, 0, 0),
       datetime.datetime(2000, 1, 3, 0, 0),
       datetime.datetime(2000, 1, 4, 0, 0),
       datetime.datetime(2000, 1, 5, 0, 0),
```

(continues on next page)

(continued from previous page)

```
       datetime.datetime(2000, 1, 6, 0, 0),
       datetime.datetime(2000, 1, 7, 0, 0),
       datetime.datetime(2000, 1, 8, 0, 0),
       datetime.datetime(2000, 1, 9, 0, 0),
       datetime.datetime(2000, 1, 10, 0, 0)], dtype=object)

In [17]: dt_array[5]
Out[17]: datetime.datetime(2000, 1, 6, 0, 0)
```

matplotlib knows how to handle `datetime.datetime` but not Timestamp objects. While I recommend that you plot time series using `TimeSeries.plot`, you can either use `to_pydatetime` or register a converter for the Timestamp type. See matplotlib documentation for more on this.

Warning: There are bugs in the user-facing API with the nanosecond datetime64 unit in NumPy 1.6. In particular, the string version of the array shows garbage values, and conversion to `dtype=object` is similarly broken.

```
In [18]: rng = pd.date_range('1/1/2000', periods=10)

In [19]: rng
Out[19]:
DatetimeIndex(['2000-01-01', '2000-01-02', '2000-01-03', '2000-01-04',
               '2000-01-05', '2000-01-06', '2000-01-07', '2000-01-08',
               '2000-01-09', '2000-01-10'],
              dtype='datetime64[ns]', freq='D')

In [20]: np.asarray(rng)
Out[20]:
array(['2000-01-01T00:00:00.000000000', '2000-01-02T00:00:00.000000000',
       '2000-01-03T00:00:00.000000000', '2000-01-04T00:00:00.000000000',
       '2000-01-05T00:00:00.000000000', '2000-01-06T00:00:00.000000000',
       '2000-01-07T00:00:00.000000000', '2000-01-08T00:00:00.000000000',
       '2000-01-09T00:00:00.000000000', '2000-01-10T00:00:00.000000000'],
      dtype='datetime64[ns]')

In [21]: converted = np.asarray(rng, dtype=object)

In [22]: converted[5]
Out[22]: Timestamp('2000-01-06 00:00:00', freq='D')
```

Trust me: don't panic. If you are using NumPy 1.6 and restrict your interaction with `datetime64` values to pandas's API you will be just fine. There is nothing wrong with the data-type (a 64-bit integer internally); all of the important data processing happens in pandas and is heavily tested. I strongly recommend that you **do not work directly with datetime64 arrays in NumPy 1.6** and only use the pandas API.

Support for non-unique indexes: In the latter case, you may have code inside a `try:... catch:` block that failed due to the index not being unique. In many cases it will no longer fail (some method like `append` still check for uniqueness unless disabled). However, all is not lost: you can inspect `index.is_unique` and raise an exception explicitly if it is `False` or go to a different code branch.

Contributors

A total of 27 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Adam Klein
- Chang She
- David Zaslavsky +
- Eric Chlebek +
- Jacques Kvam
- Kamil Kisiel
- Kelsey Jordahl +
- Kieran O'Mahony +
- Lorenzo Bolla +
- Luca Beltrame
- Marc Abramowitz +
- Mark Wiebe +
- Paddy Mullen +
- Peng Yu +
- Roy Hyunjin Han +
- RuiDC +
- Senthil Palanisami +
- Skipper Seabold
- Stefan van der Walt +
- Takafumi Arakaki +
- Thomas Kluyver
- Vytautas Jancauskas +
- Wes McKinney
- Wouter Overmeire
- Yaroslav Halchenko
- thuske +
- timmie +

5.20 Version 0.7

5.20.1 v.0.7.3 (April 12, 2012)

This is a minor release from 0.7.2 and fixes many minor bugs and adds a number of nice new features. There are also a couple of API changes to note; these should not affect very many users, and we are inclined to call them “bug fixes” even though they do constitute a change in behavior. See the *full release notes* or issue tracker on GitHub for a complete list.

New features

- New *fixed width file reader*, `read_fwf`
- New *scatter_matrix* function for making a scatter plot matrix

```
from pandas.tools.plotting import scatter_matrix
scatter_matrix(df, alpha=0.2)         # noqa F821
```

- Add `stacked` argument to Series and DataFrame’s `plot` method for *stacked bar plots*.

```
df.plot(kind='bar', stacked=True)     # noqa F821
```

```
df.plot(kind='barh', stacked=True)    # noqa F821
```

- Add log x and y *scaling options* to `DataFrame.plot` and `Series.plot`
- Add `kurt` methods to Series and DataFrame for computing kurtosis

NA Boolean comparison API change

Reverted some changes to how NA values (represented typically as `NaN` or `None`) are handled in non-numeric Series:

```
In [1]: series = pd.Series(['Steve', np.nan, 'Joe'])

In [2]: series == 'Steve'
Out[2]:
0     True
1    False
2    False
Length: 3, dtype: bool

In [3]: series != 'Steve'
Out[3]:
0    False
1     True
2     True
Length: 3, dtype: bool
```

In comparisons, NA / NaN will always come through as `False` except with `!=` which is `True`. *Be very careful* with boolean arithmetic, especially negation, in the presence of NA data. You may wish to add an explicit NA filter into boolean array operations if you are worried about this:

```
In [4]: mask = series == 'Steve'

In [5]: series[mask & series.notnull()]
Out[5]:
0    Steve
Length: 1, dtype: object
```

While propagating NA in comparisons may seem like the right behavior to some users (and you could argue on purely technical grounds that this is the right thing to do), the evaluation was made that propagating NA everywhere, including in numerical arrays, would cause a large amount of problems for users. Thus, a "practicality beats purity" approach was taken. This issue may be revisited at some point in the future.

Other API changes

When calling `apply` on a grouped Series, the return value will also be a Series, to be more consistent with the `groupby` behavior with DataFrame:

```
In [6]: df = pd.DataFrame({'A': ['foo', 'bar', 'foo', 'bar',
   ...:                          'foo', 'bar', 'foo', 'foo'],
   ...:                    'B': ['one', 'one', 'two', 'three',
   ...:                          'two', 'two', 'one', 'three'],
   ...:                    'C': np.random.randn(8), 'D': np.random.randn(8)})
   ...:

In [7]: df
Out[7]:
     A      B         C         D
0  foo    one  0.469112 -0.861849
1  bar    one -0.282863 -2.104569
2  foo    two -1.509059 -0.494929
3  bar  three -1.135632  1.071804
4  foo    two  1.212112  0.721555
5  bar    two -0.173215 -0.706771
6  foo    one  0.119209 -1.039575
7  foo  three -1.044236  0.271860

[8 rows x 4 columns]

In [8]: grouped = df.groupby('A')['C']

In [9]: grouped.describe()
Out[9]:
     count      mean       std       min       25%       50%       75%       max
A
bar    3.0 -0.530570  0.526860 -1.135632 -0.709248 -0.282863 -0.228039 -0.173215
foo    5.0 -0.150572  1.113308 -1.509059 -1.044236  0.119209  0.469112  1.212112

[2 rows x 8 columns]

In [10]: grouped.apply(lambda x: x.sort_values()[-2:])    # top 2 values
Out[10]:
A
bar  1   -0.282863
     5   -0.173215
foo  0    0.469112
```

(continues on next page)

(continued from previous page)

```
     4    1.212112
Name: C, Length: 4, dtype: float64
```

Contributors

A total of 15 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Abraham Flaxman +
- Adam Klein
- Andreas H. +
- Chang She
- Dieter Vandenbussche
- Jacques Kvam +
- K.-Michael Aye +
- Kamil Kisiel +
- Martin Blais +
- Skipper Seabold
- Thomas Kluyver
- Wes McKinney
- Wouter Overmeire
- Yaroslav Halchenko
- lgautier +

5.20.2 v.0.7.2 (March 16, 2012)

This release targets bugs in 0.7.1, and adds a few minor features.

New features

- Add additional tie-breaking methods in DataFrame.rank (GH874)
- Add ascending parameter to rank in Series, DataFrame (GH875)
- Add coerce_float option to DataFrame.from_records (GH893)
- Add sort_columns parameter to allow unsorted plots (GH918)
- Enable column access via attributes on GroupBy (GH882)
- Can pass dict of values to DataFrame.fillna (GH661)
- Can select multiple hierarchical groups by passing list of values in .ix (GH134)
- Add `axis` option to DataFrame.fillna (GH174)
- Add level keyword to `drop` for dropping values from a level (GH159)

Performance improvements

- Use khash for Series.value_counts, add raw function to algorithms.py (GH861)
- Intercept __builtin__.sum in groupby (GH885)

Contributors

A total of 12 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Adam Klein
- Benjamin Gross +
- Dan Birken +
- Dieter Vandenbussche
- Josh +
- Thomas Kluyver
- Travis N. Vaught +
- Wes McKinney
- Wouter Overmeire
- claudiobertoldi +
- elpres +
- joshuaar +

5.20.3 v.0.7.1 (February 29, 2012)

This release includes a few new features and addresses over a dozen bugs in 0.7.0.

New features

- Add `to_clipboard` function to pandas namespace for writing objects to the system clipboard (GH774)
- Add `itertuples` method to DataFrame for iterating through the rows of a dataframe as tuples (GH818)
- Add ability to pass fill_value and method to DataFrame and Series align method (GH806, GH807)
- Add fill_value option to reindex, align methods (GH784)
- Enable concat to produce DataFrame from Series (GH787)
- Add `between` method to Series (GH802)
- Add HTML representation hook to DataFrame for the IPython HTML notebook (GH773)
- Support for reading Excel 2007 XML documents using openpyxl

Performance improvements

- Improve performance and memory usage of fillna on DataFrame
- Can concatenate a list of Series along axis=1 to obtain a DataFrame (GH787)

Contributors

A total of 9 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Adam Klein
- Brian Granger +
- Chang She
- Dieter Vandenbussche
- Josh Klein
- Steve +
- Wes McKinney
- Wouter Overmeire
- Yaroslav Halchenko

5.20.4 v.0.7.0 (February 9, 2012)

New features

- New unified *merge function* for efficiently performing full gamut of database / relational-algebra operations. Refactored existing join methods to use the new infrastructure, resulting in substantial performance gains (GH220, GH249, GH267)
- New *unified concatenation function* for concatenating Series, DataFrame or Panel objects along an axis. Can form union or intersection of the other axes. Improves performance of `Series.append` and `DataFrame.append` (GH468, GH479, GH273)
- *Can* pass multiple DataFrames to *DataFrame.append* to concatenate (stack) and multiple Series to `Series.append` too
- *Can* pass list of dicts (e.g., a list of JSON objects) to DataFrame constructor (GH526)
- You can now *set multiple columns* in a DataFrame via `__getitem__`, useful for transformation (GH342)
- Handle differently-indexed output values in `DataFrame.apply` (GH498)

```
In [1]: df = pd.DataFrame(np.random.randn(10, 4))

In [2]: df.apply(lambda x: x.describe())
Out[2]:
               0          1          2          3
count  10.000000  10.000000  10.000000  10.000000
mean    0.190912  -0.395125  -0.731920  -0.403130
std     0.730951   0.813266   1.112016   0.961912
min    -0.861849  -2.104569  -1.776904  -1.469388
25%    -0.411391  -0.698728  -1.501401  -1.076610
```

(continues on next page)