- Bug in `MultiIndex.get_level_values` including `Categorical` raises `AttributeError` (GH10460)
- Bug in `pd.get_dummies` with `sparse=True` not returning `SparseDataFrame` (GH10531)
- Bug in `Index` subtypes (such as `PeriodIndex`) not returning their own type for `.drop` and `.insert` methods (GH10620)
- Bug in `algos.outer_join_indexer` when `right` array is empty (GH10618)
- Bug in `filter` (regression from 0.16.0) and `transform` when grouping on multiple keys, one of which is datetime-like (GH10114)
- Bug in `to_datetime` and `to_timedelta` causing `Index` name to be lost (GH10875)
- Bug in `len(DataFrame.groupby)` causing `IndexError` when there's a column containing only NaNs (GH11016)
- Bug that caused segfault when resampling an empty Series (GH10228)
- Bug in `DatetimeIndex` and `PeriodIndex.value_counts` resets name from its result, but retains in result's `Index`. (GH10150)
- Bug in `pd.eval` using `numexpr` engine coerces 1 element numpy array to scalar (GH10546)
- Bug in `pd.concat` with `axis=0` when column is of dtype `category` (GH10177)
- Bug in `read_msgpack` where input type is not always checked (GH10369, GH10630)
- Bug in `pd.read_csv` with kwargs `index_col=False`, `index_col=['a', 'b']` or `dtype` (GH10413, GH10467, GH10577)
- Bug in `Series.from_csv` with `header` kwarg not setting the `Series.name` or the `Series.index.name` (GH10483)
- Bug in `groupby.var` which caused variance to be inaccurate for small float values (GH10448)
- Bug in `Series.plot(kind='hist')` Y Label not informative (GH10485)
- Bug in `read_csv` when using a converter which generates a `uint8` type (GH9266)
- Bug causes memory leak in time-series line and area plot (GH9003)
- Bug when setting a `Panel` sliced along the major or minor axes when the right-hand side is a `DataFrame` (GH11014)
- Bug that returns `None` and does not raise `NotImplementedError` when operator functions (e.g. `.add`) of `Panel` are not implemented (GH7692)
- Bug in line and kde plot cannot accept multiple colors when `subplots=True` (GH9894)
- Bug in `DataFrame.plot` raises `ValueError` when color name is specified by multiple characters (GH10387)
- Bug in left and right `align` of `Series` with `MultiIndex` may be inverted (GH10665)
- Bug in left and right `join` of with `MultiIndex` may be inverted (GH10741)
- Bug in `read_stata` when reading a file with a different order set in `columns` (GH10757)
- Bug in `Categorical` may not representing properly when category contains `tz` or `Period` (GH10713)
- Bug in `Categorical.__iter__` may not returning correct `datetime` and `Period` (GH10713)
- Bug in indexing with a `PeriodIndex` on an object with a `PeriodIndex` (GH4125)
- Bug in `read_csv` with `engine='c'`: EOF preceded by a comment, blank line, etc. was not handled correctly (GH10728, GH10548)

- Reading "famafrench" data via `DataReader` results in HTTP 404 error because of the website url is changed (GH10591).
- Bug in `read_msgpack` where DataFrame to decode has duplicate column names (GH9618)
- Bug in `io.common.get_filepath_or_buffer` which caused reading of valid S3 files to fail if the bucket also contained keys for which the user does not have read permission (GH10604)
- Bug in vectorised setting of timestamp columns with python `datetime.date` and numpy `datetime64` (GH10408, GH10412)
- Bug in `Index.take` may add unnecessary `freq` attribute (GH10791)
- Bug in `merge` with empty `DataFrame` may raise `IndexError` (GH10824)
- Bug in `to_latex` where unexpected keyword argument for some documented arguments (GH10888)
- Bug in indexing of large `DataFrame` where `IndexError` is uncaught (GH10645 and GH10692)
- Bug in `read_csv` when using the `nrows` or `chunksize` parameters if file contains only a header line (GH9535)
- Bug in serialization of `category` types in HDF5 in presence of alternate encodings. (GH10366)
- Bug in `pd.DataFrame` when constructing an empty DataFrame with a string dtype (GH9428)
- Bug in `pd.DataFrame.diff` when DataFrame is not consolidated (GH10907)
- Bug in `pd.unique` for arrays with the `datetime64` or `timedelta64` dtype that meant an array with object dtype was returned instead the original dtype (GH9431)
- Bug in `Timedelta` raising error when slicing from 0s (GH10583)
- Bug in `DatetimeIndex.take` and `TimedeltaIndex.take` may not raise `IndexError` against invalid index (GH10295)
- Bug in `Series([np.nan]).astype('M8[ms]')`, which now returns `Series([pd.NaT])` (GH10747)
- Bug in `PeriodIndex.order` reset freq (GH10295)
- Bug in `date_range` when `freq` divides `end` as nanos (GH10885)
- Bug in `iloc` allowing memory outside bounds of a Series to be accessed with negative integers (GH10779)
- Bug in `read_msgpack` where encoding is not respected (GH10581)
- Bug preventing access to the first index when using `iloc` with a list containing the appropriate negative integer (GH10547, GH10779)
- Bug in `TimedeltaIndex` formatter causing error while trying to save `DataFrame` with `TimedeltaIndex` using `to_csv` (GH10833)
- Bug in `DataFrame.where` when handling Series slicing (GH10218, GH9558)
- Bug where `pd.read_gbq` throws `ValueError` when Bigquery returns zero rows (GH10273)
- Bug in `to_json` which was causing segmentation fault when serializing 0-rank ndarray (GH9576)
- Bug in plotting functions may raise `IndexError` when plotted on `GridSpec` (GH10819)
- Bug in plot result may show unnecessary minor ticklabels (GH10657)
- Bug in `groupby` incorrect computation for aggregation on `DataFrame` with `NaT` (E.g `first`, `last`, `min`). (GH10590, GH11010)
- Bug when constructing `DataFrame` where passing a dictionary with only scalar values and specifying columns did not raise an error (GH10856)

- Bug in `.var()` causing roundoff errors for highly similar values (GH10242)
- Bug in `DataFrame.plot(subplots=True)` with duplicated columns outputs incorrect result (GH10962)
- Bug in `Index` arithmetic may result in incorrect class (GH10638)
- Bug in `date_range` results in empty if freq is negative annually, quarterly and monthly (GH11018)
- Bug in `DatetimeIndex` cannot infer negative freq (GH11018)
- Remove use of some deprecated numpy comparison operations, mainly in tests. (GH10569)
- Bug in `Index` dtype may not applied properly (GH11017)
- Bug in `io.gbq` when testing for minimum google api client version (GH10652)
- Bug in `DataFrame` construction from nested `dict` with `timedelta` keys (GH11129)
- Bug in `.fillna` against may raise `TypeError` when data contains datetime dtype (GH7095, GH11153)
- Bug in `.groupby` when number of keys to group by is same as length of index (GH11185)
- Bug in `convert_objects` where converted values might not be returned if all null and `coerce` (GH9589)
- Bug in `convert_objects` where `copy` keyword was not respected (GH9589)

Contributors

A total of 112 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Alex Rothberg
- Andrea Bedini +
- Andrew Rosenfeld
- Andy Hayden
- Andy Li +
- Anthonios Partheniou +
- Artemy Kolchinsky
- Bernard Willers
- Charlie Clark +
- Chris +
- Chris Whelan
- Christoph Gohlke +
- Christopher Whelan
- Clark Fitzgerald
- Clearfield Christopher +
- Dan Ringwalt +
- Daniel Ni +
- Data & Code Expert Experimenting with Code on Data +
- David Cottrell

- David John Gagne +
- David Kelly +
- ETF +
- Eduardo Schettino +
- Egor +
- Egor Panfilov +
- Evan Wright
- Frank Pinter +
- Gabriel Araujo +
- Garrett-R
- Gianluca Rossi +
- Guillaume Gay
- Guillaume Poulin
- Harsh Nisar +
- Ian Henriksen +
- Ian Hoegen +
- Jaidev Deshpande +
- Jan Rudolph +
- Jan Schulz
- Jason Swails +
- Jeff Reback
- Jonas Buyl +
- Joris Van den Bossche
- Joris Vankerschaver +
- Josh Levy-Kramer +
- Julien Danjou
- Ka Wo Chen
- Karrie Kehoe +
- Kelsey Jordahl
- Kerby Shedden
- Kevin Sheppard
- Lars Buitinck
- Leif Johnson +
- Luis Ortiz +
- Mac +
- Matt Gambogi +

- Matt Savoie +
- Matthew Gilbert +
- Maximilian Roos +
- Michelangelo D'Agostino +
- Mortada Mehyar
- Nick Eubank
- Nipun Batra
- Ondřej Čertík
- Phillip Cloud
- Pratap Vardhan +
- Rafal Skolasinski +
- Richard Lewis +
- Rinoc Johnson +
- Rob Levy
- Robert Gieseke
- Safia Abdalla +
- Samuel Denny +
- Saumitra Shahapure +
- Sebastian Pölsterl +
- Sebastian Rubbert +
- Sheppard, Kevin +
- Sinhrks
- Siu Kwan Lam +
- Skipper Seabold
- Spencer Carrucciu +
- Stephan Hoyer
- Stephen Hoover +
- Stephen Pascoe +
- Terry Santegoeds +
- Thomas Grainger
- Tjerk Santegoeds +
- Tom Augspurger
- Vincent Davis +
- Winterflower +
- Yaroslav Halchenko
- Yuan Tang (Terry) +

- agijsberts
- ajcr +
- behzad nouri
- cel4
- chris-b1 +
- cyrusmaher +
- davidovitch +
- ganego +
- jreback
- juricast +
- larvian +
- maximilianr +
- msund +
- rekcahpassyla
- robertzk +
- scls19fr
- seth-p
- sinhrks
- springcoil +
- terrytangyuan +
- tzinckgraf +

5.11 Version 0.16

5.11.1 v0.16.2 (June 12, 2015)

This is a minor bug-fix release from 0.16.1 and includes a a large number of bug fixes along some new features (`pipe()` method), enhancements, and performance improvements.

We recommend that all users upgrade to this version.

Highlights include:

- A new `pipe` method, see *here*
- Documentation on how to use numba with *pandas*, see *here*

What's new in v0.16.2

New features

Pipe

We've introduced a new method *DataFrame.pipe()*. As suggested by the name, `pipe` should be used to pipe data through a chain of function calls. The goal is to avoid confusing nested function calls like

```
# df is a DataFrame
# f, g, and h are functions that take and return DataFrames
f(g(h(df), arg1=1), arg2=2, arg3=3)  # noqa F821
```

The logic flows from inside out, and function names are separated from their keyword arguments. This can be rewritten as

```
(df.pipe(h)                    # noqa F821
   .pipe(g, arg1=1)            # noqa F821
   .pipe(f, arg2=2, arg3=3)    # noqa F821
)
```

Now both the code and the logic flow from top to bottom. Keyword arguments are next to their functions. Overall the code is much more readable.

In the example above, the functions `f`, `g`, and `h` each expected the DataFrame as the first positional argument. When the function you wish to apply takes its data anywhere other than the first argument, pass a tuple of `(function, keyword)` indicating where the DataFrame should flow. For example:

```
In [1]: import statsmodels.formula.api as sm

In [2]: bb = pd.read_csv('data/baseball.csv', index_col='id')

# sm.ols takes (formula, data)
In [3]: (bb.query('h > 0')
   ...:    .assign(ln_h=lambda df: np.log(df.h))
   ...:    .pipe((sm.ols, 'data'), 'hr ~ ln_h + year + g + C(lg)')
   ...:    .fit()
   ...:    .summary()
   ...:  )
   ...:
Out[3]:
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                     hr   R-squared:                       0.685
Model:                            OLS   Adj. R-squared:                  0.665
Method:                 Least Squares   F-statistic:                     34.28
Date:                Wed, 17 Jun 2020   Prob (F-statistic):           3.48e-15
Time:                        17:53:25   Log-Likelihood:                -205.92
```

(continues on next page)

(continued from previous page)

```
No. Observations:                  68   AIC:                             421.8
Df Residuals:                      63   BIC:                             432.9
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -8484.7720   4664.146     -1.819      0.074   -1.78e+04     835.780
C(lg)[T.NL]    -2.2736      1.325     -1.716      0.091      -4.922       0.375
ln_h           -1.3542      0.875     -1.547      0.127      -3.103       0.395
year            4.2277      2.324      1.819      0.074      -0.417       8.872
g               0.1841      0.029      6.258      0.000       0.125       0.243
==============================================================================
Omnibus:                       10.875   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               17.298
Skew:                           0.537   Prob(JB):                     0.000175
Kurtosis:                       5.225   Cond. No.                     1.49e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly␣
↪specified.
[2] The condition number is large, 1.49e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

The pipe method is inspired by unix pipes, which stream text through processes. More recently dplyr and magrittr have introduced the popular `(%>%)` pipe operator for R.

See the *documentation* for more. (GH10129)

Other enhancements

- Added *rsplit* to Index/Series StringMethods (GH10303)
- Removed the hard-coded size limits on the `DataFrame` HTML representation in the IPython notebook, and leave this to IPython itself (only for IPython v3.0 or greater). This eliminates the duplicate scroll bars that appeared in the notebook with large frames (GH10231).

  Note that the notebook has a `toggle output scrolling` feature to limit the display of very large frames (by clicking left of the output). You can also configure the way DataFrames are displayed using the pandas options, see here *here*.
- `axis` parameter of `DataFrame.quantile` now accepts also `index` and `column`. (GH9543)

API changes

- `Holiday` now raises `NotImplementedError` if both `offset` and `observance` are used in the constructor instead of returning an incorrect result (GH10217).

Performance improvements

- Improved `Series.resample` performance with `dtype=datetime64[ns]` (GH7754)
- Increase performance of `str.split` when `expand=True` (GH10081)

Bug fixes

- Bug in `Series.hist` raises an error when a one row `Series` was given (GH10214)
- Bug where `HDFStore.select` modifies the passed columns list (GH7212)
- Bug in `Categorical` repr with `display.width` of `None` in Python 3 (GH10087)
- Bug in `to_json` with certain orients and a `CategoricalIndex` would segfault (GH10317)
- Bug where some of the nan functions do not have consistent return dtypes (GH10251)
- Bug in `DataFrame.quantile` on checking that a valid axis was passed (GH9543)
- Bug in `groupby.apply` aggregation for `Categorical` not preserving categories (GH10138)
- Bug in `to_csv` where `date_format` is ignored if the `datetime` is fractional (GH10209)
- Bug in `DataFrame.to_json` with mixed data types (GH10289)
- Bug in cache updating when consolidating (GH10264)
- Bug in `mean()` where integer dtypes can overflow (GH10172)
- Bug where `Panel.from_dict` does not set dtype when specified (GH10058)
- Bug in `Index.union` raises `AttributeError` when passing array-likes. (GH10149)
- Bug in `Timestamp`'s' `microsecond`, `quarter`, `dayofyear`, `week` and `daysinmonth` properties return `np.int` type, not built-in `int`. (GH10050)
- Bug in `NaT` raises `AttributeError` when accessing to `daysinmonth`, `dayofweek` properties. (GH10096)
- Bug in Index repr when using the `max_seq_items=None` setting (GH10182).
- Bug in getting timezone data with `dateutil` on various platforms (GH9059, GH8639, GH9663, GH10121)
- Bug in displaying datetimes with mixed frequencies; display 'ms' datetimes to the proper precision. (GH10170)
- Bug in `setitem` where type promotion is applied to the entire block (GH10280)
- Bug in `Series` arithmetic methods may incorrectly hold names (GH10068)
- Bug in `GroupBy.get_group` when grouping on multiple keys, one of which is categorical. (GH10132)
- Bug in `DatetimeIndex` and `TimedeltaIndex` names are lost after timedelta arithmetics (GH9926)
- Bug in `DataFrame` construction from nested `dict` with `datetime64` (GH10160)
- Bug in `Series` construction from `dict` with `datetime64` keys (GH9456)
- Bug in `Series.plot(label="LABEL")` not correctly setting the label (GH10119)
- Bug in `plot` not defaulting to matplotlib `axes.grid` setting (GH9792)
- Bug causing strings containing an exponent, but no decimal to be parsed as `int` instead of `float` in `engine='python'` for the `read_csv` parser (GH9565)
- Bug in `Series.align` resets `name` when `fill_value` is specified (GH10067)
- Bug in `read_csv` causing index name not to be set on an empty DataFrame (GH10184)

- Bug in `SparseSeries.abs` resets `name` (GH10241)
- Bug in `TimedeltaIndex` slicing may reset freq (GH10292)
- Bug in `GroupBy.get_group` raises `ValueError` when group key contains `NaT` (GH6992)
- Bug in `SparseSeries` constructor ignores input data name (GH10258)
- Bug in `Categorical.remove_categories` causing a `ValueError` when removing the `NaN` category if underlying dtype is floating-point (GH10156)
- Bug where infer_freq infers time rule (WOM-5XXX) unsupported by to_offset (GH9425)
- Bug in `DataFrame.to_hdf()` where table format would raise a seemingly unrelated error for invalid (non-string) column names. This is now explicitly forbidden. (GH9057)
- Bug to handle masking empty `DataFrame` (GH10126).
- Bug where MySQL interface could not handle numeric table/column names (GH10255)
- Bug in `read_csv` with a `date_parser` that returned a `datetime64` array of other time resolution than `[ns]` (GH10245)
- Bug in `Panel.apply` when the result has ndim=0 (GH10332)
- Bug in `read_hdf` where `auto_close` could not be passed (GH9327).
- Bug in `read_hdf` where open stores could not be used (GH10330).
- Bug in adding empty `DataFrames`, now results in a `DataFrame` that `.equals` an empty `DataFrame` (GH10181).
- Bug in `to_hdf` and `HDFStore` which did not check that complib choices were valid (GH4582, GH8874).

Contributors

A total of 34 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Andrew Rosenfeld
- Artemy Kolchinsky
- Bernard Willers +
- Christer van der Meeren
- Christian Hudon +
- Constantine Glen Evans +
- Daniel Julius Lasiman +
- Evan Wright
- Francesco Brundu +
- Gaëtan de Menten +
- Jake VanderPlas
- James Hiebert +
- Jeff Reback
- Joris Van den Bossche
- Justin Lecher +

- Ka Wo Chen +
- Kevin Sheppard
- Mortada Mehyar
- Morton Fox +
- Robin Wilson +
- Sinhrks
- Stephan Hoyer
- Thomas Grainger
- Tom Ajamian
- Tom Augspurger
- Yoshiki Vázquez Baeza
- Younggun Kim
- austinc +
- behzad nouri
- jreback
- lexual
- rekcahpassyla +
- scls19fr
- sinhrks

5.11.2 v0.16.1 (May 11, 2015)

This is a minor bug-fix release from 0.16.0 and includes a a large number of bug fixes along several new features, enhancements, and performance improvements. We recommend that all users upgrade to this version.

Highlights include:

- Support for a `CategoricalIndex`, a category based index, see *here*
- New section on how-to-contribute to *pandas*, see *here*
- Revised "Merge, join, and concatenate" documentation, including graphical examples to make it easier to understand each operations, see *here*
- New method `sample` for drawing random samples from Series, DataFrames and Panels. See *here*
- The default `Index` printing has changed to a more uniform format, see *here*
- `BusinessHour` datetime-offset is now supported, see *here*
- Further enhancement to the `.str` accessor to make string operations easier, see *here*

What's new in v0.16.1

- *Enhancements*
  - *CategoricalIndex*

Warning: In pandas 0.17.0, the sub-package `pandas.io.data` will be removed in favor of a separately installable package (GH8961).

Enhancements

CategoricalIndex

We introduce a `CategoricalIndex`, a new type of index object that is useful for supporting indexing with duplicates. This is a container around a `Categorical` (introduced in v0.15.0) and allows efficient indexing and storage of an index with a large number of duplicated elements. Prior to 0.16.1, setting the index of a `DataFrame/Series` with a `category` dtype would convert this to regular object-based `Index`.

```
In [1]: df = pd.DataFrame({'A': np.arange(6),
   ...:                    'B': pd.Series(list('aabbca'))
   ...:                           .astype('category', categories=list('cab'))
   ...:                    })
   ...:

In [2]: df
Out[2]:
   A  B
0  0  a
1  1  a
2  2  b
3  3  b
4  4  c
5  5  a

In [3]: df.dtypes
Out[3]:
A       int64
B    category
dtype: object

In [4]: df.B.cat.categories
Out[4]: Index(['c', 'a', 'b'], dtype='object')
```

setting the index, will create create a `CategoricalIndex`

```
In [5]: df2 = df.set_index('B')

In [6]: df2.index
Out[6]: CategoricalIndex(['a', 'a', 'b', 'b', 'c', 'a'], categories=['c', 'a', 'b'],␣
↪ordered=False, name='B', dtype='category')
```

indexing with `__getitem__/.iloc/.loc/.ix` works similarly to an Index with duplicates. The indexers MUST be in the category or the operation will raise.

```
In [7]: df2.loc['a']
Out[7]:
   A
B
a  0
a  1
a  5
```

and preserves the `CategoricalIndex`

```
In [8]: df2.loc['a'].index
Out[8]: CategoricalIndex(['a', 'a', 'a'], categories=['c', 'a', 'b'], ordered=False,␣
↪name='B', dtype='category')
```

sorting will order by the order of the categories

```
In [9]: df2.sort_index()
Out[9]:
   A
B
c  4
a  0
a  1
a  5
b  2
b  3
```

groupby operations on the index will preserve the index nature as well

```
In [10]: df2.groupby(level=0).sum()
Out[10]:
   A
B
c  4
a  6
b  5

In [11]: df2.groupby(level=0).sum().index
Out[11]: CategoricalIndex(['c', 'a', 'b'], categories=['c', 'a', 'b'], ordered=False,␣
↪name='B', dtype='category')
```

reindexing operations, will return a resulting index based on the type of the passed indexer, meaning that passing a list will return a plain-old-`Index`; indexing with a `Categorical` will return a `CategoricalIndex`, indexed according to the categories of the PASSED `Categorical` dtype. This allows one to arbitrarily index these even with values NOT in the categories, similarly to how you can reindex ANY pandas index.

```
In [12]: df2.reindex(['a', 'e'])
Out[12]:
```

(continues on next page)

(continued from previous page)

```
     A
B
a  0.0
a  1.0
a  5.0
e  NaN

In [13]: df2.reindex(['a', 'e']).index
Out[13]: pd.Index(['a', 'a', 'a', 'e'], dtype='object', name='B')

In [14]: df2.reindex(pd.Categorical(['a', 'e'], categories=list('abcde')))
Out[14]:
     A
B
a  0.0
a  1.0
a  5.0
e  NaN

In [15]: df2.reindex(pd.Categorical(['a', 'e'], categories=list('abcde'))).index
Out[15]: pd.CategoricalIndex(['a', 'a', 'a', 'e'],
                             categories=['a', 'b', 'c', 'd', 'e'],
                             ordered=False, name='B',
                             dtype='category')
```

See the *documentation* for more. (GH7629, GH10038, GH10039)

Sample

Series, DataFrames, and Panels now have a new method: `sample()`. The method accepts a specific number of rows or columns to return, or a fraction of the total number or rows or columns. It also has options for sampling with or without replacement, for passing in a column for weights for non-uniform sampling, and for setting seed values to facilitate replication. (GH2419)

```
In [1]: example_series = pd.Series([0, 1, 2, 3, 4, 5])

# When no arguments are passed, returns 1
In [2]: example_series.sample()
Out[2]:
3    3
Length: 1, dtype: int64

# One may specify either a number of rows:
In [3]: example_series.sample(n=3)
Out[3]:
2    2
1    1
0    0
Length: 3, dtype: int64

# Or a fraction of the rows:
In [4]: example_series.sample(frac=0.5)
Out[4]:
1    1
5    5
```

(continues on next page)

(continued from previous page)

```
3    3
Length: 3, dtype: int64

# weights are accepted.
In [5]: example_weights = [0, 0, 0.2, 0.2, 0.2, 0.4]

In [6]: example_series.sample(n=3, weights=example_weights)
Out[6]:
2    2
4    4
3    3
Length: 3, dtype: int64

# weights will also be normalized if they do not sum to one,
# and missing values will be treated as zeros.
In [7]: example_weights2 = [0.5, 0, 0, 0, None, np.nan]

In [8]: example_series.sample(n=1, weights=example_weights2)
Out[8]:
0    0
Length: 1, dtype: int64
```

When applied to a DataFrame, one may pass the name of a column to specify sampling weights when sampling from rows.

```
In [9]: df = pd.DataFrame({'col1': [9, 8, 7, 6],
   ...:                    'weight_column': [0.5, 0.4, 0.1, 0]})
   ...:

In [10]: df.sample(n=3, weights='weight_column')
Out[10]:
   col1  weight_column
0     9            0.5
1     8            0.4
2     7            0.1

[3 rows x 2 columns]
```

String methods enhancements

Continuing from v0.16.0, the following enhancements make string operations easier and more consistent with standard python string operations.

- Added `StringMethods` (`.str` accessor) to `Index` (GH9068)

  The `.str` accessor is now available for both `Series` and `Index`.

  ```
  In [11]: idx = pd.Index([' jack', 'jill ', ' jesse ', 'frank'])

  In [12]: idx.str.strip()
  Out[12]: Index(['jack', 'jill', 'jesse', 'frank'], dtype='object')
  ```

  One special case for the *.str* accessor on `Index` is that if a string method returns `bool`, the `.str` accessor will return a `np.array` instead of a boolean `Index` (GH8875). This enables the following expression to work naturally:

```
In [13]: idx = pd.Index(['a1', 'a2', 'b1', 'b2'])

In [14]: s = pd.Series(range(4), index=idx)

In [15]: s
Out[15]:
a1    0
a2    1
b1    2
b2    3
Length: 4, dtype: int64

In [16]: idx.str.startswith('a')
Out[16]: array([ True,  True, False, False])

In [17]: s[s.index.str.startswith('a')]
Out[17]:
a1    0
a2    1
Length: 2, dtype: int64
```

- The following new methods are accessible via `.str` accessor to apply the function to each values. (GH9766, GH9773, GH10031, GH10045, GH10052)

		Methods		
`capitalize()`	`swapcase()`	`normalize()`	`partition()`	`rpartition()`
`index()`	`rindex()`	`translate()`		

- `split` now takes `expand` keyword to specify whether to expand dimensionality. `return_type` is deprecated. (GH9847)

```
In [18]: s = pd.Series(['a,b', 'a,c', 'b,c'])

# return Series
In [19]: s.str.split(',')
Out[19]:
0    [a, b]
1    [a, c]
2    [b, c]
Length: 3, dtype: object

# return DataFrame
In [20]: s.str.split(',', expand=True)
Out[20]:
   0  1
0  a  b
1  a  c
2  b  c

[3 rows x 2 columns]

In [21]: idx = pd.Index(['a,b', 'a,c', 'b,c'])

# return Index
In [22]: idx.str.split(',')
Out[22]: Index([['a', 'b'], ['a', 'c'], ['b', 'c']], dtype='object')
```

(continues on next page)

(continued from previous page)

```
# return MultiIndex
In [23]: idx.str.split(',', expand=True)
Out[23]:
MultiIndex([('a', 'b'),
            ('a', 'c'),
            ('b', 'c')],
           )
```

- Improved `extract` and `get_dummies` methods for `Index.str` (GH9980)

Other enhancements

- `BusinessHour` offset is now supported, which represents business hours starting from 09:00 - 17:00 on `BusinessDay` by default. See *Here* for details. (GH7905)

```
In [24]: pd.Timestamp('2014-08-01 09:00') + pd.tseries.offsets.BusinessHour()
Out[24]: Timestamp('2014-08-01 10:00:00')

In [25]: pd.Timestamp('2014-08-01 07:00') + pd.tseries.offsets.BusinessHour()
Out[25]: Timestamp('2014-08-01 10:00:00')

In [26]: pd.Timestamp('2014-08-01 16:30') + pd.tseries.offsets.BusinessHour()
Out[26]: Timestamp('2014-08-04 09:30:00')
```

- `DataFrame.diff` now takes an `axis` parameter that determines the direction of differencing (GH9727)
- Allow `clip`, `clip_lower`, and `clip_upper` to accept array-like arguments as thresholds (This is a regression from 0.11.0). These methods now have an `axis` parameter which determines how the Series or DataFrame will be aligned with the threshold(s). (GH6966)
- `DataFrame.mask()` and `Series.mask()` now support same keywords as `where` (GH8801)
- `drop` function can now accept `errors` keyword to suppress `ValueError` raised when any of label does not exist in the target data. (GH6736)

```
In [27]: df = pd.DataFrame(np.random.randn(3, 3), columns=['A', 'B', 'C'])

In [28]: df.drop(['A', 'X'], axis=1, errors='ignore')
Out[28]:
          B         C
0 -0.706771 -1.039575
1 -0.424972  0.567020
2 -1.087401 -0.673690

[3 rows x 2 columns]
```

- Add support for separating years and quarters using dashes, for example 2014-Q1. (GH9688)
- Allow conversion of values with dtype `datetime64` or `timedelta64` to strings using `astype(str)` (GH9757)
- `get_dummies` function now accepts `sparse` keyword. If set to `True`, the return `DataFrame` is sparse, e.g. `SparseDataFrame`. (GH8823)
- `Period` now accepts `datetime64` as value input. (GH9054)

- Allow timedelta string conversion when leading zero is missing from time definition, ie *0:00:00* vs *00:00:00*. (GH9570)
- Allow `Panel.shift` with `axis='items'` (GH9890)
- Trying to write an excel file now raises `NotImplementedError` if the `DataFrame` has a `MultiIndex` instead of writing a broken Excel file. (GH9794)
- Allow `Categorical.add_categories` to accept `Series` or `np.array`. (GH9927)
- Add/delete `str/dt/cat` accessors dynamically from `__dir__`. (GH9910)
- Add `normalize` as a `dt` accessor method. (GH10047)
- `DataFrame` and `Series` now have `_constructor_expanddim` property as overridable constructor for one higher dimensionality data. This should be used only when it is really needed, see *here*
- `pd.lib.infer_dtype` now returns `'bytes'` in Python 3 where appropriate. (GH10032)

API changes

- When passing in an ax to `df.plot( ..., ax=ax)`, the *sharex* kwarg will now default to *False*. The result is that the visibility of xlabels and xticklabels will not anymore be changed. You have to do that by yourself for the right axes in your figure or set `sharex=True` explicitly (but this changes the visible for all axes in the figure, not only the one which is passed in!). If pandas creates the subplots itself (e.g. no passed in *ax* kwarg), then the default is still `sharex=True` and the visibility changes are applied.
- *`assign()`* now inserts new columns in alphabetical order. Previously the order was arbitrary. (GH9777)
- By default, `read_csv` and `read_table` will now try to infer the compression type based on the file extension. Set `compression=None` to restore the previous behavior (no decompression). (GH9770)

Deprecations

- `Series.str.split`'s `return_type` keyword was removed in favor of `expand` (GH9847)

Index representation

The string representation of `Index` and its sub-classes have now been unified. These will show a single-line display if there are few values; a wrapped multi-line display for a lot of values (but less than `display.max_seq_items`; if lots of items (> `display.max_seq_items`) will show a truncated display (the head and tail of the data). The formatting for `MultiIndex` is unchanged (a multi-line wrapped display). The display width responds to the option `display.max_seq_items`, which is defaulted to 100. (GH6482)

Previous behavior

```
In [2]: pd.Index(range(4), name='foo')
Out[2]: Int64Index([0, 1, 2, 3], dtype='int64')

In [3]: pd.Index(range(104), name='foo')
Out[3]: Int64Index([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,␣
↪19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39,␣
↪40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60,␣
↪61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81,␣
↪82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, ...], dtype=
↪'int64')
```

(continues on next page)

(continued from previous page)

```
In [4]: pd.date_range('20130101', periods=4, name='foo', tz='US/Eastern')
Out[4]:
<class 'pandas.tseries.index.DatetimeIndex'>
[2013-01-01 00:00:00-05:00, ..., 2013-01-04 00:00:00-05:00]
Length: 4, Freq: D, Timezone: US/Eastern

In [5]: pd.date_range('20130101', periods=104, name='foo', tz='US/Eastern')
Out[5]:
<class 'pandas.tseries.index.DatetimeIndex'>
[2013-01-01 00:00:00-05:00, ..., 2013-04-14 00:00:00-04:00]
Length: 104, Freq: D, Timezone: US/Eastern
```

New behavior

```
In [29]: pd.set_option('display.width', 80)

In [30]: pd.Index(range(4), name='foo')
Out[30]: RangeIndex(start=0, stop=4, step=1, name='foo')

In [31]: pd.Index(range(30), name='foo')
Out[31]: RangeIndex(start=0, stop=30, step=1, name='foo')

In [32]: pd.Index(range(104), name='foo')
Out[32]: RangeIndex(start=0, stop=104, step=1, name='foo')

In [33]: pd.CategoricalIndex(['a', 'bb', 'ccc', 'dddd'],
   ....:                     ordered=True, name='foobar')
   ....:
Out[33]: CategoricalIndex(['a', 'bb', 'ccc', 'dddd'], categories=['a', 'bb', 'ccc',
↪'dddd'], ordered=True, name='foobar', dtype='category')

In [34]: pd.CategoricalIndex(['a', 'bb', 'ccc', 'dddd'] * 10,
   ....:                     ordered=True, name='foobar')
   ....:
Out[34]:
CategoricalIndex(['a', 'bb', 'ccc', 'dddd', 'a', 'bb', 'ccc', 'dddd', 'a',
                  'bb', 'ccc', 'dddd', 'a', 'bb', 'ccc', 'dddd', 'a', 'bb',
                  'ccc', 'dddd', 'a', 'bb', 'ccc', 'dddd', 'a', 'bb', 'ccc',
                  'dddd', 'a', 'bb', 'ccc', 'dddd', 'a', 'bb', 'ccc', 'dddd',
                  'a', 'bb', 'ccc', 'dddd'],
                 categories=['a', 'bb', 'ccc', 'dddd'], ordered=True, name='foobar',␣
↪dtype='category')

In [35]: pd.CategoricalIndex(['a', 'bb', 'ccc', 'dddd'] * 100,
   ....:                     ordered=True, name='foobar')
   ....:
Out[35]:
CategoricalIndex(['a', 'bb', 'ccc', 'dddd', 'a', 'bb', 'ccc', 'dddd', 'a',
                  'bb',
                  ...
                  'ccc', 'dddd', 'a', 'bb', 'ccc', 'dddd', 'a', 'bb', 'ccc',
                  'dddd'],
                 categories=['a', 'bb', 'ccc', 'dddd'], ordered=True, name='foobar',␣
↪dtype='category', length=400)

In [36]: pd.date_range('20130101', periods=4, name='foo', tz='US/Eastern')
Out[36]:
```

(continues on next page)

(continued from previous page)

```
DatetimeIndex(['2013-01-01 00:00:00-05:00', '2013-01-02 00:00:00-05:00',
               '2013-01-03 00:00:00-05:00', '2013-01-04 00:00:00-05:00'],
              dtype='datetime64[ns, US/Eastern]', name='foo', freq='D')

In [37]: pd.date_range('20130101', periods=25, freq='D')
Out[37]:
DatetimeIndex(['2013-01-01', '2013-01-02', '2013-01-03', '2013-01-04',
               '2013-01-05', '2013-01-06', '2013-01-07', '2013-01-08',
               '2013-01-09', '2013-01-10', '2013-01-11', '2013-01-12',
               '2013-01-13', '2013-01-14', '2013-01-15', '2013-01-16',
               '2013-01-17', '2013-01-18', '2013-01-19', '2013-01-20',
               '2013-01-21', '2013-01-22', '2013-01-23', '2013-01-24',
               '2013-01-25'],
              dtype='datetime64[ns]', freq='D')

In [38]: pd.date_range('20130101', periods=104, name='foo', tz='US/Eastern')
Out[38]:
DatetimeIndex(['2013-01-01 00:00:00-05:00', '2013-01-02 00:00:00-05:00',
               '2013-01-03 00:00:00-05:00', '2013-01-04 00:00:00-05:00',
               '2013-01-05 00:00:00-05:00', '2013-01-06 00:00:00-05:00',
               '2013-01-07 00:00:00-05:00', '2013-01-08 00:00:00-05:00',
               '2013-01-09 00:00:00-05:00', '2013-01-10 00:00:00-05:00',
               ...
               '2013-04-05 00:00:00-04:00', '2013-04-06 00:00:00-04:00',
               '2013-04-07 00:00:00-04:00', '2013-04-08 00:00:00-04:00',
               '2013-04-09 00:00:00-04:00', '2013-04-10 00:00:00-04:00',
               '2013-04-11 00:00:00-04:00', '2013-04-12 00:00:00-04:00',
               '2013-04-13 00:00:00-04:00', '2013-04-14 00:00:00-04:00'],
              dtype='datetime64[ns, US/Eastern]', name='foo', length=104, freq='D')
```

Performance improvements

- Improved csv write performance with mixed dtypes, including datetimes by up to 5x (GH9940)
- Improved csv write performance generally by 2x (GH9940)
- Improved the performance of `pd.lib.max_len_string_array` by 5-7x (GH10024)

Bug fixes

- Bug where labels did not appear properly in the legend of `DataFrame.plot()`, passing `label=` arguments works, and Series indices are no longer mutated. (GH9542)
- Bug in json serialization causing a segfault when a frame had zero length. (GH9805)
- Bug in `read_csv` where missing trailing delimiters would cause segfault. (GH5664)
- Bug in retaining index name on appending (GH9862)
- Bug in `scatter_matrix` draws unexpected axis ticklabels (GH5662)
- Fixed bug in `StataWriter` resulting in changes to input `DataFrame` upon save (GH9795).
- Bug in `transform` causing length mismatch when null entries were present and a fast aggregator was being used (GH9697)
- Bug in `equals` causing false negatives when block order differed (GH9330)

- Bug in grouping with multiple `pd.Grouper` where one is non-time based (GH10063)
- Bug in `read_sql_table` error when reading postgres table with timezone (GH7139)
- Bug in `DataFrame` slicing may not retain metadata (GH9776)
- Bug where `TimdeltaIndex` were not properly serialized in fixed `HDFStore` (GH9635)
- Bug with `TimedeltaIndex` constructor ignoring `name` when given another `TimedeltaIndex` as data (GH10025).
- Bug in `DataFrameFormatter._get_formatted_index` with not applying `max_colwidth` to the `DataFrame` index (GH7856)
- Bug in `.loc` with a read-only ndarray data source (GH10043)
- Bug in `groupby.apply()` that would raise if a passed user defined function either returned only `None` (for all input). (GH9685)
- Always use temporary files in pytables tests (GH9992)
- Bug in plotting continuously using `secondary_y` may not show legend properly. (GH9610, GH9779)
- Bug in `DataFrame.plot(kind="hist")` results in `TypeError` when `DataFrame` contains non-numeric columns (GH9853)
- Bug where repeated plotting of `DataFrame` with a `DatetimeIndex` may raise `TypeError` (GH9852)
- Bug in `setup.py` that would allow an incompat cython version to build (GH9827)
- Bug in plotting `secondary_y` incorrectly attaches `right_ax` property to secondary axes specifying itself recursively. (GH9861)
- Bug in `Series.quantile` on empty Series of type `Datetime` or `Timedelta` (GH9675)
- Bug in `where` causing incorrect results when upcasting was required (GH9731)
- Bug in `FloatArrayFormatter` where decision boundary for displaying "small" floats in decimal format is off by one order of magnitude for a given display.precision (GH9764)
- Fixed bug where `DataFrame.plot()` raised an error when both `color` and `style` keywords were passed and there was no color symbol in the style strings (GH9671)
- Not showing a `DeprecationWarning` on combining list-likes with an `Index` (GH10083)
- Bug in `read_csv` and `read_table` when using `skip_rows` parameter if blank lines are present. (GH9832)
- Bug in `read_csv()` interprets `index_col=True` as `1` (GH9798)
- Bug in index equality comparisons using == failing on Index/MultiIndex type incompatibility (GH9785)
- Bug in which `SparseDataFrame` could not take *nan* as a column name (GH8822)
- Bug in `to_msgpack` and `read_msgpack` zlib and blosc compression support (GH9783)
- Bug `GroupBy.size` doesn't attach index name properly if grouped by `TimeGrouper` (GH9925)
- Bug causing an exception in slice assignments because `length_of_indexer` returns wrong results (GH9995)
- Bug in csv parser causing lines with initial white space plus one non-space character to be skipped. (GH9710)
- Bug in C csv parser causing spurious NaNs when data started with newline followed by white space. (GH10022)
- Bug causing elements with a null group to spill into the final group when grouping by a `Categorical` (GH9603)
- Bug where .iloc and .loc behavior is not consistent on empty dataframes (GH9964)

- Bug in invalid attribute access on a `TimedeltaIndex` incorrectly raised `ValueError` instead of `AttributeError` (GH9680)
- Bug in unequal comparisons between categorical data and a scalar, which was not in the categories (e.g. `Series(Categorical(list("abc"), ordered=True)) > "d"`. This returned `False` for all elements, but now raises a `TypeError`. Equality comparisons also now return `False` for == and `True` for `!=`. (GH9848)
- Bug in DataFrame `__setitem__` when right hand side is a dictionary (GH9874)
- Bug in `where` when dtype is `datetime64/timedelta64`, but dtype of other is not (GH9804)
- Bug in `MultiIndex.sortlevel()` results in unicode level name breaks (GH9856)
- Bug in which `groupby.transform` incorrectly enforced output dtypes to match input dtypes. (GH9807)
- Bug in `DataFrame` constructor when `columns` parameter is set, and `data` is an empty list (GH9939)
- Bug in bar plot with `log=True` raises `TypeError` if all values are less than 1 (GH9905)
- Bug in horizontal bar plot ignores `log=True` (GH9905)
- Bug in PyTables queries that did not return proper results using the index (GH8265, GH9676)
- Bug where dividing a dataframe containing values of type `Decimal` by another `Decimal` would raise. (GH9787)
- Bug where using DataFrames asfreq would remove the name of the index. (GH9885)
- Bug causing extra index point when resample BM/BQ (GH9756)
- Changed caching in `AbstractHolidayCalendar` to be at the instance level rather than at the class level as the latter can result in unexpected behaviour. (GH9552)
- Fixed latex output for MultiIndexed dataframes (GH9778)
- Bug causing an exception when setting an empty range using `DataFrame.loc` (GH9596)
- Bug in hiding ticklabels with subplots and shared axes when adding a new plot to an existing grid of axes (GH9158)
- Bug in `transform` and `filter` when grouping on a categorical variable (GH9921)
- Bug in `transform` when groups are equal in number and dtype to the input index (GH9700)
- Google BigQuery connector now imports dependencies on a per-method basis.(GH9713)
- Updated BigQuery connector to no longer use deprecated `oauth2client.tools.run()` (GH8327)
- Bug in subclassed `DataFrame`. It may not return the correct class, when slicing or subsetting it. (GH9632)
- Bug in `.median()` where non-float null values are not handled correctly (GH10040)
- Bug in Series.fillna() where it raises if a numerically convertible string is given (GH10092)

Contributors

A total of 58 people contributed patches to this release. People with a "+" by their names contributed a patch for the first time.

- Alfonso MHC +
- Andy Hayden
- Artemy Kolchinsky
- Chris Gilmer +
- Chris Grinolds +
- Dan Birken
- David BROCHART +
- David Hirschfeld +
- David Stephens
- Dr. Leo +
- Evan Wright +
- Frans van Dunné +
- Hatem Nassrat +
- Henning Sperr +
- Hugo Herter +
- Jan Schulz
- Jeff Blackburne +
- Jeff Reback
- Jim Crist +
- Jonas Abernot +
- Joris Van den Bossche
- Kerby Shedden
- Leo Razoumov +
- Manuel Riel +
- Mortada Mehyar
- Nick Burns +
- Nick Eubank +
- Olivier Grisel
- Phillip Cloud
- Pietro Battiston
- Roy Hyunjin Han
- Sam Zhang +
- Scott Sanderson +

- Sinhrks +
- Stephan Hoyer
- Tiago Antao
- Tom Ajamian +
- Tom Augspurger
- Tomaz Berisa +
- Vikram Shirgur +
- Vladimir Filimonov
- William Hogman +
- Yasin A +
- Younggun Kim +
- behzad nouri
- dsm054
- floydsoft +
- flying-sheep +
- gfr +
- jnmclarty
- jreback
- ksanghai +
- lucas +
- mschmohl +
- ptype +
- rockg
- scls19fr +
- sinhrks

5.11.3 v0.16.0 (March 22, 2015)

This is a major release from 0.15.2 and includes a small number of API changes, several new features, enhancements, and performance improvements along with a large number of bug fixes. We recommend that all users upgrade to this version.

Highlights include:

- `DataFrame.assign` method, see *here*
- `Series.to_coo/from_coo` methods to interact with `scipy.sparse`, see *here*
- Backwards incompatible change to `Timedelta` to conform the `.seconds` attribute with `datetime.timedelta`, see *here*
- Changes to the `.loc` slicing API to conform with the behavior of `.ix` see *here*
- Changes to the default for ordering in the `Categorical` constructor, see *here*

- Enhancement to the `.str` accessor to make string operations easier, see *here*
- The `pandas.tools.rplot`, `pandas.sandbox.qtpandas` and `pandas.rpy` modules are deprecated. We refer users to external packages like seaborn, pandas-qt and rpy2 for similar or equivalent functionality, see *here*

Check the *API Changes* and *deprecations* before updating.

What's new in v0.16.0

New features

DataFrame assign

Inspired by dplyr's `mutate` verb, DataFrame has a new *`assign()`* method. The function signature for `assign` is simply `**kwargs`. The keys are the column names for the new fields, and the values are either a value to be inserted (for example, a `Series` or NumPy array), or a function of one argument to be called on the `DataFrame`. The new values are inserted, and the entire DataFrame (with all original and new columns) is returned.

```
In [1]: iris = pd.read_csv('data/iris.data')

In [2]: iris.head()
Out[2]:
   SepalLength  SepalWidth  PetalLength  PetalWidth         Name
0          5.1         3.5          1.4         0.2  Iris-setosa
1          4.9         3.0          1.4         0.2  Iris-setosa
2          4.7         3.2          1.3         0.2  Iris-setosa
3          4.6         3.1          1.5         0.2  Iris-setosa
4          5.0         3.6          1.4         0.2  Iris-setosa
```

(continues on next page)