# Jupyter Data Science Workflow

### From exploratory analysis to reproducible science

*Jake VanderPlas University of Washington eScience Institute*

[2]

```
URL = 'https://data.seattle.gov/api/views/65db-xm6k/rows.csv?
accessType=DOWNLOAD'
```

[3]

```
from urllib.request import urlretrieve
urlretrieve(URL, 'Fremont.csv')
```

```
('Fremont.csv', <http.client.HTTPMessage at 0x10d6479e8>)
```

[4]

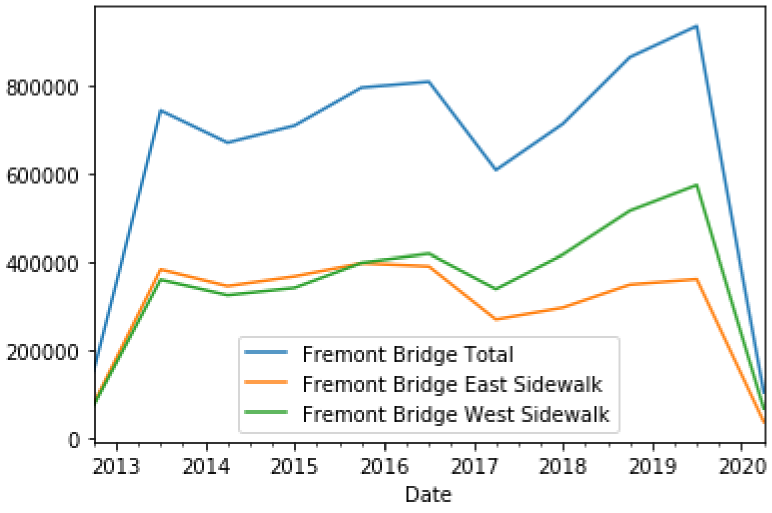
```
import pandas as pd
data = pd.read_csv('Fremont.csv', index_col='Date',
parse_dates=True)
data.head()
```

| | Fremont Bridge Total | Fremont Bridge East Sidewalk | Fremont Bridge West Sidewalk |
|---|---|---|---|
| **Date** | | | |
| 2012-10-03 00:00:00 | 13.0 | 4.0 | 9.0 |
| 2012-10-03 01:00:00 | 10.0 | 4.0 | 6.0 |
| 2012-10-03 02:00:00 | 2.0 | 1.0 | 1.0 |
| 2012-10-03 03:00:00 | 5.0 | 2.0 | 3.0 |
| 2012-10-03 04:00:00 | 7.0 | 6.0 | 1.0 |

[5]

```
%matplotlib inline
```

```
data.resample('3Q').sum().plot();
```



[6]

```
import matplotlib.pyplot as plt
plt.style.use('seaborn')

data.columns = ['West', 'East']

data.resample('W').sum().plot();
```

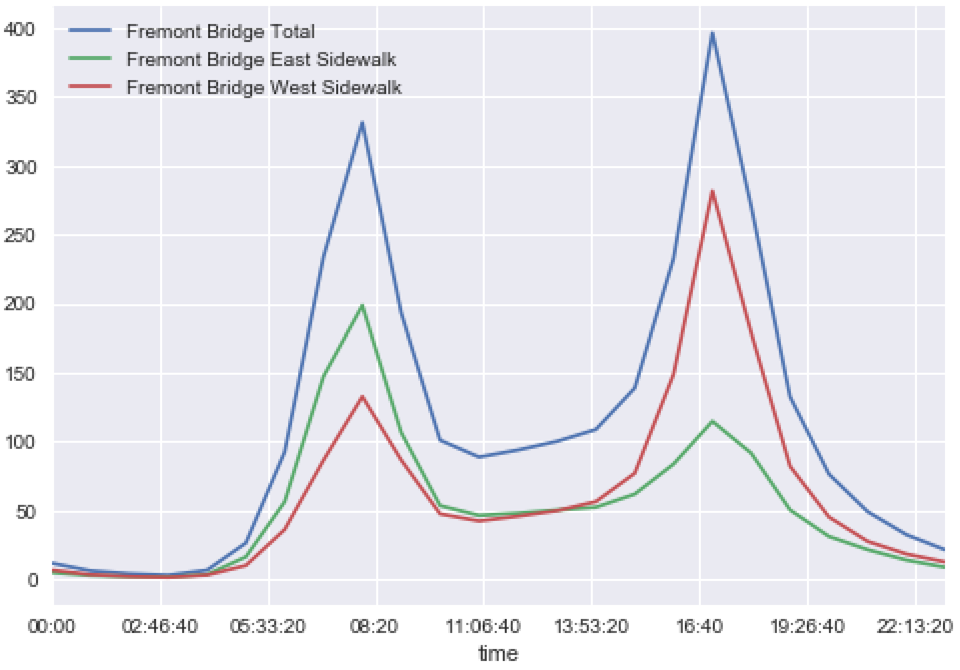
```
---------------------------------------------------------------------------
---
ValueError                                Traceback (most recent call
last)
<ipython-input-6-5bcb275115f6> in <module>
      2 plt.style.use('seaborn')
      3
----> 4 data.columns = ['West', 'East']
      5
      6 data.resample('W').sum().plot();

~/.anyenv/envs/pyenv/versions/3.7.2/lib/python3.7/site-
packages/pandas/core/generic.py in __setattr__(self, name, value)
   5190         try:
   5191             object.__getattribute__(self, name)
-> 5192             return object.__setattr__(self, name, value)
   5193         except AttributeError:
   5194             pass

pandas/_libs/properties.pyx in
pandas._libs.properties.AxisProperty.__set__()

~/.anyenv/envs/pyenv/versions/3.7.2/lib/python3.7/site-
packages/pandas/core/generic.py in _set_axis(self, axis, labels)
    688
    689     def _set_axis(self, axis, labels):
--> 690         self._data.set_axis(axis, labels)
    691         self._clear_item_cache()
    692

~/.anyenv/envs/pyenv/versions/3.7.2/lib/python3.7/site-
packages/pandas/core/internals/managers.py in set_axis(self, axis,
new_labels)
    181                 raise ValueError(
    182                     "Length mismatch: Expected axis has {old}
elements, new "
--> 183                     "values have {new} elements".format(old=old_len,
new=new_len)
    184                 )
    185

ValueError: Length mismatch: Expected axis has 3 elements, new values
have 2 elements
```

```
[7]   data['Total'] = data['West'] + data['East']

      ax = data.resample('D').sum().rolling(365).sum().plot();
      ax.set_ylim(0, None);
```

```
-----------------------------------------------------------------------
---
KeyError                                  Traceback (most recent call
last)
~/.anyenv/envs/pyenv/versions/3.7.2/lib/python3.7/site-
packages/pandas/core/indexes/base.py in get_loc(self, key, method,
tolerance)
   2896             try:
-> 2897                 return self._engine.get_loc(key)
   2898             except KeyError:

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'West'

During handling of the above exception, another exception occurred:

KeyError                                  Traceback (most recent call
last)
<ipython-input-7-402768f71e72> in <module>
----> 1 data['Total'] = data['West'] + data['East']
      2
      3 ax = data.resample('D').sum().rolling(365).sum().plot();
      4 ax.set_ylim(0, None);

~/.anyenv/envs/pyenv/versions/3.7.2/lib/python3.7/site-
packages/pandas/core/frame.py in __getitem__(self, key)
   2978             if self.columns.nlevels > 1:
   2979                 return self._getitem_multilevel(key)
-> 2980             indexer = self.columns.get_loc(key)
   2981             if is_integer(indexer):
   2982                 indexer = [indexer]

~/.anyenv/envs/pyenv/versions/3.7.2/lib/python3.7/site-
packages/pandas/core/indexes/base.py in get_loc(self, key, method,
tolerance)
   2897                 return self._engine.get_loc(key)
   2898             except KeyError:
-> 2899                 return
self._engine.get_loc(self._maybe_cast_indexer(key))
   2900         indexer = self.get_indexer([key], method=method,
tolerance=tolerance)
   2901         if indexer.ndim > 1 or indexer.size > 1:

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'West'
```

[8]

```
data.groupby(data.index.time).mean().plot();
```



```
[9]   pivoted = data.pivot_table('Total', index=data.index.time,
      columns=data.index.date)
      pivoted.iloc[:5, :5]
```

```
---------------------------------------------------------------------------
---
KeyError                                  Traceback (most recent call
last)
<ipython-input-9-06c92e263bae> in <module>
----> 1 pivoted = data.pivot_table('Total', index=data.index.time,
columns=data.index.date)
      2 pivoted.iloc[:5, :5]

~/.anyenv/envs/pyenv/versions/3.7.2/lib/python3.7/site-
packages/pandas/core/frame.py in pivot_table(self, values, index,
columns, aggfunc, fill_value, margins, dropna, margins_name, observed)
   6072             dropna=dropna,
   6073             margins_name=margins_name,
-> 6074             observed=observed,
   6075         )
   6076

~/.anyenv/envs/pyenv/versions/3.7.2/lib/python3.7/site-
packages/pandas/core/reshape/pivot.py in pivot_table(data, values,
index, columns, aggfunc, fill_value, margins, dropna, margins_name,
observed)
     70         for i in values:
     71             if i not in data:
---> 72                 raise KeyError(i)
     73
     74         to_filter = []
```

```
KeyError: 'Total'
```

```
[10] pivoted.plot(legend=False, alpha=0.01);
```

```
------------------------------------------------------------------------
---
NameError                                 Traceback (most recent call
last)
<ipython-input-10-c390894ea7ad> in <module>
----> 1 pivoted.plot(legend=False, alpha=0.01);

NameError: name 'pivoted' is not defined
```

```
[11]
```