In [2]:

```
%load_ext autoreload
%autoreload 2
```

In [3]:

```
%matplotlib inline

from fastai.imports import *
from fastai.structured import *
from pandas_summary import DataFrameSummary
from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier
from IPython.display import display
from sklearn import metrics
```

In [4]:

```
PATH='/Users/rsrivastava/Jupyter/KAGGLE/data/favorita-grocery-sales-forecasting/
'
```

In [5]:

```
!ls {PATH}
```

```
holidays_events.csv   splitFiles            train.csv
items.csv             stores.csv            trainaa.csv
oil.csv               test.csv              transactions.csv
sample_submission.csv testaa.csv
```

In [6]:

```
df_holiday=pd.read_csv(f'{PATH}holidays_events.csv')
```

In [7]:

```
df_holiday.head()
```

Out[7]:

	date	type	locale	locale_name	description	transferred
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2	2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
3	2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
4	2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

In [8]:

```
df_items=pd.read_csv(f'{PATH}items.csv')
```

In [9]:

```
df_items.head()
```

Out[9]:

	item_nbr	family	class	perishable
0	96995	GROCERY I	1093	0
1	99197	GROCERY I	1067	0
2	103501	CLEANING	3008	0
3	103520	GROCERY I	1028	0
4	103665	BREAD/BAKERY	2712	1

In [10]:

```
df_oil=pd.read_csv(f'{PATH}oil.csv')
```

In [11]:

```
df_oil.head()
```

Out[11]:

	date	dcoilwtico
0	2013-01-01	NaN
1	2013-01-02	93.14
2	2013-01-03	92.97
3	2013-01-04	93.12
4	2013-01-07	93.20

In [12]:

```
df_stores=pd.read_csv(f'{PATH}stores.csv')
```

In [13]:

```
df_stores.head()
```

Out[13]:

	store_nbr	city	state	type	cluster
0	1	Quito	Pichincha	D	13
1	2	Quito	Pichincha	D	13
2	3	Quito	Pichincha	D	8
3	4	Quito	Pichincha	D	9
4	5	Santo Domingo	Santo Domingo de los Tsachilas	D	4

In [20]:

```
df_test=pd.read_csv(f'{PATH}test.csv')
```

In [21]:

```
df_test.head()
```

Out[21]:

	id	date	store_nbr	item_nbr	onpromotion
0	125497040	2017-08-16	1	96995	False
1	125497041	2017-08-16	1	99197	False
2	125497042	2017-08-16	1	103501	False
3	125497043	2017-08-16	1	103520	False
4	125497044	2017-08-16	1	103665	False

In [22]:

```
df_test.shape
```

Out[22]:

```
(3370464, 5)
```

In [23]:

```
df_train=pd.read_csv(f'{PATH}train.csv')
```

```
/Users/rsrivastava/anaconda/lib/python3.6/site-packages/IPython/core
/interactiveshell.py:2698: DtypeWarning: Columns (5) have mixed type
s. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [25]:

```
df_train.head()
```

Out[25]:

	id	date	store_nbr	item_nbr	unit_sales	onpromotion
0	0	2013-01-01	25	103665	7.0	NaN
1	1	2013-01-01	25	105574	1.0	NaN
2	2	2013-01-01	25	105575	2.0	NaN
3	3	2013-01-01	25	108079	1.0	NaN
4	4	2013-01-01	25	108701	1.0	NaN

In [26]:

```
df_train.shape
```

Out[26]:

```
(125497040, 6)
```

In [15]:

```
df_train_sample=pd.read_csv(f'{PATH}trainaa.csv')
```

In [16]:

```
df_train_sample.head()
```

Out[16]:

	id	date	store_nbr	item_nbr	unit_sales	onpromotion
0	0	2013-01-01	25	103665	7.0	NaN
1	1	2013-01-01	25	105574	1.0	NaN
2	2	2013-01-01	25	105575	2.0	NaN
3	3	2013-01-01	25	108079	1.0	NaN
4	4	2013-01-01	25	108701	1.0	NaN

In [17]:

```
df_train_sample.shape
```

Out[17]:

```
(1199999, 6)
```

In [18]:

```
df_test_sample=pd.read_csv(f'{PATH}testaa.csv')
```

In [19]:

```
df_test_sample.head()
```

Out[19]:

	id	date	store_nbr	item_nbr	onpromotion
0	125497040	2017-08-16	1	96995	False
1	125497041	2017-08-16	1	99197	False
2	125497042	2017-08-16	1	103501	False
3	125497043	2017-08-16	1	103520	False
4	125497044	2017-08-16	1	103665	False

In [20]:

```
df_test_sample.shape
```

Out[20]:

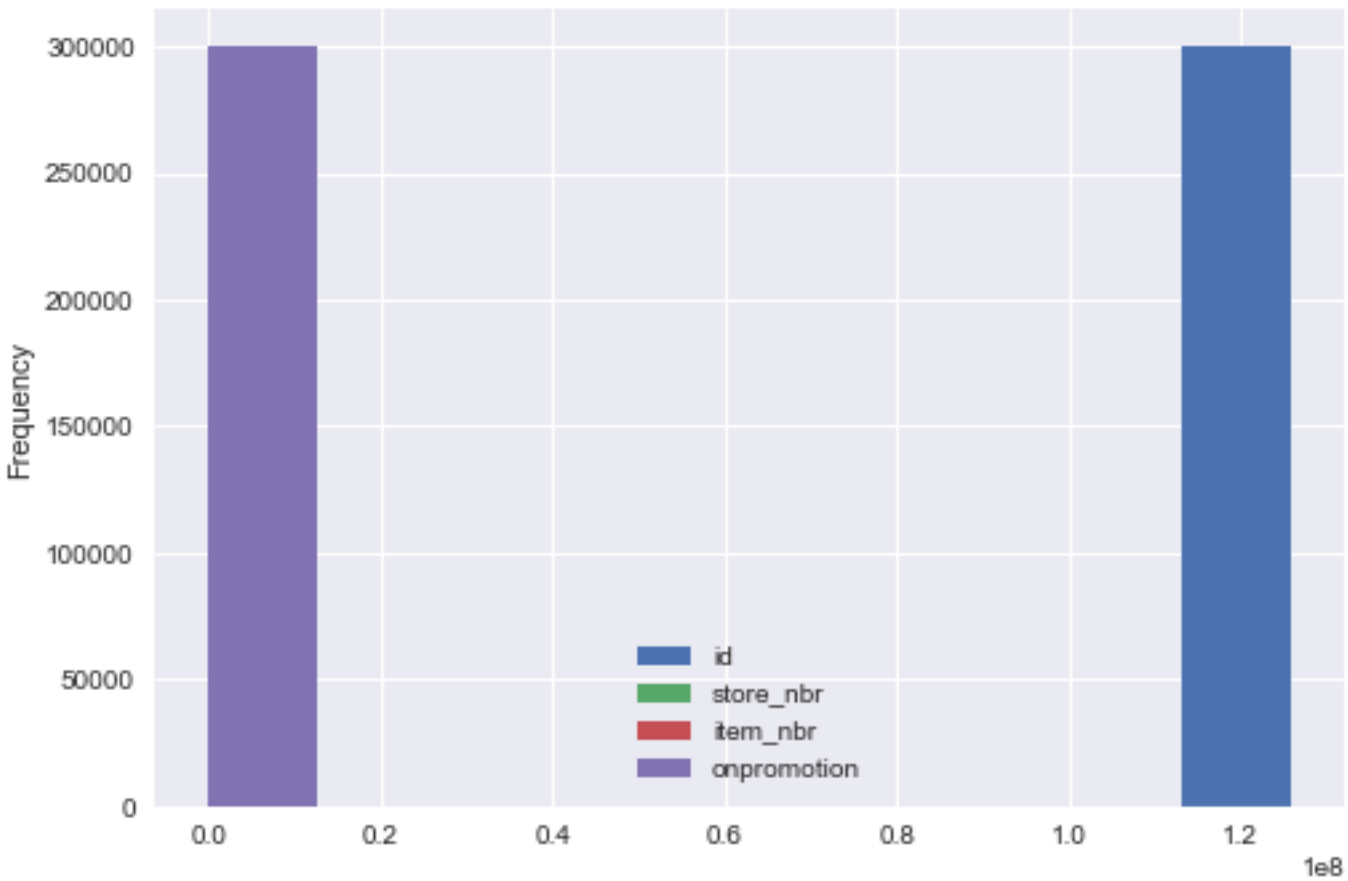
```
(299999, 5)
```

In [21]:

```
df_test_sample.plot.hist()
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11ad087f0>
```



In [44]:

```
types={'id': 'int64',
       'item_nbr': 'int32',
       'store_nbr': 'int8',
       'unit_sales': 'float32',
       'onpromotion': 'object'}
```

In [45]:

```
df_test_sample.columns
```

Out[45]:

```
Index(['id', 'date', 'store_nbr', 'item_nbr', 'onpromotion'], dtype=
'object')
```

In [46]:

```
%%time
df_all=pd.read_csv(f'{PATH}train.csv', parse_dates=['date'], dtype=types,infer_d
atetime_format=True)
```

```
CPU times: user 2min 6s, sys: 16.3 s, total: 2min 23s
Wall time: 2min 25s
```

In [47]:

```
df_all.columns
```

Out[47]:

```
Index(['id', 'date', 'store_nbr', 'item_nbr', 'unit_sales', 'onpromo
tion'], dtype='object')
```

In [48]:

```
??df_all.onpromotion
```

In [40]:

```
df_all.onpromotion.fillna(False, inplace=True)
```

In [41]:

```
df_all.onpromotion.map({'False':False,'True':True})
```

Out[41]:

```
0               NaN
1               NaN
2               NaN
3               NaN
4               NaN
5               NaN
6               NaN
7               NaN
8               NaN
9               NaN
10              NaN
11              NaN
12              NaN
13              NaN
14              NaN
15              NaN
16              NaN
17              NaN
18              NaN
19              NaN
20              NaN
21              NaN
22              NaN
23              NaN
24              NaN
25              NaN
26              NaN
27              NaN
28              NaN
29              NaN
              ...
```

```
125497010    False

125497011    False
125497012    False
125497013    False
125497014    False
125497015     True
125497016     True
125497017     True
125497018     True
125497019     True
125497020     True
125497021    False
125497022    False
125497023    False
125497024    False
125497025    False
125497026    False
125497027    False
125497028    False
125497029    False
125497030    False
125497031    False
125497032    False
125497033    False
125497034    False
125497035    False
125497036     True
125497037    False
125497038     True
125497039    False
Name: onpromotion, Length: 125497040, dtype: object
```

In [42]:

```
df_all.onpromotion.astype(bool)
```

Out[42]:

```
0            False
1            False
2            False
3            False
4            False
5            False
6            False
7            False
8            False
9            False
10           False
11           False
12           False
13           False
14           False
15           False
```

```
16           False
17           False
18           False
19           False
20           False
21           False
22           False
23           False
24           False
25           False
26           False
27           False
28           False
29           False
             ...
125497010     True
125497011     True
125497012     True
125497013     True
125497014     True
125497015     True
125497016     True
125497017     True
125497018     True
125497019     True
125497020     True
125497021     True
125497022     True
125497023     True
125497024     True
125497025     True
125497026     True
125497027     True
125497028     True
125497029     True
125497030     True
125497031     True
125497032     True
125497033     True
125497034     True
125497035     True
125497036     True
125497037     True
125497038     True
125497039     True
Name: onpromotion, Length: 125497040, dtype: bool
```

In [50]:

```
df_all.onpromotion.fillna(False,inplace=True)
```

In [51]:

```
df_all.onpromotion.astype(bool)
```

Out[51]:

```
0            False
1            False
2            False
3            False
4            False
5            False
6            False
7            False
8            False
9            False
10           False
11           False
12           False
13           False
14           False
15           False
16           False
17           False
18           False
19           False
20           False
21           False
22           False
23           False
24           False
25           False
26           False
27           False
28           False
29           False
             ...
125497010     True
125497011     True
125497012     True
125497013     True
125497014     True
125497015     True
125497016     True
125497017     True
125497018     True
125497019     True
125497020     True
125497021     True
125497022     True
125497023     True
125497024     True
125497025     True
125497026     True
```

```
125497027     True
125497028     True
125497029     True
125497030     True
125497031     True
125497032     True
125497033     True
125497034     True
125497035     True
125497036     True
125497037     True
125497038     True
125497039     True
Name: onpromotion, Length: 125497040, dtype: bool
```

In [52]:

```
%time df_all.to_feather('tmp/raw_groceries')
```

```
---------------------------------------------------------------------
-------
ValueError                                Traceback (most recent cal
l last)
<timed eval> in <module>()

~/anaconda/lib/python3.6/site-packages/pandas/core/frame.py in to_fe
ather(self, fname)
   1502         """
   1503         from pandas.io.feather_format import to_feather
-> 1504         to_feather(self, fname)
   1505
   1506     @Substitution(header='Write out column names. If a list
of string is given, \

~/anaconda/lib/python3.6/site-packages/pandas/io/feather_format.py
in to_feather(df, path)
     79         raise ValueError("feather must have string column na
mes")
     80
---> 81     feather.write_dataframe(df, path)
     82
     83

~/anaconda/lib/python3.6/site-packages/pyarrow/feather.py in write_f
eather(df, dest)
    116     writer = FeatherWriter(dest)
    117     try:
--> 118         writer.write(df)
    119     except:
    120         # Try to make sure the resource is closed

~/anaconda/lib/python3.6/site-packages/pyarrow/feather.py in write(s
elf, df)
     91                     inferred_type = infer_dtype(col[col.notn
ull()])
     92                     if inferred_type in ['mixed']:
---> 93                         raise ValueError(msg)
     94
     95                 elif inferred_type not in ['unicode', 'strin
g']:

ValueError: cannot serialize column 5 named onpromotion with dtype m
ixed


In [26]:
```

```
types={'id': 'int64',
      'item_nbr': 'int32',
      'store_nbr': 'int8',
      'unit_sales': 'float32',
      'onpromotion': 'object'}
```

In [55]:

```
%%time
df_all.dtypes
```

```
CPU times: user 233 µs, sys: 25 µs, total: 258 µs
Wall time: 262 µs
```

Out[55]:

```
id                     int64
date           datetime64[ns]
store_nbr               int8
item_nbr               int32
unit_sales           float32
onpromotion           object
dtype: object
```

In [56]:

```

```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent cal
l last)
<ipython-input-56-fbab945b1b85> in <module>()
----> 1 df_all.astype(np.bool, inplace=True)

~/anaconda/lib/python3.6/site-packages/pandas/util/_decorators.py in
wrapper(*args, **kwargs)
     89                 else:
     90                     kwargs[new_arg_name] = new_arg_value
---> 91             return func(*args, **kwargs)
     92         return wrapper
     93     return _deprecate_kwarg

~/anaconda/lib/python3.6/site-packages/pandas/core/generic.py in ast
ype(self, dtype, copy, errors, **kwargs)
   3408         # else, only a single dtype is given
   3409         new_data = self._data.astype(dtype=dtype, copy=copy,
errors=errors,
-> 3410                                      **kwargs)
   3411         return self._constructor(new_data).__finalize__(self
)
   3412

~/anaconda/lib/python3.6/site-packages/pandas/core/internals.py in a
stype(self, dtype, **kwargs)
   3222
   3223     def astype(self, dtype, **kwargs):
-> 3224         return self.apply('astype', dtype=dtype, **kwargs)
   3225
   3226     def convert(self, **kwargs):
```

```
~/anaconda/lib/python3.6/site-packages/pandas/core/internals.py in a
pply(self, f, axes, filter, do_integrity_check, consolidate, **kwarg
s)
   3089
   3090             kwargs['mgr'] = self
-> 3091             applied = getattr(b, f)(**kwargs)
   3092             result_blocks = _extend_blocks(applied,
result_blocks)
   3093

~/anaconda/lib/python3.6/site-packages/pandas/core/internals.py in a
stype(self, dtype, copy, errors, values, **kwargs)
    469     def astype(self, dtype, copy=False, errors='raise',
values=None, **kwargs):
    470         return self._astype(dtype, copy=copy, errors=errors,
values=values,
--> 471                             **kwargs)
    472
    473     def _astype(self, dtype, copy=False, errors='raise', val
ues=None,

~/anaconda/lib/python3.6/site-packages/pandas/core/internals.py in _
astype(self, dtype, mgr, **kwargs)
   2241
   2242         # delegate
-> 2243         return super(DatetimeBlock, self)._astype(dtype=dtyp
e, **kwargs)
   2244
   2245     def _can_hold_element(self, element):

~/anaconda/lib/python3.6/site-packages/pandas/core/internals.py in _
astype(self, dtype, copy, errors, values, klass, mgr, raise_on_error
, **kwargs)
    519
    520                 # _astype_nansafe works fine with 1-d only
--> 521                 values = astype_nansafe(values.ravel(),
dtype, copy=True)
    522                 values = values.reshape(self.shape)
    523

~/anaconda/lib/python3.6/site-packages/pandas/core/dtypes/cast.py in
astype_nansafe(arr, dtype, copy)
    591         elif dtype != _NS_DTYPE:
    592             raise TypeError("cannot astype a datetimelike fr
om [%s] to [%s]" %
--> 593                             (arr.dtype, dtype))
    594         return arr.astype(_NS_DTYPE)
    595     elif is_timedelta64_dtype(arr):

TypeError: cannot astype a datetimelike from [datetime64[ns]] to [bo
ol]
```

In [58]:

```
df_all.onpromotion.astype(bool, inplace=True)
```

Out[58]:

```
0            False
1            False
2            False
3            False
4            False
5            False
6            False
7            False
8            False
9            False
10           False
11           False
12           False
13           False
14           False
15           False
16           False
17           False
18           False
19           False
20           False
21           False
22           False
23           False
24           False
25           False
26           False
27           False
28           False
29           False
             ...
125497010     True
125497011     True
125497012     True
125497013     True
125497014     True
125497015     True
125497016     True
125497017     True
125497018     True
125497019     True
125497020     True
125497021     True
125497022     True
125497023     True
125497024     True
125497025     True
125497026     True
```

```
125497027     True
125497028     True
125497029     True
125497030     True
125497031     True
125497032     True
125497033     True
125497034     True
125497035     True
125497036     True
125497037     True
125497038     True
125497039     True
Name: onpromotion, Length: 125497040, dtype: bool
```

In [61]:

```
df_all['onpromotion'] = df_all['onpromotion'].astype('bool')
```

In [62]:

```
df_all.dtypes
```

Out[62]:

```
id                     int64
date          datetime64[ns]
store_nbr               int8
item_nbr               int32
unit_sales           float32
onpromotion             bool
dtype: object
```

In [64]:

```
%%time
df_all.to_feather('tmp/raw_groceries')
```

```
CPU times: user 1.16 s, sys: 2.25 s, total: 3.41 s
Wall time: 4.11 s
```

In [18]:

```
%%time
df_all = pd.read_feather('tmp/raw_groceries')
```

```
CPU times: user 1.39 s, sys: 2.29 s, total: 3.68 s
Wall time: 7.28 s
```

In [19]:

```
%%time
df_all.describe(include='all')
```

Out[19]:

	id	date	store_nbr	item_nbr	unit_sales	onpron
count	1.254970e+08	125497040	1.254970e+08	1.254970e+08	1.254970e+08	125497
unique	NaN	1684	NaN	NaN	NaN	2
top	NaN	2017-07-01 00:00:00	NaN	NaN	NaN	True
freq	NaN	118194	NaN	NaN	NaN	103839
first	NaN	2013-01-01 00:00:00	NaN	NaN	NaN	NaN
last	NaN	2017-08-15 00:00:00	NaN	NaN	NaN	NaN
mean	6.274852e+07	NaN	2.746458e+01	9.727692e+05	5.319669e+00	NaN
std	3.622788e+07	NaN	1.633051e+01	5.205336e+05	2.306714e+01	NaN
min	0.000000e+00	NaN	1.000000e+00	9.699500e+04	-1.537200e+04	NaN
25%	3.137426e+07	NaN	1.200000e+01	5.223830e+05	2.000000e+00	NaN
50%	6.274852e+07	NaN	2.800000e+01	9.595000e+05	4.000000e+00	NaN
75%	9.412278e+07	NaN	4.300000e+01	1.354380e+06	9.000000e+00	NaN
max	1.254970e+08	NaN	5.400000e+01	2.127114e+06	8.944000e+04	NaN

In [24]:

```
df_test.describe(include='all')
```

Out[24]:

	id	date	store_nbr	item_nbr	onpromotion
count	3.370464e+06	3370464	3.370464e+06	3.370464e+06	3370464
unique	NaN	16	NaN	NaN	2
top	NaN	2017-08-22	NaN	NaN	False
freq	NaN	210654	NaN	NaN	3171867
mean	1.271823e+08	NaN	2.750000e+01	1.244798e+06	NaN
std	9.729693e+05	NaN	1.558579e+01	5.898362e+05	NaN
min	1.254970e+08	NaN	1.000000e+00	9.699500e+04	NaN
25%	1.263397e+08	NaN	1.400000e+01	8.053210e+05	NaN
50%	1.271823e+08	NaN	2.750000e+01	1.294665e+06	NaN
75%	1.280249e+08	NaN	4.100000e+01	1.730015e+06	NaN
max	1.288675e+08	NaN	5.400000e+01	2.134244e+06	NaN

In [27]:

```
df_test=pd.read_csv(f'{PATH}test.csv', parse_dates=['date'], dtype=types, infer_
datetime_format=True)
```

In [28]:

```
df_test.onpromotion.fillna(False, inplace=True)
```

In [29]:

```
df_test.onpromotion = df_test.onpromotion.map({'False': False,'True':True})
```

In [30]:

```
df_test.onpromotion=df_test.onpromotion.astype(bool)
```

In [32]:

```
df_test.dtypes
```

Out[32]:

```
id                      int64
date           datetime64[ns]
store_nbr                int8
item_nbr                int32
onpromotion              bool
dtype: object
```

In [33]:

```
df_test.describe(include='all')
```

Out[33]:

	id	date	store_nbr	item_nbr	onpromotion
count	3.370464e+06	3370464	3.370464e+06	3.370464e+06	3370464
unique	NaN	16	NaN	NaN	2
top	NaN	2017-08-27 00:00:00	NaN	NaN	False
freq	NaN	210654	NaN	NaN	3171867
first	NaN	2017-08-16 00:00:00	NaN	NaN	NaN
last	NaN	2017-08-31 00:00:00	NaN	NaN	NaN
mean	1.271823e+08	NaN	2.750000e+01	1.244798e+06	NaN
std	9.729693e+05	NaN	1.558579e+01	5.898362e+05	NaN
min	1.254970e+08	NaN	1.000000e+00	9.699500e+04	NaN
25%	1.263397e+08	NaN	1.400000e+01	8.053210e+05	NaN
50%	1.271823e+08	NaN	2.750000e+01	1.294665e+06	NaN
75%	1.280249e+08	NaN	4.100000e+01	1.730015e+06	NaN
max	1.288675e+08	NaN	5.400000e+01	2.134244e+06	NaN

In [34]:

```
df_all.tail()
```

Out[34]:

	id	date	store_nbr	item_nbr	unit_sales	onpromotion
125497035	125497035	2017-08-15	54	2089339	4.0	True
125497036	125497036	2017-08-15	54	2106464	1.0	True
125497037	125497037	2017-08-15	54	2110456	192.0	True
125497038	125497038	2017-08-15	54	2113914	198.0	True
125497039	125497039	2017-08-15	54	2116416	2.0	True

In [37]:

```
df_all.unit_sales=np.log1p(np.clip(df_all.unit_sales,0,None))
```

In [38]:

```
%time
add_datepart(df_all,'date')
```

```
CPU times: user 2 µs, sys: 0 ns, total: 2 µs
Wall time: 30 µs
```

In [39]:

```
def split_vals(a,n):
    return a[:n].copy(), a[n:].copy()
```

In [40]:

```
n_valid= len(df_test)
```

In [41]:

```
n_train=len(df_all)-n_valid
```

In [42]:

```
train, valid = split_vals(df_all, n_train)
```

In [43]:

```
train.shape
```

Out[43]:

```
(122126576, 18)
```

In [44]:

```
valid.shape
```

Out[44]:

```
(3370464, 18)
```

In [46]:

```
%%time
trn, y = proc_df(train, 'unit_sales')
```

```
---------------------------------------------------------------------
-------
ValueError                                Traceback (most recent cal
l last)
<timed exec> in <module>()

ValueError: too many values to unpack (expected 2)
```

In [47]:

```
val, y_val= proc_df(valid, 'unit_sales')
```

```
---------------------------------------------------------------------
-------
ValueError                                Traceback (most recent cal
l last)
<ipython-input-47-9a50101c5c46> in <module>()
----> 1 val, y_val= proc_df(valid, 'unit_sales')

ValueError: too many values to unpack (expected 2)
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: