In [0]:

```
import pandas as pd
import os
import numpy as np
```

In [0]:

```
preprocess = True

filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
dataset = pd.read_csv(filename_read,na_values=['NA','?'])
```

In [0]:

```
dataset.head()
```

Out[0]:

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	hmax	hm
0	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 00:00:00	2007-11-06	...	29.7	16.8	25.5	10.8	35.0	58.0	32
1	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 01:00:00	2007-11-06	...	29.9	13.6	29.0	12.2	39.0	39.0	35
2	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 02:00:00	2007-11-06	...	29.0	14.0	27.4	13.6	44.0	44.0	39
3	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 03:00:00	2007-11-06	...	27.4	16.9	25.8	14.1	58.0	58.0	44
4	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 04:00:00	2007-11-06	...	26.3	17.0	25.3	16.4	57.0	58.0	56

5 rows × 31 columns

In [0]:

```
dataset.columns
```

Out[0]:

```
Index(['wsid', 'wsnm', 'elvt', 'lat', 'lon', 'inme', 'city', 'prov', 'mdct',
       'date', 'yr', 'mo', 'da', 'hr', 'prcp', 'stp', 'smax', 'smin', 'gbrd',
       'temp', 'dewp', 'tmax', 'dmax', 'tmin', 'dmin', 'hmdy', 'hmax', 'hmin',
       'wdsp', 'wdct', 'gust'],
      dtype='object')
```

In [0]:

```
data = dataset.drop(['wsid','elvt','wsnm','city','prov','inme','yr','mo','da','hr',],axis=1)
```

In [0]:

```
data = data.drop(['mdct','date'],axis=1)
```

In [0]:

```
data = data.drop(['prcp','gbrd','lat','lon','temp','dewp','tmax','tmin','dmin','hmin','wdsp','gust'
],axis=1)
```

In [0]:

```
data.head()
```

Out[0]:

	stp	smax	smin	dmax	hmdy	hmax	wdct
0	982.5	982.5	981.3	16.8	35.0	58.0	101.0
1	983.2	983.2	982.5	13.6	39.0	39.0	94.0
2	983.5	983.5	983.2	14.0	44.0	44.0	93.0
3	983.7	983.7	983.4	16.9	58.0	58.0	96.0
4	983.7	983.8	983.6	17.0	57.0	58.0	110.0

In [0]:

```
print('Number of instances = %d' % (data.shape[0]))
print('Number of attributes = %d' % (data.shape[1]))
```

```
Number of instances = 9779168
Number of attributes = 7
```

Missing Values

In [0]:

```
data = data.replace('?',np.NaN)

print('Number of instances = %d' % (data.shape[0]))
print('Number of attributes = %d' % (data.shape[1]))

print('Number of missing values:')
for col in data.columns:
    print('\t%s: %d' % (col,data[col].isna().sum()))
```

```
Number of instances = 9779168
Number of attributes = 7
Number of missing values:
 stp: 0
 smax: 0
 smin: 0
 dmax: 310
 hmdy: 0
 hmax: 12
 wdct: 0
```

Method 1

In [0]:

```
data2 = data[['dmax','hmax']]

data2 = data2.fillna(data2.median())

print('Number of missing values:')
for col in data2.columns:
    print('\t%s: %d' % (col,data2[col].isna().sum()))
```

```
Number of missing values:
 dmax: 0
 hmax: 0
```

Method 2

In [0]:

```
print('Number of rows in original data = %d' % (data.shape[0]))

data2 = data.dropna()
print('Number of rows after discarding missing values = %d' % (data2.shape[0]))
```

```
Number of rows in original data = 9779168
Number of rows after discarding missing values = 9778846
```

Outliers

In [0]:

```
%matplotlib inline

data2.head()
```

Out[0]:

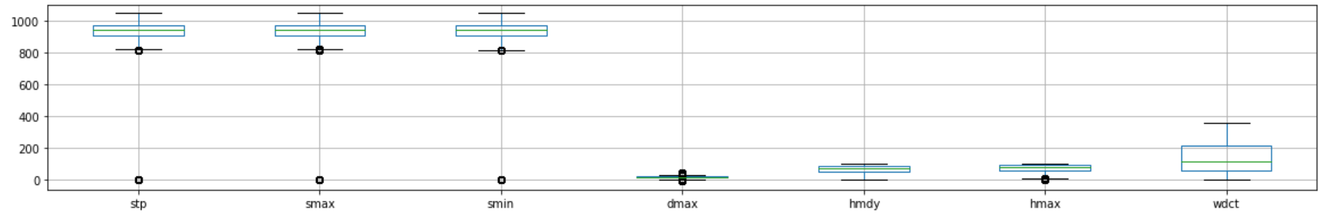
	stp	smax	smin	dmax	hmdy	hmax	wdct
0	982.5	982.5	981.3	16.8	35.0	58.0	101.0
1	983.2	983.2	982.5	13.6	39.0	39.0	94.0
2	983.5	983.5	983.2	14.0	44.0	44.0	93.0
3	983.7	983.7	983.4	16.9	58.0	58.0	96.0
4	983.7	983.8	983.6	17.0	57.0	58.0	110.0

In [0]:

```
data2.boxplot(figsize=(20,3))
```

Out[0]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11d624710>
```



In [0]:

```
z=(data2-data2.mean())/data2.std()
z[20:25]
```

Out[0]:

	stp	smax	smin	dmax	hmdy	hmax	wdct
20	0.407509	0.406947	0.407870	-0.126179	-1.404147	-1.398670	-0.309873
21	0.409523	0.408956	0.409076	0.095407	-1.027376	-1.058186	-0.347895
22	0.411940	0.411366	0.410683	0.027226	-1.441824	-1.058186	-0.148278
23	0.416773	0.416187	0.413497	-0.330721	-1.215761	-1.323007	-0.252839
24	0.419996	0.419401	0.418321	-0.023909	-1.027376	-1.133849	-0.281356

In [0]:

```
print('Number of rows before discarding outliers = %d' % (z.shape[0]))

Z2 = z.loc[((z > -3).sum(axis=1)==7) & ((z <= 3).sum(axis=1)==7),:]
```

```
print('Number of rows after discarding missing values = %d' % (Z2.shape[0]))
```

```
Number of rows before discarding outliers = 9778846
Number of rows after discarding missing values = 9069969
```

Duplicate Rows

In [0]:

```
dups = data.duplicated()
print('Number of duplicate rows = %d' % (dups.sum()))
data.loc[[1,2]]
```

```
Number of duplicate rows = 623353
```

Out[0]:

	stp	smax	smin	dmax	hmdy	hmax	wdct
1	983.2	983.2	982.5	13.6	39.0	39.0	94.0
2	983.5	983.5	983.2	14.0	44.0	44.0	93.0

Shuffling Dataframes

In [0]:

```
import os
import numpy as np
import pandas as pd


filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
df = pd.read_csv(filename_read,na_values=['NA','?'])
#np.random.seed(30) # Uncomment this line to get the same shuffle each time

df = df.reindex(np.random.permutation(df.index))
df.reset_index(inplace=True, drop=True)
df
```

Out[0]:

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	h
0	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 00:00:00	2007-11-06	...	29.7	16.8	25.5	10.8	35.0	
1	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 01:00:00	2007-11-06	...	29.9	13.6	29.0	12.2	39.0	
2	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 02:00:00	2007-11-06	...	29.0	14.0	27.4	13.6	44.0	
3	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 03:00:00	2007-11-06	...	27.4	16.9	25.8	14.1	58.0	
4	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 04:00:00	2007-11-06	...	26.3	17.0	25.3	16.4	57.0	
5	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 05:00:00	2007-11-06	...	25.4	16.4	23.8	16.0	62.0	
6	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 06:00:00	2007-11-06	...	23.8	16.7	22.0	16.2	72.0	

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	h
7	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 07:00:00	2007-11-06	...	22.0	17.8	19.5	16.6	86.0	
8	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 08:00:00	2007-11-06	...	19.7	17.3	18.3	16.9	93.0	
9	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 09:00:00	2007-11-06	...	22.9	18.3	18.2	17.1	75.0	
10	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 10:00:00	2007-11-06	...	25.1	18.4	22.9	17.0	61.0	
11	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 11:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
12	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 12:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
13	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 13:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
14	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 14:00:00	2007-11-06	...	31.8	16.0	30.0	14.3	36.0	
15	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 15:00:00	2007-11-06	...	33.0	15.4	31.0	13.6	32.0	
16	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 16:00:00	2007-11-06	...	34.0	15.6	32.5	12.9	31.0	
17	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 17:00:00	2007-11-06	...	34.7	14.6	33.4	12.2	29.0	
18	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 18:00:00	2007-11-06	...	35.2	14.2	33.9	12.6	27.0	
19	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 19:00:00	2007-11-06	...	35.1	14.5	33.7	12.6	28.0	
20	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 20:00:00	2007-11-06	...	34.7	14.5	32.2	12.8	30.0	
21	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 21:00:00	2007-11-06	...	32.7	15.8	29.9	12.5	40.0	
22	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 22:00:00	2007-11-06	...	31.7	15.4	29.4	11.3	29.0	
23	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 23:00:00	2007-11-06	...	31.5	13.3	29.8	11.4	35.0	
24	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 00:00:00	2007-11-07	...	31.0	15.1	30.2	13.4	40.0	
25	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 01:00:00	2007-11-07	...	30.3	15.1	29.3	13.1	37.0	
26	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 02:00:00	2007-11-07	...	29.3	14.0	28.1	13.1	42.0	
27	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 03:00:00	2007-11-07	...	28.1	15.5	26.5	14.0	51.0	
28	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 04:00:00	2007-11-07	...	26.6	16.4	25.1	15.5	58.0	
29	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 05:00:00	2007-11-07	...	25.2	16.4	23.7	15.3	59.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9779138	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 18:00:00	2016-09-29	...	24.6	13.0	21.7	11.6	53.0	

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	h
9779139	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 19:00:00	2016-09-29	...	22.1	12.9	20.0	11.7	61.0	
9779140	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 20:00:00	2016-09-29	...	20.2	12.6	16.8	11.7	72.0	
9779141	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 21:00:00	2016-09-29	...	16.9	12.1	15.3	11.5	79.0	
9779142	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 22:00:00	2016-09-29	...	15.3	12.4	14.2	11.6	84.0	
9779143	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 23:00:00	2016-09-29	...	14.7	11.8	14.2	10.2	75.0	
9779144	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 00:00:00	2016-09-30	...	14.9	11.4	14.5	10.3	81.0	
9779145	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 01:00:00	2016-09-30	...	14.8	11.4	14.2	9.9	75.0	
9779146	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 02:00:00	2016-09-30	...	14.7	10.2	14.3	9.2	73.0	
9779147	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 03:00:00	2016-09-30	...	14.8	11.4	14.3	9.6	80.0	
9779148	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 04:00:00	2016-09-30	...	14.9	12.3	14.7	11.3	84.0	
9779149	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 05:00:00	2016-09-30	...	14.9	12.2	14.8	10.9	77.0	
9779150	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 06:00:00	2016-09-30	...	14.9	11.6	14.6	10.9	81.0	
9779151	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 07:00:00	2016-09-30	...	14.8	11.8	14.5	11.4	81.0	
9779152	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 08:00:00	2016-09-30	...	14.9	11.8	14.6	11.3	80.0	
9779153	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 09:00:00	2016-09-30	...	14.9	11.7	14.2	11.3	82.0	
9779154	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 10:00:00	2016-09-30	...	15.8	11.4	14.3	11.1	74.0	
9779155	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 11:00:00	2016-09-30	...	17.7	12.0	15.6	11.0	69.0	
9779156	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 12:00:00	2016-09-30	...	19.3	12.0	17.1	10.6	60.0	
9779157	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 13:00:00	2016-09-30	...	20.5	12.2	18.2	10.6	58.0	
9779158	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 14:00:00	2016-09-30	...	21.4	12.5	19.4	9.8	55.0	
9779159	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 15:00:00	2016-09-30	...	21.8	12.1	19.9	10.6	54.0	
9779160	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 16:00:00	2016-09-30	...	21.4	12.8	20.2	11.5	59.0	
9779161	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 17:00:00	2016-09-30	...	21.2	12.8	19.3	11.5	64.0	
9779162	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 18:00:00	2016-09-30	...	19.5	12.8	18.0	11.8	67.0	

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	h
9779163	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 19:00:00	2016-09-30	...	18.2	12.4	16.3	11.8	76.0	
9779164	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 20:00:00	2016-09-30	...	16.8	12.5	15.3	11.7	80.0	
9779165	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 21:00:00	2016-09-30	...	15.3	11.9	14.9	11.5	79.0	
9779166	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 22:00:00	2016-09-30	...	15.0	11.7	14.4	11.4	82.0	
9779167	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 23:00:00	2016-09-30	...	14.6	11.5	14.3	11.2	82.0	

9779168 rows × 31 columns

Sorting Dataframes

In [0]:

```
df = df.sort_values(by='wsnm',ascending=True)
df
```

Out[0]:

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmd
750984	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-02-25 00:00:00	2013-02-25	...	21.6	19.2	20.6	18.8	90.
756893	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 05:00:00	2013-10-29	...	18.6	12.9	18.3	12.2	68.
756892	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 04:00:00	2013-10-29	...	18.8	13.5	18.5	12.7	69.
756891	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 03:00:00	2013-10-29	...	19.4	14.0	18.8	13.4	71.
756890	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 02:00:00	2013-10-29	...	19.8	14.1	19.4	13.8	70.
756889	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 01:00:00	2013-10-29	...	19.9	14.2	19.7	14.0	70.
756888	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 00:00:00	2013-10-29	...	20.0	14.4	19.9	14.1	70.
756887	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 23:00:00	2013-10-28	...	20.0	14.8	19.9	14.3	70.
756886	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 22:00:00	2013-10-28	...	20.1	14.8	19.9	14.6	73.
756885	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 21:00:00	2013-10-28	...	20.5	14.7	20.0	14.5	71.
756884	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 20:00:00	2013-10-28	...	20.7	15.3	20.5	14.7	69.
756894	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 06:00:00	2013-10-29	...	18.4	13.1	17.9	12.4	72.
756883	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 19:00:00	2013-10-28	...	21.7	15.6	20.6	15.0	72.
756881	311	AFONSO	507.0	-	-	A657	Afonso	ES	2013-10-28	2013-		23.2	16.7	22.3	15.7	64.

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmd
756881	311	CLAUDIO	507.0	20.104194	41.106861	A657	Cláudio	ES	28 17:00:00	10-28	...	23.2	16.7	22.3	15.7	64.
756880	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 16:00:00	2013-10-28	...	24.3	16.6	23.0	15.8	67.
756879	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 15:00:00	2013-10-28	...	25.0	17.2	23.7	16.4	62.
756878	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 14:00:00	2013-10-28	...	27.2	17.5	24.0	15.8	63.
756877	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 13:00:00	2013-10-28	...	27.3	17.4	25.2	15.6	56.
756876	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 12:00:00	2013-10-28	...	25.6	17.1	23.2	15.7	57.
756875	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 11:00:00	2013-10-28	...	23.6	16.6	22.0	15.7	64.
756874	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 10:00:00	2013-10-28	...	22.4	16.5	21.3	15.7	67.
756873	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 09:00:00	2013-10-28	...	21.4	16.6	20.3	16.3	73.
756872	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 08:00:00	2013-10-28	...	21.1	17.0	20.5	16.6	78.
756882	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-28 18:00:00	2013-10-28	...	23.1	15.9	21.6	15.4	68.
756895	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 07:00:00	2013-10-29	...	18.0	13.0	17.8	12.8	72.
756896	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 08:00:00	2013-10-29	...	18.0	12.9	17.9	12.7	72.
756897	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-29 09:00:00	2013-10-29	...	18.2	13.2	17.9	12.8	72.
756920	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-30 08:00:00	2013-10-30	...	17.2	13.2	16.9	12.3	78.
756919	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-30 07:00:00	2013-10-30	...	17.5	13.3	17.1	11.7	75.
756918	311	AFONSO CLAUDIO	507.0	-20.104194	-41.106861	A657	Afonso Cláudio	ES	2013-10-30 06:00:00	2013-10-30	...	17.4	12.8	17.1	12.1	74.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
4851210	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-27 10:00:00	2012-12-27	...	18.4	14.9	14.2	12.7	79.
4851211	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-27 11:00:00	2012-12-27	...	22.5	17.1	18.4	14.3	70.
4851212	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-27 12:00:00	2012-12-27	...	26.0	17.2	22.4	15.8	54.
4851203	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-27 03:00:00	2012-12-27	...	19.4	14.2	16.7	12.7	80.
4851191	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 15:00:00	2012-12-26	...	30.6	14.3	28.4	10.4	30.
4851190	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 14:00:00	2012-12-26	...	28.9	15.5	26.3	13.4	40.
4851189	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 13:00:00	2012-12-26	...	26.8	16.3	24.8	14.9	49.
									2012-12-							

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmd
4851168	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-25 16:00:00	2012-12-25	...	29.9	13.1	27.2	13.4	39.
4851169	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-25 17:00:00	2012-12-25	...	30.5	14.4	28.7	12.3	37.
4851170	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-25 18:00:00	2012-12-25	...	31.2	13.7	28.7	11.6	34.
4851171	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-25 19:00:00	2012-12-25	...	30.9	13.4	29.5	12.0	36.
4851172	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-25 20:00:00	2012-12-25	...	30.3	13.5	29.6	12.4	35.
4851173	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-25 21:00:00	2012-12-25	...	30.0	13.7	28.5	12.4	40.
4851174	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-25 22:00:00	2012-12-25	...	28.6	14.5	25.1	13.6	52.
4851175	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-25 23:00:00	2012-12-25	...	25.4	15.4	22.3	14.0	65.
4851176	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 00:00:00	2012-12-26	...	22.4	16.0	20.2	15.2	76.
4851177	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 01:00:00	2012-12-26	...	20.2	16.4	19.1	15.7	83.
4851178	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 02:00:00	2012-12-26	...	19.4	16.4	18.1	15.5	84.
4851179	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 03:00:00	2012-12-26	...	19.2	16.0	17.8	15.0	84.
4851180	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 04:00:00	2012-12-26	...	18.9	16.0	17.0	15.0	88.
4851181	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 05:00:00	2012-12-26	...	17.7	15.6	16.1	14.8	92.
4851182	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 06:00:00	2012-12-26	...	16.2	15.0	15.6	14.4	92.
4851183	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 07:00:00	2012-12-26	...	15.9	14.6	15.1	13.8	92.
4851184	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 08:00:00	2012-12-26	...	15.2	14.0	14.6	13.4	92.
4851185	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 09:00:00	2012-12-26	...	15.4	14.0	13.9	12.8	91.
4851186	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 10:00:00	2012-12-26	...	18.8	15.3	15.4	13.5	80.
4851187	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 11:00:00	2012-12-26	...	22.5	16.6	18.8	15.2	68.
4851188	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-26 12:00:00	2012-12-26	...	25.3	16.4	22.5	15.2	56.
4851214	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2012-12-27 14:00:00	2012-12-27	...	29.5	15.6	27.3	11.3	36.
4859823	357	ÁGUAS VERMELHAS	754.0	-15.751536	-41.457787	A549	Águas Vermelhas	MG	2013-12-21 07:00:00	2013-12-21	...	19.9	19.3	19.8	19.2	96.

9779168 rows × 31 columns

Saving a Dataframe

In [0]:

```
print("The first record is: {}".format(df['wsnm'].loc[0]))
```

```
The first record is: SÃO GONÇALO
```

In [0]:

```
import os
import pandas as pd
import numpy as np

path = "./data/"

filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
filename_write = os.path.join("/Users/surajrawat/sudeste_new.csv")
df = pd.read_csv(filename_read,na_values=['NA','?'])
df = df.reindex(np.random.permutation(df.index))
df.to_csv(filename_write,index=False)    # Specify index = false to not write row numbers
print("Done")
```

```
Done
```

Dropping Fields

In [0]:

```
import os
import pandas as pd
import numpy as np
filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
df = pd.read_csv(filename_read,na_values=['NA','?'])
print("Before drop: {}".format(df.columns))
df.drop('wsnm', axis=1, inplace=True)
print("After drop: {}".format(df.columns))
df[0:5]
```

```
Before drop: Index(['wsid', 'wsnm', 'elvt', 'lat', 'lon', 'inme', 'city', 'prov', 'mdct',
       'date', 'yr', 'mo', 'da', 'hr', 'prcp', 'stp', 'smax', 'smin', 'gbrd',
       'temp', 'dewp', 'tmax', 'dmax', 'tmin', 'dmin', 'hmdy', 'hmax', 'hmin',
       'wdsp', 'wdct', 'gust'],
      dtype='object')
After drop: Index(['wsid', 'elvt', 'lat', 'lon', 'inme', 'city', 'prov', 'mdct', 'date',
       'yr', 'mo', 'da', 'hr', 'prcp', 'stp', 'smax', 'smin', 'gbrd', 'temp',
       'dewp', 'tmax', 'dmax', 'tmin', 'dmin', 'hmdy', 'hmax', 'hmin', 'wdsp',
       'wdct', 'gust'],
      dtype='object')
```

Out[0]:

	wsid	elvt	lat	lon	inme	city	prov	mdct	date	yr	...	tmax	dmax	tmin	dmin	hmdy	hmax	hmin	v
0	178	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 00:00:00	2007-11-06	2007	...	29.7	16.8	25.5	10.8	35.0	58.0	32.0	
1	178	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 01:00:00	2007-11-06	2007	...	29.9	13.6	29.0	12.2	39.0	39.0	35.0	
2	178	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 02:00:00	2007-11-06	2007	...	29.0	14.0	27.4	13.6	44.0	44.0	39.0	
3	178	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 03:00:00	2007-11-06	2007	...	27.4	16.9	25.8	14.1	58.0	58.0	44.0	
						São		2007-11-	2007-										

	wsid	elvt	lat	lon	inme	city	prov	mdct	date	yr	...	tmax	dmax	tmin	dmin	hmdy	hmax	hmin	v
4	178	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 04:00:00	2007-11-06	2007	...	26.3	17.0	25.3	16.4	57.0	58.0	56.0	

5 rows × 30 columns

Calculated Fields

In [0]:

```
import os
import pandas as pd
import numpy as np
filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
df = pd.read_csv(filename_read,na_values=['NA','?'])
df.insert(1,'elvt_metric',(df['elvt']*0.5967).astype(int))
df
```

Out[0]:

	wsid	elvt_metric	wsnm	elvt	lat	lon	inme	city	prov	mdct	...	tmax	dmax	tmin	dmin	hm
0	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 00:00:00	...	29.7	16.8	25.5	10.8	35
1	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 01:00:00	...	29.9	13.6	29.0	12.2	39
2	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 02:00:00	...	29.0	14.0	27.4	13.6	44
3	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 03:00:00	...	27.4	16.9	25.8	14.1	58
4	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 04:00:00	...	26.3	17.0	25.3	16.4	57
5	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 05:00:00	...	25.4	16.4	23.8	16.0	62
6	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 06:00:00	...	23.8	16.7	22.0	16.2	72
7	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 07:00:00	...	22.0	17.8	19.5	16.6	86
8	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 08:00:00	...	19.7	17.3	18.3	16.9	93
9	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 09:00:00	...	22.9	18.3	18.2	17.1	75
10	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 10:00:00	...	25.1	18.4	22.9	17.0	61
11	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 11:00:00	...	0.0	0.0	0.0	0.0	0
12	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 12:00:00	...	0.0	0.0	0.0	0.0	0
13	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 13:00:00	...	0.0	0.0	0.0	0.0	0
14	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 14:00:00	...	31.8	16.0	30.0	14.3	36
15	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 15:00:00	...	33.0	15.4	31.0	13.6	32
			SÃO					São		2007-11-						

	wsid	elvt_metric	wsnm	elvt	lat	lon	inme	city	prov	mdct	...	tmax	dmax	tmin	dmin	hm
16	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 16:00:00	...	34.0	15.6	32.5	12.0	31
17	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 17:00:00	...	34.7	14.6	33.4	12.2	29
18	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 18:00:00	...	35.2	14.2	33.9	12.6	27
19	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 19:00:00	...	35.1	14.5	33.7	12.6	28
20	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 20:00:00	...	34.7	14.5	32.2	12.8	30
21	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 21:00:00	...	32.7	15.8	29.9	12.5	40
22	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 22:00:00	...	31.7	15.4	29.4	11.3	29
23	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 23:00:00	...	31.5	13.3	29.8	11.4	35
24	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 00:00:00	...	31.0	15.1	30.2	13.4	40
25	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 01:00:00	...	30.3	15.1	29.3	13.1	37
26	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 02:00:00	...	29.3	14.0	28.1	13.1	42
27	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 03:00:00	...	28.1	15.5	26.5	14.0	51
28	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 04:00:00	...	26.6	16.4	25.1	15.5	58
29	178	141	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 05:00:00	...	25.2	16.4	23.7	15.3	59
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9779138	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 18:00:00	...	24.6	13.0	21.7	11.6	53
9779139	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 19:00:00	...	22.1	12.9	20.0	11.7	61
9779140	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 20:00:00	...	20.2	12.6	16.8	11.7	72
9779141	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 21:00:00	...	16.9	12.1	15.3	11.5	79
9779142	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 22:00:00	...	15.3	12.4	14.2	11.6	84
9779143	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 23:00:00	...	14.7	11.8	14.2	10.2	75
9779144	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 00:00:00	...	14.9	11.4	14.5	10.3	81
9779145	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 01:00:00	...	14.8	11.4	14.2	9.9	75
9779146	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 02:00:00	...	14.7	10.2	14.3	9.2	73
9779147	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 03:00:00	...	14.8	11.4	14.3	9.6	80

	wsid	elvt_metric	wsnm	elvt	lat	lon	inme	city	prov	mdct	...	tmax	dmax	tmin	dmin	hm
9779148	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 04:00:00	...	14.9	12.3	14.7	11.3	84
9779149	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 05:00:00	...	14.9	12.2	14.8	10.9	77
9779150	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 06:00:00	...	14.9	11.6	14.6	10.9	81
9779151	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 07:00:00	...	14.8	11.8	14.5	11.4	81
9779152	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 08:00:00	...	14.9	11.8	14.6	11.3	80
9779153	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 09:00:00	...	14.9	11.7	14.2	11.3	82
9779154	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 10:00:00	...	15.8	11.4	14.3	11.1	74
9779155	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 11:00:00	...	17.7	12.0	15.6	11.0	69
9779156	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 12:00:00	...	19.3	12.0	17.1	10.6	60
9779157	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 13:00:00	...	20.5	12.2	18.2	10.6	58
9779158	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 14:00:00	...	21.4	12.5	19.4	9.8	55
9779159	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 15:00:00	...	21.8	12.1	19.9	10.6	54
9779160	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 16:00:00	...	21.4	12.8	20.2	11.5	59
9779161	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 17:00:00	...	21.2	12.8	19.3	11.5	64
9779162	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 18:00:00	...	19.5	12.8	18.0	11.8	67
9779163	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 19:00:00	...	18.2	12.4	16.3	11.8	76
9779164	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 20:00:00	...	16.8	12.5	15.3	11.7	80
9779165	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 21:00:00	...	15.3	11.9	14.9	11.5	79
9779166	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 22:00:00	...	15.0	11.7	14.4	11.4	82
9779167	423	463	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 23:00:00	...	14.6	11.5	14.3	11.2	82

9779168 rows × 32 columns

Feature Normalization

In [0]:

```
import os
import pandas as pd
import numpy as np
from scipy.stats import zscore

filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
df = pd.read_csv(filename_read,na_values=['NaN','?'])
df.dropna()
df['lat'] = zscore(df['lat'])
df
```

Out[0]:

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	hm
0	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 00:00:00	2007-11-06	...	29.7	16.8	25.5	10.8	35.0	5
1	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 01:00:00	2007-11-06	...	29.9	13.6	29.0	12.2	39.0	3
2	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 02:00:00	2007-11-06	...	29.0	14.0	27.4	13.6	44.0	4
3	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 03:00:00	2007-11-06	...	27.4	16.9	25.8	14.1	58.0	5
4	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 04:00:00	2007-11-06	...	26.3	17.0	25.3	16.4	57.0	5
5	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 05:00:00	2007-11-06	...	25.4	16.4	23.8	16.0	62.0	6
6	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 06:00:00	2007-11-06	...	23.8	16.7	22.0	16.2	72.0	7
7	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 07:00:00	2007-11-06	...	22.0	17.8	19.5	16.6	86.0	8
8	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 08:00:00	2007-11-06	...	19.7	17.3	18.3	16.9	93.0	9
9	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 09:00:00	2007-11-06	...	22.9	18.3	18.2	17.1	75.0	9
10	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 10:00:00	2007-11-06	...	25.1	18.4	22.9	17.0	61.0	7
11	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 11:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
12	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 12:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
13	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 13:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
14	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 14:00:00	2007-11-06	...	31.8	16.0	30.0	14.3	36.0	4
15	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 15:00:00	2007-11-06	...	33.0	15.4	31.0	13.6	32.0	3
16	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 16:00:00	2007-11-06	...	34.0	15.6	32.5	12.9	31.0	3
17	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 17:00:00	2007-11-06	...	34.7	14.6	33.4	12.2	29.0	3
18	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 18:00:00	2007-11-06	...	35.2	14.2	33.9	12.6	27.0	2
19	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 19:00:00	2007-11-06	...	35.1	14.5	33.7	12.6	28.0	3

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	hm
20	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 20:00:00	2007-11-06	...	34.7	14.5	32.2	12.8	30.0	3
21	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 21:00:00	2007-11-06	...	32.7	15.8	29.9	12.5	40.0	4
22	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 22:00:00	2007-11-06	...	31.7	15.4	29.4	11.3	29.0	4
23	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-06 23:00:00	2007-11-06	...	31.5	13.3	29.8	11.4	35.0	3
24	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-07 00:00:00	2007-11-07	...	31.0	15.1	30.2	13.4	40.0	4
25	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-07 01:00:00	2007-11-07	...	30.3	15.1	29.3	13.1	37.0	4
26	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-07 02:00:00	2007-11-07	...	29.3	14.0	28.1	13.1	42.0	4
27	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-07 03:00:00	2007-11-07	...	28.1	15.5	26.5	14.0	51.0	5
28	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-07 04:00:00	2007-11-07	...	26.6	16.4	25.1	15.5	58.0	5
29	178	SÃO GONÇALO	237.0	4.222044	-38.311583	A333	São Gonçalo	RJ	2007-11-07 05:00:00	2007-11-07	...	25.2	16.4	23.7	15.3	59.0	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9779138	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-29 18:00:00	2016-09-29	...	24.6	13.0	21.7	11.6	53.0	5
9779139	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-29 19:00:00	2016-09-29	...	22.1	12.9	20.0	11.7	61.0	6
9779140	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-29 20:00:00	2016-09-29	...	20.2	12.6	16.8	11.7	72.0	7
9779141	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-29 21:00:00	2016-09-29	...	16.9	12.1	15.3	11.5	79.0	7
9779142	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-29 22:00:00	2016-09-29	...	15.3	12.4	14.2	11.6	84.0	8
9779143	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-29 23:00:00	2016-09-29	...	14.7	11.8	14.2	10.2	75.0	8
9779144	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 00:00:00	2016-09-30	...	14.9	11.4	14.5	10.3	81.0	8
9779145	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 01:00:00	2016-09-30	...	14.8	11.4	14.2	9.9	75.0	8
9779146	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 02:00:00	2016-09-30	...	14.7	10.2	14.3	9.2	73.0	7
9779147	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 03:00:00	2016-09-30	...	14.8	11.4	14.3	9.6	80.0	8
9779148	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 04:00:00	2016-09-30	...	14.9	12.3	14.7	11.3	84.0	8
9779149	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 05:00:00	2016-09-30	...	14.9	12.2	14.8	10.9	77.0	8
9779150	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 06:00:00	2016-09-30	...	14.9	11.6	14.6	10.9	81.0	8
9779151	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 07:00:00	2016-09-30	...	14.8	11.8	14.5	11.4	81.0	8

07:00:00

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	hm
9779152	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 08:00:00	2016-09-30	...	14.9	11.8	14.6	11.3	80.0	8
9779153	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 09:00:00	2016-09-30	...	14.9	11.7	14.2	11.3	82.0	8
9779154	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 10:00:00	2016-09-30	...	15.8	11.4	14.3	11.1	74.0	8
9779155	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 11:00:00	2016-09-30	...	17.7	12.0	15.6	11.0	69.0	7
9779156	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 12:00:00	2016-09-30	...	19.3	12.0	17.1	10.6	60.0	6
9779157	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 13:00:00	2016-09-30	...	20.5	12.2	18.2	10.6	58.0	6
9779158	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 14:00:00	2016-09-30	...	21.4	12.5	19.4	9.8	55.0	6
9779159	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 15:00:00	2016-09-30	...	21.8	12.1	19.9	10.6	54.0	5
9779160	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 16:00:00	2016-09-30	...	21.4	12.8	20.2	11.5	59.0	6
9779161	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 17:00:00	2016-09-30	...	21.2	12.8	19.3	11.5	64.0	6
9779162	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 18:00:00	2016-09-30	...	19.5	12.8	18.0	11.8	67.0	6
9779163	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 19:00:00	2016-09-30	...	18.2	12.4	16.3	11.8	76.0	7
9779164	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 20:00:00	2016-09-30	...	16.8	12.5	15.3	11.7	80.0	8
9779165	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 21:00:00	2016-09-30	...	15.3	11.9	14.9	11.5	79.0	8
9779166	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 22:00:00	2016-09-30	...	15.0	11.7	14.4	11.4	82.0	8
9779167	423	BARUERI	777.0	-1.037959	-46.869450	A755	Barueri	SP	2016-09-30 23:00:00	2016-09-30	...	14.6	11.5	14.3	11.2	82.0	8

9779168 rows × 31 columns

Missing Values

In [0]:

```
import os
import pandas as pd
import numpy as np
from scipy.stats import zscore

filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
df = pd.read_csv(filename_read,na_values=['NaN','?'])
med = df['tmin'].median()
df['tmin'] = df['tmin'].fillna(med)
df
```

Out[0]:

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	h
0	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 00:00:00	2007-11-06	...	29.7	16.8	25.5	10.8	35.0	
1	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 01:00:00	2007-11-06	...	29.9	13.6	29.0	12.2	39.0	
2	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 02:00:00	2007-11-06	...	29.0	14.0	27.4	13.6	44.0	
3	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 03:00:00	2007-11-06	...	27.4	16.9	25.8	14.1	58.0	
4	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 04:00:00	2007-11-06	...	26.3	17.0	25.3	16.4	57.0	
5	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 05:00:00	2007-11-06	...	25.4	16.4	23.8	16.0	62.0	
6	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 06:00:00	2007-11-06	...	23.8	16.7	22.0	16.2	72.0	
7	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 07:00:00	2007-11-06	...	22.0	17.8	19.5	16.6	86.0	
8	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 08:00:00	2007-11-06	...	19.7	17.3	18.3	16.9	93.0	
9	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 09:00:00	2007-11-06	...	22.9	18.3	18.2	17.1	75.0	
10	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 10:00:00	2007-11-06	...	25.1	18.4	22.9	17.0	61.0	
11	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 11:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
12	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 12:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
13	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 13:00:00	2007-11-06	...	0.0	0.0	0.0	0.0	0.0	
14	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 14:00:00	2007-11-06	...	31.8	16.0	30.0	14.3	36.0	
15	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 15:00:00	2007-11-06	...	33.0	15.4	31.0	13.6	32.0	
16	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 16:00:00	2007-11-06	...	34.0	15.6	32.5	12.9	31.0	
17	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 17:00:00	2007-11-06	...	34.7	14.6	33.4	12.2	29.0	
18	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 18:00:00	2007-11-06	...	35.2	14.2	33.9	12.6	27.0	
19	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 19:00:00	2007-11-06	...	35.1	14.5	33.7	12.6	28.0	
20	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 20:00:00	2007-11-06	...	34.7	14.5	32.2	12.8	30.0	
21	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 21:00:00	2007-11-06	...	32.7	15.8	29.9	12.5	40.0	
22	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06 22:00:00	2007-11-06	...	31.7	15.4	29.4	11.3	29.0	
23	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-06	2007-11-06	...	31.5	13.3	29.8	11.4	35.0	

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	h
24	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 00:00:00	2007-11-07	...	31.0	15.1	30.2	13.4	40.0	
25	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 01:00:00	2007-11-07	...	30.3	15.1	29.3	13.1	37.0	
26	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 02:00:00	2007-11-07	...	29.3	14.0	28.1	13.1	42.0	
27	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 03:00:00	2007-11-07	...	28.1	15.5	26.5	14.0	51.0	
28	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 04:00:00	2007-11-07	...	26.6	16.4	25.1	15.5	58.0	
29	178	SÃO GONÇALO	237.0	-6.835777	-38.311583	A333	São Gonçalo	RJ	2007-11-07 05:00:00	2007-11-07	...	25.2	16.4	23.7	15.3	59.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9779138	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 18:00:00	2016-09-29	...	24.6	13.0	21.7	11.6	53.0	
9779139	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 19:00:00	2016-09-29	...	22.1	12.9	20.0	11.7	61.0	
9779140	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 20:00:00	2016-09-29	...	20.2	12.6	16.8	11.7	72.0	
9779141	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 21:00:00	2016-09-29	...	16.9	12.1	15.3	11.5	79.0	
9779142	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 22:00:00	2016-09-29	...	15.3	12.4	14.2	11.6	84.0	
9779143	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-29 23:00:00	2016-09-29	...	14.7	11.8	14.2	10.2	75.0	
9779144	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 00:00:00	2016-09-30	...	14.9	11.4	14.5	10.3	81.0	
9779145	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 01:00:00	2016-09-30	...	14.8	11.4	14.2	9.9	75.0	
9779146	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 02:00:00	2016-09-30	...	14.7	10.2	14.3	9.2	73.0	
9779147	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 03:00:00	2016-09-30	...	14.8	11.4	14.3	9.6	80.0	
9779148	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 04:00:00	2016-09-30	...	14.9	12.3	14.7	11.3	84.0	
9779149	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 05:00:00	2016-09-30	...	14.9	12.2	14.8	10.9	77.0	
9779150	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 06:00:00	2016-09-30	...	14.9	11.6	14.6	10.9	81.0	
9779151	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 07:00:00	2016-09-30	...	14.8	11.8	14.5	11.4	81.0	
9779152	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 08:00:00	2016-09-30	...	14.9	11.8	14.6	11.3	80.0	
9779153	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 09:00:00	2016-09-30	...	14.9	11.7	14.2	11.3	82.0	
9779154	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 10:00:00	2016-09-30	...	15.8	11.4	14.3	11.1	74.0	
				-	-				2016-09-	2016-							

	wsid	wsnm	elvt	lat	lon	inme	city	prov	mdct	date	...	tmax	dmax	tmin	dmin	hmdy	h
9779155	423	BARUERI	777.0	23.523890	46.869450	A755	Barueri	SP	30 11:00:00	09-30	...	17.7	12.0	15.6	11.0	69.0	
9779156	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 12:00:00	2016-09-30	...	19.3	12.0	17.1	10.6	60.0	
9779157	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 13:00:00	2016-09-30	...	20.5	12.2	18.2	10.6	58.0	
9779158	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 14:00:00	2016-09-30	...	21.4	12.5	19.4	9.8	55.0	
9779159	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 15:00:00	2016-09-30	...	21.8	12.1	19.9	10.6	54.0	
9779160	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 16:00:00	2016-09-30	...	21.4	12.8	20.2	11.5	59.0	
9779161	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 17:00:00	2016-09-30	...	21.2	12.8	19.3	11.5	64.0	
9779162	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 18:00:00	2016-09-30	...	19.5	12.8	18.0	11.8	67.0	
9779163	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 19:00:00	2016-09-30	...	18.2	12.4	16.3	11.8	76.0	
9779164	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 20:00:00	2016-09-30	...	16.8	12.5	15.3	11.7	80.0	
9779165	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 21:00:00	2016-09-30	...	15.3	11.9	14.9	11.5	79.0	
9779166	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 22:00:00	2016-09-30	...	15.0	11.7	14.4	11.4	82.0	
9779167	423	BARUERI	777.0	-23.523890	-46.869450	A755	Barueri	SP	2016-09-30 23:00:00	2016-09-30	...	14.6	11.5	14.3	11.2	82.0	

9779168 rows × 31 columns

Concatenating Rows and Columns

In [0]:

```
import os
import pandas as pd
import numpy as np
from scipy.stats import zscore

filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
df = pd.read_csv(filename_read,na_values=['NaN','?'])

col_wsnm = df['wsnm']
col_mdct = df['mdct']
result = pd.concat([col_wsnm,col_mdct],axis=1)
result
```

Out[0]:

	wsnm	mdct
0	SÃO GONÇALO	2007-11-06 00:00:00
1	SÃO GONÇALO	2007-11-06 01:00:00
2	SÃO GONÇALO	2007-11-06 02:00:00
3	SÃO GONÇALO	2007-11-06 03:00:00

	wsnm	mdct
4	SÃO GONÇALO	2007-11-06 04:00:00
5	SÃO GONÇALO	2007-11-06 05:00:00
6	SÃO GONÇALO	2007-11-06 06:00:00
7	SÃO GONÇALO	2007-11-06 07:00:00
8	SÃO GONÇALO	2007-11-06 08:00:00
9	SÃO GONÇALO	2007-11-06 09:00:00
10	SÃO GONÇALO	2007-11-06 10:00:00
11	SÃO GONÇALO	2007-11-06 11:00:00
12	SÃO GONÇALO	2007-11-06 12:00:00
13	SÃO GONÇALO	2007-11-06 13:00:00
14	SÃO GONÇALO	2007-11-06 14:00:00
15	SÃO GONÇALO	2007-11-06 15:00:00
16	SÃO GONÇALO	2007-11-06 16:00:00
17	SÃO GONÇALO	2007-11-06 17:00:00
18	SÃO GONÇALO	2007-11-06 18:00:00
19	SÃO GONÇALO	2007-11-06 19:00:00
20	SÃO GONÇALO	2007-11-06 20:00:00
21	SÃO GONÇALO	2007-11-06 21:00:00
22	SÃO GONÇALO	2007-11-06 22:00:00
23	SÃO GONÇALO	2007-11-06 23:00:00
24	SÃO GONÇALO	2007-11-07 00:00:00
25	SÃO GONÇALO	2007-11-07 01:00:00
26	SÃO GONÇALO	2007-11-07 02:00:00
27	SÃO GONÇALO	2007-11-07 03:00:00
28	SÃO GONÇALO	2007-11-07 04:00:00
29	SÃO GONÇALO	2007-11-07 05:00:00
...	...	...
9779138	BARUERI	2016-09-29 18:00:00
9779139	BARUERI	2016-09-29 19:00:00
9779140	BARUERI	2016-09-29 20:00:00
9779141	BARUERI	2016-09-29 21:00:00
9779142	BARUERI	2016-09-29 22:00:00
9779143	BARUERI	2016-09-29 23:00:00
9779144	BARUERI	2016-09-30 00:00:00
9779145	BARUERI	2016-09-30 01:00:00
9779146	BARUERI	2016-09-30 02:00:00
9779147	BARUERI	2016-09-30 03:00:00

	wsnm	mdct
9779147	BARUERI	2016-09-30 03:00:00
9779148	BARUERI	2016-09-30 04:00:00
9779149	BARUERI	2016-09-30 05:00:00
9779150	BARUERI	2016-09-30 06:00:00
9779151	BARUERI	2016-09-30 07:00:00
9779152	BARUERI	2016-09-30 08:00:00
9779153	BARUERI	2016-09-30 09:00:00
9779154	BARUERI	2016-09-30 10:00:00
9779155	BARUERI	2016-09-30 11:00:00
9779156	BARUERI	2016-09-30 12:00:00
9779157	BARUERI	2016-09-30 13:00:00
9779158	BARUERI	2016-09-30 14:00:00
9779159	BARUERI	2016-09-30 15:00:00
9779160	BARUERI	2016-09-30 16:00:00
9779161	BARUERI	2016-09-30 17:00:00
9779162	BARUERI	2016-09-30 18:00:00
9779163	BARUERI	2016-09-30 19:00:00
9779164	BARUERI	2016-09-30 20:00:00
9779165	BARUERI	2016-09-30 21:00:00
9779166	BARUERI	2016-09-30 22:00:00
9779167	BARUERI	2016-09-30 23:00:00

9779168 rows × 2 columns

In [0]:

```python
import os
import pandas as pd
import numpy as np
from scipy.stats import zscore

filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
df = pd.read_csv(filename_read,na_values=['NaN','?'])

col_wsnm = df['wsnm']
col_mdct = df['mdct']
result = pd.concat([col_wsnm,col_mdct])
result
```

Out[0]:

```
0                   SÃO GONÇALO
1                   SÃO GONÇALO
2                   SÃO GONÇALO
3                   SÃO GONÇALO
4                   SÃO GONÇALO
5                   SÃO GONÇALO
6                   SÃO GONÇALO
7                   SÃO GONÇALO
8                   SÃO GONÇALO
9                   SÃO GONÇALO
10                  SÃO GONÇALO
11                  SÃO GONÇALO
12                  SÃO GONÇALO
13                  SÃO GONÇALO
14                  SÃO GONÇALO
15                  SÃO GONÇALO
16                  SÃO GONÇALO
17                  SÃO GONÇALO
18                  SÃO GONÇALO
19                  SÃO GONÇALO
20                  SÃO GONÇALO
21                  SÃO GONÇALO
22                  SÃO GONÇALO
23                  SÃO GONÇALO
24                  SÃO GONÇALO
```

```
24                 SÃO GONÇALO
25                 SÃO GONÇALO
26                 SÃO GONÇALO
27                 SÃO GONÇALO
28                 SÃO GONÇALO
29                 SÃO GONÇALO
                  ...
9779138    2016-09-29 18:00:00
9779139    2016-09-29 19:00:00
9779140    2016-09-29 20:00:00
9779141    2016-09-29 21:00:00
9779142    2016-09-29 22:00:00
9779143    2016-09-29 23:00:00
9779144    2016-09-30 00:00:00
9779145    2016-09-30 01:00:00
9779146    2016-09-30 02:00:00
9779147    2016-09-30 03:00:00
9779148    2016-09-30 04:00:00
9779149    2016-09-30 05:00:00
9779150    2016-09-30 06:00:00
9779151    2016-09-30 07:00:00
9779152    2016-09-30 08:00:00
9779153    2016-09-30 09:00:00
9779154    2016-09-30 10:00:00
9779155    2016-09-30 11:00:00
9779156    2016-09-30 12:00:00
9779157    2016-09-30 13:00:00
9779158    2016-09-30 14:00:00
9779159    2016-09-30 15:00:00
9779160    2016-09-30 16:00:00
9779161    2016-09-30 17:00:00
9779162    2016-09-30 18:00:00
9779163    2016-09-30 19:00:00
9779164    2016-09-30 20:00:00
9779165    2016-09-30 21:00:00
9779166    2016-09-30 22:00:00
9779167    2016-09-30 23:00:00
Length: 19558336, dtype: object
```

Important methods

In [2]:

```python
import collections
from sklearn import preprocessing
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import shutil
import os


# Encode text values to dummy variables(i.e. [1,0,0],[0,1,0],[0,0,1] for red,green,blue)
def encode_text_dummy(df, name):
    dummies = pd.get_dummies(df[name])
    for x in dummies.columns:
        dummy_name = "{}-{}".format(name, x)
        df[dummy_name] = dummies[x]
    df.drop(name, axis=1, inplace=True)


# Encode text values to indexes(i.e. [1],[2],[3] for red,green,blue).
def encode_text_index(df, name):
    le = preprocessing.LabelEncoder()
    df[name] = le.fit_transform(df[name])
    return le.classes_


# Encode a numeric column as zscores
def encode_numeric_zscore(df, name, mean=None, sd=None):
    if mean is None:
        mean = df[name].mean()

    if sd is None:
        sd = df[name].std()
    df[name] = (df[name] - mean) / sd


# Convert all missing values in the specified column to the median
def missing_median(df, name):
    med = df[name].median()
    df[name] = df[name].fillna(med)


# Convert all missing values in the specified column to the default
def missing_default(df, name, default_value):
    df[name] = df[name].fillna(default_value)


# Convert a Pandas dataframe to the x,y inputs that TensorFlow needs
def to_xy(df, target):

    result = []
    for x in df.columns:
        if x != target:
            result.append(x)
    # find out the type of the target column.
    target_type = df[target].dtypes
    target_type = target_type[0] if isinstance(target_type, collections.Sequence) else target_type
    # Encode to int for classification, float otherwise. TensorFlow likes 32 bits.
    if target_type in (np.int64, np.int32):
        # Classification
        dummies = pd.get_dummies(df[target])
       try:
        return df[result].values.astype(np.float32), dummies.values.astype(np.float32)
       except :
            pass
    else:
        # Regression
       try:
        return df[result].values.astype(np.float32), df[target].values.astype(np.float32)
       except :
            pass
# Nicely formatted time string
def hms_string(sec_elapsed):
    h = int(sec_elapsed / (60 * 60))
    m = int((sec_elapsed % (60 * 60)) / 60)
    s = sec_elapsed % 60
    return "{}:{:>02}:{:>05.2f}".format(h, m, s)


# Regression chart.
def chart_regression(pred,y,sort=True):
    t = pd.DataFrame({'pred' : pred, 'y' : y.flatten()})
    if sort:
        t.sort_values(by=['y'],inplace=True)
    a = plt.plot(t['y'].tolist(),label='expected')
    b = plt.plot(t['pred'].tolist(),label='prediction')
    plt.ylabel('output')
    plt.legend()
    plt.show()

# Remove all rows where the specified column is +/- sd standard deviations
def remove_outliers(df, name, sd):
    drop_rows = df.index[(np.abs(df[name] - df[name].mean()) >= (sd * df[name].std()))]
    df.drop(drop_rows, axis=0, inplace=True)


# Encode a column to a range between normalized_low and normalized_high.
def encode_numeric_range(df, name, normalized_low=-1, normalized_high=1,
                         data_low=None, data_high=None):
    if data_low is None:
        data_low = min(df[name])
        data_high = max(df[name])

    df[name] = ((df[name] - data_low) / (data_high - data_low)) * (normalized_high - normalized_low
) + normalized_low
```

In [4]:

```
import pandas as pd
import os
import numpy as np

filename_read = os.path.join("/Users/surajrawat/sudeste.csv")
df = pd.read_csv(filename_read,na_values=['NaN','?'])
```

In [5]:

```
df = df.drop(['wsid','elvt','wsnm','inme','yr','mo','da','hr','prcp','gbrd','lat','lon','temp','dew
p','tmax','tmin','dmin','hmin','wdsp','gust','mdct','date'],axis=1)
```

Examples of label encoding, one hot encoding, and creating X/Y for TensorFlow

In [6]:

```
encode_text_index(df,"prov")    # label encoding
df
```

Out[6]:

	city	prov	stp	smax	smin	dmax	hmdy	hmax	wdct
0	São Gonçalo	2	982.5	982.5	981.3	16.8	35.0	58.0	101.0
1	São Gonçalo	2	983.2	983.2	982.5	13.6	39.0	39.0	94.0
2	São Gonçalo	2	983.5	983.5	983.2	14.0	44.0	44.0	93.0
3	São Gonçalo	2	983.7	983.7	983.4	16.9	58.0	58.0	96.0
4	São Gonçalo	2	983.7	983.8	983.6	17.0	57.0	58.0	110.0
5	São Gonçalo	2	983.7	983.8	983.6	16.4	62.0	62.0	99.0
6	São Gonçalo	2	983.7	983.7	983.6	16.7	72.0	72.0	93.0
7	São Gonçalo	2	984.6	984.6	983.7	17.8	86.0	89.0	157.0
8	São Gonçalo	2	985.7	985.7	984.6	17.3	93.0	94.0	141.0
9	São Gonçalo	2	986.7	986.7	985.7	18.3	75.0	94.0	248.0
10	São Gonçalo	2	987.2	987.2	986.7	18.4	61.0	76.0	97.0
11	São Gonçalo	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	São Gonçalo	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	São Gonçalo	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	São Gonçalo	2	986.0	987.0	986.0	16.0	36.0	42.0	97.0
15	São Gonçalo	2	984.8	986.1	984.8	15.4	32.0	37.0	103.0
16	São Gonçalo	2	983.4	984.8	983.4	15.6	31.0	34.0	78.0
17	São Gonçalo	2	982.5	983.4	982.5	14.6	29.0	31.0	102.0
18	São Gonçalo	2	981.7	982.5	981.7	14.2	27.0	29.0	94.0
19	São Gonçalo	2	981.3	981.7	981.3	14.5	28.0	30.0	93.0
20	São Gonçalo	2	981.6	981.6	981.3	14.5	30.0	33.0	106.0
21	São Gonçalo	2	982.1	982.1	981.6	15.8	40.0	42.0	102.0
22	São Gonçalo	2	982.7	982.7	982.0	15.4	29.0	42.0	123.0
23	São Gonçalo	2	983.9	983.9	982.7	13.3	35.0	35.0	112.0
24	São Gonçalo	2	984.7	984.7	983.9	15.1	40.0	40.0	109.0
25	São Gonçalo	2	985.3	985.3	984.7	15.1	37.0	41.0	109.0
26	São Gonçalo	2	985.4	985.5	985.3	14.0	42.0	42.0	120.0
27	São Gonçalo	2	985.3	985.4	985.3	15.5	51.0	51.0	114.0
28	São Gonçalo	2	985.8	985.8	985.3	16.4	58.0	58.0	100.0
29	São Gonçalo	2	985.9	985.9	985.6	16.4	59.0	59.0	111.0
...	...	...	...	...	...	...	...	...	...
9779138	Barueri	3	925.9	926.1	925.9	13.0	53.0	55.0	0.0
9779139	Barueri	3	926.3	926.3	925.8	12.9	61.0	62.0	0.0

	city	prov	stp	smax	smin	dmax	hmdy	hmax	wdct
9779140	Barueri	3	927.0	927.0	926.3	12.6	72.0	73.0	0.0
9779141	Barueri	3	927.7	927.7	927.0	12.1	79.0	79.0	0.0
9779142	Barueri	3	928.3	928.3	927.7	12.4	84.0	88.0	0.0
9779143	Barueri	3	928.9	928.9	928.4	11.8	75.0	84.0	0.0
9779144	Barueri	3	929.3	929.3	928.9	11.4	81.0	81.0	0.0
9779145	Barueri	3	929.4	929.4	929.2	11.4	75.0	81.0	0.0
9779146	Barueri	3	929.0	929.4	929.0	10.2	73.0	76.0	0.0
9779147	Barueri	3	928.3	929.1	928.3	11.4	80.0	80.0	0.0
9779148	Barueri	3	928.1	928.3	928.1	12.3	84.0	85.0	0.0
9779149	Barueri	3	927.8	928.1	927.7	12.2	77.0	84.0	0.0
9779150	Barueri	3	927.4	927.8	927.4	11.6	81.0	81.0	0.0
9779151	Barueri	3	927.8	927.8	927.3	11.8	81.0	83.0	0.0
9779152	Barueri	3	927.9	927.9	927.8	11.8	80.0	82.0	0.0
9779153	Barueri	3	928.2	928.2	927.8	11.7	82.0	83.0	0.0
9779154	Barueri	3	928.8	928.8	928.2	11.4	74.0	82.0	0.0
9779155	Barueri	3	929.1	929.1	928.8	12.0	69.0	75.0	0.0
9779156	Barueri	3	929.4	929.6	929.1	12.0	60.0	69.0	0.0
9779157	Barueri	3	929.4	929.6	929.4	12.2	58.0	64.0	0.0
9779158	Barueri	3	928.9	929.4	928.9	12.5	55.0	62.0	0.0
9779159	Barueri	3	928.0	928.9	928.0	12.1	54.0	58.0	0.0
9779160	Barueri	3	927.6	928.0	927.5	12.8	59.0	60.0	0.0
9779161	Barueri	3	927.3	927.6	927.2	12.8	64.0	65.0	0.0
9779162	Barueri	3	927.4	927.5	927.3	12.8	67.0	68.0	0.0
9779163	Barueri	3	927.6	927.7	927.4	12.4	76.0	77.0	0.0
9779164	Barueri	3	928.1	928.2	927.5	12.5	80.0	80.0	0.0
9779165	Barueri	3	928.7	928.7	928.1	11.9	79.0	81.0	0.0
9779166	Barueri	3	929.6	929.6	928.7	11.7	82.0	83.0	0.0
9779167	Barueri	3	930.5	930.5	929.5	11.5	82.0	82.0	0.0

9779168 rows × 9 columns

In [7]:

```
df1 = pd.read_csv(filename_read,na_values=['Na','?'])

df1 = df1.drop(['wsid','elvt','wsnm','city','inme','yr','mo','da','hr','prcp','gbrd','lat','lon','temp','dewp','tmax','tmin','dmin','hmin','wdsp','gust','mdct','date'],axis=1)
```

In [8]:

```
import pandas as pd

pd.set_option('display.max_columns', None)
```

In [9]:

```
print(df1)
```

```
        prov    stp   smax   smin  dmax  hmdy  hmax   wdct
0         RJ  982.5  982.5  981.3  16.8  35.0  58.0  101.0
1         RJ  983.2  983.2  982.5  13.6  39.0  39.0   94.0
2         RJ  983.5  983.5  983.2  14.0  44.0  44.0   93.0
3         RJ  983.7  983.7  983.4  16.9  58.0  58.0   96.0
4         RJ  983.7  983.8  983.6  17.0  57.0  58.0  110.0
5         RJ  983.7  983.8  983.6  16.4  62.0  62.0   99.0
6         RJ  983.7  983.7  983.6  16.7  72.0  72.0   93.0
7         RJ  984.6  984.6  983.7  17.8  86.0  89.0  157.0
8         RJ  985.7  985.7  984.6  17.3  93.0  94.0  141.0
```

```
9          RJ   986.7   986.7   985.7   18.3   75.0   94.0   248.0
10         RJ   987.2   987.2   986.7   18.4   61.0   76.0    97.0
11         RJ     0.0     0.0     0.0    0.0    0.0    0.0     0.0
12         RJ     0.0     0.0     0.0    0.0    0.0    0.0     0.0
13         RJ     0.0     0.0     0.0    0.0    0.0    0.0     0.0
14         RJ   986.0   987.0   986.0   16.0   36.0   42.0    97.0
15         RJ   984.8   986.1   984.8   15.4   32.0   37.0   103.0
16         RJ   983.4   984.8   983.4   15.6   31.0   34.0    78.0
17         RJ   982.5   983.4   982.5   14.6   29.0   31.0   102.0
18         RJ   981.7   982.5   981.7   14.2   27.0   29.0    94.0
19         RJ   981.3   981.7   981.3   14.5   28.0   30.0    93.0
20         RJ   981.6   981.6   981.3   14.5   30.0   33.0   106.0
21         RJ   982.1   982.1   981.6   15.8   40.0   42.0   102.0
22         RJ   982.7   982.7   982.0   15.4   29.0   42.0   123.0
23         RJ   983.9   983.9   982.7   13.3   35.0   35.0   112.0
24         RJ   984.7   984.7   983.9   15.1   40.0   40.0   109.0
25         RJ   985.3   985.3   984.7   15.1   37.0   41.0   109.0
26         RJ   985.4   985.5   985.3   14.0   42.0   42.0   120.0
27         RJ   985.3   985.4   985.3   15.5   51.0   51.0   114.0
28         RJ   985.8   985.8   985.3   16.4   58.0   58.0   100.0
29         RJ   985.9   985.9   985.6   16.4   59.0   59.0   111.0
...       ...     ...     ...     ...    ...    ...    ...     ...
9779138    SP   925.9   926.1   925.9   13.0   53.0   55.0     0.0
9779139    SP   926.3   926.3   925.8   12.9   61.0   62.0     0.0
9779140    SP   927.0   927.0   926.3   12.6   72.0   73.0     0.0
9779141    SP   927.7   927.7   927.0   12.1   79.0   79.0     0.0
9779142    SP   928.3   928.3   927.7   12.4   84.0   88.0     0.0
9779143    SP   928.9   928.9   928.4   11.8   75.0   84.0     0.0
9779144    SP   929.3   929.3   928.9   11.4   81.0   81.0     0.0
9779145    SP   929.4   929.4   929.2   11.4   75.0   81.0     0.0
9779146    SP   929.0   929.4   929.0   10.2   73.0   76.0     0.0
9779147    SP   928.3   929.1   928.3   11.4   80.0   80.0     0.0
9779148    SP   928.1   928.3   928.1   12.3   84.0   85.0     0.0
9779149    SP   927.8   928.1   927.7   12.2   77.0   84.0     0.0
9779150    SP   927.4   927.8   927.4   11.6   81.0   81.0     0.0
9779151    SP   927.8   927.8   927.3   11.8   81.0   83.0     0.0
9779152    SP   927.9   927.9   927.8   11.8   80.0   82.0     0.0
9779153    SP   928.2   928.2   927.8   11.7   82.0   83.0     0.0
9779154    SP   928.8   928.8   928.2   11.4   74.0   82.0     0.0
9779155    SP   929.1   929.1   928.8   12.0   69.0   75.0     0.0
9779156    SP   929.4   929.6   929.1   12.0   60.0   69.0     0.0
9779157    SP   929.4   929.6   929.4   12.2   58.0   64.0     0.0
9779158    SP   928.9   929.4   928.9   12.5   55.0   62.0     0.0
9779159    SP   928.0   928.9   928.0   12.1   54.0   58.0     0.0
9779160    SP   927.6   928.0   927.5   12.8   59.0   60.0     0.0
9779161    SP   927.3   927.6   927.2   12.8   64.0   65.0     0.0
9779162    SP   927.4   927.5   927.3   12.8   67.0   68.0     0.0
9779163    SP   927.6   927.7   927.4   12.4   76.0   77.0     0.0
9779164    SP   928.1   928.2   927.5   12.5   80.0   80.0     0.0
9779165    SP   928.7   928.7   928.1   11.9   79.0   81.0     0.0
9779166    SP   929.6   929.6   928.7   11.7   82.0   83.0     0.0
9779167    SP   930.5   930.5   929.5   11.5   82.0   82.0     0.0

[9779168 rows x 8 columns]
```

In [16]:

```
df1
```

In [11]:

```
encode_text_dummy(df1,"prov")    # One hot encoding
df1
```

Out[11]:

	stp	smax	smin	dmax	hmdy	hmax	wdct	prov-ES	prov-MG	prov-RJ	prov-SP
0	982.5	982.5	981.3	16.8	35.0	58.0	101.0	0	0	1	0
1	983.2	983.2	982.5	13.6	39.0	39.0	94.0	0	0	1	0
2	983.5	983.5	983.2	14.0	44.0	44.0	93.0	0	0	1	0
3	983.7	983.7	983.4	16.9	58.0	58.0	96.0	0	0	1	0

	stp	smax	smin	dmax	hmdy	hmax	wdct	prov-ES	prov-MG	prov-RJ	prov-SP
4	983.7	983.8	983.6	17.0	57.0	58.0	110.0	0	0	1	0
5	983.7	983.8	983.6	16.4	62.0	62.0	99.0	0	0	1	0
6	983.7	983.7	983.6	16.7	72.0	72.0	93.0	0	0	1	0
7	984.6	984.6	983.7	17.8	86.0	89.0	157.0	0	0	1	0
8	985.7	985.7	984.6	17.3	93.0	94.0	141.0	0	0	1	0
9	986.7	986.7	985.7	18.3	75.0	94.0	248.0	0	0	1	0
10	987.2	987.2	986.7	18.4	61.0	76.0	97.0	0	0	1	0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	1	0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	1	0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	1	0
14	986.0	987.0	986.0	16.0	36.0	42.0	97.0	0	0	1	0
15	984.8	986.1	984.8	15.4	32.0	37.0	103.0	0	0	1	0
16	983.4	984.8	983.4	15.6	31.0	34.0	78.0	0	0	1	0
17	982.5	983.4	982.5	14.6	29.0	31.0	102.0	0	0	1	0
18	981.7	982.5	981.7	14.2	27.0	29.0	94.0	0	0	1	0
19	981.3	981.7	981.3	14.5	28.0	30.0	93.0	0	0	1	0
20	981.6	981.6	981.3	14.5	30.0	33.0	106.0	0	0	1	0
21	982.1	982.1	981.6	15.8	40.0	42.0	102.0	0	0	1	0
22	982.7	982.7	982.0	15.4	29.0	42.0	123.0	0	0	1	0
23	983.9	983.9	982.7	13.3	35.0	35.0	112.0	0	0	1	0
24	984.7	984.7	983.9	15.1	40.0	40.0	109.0	0	0	1	0
25	985.3	985.3	984.7	15.1	37.0	41.0	109.0	0	0	1	0
26	985.4	985.5	985.3	14.0	42.0	42.0	120.0	0	0	1	0
27	985.3	985.4	985.3	15.5	51.0	51.0	114.0	0	0	1	0
28	985.8	985.8	985.3	16.4	58.0	58.0	100.0	0	0	1	0
29	985.9	985.9	985.6	16.4	59.0	59.0	111.0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
9779138	925.9	926.1	925.9	13.0	53.0	55.0	0.0	0	0	0	1
9779139	926.3	926.3	925.8	12.9	61.0	62.0	0.0	0	0	0	1
9779140	927.0	927.0	926.3	12.6	72.0	73.0	0.0	0	0	0	1
9779141	927.7	927.7	927.0	12.1	79.0	79.0	0.0	0	0	0	1
9779142	928.3	928.3	927.7	12.4	84.0	88.0	0.0	0	0	0	1
9779143	928.9	928.9	928.4	11.8	75.0	84.0	0.0	0	0	0	1
9779144	929.3	929.3	928.9	11.4	81.0	81.0	0.0	0	0	0	1
9779145	929.4	929.4	929.2	11.4	75.0	81.0	0.0	0	0	0	1
9779146	929.0	929.4	929.0	10.2	73.0	76.0	0.0	0	0	0	1
9779147	928.3	929.1	928.3	11.4	80.0	80.0	0.0	0	0	0	1
9779148	928.1	928.3	928.1	12.3	84.0	85.0	0.0	0	0	0	1
9779149	927.8	928.1	927.7	12.2	77.0	84.0	0.0	0	0	0	1
9779150	927.4	927.8	927.4	11.6	81.0	81.0	0.0	0	0	0	1
9779151	927.8	927.8	927.3	11.8	81.0	83.0	0.0	0	0	0	1
9779152	927.9	927.9	927.8	11.8	80.0	82.0	0.0	0	0	0	1
9779153	928.2	928.2	927.8	11.7	82.0	83.0	0.0	0	0	0	1
9779154	928.8	928.8	928.2	11.4	74.0	82.0	0.0	0	0	0	1
9779155	929.1	929.1	928.8	12.0	69.0	75.0	0.0	0	0	0	1
9779156	929.4	929.6	929.1	12.0	60.0	69.0	0.0	0	0	0	1
9779157	929.4	929.6	929.4	12.2	58.0	64.0	0.0	0	0	0	1
9779158	928.9	929.4	928.9	12.5	55.0	62.0	0.0	0	0	0	1
9779159	928.0	928.9	928.0	12.1	54.0	58.0	0.0	0	0	0	1
9779160	927.6	928.0	927.5	12.8	59.0	60.0	0.0	0	0	0	1
9779161	927.3	927.6	927.2	12.8	64.0	65.0	0.0	0	0	0	1

	stp	smax	smin	dmax	hmdy	hmax	wdct	prov-ES	prov-MG	prov-RJ	prov-SP
9779162	927.4	927.5	927.3	12.8	67.0	68.0	0.0	0	0	0	1
9779163	927.6	927.7	927.4	12.4	76.0	77.0	0.0	0	0	0	1
9779164	928.1	928.2	927.5	12.5	80.0	80.0	0.0	0	0	0	1
9779165	928.7	928.7	928.1	11.9	79.0	81.0	0.0	0	0	0	1
9779166	929.6	929.6	928.7	11.7	82.0	83.0	0.0	0	0	0	1
9779167	930.5	930.5	929.5	11.5	82.0	82.0	0.0	0	0	0	1

9779168 rows × 11 columns

In [26]:

```
df1 = pd.read_csv(filename_read,na_values=['Na','?'])

df1 = df1.drop(['wsid','elvt','wsnm','city','inme','yr','mo','da','hr','prcp','gbrd','lat','lon','temp','dewp','tmax','tmin','dmin','hmin','wdsp','gust','mdct','date'],axis=1)
encode_text_index(df1,"prov")

x, y = to_xy(df1, "prov")
```

In [27]:

```
x
```

Out[27]:

```
array([[982.5, 982.5, 981.3, ...,  35. ,  58. , 101. ],
       [983.2, 983.2, 982.5, ...,  39. ,  39. ,  94. ],
       [983.5, 983.5, 983.2, ...,  44. ,  44. ,  93. ],
       ...,
       [928.7, 928.7, 928.1, ...,  79. ,  81. ,   0. ],
       [929.6, 929.6, 928.7, ...,  82. ,  83. ,   0. ],
       [930.5, 930.5, 929.5, ...,  82. ,  82. ,   0. ]], dtype=float32)
```

In [28]:

```
y
```

Out[28]:

```
array([[0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       ...,
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.]], dtype=float32)
```

In [29]:

```
df1 = pd.read_csv(filename_read,na_values=['Na','?'])

df1 = df1.drop(['wsid','elvt','wsnm','city','inme','yr','mo','da','hr','prcp','gbrd','lat','lon','temp','dewp','tmax','tmin','dmin','hmin','wdsp','gust','mdct','date'],axis=1)

missing_median(df1,'stp')


# Drop outliers in stp
print("Length before stp outliers dropped: {}".format(len(df1)))
remove_outliers(df1,'stp',2)
print("Length after stp outliers dropped: {}".format(len(df1)))
```

```
Length before stp outliers dropped: 9779168
Length after stp outliers dropped: 9077445
```

Train and Test Split

In [34]:

```
import pandas as pd
import io
import numpy as np
import os
from sklearn.model_selection import train_test_split
from sklearn import preprocessing

df1 = pd.read_csv(filename_read,na_values=['Na','?'])

df1 = df1.drop(['wsid','elvt','wsnm','city','inme','yr','mo','da','hr','prcp','gbrd','lat','lon','temp','dewp','tmax','tmin','dmin','hmin','wdsp','gust','mdct','date'],axis=1)

le = preprocessing.LabelEncoder()
df1['encoded_prov'] = le.fit_transform(df1['prov'])
df1[0:5]

x_train,x_test,y_train,y_test=train_test_split(df1[['smax', 'smin', 'dmax', 'wdct']], df1['encoded_prov'], test_size=0.25, random_state=42)
```

In [36]:

```
x_train.shape
```

Out[36]:

```
(7334376, 4)
```

In [37]:

```
y_train.shape
```

Out[37]:

```
(7334376,)
```

In [38]:

```
x_test.shape
```

Out[38]:

```
(2444792, 4)
```

In [39]:

```
y_test.shape
```

Out[39]:

```
(2444792,)
```

In [41]:

```
print('Training Mean=',x_train['smax'].mean(),' ','Testing Mean',x_test['smax'].mean())
```

```
Training Mean= 880.3301358561372   Testing Mean 880.2231675332878
```

In [42]:

```
print('Training Std=',x_train['smax'].std(skipna = True),' ','Testing Std:',x_test['smax'].std(skipna = True))
```

```
Training Std= 248.88534458654908   Testing Std: 249.0125663191727
```

Train and Test split for Sequential Data

In [6]:

```
import pandas as pd
import io
import numpy as np
import os
from sklearn.model_selection import train_test_split
from sklearn import preprocessing

filename_read = os.path.join("sudeste.csv")
df1 = pd.read_csv(filename_read,na_values=['Na','?'])

df1 = df1.drop(['wsid','elvt','wsnm','city','inme','yr','mo','da','hr','prcp','gbrd','lat','lon','t
emp','dewp','tmax','tmin','dmin','hmin','wdsp','gust','mdct','date'],axis=1)

le = preprocessing.LabelEncoder()
df1['encoded_prov'] = le.fit_transform(df1['prov'])
df1[0:5]
```

Out[6]:

	prov	stp	smax	smin	dmax	hmdy	hmax	wdct	encoded_prov
0	RJ	982.5	982.5	981.3	16.8	35.0	58.0	101.0	2
1	RJ	983.2	983.2	982.5	13.6	39.0	39.0	94.0	2
2	RJ	983.5	983.5	983.2	14.0	44.0	44.0	93.0	2
3	RJ	983.7	983.7	983.4	16.9	58.0	58.0	96.0	2
4	RJ	983.7	983.8	983.6	17.0	57.0	58.0	110.0	2

splitting in sequence with some percentage part

In [7]:

```
df1_y = df1['encoded_prov']

percent70 = int(len(df1)* 0.70)
percent30 = len(df1) - percent70

x_train = df1[0:percent70]
x_test = df1[percent70:len(df1)]
y_train = df1_y[0:percent70].values
y_test = df1_y[percent70:len(df1_y)].values

print("Shape of X_Train: ", x_train.shape)
print("Shape of Y_Train: ", y_train.shape)
print("Shape of X_Test: ", x_test.shape)
print("Shape of Y_Test: ", y_test.shape)
```

```
Shape of X_Train:  (6845417, 9)
Shape of Y_Train:  (6845417,)
Shape of X_Test:  (2933751, 9)
Shape of Y_Test:  (2933751,)
```

In [8]:

```
print('Training Mean=',x_train['smax'].mean(),'  ','Testing Mean',x_test['smax'].mean())
print('Training Std=',x_train['smax'].std(skipna = True),'  ','Testing Std:',x_test['smax'].std(skip
na = True))
```

```
Training Mean= 901.4554512018825   Testing Mean 830.9486075335269
Training Std= 212.6536936512422   Testing Std: 312.30690460119354
```

Aggregration

In [35]:

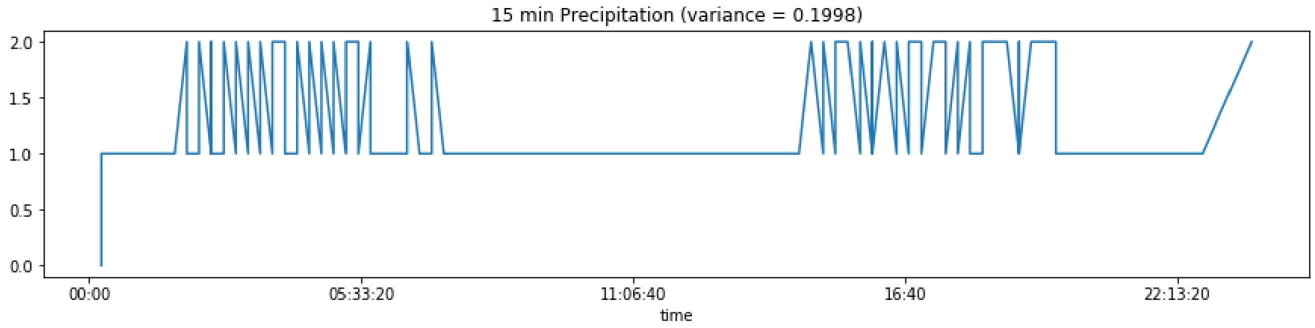
```
import pandas as pd
import io
import numpy as np
from datetime import datetime
min_15=pd.read_csv('https://www1.ncdc.noaa.gov/pub/data/cdo/samples/PRECIP_15_sample_csv.csv',head
er='infer')
nmin_15=pd.to_datetime(min_15['DATE'])
create_nmin=[]
for x in nmin_15:
    x=datetime.time(x)
    create_nmin.append(x)


min_15.index=create_nmin
min_15 = min_15['QPCP']

ax = min_15.plot(kind='line',figsize=(15,3))
ax.set_title('15 min Precipitation (variance = %.4f)' % (min_15.var()))
```

Out[35]:

```
Text(0.5,1,'15 min Precipitation (variance = 0.1998)')
```



In [18]:

```
min_15
```

Out[18]:

```
00:15:00    0
22:45:00    1
00:30:00    1
01:30:00    1
02:15:00    1
03:00:00    1
03:45:00    1
08:00:00    1
08:15:00    1
08:30:00    1
08:45:00    1
09:00:00    1
09:15:00    1
09:30:00    1
09:45:00    1
10:00:00    1
10:15:00    1
10:30:00    1
10:45:00    1
11:00:00    1
11:15:00    1
11:30:00    1
11:45:00    1
12:00:00    1
12:15:00    1
12:30:00    1
12:45:00    1
13:00:00    1
13:15:00    1
13:30:00    1
```

```
            ..
06:45:00     1
07:00:00     1
07:15:00     1
09:00:00     1
10:00:00     1
11:00:00     1
12:00:00     1
13:00:00     1
14:00:00     1
15:00:00     1
16:00:00     1
17:00:00     1
18:00:00     1
19:00:00     1
20:00:00     1
21:00:00     1
22:30:00     1
02:15:00     1
06:00:00     1
08:30:00     1
10:15:00     1
11:30:00     1
12:45:00     1
14:00:00     1
15:15:00     1
16:30:00     1
17:45:00     1
19:00:00     1
22:00:00     1
02:30:00     1
Name: QPCP, Length: 158, dtype: int64
```

In [33]:

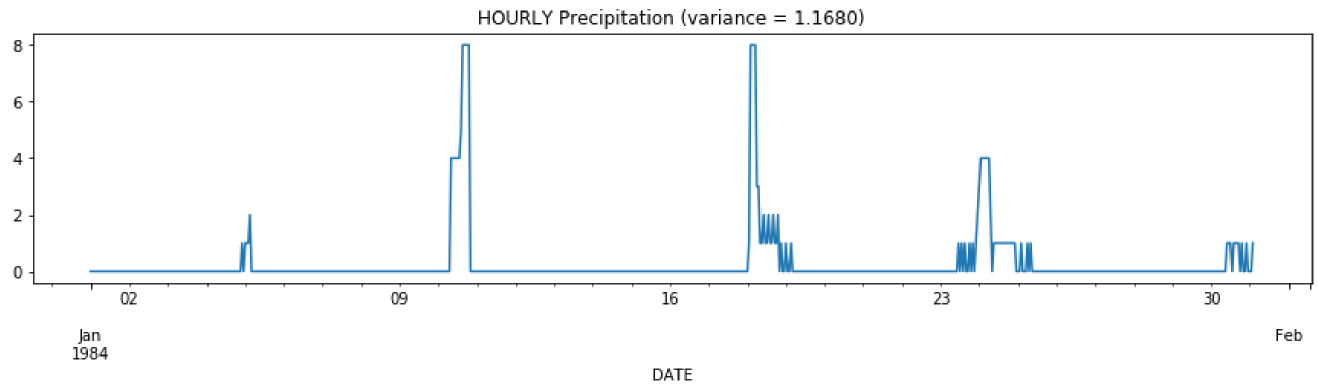
```
import pandas as pd
import io
import numpy as np
from datetime import datetime
min_15=pd.read_csv('https://www1.ncdc.noaa.gov/pub/data/cdo/samples/PRECIP_15_sample_csv.csv',head
er='infer')
nmin_15=pd.to_datetime(min_15['DATE'])


min_15.index=nmin_15
min_15 = min_15['QPCP']

HOURLY = min_15.groupby(pd.Grouper(freq='H')).sum()
ax = HOURLY.plot(kind='line',figsize=(15,3))
ax.set_title('HOURLY Precipitation (variance = %.4f)' % (HOURLY.var()))
```

Out[33]:

```
Text(0.5,1,'HOURLY Precipitation (variance = 1.1680)')
```



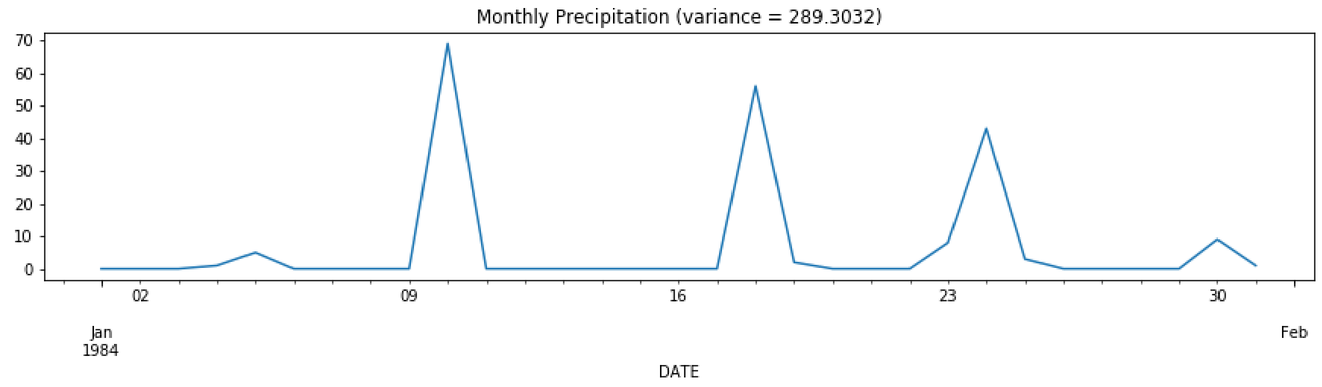
In [34]:

```
DAILY = HOURLY.groupby(pd.Grouper(freq='D')).sum()
ax = DAILY.plot(kind='line',figsize=(15,3))
```

```
ax.set_title('Monthly Precipitation (variance = %.4f)' % (DAILY.var()))
```

Out[34]:

```
Text(0.5,1,'Monthly Precipitation (variance = 289.3032)')
```



In [1]:

```
import pandas as pd
import io
import numpy as np
from datetime import datetime
data=pd.read_csv('https://www1.ncdc.noaa.gov/pub/data/cdo/samples/PRECIP_15_sample_csv.csv',header
='infer')
data.head()
```

Out[1]:

	STATION	STATION_NAME	ELEVATION	LATITUDE	LONGITUDE	DATE	QPCP	Measurement Flag	Quality Flag	Units
0	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840101 00:15	0	g		HI
1	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840104 22:45	1			HI
2	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840105 00:30	1			HI
3	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840105 01:30	1			HI
4	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840105 02:15	1			HI

Sampling

In [2]:

```
sample = data.sample(n=3)
sample
```

Out[2]:

	STATION	STATION_NAME	ELEVATION	LATITUDE	LONGITUDE	DATE	QPCP	Measurement Flag	Quality Flag	Units
103	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840124 00:30	1			HI
27	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840110 13:00	1			HI
55	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840118 01:30	1			HI

In [3]:

```
sample = data.sample(frac=0.03, random_state=1)
sample
```

Out[3]:

	STATION	STATION_NAME	ELEVATION	LATITUDE	LONGITUDE	DATE	QPCP	Measurement Flag	Quality Flag	Units
29	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840110 13:30	1			HI
107	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840124 01:30	1			HI
14	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840110 09:45	1			HI
81	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840118 12:15	1			HI
124	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840124 05:45	1			HI

In [4]:

```
sample = data.sample(frac=0.03, replace=True, random_state=1)
sample
```

Out[4]:

	STATION	STATION_NAME	ELEVATION	LATITUDE	LONGITUDE	DATE	QPCP	Measurement Flag	Quality Flag	Units
37	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840110 15:30	2			HI
140	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840124 18:00	1			HI
72	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840118 06:00	1			HI
137	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840124 15:00	1			HI
133	COOP:311564	CATALOOCHEE NC US	798.9	35.61667	-83.1	19840124 11:00	1			HI

In [19]:

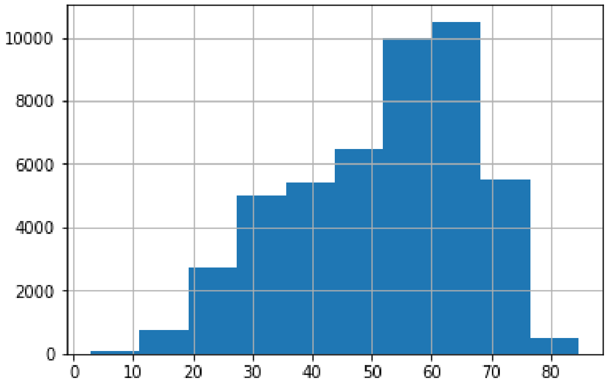
```
import os
import pandas as pd
import numpy as np
from scipy.stats import zscore

filename_read = os.path.join("/Users/surajrawat/temperature.csv")
df = pd.read_csv(filename_read,na_values=['NaN','?'])
df=df.replace('NaN',np.NaN)
df=df.dropna()
df['AverageTemperatureFahr'].hist(bins=10)
df['AverageTemperatureFahr'].value_counts(sort=False)
```

Out[19]:

```
14.0000     1
38.7734     1
56.7176     2
63.8042     2
23.4824     1
64.4450     3
70.6406     1
66.6626     5
39.9524     1
52.7774     2
66.5654     4
68.8802     2
72.2390     1
70.5434     1
47.3918     1
```

```
33.3374    2
74.5502    1
40.2170    2
72.8600    2
36.3200    3
70.1834    1
72.4010    2
39.4844    2
77.8208    1
65.3576    3
44.5712    1
66.2396    3
68.0342    2
58.7714    3
27.6386    1
          ..
37.3280    3
65.9318    1
65.5682    2
75.3440    1
62.1554    5
68.5166    4
42.3446    3
50.9828    2
56.2460    2
25.7792    1
73.8194    4
60.8234    1
72.6296    2
16.2176    1
74.7572    1
34.6640    1
74.7428    1
62.7890    1
59.4914    2
71.6540    2
22.3358    1
66.5438    2
33.3320    1
37.4180    1
79.4966    1
78.4742    1
22.3844    2
75.2612    1
66.0344    3
40.5086    1
Name: AverageTemperatureFahr, Length: 24016, dtype: int64
```



In [22]:

```
bins = pd.cut(df['AverageTemperatureFahr'],4)
bins.value_counts(sort=False)
```

Out[22]:

```
(2.774, 23.306]       1829
(23.306, 43.756]     12202
(43.756, 64.206]     21923
(64.206, 84.655]     11114
Name: AverageTemperatureFahr, dtype: int64
```

In [23]:

```
bins = pd.qcut(df['AverageTemperatureFahr'],4,duplicates='drop')
bins.value_counts(sort=False)
```

Out[23]:

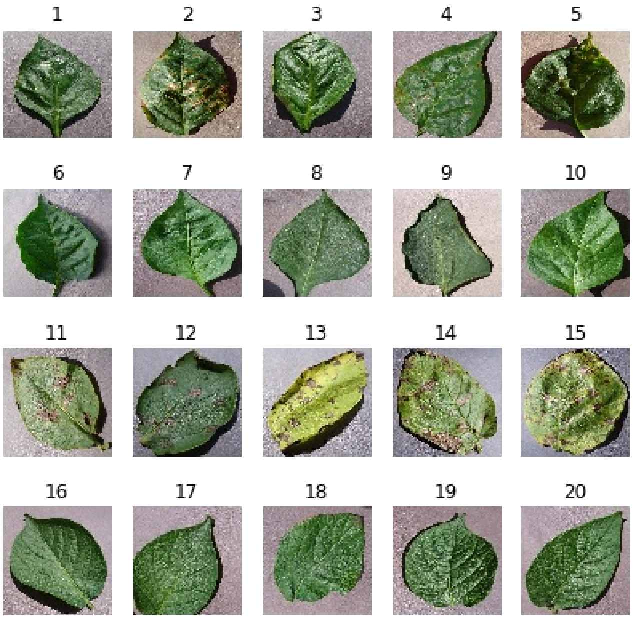
```
(2.855, 40.15]      11767
(40.15, 54.576]     11770
(54.576, 63.709]    11766
(63.709, 84.655]    11765
Name: AverageTemperatureFahr, dtype: int64
```

In [51]:

```
%matplotlib inline
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import numpy as np

numImages = 20
fig = plt.figure(figsize=(7,7))
imgData = np.zeros(shape=(numImages,196608))

for i in range(1,numImages+1):
    filename = '/Users/surajrawat/data_img/picture'+str(i)+'.JPG'
    img = mpimg.imread(filename)
    ax = fig.add_subplot(4,5,i)
    plt.imshow(img)
    plt.axis('off')
    ax.set_title(str(i))
    imgData[i-1] = np.array(img.flatten()).reshape(1,img.shape[0]*img.shape[1]*img.shape[2])
```



In [52]:

```
import pandas as pd
from sklearn.decomposition import PCA

numComponents = 2
pca = PCA(n_components=numComponents)
pca.fit(imgData)

projected = pca.transform(imgData)
projected = pd.DataFrame(projected,columns=['pc1','pc2'],index=range(1,numImages+1))
projected['leaf_disease'] = ['Pepper__bell___Bacterial_spot',
'Pepper__bell___Bacterial_spot','Pepper__bell___Bacterial_spot','Pepper__bell___Bacterial_spot','P
epper__bell___Bacterial_spot','Pepper__bell___healthy','Pepper__bell___healthy','Pepper__bell___hea
lthy',
                             'Pepper__bell___healthy', 'Pepper__bell___healthy','Potato___Early_blight','P
otato___Early_blight','Potato___Early_blight','Potato___Early_blight','Potato___Early_blight','Pot
ato___healthy','Potato___healthy','Potato___healthy','Potato___healthy','Potato___healthy']
projected
```

Out[52]:

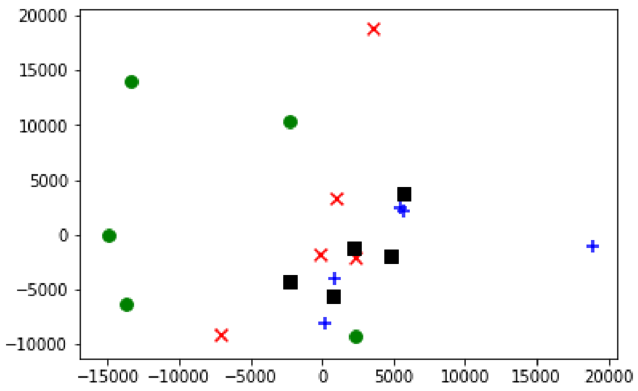
	pc1	pc2	leaf_disease
1	5686.157924	2090.425772	Pepper__bell___Bacterial_spot
2	219.219086	-8048.826779	Pepper__bell___Bacterial_spot
3	5509.697309	2409.461726	Pepper__bell___Bacterial_spot
4	820.726001	-3981.007202	Pepper__bell___Bacterial_spot
5	18879.174893	-1065.866634	Pepper__bell___Bacterial_spot
6	2299.099147	-2105.761993	Pepper__bell___healthy
7	-121.947538	-1733.916440	Pepper__bell___healthy
8	963.082746	3353.640174	Pepper__bell___healthy
9	3547.566496	18867.437268	Pepper__bell___healthy
10	-7036.142542	-9075.107330	Pepper__bell___healthy
11	-14924.839517	-65.260859	Potato___Early_blight
12	2360.939335	-9265.246364	Potato___Early_blight
13	-13349.522766	13949.922245	Potato___Early_blight
14	-2318.019444	10362.825558	Potato___Early_blight
15	-13680.789570	-6396.272638	Potato___Early_blight
16	4758.185499	-1952.454883	Potato___healthy
17	5656.251445	3703.969159	Potato___healthy
18	-2294.418772	-4214.078240	Potato___healthy
19	750.976687	-5553.279236	Potato___healthy
20	2274.603583	-1280.603304	Potato___healthy

In [53]:

```
import matplotlib.pyplot as plt

colors = {'Pepper__bell___Bacterial_spot':'b', 'Pepper__bell___healthy':'r',
'Potato___Early_blight':'g', 'Potato___healthy':'k'}
markerTypes = {'Pepper__bell___Bacterial_spot':'+', 'Pepper__bell___healthy':'x', 'Potato___Early_b
light':'o', 'Potato___healthy':'s'}

for diseaseType in markerTypes:
    d = projected[projected['leaf_disease']==diseaseType]
    plt.scatter(d['pc1'],d['pc2'],c=colors[diseaseType],s=60,marker=markerTypes[diseaseType])
```



In [ ]: