PANDAS FOUNDATIONS

Reading and cleaning the data

Case study

- Comparing observed weather data from two sources

	Temperature	DewPoint	Pressure	Date
0	46.2	37.5	1.0	20100101 00:00
1	44.6	37.1	1.0	20100101 01:00
2	44.1	36.9	1.0	20100101 02:00
3	43.8	36.9	1.0	20100101 03:00
4	43.5	36.8	1.0	20100101 04:00

	Date	Wban	...	station_pressure	sea_level_pressure
0	2011-01-01 00:53:00	13904	...	29.42	29.95
1	2011-01-01 01:53:00	13904	...	29.49	30.01
2	2011-01-01 02:53:00	13904	...	29.49	30.01
3	2011-01-01 03:53:00	13904	...	29.51	30.03
4	2011-01-01 04:53:00	13904	...	29.51	30.04

Source: National Oceanic & Atmospheric Administration, www.noaa.gov/climate

Climate normals of Austin, TX from 1981-2010

	Temperature	DewPoint	Pressure	Date
0	46.2	37.5	1.0	20100101 00:00
1	44.6	37.1	1.0	20100101 01:00
2	44.1	36.9	1.0	20100101 02:00
3	43.8	36.9	1.0	20100101 03:00
4	43.5	36.8	1.0	20100101 04:00
5	43.0	36.5	1.0	20100101 05:00
6	43.1	36.3	1.0	20100101 06:00
7	42.3	35.9	1.0	20100101 07:00
8	42.5	36.2	1.0	20100101 08:00
9	45.9	37.8	1.0	20100101 09:00

Source: National Oceanic & Atmospheric Administration, www.noaa.gov/climate

Weather data of Austin, TX from 2011

	Date	Wban	date	Time	StationType	...	relative_humidity	wind_speed	wind_direction	station_pressure	sea_level_pressure
0	2011-01-01 00:53:00	13904	20110101	5300	12	...	24.0	15.0	360	29.42	29.95
1	2011-01-01 01:53:00	13904	20110101	15300	12	...	23.0	10.0	340	29.49	30.01
2	2011-01-01 02:53:00	13904	20110101	25300	12	...	22.0	15.0	010	29.49	30.01
3	2011-01-01 03:53:00	13904	20110101	35300	12	...	27.0	7.0	350	29.51	30.03
4	2011-01-01 04:53:00	13904	20110101	45300	12	...	25.0	11.0	020	29.51	30.04
5	2011-01-01 05:53:00	13904	20110101	55300	12	...	28.0	6.0	010	29.53	30.06
6	2011-01-01 06:53:00	13904	20110101	65300	12	...	29.0	7.0	360	29.57	30.10
7	2011-01-01 07:53:00	13904	20110101	75300	12	...	29.0	11.0	020	29.59	30.12
8	2011-01-01 08:53:00	13904	20110101	85300	12	...	25.0	15.0	020	29.62	30.16
9	2011-01-01 09:53:00	13904	20110101	95300	12	...	22.0	18.0	010	29.65	30.19

Source: National Oceanic & Atmospheric Administration, www.noaa.gov/climate

Reminder: read_csv()

- Useful keyword options
  - names: assigning column labels
  - index_col: assigning index
  - parse_dates: parsing datetimes
  - na_values: parsing NaNs

PANDAS FOUNDATIONS

Let’s practice!

PANDAS FOUNDATIONS

Statistical exploratory data analysis

Reminder: time series

- Index selection by date time
- Partial datetime selection
- Slicing ranges of datetimes

```
In [1]: climate2010['2010-05-31 22:00:00'] # datetime

In [2]: climate2010['2010-06-01'] # Entire day

In [3]: climate2010['2010-04'] # Entire month

In [4]: climate2010[‘2010-09':'2010-10'] # 2 months
```

Reminder: statistics methods

- Methods for computing statistics:
  - describe(): summary
  - mean(): average
  - count(): counting entries
  - median(): median
  - std(): standard deviation

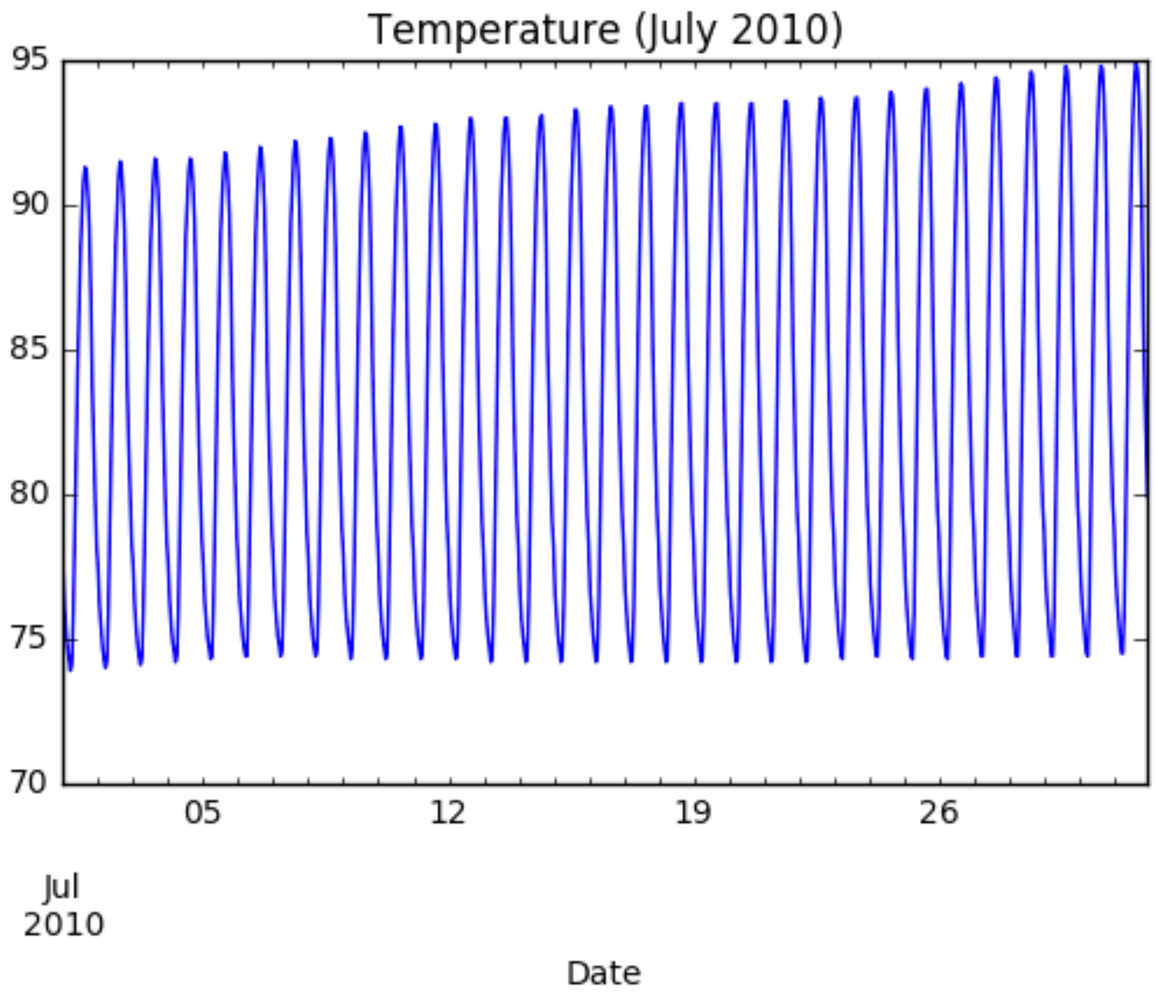
PANDAS FOUNDATIONS

Let’s practice!

PANDAS FOUNDATIONS

Visual exploratory data analysis

Line plots in pandas



Line plots in pandas

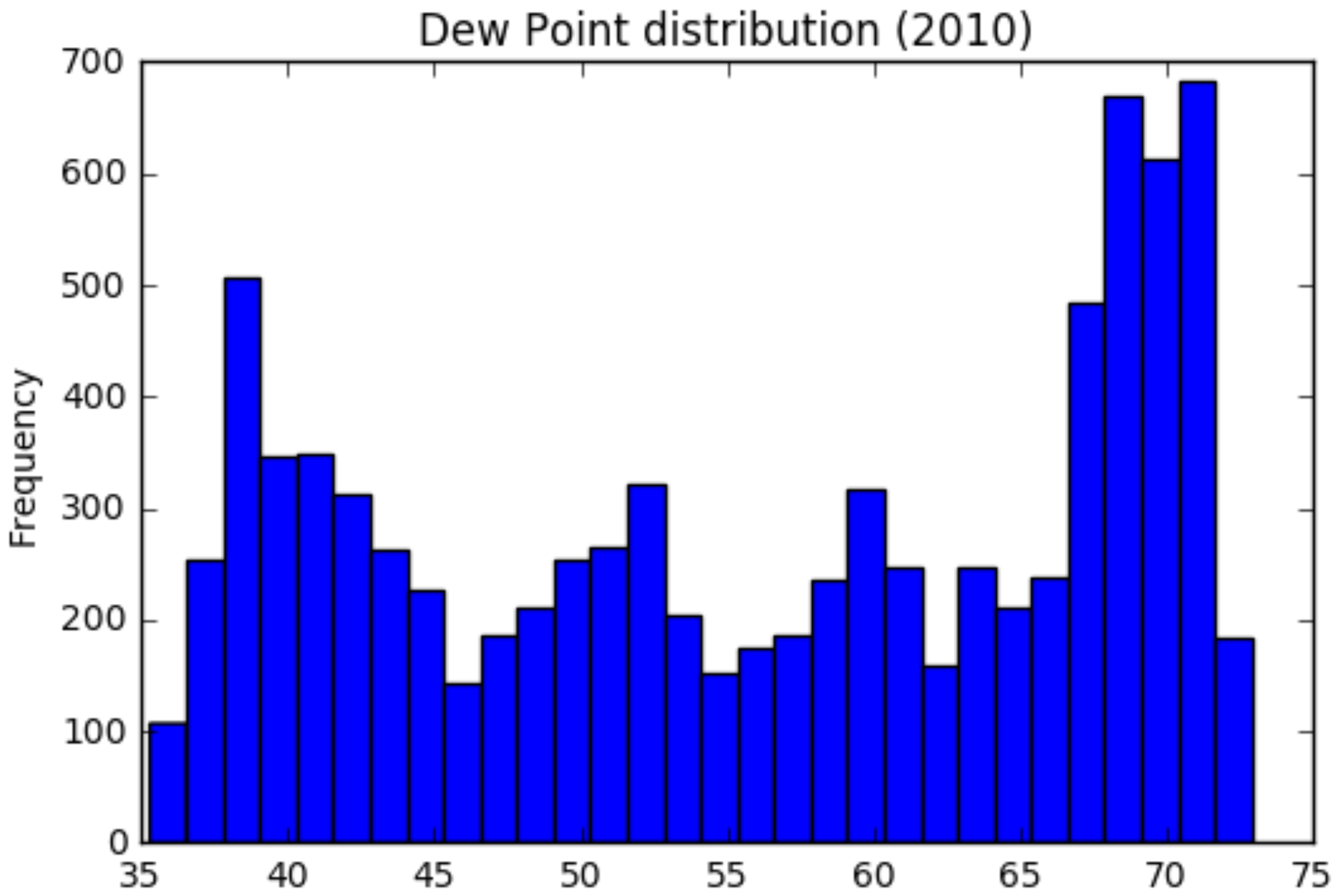
```
In [1]: import matplotlib.pyplot as plt

In [2]: climate2010.Temperature['2010-07'].plot()

In [3]: plt.title('Temperature (July 2010)')

In [4]: plt.show()
```

Histograms in pandas



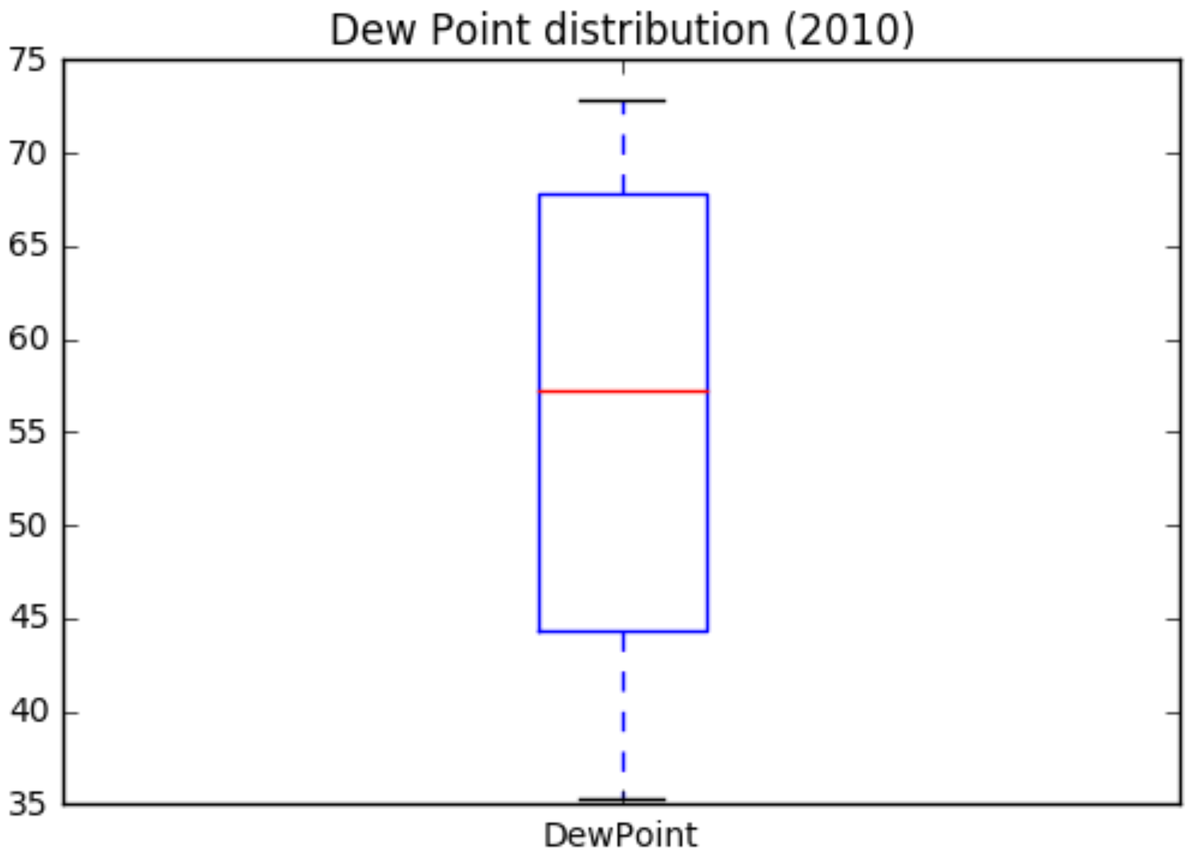
Histograms in pandas

```
In [5]: climate2010['DewPoint'].plot(kind= 'hist', bins=30)

In [6]: plt.title('Dew Point distribution (2010)')

In [7]: plt.show()
```

Box plots in pandas



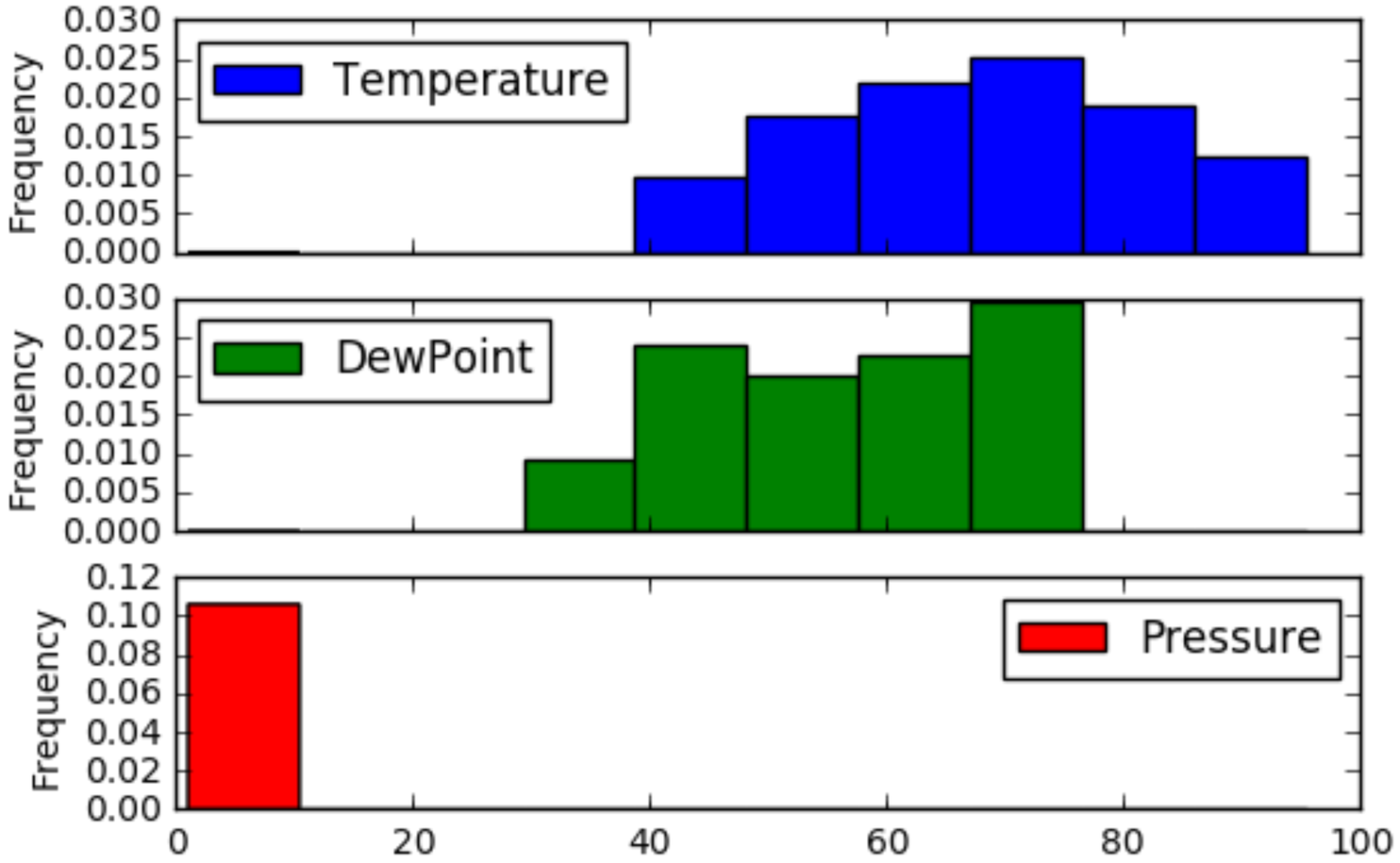
Box plots in pandas

```
In [8]: climate2010['DewPoint'].plot(kind='box')

In [9]: plt.title('Dew Point distribution (2010)')

In [10]: plt.show()
```

Subplots in pandas



Subplots in pandas

```
In [11]: climate2010.plot(kind='hist', normed=True, subplots=True)

In [12]: plt.show()
```

PANDAS FOUNDATIONS

Let’s practice!

PANDAS FOUNDATIONS

Final thoughts

You can now...

- Import many types of datasets and deal with import issues
- Export data to facilitate collaborative data science
- Perform statistical and visual EDA natively in pandas

PANDAS FOUNDATIONS

See you in the next course!