Temporal Data II (Pandas + Analysis)
SI370

March 16, 2017

Eytan Adar

Last Time: Temporal data I

- What is temporal data?
- How do we visualize it?
  - Time series
  - Clustering
  - Event-driven analysis
- Manipulation
  - Pandas examples

Case 7

- Abstract task:
  - Find repeating patterns (seasonality) in time series datasets
  - Isolate the periodicity
  - Compare aligned time series

Spiral visualizations
Carlis et al., 1998

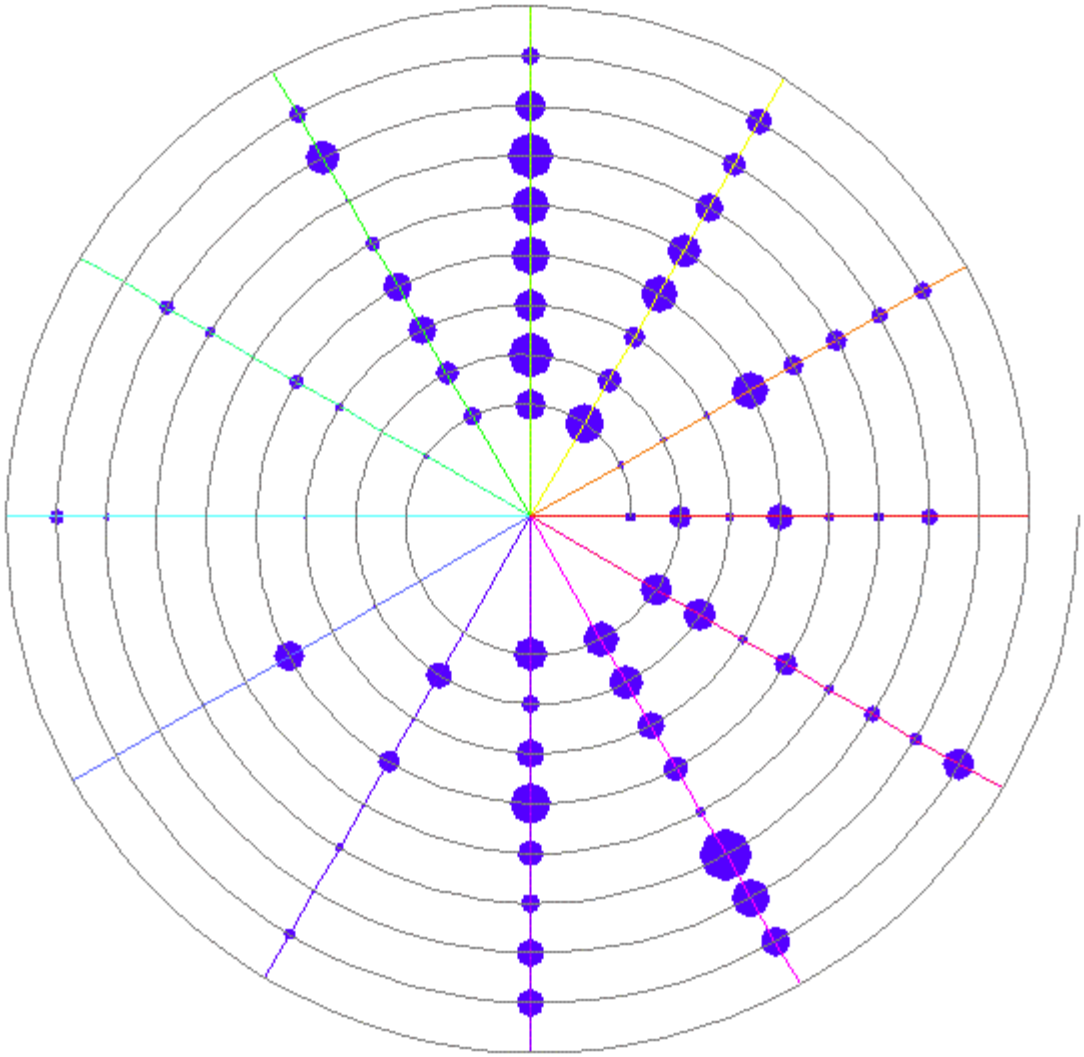


Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for Baphia Capparidifolia during the period 1980 – 1988.

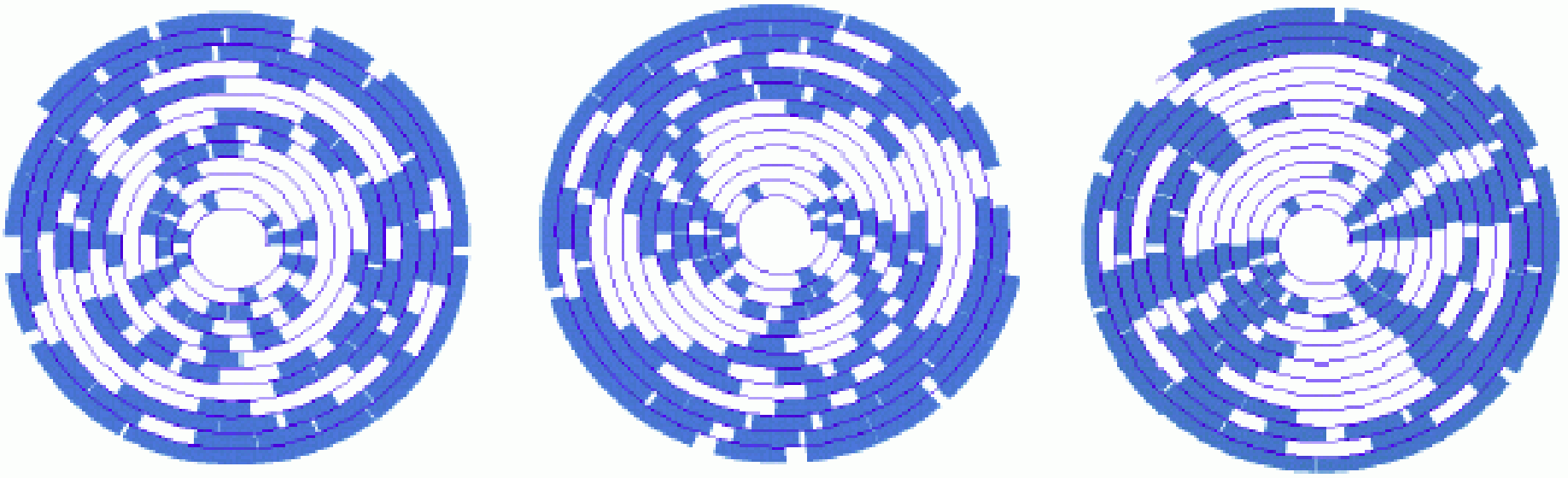


Figure 7. Tightening a spiral view of sound data from five instruments. From left to right the structure of the sound reveals itself.

Spiral visualizations
Carlis et al., 1998

https://eagereyes.org/techniques/spirals

Summary

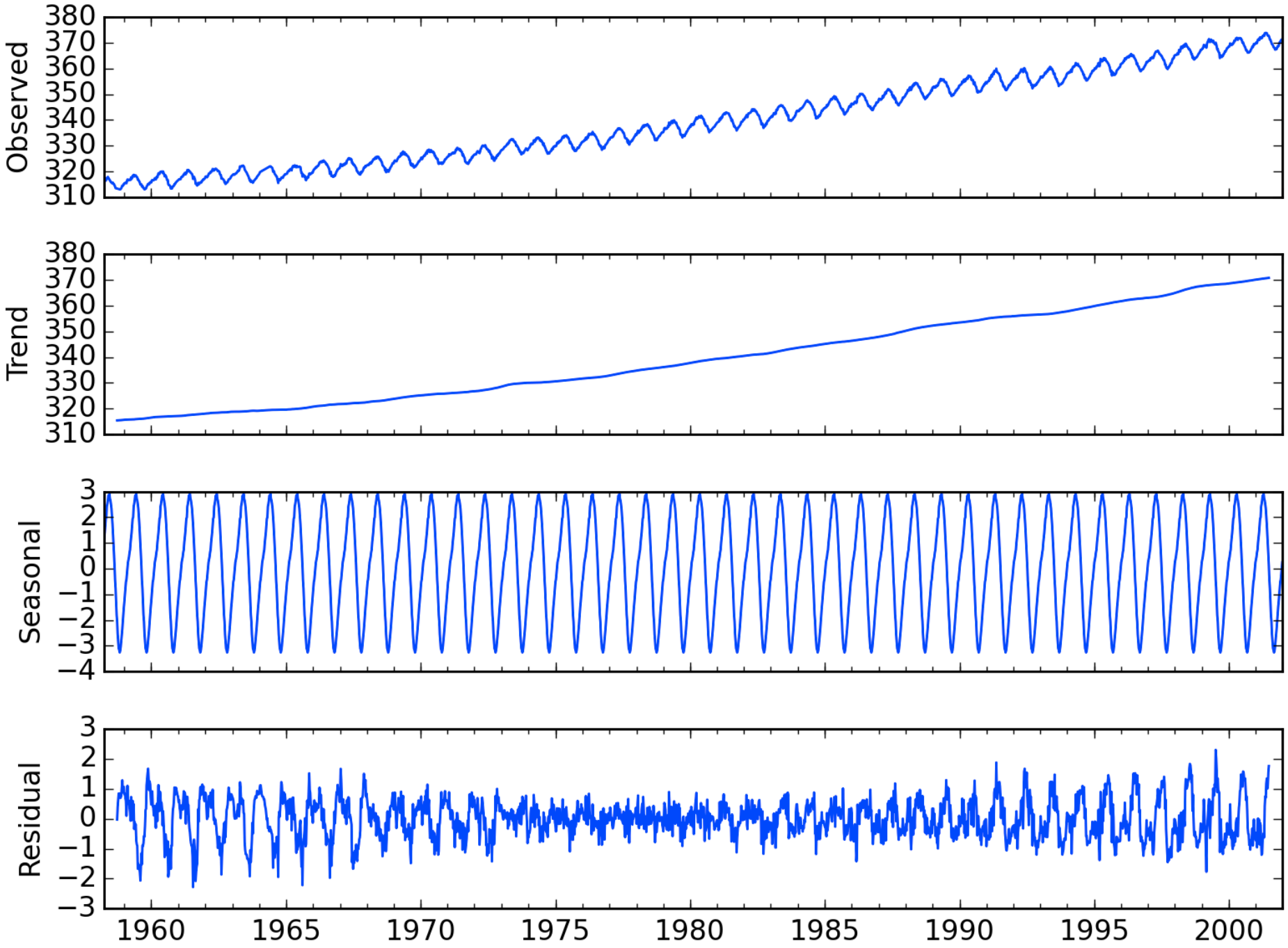
- Identify the different domain goals
- Isolate the tasks (there are many)
  - Select encodings that support the tasks
  - Standard types are a good place to start
    - Additional retinal variables often useful
  - Some less standard types might also be appropriate
- Now for some Pandas…

Today: Temporal data

- What is temporal data?
- How do we visualize it?
  - Time series
  - Clustering
  - Event-driven analysis
- Manipulation
  - Pandas examples
- Analysis
  - Trends and seasonality
  - Cross-correlation
  - Autocorrelation

A few key concepts

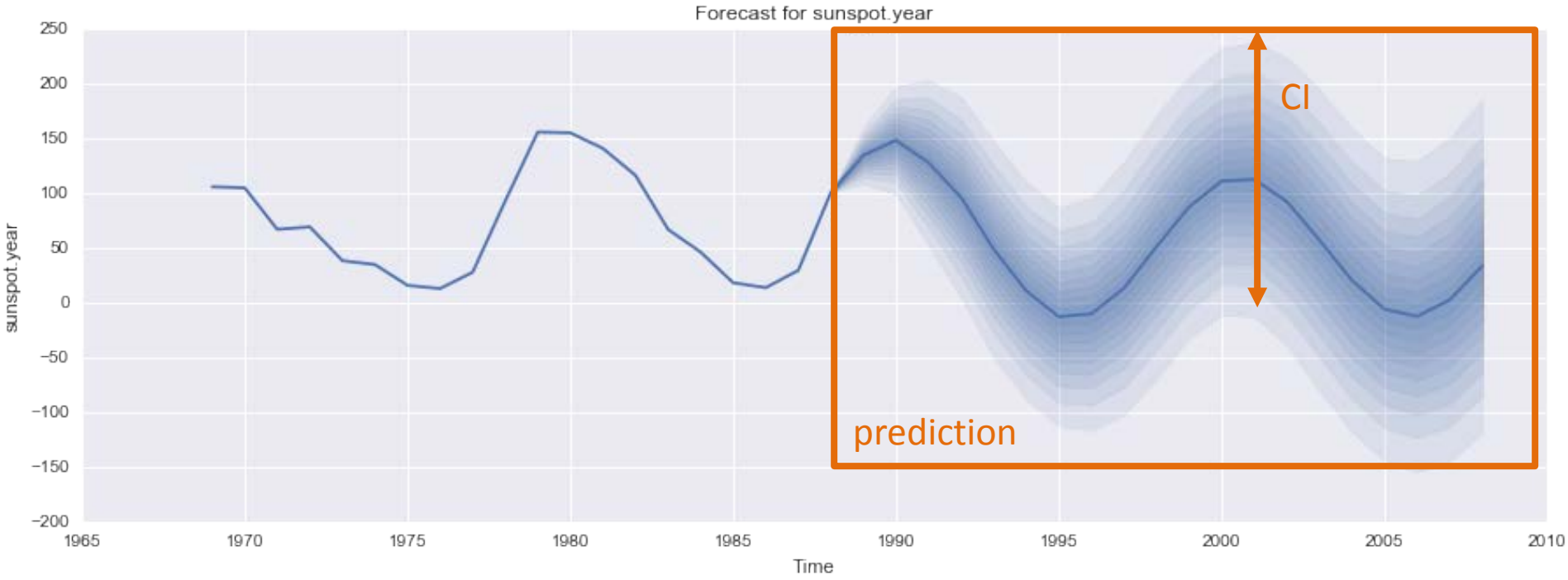
- “Univariate” time-series
  - We care about how it breaks down


Observed
380
370
360
350
340
330
320
310
Trend
380
370
360
350
340
330
320
310
Seasonal
3
2
1
0
−1
−2
−3
−4
Residual
3
2
1
0
−1
−2
−3
1960
1965
1970
1975
1980
1985
1990
1995
2000

A few key concepts

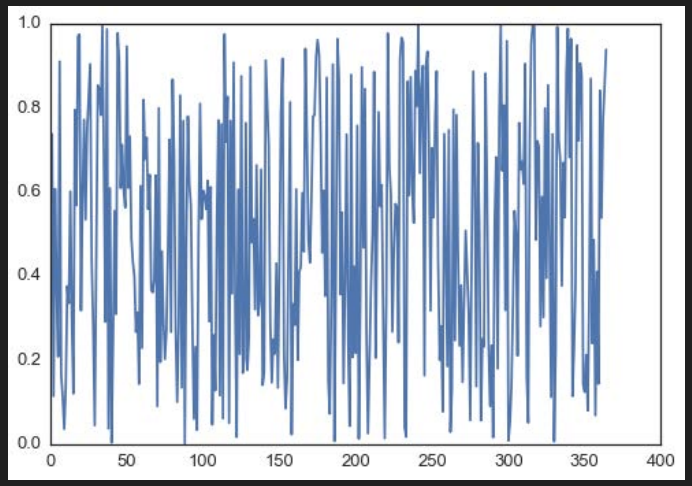
- “Univariate” time-series
  - We care about how it breaks down
    - Seasonality, trend, residuals
  - We care about what value it will take in the future

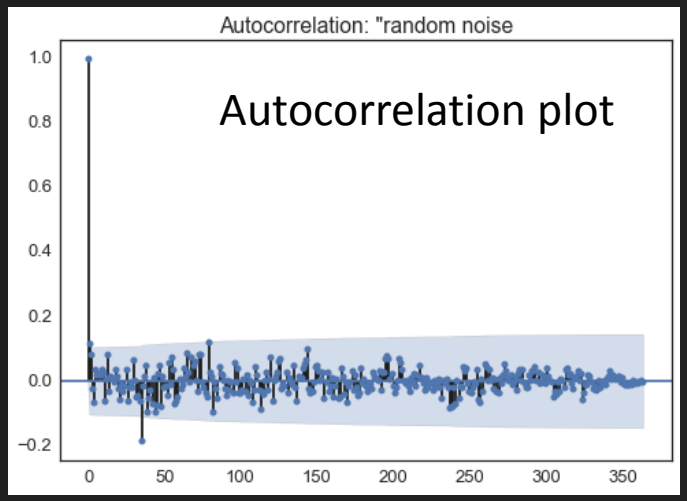
ARIMA model for sunspots

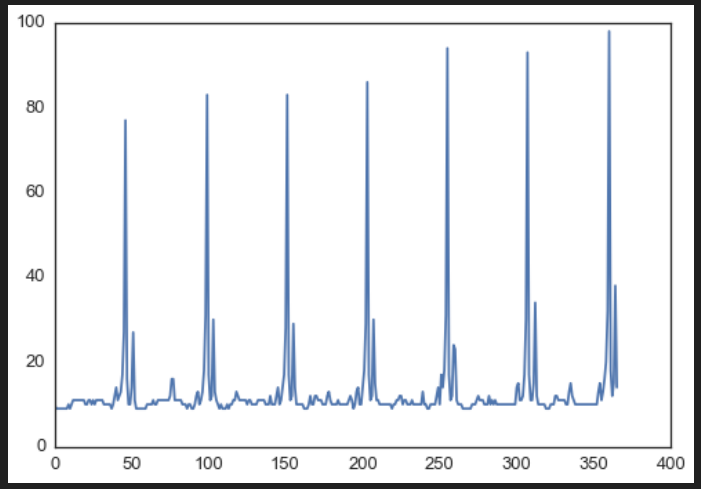


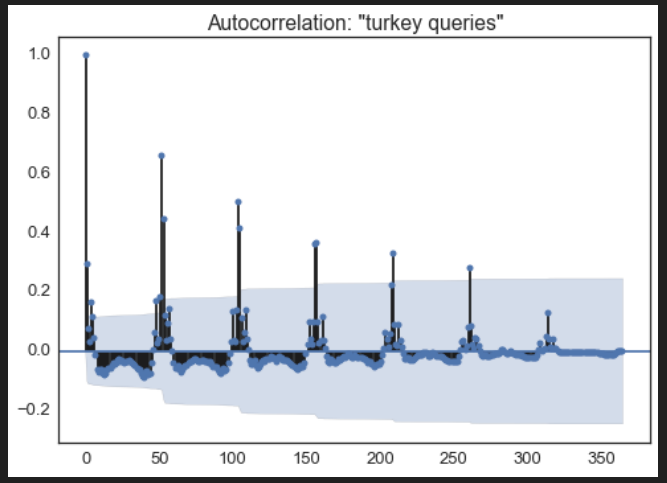
A few key concepts

- “Univariate” time-series
  - We care about how it breaks down
    - Seasonality, trend, residuals
  - We care about what value it will take in the future
    - For example: ARIMA
      - Determine parameters (ACF, PACF)


1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350
400


Autocorrelation: "random noise
Autocorrelation plot
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
50
100
150
200
250
300
350


100
80
60
40
20
0
0
50
100
150
200
250
300
350
400

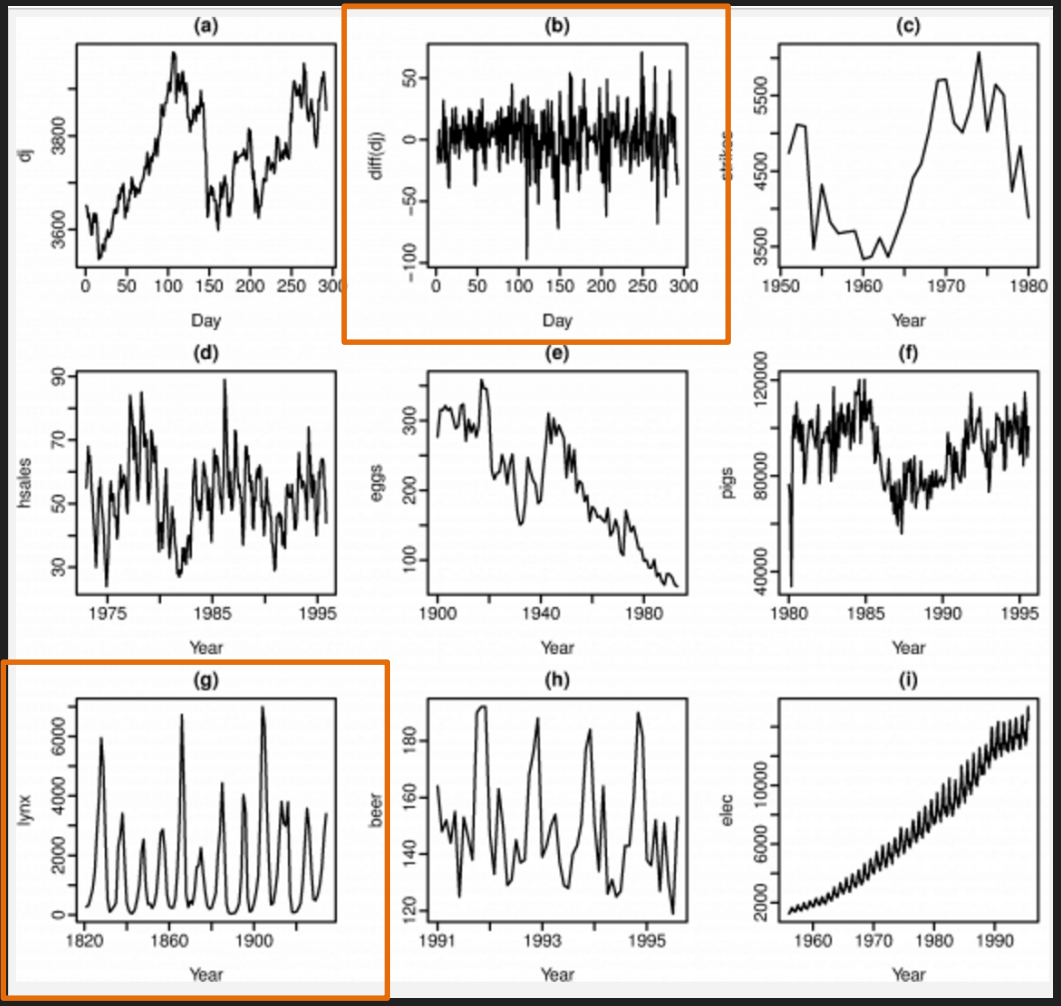

Autocorrelation: "turkey queries"
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
50
100
150
200
250
300
350

Auto-correlation

- Is the data noise?
- Compare the data to itself at different lags
  - If we only see one “peak” (basically at lag 0)
    - Random
  - If we see other peaks/hills
    - Indication of seasonality
    - Or that value at t_n “depends on value” t_{n-1}

A few key concepts

- “Univariate” time-series
  - We care about how it breaks down
    - Seasonality, trend, residuals
  - We care about what value it will take in the future
    - For example: ARIMA
    - Some assumptions about this: stationary



(a) *Dow Jones index on 292 consecutive days;* ***(b)*** *Daily change in Dow Jones index on 292 consecutive days;* ***(c)*** *Annual number of strikes in the US;* ***(d)*** *Monthly sales of new one-family houses sold in the US;* ***(e)*** *Price of a dozen eggs in the US (constant dollars);* ***(f)*** *Monthly total of pigs slaughtered in Victoria, Australia;* ***(g)*** *Annual total of lynx trapped in the McKenzie River district of northwest Canada;* ***(h)*** *Monthly Australian beer production;* ***(i)*** *Monthly Australian electricity production*

A few key concepts

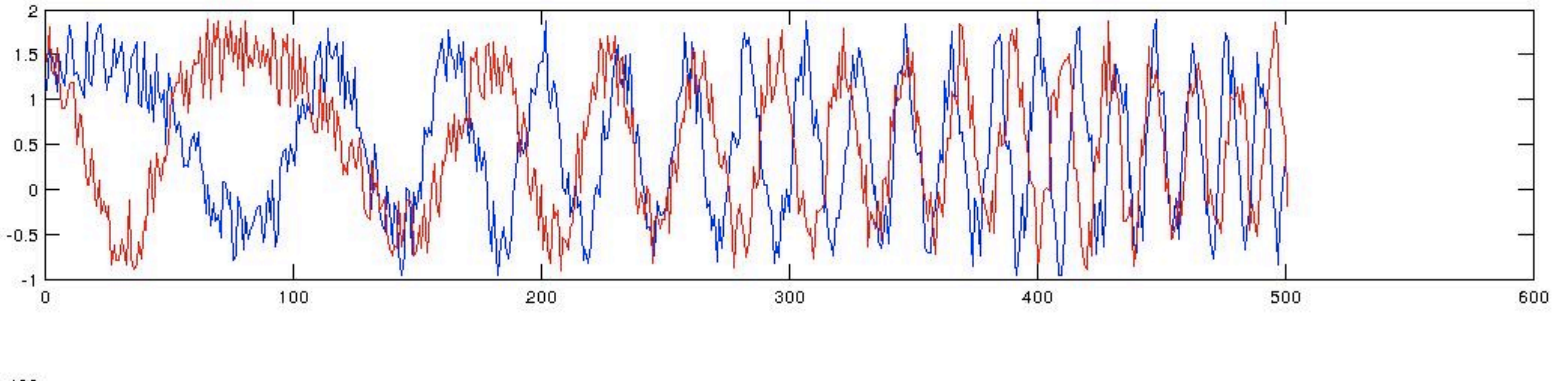
- “Univariate” time-series
  - We care about how it breaks down
    - Seasonality, trend, residuals
  - We care about what value it will take in the future
    - For example: ARIMA
    - Some assumptions about this: stationary
      - We can make data stationary

A few key concepts

- Bi-variate time-series analysis
  - How related are two time series?
    - Cross-correlation
      - A bit like autocorrelation

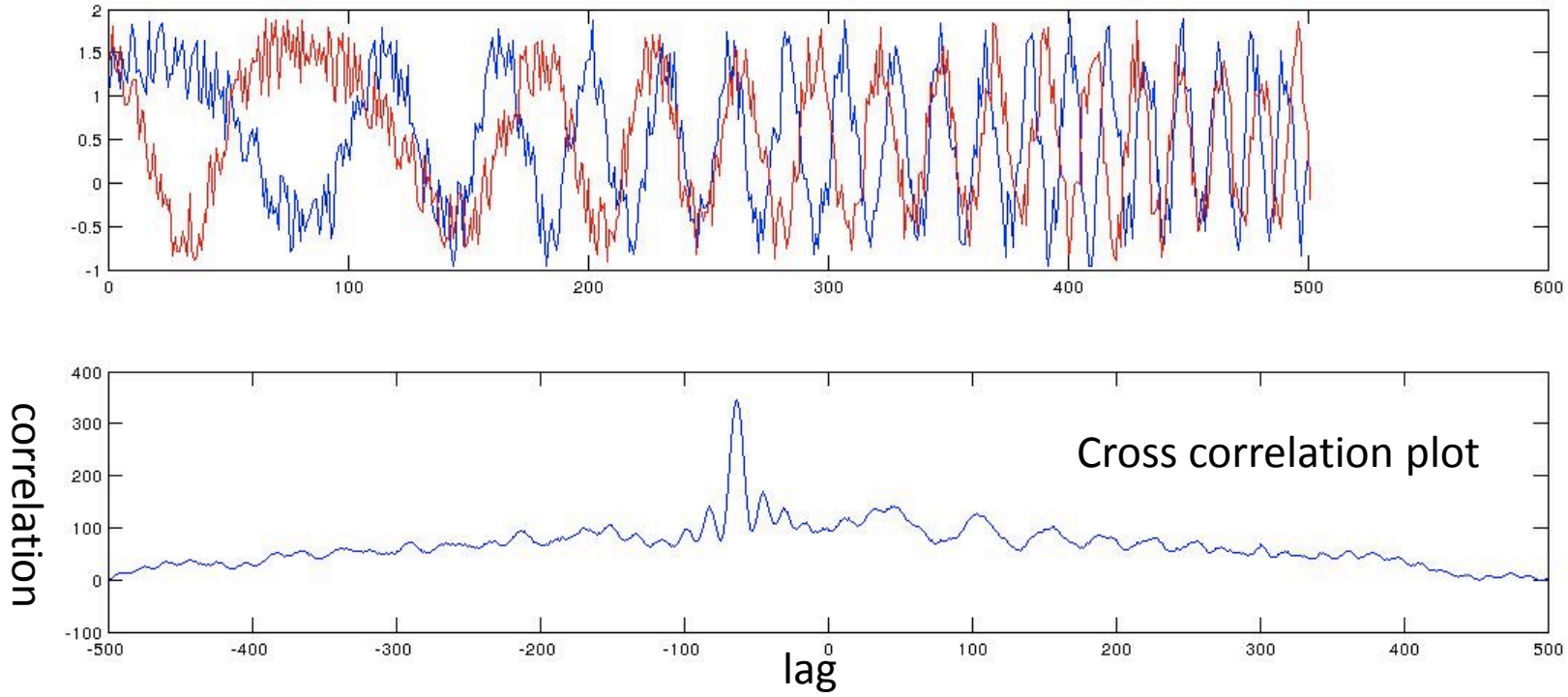
Cross-Correlation

- Works much the same as standard correlation.
  - corr(x,y) → treats each pair as “point”
  - t1 → x1,y1; t2 → x2,y2 ; …
- One exception:
  - Lags/delays
  - Often one series is in advance of other


2
1.5
1
0.5
0
-0.5
-1
0
100
200
300
400
500
600

Cross-Correlation

- Works much the same as standard correlation.
  - corr(x,y) → treats each pair as “point”
  - t1 → x1,y1; t2 → x2,y2 ; …
- One exception:
  - Lags/delays
  - Often one series is in advance of other
- We can include a “lag”
  - Often part of utility function… provide min/max lag
  - Find max value (above some critical point)


2
1.5
1
0.5
0
-0.5
-1
0
100
200
300
400
500
600
400
300
200
100
0
-100
-500
-400
-300
-200
-100
0
100
200
300
400
500
correlation
lag
Cross correlation plot

Summary

- Temporal data sets are ubiquitous
- Need to carefully consider the problem
  - Different ways to structure the data
  - Different ways to do EDA
    - Visual forms and statistical
  - Statistics are wonky… data needs to be “cleaned” differently than other datasets
- Now… more Pandas