ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 17: Time Series

William Marsh

See also introductory notebook

Outline

- Aims
  - Understand why time series data is different
  - Basic operations on time series data

- Meanings of time
- Time series and trends
  - Moving average
  - Changing period (resampling)
- Periodicity and randomness
- Auto regression and auto correlation

Introduction

Meanings of Time

- Timestamp
  - A specific instance
  - Python type ‘datetime’; Pandas ‘Timestamp’
- Interval or Period
  - The time between two instances
  - ‘A week later’ or a ‘month later’
- Duration
  - How long it takes to ...
  - Time as data (cf. time as the index)

Our concern here is with ‘time as an index’

Time Issues

- Pseudo periods
  - Year: varies in length (leap year)
  - Month: different numbers of days
- Time zone:
  - It is between 9 and 10 somewhere
  - UTC used but also local time
- Daylight saving time
  - How many hours in a day?

Date Representation and Parsing

- A data file usually has dates and times as strings
  - Different formats e.g. UK days first, US month first
- Integer representation of dates
  - Y2K panic (aka ‘millennium bug’)
  - Seconds from a baseline
  - Watch out for spreadsheets (*what a mess*)

Excel supports two date systems, the 1900 date system and the 1904 date system. Each date system uses a unique starting date from which all other workbook dates are calculated. All versions of Excel for Windows calculate dates based on the 1900 date system. Excel 2008 for Mac and earlier Excel for Mac versions calculate dates based on the 1904 date system. Excel 2016 for Mac and Excel for Mac 2011 use the 1900 date system, which guarantees date compatibility with Excel for Windows.

https://support.microsoft.com/en-us/office/date-systems-in-excel-e7fe7167-48a9-4b96-bb53-5612a800b487

Dates as Objects: Pandas

- timestamp
  - A date time representation
  - Understands calendars i.e. which dates exist
- timedelta
  - An interval
- dateoffset
  - A calendar-aware offset (e.g. add one month)

Comprehensive but complex

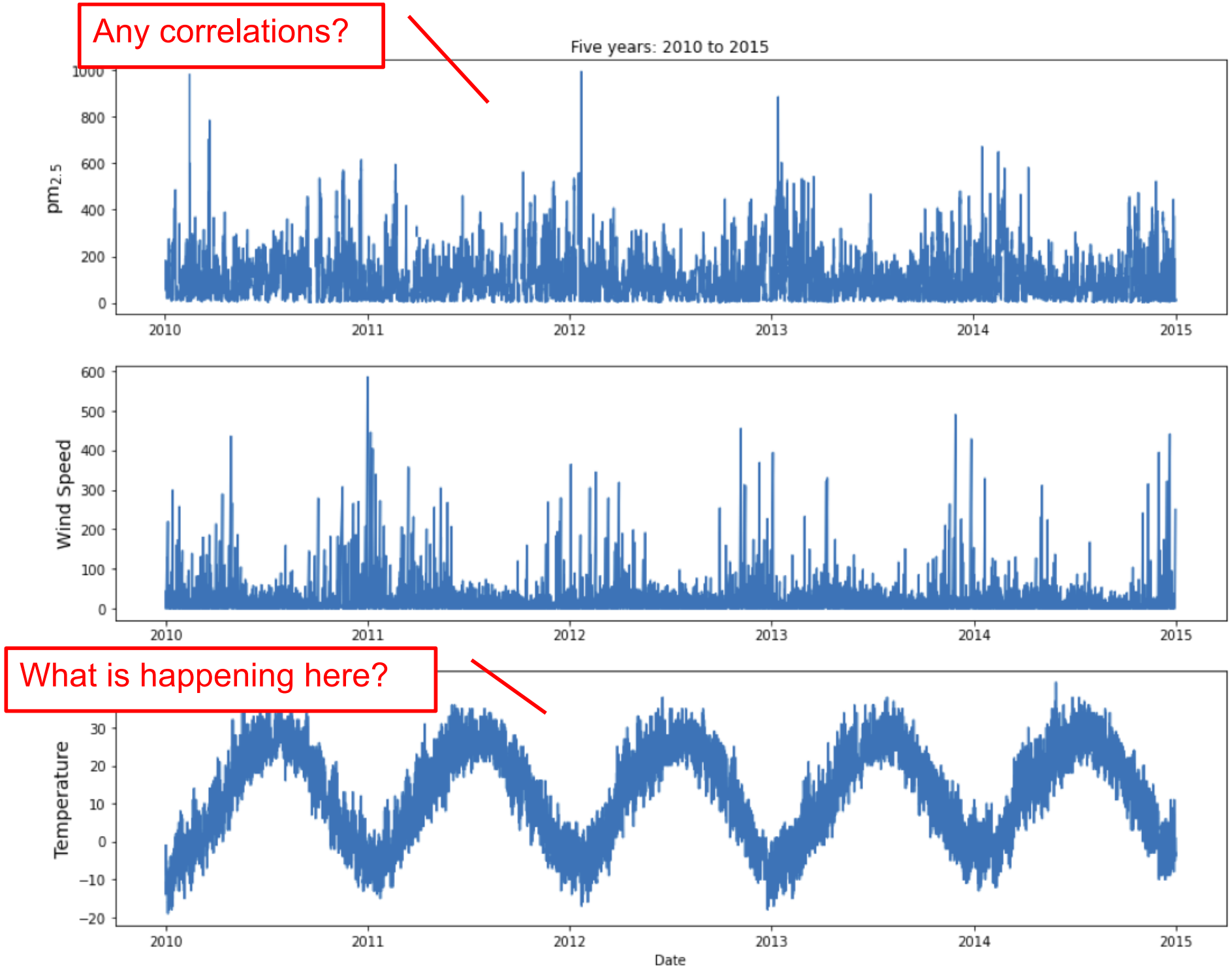
Time Series

- Data value or values indexed by time
- Typically
  - Fixed period e.g. daily or hourly or annually

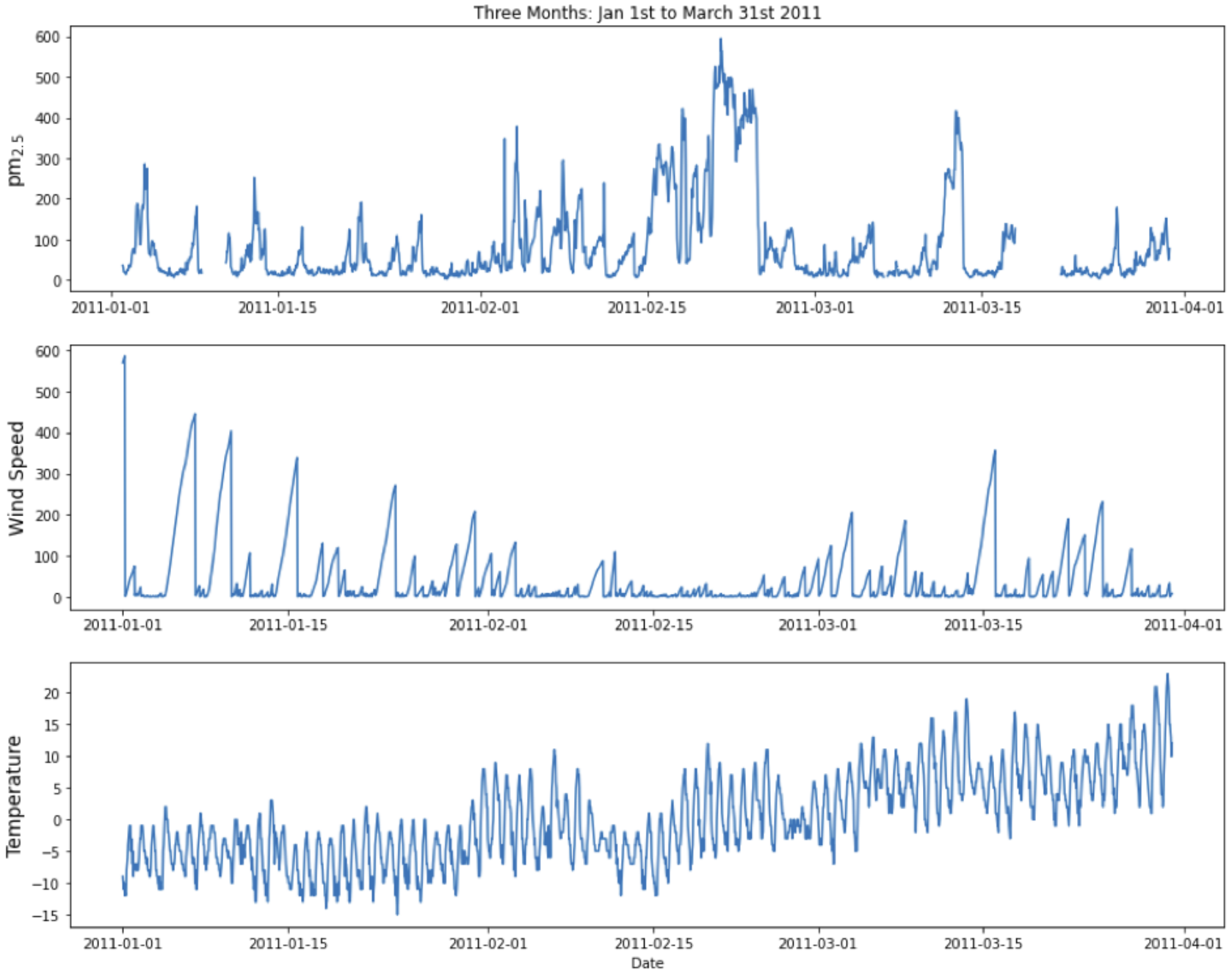
- Notebook has hourly data from Beijing, 2010-2015
  - Separate year, month, day and hour fields
  - Converted to timestamps

Trends and Operations on Time Series

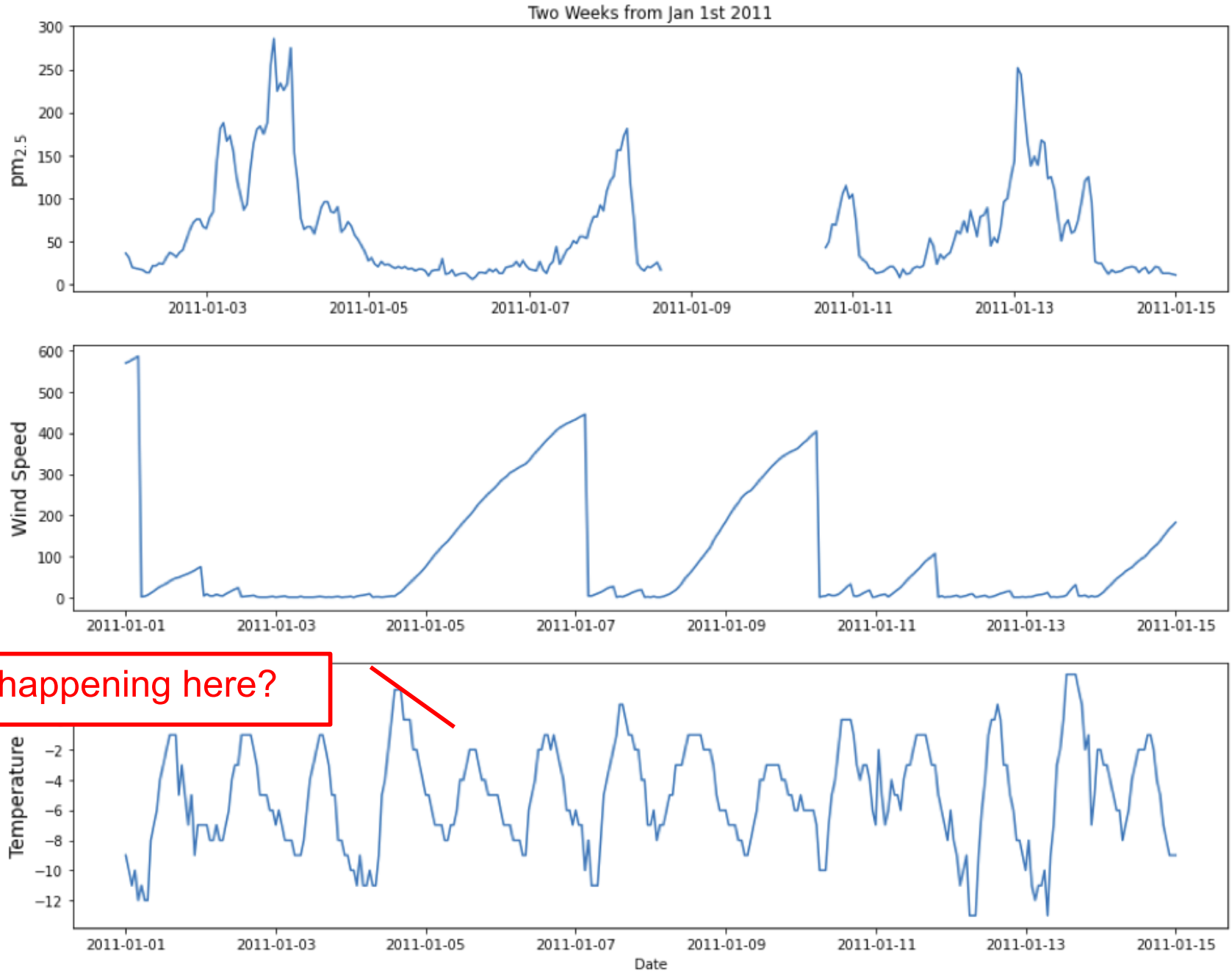
Plotting Against Time


Any correlations?
Five years: 2010 to 2015
pm2.5
1000
800
600
400
200
0
2010
2011
2012
2013
2014
2015
Wind Speed
600
500
400
300
200
100
0
What is happening here?
Temperature
30
20
10
0
−10
−20
Date

Selecting a Range: 3 Months

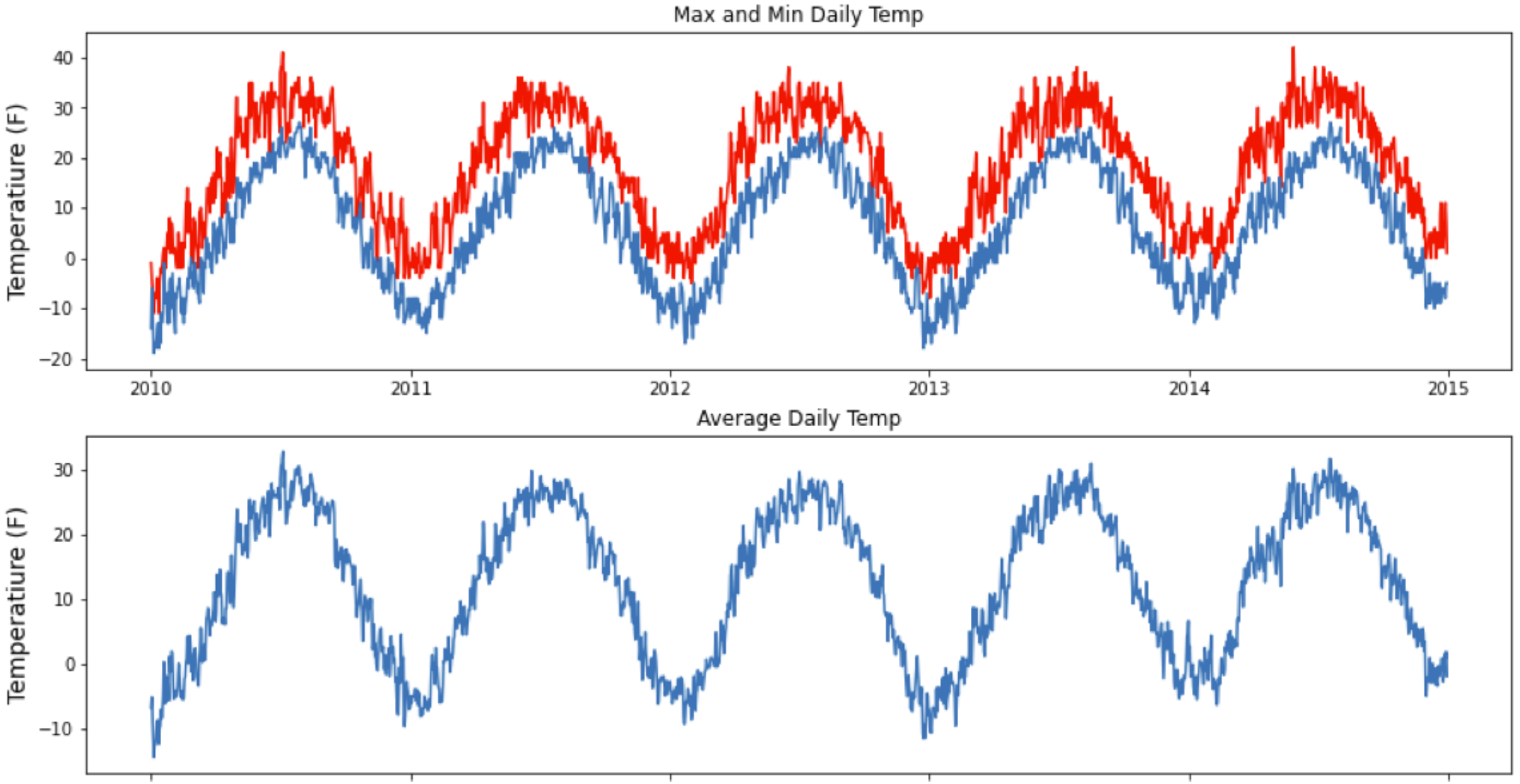


Selecting a Range: 2 Weeks



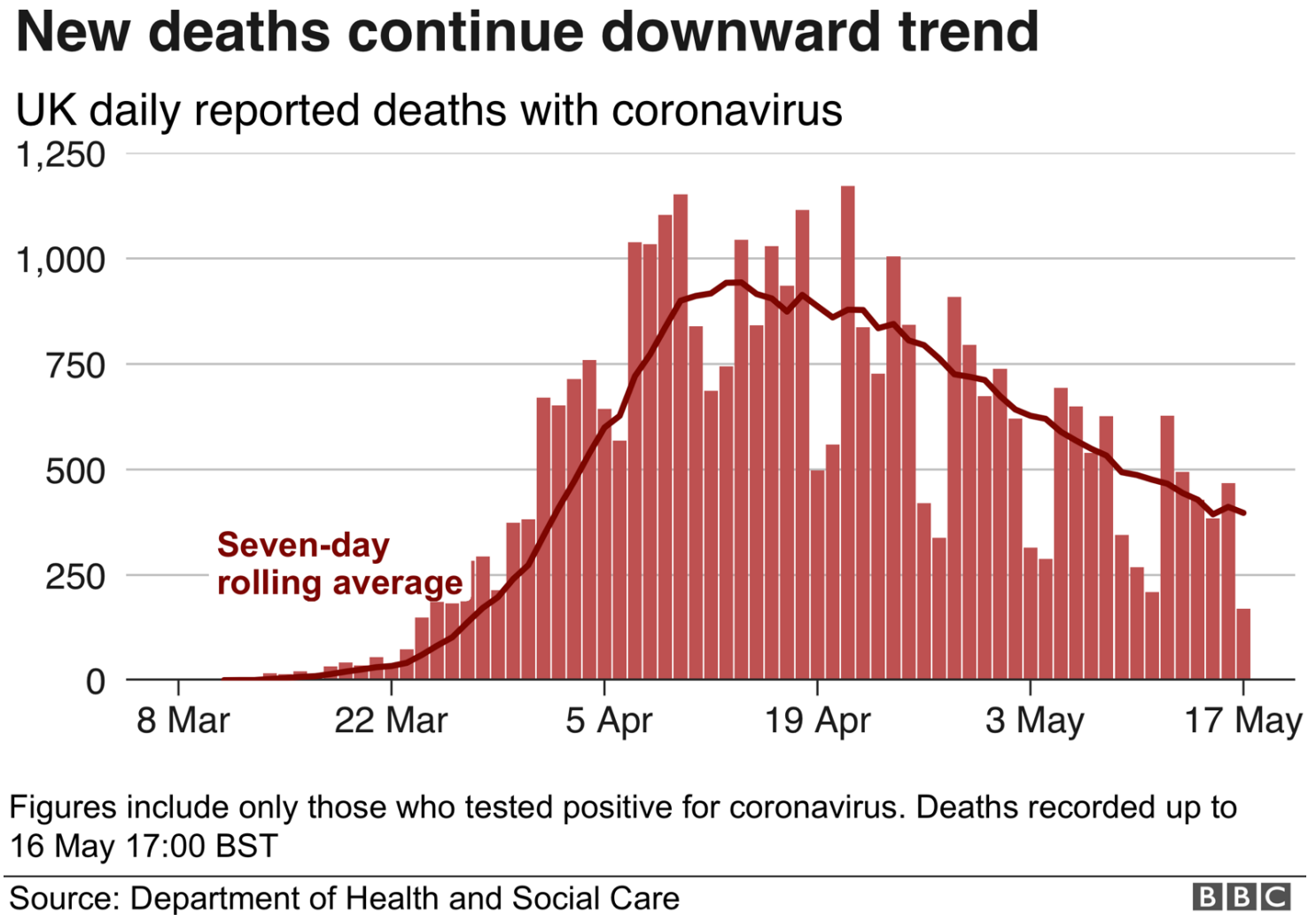
Resample – Frequency Conversion

- Change e.g. hourly data to daily
  - Combine data values: e.g. max, min, mean

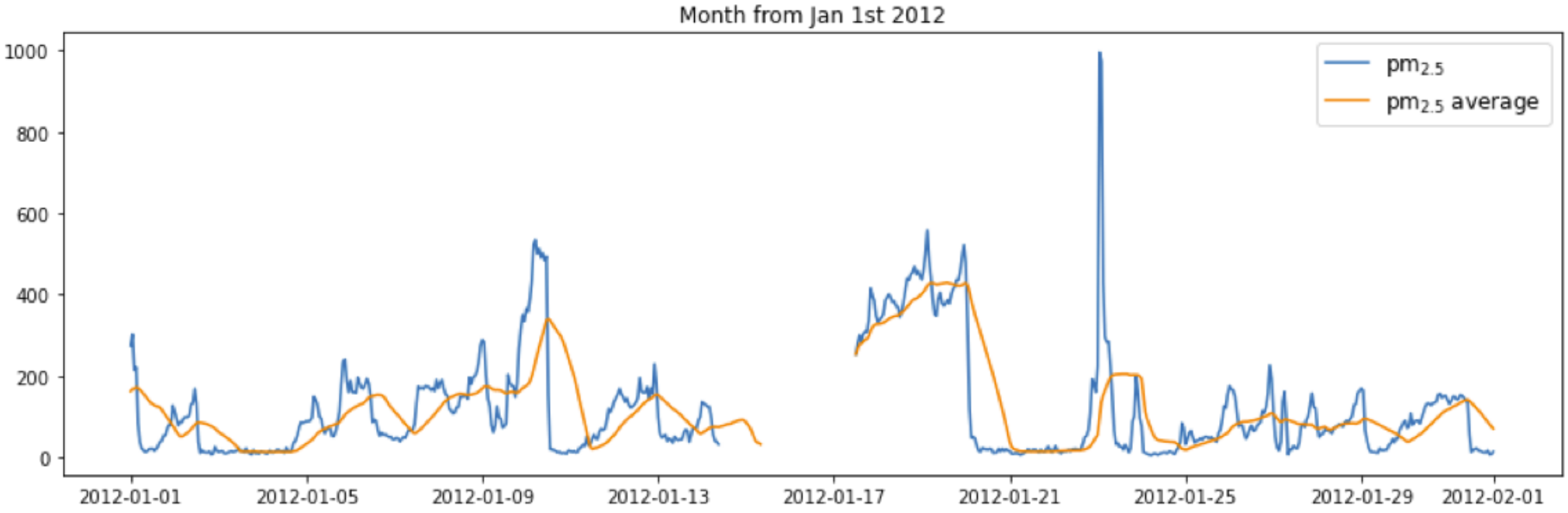


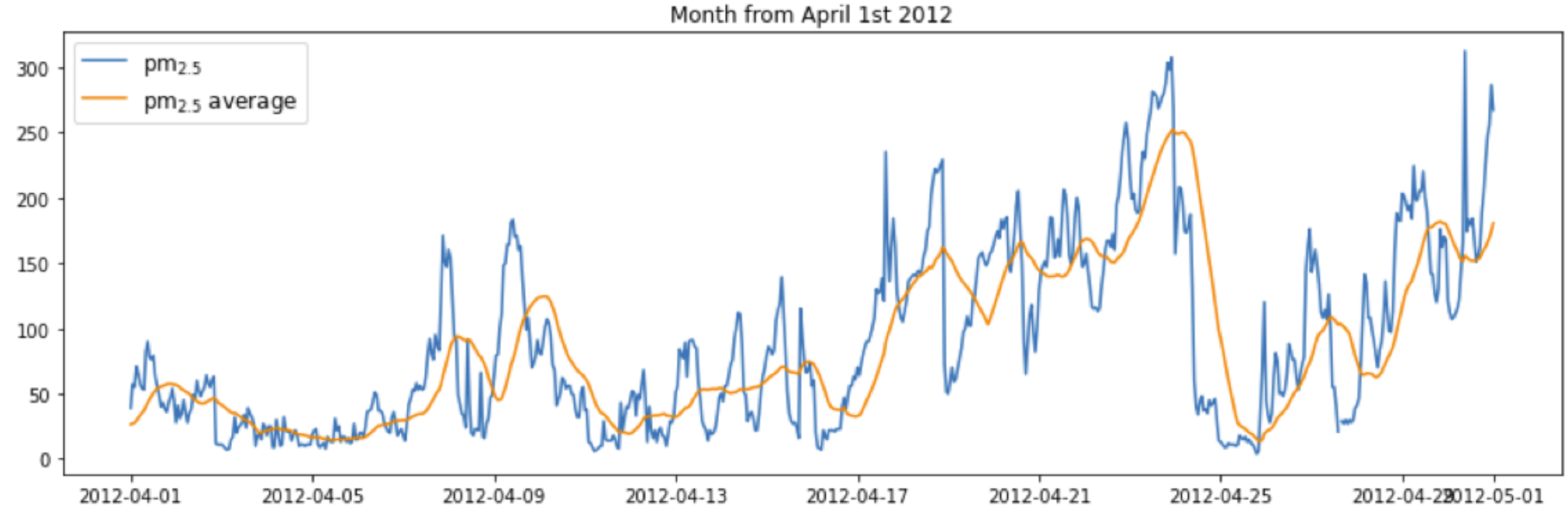
Rolling (Moving) Average

- Average over a window of past values



Rolling Average: Daily





What is a Mean?

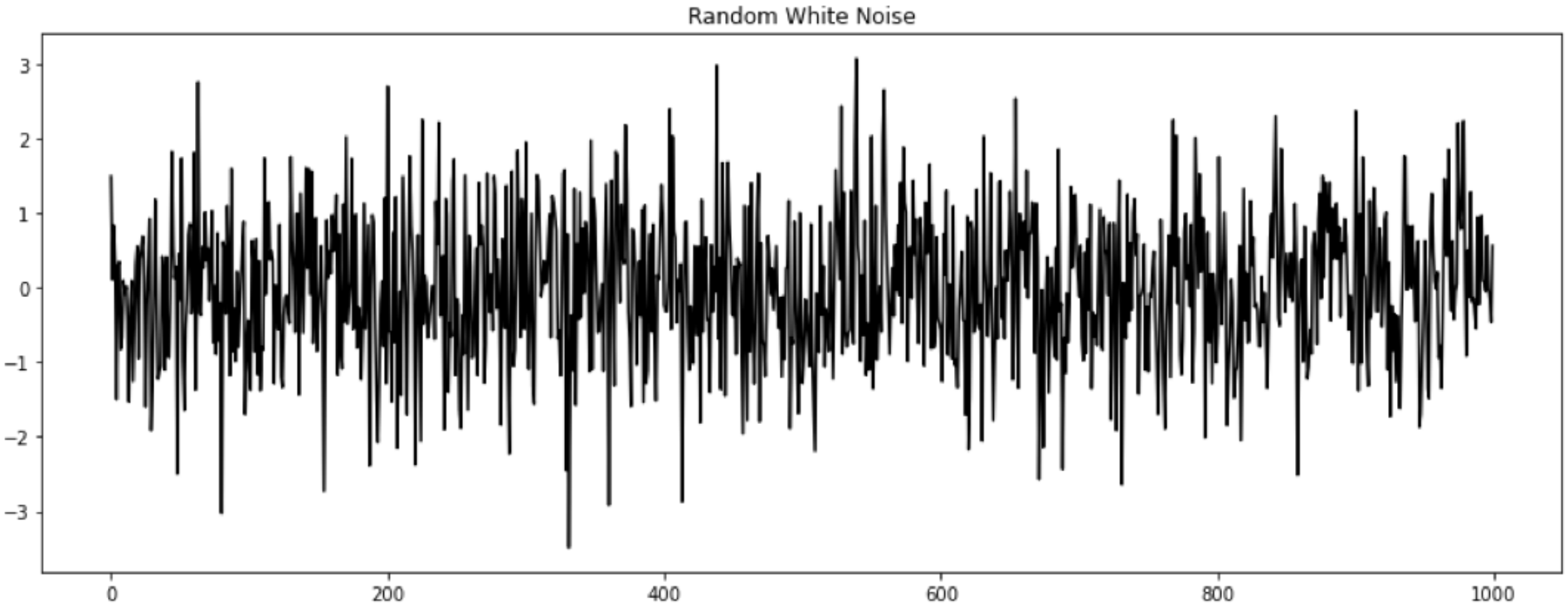
- Since values change with time, the overall mean less useful
  - Daily mean value
  - Monthly mean value
- Same issues apply to other statistics

Idea of a Stochastic Process

Analogous to a distribution

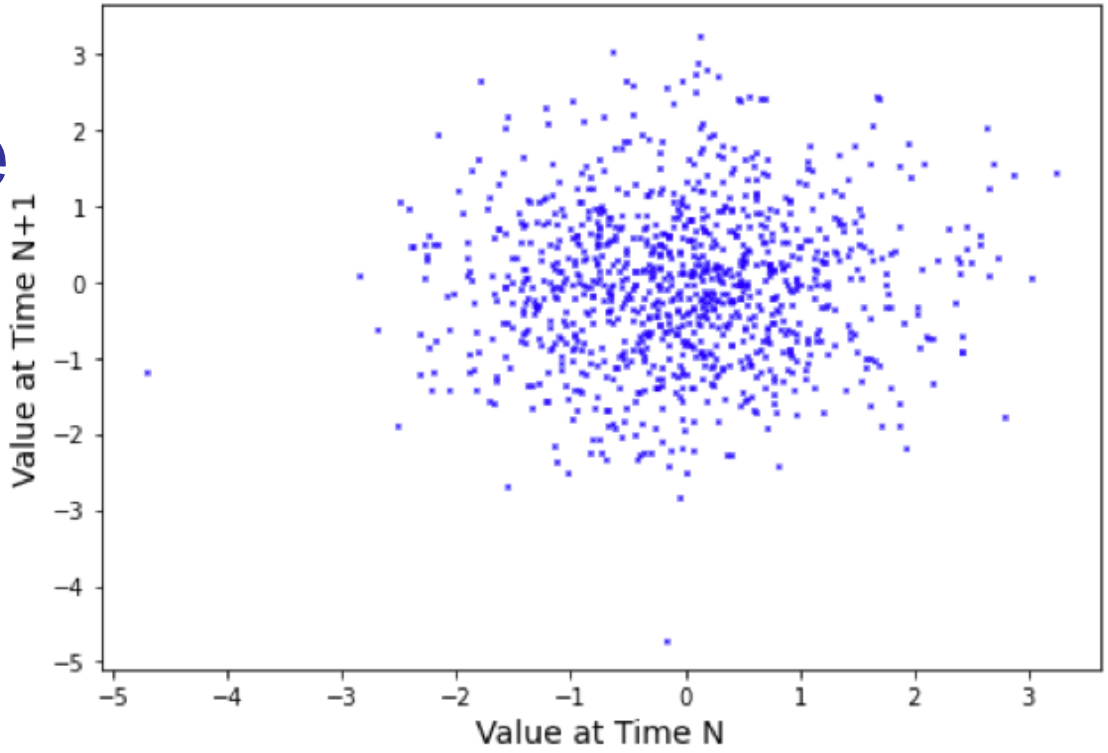
Gaussian Noise

- Data at each time is independent of the previous time(s)
  - $t_n \sim \text{norm}(0, \boldsymbol{\sigma})$

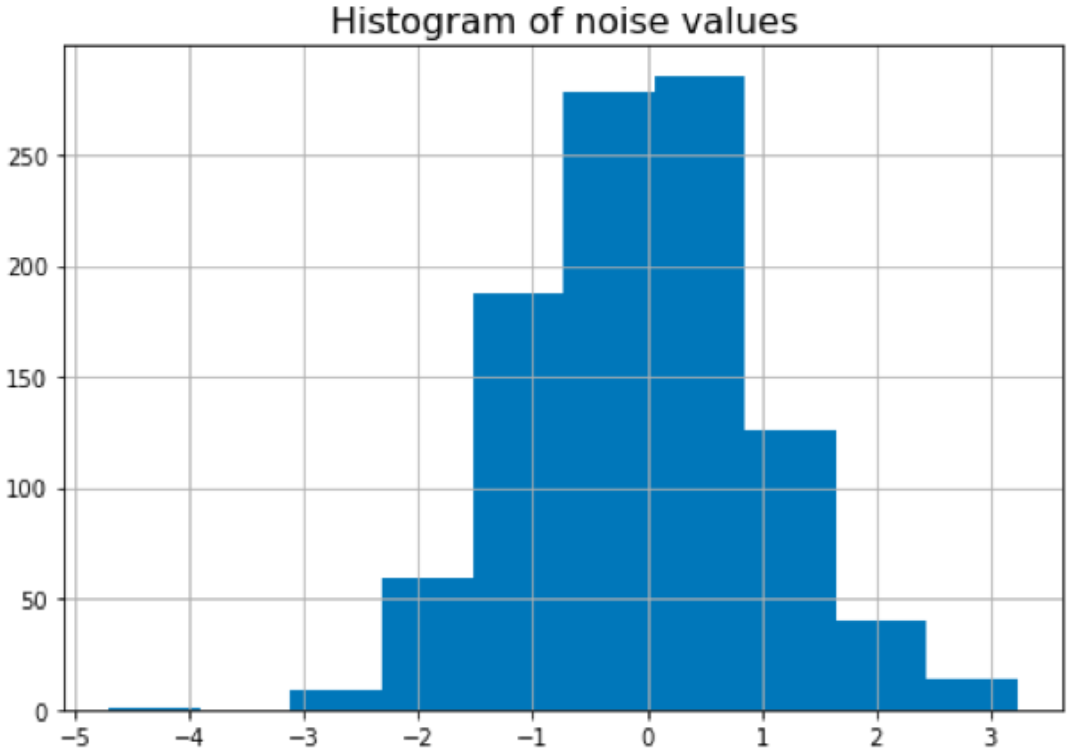


Gaussian Noise

- Scatter plot
  - Successive values unrelated

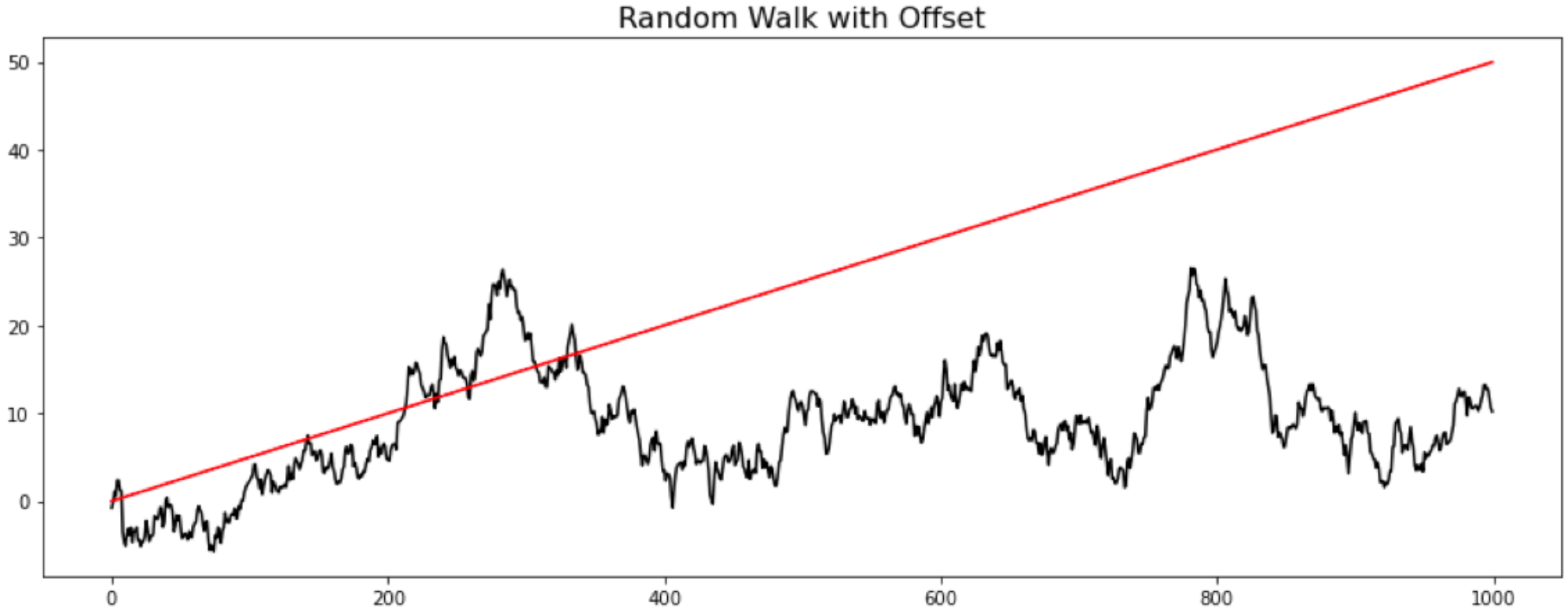


- Value distribution Gaussian



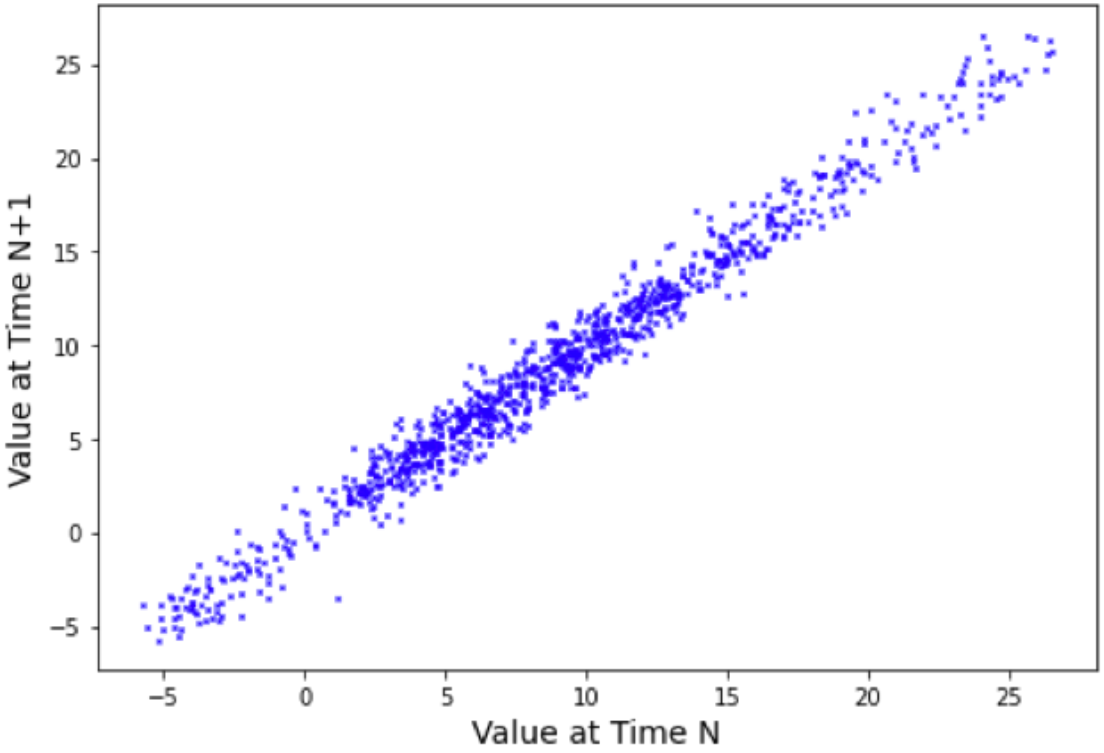
Random Walk (Weiner process)

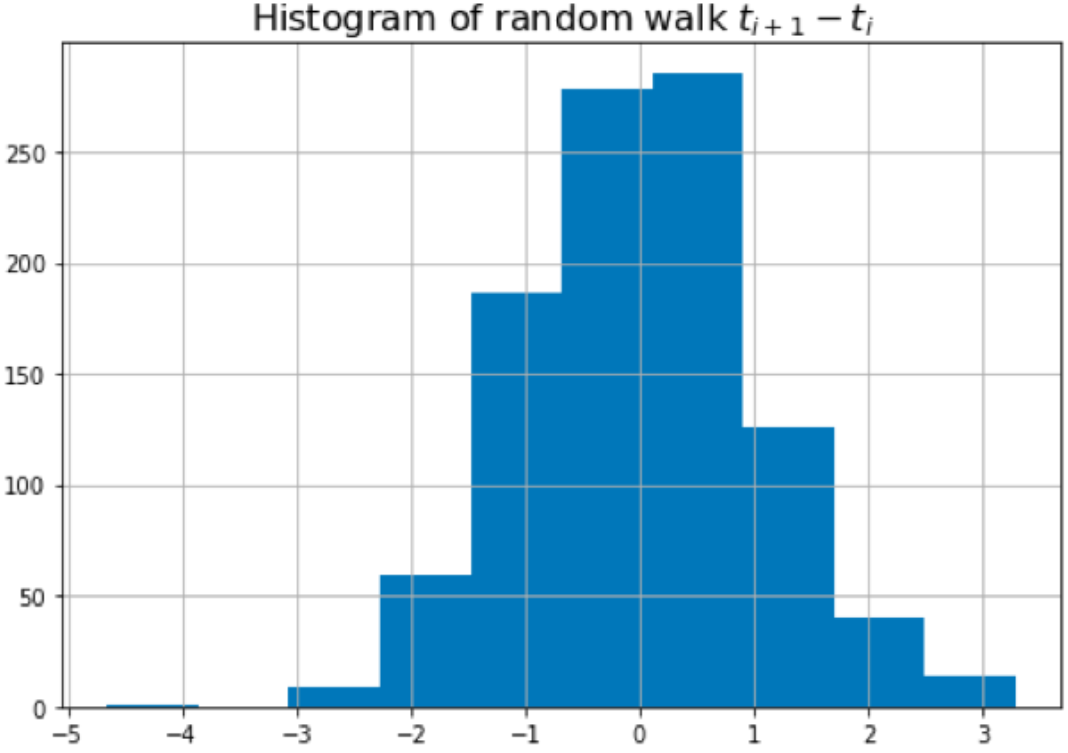
- Each value has fixed and random offset from previous
  - $t_n = t_{n-1} + \delta + \varepsilon_n$ where, $\varepsilon_n \sim \text{norm}(0, \sigma)$



Random Walk

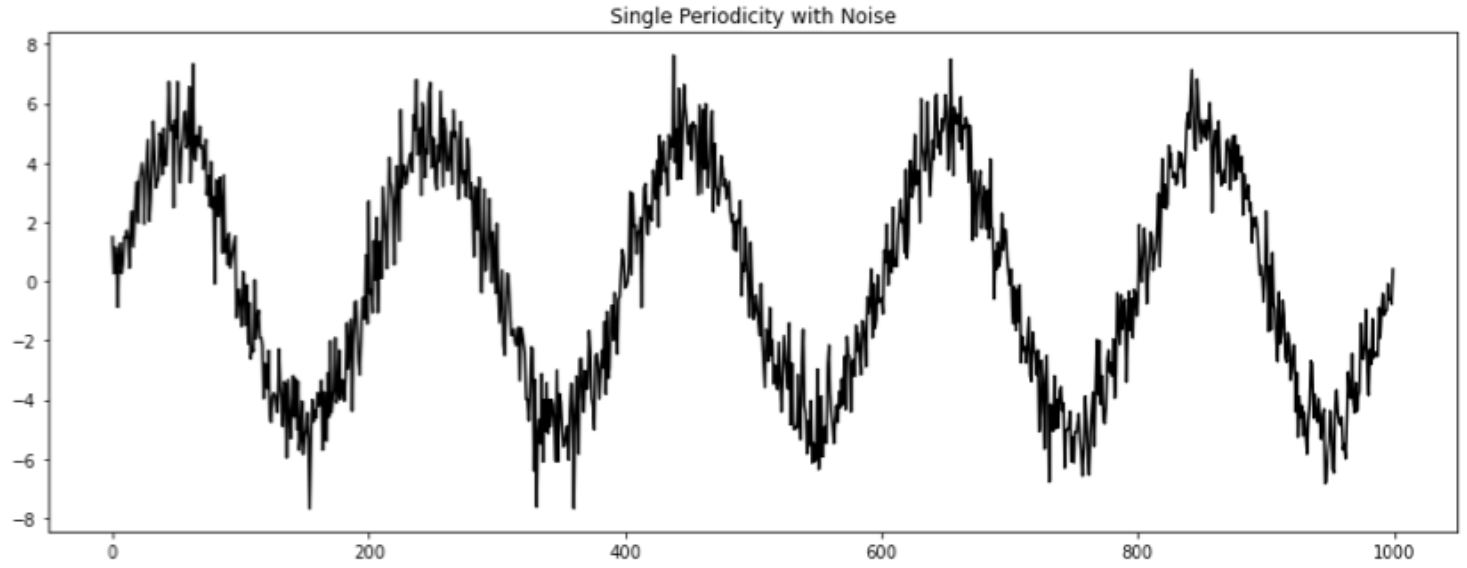
- Successive values related
  - Randomness disguised
- Distribution of difference
  - Gaussian
  - Mean > 0

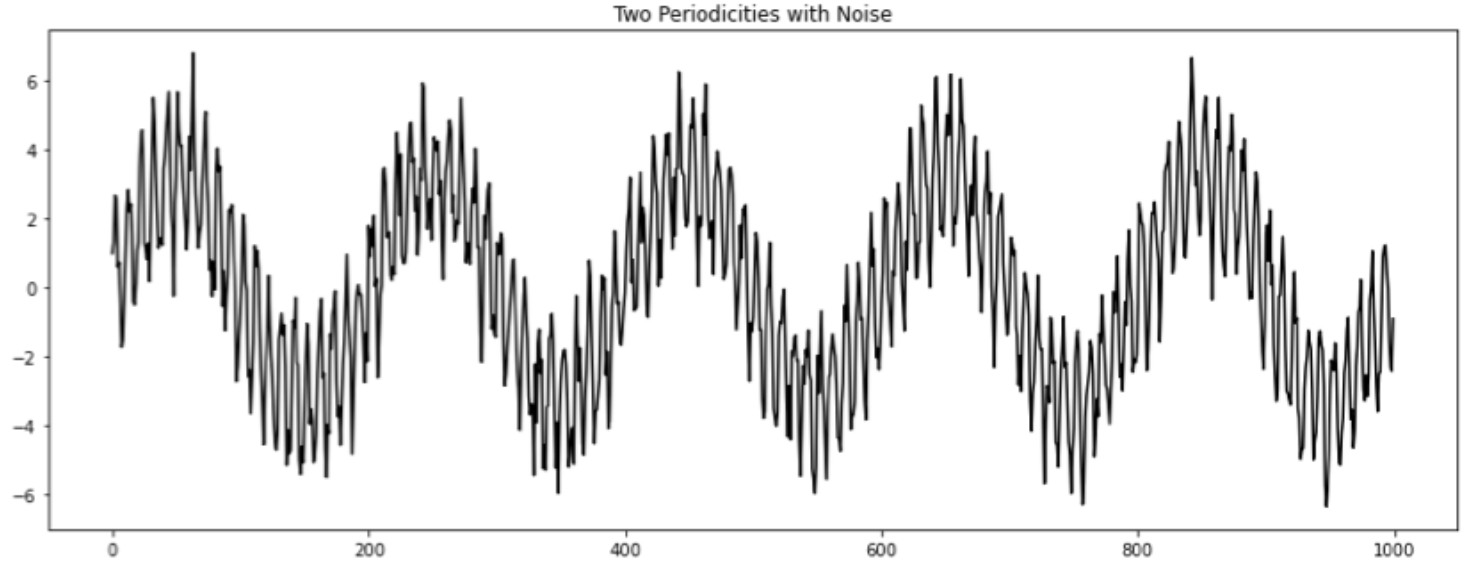




Periodic Processes

- Daily, weekly and seasonal periods possible
- Combined





Going Further

- Stochastic processes can be
  - Discrete time (e.g. daily, hourly)
  - Continuous time

Models of change: not just change in time

- Bernoulli process
  - Discrete
  - Value true or false at any time point
- Poisson process
  - Discrete or continuous
  - Cumulative number of events (e.g. arrivals)
  - Event rate λ – number of arrivals has Poisson distribution

Preview of C/W 2

Average Property prices

- The data

Name	Description
Date	A date, which is the first of the month, between September 1st 2016 and August 1st 2019. 36 months in total.
Area	The name of an area (or region – see below)
Code	The code for the area (or region – see below)
Detached	Average sale price of a detached property in this area in the month
Semi	Same, for semi-detached property.
Terraced	Same, for a terraced property
Flat	Same, for a flat.

Two Data Files

- Price data
- Data to lookup region

Code Prefix	Classification	Level	Description
E12	English Region	1	One of 9 different English regions
E10	County	2	Local government area. Parts of a region.
E09	London Borough	2	
E08	Metropolitan Boroughs	2	
E06	English unitary authority	2	

Main Requirements

- Data preparation
  - *Always look at data before changing it*
- Trends
- Price changes
- Statistical analysis

Time Series Regression

Goals of Time Series Analysis

- Is there a trend?
  - Is time a predictor
- Causal patterns (explaining)
  - What other time-varying values predict outcome
  - Time ordering: cause before effect
- Forecasting
  - Can we predict a future value?
  - Lag: value today predicts

Trend

- Regression of the form

$$y_t = \beta_0 + \beta_1 t + \epsilon$$

- Combined with other terms

Auto-Regression

- Use an earlier value (lagged value) to predict later value

$$y_t = \beta_0 + \beta_1 y_{t-T} + \epsilon$$

- Used for prediction

Regression with Multiple Time Series

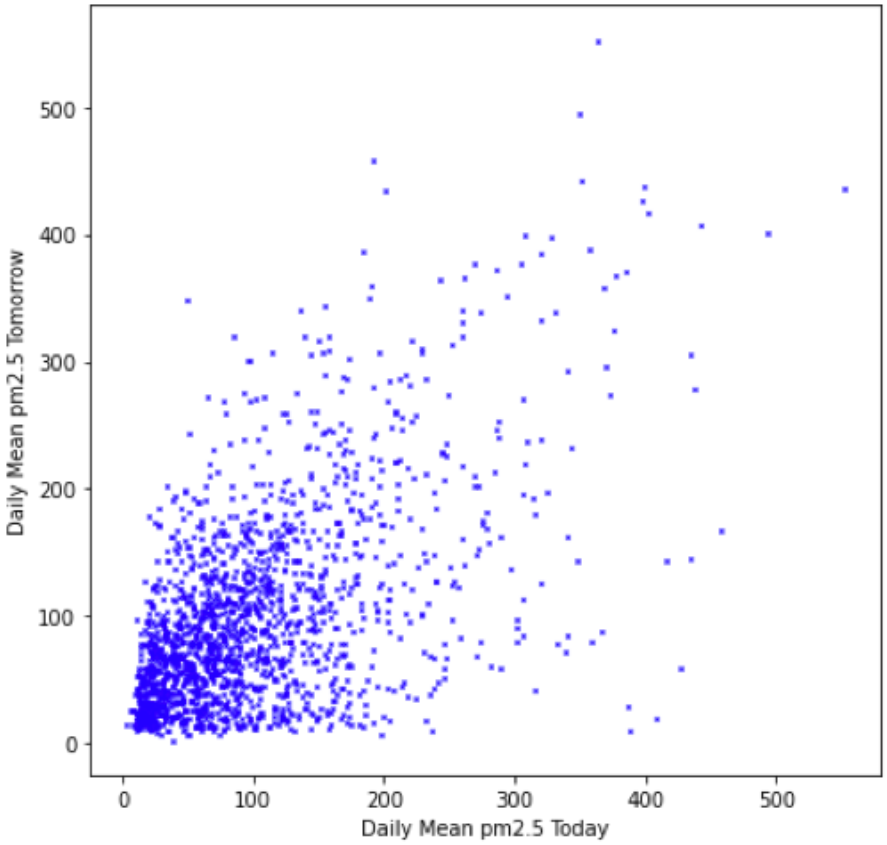
- Values of some time series predict another time series

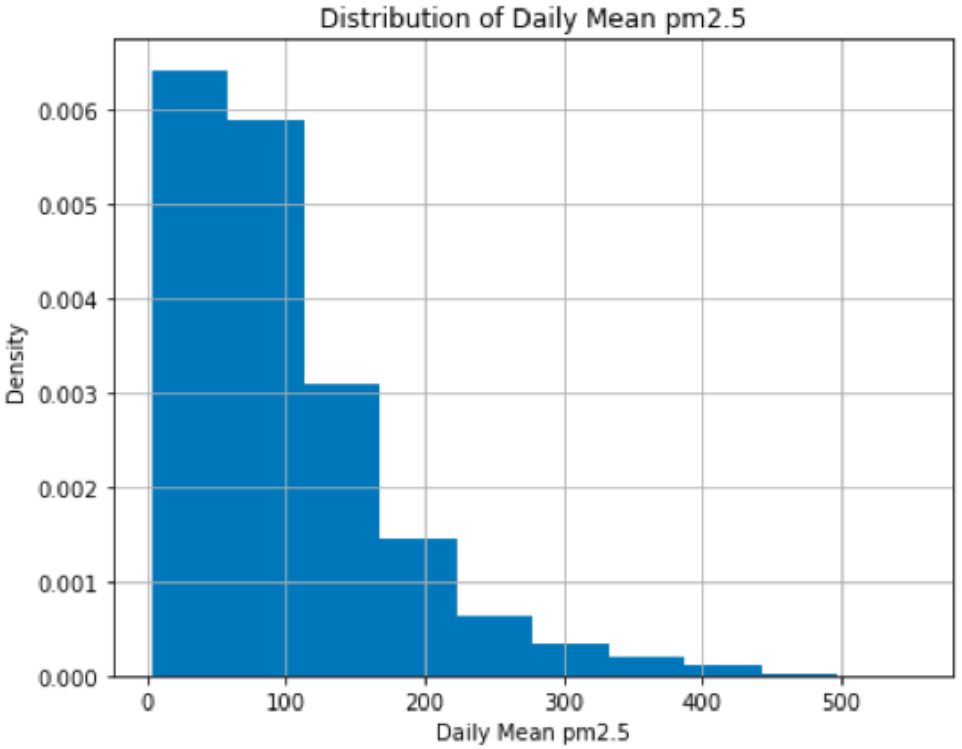
$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \epsilon$$

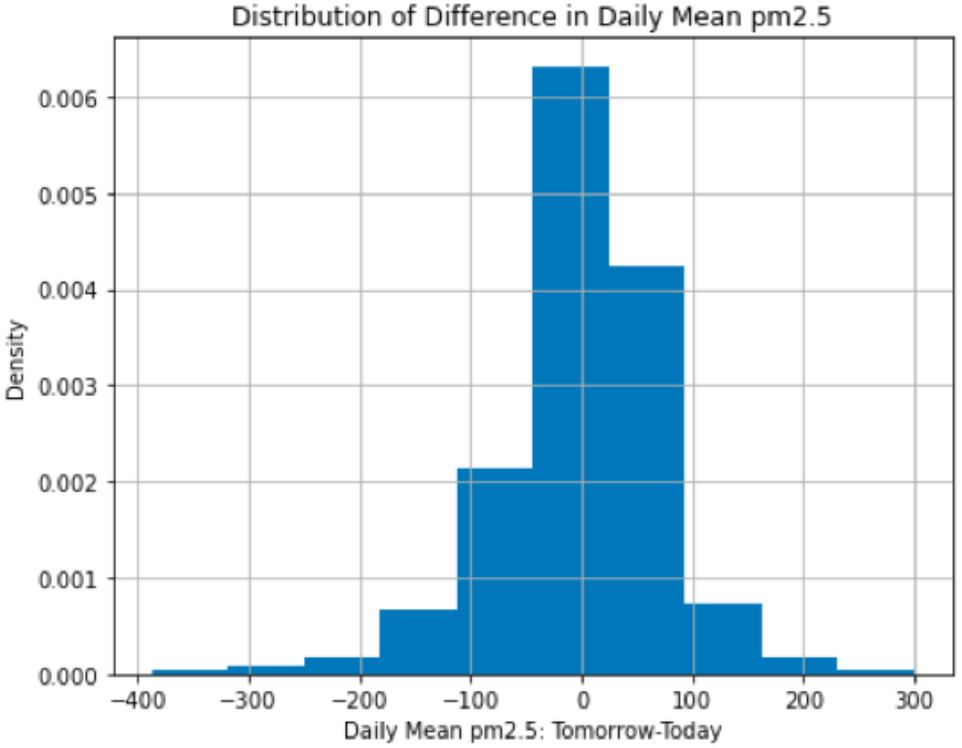
- Without lag, not a prediction
- Used to understand causal patterns

Example: pm2.5 Tomorrow

- 24 hours means
- Can we predict mean tomorrow?







Example Regression

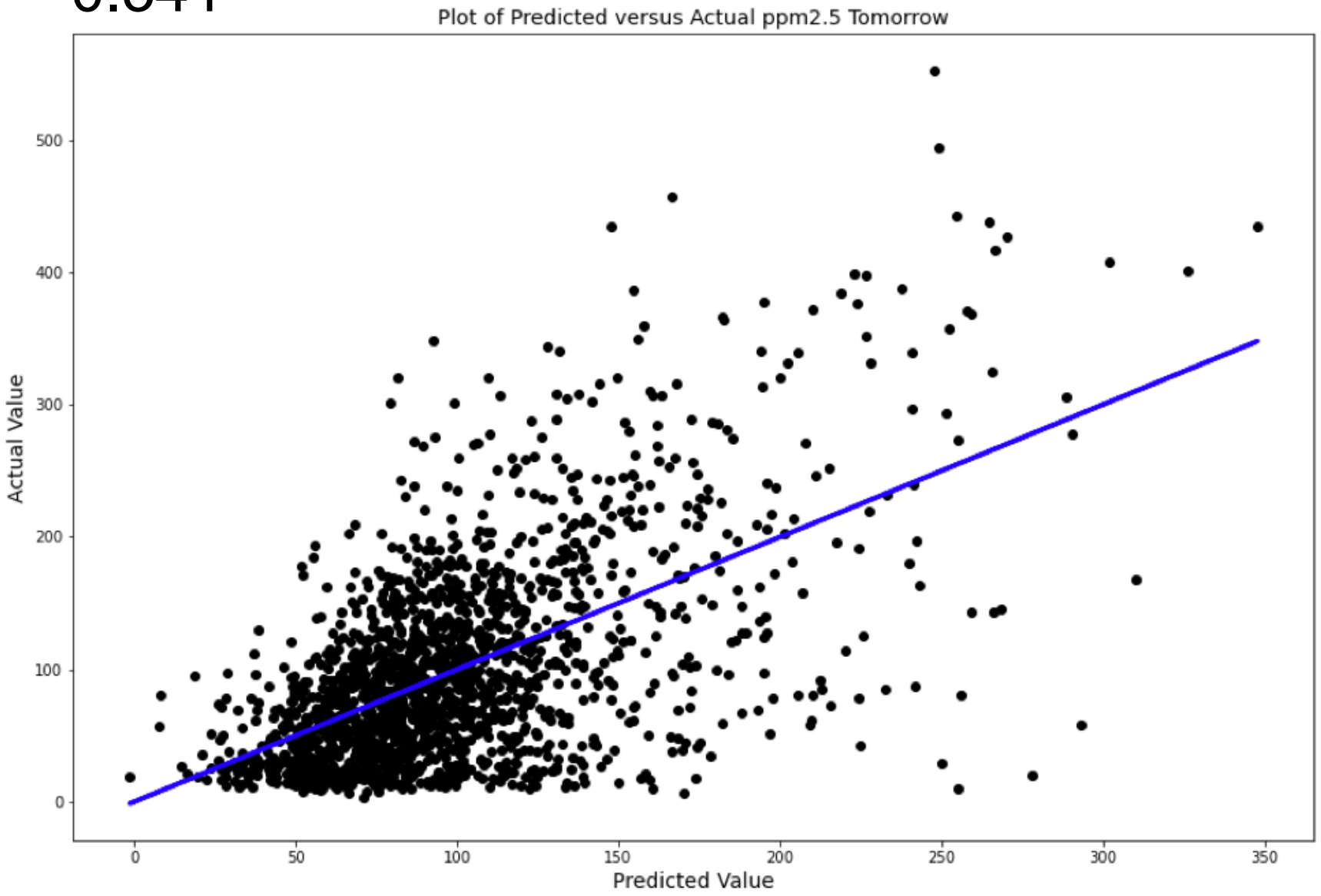
```
X = np.column_stack(
        (bdaily.pm_av_tdy,
         bdaily.meanTemp,
         bdaily.meanPres,
         bdaily.meanIws))
y = bdaily.pm_av_tmw
reg = LinearRegression().fit(X,y)
```

pm2.5 today
mean temp
mean pressure
mean wind speed

- What causes high pm2.5?
- Could try lags

Example Regression

- R2 = 0.341

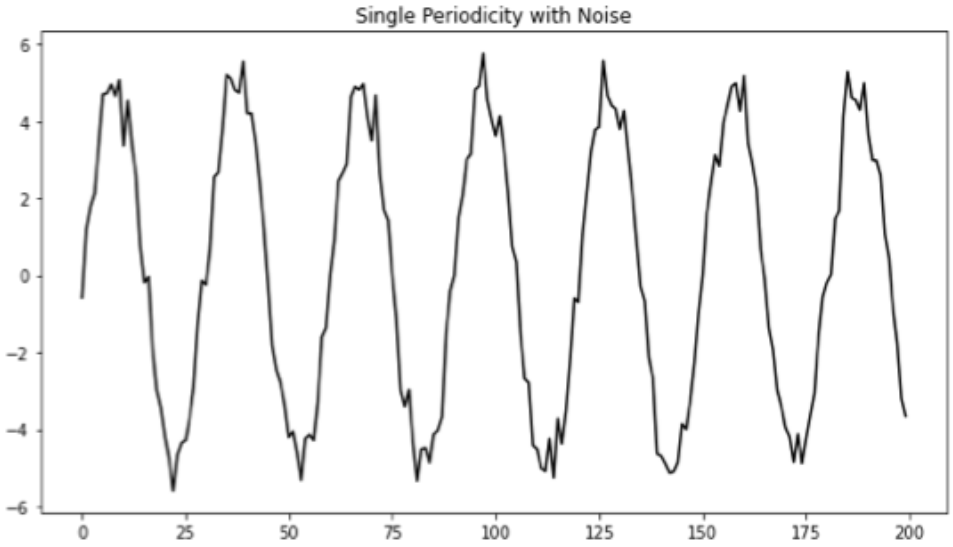


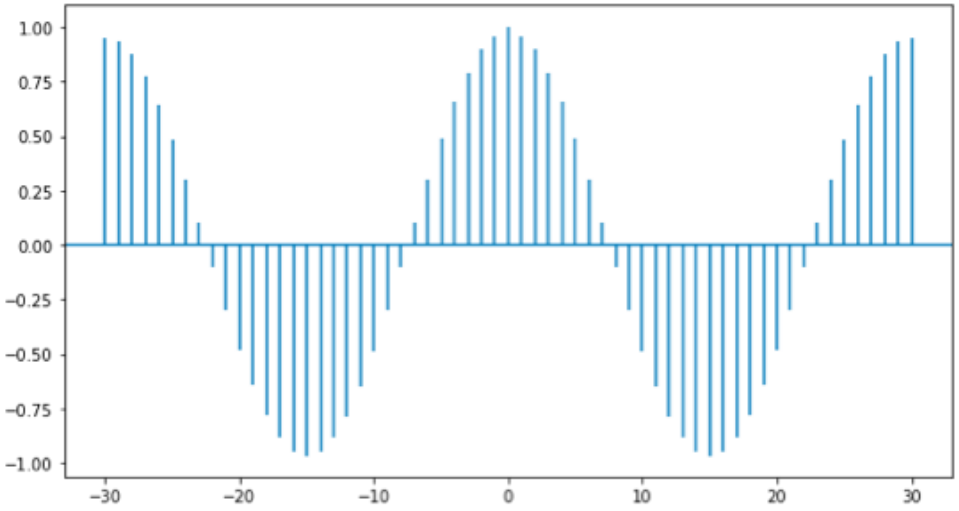
Auto Correlation

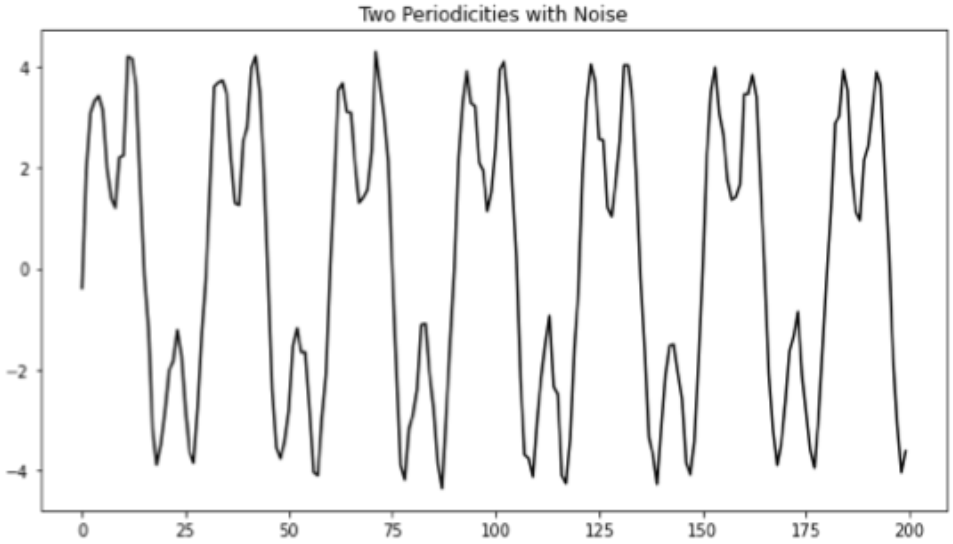
Auto Correlation

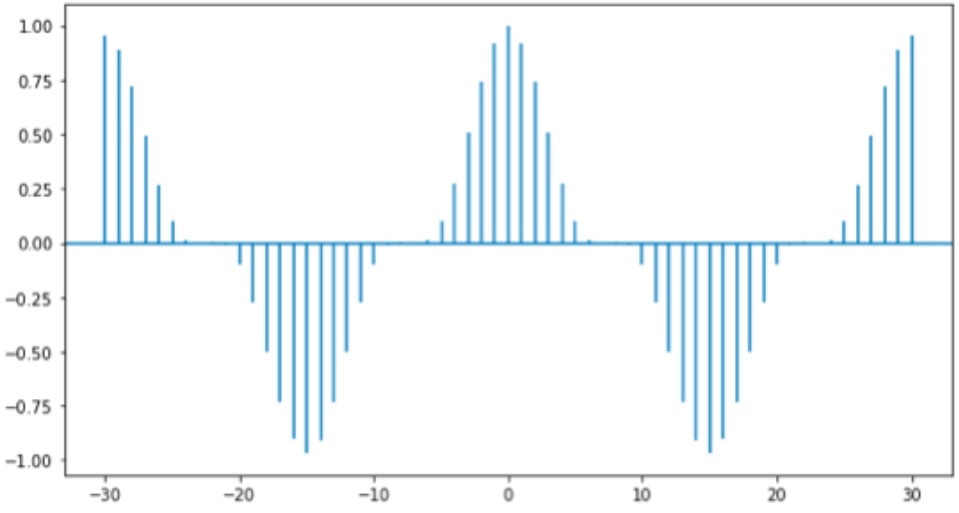
- Idea that earlier value predicts later value
- Calculate
  - Corr(y_t, y_{t-k}) for different lag values k = 1, 2, 3
  - One way to uncover periods

Auto Correlation Example









Summary

- Time series data is indexed by a date or time
- Use date/time representation with arithmetic
  - Instants and intervals
- Sources of variation
  - Random
  - Periodic changes
  - Trend
- Basic operations
  - Moving average
  - Resampling
- Varieties of regression