Time Series Data

Ivan Corneillet

Data Scientist

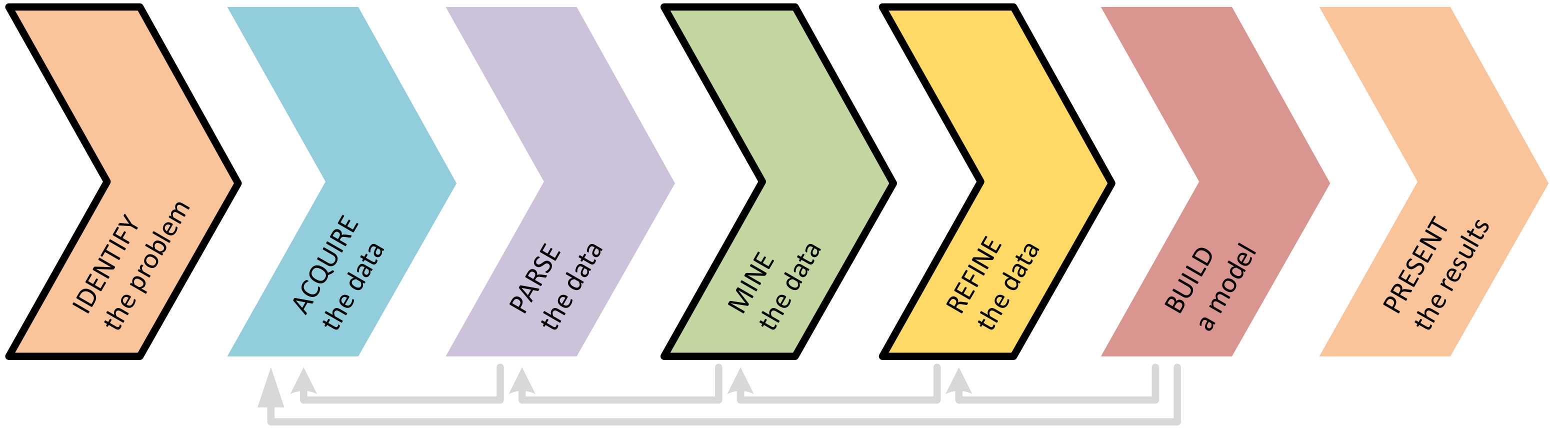
GENERAL ASSEMBLY

Final Project Countdown (you can see the light at the end of the tunnel…)

Final Project, Part 3	April 19; due next session
Final Project, Part 4	April 26; due in 1.5 weeks
Final Project, Part 5	April 28; due in 2 weeks

Today, we will focus on Identifying problems related to time series and discuss the unique aspects of Mining and Refining time series data

Unit 1 – Research Design and Data Analysis	*Research Design*	*Data Visualization in Pandas*	*Statistics*	*Exploratory Data Analysis in Pandas*
Unit 2 – Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Unit 3 – Data Science in the Real World	*Decision Trees and Random Forests*	*Time Series Data*	*Natural Language Processing*	*Databases*



Learning Objectives

After this lesson, you should be able to:

- Understand what time series data is and what is unique about it
- Perform time series analysis in *pandas* including rolling mean/median and autocorrelation

Outline

- Review
- Time Series Analysis
  - Codealong – Data Exploration
- Seasonality, Trends, and Cycles
  - Codealong – Seasonality, Trends, and Cycles
- Moving Averages; Rolling Mean and Median
  - Codealong – Rolling Averages; pandas Window and Expending Functions
- Weighted Moving Averages
- Autocorrelation
  - Codealong – Autocorrelation
- Office hours in class for final projects
- Review

Review

Review

- Latent variable models attempt to uncover structure from text
- Dimensionality reduction is focused on replacing correlated columns
- Topic modeling (or LDA) uncovers the topics that are most common to each document and then the words most common to those topics
- Word2Vec builds a representation of a word from the way it was used originally
- Both techniques avoid learning grammar rules and instead rely on large datasets. They learn based on how the words are used, making them very flexible

Pre-Work

Pre-Work

Before the next lesson, you should already be able to:

- Load data with *pandas*
- Plotting data with *seaborn*
- Understand correlation

Time Series Analysis

Time Series Analysis

- In most of our previous examples, we assumed that the data was not changing over time and we didn't care which data points were collected earlier or later than others

- In this class, we will discuss analyzing data that is changing over time (e.g., S&P 500) with a focus on statistics around data that is changing over time and how to measure that change

A time series is an ordered sequence of values of a variable at equally spaced time intervals

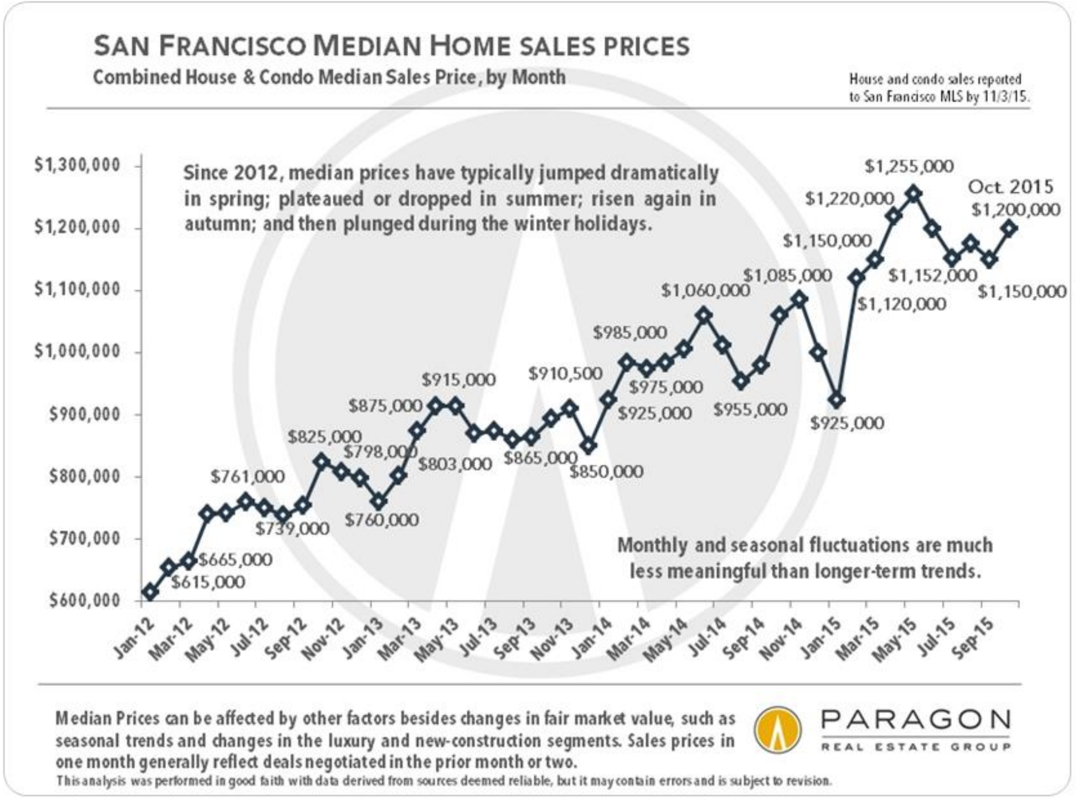
- The usage of time series models is twofold:
  - Understand the underlying forces and structure that produced the observed data
  - Fit a model and proceed to forecasting, monitoring, or even feedback and feedforward control

Time Series Analysis is used for many applications

- E.g.,
  - Stock Market Analysis
  - Sales Forecasting
  - Yield Projections

Time Series (cont.)

- Most datasets are likely to have a time component. E.g., if we were analyzing real estate prices, it is clear that prices shift over time and vary with economic periods
- But typically we assume the time component is fairly minimal. E.g., if we are examining real estate prices within a small time period, the time effect on prices is much smaller than other factors, like number of bedrooms and bathrooms

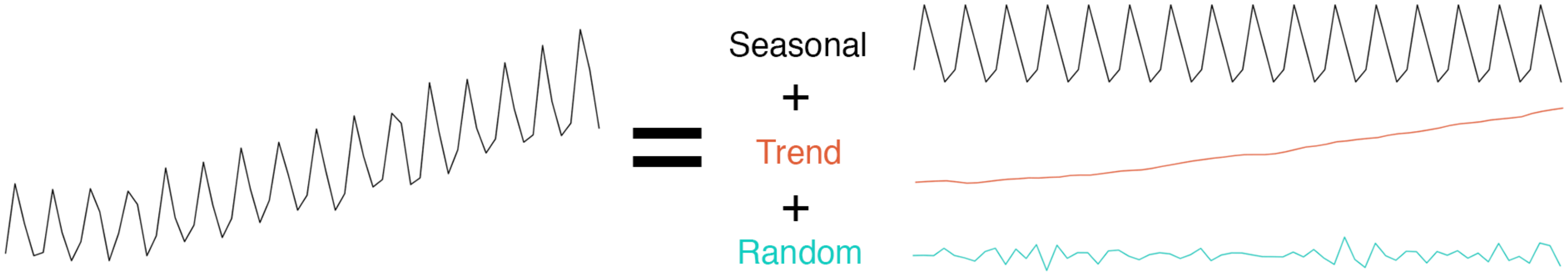


Codealong – Part A
Data Exploration

Seasonality, Trends, and Cycles

Typically, we are interested in separating the effects of time into two components

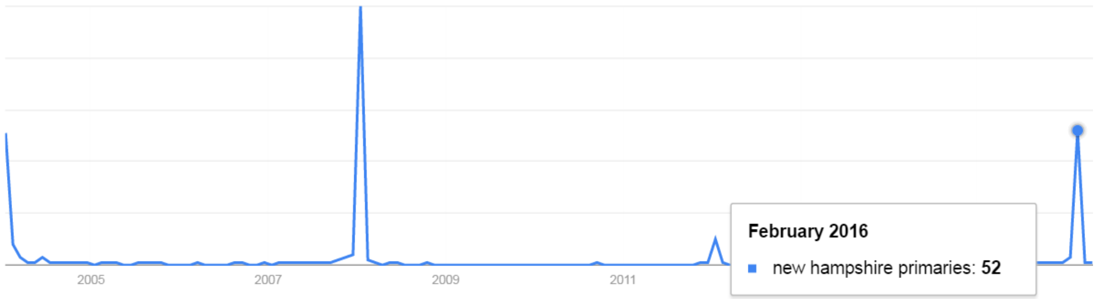
- Seasonality
  - A seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week). Seasonality is always of a fixed and known period
- Trends
  - A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear

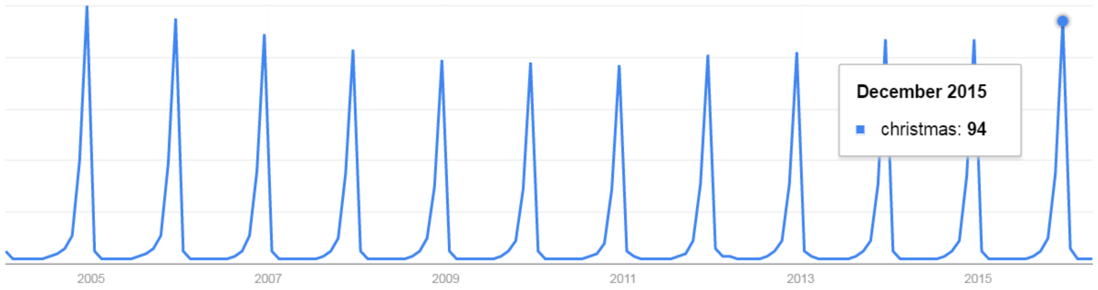


Source: anomaly.io

Seasonality is always of a fixed and known period

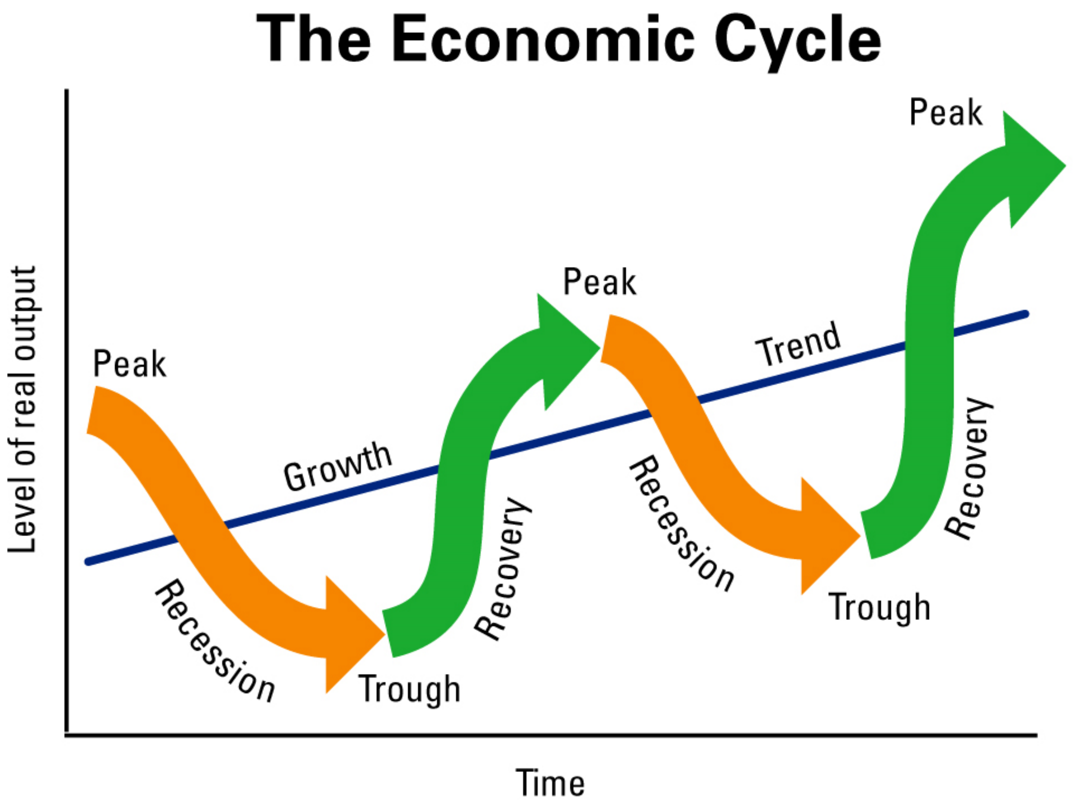
- Searches for "New Hampshire Primary" has a clear seasonal component – It peaks every four years and on election years
- Similarly, searches for "Christmas" spike every year around the holiday season.
- These spikes recur on a fixed time-scale, making them seasonal patterns





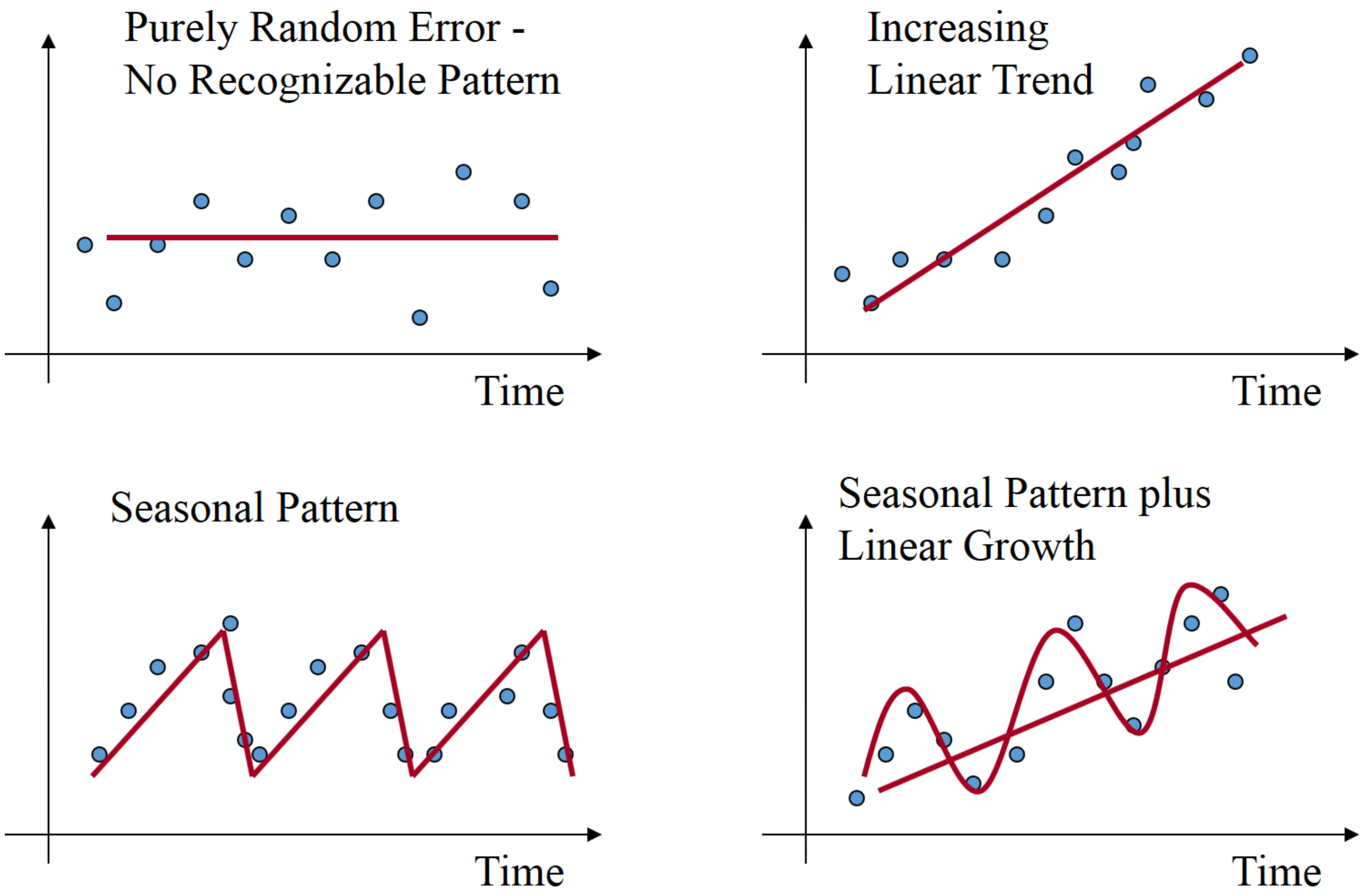
A cyclic pattern exists when data exhibit rises and falls that are *not of fixed period*

- Many people confuse cyclic behavior with seasonal behavior, but they are really quite different
- If the fluctuations are not of fixed period then they are cyclic; if the period is unchanging and associated with some aspect of the calendar, then the pattern is seasonal
- In general, the average length of cycles is longer than the length of a seasonal pattern, and the magnitude of cycles tends to be more variable than the magnitude of seasonal patterns

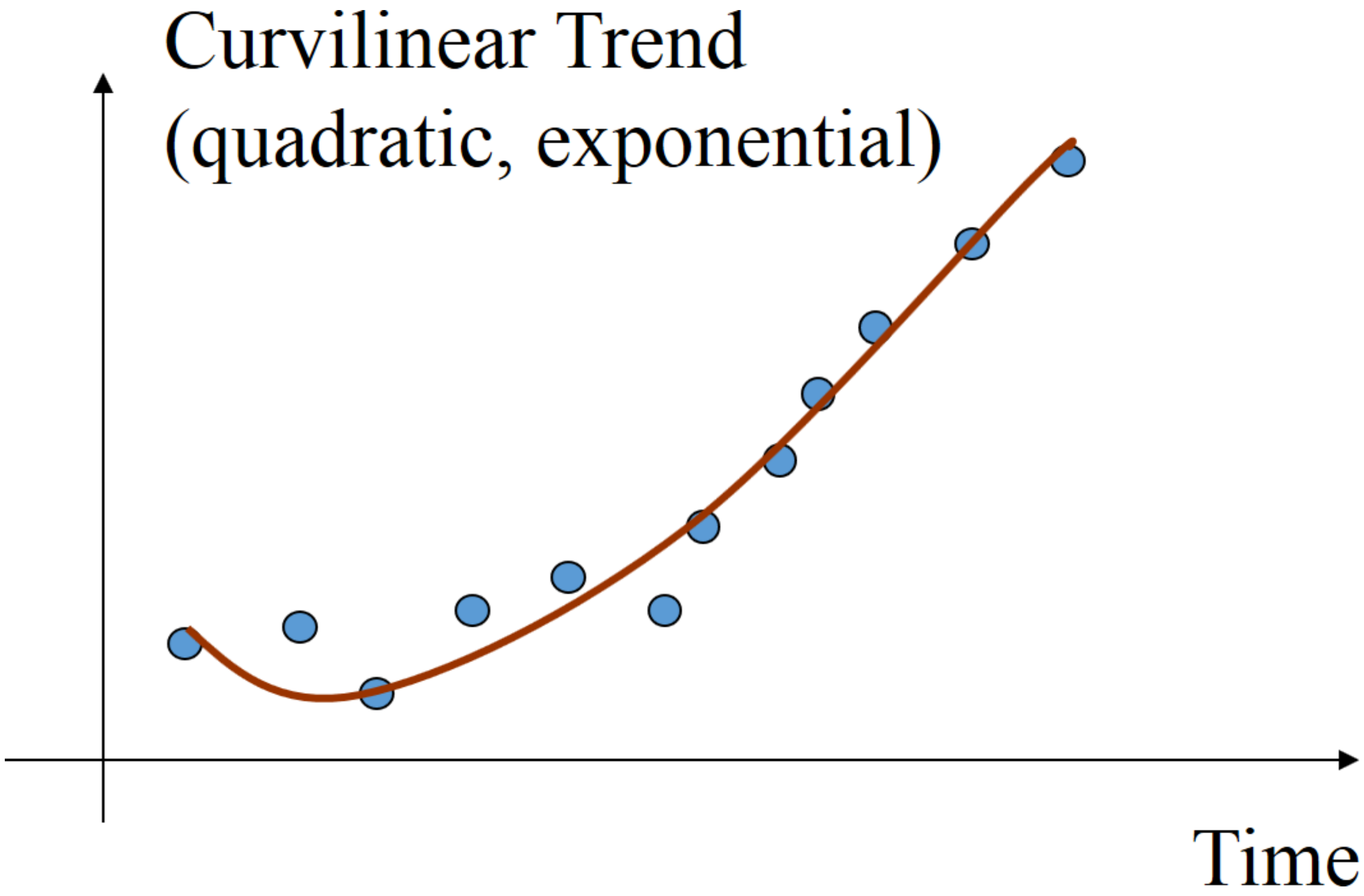


Source: zerohedge.com

Common Time Series Patterns



Common Time Series Patterns (cont.)



Activity: Trends, Seasonality, and Cycles

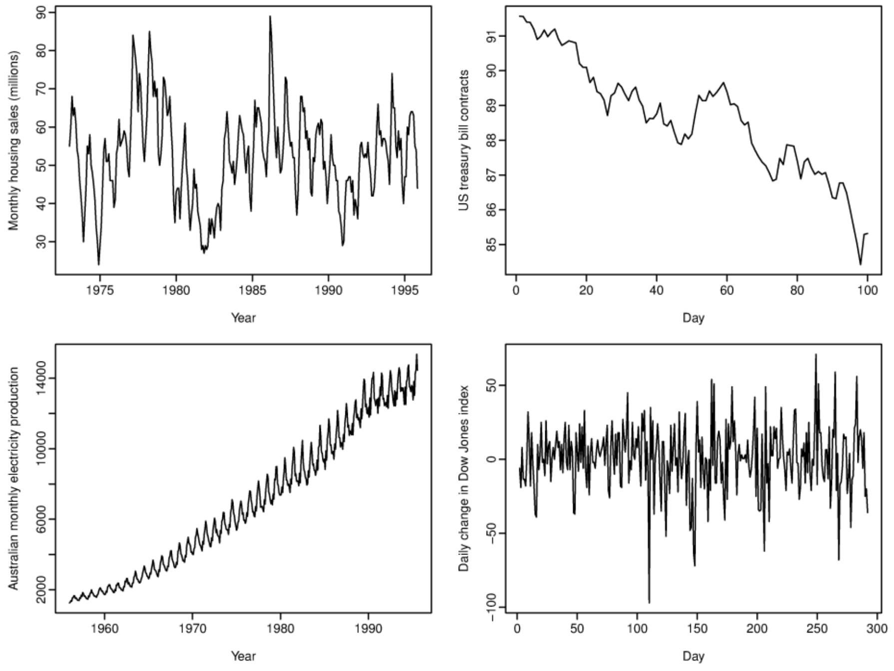
EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. The four time series on the right exhibit different types of time series patterns (trends, seasonality, and cycles). Which time series have what patterns?
2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions



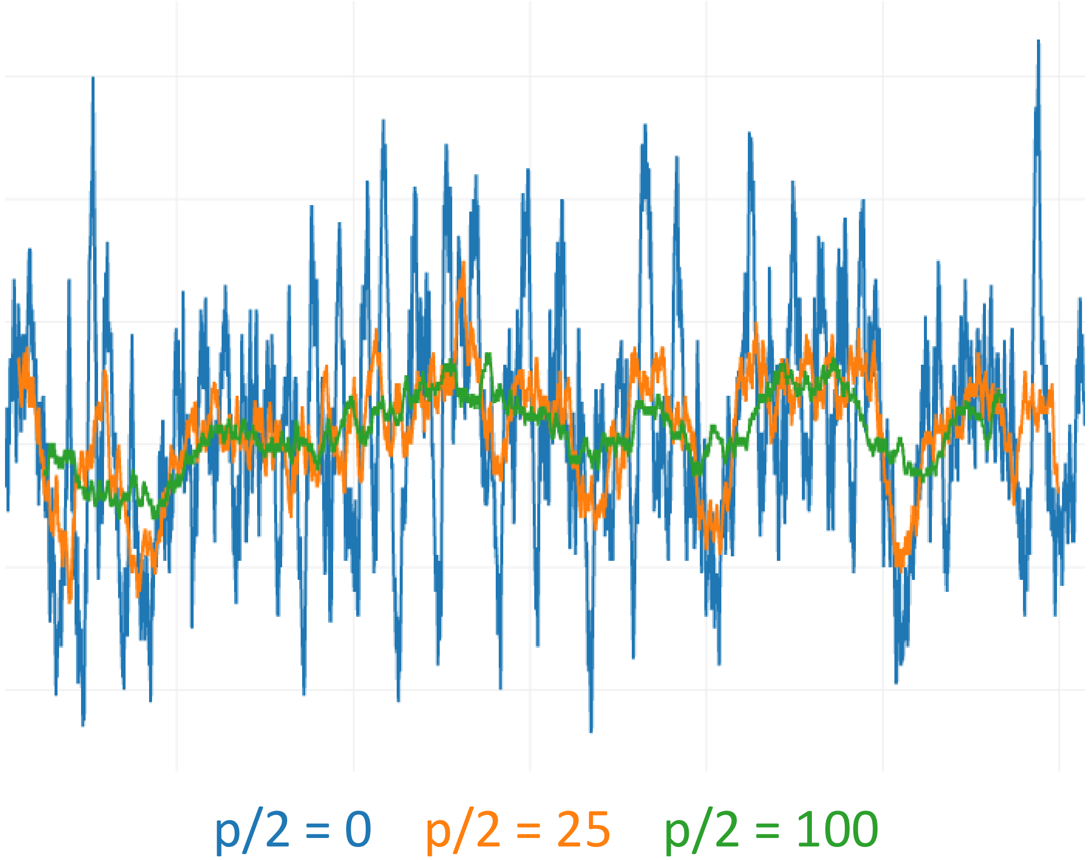
Codealong – Part B Seasonality, Trends, and Cycles

Moving Averages; Rolling Mean and Median

A moving average replaces each data point with an average of k consecutive data points in time

- This could be using the $p/2$ data points prior to and following a given time point; it could also be the p preceding points
- These are often referred to as the "rolling" average
- The measure of average could be mean or median
- The *rolling mean* is

$$F_t = \frac{1}{p}\sum_{k=-\frac{p}{2}}^{\frac{p}{2}} Y_{t+k} \;\; or \; F_t = \frac{1}{p}\sum_{k=0}^{p} Y_{t+k}$$



Rolling means and median (cont.)

Rolling mean

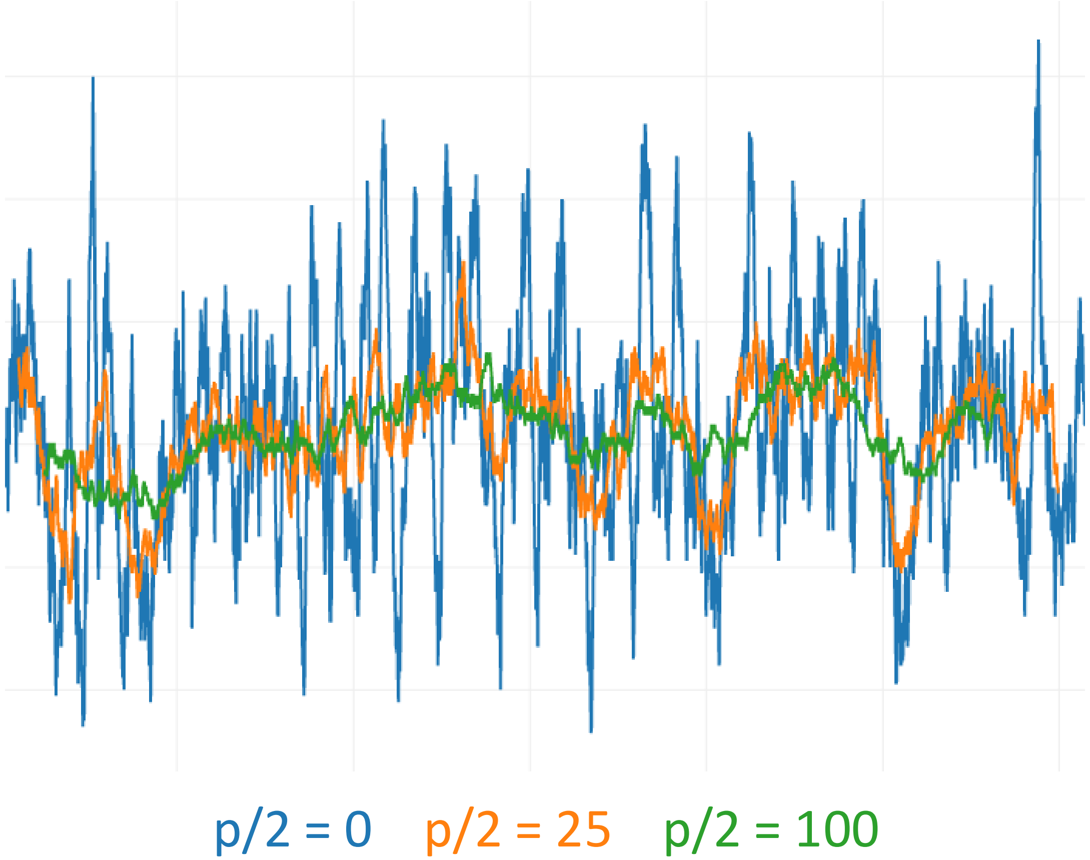
- A rolling mean averages all values in its window, but can be skewed by outliers
  - This may be useful if we are looking to identify atypical periods or we want to evaluate these odd periods
  - E.g., this would be useful if we are trying to identify particularly successful or unsuccessful sales days

Rolling median

- The rolling median would provide the 50 percentile value for the period and would possibly be more representative of a “typical” day

Rolling means and median (cont.)

- Plotting the moving average allows us to more easily visualize trends by smoothing out random fluctuations and outliers

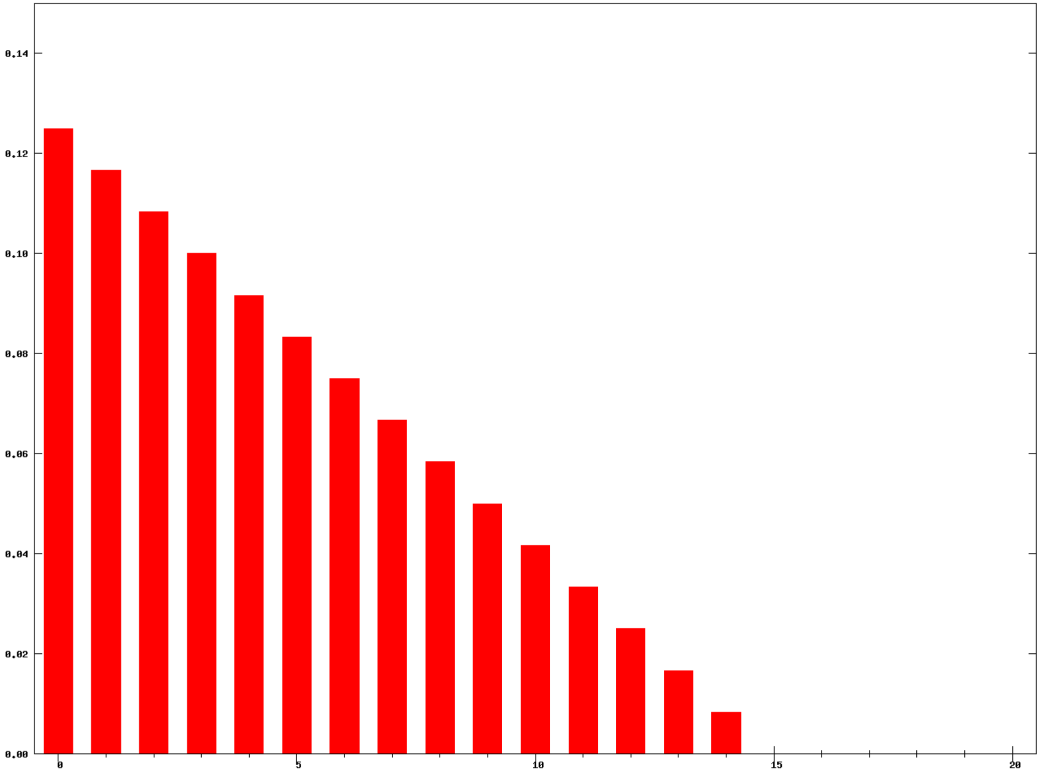


Codealong – Part C
Rolling Averages; *pandas* Window and Expending Functions

Weighted Moving Averages

Weighted Moving Average

- While rolling means and medians weights all data evenly, it may make sense to weight data closer to our date of interest higher
- We do this by taking a *weighted moving average*, where we assign particular weights to certain time points
- Various formulas or schemes can be used to weight the data points



Weights decreasing in arithmetical progression

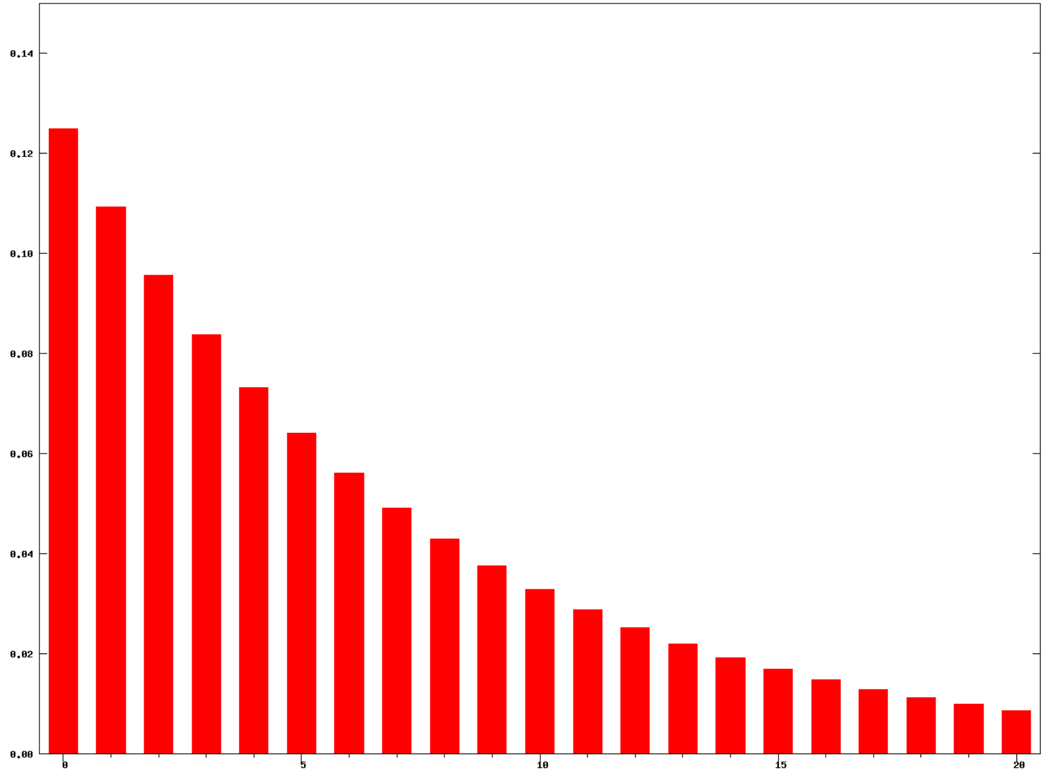
Exponential Weighted Moving Average (EWMA)

- A common weighting scheme is an *exponential weighted moving average (EWMA)* where we add a *decay* term to give lesser and lesser weight to older data points

- The EWMA can be calculated recursively for a series Y

$$EWMA_1 = Y_1 \ for \ t = 1$$

$$EWMA_t = \alpha \cdot Y_t + (1 - \alpha) \cdot EWMA_{t-1} \ for \ t > 1$$



Weights decreasing exponentially

Autocorrelation

Autocorrelation

- In previous classes, we have been concerned with how two variables are correlated (e.g., height and weight, education and salary)
- *Autocorrelation* is how correlated a variable is with itself. Specifically, how related are variables earlier in time with variables later in time
- To compute autocorrelation, we fix a "lag" k denoting how many time points earlier we should use to compute the correlation
- A lag of $k = 1$ computes how correlated a value is with the prior one. A lag of $k = 10$ computes how correlated a value is with one 10 time points earlier

$$r_k = \frac{\sum_{i=1}^{N-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

with N observations and $\bar{x}$ the overall mean

Codealong – Part D
Autocorrelation

Review

Review

- We use time series analysis to identify changes in values over time
- We want to identify whether changes are true trends or seasonal changes
- Rolling means give us a local statistic of an average in time, smoothing out random fluctuations and removing outliers
- Autocorrelations are a measure of how much a data point is dependent on previous data points

Pre-Work

Pre-Work

Before the next lesson, you should already be able to:

- Prior definition and Python functions for moving averages and autocorrelation
- Prior exposure to linear regression with discussion of coefficients and residuals

Q & A

Exit Ticket

Don’t forget to fill out your exit ticket here