GENERAL ASSEMBLY

INTRODUCTION TO TIME SERIES

Alex Cave

LEARNING OBJECTIVES

- Understand time series data is and how it differs from other data types
- Understand the concept of autocorrelation
- Perform time series analysis in Pandas including rolling mean/median

COURSE

REVIEW

REVIEW

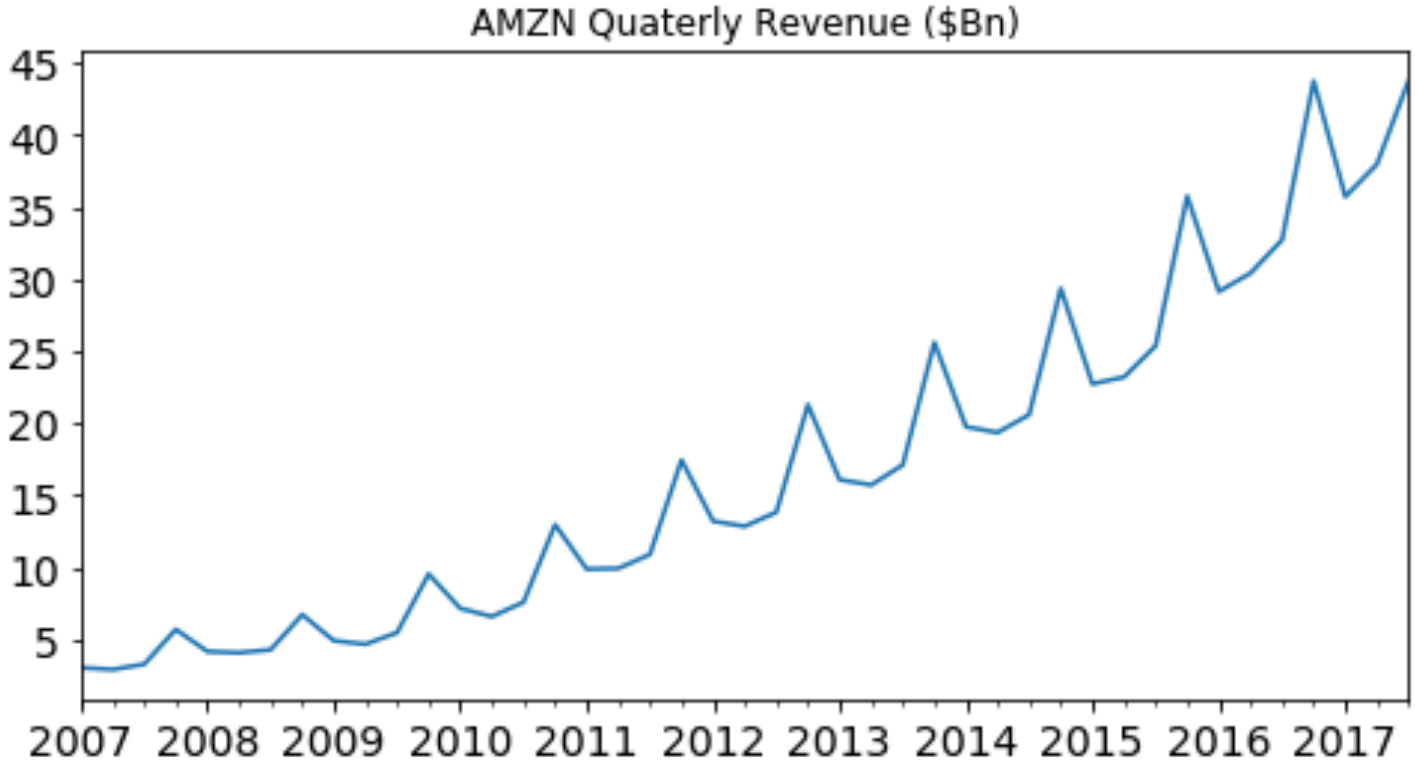
- Up until now, we have only looked into cross-sectional data for building models
- WHY?!?!?!
- Lets examine the differences between cross-sectional data and time series data

EXAMPLE: FORECASTING AMAZON EARNINGS

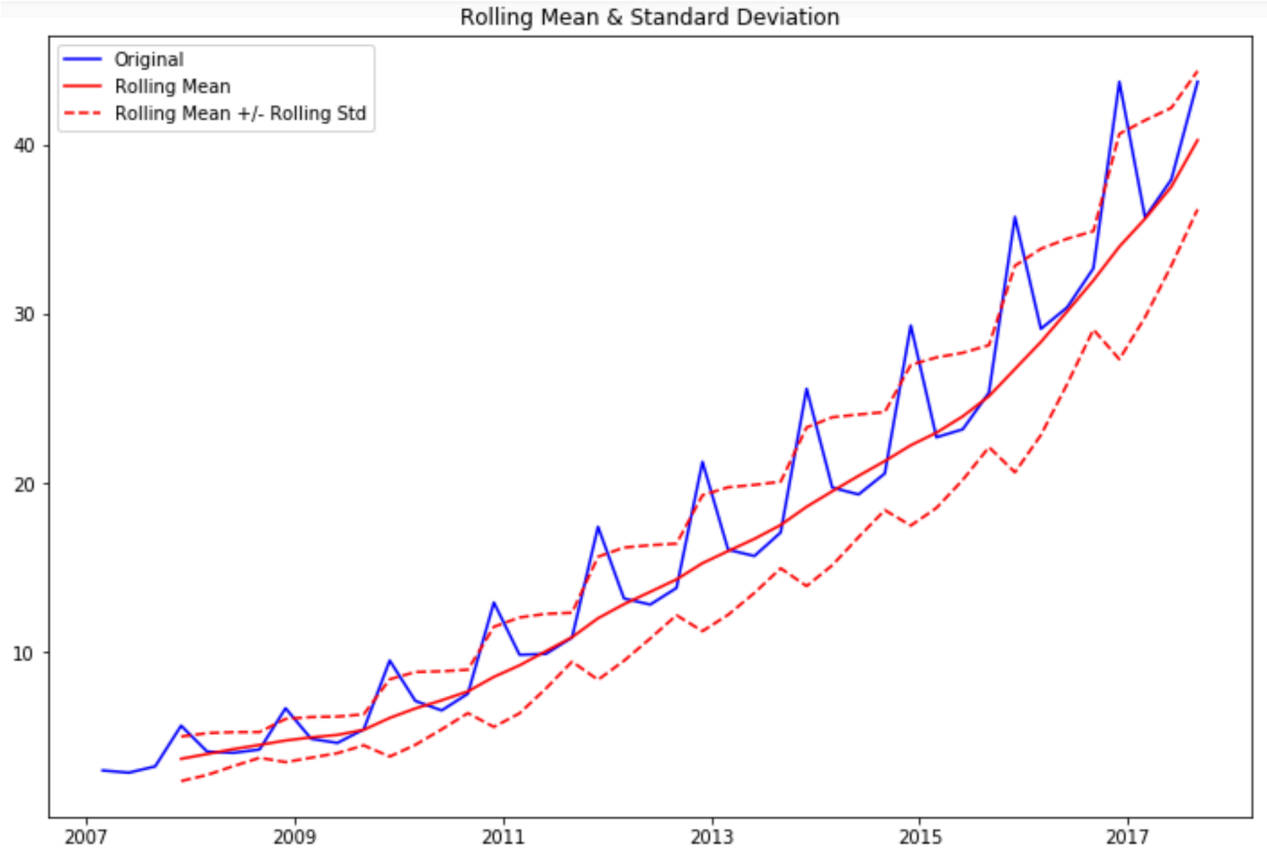
- A hedge fund asked to predict Amazon.com, Inc.'s revenue for fiscal year 2017 (4^{th} Qtr Earnings)

- They expected to see time series analysis, so I gave them what they asked for

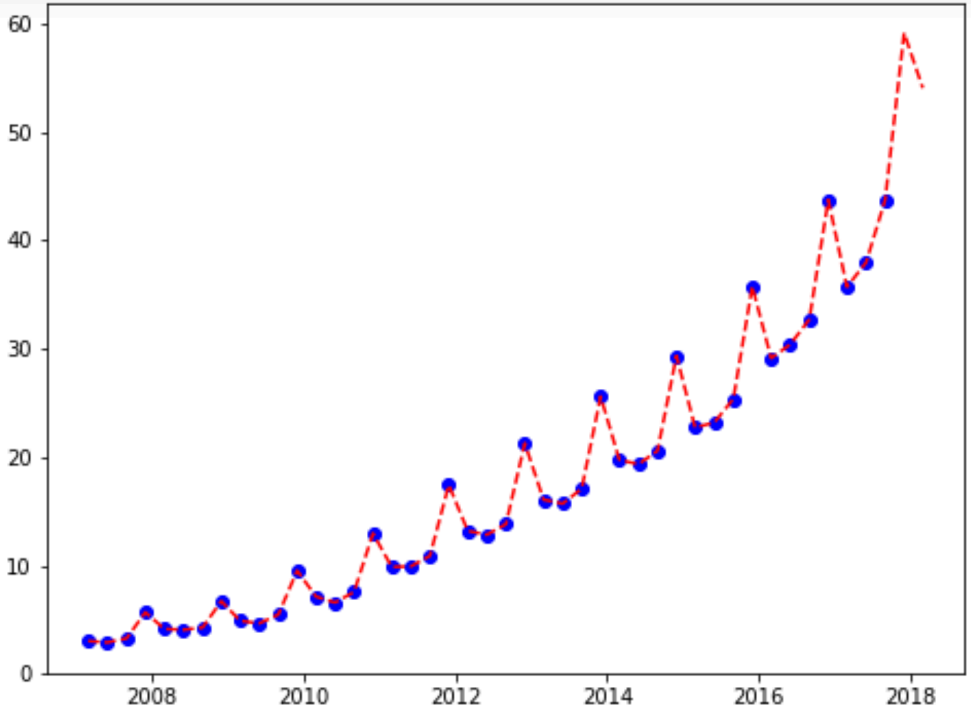
EXAMPLE: FORECASTING AMAZON EARNINGS



EXAMPLE: FORECASTING AMAZON EARNINGS



EXAMPLE: FORECASTING AMAZON EARNINGS




```
In [142]: preds

Out[142]: 2017-12-01    59.149130
          2018-03-01    54.070608
```

EXAMPLE: FORECASTING AMAZON EARNINGS

Revenue Estimate	Current Qtr. (Dec 2017)	Next Qtr. (Mar 2018)	Current Year (2017)	Next Year (2018)
No. of Analysts	37	30	43	43
Avg. Estimate	59.8B	48.65B	177.23B	228.69B
Low Estimate	57.95B	47.27B	175.36B	220.02B
High Estimate	60.63B	50.06B	178.3B	236.56B
Year Ago Sales	43.74B	35.71B	135.99B	177.23B
Sales Growth (year/est)	36.70%	36.20%	30.30%	29.00%

EXAMPLE: FORECASTING AMAZON EARNINGS

Amazon blew past street estimates for its fourth quarter earnings, reflecting strong holiday sales and growth in its cloud business.

Amazon's stock went up more than 6 percent in after hours trading.

Here are the most important numbers:

- **Revenue:** $60.5 billion vs. $59.83 billion, as estimated, according to Thomson Reuters
- **EPS:** $3.75 per share*
- **AWS revenue:** $5.11 billion vs. $4.97 billion, as estimated, according to FactSet

Amazon's revenue, which includes sales from Whole Foods, jumped 38 percent year-over-year. Its North America revenue jumped 42 percent to $37 billion, while international sales grew 29 percent to $18 billion.

REVIEW - CONCLUSION

- Usually, no dataset is an island on its own.
- There are interactions between different datasets which may explain some, or most of the variation in your target.
- Machine learning is focused on identifying relationships between features and providing inferences based on these other datasets.
- Time Series largely focuses on variations that are explained by itself.
- Researchers are paid a lot to forecast in this fashion, and frequently get it wrong (takes a long time to tune parameters)
- Time Series problems can be turned into machine learning problems using feature engineering and adding potentially related features

INTRODUCTION

WHAT IS TIME SERIES DATA?

WHAT IS TIME SERIES DATA?

- Time series data is any data where the individual data points change over time.
- This is fairly common in sales and other business cases where data would likely change according to seasons and trends.
- Time series data is also useful for studying social phenomena. For instance, there is statistically more crime in the summer, which is a seasonal trend.

WHAT IS TIME SERIES DATA?

- Most datasets are likely to have an important time component, but typically we assume that it's fairly minimal.
- For example, if we were analyzing salaries in an industry, it's clear that salaries shift over time and vary with the economic period.
- However, if we are examining the problem on a smaller scale (e.g. 3-5 years), the effect of time on salaries is much smaller than other factors, like industry or position.

WHAT IS TIME SERIES DATA?

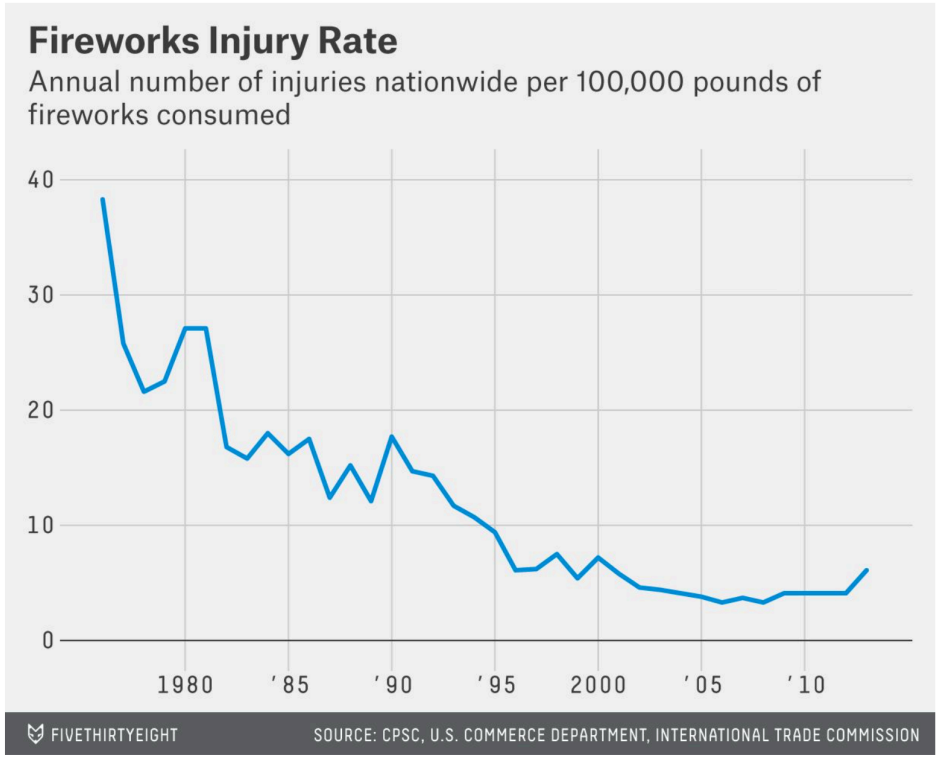
- When the time component *is* important, we need to focus on identifying the aspects of the data that are influenced by time and those that aren't.

- Typically, time series data will be a sequence of values. We will be interested in studying the changes to this series and how related individual values are.

- For example, how much does this week's sales affect next week's? How much does today's stock price affect tomorrow's?

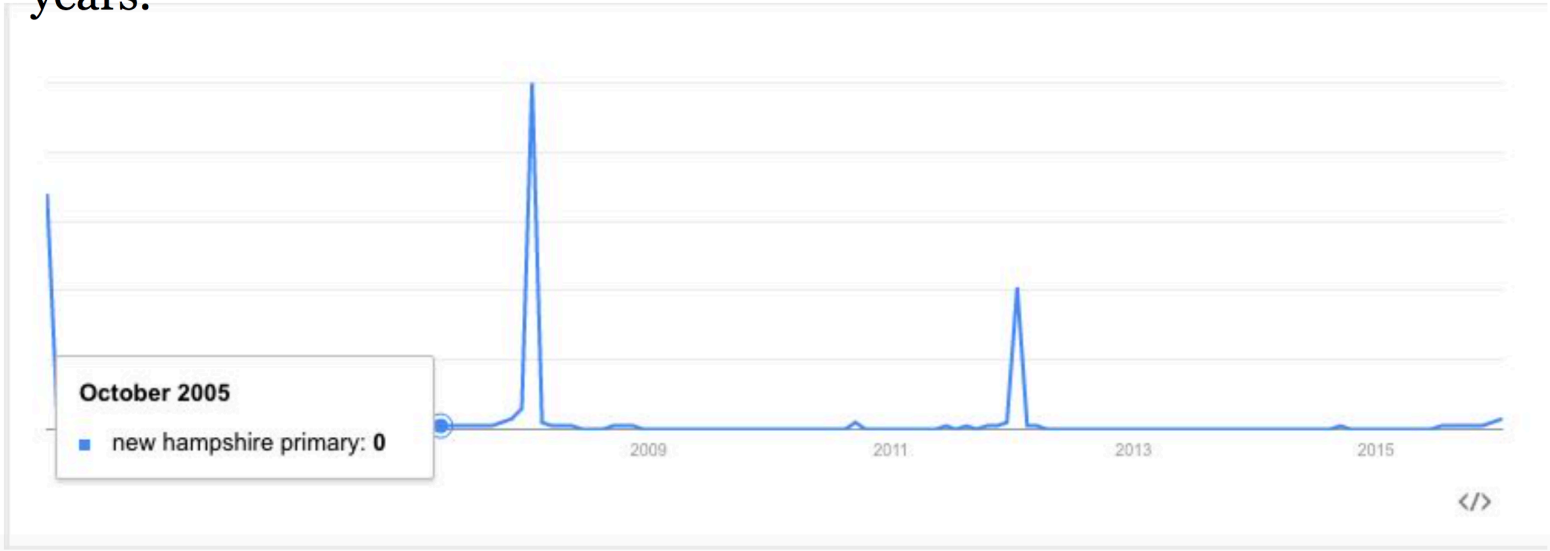
WHAT IS TIME SERIES DATA?

- This plot of fireworks injury rates has an overall *trend* of fewer injuries with no *seasonal* pattern.



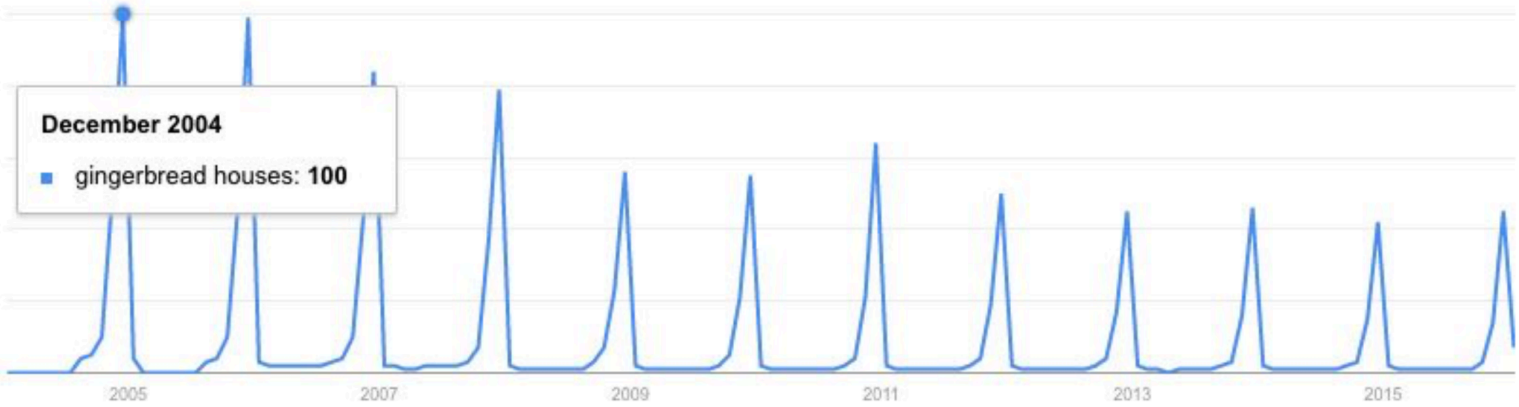
WHAT IS TIME SERIES DATA?

- Meanwhile, the number of searches for the New Hampshire Primary has a clear *seasonal* component - it peaks every four years and on election years.



WHAT IS TIME SERIES DATA?

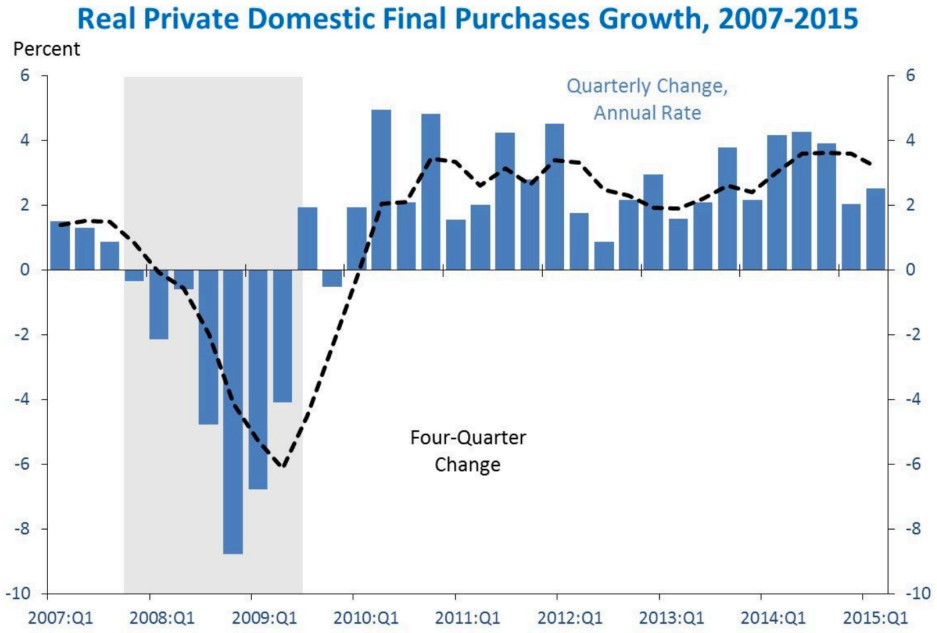
- Similarly, searches for ‘gingerbread houses’ spike every year around the holiday season.



- These spikes recur on a fixed time-scale, making them *seasonal* patterns.

WHAT IS TIME SERIES DATA?

- Many other types of regularly occurring up or down swings may occur without a fixed timescale or *period* (e.g. growth vs. recession for economic trends).



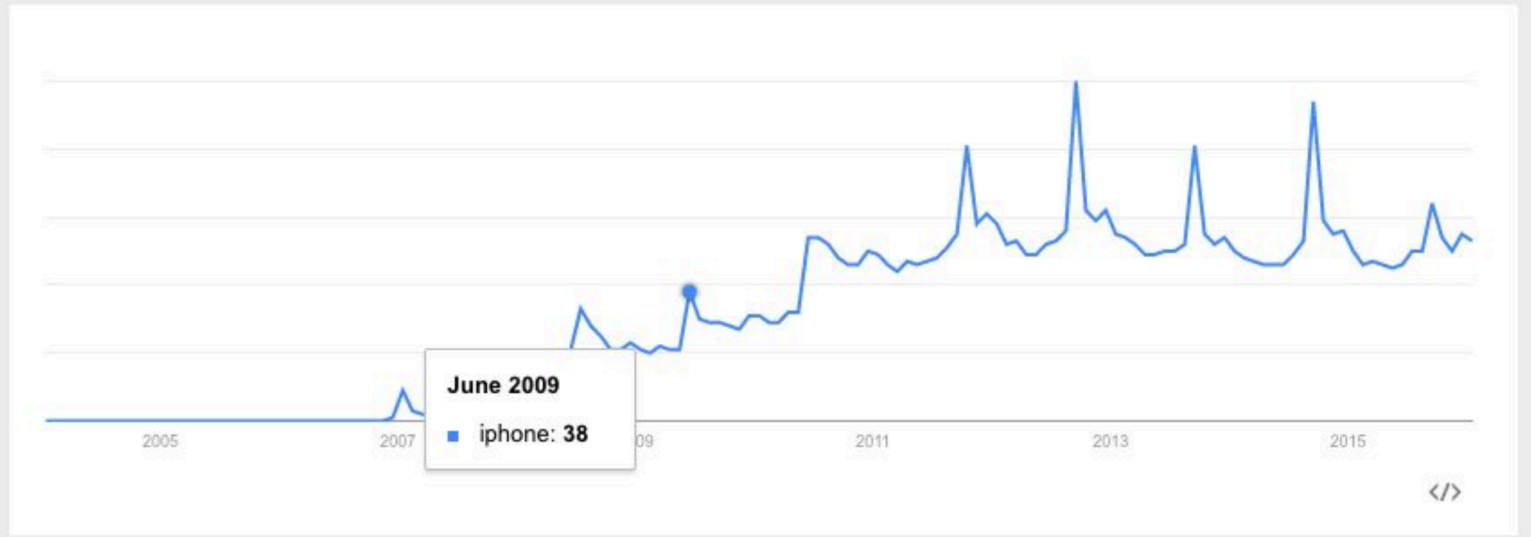
WHAT IS TIME SERIES DATA?

- These aperiodic patterns are called *cycles*.

- While identifying aperiodic cycles is important, they are often treated differently than seasonal effects. Seasonal effects are useful for their consistency, since prior data is useful as a predictor.

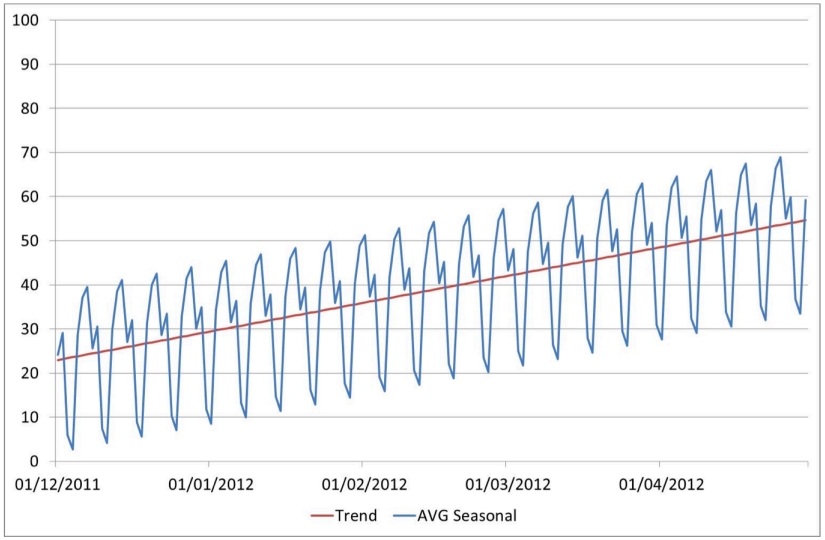
WHAT IS TIME SERIES DATA?

- Searches for “iphone” have both a general trend upwards (indicating more popularity for the phone) as well as a seasonal spike in September (which is when Apple typically announces new versions).



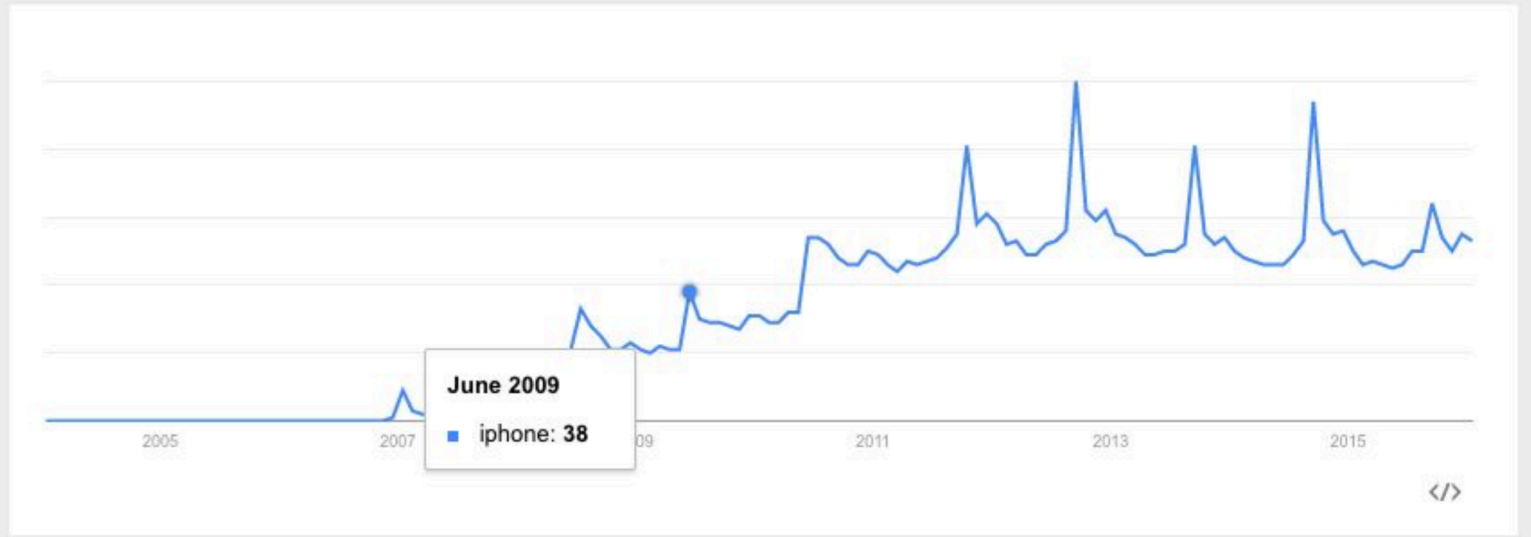
WHAT IS TIME SERIES DATA?

- Most often, we're interested in studying the *trend* and not the *seasonal* fluctuations.
- Therefore it is important to identify whether we think a change is due to an ongoing trend or seasonal change.



WHAT IS TIME SERIES DATA?

- Searches for “iphone” have both a general trend upwards (indicating more popularity for the phone) as well as a seasonal spike in September (which is when Apple typically announces new versions).



INTRODUCTION

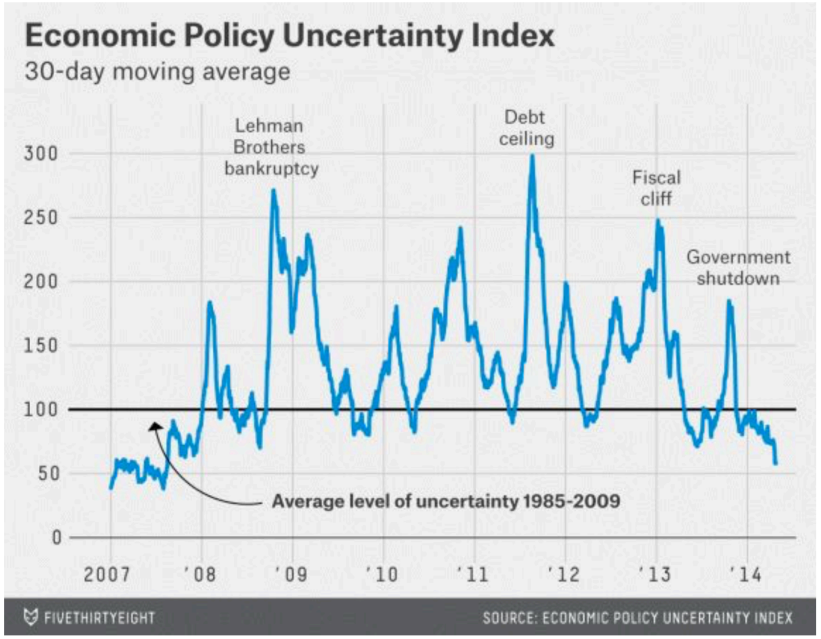
COMMON ANALYSES FOR TIME SERIES DATA

MOVING AVERAGES

- A *moving average* replaces each data point with an average of *k* consecutive data points in time.
- Typically, this is *k/2* data points prior to and following a given time point, but it could also be the *k* preceding points.
- These are often referred to as the "rolling" average.
- The measure of average could be mean or median.
- The formula for the rolling *mean* is $F_t = \frac{1}{p} \sum_{k=t}^{t-p+1} Y_k$

MOVING AVERAGES

- This plot shows the 30-day moving average of the Economic Uncertainty Index.
- Plotting the average allows us to more easily visualize trends by smoothing out random fluctuations and removing outliers.



MOVING AVERAGES

- While this statistic 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.

MOVING AVERAGES

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

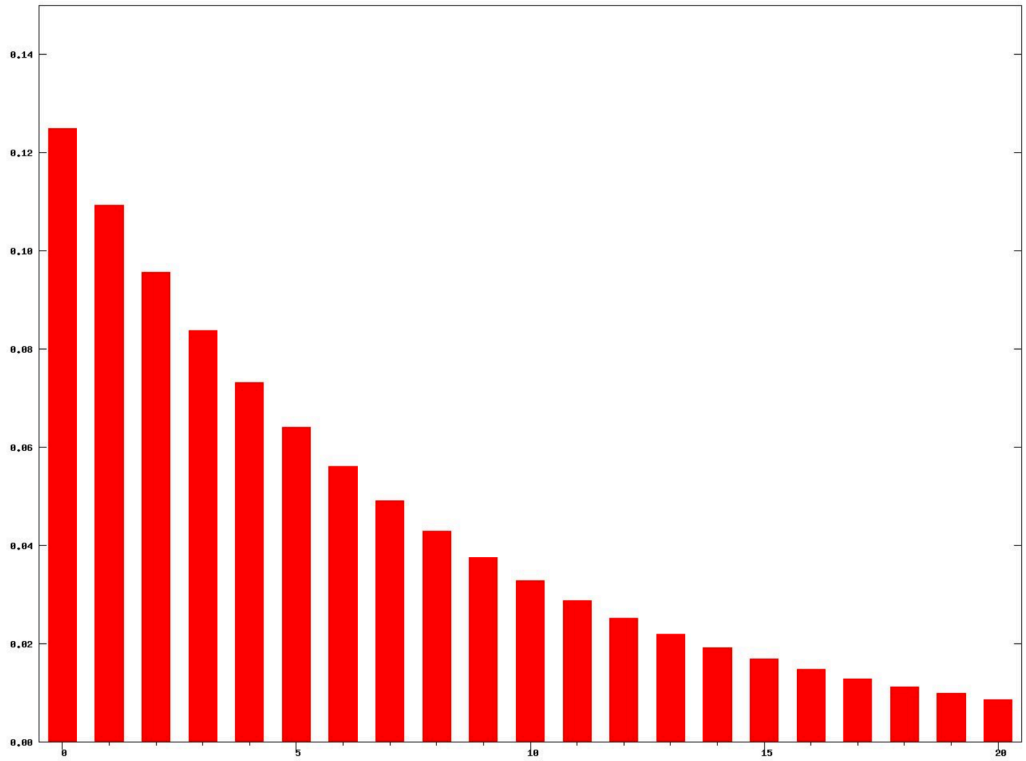
- The EWMA can be calculated recursively for a series Y.

For $t = 1$, $EWMA_1 = Y_1$

For $t > 1$, $EWMA_t = \alpha \bullet Y_t + (1 - \alpha) \bullet EWMA_{t-1}$

MOVING AVERAGES

- The weights for an exponential weighted moving average with k = 15.



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.

AUTOCORRELATION

- To compute autocorrelation, we fix a "lag" k. This is how many time points earlier we should use to compute the correlation.

- A lag of 1 computes how correlated a value is with the prior one. A lag of 10 computes how correlated a value is with one 10 time points earlier.

AUTOCORRELATION

- The following formula can be used to calculate autocorrelation.

$$r_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$$