Time Series Data

Ivan Corneillet

Data Scientist

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

Announcements and Exit Tickets

Q & A

Review

Review

Latent Variables and Natural Language Processing

Today

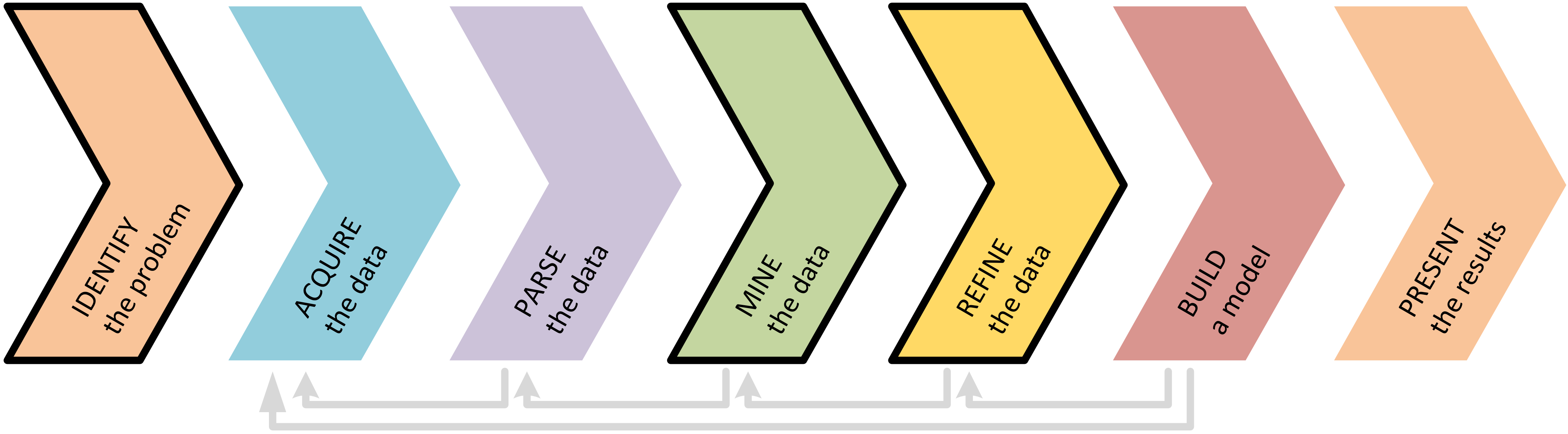
Here's what's happening today:

- Announcements and Exit Tickets
- Review
- Time Series Analysis
  - Codealong – Data Exploration
- Seasonality, Trends, and Cycles
  - Codealong – Seasonality, Trends, and Cycles
- Moving Averages; Rolling Means and Medians
  - Codealong – Rolling Averages; *pandas* Windows and Expending Functions
- Weighted Moving Averages
- Autocorrelation
  - Codealong – Autocorrelation
- Office hours in class for final projects
- Review
- Exit Tickets

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

Research Design and Data Analysis	Research Design	Data Visualization in *pandas*	Statistics	Exploratory Data Analysis in *pandas*
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and Random Forests	$\boxed{\text{Time Series Data}}$	Natural Language Processing	Databases

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



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

Codealong | Part A – Data Exploration

Seasonality, Trends, and Cycles

Seasonality, Trends, and Cycles

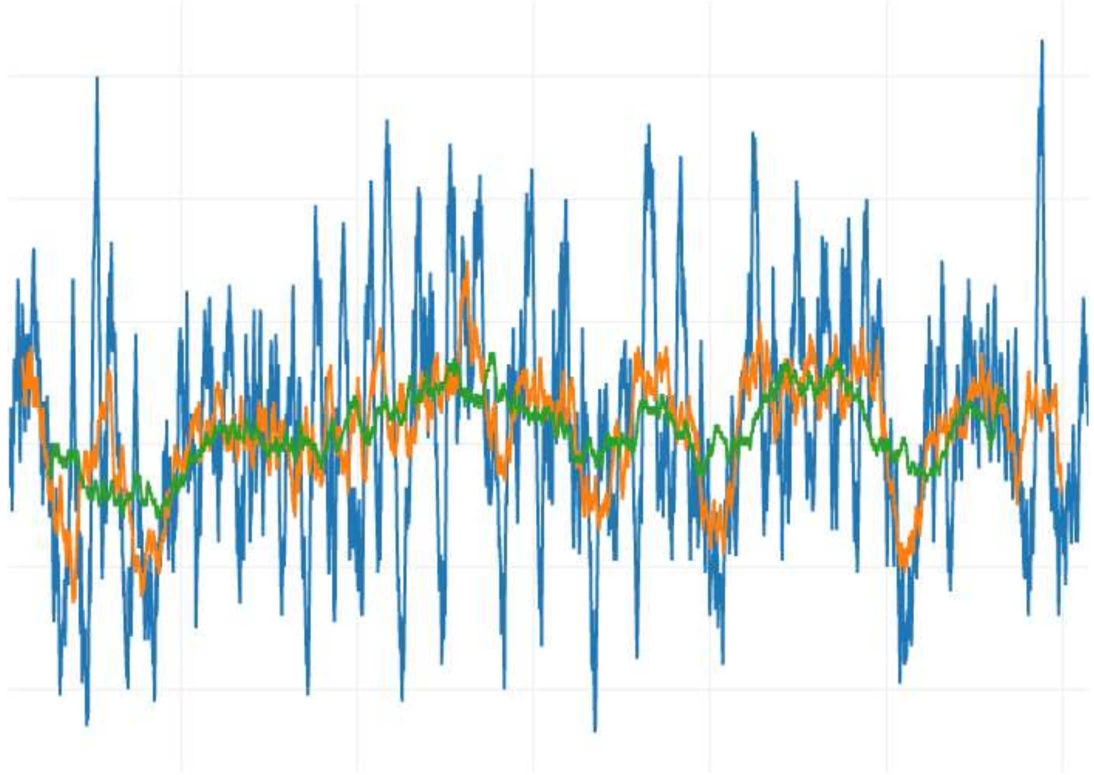
Codealong | Part B – Seasonality, Trends, and Cycles

Moving Averages; Rolling Means and Medians

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

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



p/2 = 0 p/2 = 25 p/2 = 100

Rolling means and median (cont.)

Rolling mean

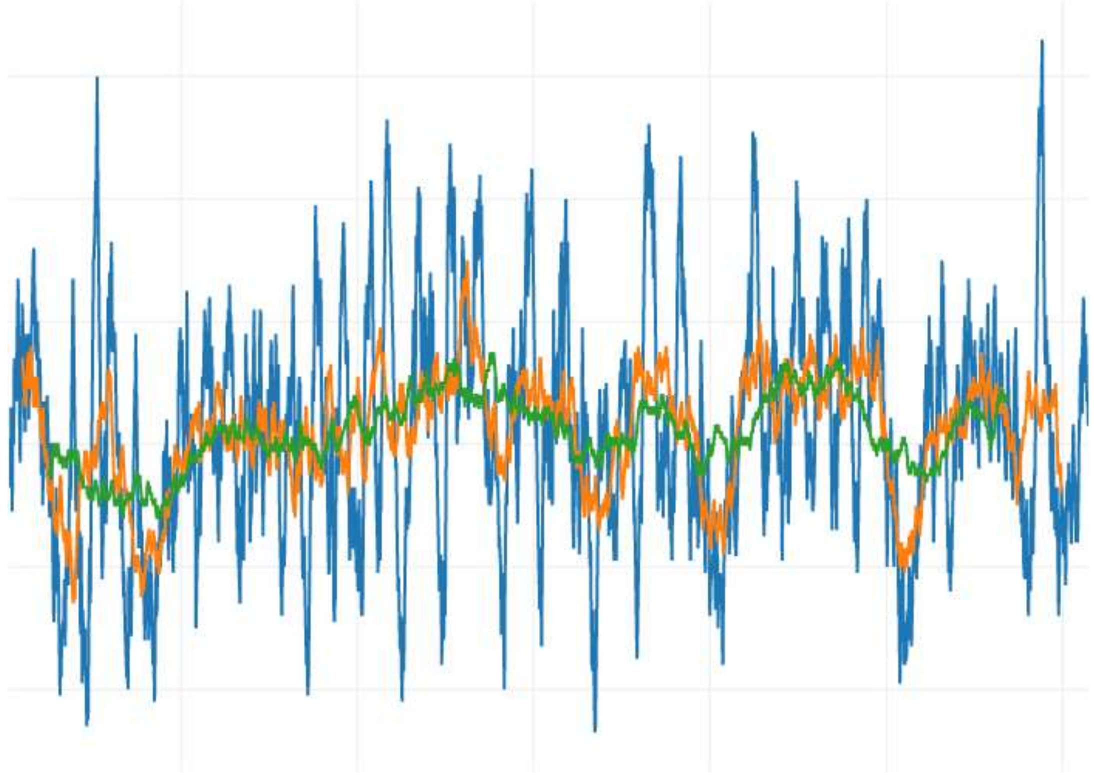
- A rolling mean averages all values in its window, but can be skewed by outliers
  - This may be useful if we are looking to identify atypical periods or we want to evaluate these odd periods
  - E.g., this would be useful if we are trying to identify particularly successful or unsuccessful sales days

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



p/2 = 0 p/2 = 25 p/2 = 100

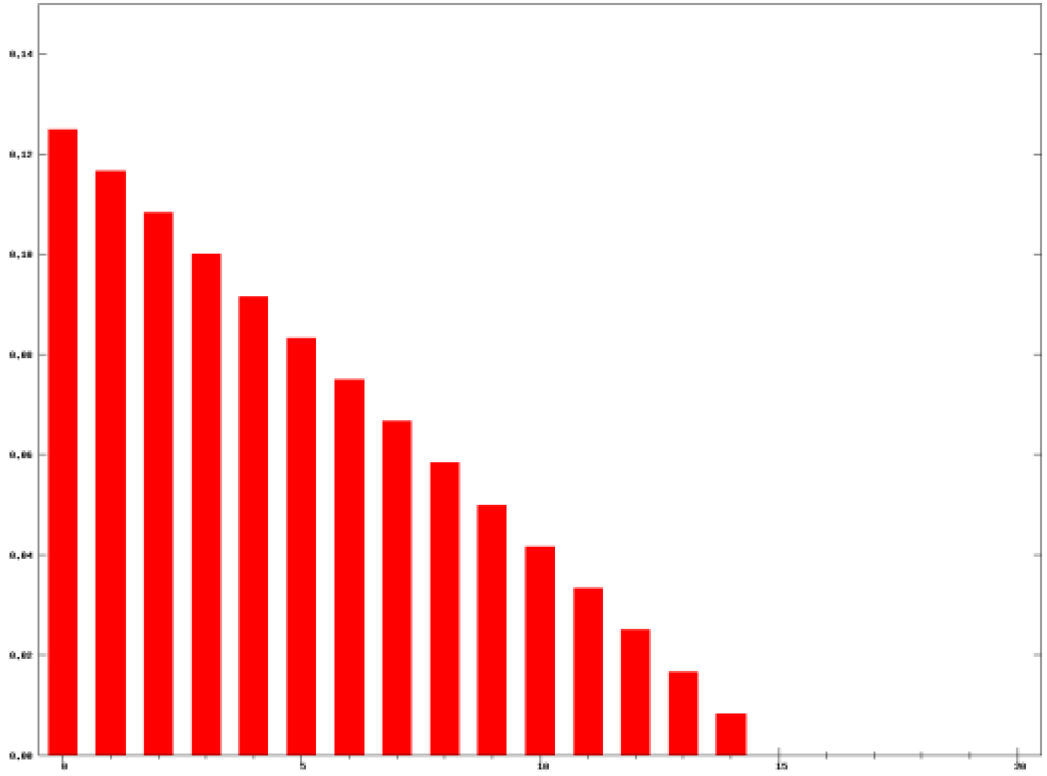
Moving Averages; Rolling Means and Medians

Codealong | Part C – Rolling Averages; pandas Windows and Expanding 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

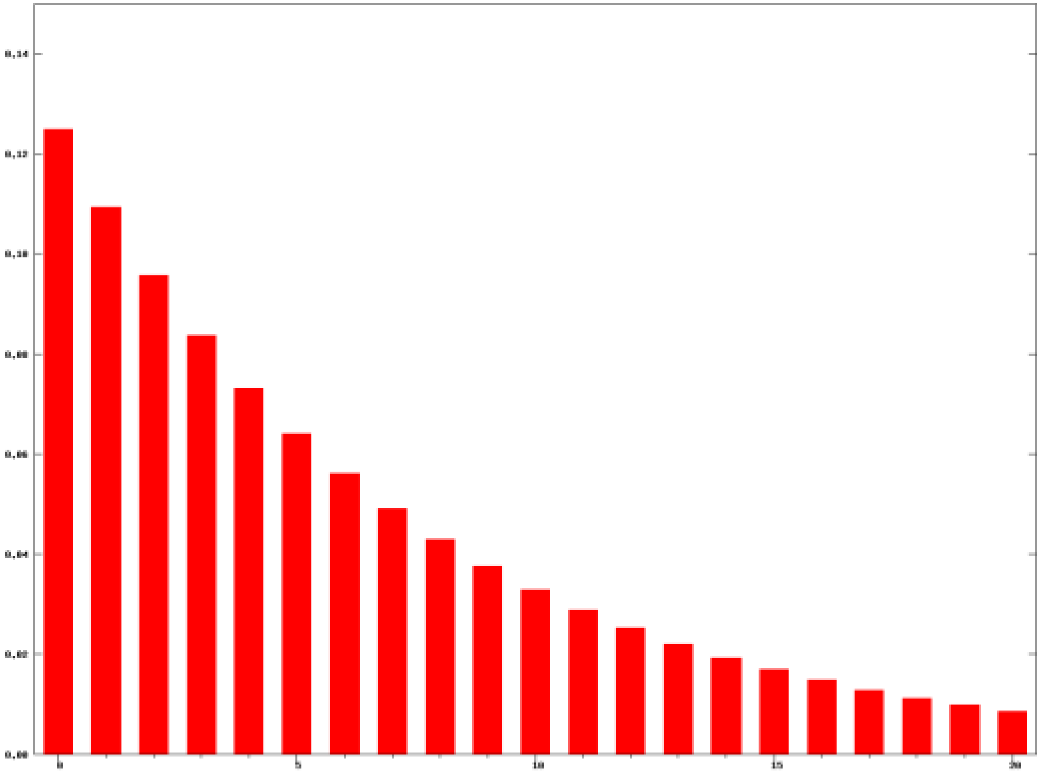
Exponential Weighted Moving Average (EWMA)

- 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

Autocorrelation

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

Q & A

Before Next Class

Before Next Class

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

Next Class

Time Series, Part 2

Learning Objectives

After the next lesson, you should be able to:

- Model and predict from time series data using AR, ARMA, or ARIMA models
- Specifically, coding these models in *statsmodels*

Exit Ticket

Don't forget to fill out your exit ticket here