UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3250 Foundations of Data Science

Module 8: Time Series and Forecasting with Pandas

Course Plan

Module Titles
Module 1 – Introduction to Data Science
Module 2 – Introduction to Python
Module 3 – NumPy
Module 4 – Pandas
Module 5 – Data Collection and Cleaning
Module 6 – Descriptive Statistics and Visualization
Module 7 – Workshop
Current Focus: Module 8 – Time Series
Module 9 – Introduction to Regression and Classification
Module 10 – Databases and SQL
Module 11 – Data Privacy and Security
Module 12 – Term Project Presentations (no content)

Learning Outcomes for this Module

- Develop familiarity with basic forecasting techniques and methods
- Understand how Pandas supports working with time series data
- Gain experience working with time series data in Pandas
- Practice downloading stock information and calculating returns

Topics for this Module

- **8.1** Time Series and Forecasting
- **8.2** Pandas for Time Series
- **8.3** Resources and Homework

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 8 – Section 1

Time Series and Forecasting

Time Series

- A ***time series*** is a set of observations taken at different points in time

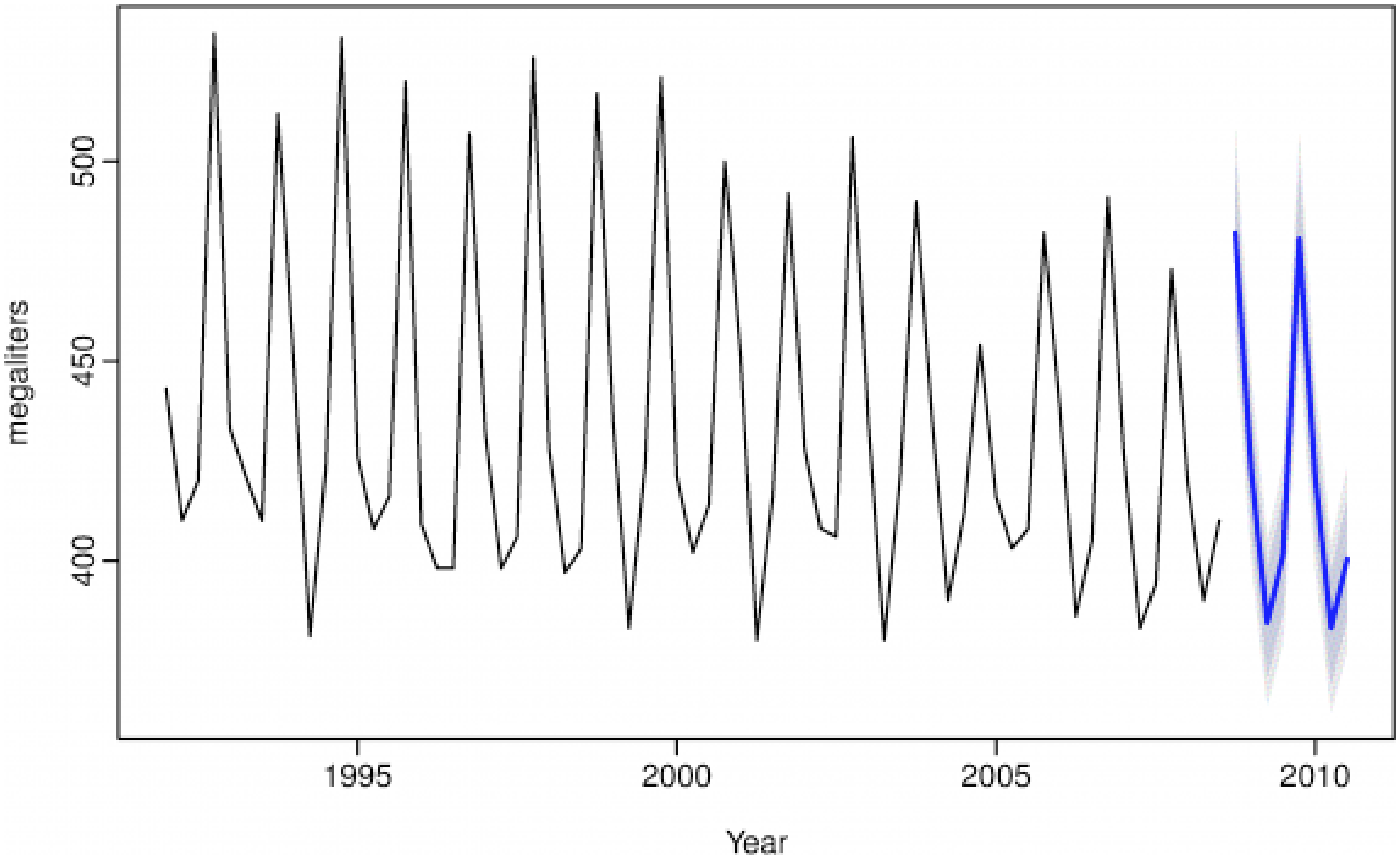
- Time series can be:
  - Fixed frequency
  - Irregular

- Particularly important in Finance and Economics

Forecasting

- Prediction where we have data sets that are in the form of a time series
- People who specialize in forecasting would call the kind of predictive models we've been talking about so far (not involving a time element) "cross-sectional forecasting"

Forecasting Time Series



What can be Forecast?

- Predictability depends on:
  - Our understanding of the predictive factors
  - The quantity of data available
  - The quality of data available
  - Whether taking measurements or making predictions that will influence the future outcomes

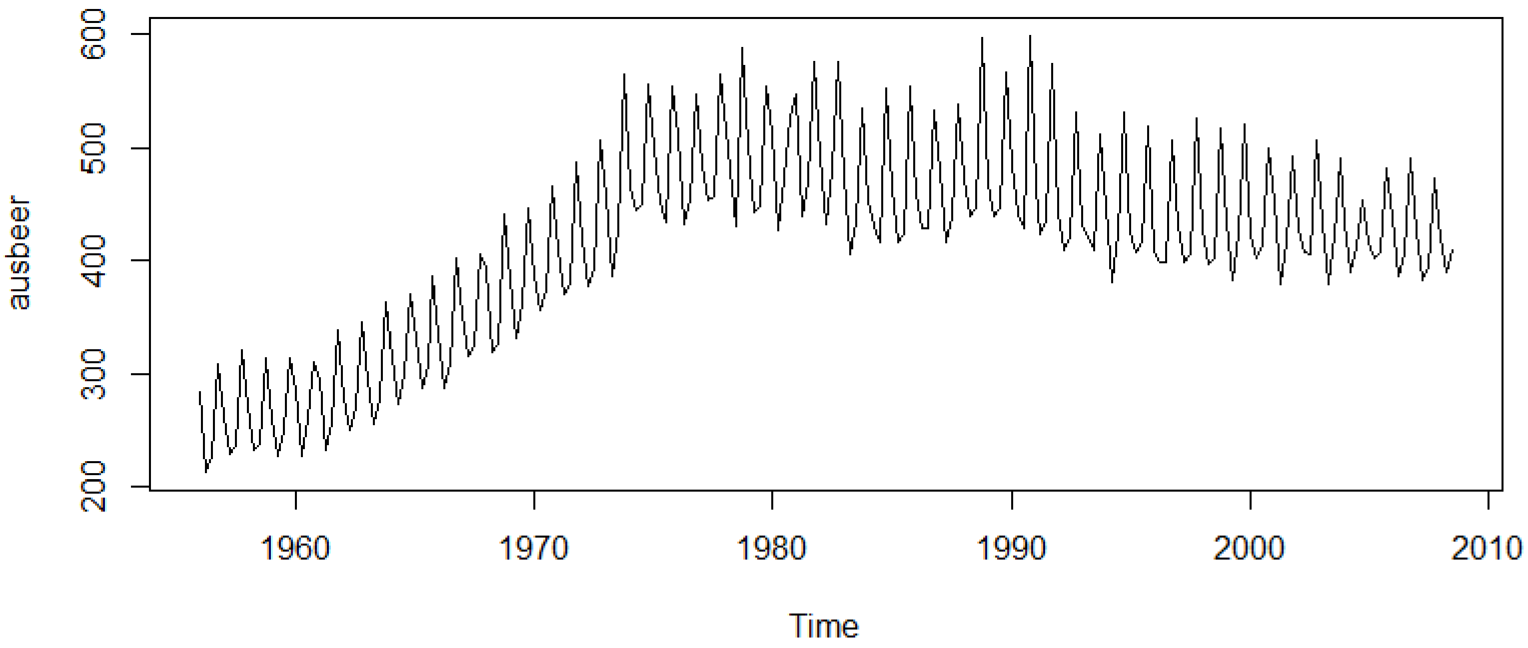
What’s Different with Time Series Prediction?

Prices						
Date	Open	High	Low	Close	Volume	Adj Close*
Oct 22, 2014	102.84	104.11	102.60	102.99	68,159,000	102.99
Oct 21, 2014	103.02	103.02	101.27	102.47	94,492,300	102.47
Oct 20, 2014	98.32	99.96	98.22	99.76	77,041,900	99.76
Oct 17, 2014	97.50	99.00	96.81	97.67	68,032,200	97.67
Oct 16, 2014	95.55	97.72	95.41	96.26	72,110,700	96.26
Oct 15, 2014	97.97	99.15	95.18	97.54	100,875,400	97.54
Oct 14, 2014	100.39	100.52	98.57	98.75	63,662,200	98.75
Oct 13, 2014	101.33	101.78	99.81	99.81	53,485,500	99.81
Oct 10, 2014	100.69	102.03	100.30	100.73	66,270,200	100.73
Oct 9, 2014	101.54	102.38	100.61	101.02	77,312,200	101.02
Oct 8, 2014	98.76	101.11	98.31	100.80	57,364,800	100.80
Oct 7, 2014	99.43	100.12	98.73	98.75	42,068,200	98.75
Oct 6, 2014	99.95	100.65	99.42	99.62	36,974,800	99.62
Oct 3, 2014	99.44	100.21	99.04	99.62	43,445,800	99.62
Oct 2, 2014	99.27	100.22	98.04	99.90	47,681,000	99.90
Oct 1, 2014	100.59	100.69	98.70	99.18	51,404,400	99.18

Exploratory Analysis

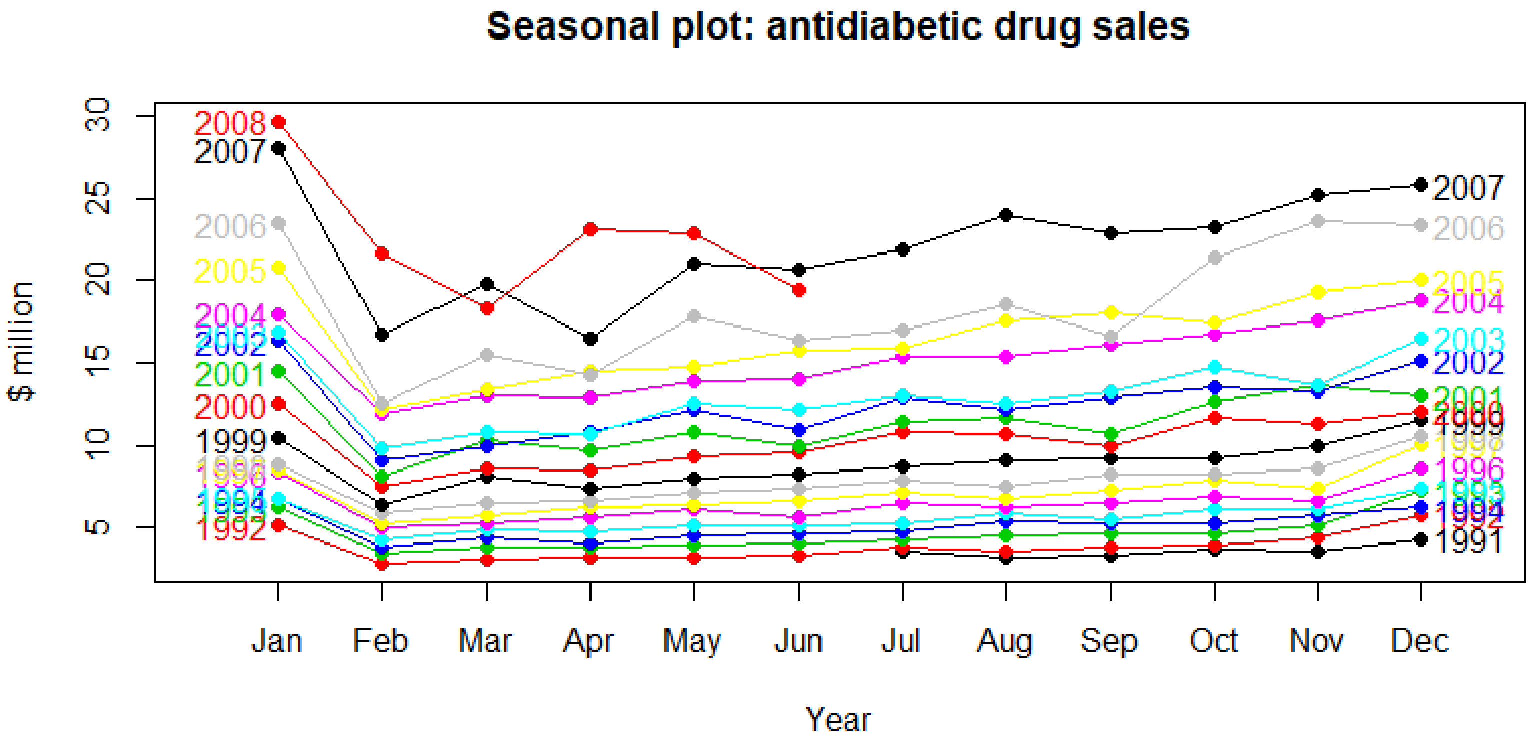
- Time Plot
- Seasonal Plot
- Lag Plot

Time Plot

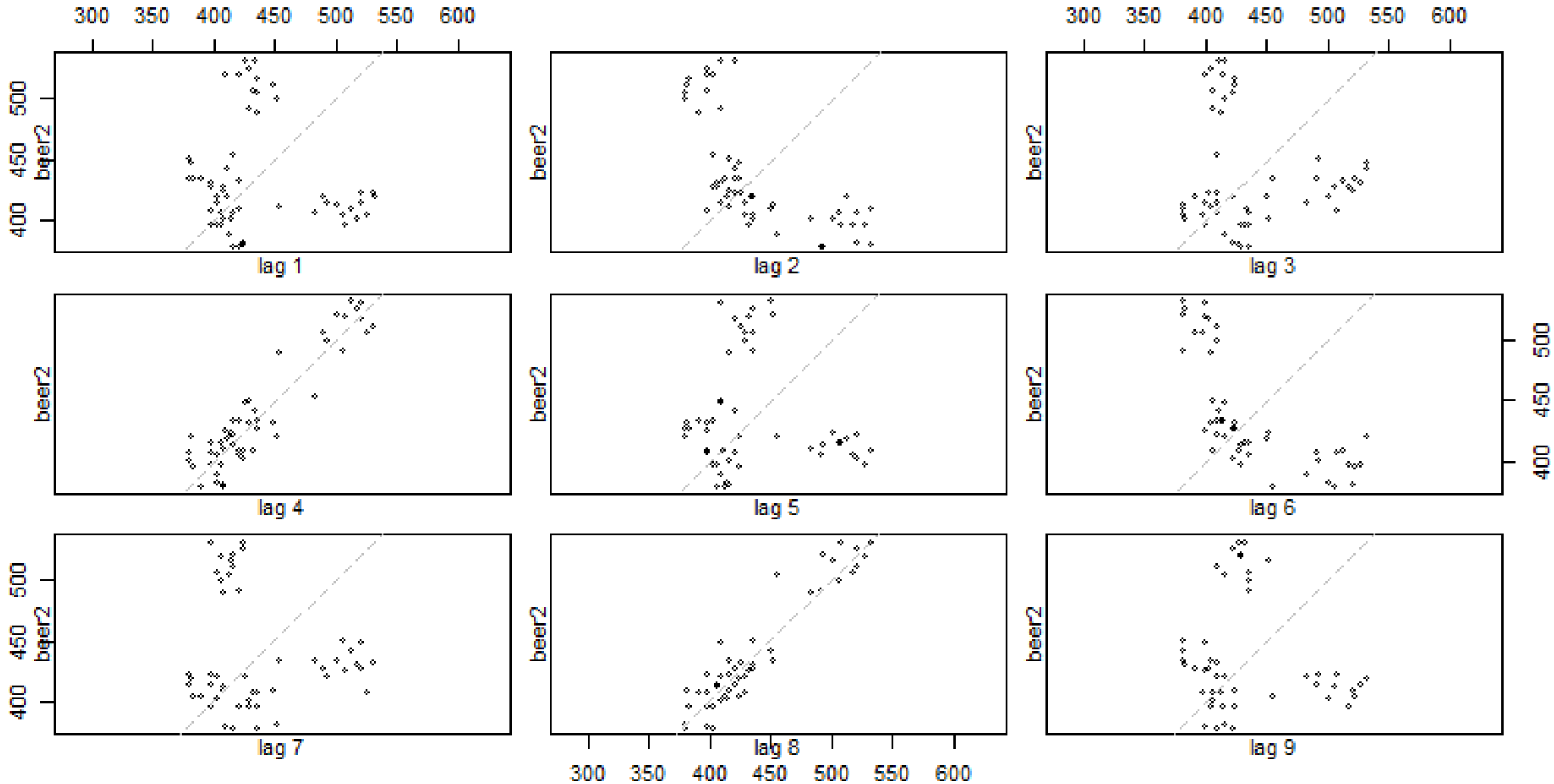


ausbeer
200
300
400
500
600
1960
1970
1980
1990
2000
2010
Time

Seasonal Plot



Lag Plot



300 350 400 450 500 550 600
400 450 500
beer2
lag 1
lag 2
lag 3
lag 4
lag 5
lag 6
lag 7
lag 8
lag 9

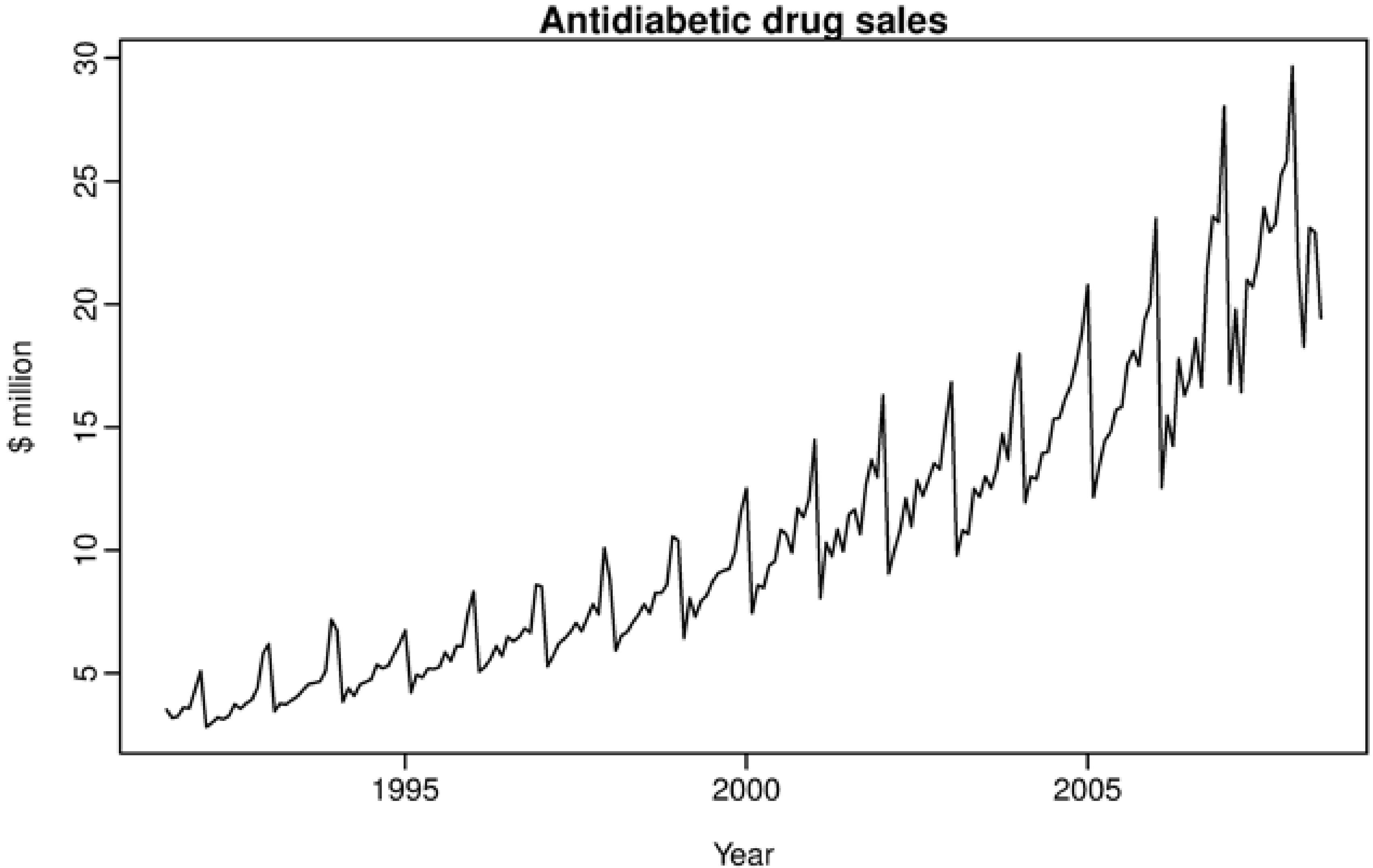
Typical Patterns in Time Series Data

- **Trends**: Long-term increase or decrease

- **Seasonality**: Where there is an influence that varies with the time of year or other calendar period

- **Cycles**: Patterns of repeated increase and decrease of *varying period*

Trend and Seasonality



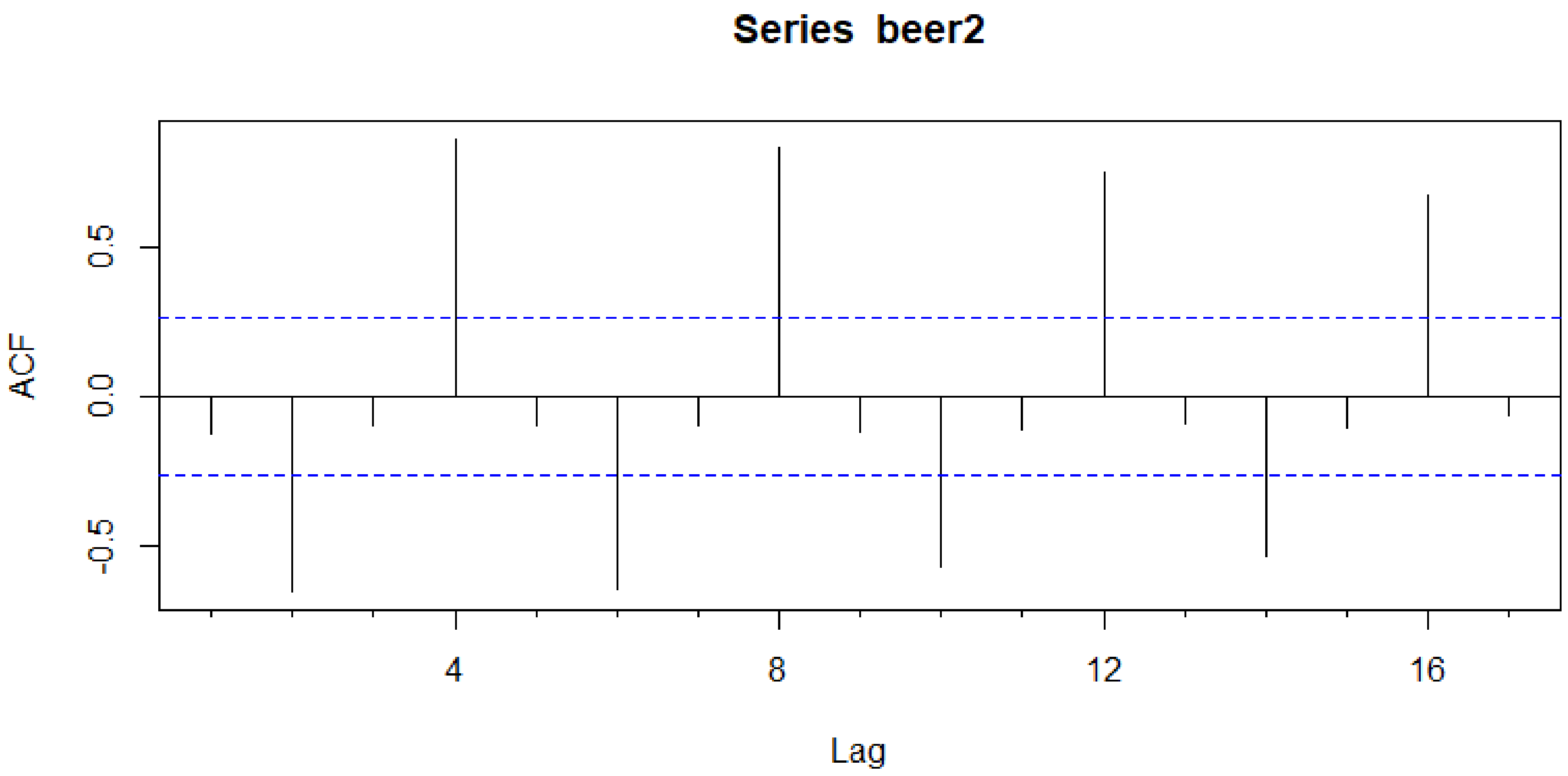
Stationarity

- Most time series methods make a simplifying assumption: that its statistical properties (mean, variance, growth rate) are not varying over time

Autocorrelation

- Correlation of a time series with lagged values of itself
- How much lag? Up to us: it’s a parameter

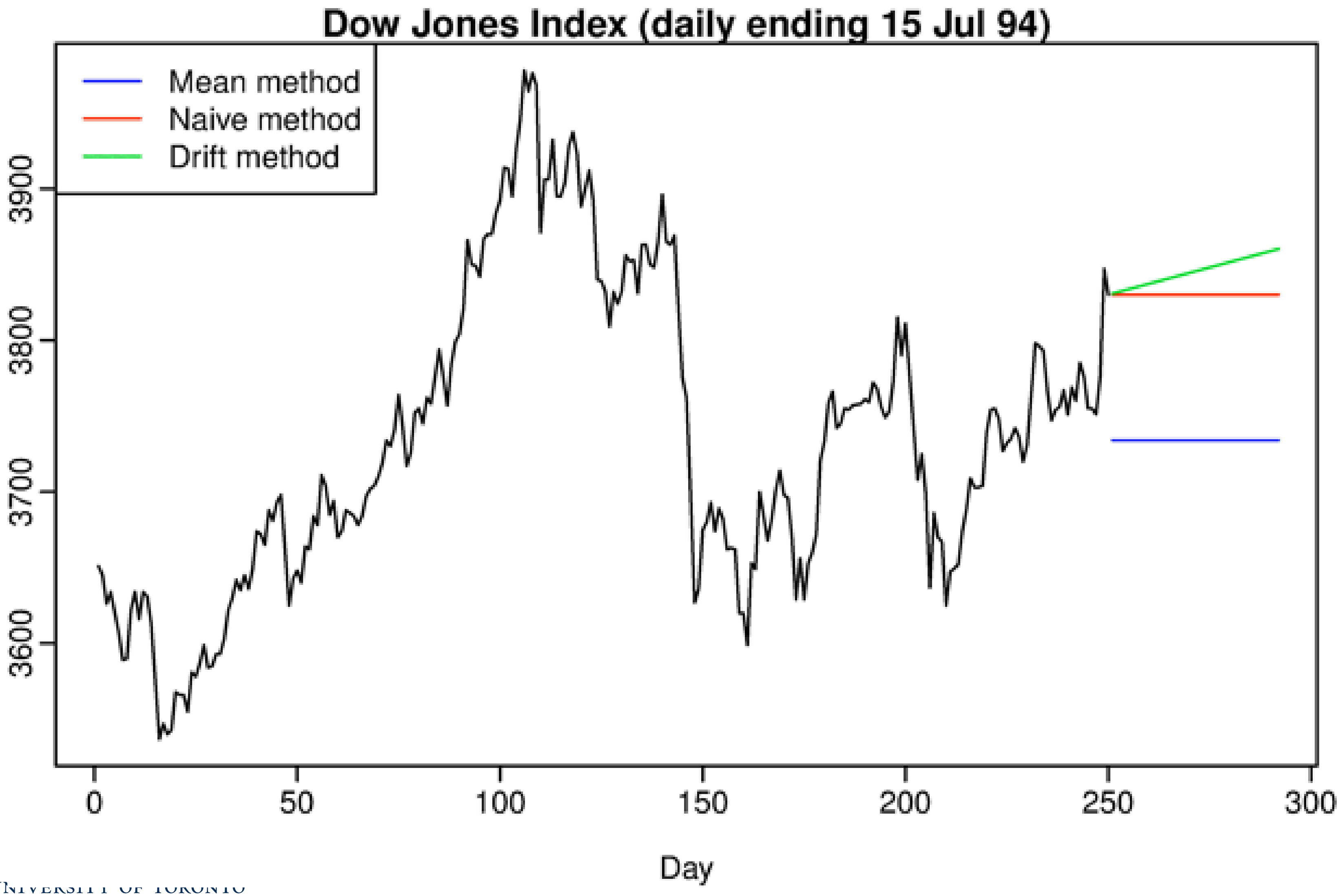
Autocorrelation Function



Some (Very) Simple Forecasting Methods

- **Average method**: Use average of data as forecast
- **Naïve method**: Use last data point as forecast
- **Seasonal naïve method**: Use data point from last corresponding season
- **Drift method**: Variation on naïve where we extrapolate the trend by drawing a line through the first and last observations

Forecast Methods Example



Common Transformations and Adjustments

- Use logarithms (or powers)
- Calendar adjustments
- Population adjustments
- Inflation adjustments

Model Evaluation

- Measuring error
- Training and test sets
- Cross-validation
- Overfitting

Model Evaluation (cont'd)

- A good forecasting model will have residuals that are:
  - Uncorrelated
  - Zero mean
- Better, but not necessary if they also:
  - Have constant variance
  - Are normally distributed

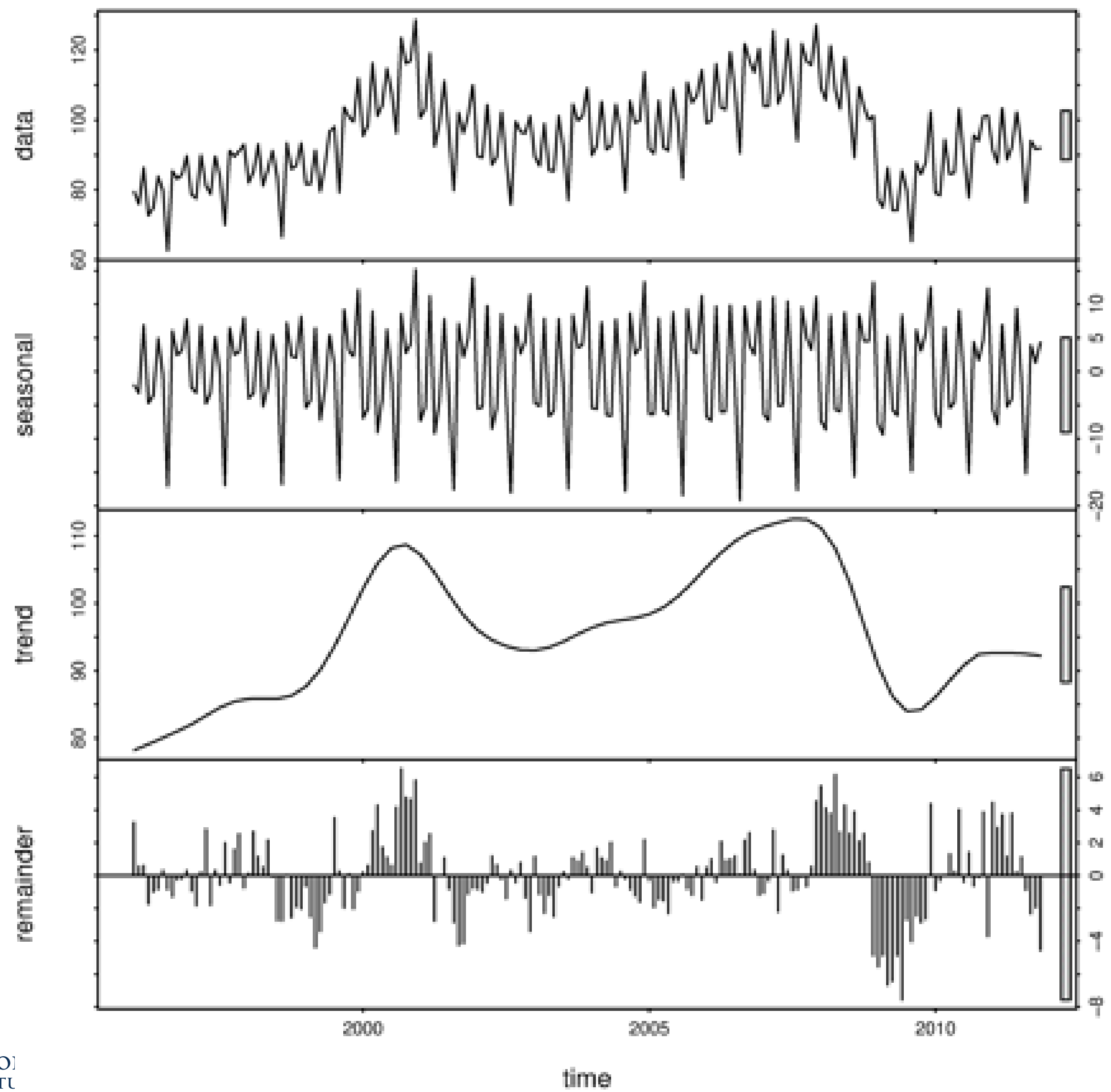
Regression-based Techniques

- Linear regression
  - To find trend line
- Multiple regression
  - Use dummy variables for seasons
  - Incorporate other predictors

Time Series Decomposition

- Time series can be decomposed into seasonal, trend-cycle and remainder components
- Additive, multiplicative and log-additive models are all common
- Moving averages
  - Smooth out variation to find non-linear trends
  - Can take moving averages of moving averages
  - Common to give recent observations higher weights

Time Series Decomposition (cont'd)



UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 8 – Section 2

Pandas for Time Series

Pandas Core Object Types

- Series
- DataFrame

The Time Dimension

- The time dimension in Pandas objects can be marked with:
  - Timestamps e.g. December 13, 2017 at 11:22 EST
  - Fixed periods e.g. monthly
  - Intervals e.g. 2015-04-03 03:12 to 2015-04-14 11:11
  - Elapsed time e.g. 45 mins. 32:05 secs.

Dates and Times in Python

- The main type in Python for dates and times is: `datetime`
- Stores time to the microsecond
- Can add or subtract times using a `timedelta` object
- Can convert back and forth between datetimes and strings

Series and Timeseries

- Most basic Pandas time series object is `Series`
- If a series is created where the index is made from a list of datetime objects, the Series will become a `Timeseries`
- Arithmetic between differently-indexed time series automatically align on the dates
- Indexing, selection, subsetting work the way we've seen for DataFrames
- Duplicate index timestamps are allowed

Fixed Frequency Data

- Generic time series in Pandas are assumed to be irregular
- Pandas has powerful capabilities for working with fixed frequency time series
- Use `.resample(period)` to convert an irregular time series to a fixed frequency one e.g. `ts.resample('D')`
- Newly created observations for times where there was no data will get values of NaN

Date and Time Ranges

- Use `pd.daterange()`
- Specify start and either end or number of periods
- Time ranges don’t exist as something independent of dates

Frequencies and Date Offsets

- Frequencies are expressed as a base frequency and a multiplier
- Base frequency identifiers have a lot of built-in knowledge about business calendars

Base Time Series Frequencies

Alias	Offset Type	Description
D	Day	Calendar daily
B	BusinessDay	Business daily
H	Hour	Hourly
M	Minute	Minutely
S	Second	Secondly
W-MON, W-TUE, etc.	Week	Weekly on given day of month
BQ-JAN BQ-FEB, etc.	BusinessQuarterEnd	

More complete list at PDA p. 300

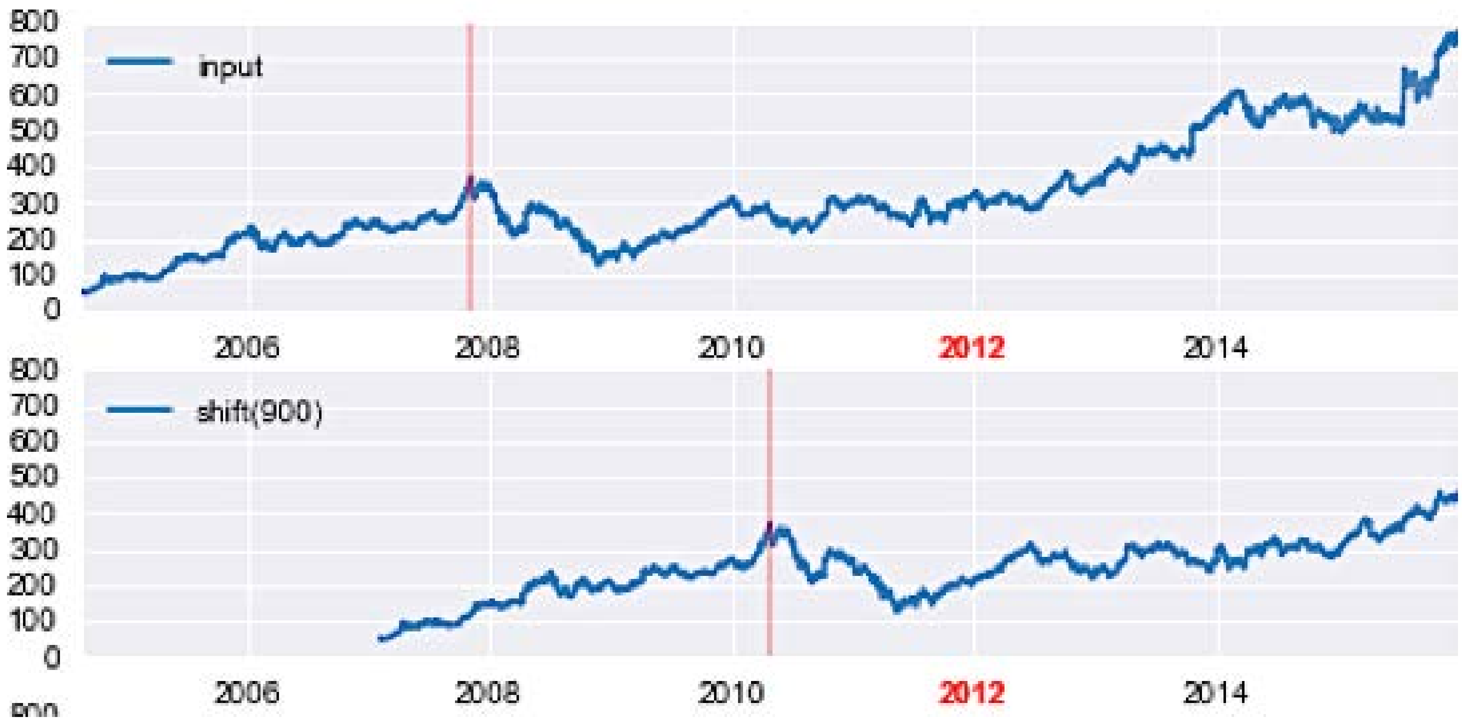
UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Shifting

- Both Series and DataFrame have a `.shift()` method to shift data without changing the index e.g.:
  `ts / ts.shift(1) – 1`

- Shift is specified in multiples of the frequency

Shifting (Cont'd)



- In this example, **`shift(900)`** shifts the data by 900 days, pushing some of it off the end of the graph (and leaving NA values at the other end)

Time Zone Handling

- Timestamps are usually in the form of UTC time plus an offset for the time zone
- This is a nightmare to work with directly
- Fortunately Pandas has access to a detailed database of world time zone information

Periods and Period Arithmetic

- Periods represent time spans

- Pandas has classes and methods for this:
```
Period(start, freq)
PeriodIndex(values, freq)
.period_range(start, end, freq)
.asfreq()
```

Resampling and Frequency Conversions

- **Resampling** is converting from one frequency to another
- Aggregating data from a high frequency to a lower one is called **downsampling**
- Converting from a lower frequency to a higher one is called **upsampling**

Time Series Plotting

- Pandas improves on Matplotlib’s date formatting

Moving Window Functions for Series

- Number of non-NA observations in a window: `rolling_count`
- Moving window sum: `rolling_sum`
- Moving window average: `Series.rolling(window=250, center=False).mean()`
- Moving window correlation: `Series.rolling(min_periods=100, window=125).corr(other=<Series>)`
- Apply function to a window: `Series.rolling(center=False,window=250).apply(args=<tuple>,func=<function>,kwargs=<dict>)`
- etc.

Module 8 – Section 3

Resources and Homework

Resources

- Hyndman & Athanasopoulos. Forecasting Principles and Practices. OTexts. 2013.
- Complete Time Series Modeling Tutorial
- Shumway & Stoffer. Time Series and Its Applications. Free Texts in Statistics.

Resources (cont'd)

- Bayesian causal impact analysis in time series (CausalImpact package in R and paper):
- Online course in quantitative economics:

Resources (cont'd)

- Blog on algorithmic trading using free and open source software:
- Autocorrelation Plot
- Hilpisch, Yves. Python for Finance. O'Reilly. 2014.

Time Series Assignment

1. In a command window: `conda install pandas-datareader`
2. Download the adjusted close price for AAPL, BBRY, LULU and AMZN using the following code:
   ```
   import pandas_datareader.data as web
   import datetime
   start = datetime.datetime(2012, 7, 31)
   end = datetime.datetime(2017, 6, 30)
   aapl = web.DataReader(‘WIKI/AAPL', ‘quandl', start, end)
   ```
3. Get the data for the last 60 months, select the adjusted monthend close for each.
4. Use pandas autocorrelation_plot to plot the autocorrelation of the adjusted monthend close of each of the stocks. Are they autocorrelated? Why or why not?

Time Series Assignment (cont'd)

5. Calculate the monthly return over the period for each stock using the "shift trick" on the lecture slide titled *Shifting* (Note: you should end up with a time series 59 months long)
6. Use pandas autocorrelation_plot to plot the autocorrelation of the monthly returns. Are they autocorrelated? Why or why not?
7. OPTIONAL: Visualize the correlation between the returns of all pairs of stocks using a scatterplot matrix (1 bonus mark)
8. OPTIONAL: Following the instructions in the Glowing Python blog visualize the correlation of the returns of all pairs of stocks (2 bonus marks)

Optional Homework Exercises

- Homework, ch11.ipynb

Optional Homework Exercises (cont'd)

Exercise with Matching Homework Video:

- Exercise
- Note:
  - !head will only work on Linux
  - GOOG is now GOOGL

Next Class

- Introduction to Regression and Classification

Follow us on social

Join the conversation with us online:

facebook.com/uoftscs

@uoftscs

linkedin.com/company/university-of-toronto-school-of-continuing-studies

@uoftscs

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Any questions?

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Thank You

Thank you for choosing the University of Toronto School of Continuing Studies