DSC 201: Data Analysis & Visualization

Exploratory Data Analysis

Dr. David Koop
Python Support for Time

• The `datetime` package
  - Has date, time, and datetime classes
  - `now()` method: the current datetime
  - Can access properties of the time (year, month, seconds, etc.)

• Converting from strings to datetimes:
  - `datetime.strptime`: good for known formats
  - `dateutil.parser.parse`: good for unknown formats

• Converting to strings
  - `str(dt)` or `dt.strftime(<format>)`

• Differences between times
  - `datetime.timedelta`: can get number of days/hours/etc.
  - Deal with issues with different length months, etc.
Time Zones

• Why?

• Coordinated Universal Time (UTC) is the standard time (basically equivalent to Greenwich Mean Time (GMT))

• Other time zones are UTC +/- a number in [1,12]

• Dartmouth is UTC-5 (aka US/Eastern)
Python, Pandas, and Time Zones

• Time series in pandas are **time zone native**
• The pytz module keeps track of all of the time zone parameters  
  - even Daylight Savings Time
• Localize a timestamp using `tz_localize`
  
  - `ts = pd.Timestamp("1 Dec 2016 12:30 PM")`
  
  `ts = ts.tz_localize("US/Eastern")`
• Convert a timestamp using `tz_convert`
  
  - `ts.tz_convert("Europe/Budapest")`
• Operations involving timestamps from different time zones become UTC
Resampling

• resample method
• Can downsample or upsample data
• If downsample, need to provide method for combining data (mean/sum)
• If upsample, need to fill in missing data
• Interpolation is useful here
Downsampling

- `rng = pd.date_range('1/1/2010', periods=100, freq='D')`
- `ts = pd.Series(np.random.randn(len(rng)), index=rng)`

<table>
<thead>
<tr>
<th>Date</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-01-01</td>
<td>0.649916</td>
</tr>
<tr>
<td>2010-01-02</td>
<td>-0.856086</td>
</tr>
<tr>
<td>2010-01-03</td>
<td>0.225041</td>
</tr>
<tr>
<td>2010-01-04</td>
<td>-0.510606</td>
</tr>
</tbody>
</table>

- `ts.resample('M').mean()`

<table>
<thead>
<tr>
<th>Date</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-01-31</td>
<td>-0.152096</td>
</tr>
<tr>
<td>2010-02-28</td>
<td>0.247770</td>
</tr>
<tr>
<td>2010-03-31</td>
<td>-0.008501</td>
</tr>
<tr>
<td>2010-04-30</td>
<td>0.432483</td>
</tr>
</tbody>
</table>

Freq: M, dtype: float64
Upsampling

- `ts.resample('12h').mean()`
  
  - 2010-01-01 00:00:00    0.649916
  - 2010-01-01 12:00:00    NaN
  - 2010-01-02 00:00:00    -0.856086
  - 2010-01-02 12:00:00    NaN
  - 2010-01-03 00:00:00    0.225041

- Now, we need to fill in missing data (NaN)

- Options: `fillna()`, `ffill()`, `bfill()`

- Interpolation!

- `ts.resample('12h').sum().interpolate()`
  
  - 2010-01-01 00:00:00    0.649916
  - 2010-01-01 12:00:00    -0.103085
  - 2010-01-02 00:00:00    -0.856086
  - 2010-01-02 12:00:00    -0.315523
  - 2010-01-03 00:00:00    0.225041
Window Functions

• Idea: want to aggregate over a window of time, calculate the answer, and then slide that window ahead. Repeat.

• rolling: smooth out data

• In old versions of pandas (like the book uses), this used to be rolling_count, rolling_sum, rolling_mean

• Specify the window size, then an aggregation method

• Can also specify the window

• Result is set to the right edge of window (change with center=True)
Food Inspections Example

• Questions:
  - When was a location was last inspected?
  - How many inspections are done per month/year/etc.?
  - Is there any trend in the number of inspections over time?

• Notebook posted on course web page
Assignment 5

- [http://www.cis.umassd.edu/~dkoop/dsc201/assignment5.html](http://www.cis.umassd.edu/~dkoop/dsc201/assignment5.html)
- Aggregation, resampling, and visualization of time series data
- Part 3: Multi-level index and groupbys
What is Exploratory Data Analysis?

• "Detective work" to summarize and explore datasets
• Includes:
  - Data acquisition and input
  - Data cleaning and wrangling ("tidying")
  - Data transformation and summarization
  - Data visualization
Exploratory Data Analysis

- John W. Tukey
  - Born in New Bedford
  - 1977: Highly influential book
- Emphasis on value of visualization in discovering trends, relationships
- From a review of the book:
  “Tukey favors analysis of data with little more than pencil and paper. Specifically, there is no need for a calculator, a computer, or a lettering guide to do the analyses he proposes” [R.M. Church, 1979]
“Reflective thought requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. The process takes time.”

– Donald A. Norman
Comparison with Older Statistical Methods

• Older method:
  - Take data and a probable model and learn model parameters
  - Good models are useful, help understand phenomena
  - What happens when we pick the wrong model, or don't know which one to pick?

• EDA:
  - Postpone the model assumptions, let the data speak first
  - Usually involves graphical techniques
  - Tukey used pen-and-paper approaches
  - Today, we can do much of this via computer, but insight may still take time
Types of EDA

- Univariate vs. multivariate
- Non-graphical vs. graphical
Univariate Non-graphical EDA

- Categorical Data:

- Quantitative Data:
Univariate Non-graphical EDA

• Categorical Data:
  - Frequency counts, proportions
  - Groupings

• Quantitative Data:
  - Distribution
  - Summary statistics: mean, median, mode, variance, standard deviation, quantiles
Univariate Graphical EDA
Univariate Graphical EDA

- **Histograms**
  - Aggregation of data
  - Choose number of bins
  - Bin width makes a difference!

- **Stem-and-leaf plots:**
  - 1|710340
  - 2|06
  - 3|030223459
  - 4|028907
  - 5|00798273487
  - 6|128
  - 7|7897
  - 8|345
  - 9|1
Univariate Graphical EDA

- **Boxplots**
  - Show distribution
  - Multiple summary statistics can be read from the chart
  - Also provides a general shape of the data
  - Best for *unimodal* data
Multivariate non-graphical EDA
Multivariate non-graphical EDA

- Crosstabs and Pivot Tables
  - What is in the data? Count
- Correlation and covariance
  - How related are different columns?
  - How do they change together?
Multivariate graphical EDA
Multivariate Graphical EDA

- Scatterplots: look for correlation
  - Usually put outcome on y-axis
  - Can encode other variables
- Side-by-side boxplots
- Parallel coordinates
- Grouped bar charts