SI330: LECTURE 5

PANDAS PART I

Overview of Today's Class

- **Welcome**
- Announcements
- Questions, Concerns and Comments
- Pandas: building on the readings (!)

Course Roadmap

Week	Topics
1	Course introduction & review of python basics
2	Basic and compound data structures
3	Extracting patterns from text with regular expressions
4	Natural Language Processing
5	JSON, APIs, AWS
6	Pandas I
7	Pandas II
8	Pandas III
9	Break
10	SQL

Amazon Web Services

- AWS Educate finally approved for me: 8 days!
- Questions?
- Concerns?
- Homework?

PANDAS: SIMPLIFYING DATA ANALYSIS IN PYTHON

- high-level library to support data manipulation and analysis
- DataFrame is the primary object we'll be dealing with
  - similar to R's dataframe
  - maps onto tabular structure
- good for time series and econometric data

How to cope with the "Readings"

- don't read them like a novel
- try the examples
- try changing the examples

From "Pythonic" to "Pandorable"

Pandas: Building on NumPy: Series and DataFrames

list

```
names=['September','Jarod','Digna','Minda','Lan']
```

```
enthusiasmScores = [3,5,4,4,5]
```

numpy.ndarray

September
Jarod
Digna
Minda
Lan

3
5
4
4
5

panda.Series

September	3
Jarod	5
Digna	4
Minda	4
Lan	5

panda.DataFrame

	enthusiasm	worry
September	3	4
Jarod	5	2
Digna	4	4
Minda	4	4
Lan	5	3

Pandas: Building on NumPy: Series and DataFrames

list

```
names=['September','Jarod','Digna','Minda','Lan']
```

```
enthusiasmScores = [3,5,4,4,5]
```

numpy.ndarray

September
Jarod
Digna
Minda
Lan

3
5
4
4
5

panda.Series

September	3
Jarod	5
Digna	4
Minda	4
Lan	5

panda.DataFrame

	enthusiasm	worry
September	3	4
Jarod	5	2
Digna	4	4
Minda	4	4
Lan	5	3

NUMPY.NDARRAY

- for our purposes:
  - array == numpy.ndarray == ndarray == “NumPy array”
- underlying type for panda Series (and therefore DataFrame)
- consists of elements of the same dtypes (e.g. int8, int16, int64, float32, bool, object, string_, etc.)

Why do we care about ndarray?

- mostly because they support Universal Functions (ufuncs)
- ufuncs perform elementwise operations on ndarrays
- e.g. sqrt, maximum, see McKinney p. 96 for complete list, also p. 101 for list of statistical functions

Pandas: Building on NumPy: Series and DataFrames

list

```
names=['September','Jarod','Digna','Minda','Lan']
```

```
enthusiasmScores = [3,5,4,4,5]
```

numpy.ndarray

September
Jarod
Digna
Minda
Lan

3
5
4
4
5

panda.Series

September	3
Jarod	5
Digna	4
Minda	4
Lan	5

panda.DataFrame

	enthusiasm	worry
September	3	4
Jarod	5	2
Digna	4	4
Minda	4	4
Lan	5	3

Pandas: Building on NumPy: Series and DataFrames

list

```
names=['September','Jarod','Digna','Minda','Lan']
```

```
enthusiasmScores = [3,5,4,4,5]
```

numpy.ndarray

September
Jarod
Digna
Minda
Lan

3
5
4
4
5

panda.Series

September	3
Jarod	5
Digna	4
Minda	4
Lan	5

panda.DataFrame

	enthusiasm	worry
September	3	4
Jarod	5	2
Digna	4	4
Minda	4	4
Lan	5	3

DataFrame: our pandas workhorse

- “tabular, spreadsheet-like data structure” (McKinney, p 115)
- think of it like a collection of Series aligned on the same index
- has indexes on both rows (called the “index”) and on columns (called the “columns”)

To the Jupyter Notebooks for today's lecture...